UNIVERSITY OF CALIFORNIA
SANTA CRUZ

NOVEL GENE EXPRESSION ANALYSES TO ACCELERATE PRECISION PEDIATRIC ONCOLOGY RESEARCH

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Jacob J. Pfeil

March 2020

The Dissertation of Jacob J. Pfeil
is approved:

Professor Richard E Green, Chair

Professor David Haussler

Professor Joshua Stuart

Professor Olena Morozova Vaske

Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

Abstract

Novel Gene Expression Analyses to Accelerate Precision Pediatric Oncology Research

by

Jacob J. Pfeil

Cancer is the second leading cause of death in the United States. While there have been medical advances in treating cancer, the standard of care has not changed significantly in recent decades. Chemotherapy, radiation, and surgery are the clinician's first line of defense against cancer progression, but new therapeutic strategies such as precision oncology are being developed that personalize cancer therapy to individuals. Precision oncology has primarily relied on coding mutations as biomarkers of response to therapies. Numerous challenges have arisen in the incorporation of transcriptome analysis into precision oncology workflows. One such challenge is in the necessary consideration of relative rather than absolute gene expression level, requiring differential expression analysis across samples. However, expression programs related to the cell-of-origin and tumor microenvironment effects confound the search for cancer-specific expression changes. To address these challenges, we developed an unsupervised clustering approach for discovering differential pathway expression within cancer cohorts using gene expression measurements. The hydra approach uses a Dirichlet process mixture model to automatically detect multimodally distributed genes and expression signatures. This led to the identification of recurrent tumor microenvironment signatures across pediatric cancers as well as a relationship between transposable element expression and immune infiltration.

I then developed the vaccinaTE software toolkit to further characterize transposable elements as potential immunotherapy targets. Using RNA-seq and mass spectrometry analysis, I found expression and MHC-bound peptides uniquely mapping to transposable element loci. This led to the creation of a novel process for prioritizing TE vaccine targets as well as a microarray technology for personalizing TE vaccine therapy. To address the need for accurate preclinical models to accelerate drug development for pediatric cancers, I then created a Bayesian hierarchical modeling framework for evaluating patient-derived xenografts. I generated a database of PDX-specific pathway expression to facilitate validation studies that attempt to target differentially expressed pathways. This thesis has sought to improve the treatment of pediatric cancers through the identification of tumor subtypes that respond to specific therapies, identify novel immunotherapy targets based on tumor microenvironment states, and use gene expression analysis to optimize preclinical validation experiments. These methods have been developed for pediatric cancers, but can be modified for adult cancers as well as other diseases for which gene expression data is available.

I dedicate my dissertation to my parents, Brian & Nancy Pfeil, who instilled in me a reverence for life that fueled this work. I would also like to thank my wife Alison Roozeboom for supporting me throughout this endeavor and encouraging me to follow my research interests. I also want to thank my son Rowan Pfeil for motivating me to complete my dissertation.

Acknowledgments

I would like to acknowledge the following people for encouraging me and mentoring me throughout my graduate training:

- Lauren Sanders
- Alana Weinstein
- Geoff Lyle
- Jason Fernandes
- Sofie Salama
- Holly Beale
- Chris Vollmers
- Nik Sgourakis
- Isabel Bjork
- Olena Vaske
- Alejandro Sweet-Cordero
- Ed Green
- Josh Stuart
- David Haussler

Part I

UCSC Treehouse Gene Expression Analysis

Childhood cancer patients need therapies that cure disease while also safeguarding development and future health. Approximately, 16,000 children are diagnosed with cancer each year in the United States. Despite significant improvements in childhood cancer therapies, one in eight children will die of cancer. Some forms of childhood cancer respond better to standard of care therapies than others (Figure 0.1). There are forms of pediatric brain tumors that have survival rates around ~10 %.

The standard of care therapies are also harmful to the long-term health of childhood cancer survivors. For instance, children respond well to high-dose chemotherapy, but chemotherapeutic agents are toxic and damage healthy tissue. Life-long side effects develop in ~60 % of the childhood cancer survivors. Childhood cancer survivors are more likely to develop other forms of cancer, heart and lung problems, stunted growth, and learning disabilities [1, 23, 39]. There are ~380,000 childhood cancer survivors in the United States and 60% of them are facing life-long disabilities as a result of their cancer therapy.

A more personalized approach may overcome the shortcomings of current standard of care therapies. Molecularly targeted therapies identify rare alterations within a patient's cancer that can be specifically inhibited to prevent cancer progression. Targeted therapies are biologically active at a lower dose than many standard of care therapies which makes them less toxic. While targeted therapies have induced tumor remissions, cancer cells are prone to become resistant to targeted therapies and the cancer returns. Research into the molecular mechanisms of drug resistance as well as development of more pediatric targeted inhibitors may yield novel therapeutic directions that yield better outcomes for patients with fewer harmful side effects [32].

Cancer cells divide at an uncontrolled rate and rely on DNA replication to sustain growth. One of the first applications of chemotherapy targeted DNA replication. During the 1950s, the pediatric oncologist Sydney Farber was experimenting with folic acid (vitamin B) as a potential cancer therapy. Folic acid is an important starting material for synthesizing DNA and RNA, and leukemia cells use folic acid to proliferate. When exploring folic acid analogues, Sydney Farber stumbled upon the folic acid antagonist amethopterin. Amethopterin works by inhibiting the cell's ability to use folic acid. Sydney Farber used amethopterin to induce remissions in childhood leukemia. This was the first successful application of chemical therapies for pediatric cancer and amethopterin remains part of the standard of care for childhood leukemia [31, 5].

A more targeted approach identifies specific molecular alterations that make cancer cells susceptible to targeted therapies. An example of a successful targeted therapy is imatinib (Gleevec) for BCR-ABL driven leukemia. BCR-ABL is a fusion protein that couples the oncogenic ABL1 gene with a constitutively expressed BCR gene. This increases the concentration of the oncogenic ABL1 gene to drive cancer progression. Imatinib can correct for this alteration by binding to the ABL active site and preventing ABL's biological function. BCR-ABL positive cancer cells depend on the ABL protein to proliferate, so inhibition of ABL's function halts cancer progression. The BCR-ABL fusion occurs in a fraction of leukemia patients, but application of imatinib to BCR-ABL positive leukemias has been proven to improve treatment outcomes [44, 2].

Each patient's cancer evolves from a single cell that gradually accumulated cancer features [38]. Cancer cells evolve through Darwinian selection such that the cancer cell pop-

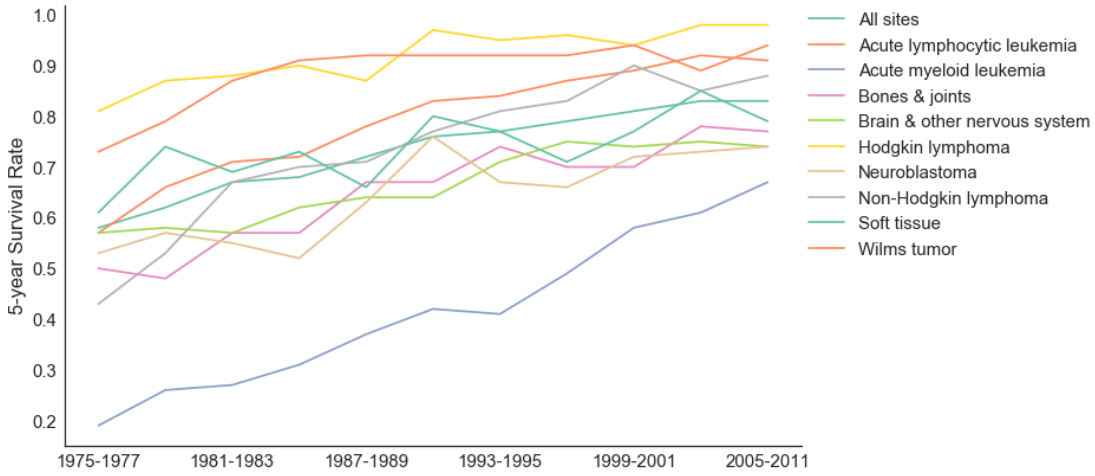


Figure 0.1: Pediatric 5-year survival rates (Birth to 14 years) collected by the National Cancer Institute SEER program [36].

ulation adapts to survive. This process is random and depends on the patient's genetics and environment. Therefore, each cancer is unique and requires a personalized approach to identifying drug targets. The current treatment paradigm uses a one-size-fits-all approach that may not be appropriate for some cancer patients. A personalized approach learns the molecular features of each patient and identifies potential drug targets. The personalized approach maximizes the use of effective therapies and improves treatment outcomes.

Gene Expression Analysis for Pediatric Cancer

Gene expression analysis is a relatively new approach for identifying drug targets for cancer. Testing for specific DNA-level alterations has been better developed and has become routine for several forms of cancer. Private diagnostics companies like Foundation Medicine and Quest Diagnostics routinely test for genetic variants and report findings to clinicians. Clinicians use the genetic testing results to direct the treatment of their patients. Insurance companies

will often cover the cost for genetic testing if there is an actionable variant for the disease. For example, women with a family history of breast cancer can test for mutations in the BRCA1 or BRCA2 DNA repair genes. Women with pathogenic BRCA mutations have a higher probability of developing breast cancer. Identification of BRCA mutations also predicts sensitivity to a targeted inhibitor. Mutations that disable BRCA genes correlate with sensitivity to PARP inhibition. Cancer cells are often sensitive to loss of PARP and BRCA proteins [8]. Clinical genetic testing is an important tool for identifying patients who could benefit from targeted therapies.

Genetic testing has proven clinical utility, but genetic testing depends on well characterized variants. However, many patients who have genetic testing done receive a variant prediction of unknown significance. Variants of unknown significance do not benefit the patient's treatment, and pediatric cancer has fewer somatic mutations overall [43, 46, 15].

Pediatric cancer may also have a strong epigenetic component that cannot be detected with genetic testing. Epigenetics consists of regulatory mechanisms that control gene expression. Examples of epigenetic modifications include DNA methylation and histone post-translational modifications. While epigenetic modifications are not as long-lasting as DNA-level alterations, epigenetic modifications are inheritable and can promote tumorigenesis [7]. An example of a pediatric genetic variant that has epigenetic implications is a recurrent mutation in the histone tails of diffuse intrinsic pontine glioma patients. A recurrent histone H3 tail mutation occurs for ~80% of diffuse intrinsic pontine glioma patients [17]. Histone H3 at Lysine 27 is substituted for a methionine. The lysine residue can be post-translationally modified through methylation. Methylation of lysine 27 turns off the expression of neighboring genes. The methionine substitution prevents methylation and leads to over-expression of some genes which is

implicated in pediatric brain tumors.

Genetic variants and epigenetic modifications influence the expression of other genes which can be measured by genome-wide gene expression profiling. Gene expression analysis can therefore be used to identify the combined effects of genetic and epigenetic alterations in cancer. Gene expression analysis has been used to identify cancer biomarkers and predict drug sensitivity [4]. Gene expression analysis has not been validated for clinical applications, but medical research institutions are currently developing these tools to identify drug targets for cancer.

Overcoming Barriers to Genomic Medicine Approaches

Genomics is the study of the structure and function of all coding and non-coding elements in the genome. DNA sequencing technology is used to study the genome sequence, patterns of gene expression, and genome-wide epigenetic modifications. These methods are now being translated into clinical tools and used in medicine to inform clinical decisions. Genomic medicine has a lot of potential, but several challenges are currently being addressed to facilitate wide-spread adoption of genomic approaches.

One of the major milestones of genomic research was the completion of the Human Genome Project which generated the first draft of the human genome. The Human Genome Project initiated a new era of biological research with the hope of reinventing medicine and providing new cures for human disease. The United States has committed to supporting genomic medicine research since the Human Genome Project. In 2015, President Obama announced additional precision medicine funding through the Precision Medicine Initiative. The Precision

Medicine Initiative provides $216 million for developing genomic approaches for cancer research. One of the goals of the Precision Medicine Initiative is to collect genomic data for at least one million US citizens.

In addition to public support, technological advances have also paved the way for genomic medicine. Innovations in DNA sequencing technology have lowered the cost to allow for routine sequencing. Massively parallel DNA sequencing technology breaks the genome into small fragments and uses fluorescence chemistry to discern the nucleic acid sequence. Rapid development of DNA sequencing chemistry has driven the cost down, but the clinical utility of DNA sequencing approaches needs to be proven. Several FDA approved drugs are linked to a genomic alteration, which supports the utility of genomic methods, but DNA sequencing methods must be tested through regulatory channels and approved for clinical testing [28]. Medical research institutions are validating and integrating DNA sequencing technology into pathology departments.

The human genome contains three billion base pairs and 20,000 protein coding genes. Genomic data is high-dimensional and requires many samples and sophisticated computational resources to process and learn from the data. While some batch effects exist, genomic data can be shared to increase the power of statistical analyses. The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) are large cancer genome sequencing projects that generated high-quality sequencing data for cancer researchers.

Genomic medicine requires computational infrastructure to analyze large high-dimensional data sets. Fortunately, the cloud computing market gives medical institutions the flexibility to

scale computation to community needs. The major cloud computing companies are Amazon, Microsoft, and Google, and many Silicon Valley companies are taking part in the growing genomic healthcare market. Although most hospitals would benefit from applying precision medicine techniques, few have the resources to support a computer cluster. The cloud computing market opens scientific computing to the general public. Cloud computing also creates a market for bioinformatic application development. Clinicians will soon be able to select an appropriate analysis, upload patient data, and download the results without being familiar with the technical challenges of running bioinformatics software on a computer cluster. Research into cloud-based bioinformatics tools will therefore facilitate adoption of precision medicine approaches.

Tools For Gene Expression Analysis

Gene expression analysis measures the abundance of all coding and non-coding RNA transcripts in a biological sample. The pattern of gene expression describes the transcriptional activity of cells at the time the sample was collected. Cancer tissue is a complex mixture of cancer cells, infiltrating immune cells, stromal cells, and cells from healthy tissue. Extracting RNA from a heterogeneous tissue sample effectively averages gene expression estimates across all cell types in the sample. The average is skewed towards the most common cell population and cells expressing the most RNA transcripts.

The two main technologies for studying transcript abundance are gene expression microarrays and RNA sequencing. Gene expression microarrays are glass sides with thousands of short DNA probes printed on the surface. RNA transcripts are reverse transcribed

into complementary DNA and labeled with a fluorescent tag. The labeled cDNA is then allowed to bind to the array of DNA probes. The number of bound cDNA is approximated using fluorescence. Relative expression between a control and experimental group is measured by labeling the groups with different color fluorophores. Microarrays have largely been replaced with RNA sequencing, which is cheaper, more accurate, and can detect a larger range of transcripts. Massively parallel DNA sequencing technology has been expanded to quantify RNA transcript abundance. RNA transcripts are reverse-transcribed into cDNA and put into a DNA sequencing library. RNA sequencing is quantitative such that the number of sequencing reads for a transcript is proportional to the concentration of the transcript in the sample.

Raw RNA sequencing data is in FASTQ format. FASTQ format is a simple text format that lists each sequence with the sequencer's confidence score for calling each base in the sequence. After preprocessing and quality control, the next step in gene expression analysis is to map the sequencing data to a reference genome or transcriptome using sequence similarity. The human genome is well-annotated, and the annotation is used to assign sequencing data to specific genes. There are many algorithms for mapping sequencing data to reference genomes, but one of the most widely adopted algorithms is called STAR [6]. After alignment, gene quantification algorithms count the number of reads that mapped to each gene or transcript. To improve transcript-level quantifications, some algorithms like the RSEM algorithm try to maximize the likelihood of observing the data and estimate an expected count for each gene [25].

Absolute gene expression is difficult to analyze, so a common analysis method is to compare absolute gene expression of two groups of data and identify differences in expression.

Differential expression analysis for cancer studies typically estimate gene expression in two groups of samples, typically a healthy control and disease group, and identifies differences in gene expression. Differential expression analysis can be used to find cancer genes by comparing tumor expression to matched healthy tissue expression. When a tumor is biopsied or resected, the surgeon often takes a sample of healthy tissue for comparison. For many cancer types, it is not feasible to take a matched normal sample. In our experience, pediatric gene expression data rarely has matched normal data, so other methods are needed to identify differentially expressed genes.

One approach to interpreting gene expression results is to integrate genomic data into functional pathways. Pathways describe mechanistic relationships between genes. Synthesizing differentially expressed genes into pathways provides a system-level view of cell function. Examples of pathway databases include the Kyoto Encyclopedia of Genes and Genomes and Reactome [22, 19]. Trained scientists curate pathway databases using scientific literature. One challenge with pathway analysis is that genes interact in a tissue-specific manner and well-curated pathways may not describe subtle changes in biological mechanisms. For this reason, pathway analysis suffers from poor sensitivity.

UCSC Treehouse Approach To Finding Drug Targets

Treehouse is a UCSC pediatric cancer research initiative working to improve childhood cancer therapies using genomic data. Treehouse collaborates with several children's hospitals in California and presents findings at tumor boards at Stanford University, Children's Hospital of Orange County, BC Children's Hospital, and University of California, San Fran-

cisco (UCSF). While other pediatric research programs have focused on genetic variants, Treehouse prioritizes gene expression analysis because there are so few actionable genetic variants. The Treehouse analysis consists of classifying patients based on gene expression profiles and predicting drug sensitivity using gene expression outlier analysis.

Treehouse advocates for open data sharing policies and has built one of the largest cancer gene expression databases called the Treehouse compendium. The Treehouse compendium includes public data from TARGET, TCGA, and the Short Read Archive. The compendium also includes pediatric data obtained through collaboration with children's hospitals and clinical trials. Pediatric gene expression data is relatively rare, so most of the samples in the compendium are from adults. Treehouse compendium V4 has over 11,000 samples representing 77 different cancer types. There are 1,558 pediatric and young adult samples in the compendium. Each version of the Treehouse compendium is processed using the same bioinformatic pipeline to reduce batch effects.

Treehouse has adopted docker containerization as a standard for bioinformatic pipeline development. Docker is software that manages and builds light-weight virtual machines that can run on any computer with docker software installed. Treehouse docker containers ensure that partner institutions are able to run Treehouse methods in a consistent way. This is particularly helpful in an environment where sharing raw data is difficult. For instance, some institutions are unable to share raw sequencing data, so Treehouse can instead send the computation to the data by deploying a dockerized version of the Treehouse pipeline.

The Treehouse workflow begins when clinicians submit RNA sequencing data for analysis (Figure 0.2). Preprocessing and quality control steps ensure that reads are properly

paired and that there is a sufficient number of RNA transcripts for analysis. Treehouse researchers developed a novel QC metric for RNA sequencing data that quantifies the total number of uniquely mapped, exonic, and non-duplicate (UMEND) reads. The number of UMEND reads estimates the total amount of gene-level information in an RNA-sequencing run. A threshold of ten million UMEND reads is used to filter low-quality RNA sequencing data.

Preprocessed RNA sequencing data are then submitted to the UCSC Genomics Core for alignment and gene expression quantification. The major steps in the Genomics Core RNA-seq pipeline are alignment using the STAR algorithm [6] and gene quantification using the RSEM algorithm [25]. The Genomics Core RNA-seq pipeline outputs several normalized gene-level expression estimates, but Treehouse currently uses transcripts per million mapped reads (TPM) normalization.

The Treehouse tertiary analysis pipeline classifies patients into disease cohorts, detects gene expression outliers, identifies enriched pathways, and nominates therapeutic targets. The results of Treehouse tertiary analysis are sent to a trained Treehouse analyst who synthesizes the information and reports findings to clinicians.

Genomic data is high-dimensional and therefore difficult to visualize. Dimensionality reduction methods have been developed to aid in identifying patterns in high-dimensional data. Treehouse uses a method developed in Joshua Stuarts lab at UCSC called TumorMap. TumorMap is a data clustering algorithm that uses the Google Maps API for visualization [3]. The TumorMap visualization for the Treehouse compendium shows that samples tend to cluster by cancer subtype (Figure 0.3). An unexpected TumorMap placement occurs when a patient places with a cancer subtype that is different than the patient's original cancer diagnosis. Unex-

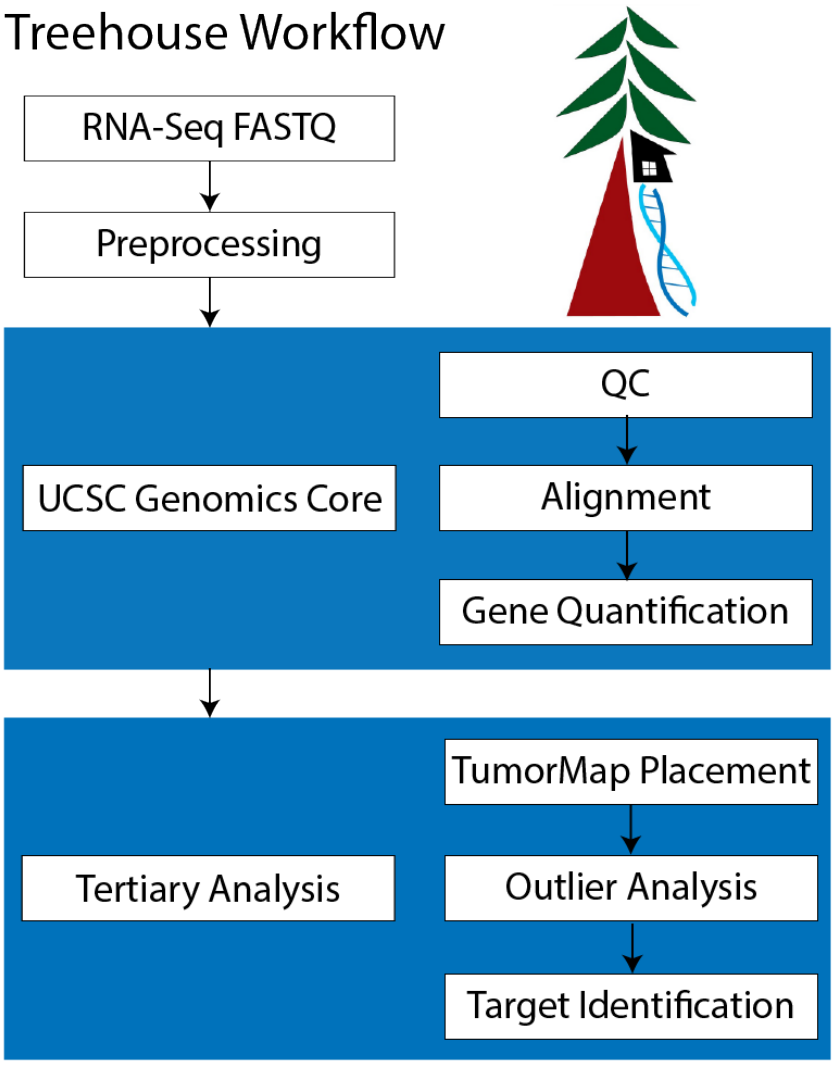


Figure 0.2: The foundation for the Treehouse analysis is the Treehouse compendium. The Treehouse compendium is an ongoing project to improve representation of pediatric samples through collaborations. Samples that are processed through the Treehouse workflow are also placed within the compendium. After receiving Tumor RNA-seq data, the data is preprocessed and the quality of the data is assessed. The preprocessing step checks for common errors in sequencing data, removes adapter sequences, and submits the data for alignment and gene quantification. The tumor gene expression profile is placed on the TumorMap and analyzed using outlier analysis. These results are reviewed by a trained analyst and presented to clinicians for further review.

pected TumorMap placements occur in approximately 20% of Treehouse cases and may suggest a refinement in the molecular diagnosis.

TumorMap also describes the patient's six most similar gene expression profiles or nearest neighbors. The nearest neighbor cancer types are used to define the patients disease cohort for outlier analysis. The disease cohort can range from a single cancer to a mixture of six different cancer types. All compendium samples that belong to the disease cohort are aggregated to estimate the patients expected gene expression profile. The expected gene expression profile is used to find abnormally expressed genes.

Treehouse analysis identifies genes that are over- or under-expressed in a given tumor. Cells over-expressing cancer genes are sensitive to targeted inhibitors. Kothari et al. identified sensitivity to ERBB2 inhibition by trastuzumab (herceptin) in breast cancer cell lines over-expressing ERBB2. These cells were also over-expressing the FGFR4 genes, and combination trastuzumab and FGFR4 inhibition by PD173074 showed an additive decrease in cell viability [24]. Gene expression outlier analysis has also been used to inform clinical decisions. Jones et al., used over-expression of RET and under-expression of PTEN to infer up-regulation of the MAPK pathway. The patient consented to targeted inhibition of RET using sunitinib and the patient's disease stabilized for four months [18].

There are two kinds of Treehouse outlier analyses. The first is pan-cancer outlier analysis which averages over all cancer types in the Treehouse compendium. Pan-cancer analysis highlight tissue-specific expression features. The second kind of outlier analysis is pan-disease outlier analysis, which uses the TumorMap disease cohort to calculate outlier expression thresholds. The list of gene expression outliers are then used for pathway analysis.

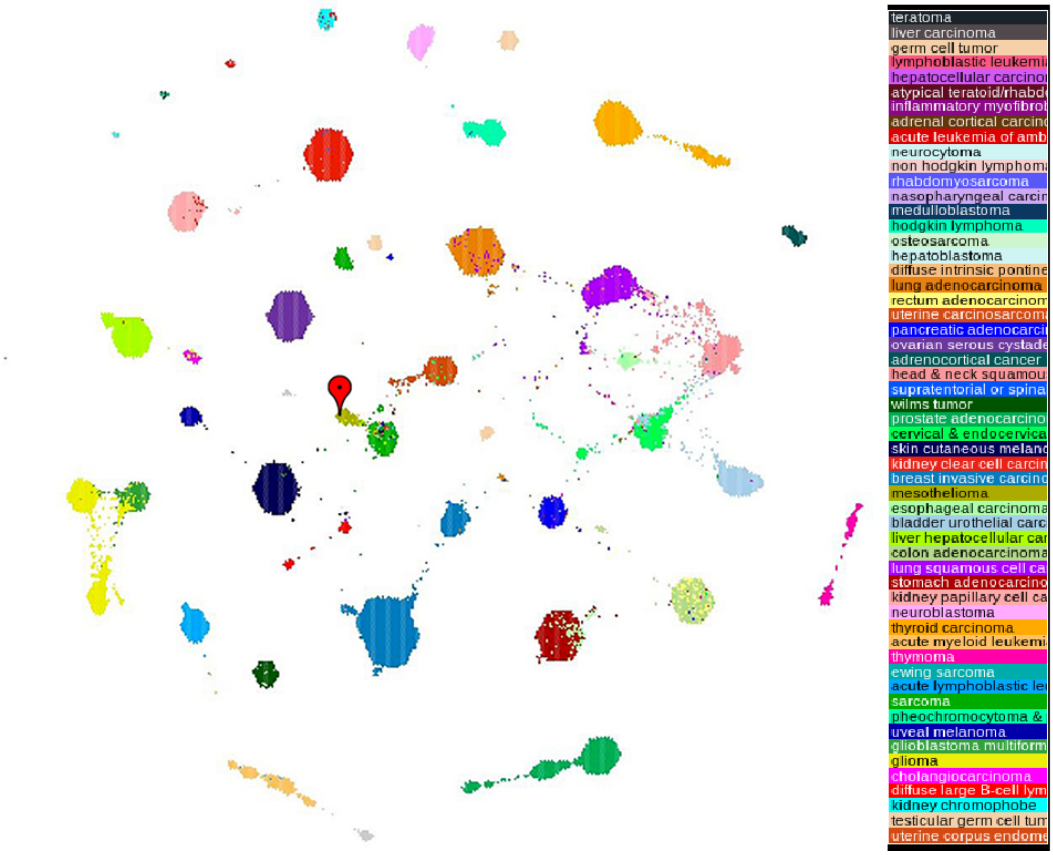


Figure 0.3: TumorMap representation of Treehouse compendium *version*1 shows distinct clustering of gene expression profiles by cancer diagnosis. TumorMap uses the Google Maps API to visualize relationships between genomic features. Each hexagon in the TumorMap represents a sample in the Treehouse compendium and is colored by the patient's cancer diagnosis. Each gene expression profile is grouped by the six most similar gene expression profiles in the compendium. A patient's gene expression profile places with a surprising cancer cluster in 20% of cases.

Treehouse identifies gene expression outliers using the standard Tukey method for univariate data [14]. Over-expression outliers are expressed in the top 5% of all genes and have gene expression levels greater than $Q3 + 1.5 \cdot IQR$, where Q3 is the third quartile and IQR is the interquartile range. Likewise, under-expression outliers have gene expression levels lower than $Q1 - 1.5 \cdot IQR$. The Tukey method sets thresholds for labeling genes expression outliers. The method is analogous to using an outlier threshold of three standard deviations.

Treehouse outlier analysis was developed because current tools for identifying differentially expressed genes are not designed for single sample applications. Differential expression analysis requires replicate expression profiles to control for technical noise. While replicate measurements are important for making accurate statistical inferences, the cost of RNA sequencing and the limited amount of cancer tissue per patient make it difficult to generate replicate gene expression profiles.

Differential expression analysis also requires defining two conditions. For cancer, the two conditions are usually cohort of paired healthy tissue, or normal samples, and the second condition is a cohort of disease samples. In addition to having limited cancer tissue, in our experience, it is more difficult to obtain paired normal pediatric tissue. Therefore, there is not a control group to compare pediatric cancer expression to. This is one reason to assemble the Treehouse compendium of adult and pediatric cancer because we can use other pediatric cancer samples to identify patterns in expression for pediatric tissue.

Pathway analysis is used to interpret gene expression outlier lists. Pathway analysis uses prior knowledge of molecular biology to interpret genomic data. Pathways are often represented as lists of related genes called gene sets. Gene set enrichment analysis is used to find

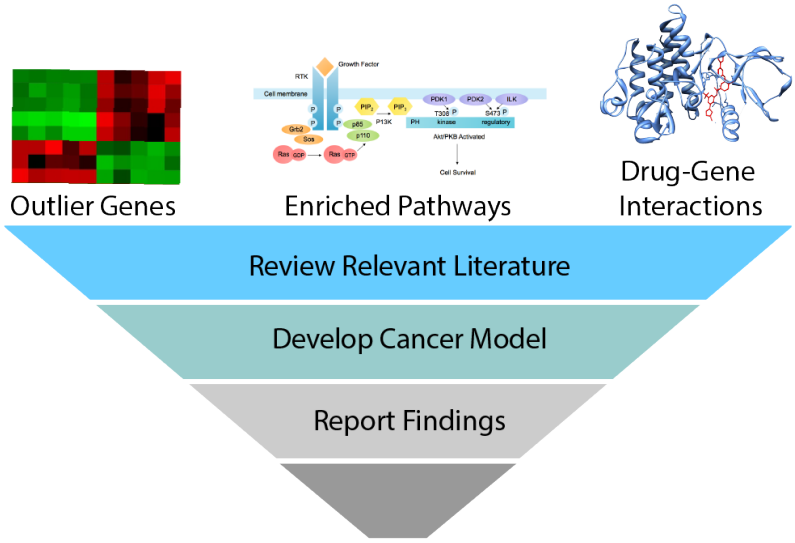


Figure 0.4: Process of narrowing down Treehouse tertiary analysis results. Treehouse tertiary analysis produces outlier genes, enriched pathways, and lists of known drug-gene interactions. Gene expression outliers are prioritized if there is pathway level evidence and the outlier is druggable. Relevant literature is used to refine the model and provide evidence for gene interactions. Clinical information, including genetic testing, is used to supplement the cancer model. Finally, the results are discussed with clinicians who provide more evidence for the Treehouse drug targets.

statistically significant overlap between gene expression outliers and pathway gene sets. Treehouse uses the MSigDB website for gene set enrichment analysis [27]. Hallmark and Canonical pathway gene sets are used to interpret gene expression outlier results [27, 26]. Hallmark gene sets annotate gene expression programs under specific biological conditions. For example, the MYC Targets V1 gene set contains genes expressed at high levels when the oncogenic MYC protein is active. Canonical gene sets describe well characterized protein interactions that may not be reflected in gene expression data.

In order to identify potential druggable targets, over-expressed genes from pan-cancer and pan-disease analysis are used as input data into the Drug Gene Interaction Database (DGIdb) [12]. DGIdb pulls data from publications to find the relationship between genes and their potential drug inhibitors. For our CKCC analysis, we set DGIdb to query for drug-gene interactions among four cancer databases: CIVic, CancerCommons, MyCancerGenome, and MyCancerGenomeClinicalTrail, thus limiting our findings to only cancer therapies. DGIdb does not contain all known drug-gene interactions nor does it guarantee gene druggability. As a result, literature searches are used to find rational targeted inhibitors for over-expressed genes.

Treehouse analysis ends with synthesizing gene expression outlier results, enriched pathway information, drug-gene interaction data, and relevant literature. The goal of the analysis is to build a descriptive model for the patients cancer and identify targeted inhibitors that could impede tumor growth. Treehouse therapeutic directions consist of FDA-approved drugs, off-label use of FDA-approved drugs for adults, and targeted therapies currently in pediatric clinical trials. This information is presented back to clinicians for review.

Alternative methods to differential expression analysis include GFOLD and Cancer

Outlier Profile Analysis (COPA). GFOLD is the state-of-the-art method for ranking genes based on fold-change. GFOLD prioritizes genes that have high fold change relative to controls and a large number of read counts. GFOLD performs better than differential expression algorithms when working with a single biological replicate [9]. The Treehouse algorithm is similar to GFOLD in that a gene expression outlier needs to be expressed at a much higher level than the median and be in the top 5% of all expressed genes.

Many differential expression tools are based on a t-test for comparing two means. One challenge with this approach is that some samples in a cohort may have differential gene expression that is not consistent with the overall population. For a particular disease, patient A may have MYC over-expression and normal levels of CDK4, but patient B may have CDK4 over-expression and normal levels of MYC. The COPA method was designed to find subtle patterns of differential expression compared to a normal cohort. The COPA method assumes that the healthy cohort will not have pathogenic expression, but samples within the experimental disease cohort will show mutually exclusive expression for pairs of genes [29, 42]. This approach fails for Treehouse analysis because our control cohort includes cancer samples that will likely have over-expression of oncogenic genes.

Chapter 1

Comparative Tumor RNA Sequencing Analysis for Difficult-to-Treat Pediatric and Young Adult Patients With Cancer

Introduction

Innovation in the treatment of pediatric cancers has lagged behind that of adult cancers, although many of the FDA-approved therapies for adults likely have efficacy in pediatric cancers. In order to repurpose available cancer therapies for pediatric cancers, the UCSC Treehouse Childhood Cancer Initiative developed a gene expression analysis called Treehouse outlier analysis to match individual patients to FDA-approved drugs. The use of gene expression data is particularly important since many pediatric cancers have low mutation burdens with some tumors lacking a single somatic mutation. Current research suggests that pediatric cancers are driven by epigenetic dysregulation, so an analysis of gene expression may yield leads

for more cases. As described above, the Treehouse outlier analysis uses the Tukey box-and-whisker plot thresholds for defining overexpression. Using our gene expression approach, we found more actionable leads than those found by a strictly DNA-level analysis of pediatric tumors.

As a Treehouse case analyst, I learned to apply the Treehouse gene expression analysis to pediatric cancer cases. I was responsible for cases from UCSF where I regularly analyzed cases that were presented at UCSF molecular tumor boards. My investigation of individual pediatric cases helped to shape the Treehouse approach and I identified several improvements that have been implemented in Treehouse case analysis. I also trained several people in the Treehouse group to apply Treehouse outlier analysis which has led to improvements the overall analysis. I provided bioinformatic and statistical expertise for the Treehouse analysis, including the theoretical background underlying the Tukey outlier method. I also generated figure 5 for the [40] manuscript, which detailed the benefits of the Treehouse expression outlier approach compared to the current clinical practice of focusing on mutated genes.

JAMA Network Open

Original Investigation | Genetics and Genomics

Comparative Tumor RNA Sequencing Analysis for Difficult-to-Treat Pediatric and Young Adult Patients With Cancer

Olena M. Vaske, PhD; Isabel Bjork, JD, MSc; Sofie R. Salama, PhD; Holly Beale, PhD; Avanthi Tayi Shah, MD; Lauren Sanders, BSc; Jacob Pfeil, BSc; Du L. Lam, BSc; Katrina Learned, BSc; Ann Durbin, BSc; Ellen T. Kephart, BSc; Rob Currie, MBA; Yulia Newton, PhD; Teresa Swatloski, BSc; Duncan McColl, BSc; John Vivian, PhD; Jingchun Zhu, PhD; Alex G. Lee, PhD; Stanley G. Leung, BSc; Aviv Spillinger, BSc; Heng-Yi Liu, BSc; Winnie S. Liang, PhD; Sara A. Byron, PhD; Michael E. Berens, PhD; Adam C. Resnick, PhD; Norman Lacayo, MD; Sheri L. Spunt, MD; Arun Rangaswami, MD; Van Huynh, MD; Lilibeth Torno, MD; Ashley Plant, MD; Ivan Kirov, MD; Keri B. Zabokrtsky, MSc; S. Rod Rassekh, MD; Rebecca J. Deyell, MD; Janessa Laskin, MD; Marco A. Marra, PhD; Leonard S. Sender, MD; Sabine Mueller, MD, PhD; E. Alejandro Sweet-Cordero, MD; Theodore C. Goldstein, PhD; David Haussler, PhD

Abstract

IMPORTANCE Pediatric cancers are epigenetic diseases; therefore, considering tumor gene expression information is necessary for a complete understanding of the tumorigenic processes.

OBJECTIVE To evaluate the feasibility and utility of incorporating comparative gene expression information into the precision medicine framework for difficult-to-treat pediatric and young adult patients with cancer.

DESIGN, SETTING, AND PARTICIPANTS This cohort study was conducted as a consortium between the University of California, Santa Cruz (UCSC) Treehouse Childhood Cancer Initiative and clinical genomic trials. RNA sequencing (RNA-Seq) data were obtained from the following 4 clinical sites and analyzed at UCSC: British Columbia Children's Hospital (n = 31), Lucile Packard Children's Hospital at Stanford University (n = 80), CHOC Children's Hospital and Hyundai Cancer Institute (n = 46), and the Pacific Pediatric Neuro-Oncology Consortium (n = 24). The study dates were January 1, 2016, to March 22, 2017.

EXPOSURES Participants underwent tumor RNA-Seq profiling as part of 4 separate clinical trials at partner hospitals. The UCSC either downloaded RNA-Seq data from a partner institution for analysis in the cloud or provided a Docker pipeline that performed the same analysis at a partner institution. The UCSC then compared each participant's tumor RNA-Seq profile with more than 11 000 uniformly analyzed tumor profiles from pediatric and young adult patients with cancer, downloaded from public data repositories. These comparisons were used to identify genes and pathways that are significantly overexpressed in each patient's tumor. Results of the UCSC analysis were presented to clinical partners.

MAIN OUTCOMES AND MEASURES Feasibility of a third-party institution (UCSC Treehouse Childhood Cancer Initiative) to obtain tumor RNA-Seq data from patients, conduct comparative analysis, and present analysis results to clinicians; and proportion of patients for whom comparative tumor gene expression analysis provided useful clinical and biological information.

RESULTS Among 144 samples from children and young adults (median age at diagnosis, 9 years; range, 0-26 years; 72 of 118 [61.0%] male [26 patients sex unknown]) with a relapsed, refractory, or rare cancer treated on precision medicine protocols, RNA-Seq–derived gene expression was potentially useful for 99 of 144 samples (68.8%) compared with DNA mutation information that was potentially useful for only 34 of 74 samples (45.9%).

(continued)

Key Points

Question Is it feasible and useful to compare the tumor RNA sequencing data of a child or young adult with the tumor RNA sequencing data of thousands of other patients (of all ages) in a research setting?

Findings Among 144 tumor samples from children and young adults, comparative RNA sequencing analysis, conducted across 4 precision medicine studies in the United States and Canada, was feasible and potentially useful for 99 of 144 pediatric and young adult cancer samples. In contrast, DNA mutation information was potentially useful for only 34 of 74 samples.

Meaning This study's findings suggest that open sharing and combined analysis of tumor RNA sequencing data from pediatric and young adult patients treated on different clinical trials may represent a feasible approach and may produce useful clinical and biological information for individual patients.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

CONCLUSIONS AND RELEVANCE This study's findings suggest that tumor RNA-Seq comparisons may be feasible and highlight the potential clinical utility of incorporating such comparisons into the clinical genomic interpretation framework for difficult-to-treat pediatric and young adult patients with cancer. The study also highlights for the first time to date the potential clinical utility of harmonized publicly available genomic data sets.

JAMA Network Open. 2019;2(10):e1913968. doi:10.1001/jamanetworkopen.2019.13968

Introduction

We present a framework for comparative RNA sequencing (RNA-Seq) analysis of pediatric tumors across multiple precision medicine studies. Our framework uses public genomic data sets of more than 11 000 tumor RNA-Seq samples that we consolidated and released to the community. We describe an application of our framework and the data compendium to the analysis of 144 tumors from children and young adults with a relapsed, refractory, or rare cancer, studied on 4 separate precision medicine trials in the United States and Canada.

While genomic profiling of tumors is becoming the standard of care in oncology, many tumors, especially in children, do not harbor actionable DNA aberrations. Tumor gene expression information may increase the number of actionable aberrations detected in tumors, and its utility is being evaluated in adults (eg, the WINTHER trial[1]). Results of several studies suggested the possible clinical utility of RNA-Seq for children. The Michigan Oncology Sequencing Center's Peds-MiOncoSeq study[2] evaluated 92 patients with relapsed or refractory tumors using a combination of whole-exome sequencing (WES) and RNA-Seq and reported that 46% of samples had actionable findings, including 36% of this subset that had gene fusions with a known or suspected role in tumorigenesis identified through RNA-Seq analysis. In another study[3] of 59 children, most with relapsed or refractory cancers, analysis revealed actionable findings, including RNA fusions, in 51% of cases. The Individualized Therapy for Relapsed Malignancies in Childhood (INFORM) consortium[4] studied 57 patients with WES, low-coverage whole-genome sequencing, RNA-Seq, methylation, and gene expression microarrays and reported a 50% rate of actionable findings that included overexpression of druggable oncogenes. Several patients whose tumors exhibited oncogene overexpression were placed on targeted therapies against these alterations.[4] Finally, the Precision in Pediatric Sequencing (PIPseq) program[5] profiled 65 patients using a combination of tumor or normal WES and tumor RNA-Seq. Tumor RNA-Seq identified therapeutic targets in 23% of the patients; these targets included overexpression of druggable oncogenes, defined based on comparisons of tumor RNA-Seq expression with the RNA-Seq expression levels in a panel of normal tissues. While results of these studies suggested that RNA-Seq expression may be clinically beneficial, they did not provide reproducible methods that could be applied across different precision medicine trials.

Our group recently developed a reproducible and scalable approach for performing outlier analysis for pediatric patients with cancer by using large publicly available cancer RNA-Seq data sets.[6] The objective of the present study was to evaluate the feasibility and potential utility of our approach for cancer samples collected prospectively from multiple precision medicine trials in difficult-to-treat pediatric and young adult patients with cancer.

Methods

Study Design

Among 144 tumors from children and young adults, this cohort study was conducted as a consortium of the following 4 clinical sites: British Columbia Children's Hospital (BCCH), Vancouver, British Columbia, Canada; Lucile Packard Children's Hospital at Stanford University (LPCH), Stanford,

California; CHOC Children's Hospital and Hyundai Cancer Institute, Orange, California; and the Pacific Pediatric Neuro-Oncology Consortium (PNOC), San Francisco, California. During the period from January 1, 2016, to March 22, 2017, the University of California, Santa Cruz (UCSC) obtained and processed tumor RNA-Seq data, as well as deidentified clinical and molecular information, for 181 tumors from 161 children and young adults with a relapsed, refractory, or rare cancer treated on precision medicine protocols. Tumor RNA-Seq data were obtained from the following 4 clinical sites: BCCH (n = 31), LPCH (n = 80), CHOC (n = 46), and PNOC (n = 24). Each clinical site had its own precision medicine protocol in place, and UCSC Treehouse Childhood Cancer Initiative served as a third-party institution conducting secondary analysis of each site's tumor RNA-Seq data. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline.

The BCCH study was approved by the University of British Columbia Research Ethics Committee. The LPCH protocol "Clinical Implementation of Genomic Analysis in Pediatric Malignancies" was approved by the Stanford University Institutional Review Board. The CHOC study "Pilot Project: Molecular Profiles of Newly Diagnosed, Refractory and Recurrent Childhood, Adolescent, and Young Adult Cancers" was approved by the CHOC Children's Hospital and Hyundai Cancer Institute Institutional Review Board. The PNOC-003 protocol has been previously described.[7] The UCSC Treehouse Childhood Cancer Initiative protocol was approved by the UCSC Institutional Review Board.

Because this study involved the sharing of deidentified data, UCSC was not required by our institutional review board to obtain informed consent from study participants; however, clinical partners obtained written informed consent from their participants as per their individual study protocols. All study participants were informed that their deidentified data would be shared with research partners, including UCSC.

Statistical Analysis

Comparative RNA-Seq Analysis

All RNA-Seq data (11 340 compendium samples and 144 samples from clinical partners) were first uniformly processed using the RNA-Seq pipeline version 3.2 developed by the UCSC Computational Genomics Lab[8] (eMethods in the Supplement). The UCSC either downloaded RNA-Seq data from a partner institution for analysis in the cloud or provided a Docker pipeline composed of gene-level expression calculation, which was run at the partner institution; gene expression outlier analysis and identification of druggable genes and pathways was then run on each of the 144 samples at UCSC.

Gene Expression Outlier Analysis

Gene-level transcript per million data were used to perform gene expression outlier analysis[9] to identify transcripts significantly enriched in each patient's tumor compared with either all 11 340 tumors or tumor types identified as most similar (pan-disease analysis). For pan-cancer analysis, we used the filtered set of 27 084 genes; for pan-disease analysis, we used the unfiltered set of 58 581 unique GENCODE Human Release 23 genes (eMethods in the Supplement) to make sure we did not miss genes whose expression is specific to certain tumor subtypes.

Identification of Druggable Overexpressed Genes and Gene Sets

We obtained the following 3 lists of overexpressed genes: one list from pan-disease outlier analysis, a second list from pan-cancer outlier analysis, and a third list from overlapping genes in pan-disease and pan-cancer lists. For each list, we identified potential druggable genes and statistically enriched pathways.

Drug-Gene Interaction Analysis

We used the Drug-Gene Interaction Database to assess which of the overexpressed genes can be considered actionable by available therapies.[10] The database programmatically searches through

publications and other curated databases for reported associations between human genes and available inhibitors. To refine our findings to only existing cancer therapies, we set the Drug-Gene Interaction Database to query for drug-gene interactions among the following 4 curated cancer databases (all part of the Drug-Gene Interaction Database[10]): CIViC, Cancer Commons, My Cancer Genome, and My Cancer Genome Clinical Trial. The Drug-Gene Interaction Database does not contain all known drug-gene interactions, nor does it guarantee a gene's druggability. As a result, we performed additional literature searches and consulted published clinical cancer genomic studies. We prioritized studies, such as INFORM,[4] in which gene expression information was considered in assessing the actionability of each gene. The 92 genes for which overexpression was considered directly or indirectly actionable in this study are listed in eTable 1 in the Supplement.

Gene Set Overlap Analysis

In parallel to identifying druggable genes, we used the Molecular Signature Database[11] to identify overexpressed cancer pathways in the tumor sample. Gene set overlap analysis computes statistically significant pathways by evaluating the overlap between the input gene list of overexpressed genes and the gene sets from the Molecular Signature Database[11] collections "Hallmark Gene Sets" and "Canonical Pathways." In this analysis, for each input gene list, we looked at the first 100 reported gene sets that have the false discovery rate (false discovery rate *q* value) below 0.05.

DNA Mutation Analysis

DNA mutation data were obtained from the following platforms: Foundation Medicine gene panel (LPCH), whole-genome sequencing as part of the Personalized Onco-Genomics Program (POG) (BCCH), NantOmics whole-genome sequencing (CHOC), or Ashion Analytics whole-exome sequencing (PNOC). We used the National Cancer Institute (NCI) Pediatric Molecular Analysis for Therapeutic Choice (hereinafter the NCI Pediatric MATCH) considerations to curate the mutation data reported by the DNA platforms and to classify samples into treatment arms based on the DNA aberrations.[12]

Results

Patient Characteristics

To evaluate the feasibility of comparative RNA-Seq analysis across multiple precision medicine studies, we obtained RNA-Seq data from 181 samples from 161 pediatric and young adult patients (age range, 0-29 years; 65 of 108 [60.2%] male) with a relapsed, refractory, or rare cancer treated at the following 4 clinical sites: BCCH (n = 31), LPCH (n = 80), CHOC (n = 46), and PNOC (n = 24). The age at diagnosis was available for 126 individuals: the median age at diagnosis was 9 years, and the range was 0 to 26 years. Among 144 tumor samples, 46 were from female patients, while 72 were male patients; sex was not reported for 26 samples. RNA sequencing quality control analysis (eMethods in the Supplement) was applied to all 181 samples; of these, 144 samples from 128 patients were of sufficient quality for further analysis. For each case, gene-level transcript per million measurements were computed[8] from tumor RNA-Seq data, which were used in 2 types of analyses to identify expression features of potential clinical relevance (**Figure 1**).

Reference Compendium for Tumor Comparisons

To provide a robust reference for tumor comparisons and gene expression outlier detection, we assembled a compendium of 11 340 uniformly analyzed adult, pediatric, and young adult tumor profiles (eTable 2 and eFigure 1 in the Supplement). Of 11 340 samples in the compendium, 1859 (16.4%) were from pediatric, adolescent, and young adult patients with cancer who were younger than 30 years.

Gene Expression Outlier Analysis

Gene expression outlier analysis is a promising method for identifying druggable overexpressed oncogenes in adult tumors.[9,13] We performed gene expression outlier analysis against similar tumors (pan-disease analysis) and against all cancers in our compendium (pan-cancer analysis) (eMethods in the Supplement).

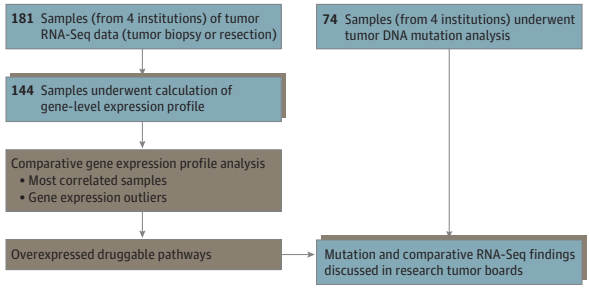
The gene expression outliers were analyzed for the presence of genes whose products could be targeted by small molecules directly or indirectly by targeting the downstream signaling pathway (eTable 1 in the Supplement). This list is based on a similar list prepared by the INFORM study[4] and contains 37 genes whose protein products can be targeted directly and 55 genes whose products cannot be targeted but that function in a pathway that can be targeted by a therapy. We hypothesized that aberrant gene dosage of these directly or indirectly actionable genes could be detected by gene expression outlier analysis. We also sought to assess whether multiple members of the same pathways were highly expressed in concert in the same tumor.

Of 144 high-quality RNA-Seq data sets, 99 (68.8%) harbored outlier gene expression of 1 of 92 actionable genes. In 75 samples, both an actionable gene and the corresponding pathway were overexpressed using outlier analysis. The most common gene expression outlier was *FLT3* (OMIM 136351), overexpressed in 16 samples, all from hematopoietic tumors. This was followed by *BTK* (OMIM 300300) and *CDK6* (OMIM 603368), overexpressed in 14 samples each. While *BTK* was overexpressed in 14 hematopoietic tumors, *CDK6* was overexpressed in both hematopoietic and nonhematopoietic tumors, including neuroblastoma and glioma. The most common gene expression outlier in nonhematopoietic tumors was *PTCH1* (OMIM 601309), overexpressed in 11 samples from craniopharyngioma, neurofibroma, sarcoma, glioma, medulloblastoma, and osteosarcoma. The most common overrepresented gene set was receptor tyrosine kinases, overexpressed in 55 samples from all diagnostic categories (**Figure 2**). Among these, *FLT3* was most commonly overexpressed, followed by *FGFR1* (OMIM 136350) and *PDGFRA* (OMIM 173490). While *FGFR1* was overexpressed in a variety of nonhematopoietic tumor types, *PDGFRA* was exclusively overexpressed in brain tumors, and *FLT3* was exclusively overexpressed in acute leukemias. Of the 92 actionable genes, 47 were overexpressed in 2 or more samples (**Figure 3**). For the remaining 45 of the 144 samples (31.3%), our comparative RNA-Seq analysis did not identify any actionable outliers (eTable 3 in the Supplement). An example of Treehouse analysis is provided in eFigure 2 in the Supplement.

Comparison of RNA-Seq Findings With DNA Mutation Analysis

A small number of childhood tumors contain DNA alterations that may forecast response to molecularly targeted therapies.[14] Children's Oncology Group NCI Pediatric MATCH[12] is a nationwide basket trial for children and adolescents with relapsed or refractory solid tumors evaluating the use of DNA analysis to match patients to therapies. We had mutation data available for 74 of the 144 samples in our cohort; 52 of 74 were solid tumors.

Figure 1. Treehouse Workflow



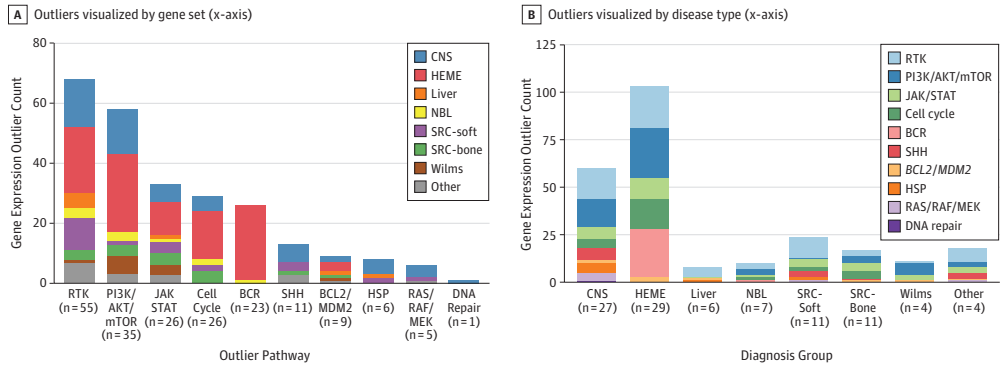
The components in brown are performed by the University of California, Santa Cruz bioinformatics team, while the components in gray are performed by the clinical partners. Calculation of gene-level expression profiles can occur at the University of California, Santa Cruz or at a partner site through the use of portable software. Both the University of California, Santa Cruz and clinical partners participate in research discussions about cases. RNA-Seq indicates RNA sequencing.

Of 74 solid tumor and leukemia samples, 34 (45.9%) had an actionable abnormality as defined by the NCI Pediatric MATCH study[12] detected by DNA analysis. Fifty-five of 74 samples (74.3%) had an actionable gene expression outlier (eTable 3 in the Supplement) detected by RNA-Seq, 28 (37.8%) had abnormalities detected by both DNA and RNA analysis, 6 (8.1%) had only DNA abnormalities, and 13 (17.6%) had no DNA or RNA abnormalities. Remarkably, 27 samples (36.5%) had only a gene expression dosage abnormality, highlighting the potential utility of comparative RNA-Seq for nominating molecular targets for patients with no DNA findings (**Figure 4** and **Figure 5**).

To assess the consistency of DNA and RNA findings, we reviewed 28 samples that had both types of findings. In 11 of 28 samples, at least 1 of the genes with a targetable DNA mutation was identified as a gene expression outlier, suggesting that actionable DNA mutations are often associated with the overexpression of the mutated gene. In 17 of 28 samples, however, none of the genes with a targetable DNA abnormality were identified as a gene expression outlier. Because we do not necessarily expect all mutant genes to be abnormally expressed themselves, we then reviewed the 17 samples to see if there was expression support of the DNA abnormality downstream of the mutated gene.

DNA analysis of 2 acute lymphoblastic leukemia samples (TH01_0122_S01 and TH01_0130_S01) revealed a *PAX5* (OMIM 167414)–*JAK2* (OMIM 147796) fusion, which was previously shown to activate Janus kinase and signal transducer and activator of transcription (JAK/STAT) signaling and promote a progenitor phenotype in leukemia cells.[15] Our comparative gene expression analysis did not reveal the overexpression of the JAK/STAT pathway in these tumors but instead identified overexpression of phosphatidylinositol-3-kinase (PI3K)/AKT and the mammalian target of rapamycin (mTOR) (PI3K/AKT/mTOR) signaling pathway and B-cell receptor signaling pathways in both tumors and overexpression of *FLT3* in TH01_0130_S01. The overexpression of PI3K/AKT/mTOR and B-cell receptor signaling pathway genes may be indicative of a progenitor B-cell state assumed by the leukemia cells.[16] Similarly, another acute lymphoblastic leukemia sample (TH01_0129_S01) harbored a *BCR–ABL* (OMIM 151410) fusion. RNA sequencing revealed outlier expression of PI3K/AKT/mTOR and B-cell receptor signaling pathways; PI3K/AKT/mTOR activation is known to be downstream of the *BCR-ABL* fusion signaling,[17] suggesting that this overexpression is consistent with the DNA finding of the gene fusion. DNA analysis of 5 leukemia samples (TH01_0124_S01, TH01_0134_S01, TH03_0010_S01, TH03_0010_S02, and TH03_0011_S01) identified an activating mutation in *NRAS* (OMIM 164790). Activation of *NRAS* has been associated with proliferation and self-renewal in leukemia via the activation

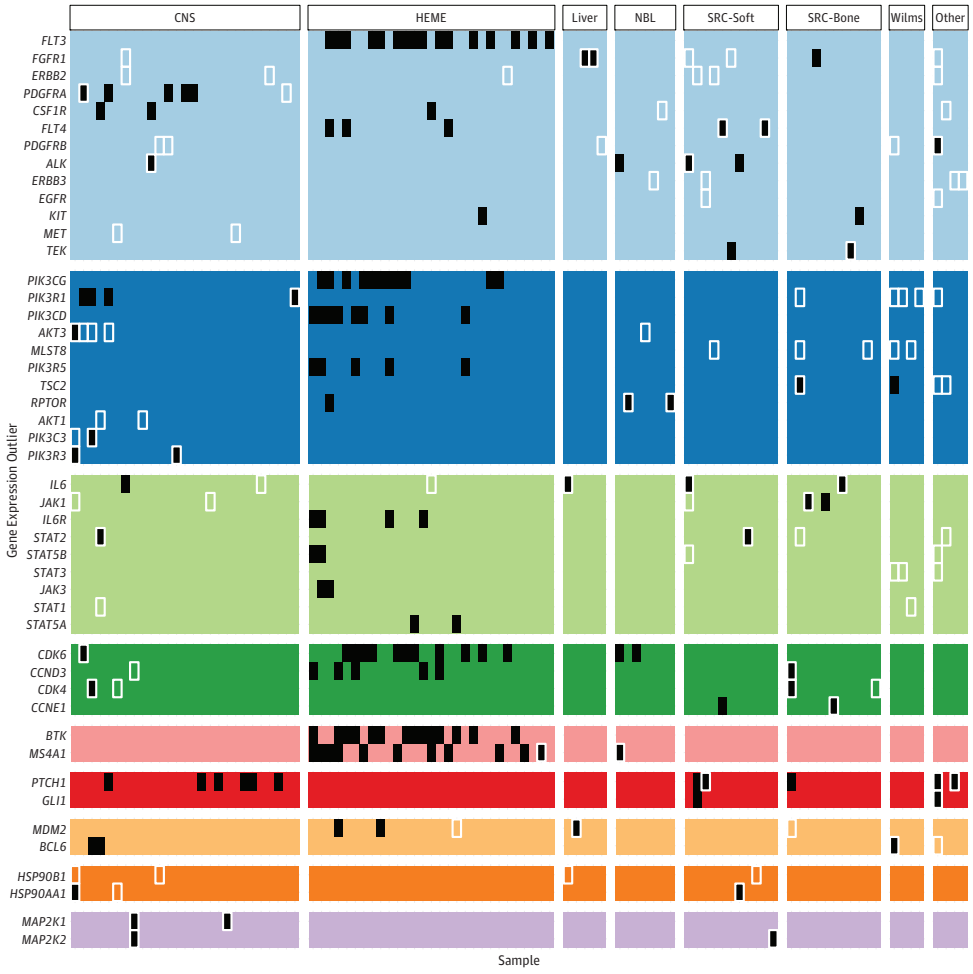
Figure 2. Actionable Gene Expression Outliers Identified Through Comparative RNA Sequencing Analysis of the Cohort



The details of findings in each sample are listed in eTable 3 in the Supplement. BCR indicates B-cell receptor; CNS, central nervous system tumors; HEME, hematopoietic tumors; HSP, heat-shock proteins; JAK/STAT, Janus kinase and signal transducer and activator of transcription signaling pathway; NBL, neuroblastomas; PI3K/AKT/mTOR, phosphatidylinositol-3-kinase (PI3K)/AKT and the mammalian target of rapamycin (mTOR) signaling pathway; RAS/RAF/MEK, mitogen-activated protein kinase RAS/RAF/MEK/ERK pathway; RTK, receptor tyrosine kinases; SHH, sonic hedgehog; and SRC, sarcomas.

of MEK and mTOR signaling pathways.[18] Our RNA-Seq analysis revealed overexpression of cell cycle or *BCL2* (OMIM 603167)–*MDM2* (OMIM 164785) pathways in TH01_0134_S01, TH03_0010_S01, TH03_0010_S02, and TH03_0011_S01; these pathways are downstream of activated RAS signaling, and their overexpression is thus consistent with the activating *NRAS* mutation. Notably, TH01_0124_S01 harbored subclonal activating mutations in both *KRAS* (OMIM 190070) and *NRAS* (20.6% and 29.1% mutant allele frequency based on RNA-Seq, respectively). While gene expression analysis revealed overexpression of *FLT3*, outlier expression associated with pathways downstream of activated RAS signaling was not found. These findings may represent either discordance between the DNA and RNA analysis or intratumor heterogeneity in this leukemia sample, already suspected based on the presence of 2 subclonal RAS mutations.

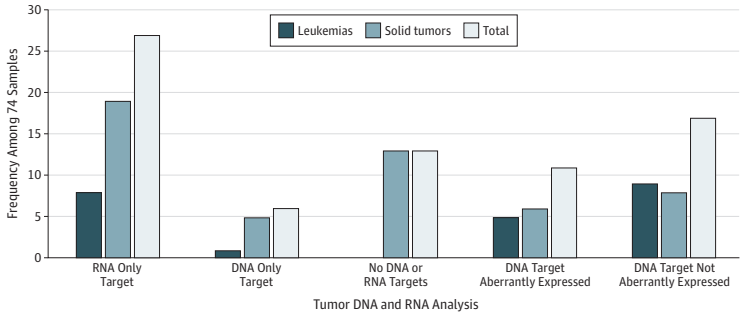
Figure 3. Recurrent Actionable Gene Expression Outliers



Recurrent actionable gene expression outliers (y-axis), colored by gene sets as in Figure 2B, organized by disease (x-axis). Filled black squares denote outliers identified using the pan-cancer analysis approach, while unfilled white squares denote outliers identified by the pan-disease analysis approach. CNS indicates central nervous system tumors; HEME, hematopoietic tumors; NBL, neuroblastoma; and SRC, sarcoma.

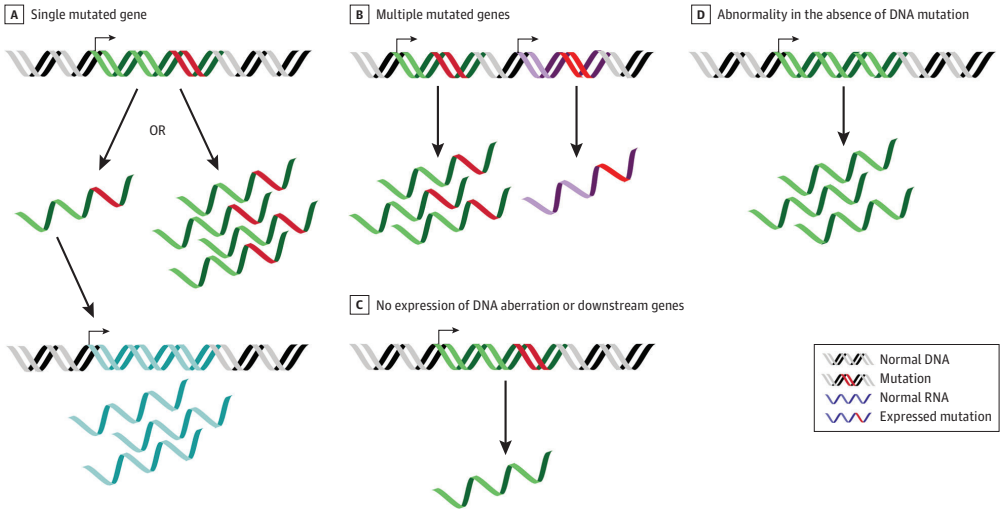
DNA analysis of a diffuse intrinsic pontine glioma (DIPG), TH02_0092_S01, revealed copy number gains of *KDR* (OMIM 191306), *KIT* (OMIM 164920), and *PDGFRA*, located on 4q12. Notably, while *KIT* and *PDGFRA* were highly expressed but not meeting the outlier threshold (84th and 93rd percentiles in the compendium), *KDR* was expressed at a much lower level, in the 54th percentile. Therefore, considering expression information alongside the copy number information may be useful for prioritizing druggable targets within copy number amplicons.[19] In another DIPG sample, TH02_0091_S01 with a *BRAF* (OMIM 164757) p.V600E mutation, gene expression analysis revealed outlier expression of *CSF1R* (OMIM 164770). Recent work in melanoma showed that overexpression of *CSF1R* can occur in melanomas with activating *BRAF* or MAPK mutations and is associated with resistance to BRAF inhibitors.[20] Because the interaction of these 2 pathways in DIPG is not known, we did not consider these concordant DNA and RNA findings.

Figure 4. Comparison of DNA and RNA Analysis Results



DNA and RNA analysis results were reviewed for 74 samples with both types of data available.

Figure 5. Utility of RNA Sequencing (RNA-Seq) Analysis



A, RNA-Seq analysis can be used as additional support for DNA aberrations when a single mutated gene is itself highly expressed or downstream genes are highly expressed as a result of the mutation. B, With multiple mutated genes, RNA-Seq analysis can be used to prioritize among them based on high expression of the mutated gene itself or downstream targets. C, If DNA aberration is not expressed, nor are downstream genes, RNA-Seq analysis can be used to deprioritize DNA abnormalities with no evidence of effectiveness at the level of RNA. D, RNA-Seq analysis can reveal an abnormality in the absence of DNA mutation.

An atypical teratoid rhabdoid tumor (TH03_0016_S01) and myoepithelial carcinoma (TH03_0113_S01) harbored loss of *SMARCB1* (OMIM 601607) (INI1) through a frameshift mutation or protein loss of unknown mechanism detected by immunohistochemistry, respectively. Comparative gene expression analysis of both tumors revealed outlier expression of *FGFR1*, a promising target in rhabdoid tumors deficient in *SMARCB1* (INI1).[21] Gene expression analysis of DIPG tumor TH02_0087_S01 with a loss-of-function mutation of *PIK3R1* (OMIM 171833) activating the PI3K/AKT/mTOR pathway revealed overexpression of the JAK/STAT pathway. While it is unknown whether PI3K/AKT/mTOR and JAK/STAT pathways interact in DIPG, these pathways may be coactivated as a result of PI3K mutations in meningiomas.[22] Because the interaction of these 2 pathways in DIPG is not known, we did not count this sample as having concordant DNA and RNA findings. Comparative gene expression analysis of a malignant peripheral nerve sheath tumor TH06_0645_S01 and neurofibroma TH06_0646_S01 with loss of *NF1* (OMIM 162200) revealed overexpression of sonic hedgehog signaling present in this tumor type.[23] We also identified overexpression of receptor tyrosine kinases *ERBB3* (OMIM 190151) and *EGFR* (OMIM 131550) in these tumors.

Finally, in a glioma TH03_0290_S01 with a *BRAF* p.V600E mutation, the mutation was not expressed in the RNA. In an additional case (TH01_0131_S01), an activating *JAK2* mutation was supported by only a few reads, with more than 100 total read coverage in both the DNA and RNA, suggesting that the mutation may represent a subclonal event or a technical artifact.

Overall, our review of 17 samples with mutated genes not themselves overexpressed by RNA-Seq analysis revealed that in 12 of the 17 samples the overexpressed genes and pathways were consistent with the detected DNA mutations, even though the mutant genes themselves were not overexpressed. In the remaining 5 samples, outlier expression was not consistent with an activating mutation detected in the sample (including the lack of a *BRAF* p.V600E mutant allele in the RNA in TH03_0290_S01; ambiguous evidence in TH02_0087_S01, TH01_0124_S01, and TH02_0091_S01; and possible technical issues in TH01_0131_S01).

Discussion

DNA sequencing is increasingly integrated in clinical trials to identify new molecular targets for children with incurable cancers. However, molecular targets are found for only a small number of patients, and the yield is much lower than that of similar adult cancer trials.[24] Studies focusing on pediatric cancers have shown that the percentage of patients with potentially actionable findings increases to 40% to 50% when RNA-Seq data are considered alongside DNA mutation information.[4] Herein, we described a framework for including RNA-Seq–derived gene expression information into precision medicine studies. Most notably, we show for the first time to date that such a framework can be used consistently across separate precision medicine clinical trials.

To our knowledge, our work represents the first report of a translational cancer genomic analysis in which prospective patient data are analyzed by a third-party computational group, with results returned to clinicians and researchers. We found that this comparative analysis is feasible and can produce new information of potential clinical relevance in 68.8% of samples. In 36.5% of samples (27 of 74), druggable overexpressed genes and pathways were identified based on RNA analysis alone and were not apparent in the tumor DNA analysis. Our work suggests that direct investigations of the clinical utility and effectiveness of tumor RNA-Seq–derived gene expression information will be valuable, and the next phase of our project will focus on defining the incremental benefit of this approach. The findings from our work also suggest that open sharing of cancer genomic data can benefit each pediatric and young adult patient with cancer so that every family's struggle contributes to the advancement of clinical care for the families that follow.

Clinical Implications

Although this study was not designed to assess clinical consequences, we noted associations of comparative RNA-Seq analysis findings and clinical features. For example, our analysis of a high-risk

neuroblastoma sample revealed outlier expression of the *ALK* (OMIM 105590) kinase and *CDK6* kinase (eFigure 2 in the Supplement). The outlier expression of *CDK6*, as well as several other cell cycle genes, was consistent with a known DNA amplification of *CDK6* in this sample; however, the potential activation of *ALK* (OMIM 191175) was not evident before the RNA analysis. In another example, a 2-year-old boy with multifocal stage 4 hepatoblastoma metastatic to the lungs, was initially treated in the Childhood Liver Tumour Strategy Group of the International Society of Paediatric Oncology (SIOPEL-4) study,[25] followed by surgery, 2 cycles of HEP0731 regimen T protocol, then salvage therapy with 3 cycles of vincristine, irinotecan, and temozolomide and 1 cycle of gemcitabine-oxaliplatin with bevacizumab. The patient had disease progression despite these therapies. Pathological analysis showed well and poorly differentiated hepatoblastoma with fetal and embryonal elements, and immunostaining showed retention of INI1 staining and diffuse nuclear and cytoplasmic β-catenin. Foundation Medicine testing revealed the p.G34V variant in CTNNB1, previously reported in hepatocellular carcinoma as an activating mutation.[26] Comparative RNA-Seq analysis of the liver sample (TH03_0004_S04) uncovered gene expression similar to the proliferation subtype of hepatocellular carcinoma[27,28] as well as outlier expression of HSP90B1, interleukin 6, and 4 other members of the JAK/STAT pathway. The overexpression of HSP90B was previously noted in hepatocellular carcinoma.[29] The proliferative subtype of hepatocellular carcinoma is characterized by increased proliferation, high levels of serum α-fetoprotein (AFP), and chromosomal instability[27]; tumors with chromosomal instability are potentially sensitive to Aurora kinase inhibitors.[30] Consistent with the similarity of the tumor to the proliferative subtype of hepatocellular carcinoma, the patient with the TH03_0004_S04 tumor had a response to the pan-kinase inhibitor pazopanib hydrochloride, with activity against Aurora kinase A.[31] Based on the present study, after initiation of this treatment, the patient had a decline in his AFP levels from 14 036 to 1052 ng/mL at 7 weeks after initiation of the therapy (to convert AFP level to micrograms per liter, multiply by 1.0). At 10 weeks into this therapy, restaging studies showed progressive disease, and the patient was switched to therapy with ruxolitinib phosphate, without objective response by AFP levels or by imaging criteria.

Limitations

Our study has some limitations. The heterogeneous nature of the patients analyzed in this study (all types of relapsed, refractory, and rare cancers) made drawing general statements difficult. The study was not designed to directly evaluate clinical utility of comparative RNA-Seq analysis, and clinical follow-up data on these patients were not readily available.

Conclusions

Our experience suggests that it is feasible to include RNA-Seq–derived gene expression analysis in precision medicine studies and that this analysis can be harmonized across studies. We showed that RNA-Seq–derived gene expression was potentially useful for 68.8% of 144 samples compared with DNA mutation information, which was potentially useful for only 45.9% of 74 samples. Our study also highlights for the first time to date the potential clinical utility of harmonized publicly available genomic data sets. Open sharing and combined analysis of tumor RNA-Seq data from pediatric and young adult patients treated on separate clinical trials represent a feasible approach and can produce useful clinical and biological information for individual patients.

ARTICLE INFORMATION

Accepted for Publication: September 6, 2019.

Published: October 25, 2019. doi:10.1001/jamanetworkopen.2019.13968

Corresponding Author: Olena M. Vaske, PhD, Department of Molecular, Cell, and Developmental Biology, University of California, Santa Cruz, 1156 High St, Santa Cruz, CA 95060 (olena@ucsc.edu).

Author Affiliations: Department of Molecular, Cell, and Developmental Biology, University of California, Santa Cruz (Vaske, Beale); University of California, Santa Cruz Genomics Institute, Santa Cruz (Vaske, Bjork, Salama, Beale, Sanders, Pfeil, Lam, Learned, Durbin, Kephart, Currie, Newton, Swatloski, McColl, Vivian, Zhu, Goldstein, Haussler); Howard Hughes Medical Institute, University of California, Santa Cruz (Salama, Haussler); Division of Hematology and Oncology, Department of Pediatrics, University of California, San Francisco (Tayi Shah, Lee, Leung, Spillinger, Liu, Sweet-Cordero); Integrated Cancer Genomics Division, Translational Genomics Research Institute (TGen), Phoenix, Arizona (Liang, Byron); Cancer and Cell Biology Division, TGen, Phoenix, Arizona (Berens); Center for Data Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania (Resnick); Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California (Lacayo, Spunt, Rangaswami); CHOC Children's Hospital, Hyundai Cancer Institute, Orange, California (Huynh, Torno, Plant, Kirov, Zabokrtsky, Sender); British Columbia Children's Hospital Research Institute, British Columbia Children's Hospital, Vancouver, British Columbia, Canada (Rassekh, Deyell); BC Cancer, Vancouver, British Columbia, Canada (Laskin); Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, British Columbia, Canada (Marra); Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada (Marra); Department of Neurology, University of California, San Francisco (Mueller); Department of Neurosurgery, University of California, San Francisco (Mueller); Department of Pediatrics, University of California, San Francisco (Mueller); Now with Anthem, Inc, Palo Alto, California (Goldstein).

Author Contributions: Drs Sender, Mueller, Sweet-Cordero, Goldstein, and Haussler are co–senior authors. Dr Vaske had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Vaske, Bjork, Salama, Pfeil, Newton, Resnick, Spunt, Deyell, Laskin, Mueller, Goldstein, Haussler.

Acquisition, analysis, or interpretation of data: Vaske, Beale, Tayi Shah, Sanders, Lam, Learned, Durbin, Kephart, Currie, Swatloski, McColl, Vivian, Zhu, Lee, Leung, Spillinger, Liu, Liang, Byron, Berens, Resnick, Lacayo, Spunt, Rangaswami, Huynh, Torno, Plant, Kirov, Zabokrtsky, Rassekh, Deyell, Marra, Sender, Mueller, Sweet-Cordero, Goldstein.

Drafting of the manuscript: Vaske, Bjork, Beale, Pfeil, Lam, Currie, Swatloski, Resnick, Lacayo, Torno, Zabokrtsky, Marra, Goldstein.

Critical revision of the manuscript for important intellectual content: Vaske, Salama, Beale, Tayi Shah, Sanders, Pfeil, Learned, Durbin, Kephart, Newton, McColl, Vivian, Zhu, Lee, Leung, Spillinger, Liu, Liang, Byron, Berens, Lacayo, Spunt, Rangaswami, Huynh, Plant, Kirov, Rassekh, Deyell, Laskin, Sender, Mueller, Sweet-Cordero, Goldstein, Haussler.

Statistical analysis: Vaske, Beale, Pfeil, Lam, Resnick, Goldstein.

Obtained funding: Vaske, Bjork, Resnick, Spunt, Rassekh, Deyell, Laskin, Sender, Mueller, Goldstein, Haussler.

Administrative, technical, or material support: Bjork, Sanders, Learned, Durbin, Kephart, Currie, Swatloski, McColl, Lee, Spillinger, Liu, Liang, Resnick, Spunt, Torno, Plant, Zabokrtsky, Rassekh, Deyell, Laskin, Sender, Sweet-Cordero, Goldstein.

Supervision: Vaske, Bjork, Salama, Pfeil, Berens, Resnick, Lacayo, Laskin, Goldstein, Haussler.

Conflict of Interest Disclosures: Drs Vaske, Beale, and Haussler, Mss Sanders, Lam, Learned, Durbin, and Kephart, and Mr Pfeil reported receiving grants from the State of California's California Initiative to Advance Precision Medicine (CIAPM), St Baldrick's Foundation, Alex's Lemonade Stand Foundation, Unravel Pediatric Cancer, Team G Childhood Cancer Foundation, and Live for Others Foundation. Dr Vaske disclosed that her spouse is an employee of ImmunityBio Inc (formerly NantOmics) and has equity interests in NantHealth. Ms Bjork reported receiving grants from the CIAPM. Dr Newton reported receiving funding from ImmunityBio Inc (formerly NantOmics). Ms Swatloski reported receiving grants from the American Association for Cancer Research, California Initiative to Advance Precision Medicine, National Institutes of Health (NIH)/National Cancer Institute, NIH/National Heart, Lung, and Blood Institute, Prostate Cancer Foundation, and Northern California California Institute for Regenerative Medicine (CIRM) Genomics Center of Excellence. Drs Byron and Berens reported receiving grants from TGen Foundation. Dr Spunt reported receiving grants from University of California, Santa Cruz, F. Hoffman–La Roche & Co, Novartis, Alex's Lemonade Stand Foundation, Cookies for Kids' Cancer, Bayer HealthCare Pharmaceuticals, Sanofi US Services, Inc, Loxo Oncology, Incyte Corporation, Bristol-Myers Squibb, St Baldrick's Foundation, and Pfizer. Ms Zabokrtsky reported being supported by grants from Hyundai Motor America/Hyundai Hope on Wheels. Dr Laskin reported receiving grants from Roche Canada and AstraZeneca and receiving honoraria for academic talks from Boehringer Ingelheim, Roche Canada, and Pfizer. Dr Marra reported receiving grants from British Columbia Cancer Foundation and Genome British Columbia. Dr Goldstein reported disclosing that this work

was completed before he joined Anthem, Inc, while he was on the faculty of University of California, Santa Cruz. Dr Haussler reported receiving grants from Howard Hughes Medical Institute, having a patent to BAMBAM issued and a patent to PARADIGM issued, and disclosing that Five3 Genomics, LLC and NantOmics data were used in this article. No other disclosures were reported.

Funding/Support: This study was funded by St Baldrick's Foundation Consortium Award and Emily Beazley Kures for Kids Fund Hero Award, CIAPM, Alex's Lemonade Stand Foundation for Childhood Cancer Research, Unravel Pediatric Cancer, Team G Childhood Cancer Foundation, and Live for Others Foundation. Dr Goldstein was supported by NIH grant U24 CA195858 from the National Cancer Institute Oncology Models Forum. Dr Haussler is a Howard Hughes Medical Institute Investigator. The Personalized Onco-Genomics (POG) team acknowledges the generous support of the British Columbia Cancer Foundation and Genome British Columbia (project B20POG), as well as contributions toward equipment and infrastructure from Genome Canada and Genome British Columbia (projects 202SEQ, 212SEQ, and 12002), Canada Foundation for Innovation (projects 20070, 30981, 30198, and 33408), and the B.C. Knowledge Development Fund.

Role of the Funder/Sponsor: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: Jacquelyn M. Roger, University of California, Santa Cruz, helped in evaluating the significance of outlier genes. A. Geoffrey Lyle, MSc, University of California, Santa Cruz, assisted in addressing editorial queries. They received no compensation for their contributions. We thank the patients and their parents for participating in this study.

Additional Information: Dr Vaske holds the Colligan Presidential Chair in Pediatric Genomics. The Treehouse data compendium (eTable 2 in the Supplement) and processed RNA-Seq data from the 144 patient samples discussed in this study are publicly available (https://treehousegenomics.ucsc.edu/p/vaske-2019-comparative-tumor-RNA). A previously published case (https://www.ncbi.nlm.nih.gov/pubmed/31511612) was included in the cohort of 144 samples. The POG pediatrics program (https://www.personalizedoncogenomics.org/; poginfo@bcgsc.ca) does not consent for data release, as such raw data are not posted publicly. The CHOC raw data have not been consented for data release and are not posted publicly. Raw data from the Stanford trial are available through the European Genome-Phenome Archive (accession No. EGAS00001003900). Raw data from PNOC-003 are available on CAVATICA (https://cavatica.sbgenomics.com/p/datasets#cavatica/cbttc-mixed-pa-01).

REFERENCES

1. Rodon J, Soria JC, Berger R, et al. Challenges in initiating and conducting personalized cancer therapy trials: perspectives from WINTHER, a Worldwide Innovative Network (WIN) Consortium trial. *Ann Oncol*. 2015;26(8):1791-1798. doi:10.1093/annonc/mdv191

2. Mody RJ, Wu YM, Lonigro RJ, et al. Integrative clinical sequencing in the management of refractory or relapsed cancer in youth. *JAMA*. 2015;314(9):913-925. doi:10.1001/jama.2015.10080

3. Chang W, Brohl AS, Patidar R, et al. Multidimensional clinomics for precision therapy of children and adolescent young adults with relapsed and refractory cancer: a report from the Center for Cancer Research. *Clin Cancer Res*. 2016;22(15):3810-3820. doi:10.1158/1078-0432.CCR-15-2717

4. Worst BC, van Tilburg CM, Balasubramanian GP, et al. Next-generation personalised medicine for high-risk paediatric cancer patients: the INFORM pilot study. *Eur J Cancer*. 2016;65(65):91-101. doi:10.1016/j.ejca.2016.06.009

5. Oberg JA, Glade Bender JL, Sulis ML, et al. Implementation of next generation sequencing into pediatric hematology-oncology practice: moving beyond actionable alterations. *Genome Med*. 2016;8(1):133. doi:10.1186/s13073-016-0389-6

6. Newton Y, Rassekh SR, Deyell RJ, et al. Comparative RNA-sequencing analysis benefits a pediatric patient with relapsed cancer [published oline April 19, 2018]. *JCO Precision Oncol*. doi:10.1200/PO.17.00198

7. Mueller S, Jain P, Liang WS, et al. A pilot precision medicine trial for children with diffuse intrinsic pontine glioma–PNOC003: a report from the Pacific Pediatric Neuro-Oncology Consortium. *Int J Cancer*. 2019;145(7):1889-1901. doi:10.1002/ijc.32258

8. Vivian J, Rao AA, Nothaft FA, et al. Toil enables reproducible, open source, big biomedical data analyses. *Nat Biotechnol*. 2017;35(4):314-316. doi:10.1038/nbt.3772

9. Jones SJ, Laskin J, Li YY, et al. Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome Biol*. 2010;11(8):R82. doi:10.1186/gb-2010-11-8-r82

10. Wagner AH, Coffman AC, Ainscough BJ, et al. DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res*. 2016;44(D1):D1036-D1044. doi:10.1093/nar/gkv1165

11. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-1740. doi:10.1093/bioinformatics/btr260

12. Allen CE, Laetsch TW, Mody R, et al; Pediatric MATCH Target and Agent Prioritization Committee. Target and Agent Prioritization for the Children's Oncology Group–National Cancer Institute Pediatric MATCH trial. *J Natl Cancer Inst*. 2017;109(5). doi:10.1093/jnci/djw274

13. Kothari V, Wei I, Shankar S, et al. Outlier kinase expression by RNA sequencing as targets for precision therapy. *Cancer Discov*. 2013;3(3):280-293. doi:10.1158/2159-8290.CD-12-0336

14. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014;505(7484):495-501. doi:10.1038/nature12912

15. Schinnerl D, Fortschegger K, Kauer M, et al. The role of the Janus-faced transcription factor PAX5-JAK2 in acute lymphoblastic leukemia. *Blood*. 2015;125(8):1282-1291. doi:10.1182/blood-2014-04-570960

16. Bertacchini J, Heidari N, Mediani L, et al. Targeting PI3K/AKT/mTOR network for treatment of leukemia. *Cell Mol Life Sci*. 2015;72(12):2337-2347. doi:10.1007/s00018-015-1867-5

17. Dinner S, Platanias LC. Targeting the mTOR pathway in leukemia. *J Cell Biochem*. 2016;117(8):1745-1752. doi:10.1002/jcb.25559

18. Sachs Z, LaRue RS, Nguyen HT, et al. *NRAS*^{G12V} oncogene facilitates self-renewal in a murine model of acute myelogenous leukemia. *Blood*. 2014;124(22):3274-3283. doi:10.1182/blood-2013-08-521708

19. Ohshima K, Hatakeyama K, Nagashima T, et al. Integrated analysis of gene expression and copy number identified potential cancer driver genes with amplification-dependent overexpression in 1,454 solid tumors. *Sci Rep*. 2017;7(1):641. doi:10.1038/s41598-017-00219-3

20. Giricz O, Mo Y, Dahlman KB, et al. The RUNX1/IL-34/CSF-1R axis is an autocrinally regulated modulator of resistance to BRAF-V600E inhibition in melanoma. *JCI Insight*. 2018;3(14):120422. doi:10.1172/jci.insight.120422

21. Wong JP, Todd JR, Finetti MA, et al. Dual targeting of PDGFRα and FGFR1 displays synergistic efficacy in malignant rhabdoid tumors. *Cell Rep*. 2016;17(5):1265-1275. doi:10.1016/j.celrep.2016.10.005

22. El-Habr EA, Levidou G, Trigka EA, et al. Complex interactions between the components of the PI3K/AKT/mTOR pathway, and with components of MAPK, JAK/STAT and Notch-1 pathways, indicate their involvement in meningioma development. *Virchows Arch*. 2014;465(4):473-485. doi:10.1007/s00428-014-1641-3

23. Lévy P, Bièche I, Leroy K, et al. Molecular profiles of neurofibromatosis type 1–associated plexiform neurofibromas: identification of a gene expression signature of poor prognosis. *Clin Cancer Res*. 2004;10(11):3763-3771. doi:10.1158/1078-0432.CCR-03-0712

24. Rahal Z, Abdulhai F, Kadara H, Saab R. Genomics of adult and pediatric solid tumors. *Am J Cancer Res*. 2018;8(8):1356-1386.

25. Zsiros J, Brugieres L, Brock P, et al; International Childhood Liver Tumours Strategy Group (SIOPEL). Dose-dense cisplatin-based chemotherapy and surgery for children with high-risk hepatoblastoma (SIOPEL-4): a prospective, single-arm, feasibility study. *Lancet Oncol*. 2013;14(9):834-842. doi:10.1016/S1470-2045(13)70272-9

26. Taniguchi K, Roberts LR, Aderca IN, et al. Mutational spectrum of β-catenin, AXIN1, and AXIN2 in hepatocellular carcinomas and hepatoblastomas. *Oncogene*. 2002;21(31):4863-4871. doi:10.1038/sj.onc.1205591

27. Chiang DY, Villanueva A, Hoshida Y, et al. Focal gains of *VEGFA* and molecular classification of hepatocellular carcinoma. *Cancer Res*. 2008;68(16):6779-6788. doi:10.1158/0008-5472.CAN-08-0742

28. European Association for the Study of the Liver; European Organisation for Research and Treatment of Cancer. EASL-EORTC Clinical Practice Guidelines: management of hepatocellular carcinoma [published correction appears in *J Hepatol*. 2012;56(6):1430]. *J Hepatol*. 2012;56(4):908-943. doi:10.1016/j.jhep.2011.12.001

29. Gotoh K, Nonoguchi K, Higashitsuji H, et al. Apg-2 has a chaperone-like activity similar to Hsp110 and is overexpressed in hepatocellular carcinomas. *FEBS Lett*. 2004;560(1-3):19-24. doi:10.1016/S0014-5793(04)00034-1

30. Jeng YM, Peng SY, Lin CY, Hsu HC. Overexpression and amplification of Aurora-A in hepatocellular carcinoma. *Clin Cancer Res*. 2004;10(6):2065-2071. doi:10.1158/1078-0432.CCR-1057-03

31. Isham CR, Bossou AR, Negron V, et al. Pazopanib enhances paclitaxel-induced mitotic catastrophe in anaplastic thyroid cancer. *Sci Transl Med*. 2013;5(166):166ra3. doi:10.1126/scitranslmed.3004358

SUPPLEMENT.

eMethods. Supplemental Methods

eTable 1. Directly and Indirectly Actionable Genes Used to Prioritize Gene Expression Outlier Findings

eTable 2. Published Repository Data Sets Included in the Treehouse Reference Compendium v5

Chapter 2

Toil enables reproducible, open source, big biomedical data analyses

Introduction

My first role in the UCSC Genomics Institute was to develop bioinformatics pipelines uing the Toil framework. While working on Toil pipelines, I found the hardcoding of server resources to be a significant limitation of this software. I developed an innovative solution to this problem that allows the user to dynamically allocate server resources. This work was incorporated into the Toil source code and has been adopted by many UCSC bioinformatic pipelines.

open sharing of protocols. With a precise ontology to describe standardized protocols, it may be possible to share methods widely and create community standards.

We envisage that in future individual research laboratories, or clusters of co-located laboratories, will have in-house, low-cost automation work cells but will access DNA foundries via the cloud to carry out complex experimental workflows. Technologies enabling this from companies such as Emerald Cloud Lab (S. San Francisco, CA, USA), Synthace (London) and Transcriptic (Menlo Park, CA, USA) could, for example, send experimental designs to foundries and return output data to a researcher. This 'mixed economy' should accelerate the development and sharing of standardized protocols and metrology standards and shift a growing proportion of molecular, cellular and synthetic biology into a fully quantitative and reproducible era.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

David W McClymont[1] **& *Paul S Freemont***[1,2]

[1]The London DNA Foundry, UK Synthetic Biology Innovation, Commercialisation and Industrial Translation Centre, London, UK. [2]Centre for Synthetic Biology and Innovation, Department of Medicine, South Kensington Campus, Imperial College London, London, UK. e-mail: d.mcclymont@imperial.ac.uk or p.freemont@imperial.ac.uk.

1. Baker, M. *Nature* **533**, 452–454 (2016).
2. Yachie, N. *et al. Nat. Biotechnol.* **35**, 310–312 (2017).
3. Hadimioglu, B., Stearns, R. & Ellson, R. *J. Lab. Autom.* **21**, 4–18 (2016).
4. ANSI SLAS 1–2004: Footprint dimensions; ANSI SLAS 2–2004: Height dimensions; ANSI SLAS 3–2004: Bottom outside flange dimensions; ANSI SLAS 4–2004: Well positions; (ANSI SLAS, 2004).
5. Mckernan, K. & Gustafson, E. in *DNA Sequencing II: Optimizing Preparation and Cleanup* (ed. Kieleczawa, J.) 9.128 (Jones and Bartlett Publishers, 2006).
6. Storch, M. *et al.* BASIC: a new biopart assembly standard for idempotent cloning provides accurate, single-tier DNA assembly for synthetic biology. *ACS Synth. Biol.* **4**, 781–787 (2015).

Toil enables reproducible, open source, big biomedical data analyses

To the Editor:

Contemporary genomic data sets contain tens of thousands of samples and petabytes of sequencing data[1–3]. Pipelines to process genomic data sets often comprise dozens of individual steps, each with their own set of parameters[4,5]. Processing data at this scale and complexity is expensive, can take an unacceptably long time, and requires significant engineering effort. Furthermore, biomedical data sets are often siloed, both for organizational and security considerations and because they are physically difficult to transfer between systems, owing to bandwidth limitations. The solution to better handling these big data problems is twofold: first, we need robust software capable of running analyses quickly and efficiently, and second, we need the software and pipelines to be portable, so that they can be reproduced in any suitable compute environment.

Here, we present Toil, a portable, open-source workflow software that can be used to run scientific workflows on a large scale in cloud or high-performance computing (HPC) environments. Toil was created to include a complete set of features necessary for rapid large-scale analyses across multiple environments. While several other scientific workflow software packages[6–8] offer some subset of fault tolerance, cloud support and HPC support, none offers these with the scale and efficiency to process petabyte and larger-scale data sets efficiently. This sets Toil apart in its capacity to produce results faster and for less cost across diverse environments.

We demonstrate Toil by processing >20,000 RNA-seq samples (**Fig. 1**). The resulting meta-analysis of five data sets is available to readers[9]. The large majority (99%) of these samples were analyzed in under 4 days using a commercial cloud cluster of 32,000 preemptable cores.

To support the sharing of scientific workflows, we designed Toil to execute common workflow language (CWL; **Supplementary Note 1**) and provide draft support for workflow description language (WDL). Both CWL and WDL are standards for scientific workflows[10,11]. A workflow comprises a set of tasks, or 'jobs', that are orchestrated by specification of a set of dependencies that map the inputs and outputs between jobs. In addition to CWL and draft WDL support, Toil provides a Python application program interface (API) that allows workflows to be declared statically, or generated dynamically, so that jobs can define further jobs during execution and therefore as needed (**Supplementary Note 2** and **Supplementary Toil Documentation**). The jobs defined in either CWL or Python can consist of Docker containers, which permit sharing of a program without requiring individual tool installation or configuration within a specific environment. Open-source workflows that use containers can be run regardless of environment. We provide a repository of genomic workflows as examples[12]. Toil supports services, such as databases or servers, that are defined and

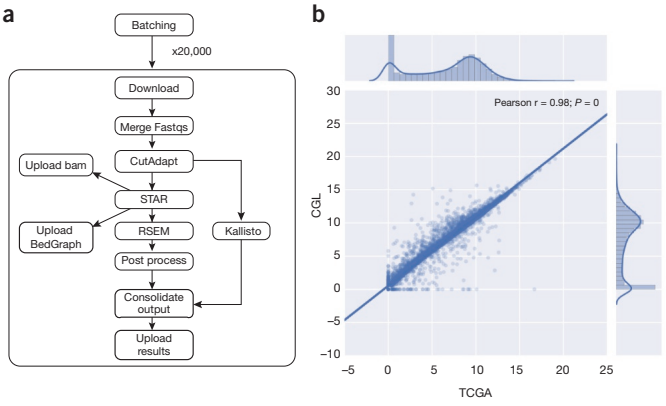


Figure 1 RNA-seq pipeline and expression concordance. (**a**) A dependency graph of the RNA-seq pipeline we developed (named CGL). CutAdapt was used to remove extraneous adapters, STAR was used for alignment and read coverage, and RSEM and Kallisto were used to produce quantification data. (**b**) Scatter plot showing the Pearson correlation between the results of the TCGA best-practices pipeline and the CGL pipeline. 10,000 randomly selected sample and/or gene pairs were subset from the entire TCGA cohort and the normalized counts were plot against each other; this process was repeated five times with no change in Pearson correlation. The unit for counts is: $\log_2$(norm_counts+1).

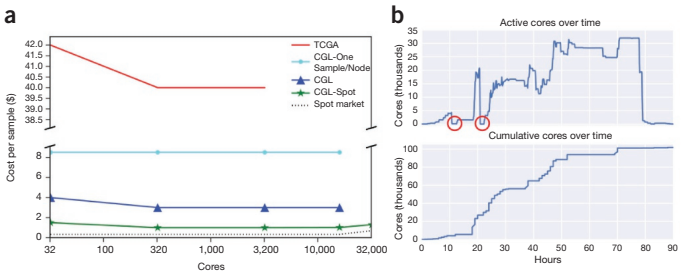


Figure 2 Costs and core usage. (**a**) Scaling tests were run to ascertain the price per sample at varying cluster sizes for the different analysis methods. TCGA (red) shows the cost of running the TCGA best-practices pipeline as re-implemented as a Toil workflow (for comparison). CGL-One-Sample/Node (cyan) shows the cost of running the revised Toil pipeline, one sample per node. CGL (blue) denotes the pipeline running samples across many nodes. CGL-Spot (green) is the same as CGL, but denotes the pipeline run on the Amazon spot market. The slight rise in cost per sample at 32,000 cores was due to a couple of factors: aggressive instance provisioning directly affected the spot price (dotted line), and saving *bam* and *bedGraph* files for each sample. (**b**) Tracking number of cores during the recompute. The two red circles indicate where all worker nodes were terminated and subsequently restarted shortly thereafter.

managed within a workflow. Through this mechanism it integrates with Apache Spark[13] (**Supplementary Fig. 4**), and can be used to rapidly create containerized Spark clusters[14] (**Supplementary Note 3**).

Toil runs in multiple cloud environments including those of Amazon Web Services (AWS; Seattle, WA, USA), Microsoft Azure (Seattle, WA, USA), Google Cloud (Mountain View, CA, USA), OpenStack, and in HPC environments running GridEngine or Slurm and distributed systems running Apache Mesos[15–17] (Forest Hill, MD, USA). Toil can run on a single machine, such as a laptop or workstation, to allow for interactive development, and can be installed with a single command. This portability stems from pluggable backend APIs for machine provisioning, job scheduling and file management (**Supplementary Note 4**). Implementation of these APIs facilitates straightforward extension of Toil to new compute environments. Toil manages intermediate files and checkpointing through a 'job store', which can be an object store like AWS's S3 or a network file-system. The flexibility of the backend APIs allow a single script to be run on any supported compute environment, paired with any job store, without requiring any modifications to the source code.

Toil includes numerous performance optimizations to maximize time and cost efficiencies (**Supplementary Note 5**). Toil implements a leader/worker pattern for job scheduling, in which the leader delegates jobs to workers. To reduce pressure on the leader, workers can decide whether they are capable of running jobs immediately downstream to their assigned task (in terms of resource requirements and workflow dependencies). Frequently, next-generation sequencing workflows are I/O bound, owing to the large volume of data analyzed. To mitigate this, Toil uses file caching and data streaming. Where possible, successive jobs that share files are scheduled on a single node, and caching prevents the need for repeated transfers from the job store. Toil is robust to job failure because workflows can be resumed after any combination of leader and worker failures. This robustness enables workflows to use low-cost machines that can be terminated by the provider at short notice and are currently available at a significant discount on AWS and Google Cloud. We estimate the use of such preemptable machines on AWS lowered the cost of our RNA-seq compute job 2.5-fold, despite encountering over 2,000 premature terminations (**Fig. 2**). Toil also supports fine-grained resource requirements, enabling each job to specify its core, memory and local storage needs for scheduling efficiency.

Controlled-access data requires appropriate precautions to ensure data privacy and protection. Cloud environments offer measures that ensure stringent standards for protected data. Input files can be securely stored on object stores, using encryption, either transparently or with customer managed keys. Compute nodes can be protected by SSH key pairs. When running Toil, all intermediate data transferred to and from the job store can be optionally encrypted during network transmission and on the compute nodes' drives using Toil's cloud-based job store encryption. These and other security measures help ensure protection of the input data, and as part of a broader security plan, can be used to ensure compliance with strict data security requirements.

To demonstrate Toil, we used a single script to compute gene- and isoform-level expression values for 19,952 samples from four studies: The Cancer Genome Atlas (TCGA)[1], Therapeutically Applicable Research To Generate Effective Treatments (TARGET; https://ocg.cancer.gov/programs/target), Pacific Pediatric Neuro-Oncology Consortium (PNOC; http://www.pnoc.us/), and the Genotype Tissue Expression Project (GTEx)[18]. The data set comprised 108 terabytes. The Toil pipeline uses STAR[19] to generate alignments and read coverage graphs, and performs quantification using RSEM[20] and Kallisto[21] (**Fig. 1** and **Supplementary Note 6**). Processing the samples in a single batch on ~32,000 cores on AWS took 90 h of wall time, 368,000 jobs and 1,325,936 core hours. The cost per sample was $1.30, which is an estimated 30-fold reduction in cost, and a similar reduction in time, compared with the TCGA best-practices workflow[5]. We achieved a 98% gene-level concordance with the previous pipeline's expression predictions (**Figs. 1**,**2** and **Supplementary Fig. 1**). Notably, we estimate that the pipeline, without STAR and RSEM, could be used to generate quantifications for $0.19/sample with Kallisto. To illustrate portability, the same pipeline was run on the I-SPY2 data set[22] (156 samples) using a private HPC cluster, achieving similar per sample performance (**Supplementary Table 1**). Expression-level signal graphs (read coverage) of the GTEx data (7,304 samples from 53 tissues, 570 donors) are available from a UCSC Genome Browser[23] public track hub (**Supplementary Fig. 2**). Gene and isoform quantifications for this consistent, union data set are publicly hosted on UCSC Xena[9] and are available for direct access through a public AWS bucket (**Supplementary Fig. 3** and **Supplementary Note 7**).

Although there is an extensive history of open-source workflow-execution software[6–8], the shift to cloud platforms and the advent of standard workflow languages is changing the scale of analyses. Toil is a portable workflow software that supports open community standards for workflow specification and enables researchers to move their computation according to cost, time and data location. For example, in our analysis the sample data were intentionally co-located in the same region as the compute servers in order to provide optimal bandwidth when scaling to thousands of simultaneous jobs (**Supplementary Note 8**). This type of flexibility enables larger, more

comprehensive analyses. Further, it means that results can be reproduced using the original computation's set of tools and parameters. If we had run the original TCGA best-practices RNA-seq pipeline with one sample per node, it would have cost ~$800,000. Through the use of efficient algorithms (STAR and Kallisto) and Toil, we were able to reduce the final cost to $26,071 (**Supplementary Note 9**).

We have demonstrated the utility of Toil by creating one of the single largest, consistently analyzed, public human RNA-seq expression repositories, which we hope the community will find useful.

Editor's note: This article has been peer-reviewed.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by (BD2K) the National Human Genome Research Institute of the National Institutes of Health award no. 5U54HG007990 and (Cloud Pilot) the National Cancer Institute of the National Institutes of Health under the Broad Institute subaward no. 5417071-5500000716. The UCSC Genome Browser work was supported by the NHGRI award 5U41HG002371 (Corporate Sponsors). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or our corporate sponsors.

AUTHOR CONTRIBUTIONS

J.V., A.A.R. and B.P wrote the manuscript. J.V., A.A.R., A.N., J.A., C.K., J.N., H.S., P.A., J.P., A.D.D., B.O. and B.P. contributed to Toil development. F.A.N. and A.M. contributed to Toil-Spark integration. J.V. wrote the RNA-seq pipeline and automation software. M.H. and C.B. contributed WDL and cloud support. P.A. and S.Z. contributed CWL support. J.Z., B.C. and M.G. hosted quantification results on UCSC Xena. K.R. hosted GTEx results in UCSC Genome Browser. W.J.K., J.Z., S.Z., G.G., D.A.P., A.D.J., M.C., D.H. and B.P. provided scientific leadership and project oversight.

Data availability. Data are available from this project at the Toil xena hub (https://genome-cancer.soe.ucsc.edu/proj/site/xena/datapages/?host=https://toil.xenahubs.net).

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

John Vivian[1], Arjun Arkal Rao[1], Frank Austin Nothaft[2,3], Christopher Ketchum[1], Joel Armstrong[1], Adam Novak[1], Jacob Pfeil[1], Jake Narkizian[1], Alden D Deran[1], Audrey Musselman-Brown[1], Hannes Schmidt[1], Peter Amstutz[4], Brian Craft[1], Mary Goldman[1], Kate Rosenbloom[1], Melissa Cline[1], Brian O'Connor[1], Megan Hanna[5], Chet Birger[5], W James Kent[1], David A Patterson[2,3], Anthony D Joseph[2,3], Jingchun Zhu[1], Sasha Zaranek[4], Gad Getz[5], David Haussler[1] & Benedict Paten[1]

[1]Computational Genomics Lab, UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, California, USA. [2]AMP Lab, University of California Berkeley, Berkeley, California, USA. [3]UC Berkeley ASPIRE Lab, Berkeley, California, USA. [4]Curoverse, Somerville, Massachusetts, USA. [5]Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA.
e-mail: benedict@soe.ucsc.edu

1. Weinstein, J.N. *et al. Nat. Genet.* **45**, 1113–1120 (2013).
2. Zhang, J. et al. Database. http://dx.doi.org/10.1093/database/bar026 (2011)
3. Siva, N. *Lancet* **385**, 103–104 (2015).
4. McKenna, A. *et al. Genome Res.* **20**, 1297–1303 (2010).
5. UNC Bioinformatics. TCGA mRNA-seq pipeline for UNC data. https://webshare.bioinf.unc.edu/public/mRNAseq_TCGA/UNC_mRNAseq_summary.pdf (2013).
6. Albrecht, M., Michael, A., Patrick, D., Peter, B. & Douglas, T. in *Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies* (SWEET '12) 1. ACM (Association of Computing Machinery. http://dx.doi.org/10.1145/2443416.2443417 (2012).
7. Bernhardsson, E. & Frieder, E. Luigi. *Github* https://github.com/spotify/luigi (2016).
8. Goecks, J., Nekrutenko, A. & Taylor, J. *Genome Biol.* **11**, R86 (2010).
9. UCSC. Xena http://xena.ucsc.edu (2016).
10. Amstutz, P. Common workflow language. *Github* https://github.com/common-workflow-language/common-workflow-language (2016).
11. Frazer, S. Workflow description language. *Github* https://github.com/broadinstitute/wdl (2014).
12. Vivian, J. Toil scripts. *Github* https://github.com/BD2KGenomics/toil-scripts/tree/master/src/toil_scripts (2016).
13. Apache Software Foundation. Apache Spark http://spark.apache.org/ (2017).
14. Massie, M. *et al.* ADAM: genomics formats and processing patterns for cloud scale computing. University of California, Berkeley, Technical Report No. UCB/EECS-2013-207 (2013).
15. Gentzsch, W. in *Proceedings First IEEE/ACM International Symposium on Cluster Computing and the Grid* 35–36 http://dx.doi.org/10.1109/ccgrid.2001.923173 (IEEE, 2001).
16. Yoo, A.B., Jette, M.A. & Mark, G. in *Lecture Notes in Computer Science* 44–60 (2003) Springer, Berlin, Heidelberg.
17. Apache Software Foundation. Apache Mesos http://mesos.apache.org/
18. GTEx Consortium. *Science* **348**, 648–660 (2015).
19. Dobin, A. *et al. Bioinformatics* **29**, 15–21 (2013).
20. Li, B. & Dewey, C.N. *BMC Bioinformatics* **12**, 323 (2011).
21. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. *Nat. Biotechnol.* **34**, 525–527 (2016).
22. Barker, A.D. *et al. Clin. Pharmacol. Ther.* **86**, 97–100 (2009).
23. Kent, W.J. *et al. Genome Res.* **12**, 996–1006 (2002).

Nextflow enables reproducible computational workflows

To the Editor:

The increasing complexity of readouts for omics analyses goes hand-in-hand with concerns about the reproducibility of experiments that analyze 'big data'[1–3]. When analyzing very large data sets, the main source of computational irreproducibility arises from a lack of good practice pertaining to software and database usage[4–6]. Small variations across computational platforms also contribute to computational irreproducibility by producing numerical instability[7], which is especially relevant to high-performance computational (HPC) environments that are routinely used for omics analyses[8]. We present a solution to this instability named Nextflow, a workflow management system that uses Docker technology for the multi-scale handling of containerized computation.

In silico workflow management systems are an integral part of large-scale biological analyses. These systems enable the rapid prototyping and deployment of pipelines that combine complementary software packages. In genomics the simplest pipelines, such as Kallisto and Sleuth[9], combine an RNA-seq quantification method with a differential expression module (**Supplementary Fig. 1**). Complexity rapidly increases when all aspects of a given analysis are included. For example, the Sanger Companion pipeline[10] bundles 39 independent software tools and libraries into a genome annotation suite. Handling such a large number of software packages, some of which may be incompatible, is a challenge. The conflicting requirements of frequent software updates and maintaining the reproducibility of original results provide another unwelcome wrinkle. Together with these problems, high-throughput usage of complex pipelines can also be burdened by the hundreds of intermediate files often produced by individual tools. Hardware fluctuations in these types of pipelines, combined with poor error handling, could result in considerable readout instability.

Nextflow (http://nextflow.io; **Supplementary Methods**, **Supplementary Note** and **Supplementary Code 1**) is designed to address numerical instability, efficient parallel execution, error tolerance, execution provenance and traceability. It is a domain-specific language that enables rapid pipeline development through the adaptation of existing pipelines written in any scripting language.

We present a qualitative comparison between Nextflow and other similar tools in **Table 1** (ref. 11). We found that multi-scale containerization, which makes it possible to

Part II

Nonparametric Bayesian Models for Precision Pediatric Oncology

Introduction

The Treehouse compendium consists of cancer gene expression profiles. One of the limitations of gene expression outlier analysis is that it is not able to identify cancer gene expression that is common to many of the samples in the compendium. For example, during the PDX analysis, I found that the model overestimated the threshold for FOXM1 outliers because many of the samples in the compendium were expressing FOXM1 at an elevated level. Every patient that is analyzed through the Treehouse approach is added to the Treehouse compendium. Therefore, as the number of outliers increases, the threshold for identifying new outliers also increases (Figure 2.1). As this method is not sustainable, I propose a mixture model to classify patients into normal and elevated expression distribution.

I observed that many of the genes in the Treehouse outlier analysis results had several peaks and could be modeled as a mixture of Gaussian distributions. After reviewing these genes further, I found that many of these genes were not detectable by Treehouse outlier analysis because the multimodal pattern led to a high variance when modeled as a single Gaussian distribution. By applying a Dirichlet process mixture model to the data, I was able to resolve complex gene expression distributions to identify clinically relevant pathways. The pipeline is named hydra because it reveals the many ways in which a cancer subtype may present itself. Identifying these manifestations of cancer gene expression data will accelerate the clinical application of this data because it highlights biological signals and reduces unexplained variance.

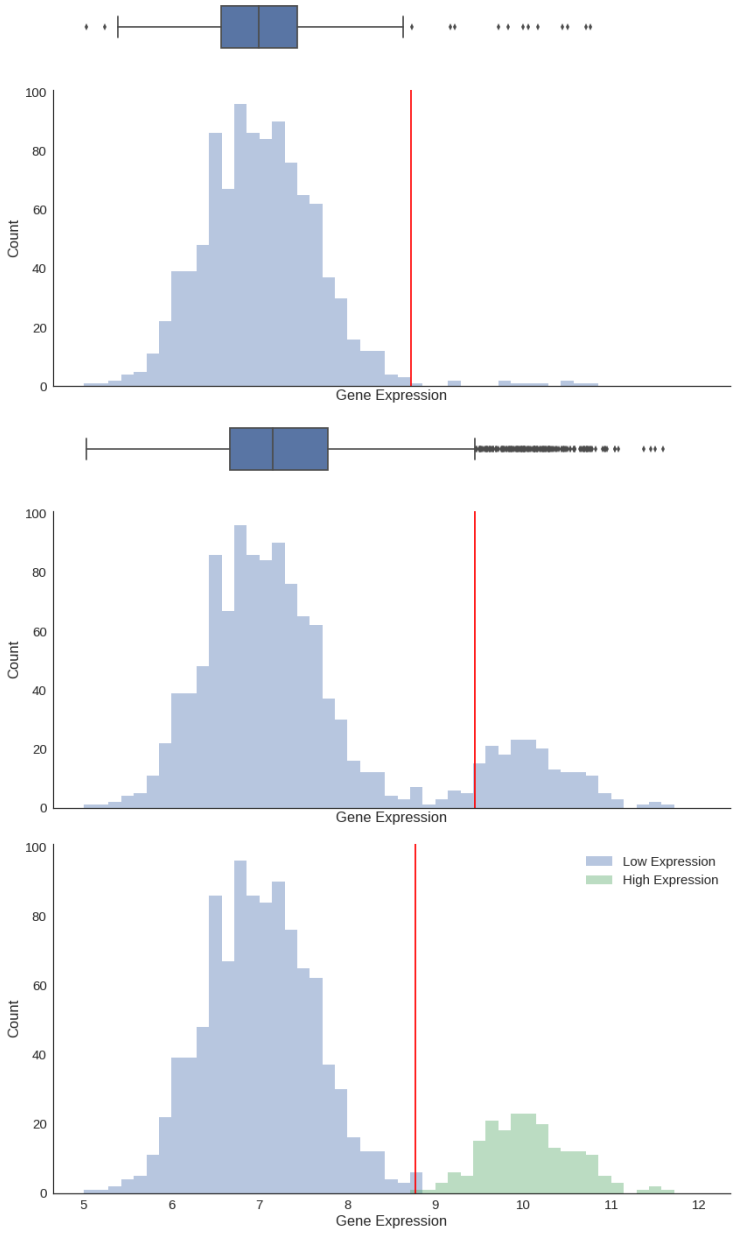


Figure 2.1: Outlier analysis becomes less sensitive as you increase the number of outliers in the compendium. One of the goals of the Treehouse initiative is to increase the size of the compendium, but as you increase the number of outliers, the sensitivity for identifying additional outliers decreases. The red line marks the threshold for identifying abnormal gene expression. The first two distributions use Treehouse outlier analysis, but the last distribution uses a two-component Gaussian mixture model to infer the normal and over-expression distributions.

Chapter 3

UCSC Treehouse Outlier Analysis Leads to the Discovery of Multimodal Expression Distributions

Analysts responsible for reviewing UCSC Treehouse outlier results have many opportunities to investigate expression distributions, especially distributions for druggable genes. As part of the UCSC Treehouse group, I personally reviewed many genes that did not have a unimodal Gaussian distribution, violating the underlying assumption of Treehouse outlier analysis. This led me to study other statistical models for analyzing gene expression for precision pediatric oncology research. After reviewing the literature, I found that mixture models have the flexibility to infer complex expression distributions and recent advances in Bayesian inference algorithms makes it possible to apply these models to large gene expression datasets.

The Treehouse analysis estimates pan-cancer and pan-disease outlier thresholds. The

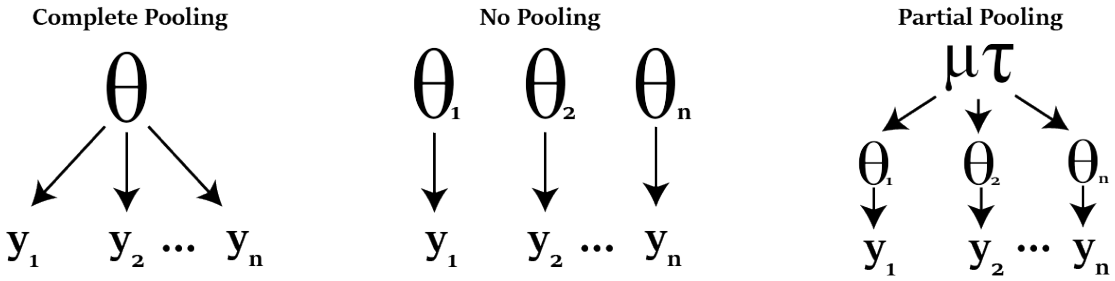


Figure 3.1: Models for the Treehouse analysis. Treehouse pan-cancer analysis is an example of the complete pooling model. In a complete pooling model, distinct groups of data are not modeled individually. Pan-cancer analysis does not account for different data features like the age, cancer type, and gender. Pan-disease analysis is a form of no-pooling model where each disease is modeled separately without considering information learned from other cancer types. A hierarchical models is a compromise between the complete and no pooling model. In a hierarchical model, separate parameters are learned for each data group while also sharing information through prior distributions on the group specific parameters.

pan-cancer analysis uses the entire compendium to estimate the outlier threshold. Pan-disease analysis uses expression from samples that share a diagnosis with the patient's TumorMap cohort to estimate the outlier threshold. I propose a hierarchical model which uses the entire compendium to estimate the predicted distribution of gene expression values for a specific disease. Additionally, the hierarchical model will be expanded to include predictors that identify differences in gene expression related to biological features of the data.

Pan-cancer analysis is a form of complete pooling model. Complete pooling models use one set of parameters to describe all the variation in the data (Figure 6.1). Complete pooling models overestimate the uncertainty in the distribution because it does not account for known variation in the data. For example, some genes are expressed in a tissue-specific manner, so subsetting the data by tissue will improve estimates for that gene. Pan-cancer analysis overestimates the uncertainty in gene expression for some genes, which makes it more difficult to

identify gene expression outliers. Pan-cancer analysis is appropriate for tightly controlled genes that are uniformaly expressed in different tissues, but is may be less sensitive to tissue-specific expression.

Pan-disease analysis uses the TumorMap defined disease cohort to find disease-specific gene expression outliers. The goal of pan-disease analysis is to find abnormal gene expression relative to patients with similar overall gene expression profiles. Pan-disease analysis is a form of no-pooling model because samples outside of the disease cohort are not used in the analysis (Figure 6.1). In the no-pooling model, each data cluster is modeled separately, so information is not shared across data clusters. The no-pooling model ignores similarities across data clusters that can be used to make more accurate estimates. While gene expression is tissue-specific, the range of possible gene expression levels across all tissues is constrained by the limits of human biology. Therefore, information from distinct cancers can still inform a disease-specific analysis. This is particularly important for pediatric data because estimates of pediatric expression are susceptible to large errors due to the lack of samples in the compendium [11, 37].

Complete pooling maximally underfits and no-pooling maximally overfits data, but hierarchical modeling strikes a balance between the two [30]. In a hierarchical model, each data cluster is modeled separately, but information is shared across levels of the hierarchy (Figure 6.1). Hierarchical models encode the collective knowledge about the system and can be used to study macro-level parameters at the population level and micro-level parameters about specific data clusters [30, 37]. These are ideal features for modeling and learning from the Treehouse compendium data.

A mixture model is an extension of a hierarchical model where the labels that encode

distribution membership are learned from the data. In a traditional hierarchical model, the data groups are known *a priori*, but in the mixture model framework, the labels need to be inferred from the data. The labels are identified by maximizing the probability that each sample belongs to either the normal or elevated gene expression distributions. The additional complexity is in inferring the latent variables that encode distribution membership for each sample. MCMC sampling tools, including STAN, provide tools for modeling mixture distributions. Future work will unify the Treehouse hierarchical model with a mixture modeling component to identify unobserved variables associated with cancer expression.

As a preliminary investigation into the different types of gene expression distributions represented in the Treehouse compendium, I clustered gene expression distributions using K-means clustering for a cohort of acute lymphoblastic leukemia patients (Figure 3.2). Using 10 clusters, I identified five different distributions. I found approximately normal, bi-modal, exponential, left-skewed, and right-skewed distributions. Bi-modal expression included important genes commonly identified in Treehouse gene expression analysis. Examples of genes with bi-modal distributions for acute lymphoblastic leukemia are AKT1, BTK, CREB1, FLI1, JAK2/3, and MYC. These distributions require special modeling considerations to properly identify biologically meaningful expression differences.

A Bayesian hierarchical mixture model is able to identify over-expression of FOXM1 for PDX PSS078 when outlier detection was not (Figure 3.3). The mixture model was able to decompose the pan-cancer distribution into a low and high expression distribution. The mixture model would have classified PSS078 FOXM1 expression into the high-expression component. This would have identified FOXM1 as a potential drug target, which would have been helpful

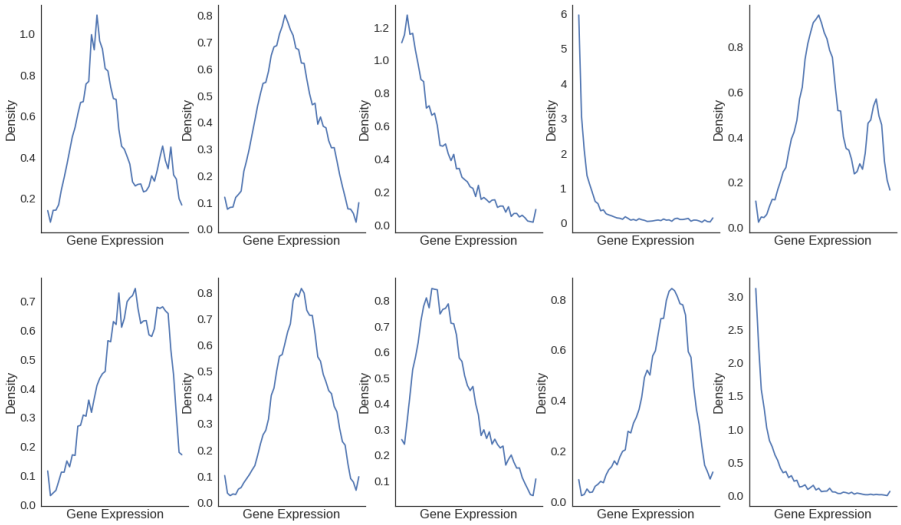


Figure 3.2: Clusters of gene expression profiles for the Treehouse acute lymphoblastic leukemia patients. Gene expression was centered and normalized by two standard deviations. The histograms were then clustered using K-means clustering (k=10). The Treehouse compendium includes uni- and bi-modal distributions as well as exponential distributions. Careful modeling of these distributions may yield biological insight.

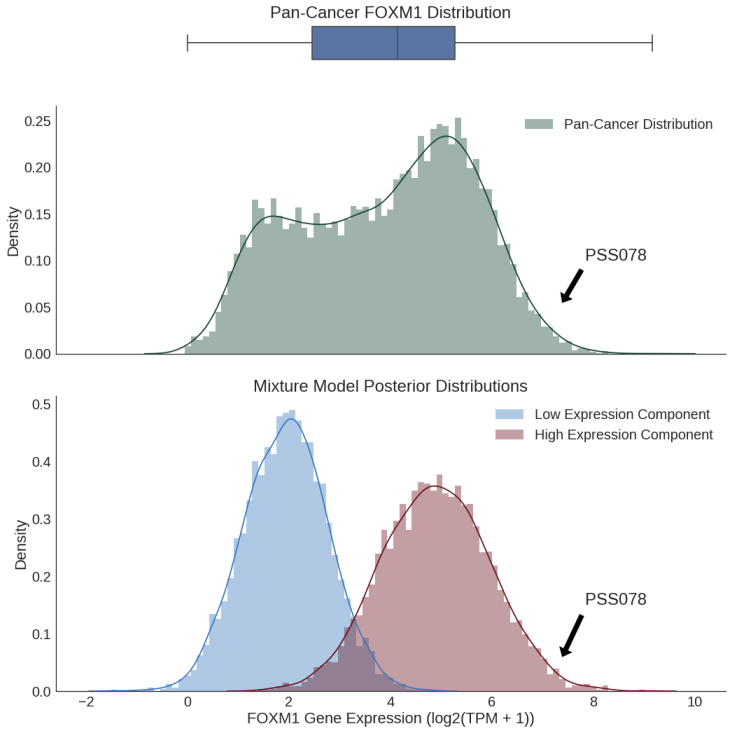
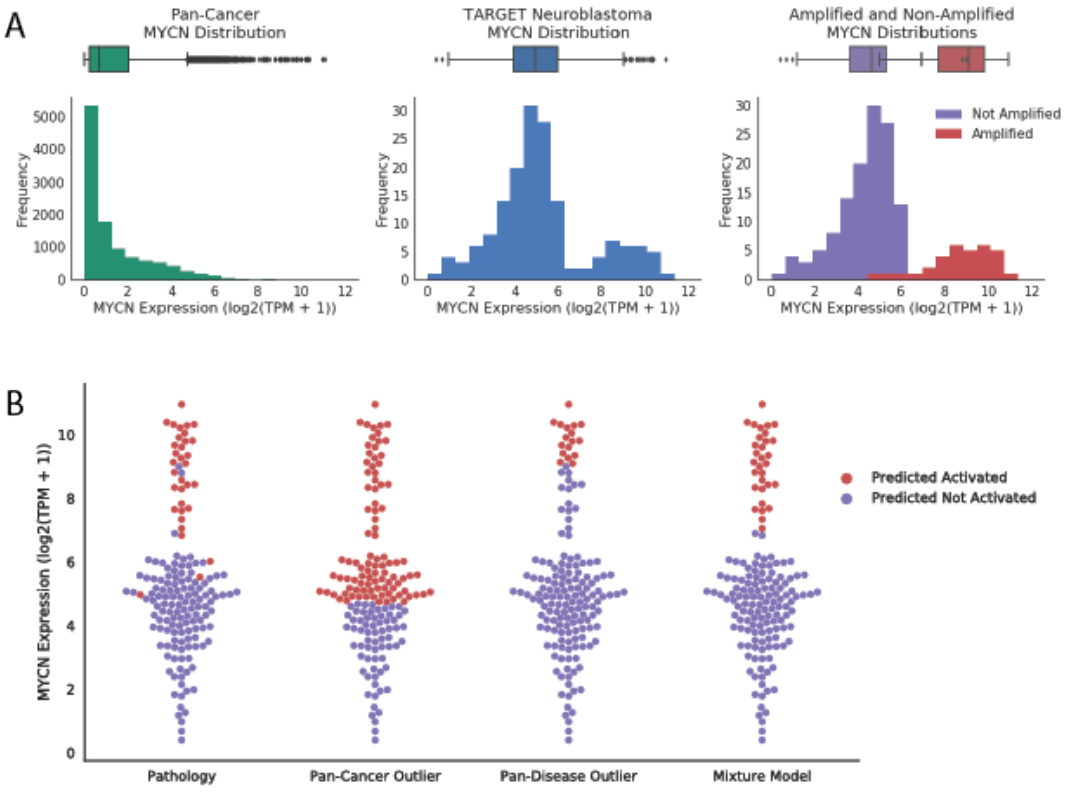


Figure 3.3: Known cancer genes, such as FOXM1, have a bi-modal distribution and are difficult to detect by outlier analysis. A hierarchical mixture model learns which samples come from the low expressed or high expressed modes and can be used to classify FOXM1 over-expression. The PDX PSS078 had a FOXM1 amplification that was not detected by outlier analysis, but the mixture model classifies PSS078 expression with the high expression component of the distribution.

for identifying drug targets and activated pathways.

Hierarchical modeling provides an opportunity to use all of the Treehouse compendium to learn tissue specific expression, biological effects on gene expression, and predict expected gene expression for new patients. A hierarchical model is a tool for detecting drug targets but it is also an interpretable model for learning about general molecular features of pediatric cancers. Hierarchical modeling is a well developed statistical framework that can be applied to childhood cancer research and adapted into a diagnostic tool.

When compared to outlier anlaysis, the mixture modeling approach achieved a significant improvement in performance. UCSC Treehouse outlier analysis has two modes: pan-cancer and pan-disease analysis. We used this approach to study the MYCN transcription factor, which is an important biomarker in neuroblastoma (3.4). We found that pan-cancer analysis underestimates the threshold for overexpression, leading to false positives. The pan-disease analysis overestimates the threshold for overexpression leading to false negatives. The mixture modeling approach is better able to discern the threshold for overexpression, leading to a 20-30% increase in F1 score.



C

	Pan-Cancer	Pan-Disease	Mixture Model
Accuracy	62%	90%	96%
Precision	35%	100%	94%
Recall	100%	52%	88%
F1	52%	68%	91%

Figure 3.4: Differential expression of cancer biomarkers yields multimodal distributions. Application of a Gaussian mixture model performs better at isolating expression subtypes than pan-cancer and pan-disease outlier analysis.

Chapter 4

Hydra: A Bayesian Nonparametric Approach for Identifying Cancer Gene Expression Subtypes

Introduction

The hydra pipeline runs in two modes. The first mode applies the Dirichlet process mixture model analysis to curated gene sets. This analysis is useful for identifying known gene expression signatures. Comparing the hydra method to widely-used pathway enrichment tools found that the hydra approach is more sensitive at detecting clusters driven by multimodal expression patterns. The second mode identifies the enrichment of pathways across all multimodally expressed genes and can be used to identify novel pathway expression that may not have been discovered before. This approach has been helpful for identifying important signals associated with cancer subtypes, tumor microenvironment expression, and complex tissue sam-

ples. We have found that multimodally expressed genes better separate known clinical subtypes of neuroblastoma using the UCSC TumorMap tool. We have further subtyped a cohort of pediatric neuroblastoma samples and identified differential expression of tumor microenvironment signatures, including markers of the adaptive immune response and fibroblasts. This information is important for identifying opportunities for eradicating tumors with an immunotherapy treatment approach. Lastly, we have been able to use this tool to identify complex tissue samples that can influence the interpretation of cancer gene expression data. Osteosarcoma is a pediatric bone tumor, but hydra analysis revealed a strong skeletal muscle signature in a subset of samples. Through collaborations at UCSF, I validated this signal and confirmed that bone samples with this signature contain contaminating muscle tissue. The hydra method is a flexible analysis tool that combines the capability to detect differential pathway expression with the capability to perform clustering to identify the biological signals that differentiate cancer gene expression profiles.

Hydra: A mixture modeling framework for subtyping pediatric cancer cohorts using multimodal gene expression signatures

Jacob Pfeil[1,2*], Lauren M. Sanders[1,2,3], Ioannis Anastopoulos[1,2], A. Geoffrey Lyle[2,3], Alana S. Weinstein[1,2], Yuanqing Xue[1,2], Andrew Blair[1,2], Holly C. Beale[2,3], Alex Lee[4], Stanley G. Leung[4], Phuong T. Dinh[4], Avanthi Tayi Shah[4], Marcus R. Breese[4], W. Patrick Devine[5], Isabel Bjork[2], Sofie R. Salama[1,2,6‡], E. Alejandro Sweet-Cordero[4‡], David Haussler[1,2,6‡], Olena Morozova Vaske[2,3‡]

1 Department of Biomolecular Engineering, University of California, Santa Cruz, California, United States of America
2 Genomics Institute, University of California, Santa Cruz, California, United States of America
3 Department of Molecular, Cell and Developmental Biology, University of California, Santa Cruz, California, United States of America
4 Department of Pediatrics, Division of Hematology and Oncology, University of California, San Francisco, California, United States of America
5 Department of Anatomic Pathology, University of California, San Francisco, California, United States of America
6 Howard Hughes Medical Institute, University of California, Santa Cruz, California, United States of America

‡Senior authorship
* Corresponding author: jpfeil@ucsc.edu

Abstract

Precision oncology has primarily relied on coding mutations as biomarkers of response to therapies. While transcriptome analysis can provide valuable information, incorporation into workflows has been difficult. For example, the relative rather than absolute gene expression level needs to be considered, requiring differential expression analysis across samples. However, expression programs related to the cell-of-origin and tumor microenvironment effects confound the search for cancer-specific expression changes. To address these challenges, we developed an unsupervised clustering approach for discovering differential pathway expression within cancer cohorts using gene expression measurements. The hydra approach uses a Dirichlet process mixture model to automatically detect multimodally distributed genes and expression signatures without the need for matched normal tissue. We demonstrate that the hydra approach is more sensitive than widely-used gene set enrichment approaches for detecting multimodal expression signatures. Application of the hydra analysis framework to small blue round cell tumors (including rhabdomyosarcoma, synovial sarcoma, neuroblastoma, Ewing sarcoma, and osteosarcoma) identified expression signatures associated with changes in the tumor microenvironment. The hydra approach also identified an association between *ATRX* deletions and elevated immune marker expression in high-risk neuroblastoma. Notably, hydra analysis of all small blue round cell tumors revealed similar subtypes, characterized by changes to infiltrating immune and stromal expression signatures.

Author summary

Pediatric cancers generally have few somatic mutations. To increase the number of actionable treatment leads, precision pediatric oncology initiatives also analyze tumor gene expression patterns. However, currently available approaches for gene expression data analysis in the clinical setting often use arbitrary thresholds for assessing overexpression and assume gene expression is normally distributed. These methods also rely on reference distributions of related cancer types or normal samples for assessing expression distributions. Often adequate normal samples are not available, and comparing matched cancer cohorts without accounting for subtype expression overestimates the uncertainty in the analysis. We developed a computational framework to automatically detect multimodal expression distributions within well-defined disease populations. Our analysis of small blue round cell tumors (including rhabdomyosarcoma, synovial sarcoma, neuroblastoma, Ewing sarcoma and osteosarcoma) discovered a significant number of multimodally expressed genes. Multimodally expressed genes were associated with proliferative signaling, extracellular matrix organization, and immune signaling pathways across cancer types. Expression signatures correlated with differences in patient outcomes for *MYCN* non-amplified neuroblastoma, osteosarcoma, and synovial sarcoma. The low mutation rate in pediatric cancers has led some to suggest that pediatric cancers are less immunogenic. However, our analysis suggests that immune infiltration can be identified across small blue round cell tumors. Thus, further research into modulating immune cells for patient benefit may be warranted.

Introduction

Large cancer sequencing projects, including The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET), have facilitated the development of cancer gene expression compendia [1–6], but these compendia often lack expression data from corresponding normal tissue. Without the normal comparator, Hoadley et al. (2018) found that cell-of-origin signals drive integrative clustering of TCGA data. Strong cell-of-origin and tumor microenvironment (TME) signals may also complicate the interpretation of gene expression results for precision oncology applications, so careful modeling of the data is necessary to infer accurate conclusions.

The TME includes tumor cells, stromal fibroblasts, immune cells, and vasculature [7]. Similarities in TME composition across tumor samples have led to the identification of TME states (e.g. inflamed, immune-excluded, immune-desert). While these states are dynamic, they can still shed light on the immunogenicity of tumor cells and correlate with response to cancer immunotherapies [8]. The TME cellular composition can be inferred from tumor RNA-Seq data since host cell RNA is sequenced along with the cancer cell RNA. Tumor progression and response to therapies is associated with features of the TME. Therefore, targeting the TME therapeutically may improve treatment outcomes in some cancers.

Immunotherapies that activate the host immune system to eradicate tumors have been effective in treating several cancer types, particularly cancers with a high mutation burden [9, 10]. Pediatric cancers tend to have fewer mutations than adult cancers, and while there has been limited testing of immunotherapies in pediatric cancer patients, the currently available data suggest lower response rates than adult cancers [11, 12]. However, improved immune subtyping of pediatric cancers may identify subsets of patients that are candidates for powerful immunotherapies. In addition to infiltrating immune cells, cancer-associated fibroblasts (CAFs) assist in extracellular matrix remodeling and activation of growth factor signaling. CAFs facilitate tumor growth,

metastasis, and resistance to some therapies, so identification of CAF functions within a tumor may also facilitate clinical decision making. Methods are needed to both infer and characterize gene expression subtypes that correlate with tumor microenvironment states to accelerate the development of personalized therapies for pediatric cancers.

Tumor/normal differential expression analysis in which a cohort of tumor tissues is compared to corresponding normal tissue samples is an effective approach for identifying gene expression biomarkers [13–15], but it is often not possible to conduct this analysis in a clinical setting. Sufficient biological and technical replicates are limited by tumor tissue availability, and healthy neighboring tissue often cannot be isolated. In addition, for many pediatric cancers, the cell-of-origin, and thus the appropriate reference normal tissue, is not known. Besides differential expression analysis, single-sample pathway analysis can be used to identify upregulation of biological gene sets in tumor subtypes. Among the most widely used pathway analysis approaches is gene set enrichment analysis (GSEA) [16, 17]. GSEA identifies coordinated expression of pathway genes using gene ranks and a Kolmogorov-Smirnov-like test statistic. GSEA is usually performed on differentially expressed genes to compare two cohorts or phenotypes, but single-sample GSEA is also available when there is not an obvious comparator. GSEA uses curated pathway gene sets like those in the Molecular Signatures Database (MSigDB) [18].

Cancer gene expression subtypes are traditionally identified using unsupervised clustering methods such as consensus clustering analysis [19–21]. These methods are generally underpowered because the number of genes greatly exceeds the number of samples. Dimensionality reduction approaches such as Principal Component Analysis (PCA) have been found to underestimate the dimensionality of gene expression data [22]. Lenz at al. (2016) found two cases in which PCA fails to identify a biological signal: when the size of the cluster is small and when the effect size is small. Lenz et al. (2016) suggests investigating multimodally expressed genes to improve identification of cancer subtypes. Cancer subtypes naturally lead to multimodal expression patterns because each subtype expresses a correlated set of genes at different expression levels. Expression subtypes may result from dysregulated pathway expression within cancer cells, but another source of multimodal expression comes from varying amounts of infiltrating immune and stromal cells in the TME.

Gaussian mixture models are a powerful class of unsupervised clustering algorithms that can be used to detect multimodally expressed genes [23–25]. A Gaussian mixture model is appropriate when the expression data can be modeled as a mixture of two or more Gaussian distributions [26]. One limitation of Gaussian mixture models in this context is that the number of clusters in the data is often not known beforehand, so a parameter search must be used to identify the best-performing model. However, this is a computationally expensive approach. This problem can be overcome by placing a Dirichlet process prior on the number of expression clusters. The number of clusters is then inferred while fitting the mixture model using Markov chain Monte Carlo (MCMC) sampling [26]. This approach has not been widely used in clinical cancer research because these algorithms are still computationally expensive, but recent advances in Bayesian variational inference have made this approach scalable for precision oncology applications [27].

Here, we present the hydra framework for identifying clinically relevant expression subtypes and classifying N-of-1 tumor samples using learned models. We provide an overview of the hydra framework, assess performance for detecting differential pathway expression, and apply the framework to better understand expression patterns in high-risk neuroblastoma and other small blue round cell tumors. We apply the learned models trained on publicly available cancer gene expression data to the N-of-1 setting and show that this framework can identify distinct immune and stromal expression

signatures that differentiate pediatric cancer samples. Finally, we identify recurrent tumor microenvironment signatures across pediatric cancer types associated with differences in patient outcomes.

Materials and methods

Dirichlet process gaussian mixture model

Traditional parametric models, like the finite mixture model, use a fixed number of parameters (i.e. number of clusters). Over- or underfitting can occur when the parametric model does not reflect the underlying data [28]. Unlike the finite mixture model, the Dirichlet process mixture model (DPMM) represents a theoretically infinite number of clusters and can adapt the number of clusters based on prior belief and the data [26, 28, 29].

The Dirichlet process (DP) is an infinite dimensional extension of the Dirichlet distribution [30] and is commonly used as a prior distribution for infinite mixture models [31, 32]. The Dirichlet process has two parameters: the concentration parameter α and centering distribution H. The concentration parameter α, where $\alpha \in \mathbb{R}^+$, controls the extent to which samples from the DP resemble the centering distribution H. We model gene expression as a multivariate Gaussian distribution, so our centering distribution is a normal-Wishart distribution ($\mathcal{NW}_0$).

We briefly describe the stick-breaking construction of the Dirichlet process $G \sim DP(\alpha, H)$. Consider a stick of unit length. To generate an infinite number of mixing weights $\pi_1, \pi_2, ..., \pi_k$ for the DPMM, first break a stick of unit length at $\nu \in [0, 1]$ where ν is sampled from a Beta distribution, and set π_1 to be the length of the first piece. We repeat this process using the remainder of the stick for each π_k. The DP is truncated to the number of clusters K, which was shown to accurately approximate the infinite posterior for large K [26, 28, 30, 33–35].

$$\nu \sim \text{Beta}(1, \alpha) \tag{1}$$

$$\pi_k = \nu_k \prod_{l=1}^{k-1} (1 - \nu_l) \tag{2}$$

Next, we sample the parameters from the centering distribution H weighted by the mixing components. If we consider a probability space Θ where $\theta_k^* \in \Theta$, then H is a measure on the partitions of Θ. For our application, we will partition the parameter space Θ into finite, measurable partitions $B_1, B_2, ..., B_k$.

$$\theta_k^* \sim H \tag{3}$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \tag{4}$$

$$(G(B_1), G(B_2), ..., G(B_k)) \sim \text{Dir}(\alpha H(B_1), \alpha H(B_2), ..., \alpha H(B_k)) \tag{5}$$

This construction generates the marginal of the Dirichlet process, which follows a Dirichlet distribution. Samples from the marginal distribution are finite, discrete, and sum to 1 [30]. Next, we outline how the DPMM groups gene expression samples x_i under cluster-specific parameters μ_{z_i} and Σ_{z_i} where $z_i \in 1, 2, ..., K$ is the cluster index.

$$x_i|\mu_{z_i}, \Sigma_{z_i} \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i}) \quad (6)$$
$$z_i|\pi \sim \text{Categorical}(\pi_1, \pi_2, ..., \pi_k) \quad (7)$$
$$\mu_{z_i}, \Sigma_{z_i}|G \sim G \quad (8)$$
$$G|\alpha, \mathcal{NW}_0 \sim DP(\alpha, \mathcal{NW}_0) \quad (9)$$

To improve our methods ability to scale to larger datasets, we incorporated the bnpy memoized online variational inference algorithm (moVB) [33] into our analysis framework. The moVB algorithm uses variational inference to approximate the posterior distribution and interleaves birth, merge, and delete moves to avoid local optima and remove redundant clusters [36]. We found that the moVB algorithm accurately identified the number of clustering on validation datasets (S1 Fig), whereas standard MCMC sampling procedures tended to overestimate the number of clusters.

Hydra method

We developed a Bayesian non-parametric clustering framework for identifying biological and technical variation in large cancer gene expression datasets without the need for a reference normal dataset. To our knowledge, this is the first reproducible and widely deployable implementation of a non-parametric mixture model framework designed to overcome the challenges of precision oncology gene expression analysis. The hydra pipeline is an open source software tool hosted on GitHub (`www.github.com/jpfeil/hydra`). A Docker container is available for deployment across environments (`https://hub.docker.com/r/jpfeil/hydra`).

The hydra framework contains three main command-line tools: *filter*, *enrich*, and *sweep* (Fig 1). The *filter* command is run first to isolate the multimodally expressed genes using a univariate Dirichlet Process Gaussian Mixture Model (DP-GMM). There are two methods for analyzing the resulting set of multimodally expressed genes. The *enrich* method, which subsets to the genes found to be significantly enriched in biological pathways, and the *sweep* method, which searches within user-defined gene sets for multimodal expression signatures. The underlying analysis routines can be accessed within the Docker using Jupyter notebooks to facilitate the development of user-defined workflows.

The *filter* command (Fig 1B) takes an expression matrix and filters the genes down to the multimodally expressed genes using the DP-GMM described above. We apply a DP-GMM to each gene, saving the model for genes with two or more expression clusters. This creates a directory of multimodally expressed gene models which can be used to predict differential expression in new samples. This analysis framework is a novel contribution to the precision medicine research community. Our approach has several beneficial properties. For example, training models on curated data sets and applying the models to new samples avoids the use of reference distributions, which overestimate the uncertainty in the analysis by not accounting for subtype expression. Furthermore, this approach identifies the set of most strongly differentially expressed genes within a disease context, which may enrich for potential biomarkers for precision medicine applications. The multimodally expressed genes are also used in downstream clustering analysis.

The *enrich* (Fig 1C) and *sweep* (Fig 1D) routines are two independent analyses to explore multimodal expression in cancer gene expression cohorts. In addition to identifying expression variation within a disease context, we also found that multimodally expressed genes that participate in a biological pathway tend to have correlated expression distributions. This insight facilitates the detection of multimodal

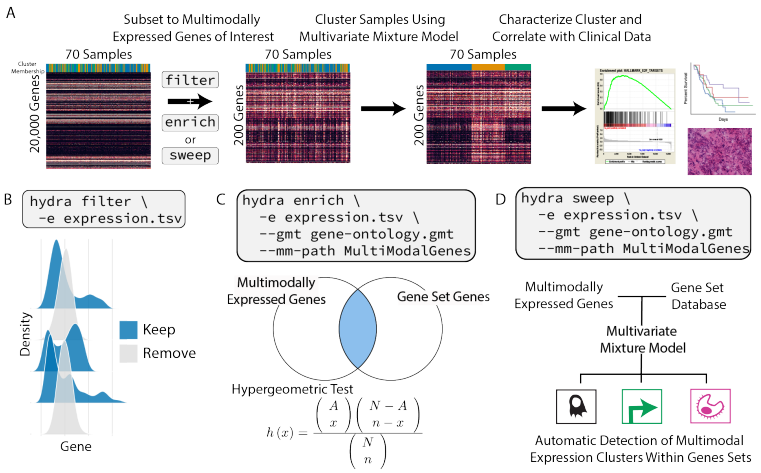


Fig 1. Overview of the hydra framework tools. A: Suggested workflow for applying hydra framework tools to identify clinically relevant gene expression subtypes. B: The hydra *filter* command removes unimodally distributed genes which greatly reduces the number of genes in downstream clustering analysis. C: The hydra *enrich* command takes the multimodally expressed genes and returns enriched gene sets. The enriched gene set genes are used for multivariate clustering of samples. D: The hydra *sweep* command looks for multivariate normal clusters within user-defined gene sets. This can be used for the automatic detection of clusters in large gene set databases.

expression signatures by enriching for genes that have multimodal expression distributions and participate in known biological processes. The hydra software comes prepackaged with popular gene sets, including the Molecular Signatures Database (MSigDB) [18], the Gene Ontology terms [37, 38], and the EnrichmentMap gene sets [39]. The gene set database is configurable, so additional gene sets can be added at runtime.

The *enrich* command uses a hypergeometric test [40] to discover enrichment of multimodally expressed genes within a user-defined database of gene sets. This creates a list of gene sets and a list of enriched gene set genes. The *enrich* method outputs a table of enriched gene sets while also clustering samples across the genes that participate in the enriched gene sets. The table of enriched gene sets may reveal surprising expression patterns and generate hypotheses for further investigation of tumor subtypes.

The implementation of the *enrich* method includes an important parameter known as the minimum component probability. The minimum component probability is the probability of placing a sample within the smallest expression cluster. This is an additional filter to remove multimodally expressed genes that influence a relatively small subset of tumor samples. This parameter gives the user the ability to subset the enriched genes to those that influence a greater number of patients. To aid in the exploration of minimum component thresholds, we implemented a *scan* sub-routine. The *scan* routine tunes the analysis with respect to the constraints of the available data (e.g. number of samples and number of genes), which is an important factor in pediatric cancer research since data is often difficult to obtain and so datasets are relatively small. We recommend setting this threshold such that the number of genes is less than the number of samples because otherwise the inference may become unstable [41].

The *sweep* routine identifies differentially expressed gene sets and can be used as an alternative to single-sample GSEA [16]. For each gene set, a multivariate DP-GMM is applied to determine if more than one expression cluster is present within the gene set. This approach is useful when curated gene sets are available for the disease of interest, but manual inspection of each gene set is not feasible. Reducing the genes to multimodally expressed genes facilitates the detection of differentially expressed gene sets. Existing gene set enrichment tools are known to under-perform when the expression is correlated [42], but our approach is designed to identify distinct correlation structures within gene expression datasets.

We have also implemented routines for cluster profiling and N-of-1 tumor analysis. These routines are accessible within the docker container using the Jupyter notebook command. Cluster profiling analysis of clusters derived from the *enrich* or *sweep* routines includes GSEA [43] to identify the pathway expression that characterizes each cluster. GSEA uses all available genes since it requires non-differentially expressed genes to assess the significance of an enrichment score. A t-statistic is calculated for each gene, comparing gene expression values of samples inside to those outside of a cluster. Cluster profiling GSEA uses the ranked gene-level t-statistics to determine gene set enrichment.

The N-of-1 tumor analysis routine classifies a new gene expression profile into one of the inferred clusters, calculates a gene-level z-score for that sample relative to the normalized expression distribution, and performs standard GSEA using a preranked list of z-score values [43]. This procedure can identify new gene expression signatures that may not be detectable using the entire expression cohort as a background reference distribution. This approach is another novel contribution to the field and may facilitate the identification of clinically relevant signatures that are being overlooked in current gene expression analyses.

Synthetic data generation and validation

We first tested the hydra framework's ability to detect differential pathway expression using synthetic cancer data. We compared hydra *sweep* to two widely used gene set enrichment tools: single-sample gene set enrichment analysis (ssGSEA) and gene set variation analysis (GSVA) [44–46]. Both methods are implemented in the GSVA R package [45]. In order to accurately model correlation structures within cancer cohorts, we modeled the synthetic cancer gene expression data as a multivariate Gaussian distribution. We used the TCGA glioblastoma multiforme (GBM) cohort (N=166) to model a background mean and covariance matrix for the synthetic data analysis. We chose TCGA GBM, a very different disease from those analyzed in the remainder of this manuscript, to avoid overfitting the hydra method to diseases of interest. This also enables us to demonstrate the flexibility of our method to analyze data from a variety of cancer genome sequencing projects.

This approach allowed us to model cancer gene expression data while also controlling for subtype-related expression variation. We downloaded the RSEM-quantified TPM normalized gene expression measurements from the UCSC Xena Browser [3]. We focus our analysis on normalized gene expression data because this data is more widely used in the cancer research community and fewer methods are available to analyze normalized counts. To reduce heteroscedasticity and the effect of outlier expression levels, we transformed the expression data to log2(TPM + 1) [47].

We defined an expression subtype as a subset of samples with a distinct expression mean and correlation structure compared to other samples within the disease cohort. To avoid biases in the synthetic data generation process, we used random sampling to select MSigDB gene sets for each subtype, the size of the subtype, and the correlation structure within the subtype. We randomly generated a covariance matrix for the cancer subtype expression data, but used the underlying covariance matrix of the

TCGA glioblastoma multiforme dataset for the background samples. We tested the effect of having 10% and 25% of genes within a gene set being differentially expressed (%DEG). In addition to these parameters, we tested a range of effect sizes: 0.25 (least different), 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, and 3.0 (most different). This process was repeated twice for each gene set to create synthetic training and test data, which resulted in the generation of 640 synthetic datasets.

We then applied the hydra framework using the hydra *sweep* command (Fig 1C), since this method is directly comparable to the single-sample GSEA methods. The mean expression filter removed any genes with a mean expression of fewer than 1.0 log2(TPM + 1). This avoids lowly-expressed genes that may have particularly noisy expression measurements. The prior on the hydra covariance matrix was the identity scaled by 2.0 and the prior on the number of clusters was set to 2 because we expect there to be an activated cluster and a baseline expression cluster. We set the over-expressing cluster to be the cluster with the largest L1 norm.

Pediatric cancer gene expression data

We downloaded pediatric cancer RNA-Seq data for neuroblastoma, osteosarcoma, Ewing sarcoma, alveolar rhabdomyosarcoma, and embryonal rhabdomyosarcoma from the UCSC Treehouse Compendium (`https://treehousegenomics.soe.ucsc.edu/public-data/`). This data was produced using the same RNA-seq pipeline, so potential computational batch effects are minimized [1, 6]. Clinical data for the TARGET neuroblastoma and osteosarcoma samples were obtained from the TARGET Data Matrix (https://ocg.cancer.gov/programs/target/data-matrix). We also analyzed a set of 58 synovial sarcoma microarray profiles with matching metastasis rate data [48].

TARGET neuroblastoma analysis

We applied each hydra tool to the TARGET *MYCN*-NA neuroblastoma cohort. We first obtained the multimodal gene models using the hydra *filter* tool. The hydra *filter* tool identified all genes with a multimodal expression pattern. We used the mean expression filter to remove genes that may have unstable measurements due to low transcript abundances. We excluded all genes with a mean expression value less than 1 log2(TPM + 1).

The hydra *sweep* command was applied to search for subtype expression within curated MSigDB gene sets. We included the hallmark (n=50), BioCarta (n=289), KEGG (n=186), PID (n=196), and Reactome (n=1499) genesets [18]. We include all signatures with a minimum component probability of 10%. For example, the smallest subtype cluster considered in this analysis had 7 samples, since the total number of samples was 70. We investigated relationships among differentially expressed gene sets by clustering the gene sets by their pairwise Jaccard index. This created a similarity network that was then visualized using the Gephi software tool [49].

The hydra *enrich* command identified correlated expression signatures using the enriched GO term genes (FDR < 0.01). The multivariate mixture model α concentration parameter was set to 5.0; the prior on the covariance matrix was set to the identity scaled by 2.0. The prior parameter for the number of clusters was set to 5. Our synthetic data analysis found that the signal decreases below an effect size of 1.0, so we use this parameter value for all following analyses. We used the hydra *scan* routine to search a range of minimum component probability thresholds (see Results) and found that a threshold/probability of 20% yielded the most clusters while keeping the number of genes ($p = 42$) below the number of samples ($n = 70$).

To validate tumor microenvironment expression subtypes, we correlated the hydra *enrich* expression clusters with the results of tumor microenvironment profiling tools xCell [50], CIBERSORT [51], and ESTIMATE [52]. We also compared the hydra *enrich* approach to state-of-the-art consensus clustering methods M3C [20] and k-means clustering using the Gap statistic to select the number of clusters [53]. Since these methods are influenced by the number of input genes, we tested a range of median absolute deviation (MAD) thresholds. The number of clusters was assumed to be the smallest statistically significant value.

Small blue round cell tumor analysis

We then compared the clustering patterns across *MYCN*-NA neuroblastoma, osteosarcoma, Ewing sarcoma, embryonal rhabdomyosarcoma, alveolar rhabdomyosarcoma, and synovial sarcoma. We applied the TumorMap dimensionality reduction method [5] to visualize clustering of the full small blue round cell tumor gene expression matrix. We then applied the hydra framework to explore expression variation within each disease. Each disease expression matrix had unique statistical properties including sample size and subtype variation. This required us to adapt the minimum probability threshold for each disease dataset using the *scan* routine. The Jupyter notebooks for exploring these datasets can be found on GitHub (`www.github.com/jpfeil/hydra-paper/analysis`). We used agglomerative clustering to investigate patterns in the top 10 enriched gene sets for each disease's expression subtypes.

Statistical analysis

A Kruskal-Wallis test was used to identify statistically significant differences across two or more groups, and a Mann-Whitney U test was used for pairwise tests using a Holm-Sidak correction for multiple hypothesis testing [54, 55]. We used the scipy [56] stats implementation of the Kruskal-Wallis test and the scikit-learn post hoc processing [57] implementation of pairwise Mann-Whitney U tests. Spearman rank and Pearson correlation values were calculated using the scipy library [55]. Survival analysis was done using the survminer package [58].

H&E slide preparation and pathologist review

Pediatric tumor samples were flash frozen, embedded in OCT, and 5μm cryosections were collected. Slides were hematoxylin and eosin (H&E) stained and imaged on a Leica DMi8, equipped with a HC PL APO 40x/0.85 NA objective and DFC7000T camera. H&E slides were reviewed by a licensed pathologist. Morphologic analysis was performed and the degree and type of inflammation estimated from the histologic sections. Grading of inflammation was either minimal (<10% of total nuclei consist of inflammatory cells) or moderate (20-30% of total nuclei consist of inflammatory cells). The type of inflammation (predominantly small mature lymphocytes or mixed inflammation consisting of small mature lymphocytes along with plasma cells and/or eosinophils) was noted for each tumor sample.

Results

Performance assessment using synthetic gene expression data

To assess how well hydra detects differentially expressed pathways as compared to common pathway enrichment approaches, we applied the hydra framework to

synthetically-generated cancer gene expression data. We generated synthetic cancer gene expression data based on the TCGA glioblastoma multiforme and the MSigDB Hallmark gene sets as described above. We tested a range of effect sizes and percent differentially expressed genes (%DEG) within the MSigDB gene sets. We generated receiver operator curves (ROC) and calculated the area under the receiver operator curve (AUC) for each analysis. Overall, the hydra pipeline outperformed the single-sample GSEA approaches with a mean AUC of 0.93 (95% CI: 0.91 - 0.95). ssGSEA had a mean AUC of 0.72 (95% CI: 0.71 - 0.74) and GSVA had a mean AUC of 0.67 (95% CI: 0.66 - 0.68) (Fig 2A).

We further investigated the performance of these methods by plotting the AUC against the effect size at 10 and 25%DEG (Fig 2B). The hydra method performed better across all effect sizes, achieving near perfect performance above an effect size of 2.0 and 0.75 at 10 and 25%DEG, respectively. ssGSEA and GSVA performed similarly at low effect sizes, but ssGSEA performed better than GSVA as the effect size increased. Overall, the hydra framework performed significantly better than these standard gene set enrichment approaches, particularly at low effect sizes. Therefore, the hydra approach is better suited for subtyping within a disease cohort when the effect sizes are smaller and fewer genes are differentially expressed.

We performed a runtime analysis comparing hydra *sweep*, ssGSEA, and GSVA for identifying a single differentially expressed gene set, since these methods are directly comparable. Training the hydra model was the most computationally expensive step, but the classification of new samples was very fast. The average runtime for the hydra *sweep* algorithm was similar to ssGSEA, but the hydra runtimes were more variable across effect-sizes and number of differentially expressed genes. The GSVA approach was faster than hydra *sweep* and ssGSEA, but GSVA performed worse on the synthetic data analysis than ssGSEA and hydra. We repeated the above analysis with an effect size of 1.0, a %DEG of 25%, and a range of sample sizes, including 50, 100, 200, 300, 400, 500, 1000 samples. The hydra *sweep* and GSVA methods scaled well to large sample sizes, but the ssGSEA runtime increased exponentially as the sample size increased (Fig 2C & D).

Hydra analysis of high-risk neuroblastoma

High-risk neuroblastoma is an aggressive disease and is resistant to intensive therapy. Further subtyping of high-risk neuroblastoma may identify novel therapeutic targets and improve risk stratification. We hypothesized that unsupervised clustering of multimodally expressed genes associated with enriched Gene Ontology terms would identify expression subtypes of high-risk neuroblastoma tumors. TumorMap analysis [5] showed that the *MYCN*-non-amplified (*MYCN*-NA) neuroblastoma samples clustered separately from *MYCN*-amplified (*MYCN*-A) and stage 4S neuroblastoma samples (S2 Fig). We focused on the *MYCN*-NA neuroblastoma tumor samples because this is the largest set of samples (N=70) and variation within *MYCN*-NA tumors is not well understood [59].

We applied the hydra *filter* analysis to the TARGET high-risk neuroblastoma cohort as described above. This analysis identified 931 genes within the *MYCN*-NA neuroblastoma cohort with a multimodal expression distribution. Of the 931 multimodally expressed genes, 358 genes were found to be potentially druggable by the Drug Gene Interaction Database (S1 File) and 60 genes were associated with an FDA-approved, anti-neoplastic drug [60].

We next examined whether unsupervised clustering of multimodally expressed genes revealed coordinated expression of annotated gene sets within the MSigDB database. Applying the hydra *sweep* command to the *MYCN*-NA neuroblastoma cohort discovered 105 gene sets with multimodal expression patterns. Each gene set sheds light

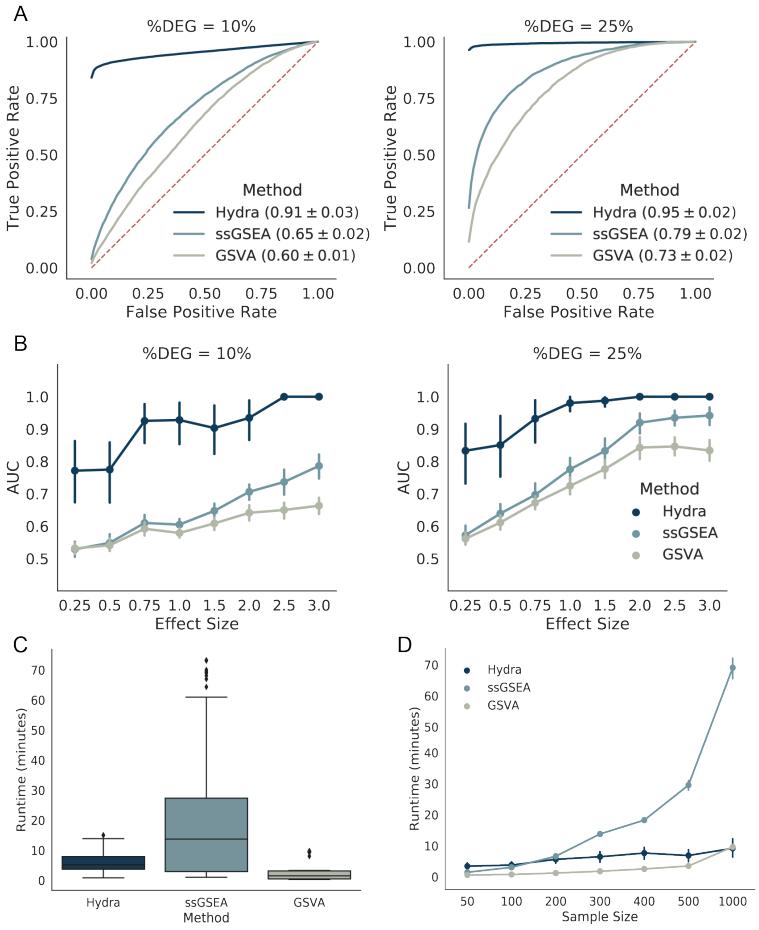


Fig 2. Hydra *sweep* is more sensitive than existing gene set enrichment approaches for detecting differential pathway expression in synthetic data and scales well to large datasets. A: Mean receiver operator curves across effect sizes, percent differentially expressed genes (%DEG), and MSigDB Hallmark gene sets. A larger area under the curve (AUC) indicates better performance. The average AUC and 95% confidence interval for each method are in the ROC plot figure legends. B: Line plots comparing the mean AUC across a range of effect sizes and %DEG values. C: Box plot showing mean runtimes for differential pathway analysis where the effect size is fixed but the sample size varies. D: Line plot comparing the mean runtimes for differential pathway analysis across a range of sample sizes.

on biological themes that are differentially expressed within the *MYCN*-NA neuroblastoma cohort. We clustered the differentially expressed gene sets to reveal these

biological themes (S4 Fig). We found 6 major themes, including annotated cancer functions, cell cycle regulation, cell signaling pathways, immune functions, extracellular matrix reorganization, and metabolic pathway gene sets.

We applied the hydra *enrich* analysis to the MYCN-NA cohort to identify how the most highly enriched gene sets interact to form expression subtypes. This analysis found 428 genes with a minor component probability greater than 20% (S1 File). Gene Ontology analysis found enrichment for the following GO terms (FDR: $q < 0.01$): adaptive immune response (24 genes), mesenchyme development (12 genes), steroid hormone secretion (4 genes), and response to corticosterone (4 genes). DP-GMM analysis of the 44 enriched GO term genes identified three *MYCN*-NA neuroblastoma clusters (Fig 3A). The posterior probability for belonging to each cluster was 42%, 34%, and 17% for clusters 1, 2, and 3, respectively. The posterior probability for a sample belonging to a new cluster was about 6% in our analysis.

We next investigated cluster-specific expression signatures using GSEA (see Hydra Method section). Cluster 1 was enriched for adaptive immune response gene sets, cluster 2 was enriched for proliferative signaling gene sets, and cluster 3 was enriched for cancer-associated fibroblast gene sets (Fig 3B). Cluster 3 shares several features of a wound healing response, including fibroblast recruitment, extracellular matrix organization, and infiltration of immune cells [61].

Clusters 1 and 3 were enriched for tumor microenvironment-associated gene expression. To further validate this signal, we correlated the hydra clusters with enrichment scores from the tumor microenvironment profiling tools xCell [50] and ESTIMATE [52]. Cluster 1 had high average xCell enrichment scores associated with adaptive immune cell types including B-cells, CD4+ naive T-cells, and CD8+ naive T-cells (Kruskal-Wallis: $p < 0.001$). Cluster 2 was characterized by the absence of immune and stromal expression and higher tumor purity scores than clusters 1 and 3. The average ESTIMATE tumor purity was 88%, 96% and 82% for clusters 1, 2, and 3, respectively. Cluster 3 was enriched for fibroblast-associated expression by xCell analysis (Kruskal-Wallis: $p < 0.001$). Clusters 1 and 3 had higher ESTIMATE immune-associated expression levels than cluster 2 (average ImmuneScore per cluster: 58, -612, 56), but cluster 3 had the highest stromal expression signature score (average StromalScore per cluster: -1027, -1310, -135). Comparing ESTIMATE enrichment scores across clusters reveals clear trends in broad immune and stromal expression signatures. Lastly, we found a correlation between the hydra-identified tumor microenvironment subtype and *CD274* and *CTLA4* expression (S6 Fig)

We next correlated clusters with clinical features. We found no difference in patient survival outcomes across clusters (log-rank test, $p > 0.05$). Notably, cluster 1, which had the highest adaptive immune expression signal in MYCN-NA neuroblastoma, over-expresses cell-cycle regulation genes, which was not observed in other small blue cell tumors. We investigated associations with clinical covariates, including mutation burden, age, and tumor content as assessed by a clinical pathologist, but found no statistically significant differences (Kruskal-Wallis: $p > 0.05$). We then investigated associations between the hydra clusters and neuroblastoma-associated molecular aberrations and clinical features (S1 File). *ATRX* gene deletions were enriched in cluster 1 (Fisher's Exact Test: $p < 0.05$). MKI low tumors were enriched in cluster 2 and 3 (Fisher's Exact Test: $p < 0.01$). Chromosome 17 wild-type tumors were enriched in clusters 2 and 3 (Fisher's Exact Test: $p < 0.01$). Analysis on a larger dataset may reveal additional clusters and correlations with clinical features.

Consensus clustering is a widely used approach for identifying tumor subtypes using gene expression data. We applied the M3C consensus clustering method, which is a more sophisticated version of consensus clustering that uses a null distribution to assess the statistical significance of the clustering [20, 21]. We used the top 5000 genes with the

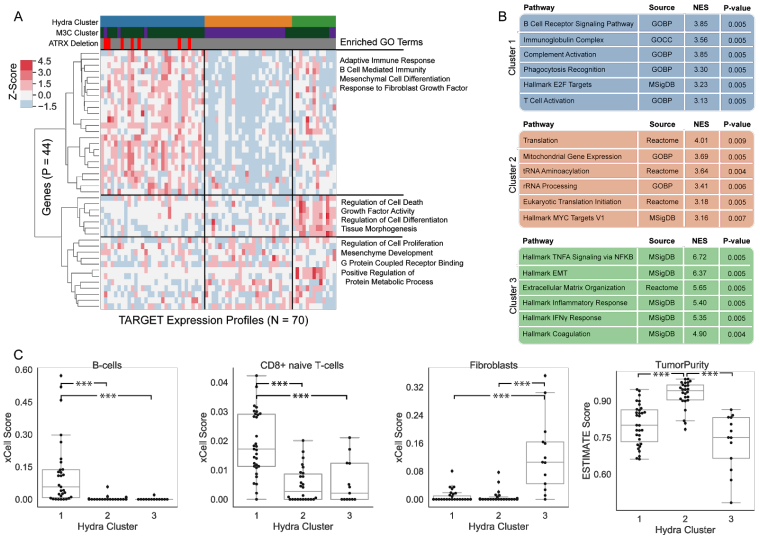


Fig 3. Hydra analysis identifies three distinct tumor microenvironment expression subtypes in *MYCN* non-amplified neuroblastoma samples. A: Gene expression heatmap displaying expression profiles of hydra clusters. Heatmap columns (samples) are ordered by hydra cluster membership. Ward hierarchical clustering applied to rows (genes) identified coordinated expression of GO term genes. These GO term genes were originally identified by the hydra *enrich* command. B: GSEA performed on each cluster identified enrichment of tumor microenvironment and proliferative signaling gene sets. C: xCell enrichment score distributions for B-cells, CD8+ naive T-cells, and Fibroblasts, and the ESTIMATE TumorPurity score distributions for each cluster; enrichments for all cell types are available in S1 File. Abbreviations: Normalized Enrichment Score (NES), Epithelial to Mesenchymal Transition (EMT), Gene Ontology Biological Process (GOBP).

largest median absolute deviation (MAD) because this threshold is routinely used in unsupervised clustering of cancer gene expression data [62–64].

The M3C analysis resulted in the identification of two statistically significant clusters. One M3C cluster correlated with hydra clusters 1 and 3 and the other M3C cluster correlated with hydra cluster 2. Therefore, M3C clustering detected the tumor purity signal in the expression data, but was not able to separate the adaptive immune cell and fibroblast infiltrated clusters (hydra clusters 1 and 3). We also applied k-means clustering using the gap statistic approach [53, 65] for estimating the number of clusters, but this approach grouped all samples into a single cluster. We tested a range of MAD thresholds based on the median absolute deviation, but found similar results across thresholds (S3 Fig). Overall, the hydra approach was more sensitive at detecting distinct tumor microenvironment states than these other popular clustering methods.

To further investigate expression patterns within the hydra-identified tumor microenvironment subtypes, we performed GSEA by z-score normalizing each tumor's gene expression data to its tumor microenvironment cluster. This is a novel GSEA approach that uses the tumor microenvironment state discovered by the hydra method

to identify additional gene expression signals for individual samples. This approach revealed signals not present at the cohort level analysis (Fig 4). For example, enrichment of immune expression signatures within cluster 2 predicted differences in overall survival such that patients with higher immune expression had a better overall survival rate. Similarly, an elevated cell cycle signal within cluster 3 predicted worse survival compared to other cluster 3 samples with lower cell cycle expression. A metastatic expression signal was identified in the analysis of cluster 1 samples, but this signature did not correlate with a difference in survival. This approach may therefore provide appropriate background distributions for revealing and evaluating the significance of gene expression patterns and survival statistics within tumor subtypes.

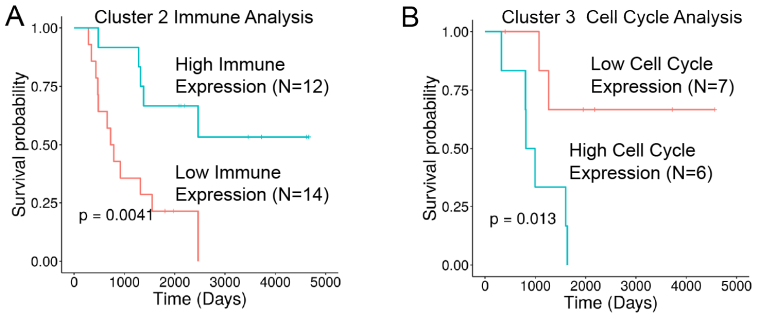


Fig 4. Gene set enrichment analysis (GSEA) of *MYCN*-NA neuroblastoma identifies overall survival differences within hydra cluster 2 and cluster 3. Cluster-level GSEA separated cluster 2 into high and low immune expression subtypes and cluster 3 into high and low cell cycle expression subtypes. A: Kaplan-Meier plot for immune expression subtypes within cluster 2. B: Kaplan-Meier plot comparing cell cycle expression subtypes within cluster 3.

N-of-1 tumor analysis for pediatric neuroblastoma

The command-line interface of the hydra toolkit includes a *predict* function for labeling samples using a pre-fit model. The *MYCN*-NA neuroblastoma model described above was used to predict expression subtypes on a new set of samples. We obtained tumor gene expression data from six stage 4, *MYCN*-NA neuroblastoma samples from the UCSC Treehouse gene expression compendium [5, 6]. The age at diagnosis ranged from 2 to 6 years. Four out of six samples had a deletion in the *ATRX* gene.

Application of the hydra N-of-1 analysis framework clustered 4 out of the 6 samples into cluster 1, which is characterized by adaptive immune cell expression. Three of the *ATRX*-deleted samples clustered with the high adaptive immune cell expression cluster (cluster 1) and one clustered in the low immune, high proliferative signaling cluster (cluster 2). We showed earlier that tumors with *ATRX* deletions tend to have higher adaptive immune expression, and we found a similar pattern in an independent set of *MYCN*-NA neuroblastoma samples.

Two of the samples with loss of *ATRX* came from the same patient but at different timepoints. The first sample (diagnostic sample) clustered with high adaptive immune cell expression (cluster 1), but the resection sample clustered with the low immune expression, high proliferative signaling cluster (cluster 2). We investigated possible explanations for the change in tumor microenvironment state. We performed GSEA

comparing the samples from different timepoints to investigate potential mechanisms leading to immune evasion in these samples. GSEA found downregulation of the MHC Class I Antigen Processing & Presentation GO term in the resection sample (adjusted p-value < 0.002). Loss of antigen processing functions is a common mechanism of immune evasion across cancer types [66].

We obtained H&E stained sections for each of the hydra-identified clusters (S5 Fig). The cluster 1 sample had moderate levels of inflammation (30-50%) consisting of mature mononuclear cells, plasma cells, and eosinophils. The cluster 2 sample had minimal levels of inflammation (<10%) with some scattered mature mononuclear cells throughout the tumor. The cluster 3 sample looked similar to the cluster 1 slide with moderate levels of inflammation (30-50%), but also had regions of apparent necrosis. The inflammation and necrosis in the cluster 3 sample may correlate with the tissue remodeling/wound healing signature identified in the expression data.

Hydra analysis discovers complex tissue signatures

While the *MYCN*-NA neuroblastoma analysis above focused on immune and wound healing expression signatures, the hydra *enrich* method is unsupervised and can therefore detect any type of expression signature. To illustrate this, we applied the hydra *filter*/*enrich* analysis to the TARGET osteosarcoma cohort (N=74) and discovered enrichment of the GO striated muscle contraction term (FDR < 0.01, Fig 5). Multivariate clustering for the GO striated muscle contraction gene set using the *sweep* routine identified two clusters. xCell analysis of the osteosarcoma cohort found significant enrichment of skeletal muscle expression in the second cluster (Mann-Whitney U test, $p < 0.001$). Surprisingly, the M3C clustering approach was not able to detect the strong muscle signature using the 5000 genes with the largest MAD ($p > 0.05$). We used the muscle expression signature to identify osteosarcoma tumors in the UCSC Treehouse Compendium which also contained a similar expression signature. We subsequently confirmed with a licensed pathologist that one of the muscle-expression positive tumor samples did contain significant muscle tissue infiltration. The hydra *enrich* analysis revealed expression signatures not routinely investigated when analyzing osteosarcoma data. Nevertheless, these signals contribute significantly to the tumor expression profile, so explaining these sources of variation is necessary to derive clinically relevant conclusions from gene expression data.

We applied the *filter* method to Ewing sarcoma and discovered multimodal expression of an important druggable gene, JAK1. Applying the multimodal expression model allowed us to deconstruct the Ewing sarcoma distribution into three components (S7 FigA). We found that the expression component with the highest JAK1 expression was also enriched for mast cell expression (S7 FigB). Therefore, overexpression of JAK1 may not correspond to activation of the JAK/STAT signaling pathway in cancer cells but rather to the presence of mast cells within the tumor microenvironment. Furthermore, targeted inhibition of JAK1 using ruxolitinib was shown to inhibit essential mast cell functions, including degranulation [67]. Therefore, therapeutic intervention intending to inhibit JAK1 expression in cancer cells may inadvertently inhibit the patient's mast cell functions. Overexpression analysis using the Ewing sarcoma JAK1 expression distribution may identify JAK1 as an actionable lead, but further investigation into the effect of inhibiting off-target JAK1 expression in mast cells is needed. The hydra framework facilitates the identification of important expression signatures which can be used to deconstruct complex tumor expression subtypes and identify potentially confounding expression signals.

We next quantified the number of multimodal druggable genes from the *MYCN*-NA neuroblastoma dataset that correlated with at least one xCell cell type signature. Out of the 358 druggable genes, we found that 77 correlated with a non-cancer cell type

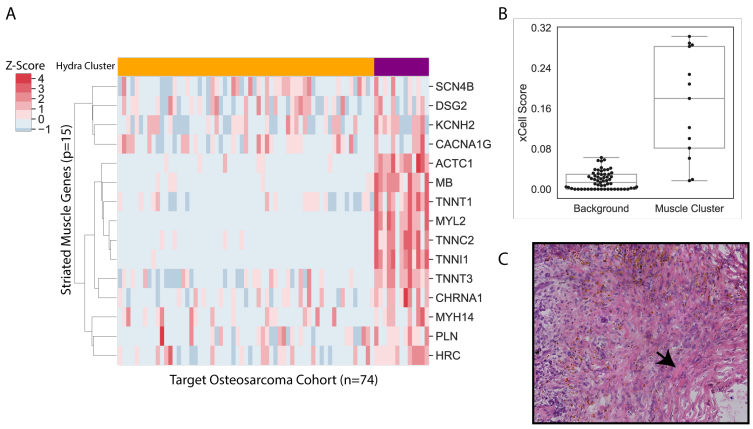


Fig 5. Hydra analysis of TARGET osteosarcoma cohort reveals skeletal muscle signature. Hydra enrichment analysis on the TARGET osteosarcoma cohort revealed a subset of patients with high skeletal muscle expression. A: Clustered heatmap shows the muscle signature genes identified by hydra unsupervised enrichment analysis (purple: enriched for muscle signature; yellow: not enriched for muscle signature). B: xCell tumor microenvironment profiling identified significant differences in skeletal muscle expression compared to background ($p < 0.001$). C: H&E stained tumor slide confirms presence of striated muscle tissue within the tumor sample.

(Kruskal-Wallis test: Holm-Sidak adjusted p-value < 0.05, S1 File). Some of the druggable genes were expected to correlate with non-cancer cells, including the cytokines *IL6* and TGFB2, which correlated with epithelial cells and fibroblasts, respectively. Other druggable genes were surprising, like *AURKA* and *AURKB*, which correlated with higher Th2 cell expression. Aurora kinases play essential roles in spindle formation during mitosis and the overexpression of these genes is associated with evading spindle formation checkpoints in cancer [68], but little is known in how these genes correlate with infiltrating immune cells. Aurora kinase inhibitors show limited clinical activity in solid tumors, but have been shown to have a greater effect in leukemias [68, 69].

Hydra analysis reveals recurrent expression subtypes across small blue round cell tumors

We next investigated whether similar hydra clusters could be identified across other small blue round cell tumors. We chose to focus on extracranial solid tumors because they are among the most common pediatric cancers, making up 20% of all pediatric cancer diagnoses [70], and while survival rates have improved, there are few effective treatment options for the subset of patients with relapse or refractory disease [71]. Identifying expression subtypes for these diseases may improve risk stratification and discover opportunities for new therapies. These tumors also share similar histopathological features, so we hypothesized that these tumors may share similar gene expression subtypes, despite significant differences in the raw expression profiles (Fig 6A).

We first performed TumorMap analysis, which is a dimensionality reduction approach for visualizing genomic data on a 2D surface [5]. We found that small blue

round cell tumor types — *MYCN*-NA neuroblastoma, osteosarcoma, Ewing sarcoma, synovial sarcoma, alveolar rhabdomyosarcoma, and embryonal rhabdomyosarcoma — all form separate TumorMap clusters (Fig 6A). This suggests there is a strong cell-of-origin signal driving the clustering of these cancer types, which is an observation that was recently made in the larger TCGA dataset of adult cancers [72]. While pan-cancer analysis emphasized the differences across small blue round cell tumors, we hypothesized that expression subtypes within cancer types would participate in shared biological themes.

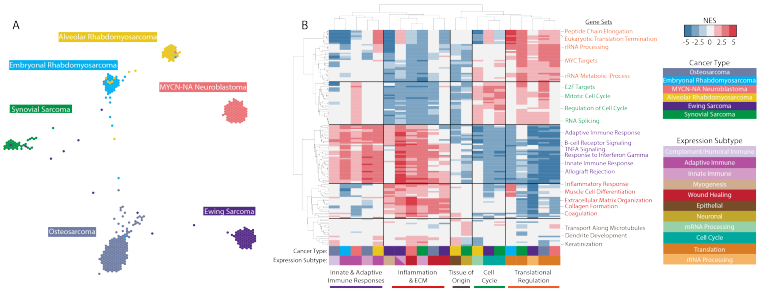


Fig 6. Hydra *enrich* analysis of small blue round cell tumors reveals similar expression subtypes across cancer types. A: TumorMap visualization of 6 small blue round cell tumor types. B: Hierarchically clustered heatmap for the top 10 enriched gene sets across the 21 small blue round cell tumor expression subtypes. Each column corresponds to a cancer type and an expression subtype (x-axis). Each row corresponds to a gene set. The expression subtype was manually assigned after reviewing the most highly enriched gene sets for each cancer expression subtype.

We next performed hydra *enrich* analysis within each small round blue cell cancer type and found shared biological themes across all six small blue round tumor types. Hierarchical clustering of the top 10 statistically significant gene sets for each cancer type resulted in clustering by expression subtype and not the cancer type (Fig 6B). Common themes emerged across diseases including translational regulation, cell cycle regulation, immune effector cell signaling, inflammation, extracellular matrix organization, and tissue-of-origin signals. Furthermore, these signals predicted differences in patient outcomes in osteosarcoma and synovial sarcoma (Fig 7). In both cases, the presence of immune-associated expression correlated with better patient outcomes compared to tumors with proliferative signaling pathways associated with translation initiation and cell cycle regulation. Other osteosarcoma clusters were not included in the survival analysis due to insufficient number of samples with survival data ($n < 5$). Survival data were not available for the rhabdomyosarcoma and Ewing sarcoma expression datasets.

Discussion

The hydra framework uses model-based clustering to facilitate the discovery of recurrent expression patterns within cancer gene expression cohorts. We leveraged recent improvements in model-based clustering algorithms to identify differentially expressed genes without a matched normal distribution. We modeled differential expression as a multimodal Gaussian distribution using nonparametric Bayesian statistics. We then enriched for biologically-annotated Gene Ontology terms and performed multivariate

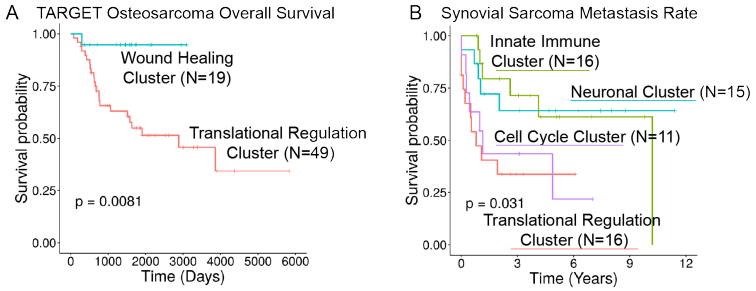


Fig 7. Hydra analysis identifies tumor microenvironment expression subtypes that correlate with patient outcomes in osteosarcoma and synovial sarcoma. A: Kaplan-Meier plot showing overall survival curves for osteosarcoma wound healing and translation clusters. B: Kaplan-Meier plot showing metastasis survival curves for synovial sarcoma clusters.

clustering to reveal expression subtypes. The hydra framework can be used for both identifying expression subtypes within large cohorts and classifying new tumor gene expression profiles using the trained models. The hydra framework outperformed standard gene set enrichment tools for identifying overexpression of the MSigDB Hallmark cancer gene sets in synthetic data. Application of this framework to small blue round cell tumors identified shared biological themes associated with the tumor microenvironment.

Multivariate gene expression analysis is typically underpowered because the number of genes greatly exceeds the number of samples. To address this limitation, we propose selecting for multimodally expressed genes before performing multivariate analysis. The hydra *filter* method reduces the number of genes and enriches for genes that participate in known biological processes, including those curated in the Gene Ontology and MSigDB databases. Selecting for multimodally expressed genes improves separation of known clinical subtypes better than the standard approach of using all expressed genes according to TumorMap analysis (S2 Fig). We also showed that the hydra approach of subsetting to multimodal genes improves detection of differential pathway expression, including the identification of expression subtypes associated with the TME.

Significant progress has been made in subtyping neuroblastomas and adapting therapy for aggressive subtypes, but unexplained heterogeneity remains [59]. Failure to account for this heterogeneity decreases the power of standard methods to detect important expression patterns. Identifying biomarkers using genome-wide technology may lead to improved risk stratification and the discovery of novel drug targets. Hydra analysis of the TARGET *MYCN*-NA neuroblastoma cohort found differential expression of tumor microenvironment markers, including markers of the adaptive immune response. Pediatric cancers are generally thought to be less immunogenic because they have lower mutation burdens than adult cancers, but the immunogenicity of pediatric cancer has not been sufficiently investigated [11, 12].

Our analysis found significant variation in immune marker expression, including markers of response to checkpoint blockade therapy, and identified *ATRX* deletions as a potential biomarker of immune infiltrated tumors in *MYCN*-NA neuroblastoma. Analysis of other small blue round cell tumors revealed similar expression signatures across tumor types, despite samples clustering by their histology in a pan-cancer TumorMap analysis. Identification of shared expression signatures across cancer types

may suggest that these patients would respond similarly to therapies that target these pathways. In particular, the identification of a cross-disease subtype associated with high expression of immune markers may warrant further investigation of immunotherapies in small blue round cell tumors using a basket clinical trial design [73].

Hydra analysis found significant differences in tumor immune and stromal expression that may inform precision medicine applications. The tumor microenvironment has become an important therapeutic consideration, but few methods account for the tumor microenvironment directly. Tumor purity has been identified as a confounding factor in cancer gene expression subtyping efforts [74]. For example, tumor purity and tumor microenvironment expression have been shown to correlate with pancreatic cancer subtypes [75]. Furthermore, Aran et al. (2018) found that tumor purity was correlated with the mesenchymal glioblastoma subtype and recommended a differential expression approach to computationally remove the tumor purity signal. However, standard approaches for subtracting the tumor purity effect may not be ideal because several mechanisms may influence tumor purity, and each mechanism may result in a different expression pattern. For instance, our analysis of *MYCN*-NA neuroblastoma identified two gene expression signatures that correlated with lower predicted tumor purity. Cluster 1 had an adaptive immune expression signature and cluster 3 had a cancer-associated fibroblast signature. Therefore, the estimated tumor purity signal should not be subtracted without first accounting for the different mechanisms influencing tumor purity.

We also found shared biological pathway enrichment across small blue round cell tumors. While these diseases are related and may derive from similar cell lineages, current expression methods often emphasize difference across these diseases (Fig 6A). Unsupervised clustering of adult cancer types found that cell-of-origin signals strongly influence clustering of cancer gene expression data [72]. Although these diseases have distinct expression patterns on the surface, we discovered common themes once we subset the data to the cell-of-origin signal and applied the hydra analysis tools.

We found at least three shared TME states: immune silent, immune infiltrated, and wound healing subtypes. The wound healing subtypes predicted better overall survival in osteosarcoma and delayed metastases in synovial sarcoma tumors, which suggests the involvement of the host immune response limits the progression of these tumors. Amplification of the host immune response may further limit tumor growth and lead to immune-mediated tumor cell death. Additional research into immune modulating therapies is warranted in small blue round cell tumors and may lead to improved outcomes for some patients.

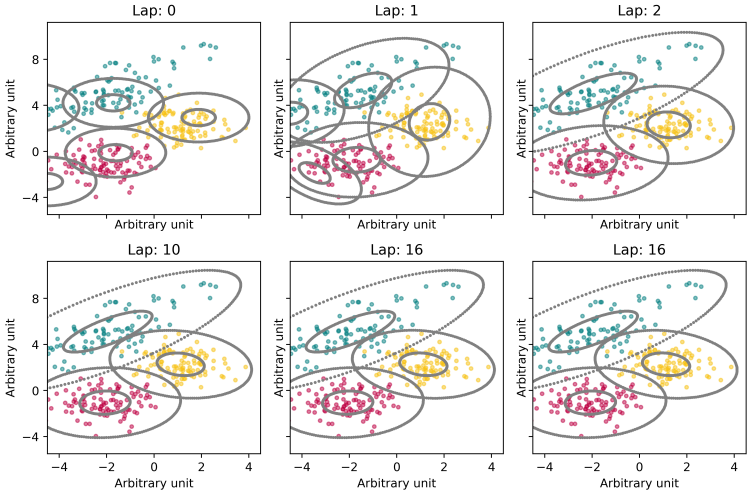
Conclusion

Precision oncology aims to differentiate tumors of the same diagnosis in order to match patients with the best treatment. We have developed the hydra framework to discover subtle but recurrent expression patterns within a cohort of samples with the same diagnosis, which is a novel strategy for pediatric precision oncology research. Our approach may help to uncover the biology underlying tumor progression and response to therapy. We have shown that hydra is more sensitive than standard gene set enrichment approaches for detecting differential pathway expression. Additionally, our framework provides tools to conduct unsupervised clustering analysis to discover expression subtypes. We applied the unsupervised hydra analysis to small blue round cell tumors and discovered distinct tumor microenvironment (TME) states. This shows that one of the strongest signals in clinical gene expression data comes from the TME, so careful modeling of the TME is required to maximize the impact of clinical gene expression analysis. The hydra framework provides unbiased clustering tools to characterize these

sources of variation in specific disease populations and identify shared biological themes that can potentially be targeted therapeutically.

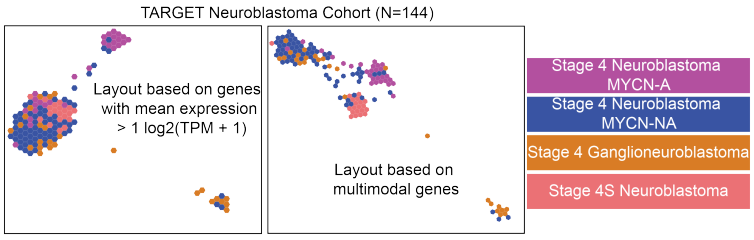
Supporting information

S1 Fig.

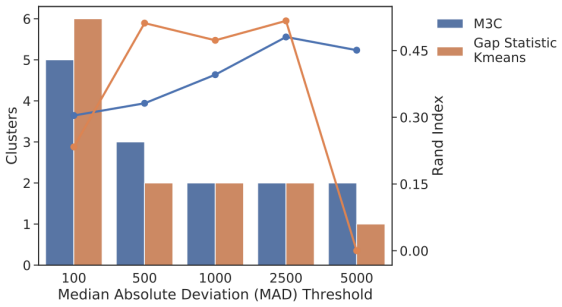


Example of bnpy memoized online variational inference clustering on toy data. We used the bnpy moVB algorithm to infer the number of clusters from synthetic data. The model first randomly assigns clusters. Then, the model iteratively improves the model fit, creating and destroying clusters until the model converges on the correct number of clusters at lap 16 [36].

S2 Fig.

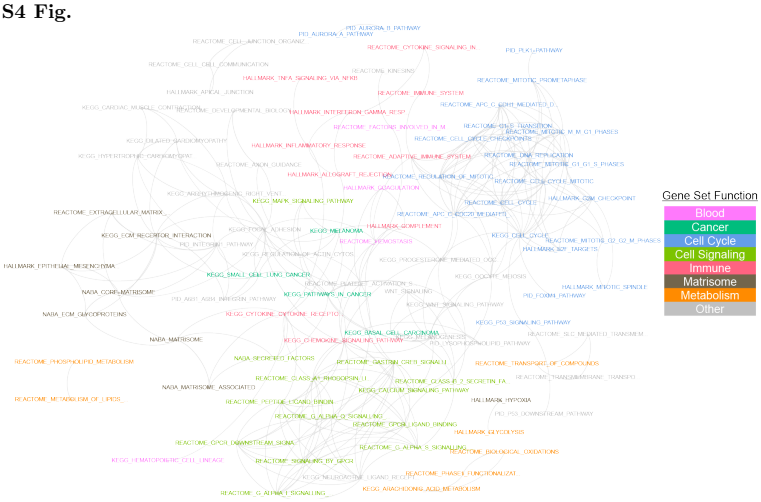


Enriching for multimodally expressed genes improves clustering of established neuroblastoma subtypes. Standard TumorMap analysis of the TARGET neuroblastoma dataset resulted in stage 4S samples clustering with stage 4 neuroblastoma samples (left). An alternative TumorMap based solely on 1,498 multimodally expressed genes separated the stage 4S samples into a distinct cluster (right).



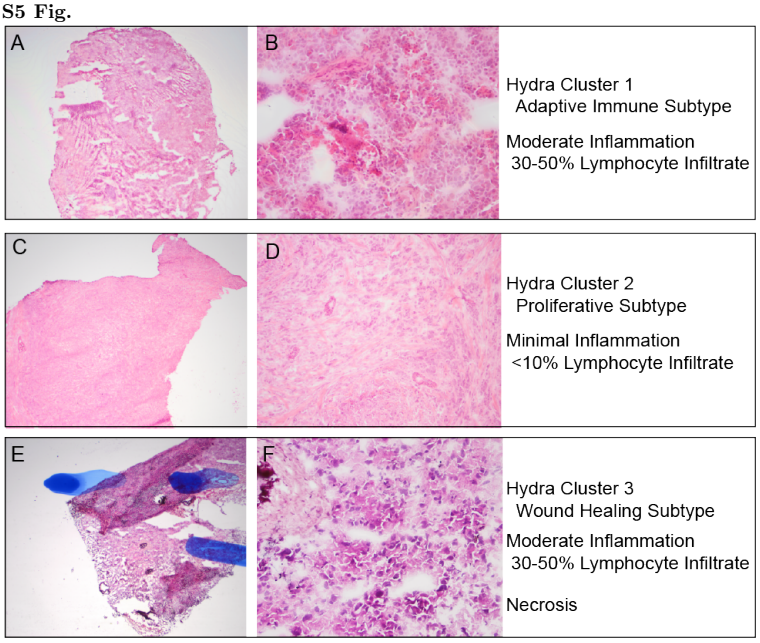
S3 Fig.
Consensus and k-means clustering applied to TARGET *MYCN*-NA dataset. We tested a range of gene expression variation thresholds based on the median absolute deviation, but found that the clusters identified by this approach could not resolve the same clusters as the hydra approach. The barplot shows the number of clusters and the lineplot tracks the Rand index comparing the M3C and k-means clusters and the hydra clusters.

S4 Fig.



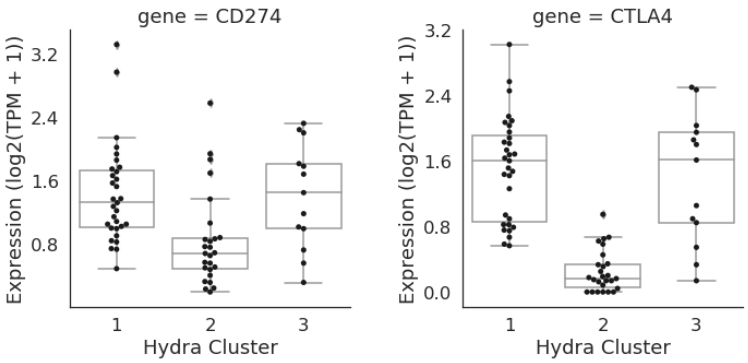
Hydra *sweep* analysis reveals differential pathway expression within *MYCN*-NA neuroblastoma without a matched cohort of normal tissue. Unsupervised clustering of multimodal gene sets revealed biological themes associated with hallmark cancer functions, including cell cycle, immune cell signaling, extracellular matrix organization, and metabolism.

S5 Fig.



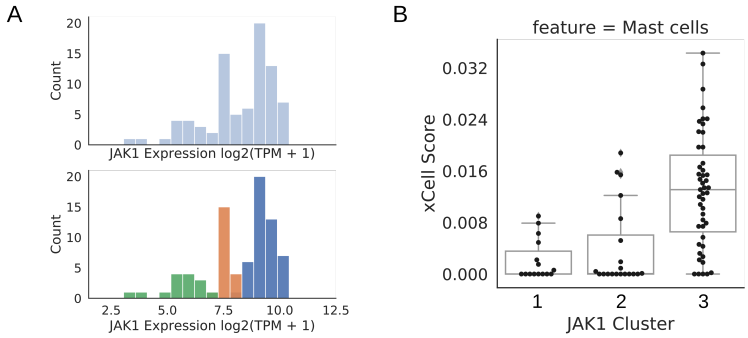
Hydra method correlates with distinct tumor features as assessed by licensed pathologist review of tumor H&E slides. A-B: H&E sections from fresh frozen tumor tissue from *MYCN*-NA neuroblastoma sample at A: 2X magnification and B: 20X magnification. Tumor cells are medium to large with moderate amounts of cytoplasm and areas of rhabdoid appearing undifferentiated cells. There is a moderate amount of mixed inflammation present (30-50%) consisting mostly of mature mononuclear cells with some plasma cells and scattered eosinophils. C-D: H&E sections from fresh frozen tumor tissue from *MYCN*-NA neuroblastoma at C: 2X magnification and D: 20X magnification. Tumor cells are moderate to large in size with moderate amounts of cytoplasm. There is a minimal amount (<10%) of apparent mononuclear inflammation scattered throughout the tumor. E-F: H&E sections from fresh frozen tumor tissue from *MYCN*-NA neuroblastoma sample at (E) 2X magnification and (F) 20X magnification. Tumor cells are medium to large with moderate amounts of cytoplasm and areas of rhabdoid appearing undifferentiated cells. There are also areas of apparent necrosis. There is a moderate amount of inflammation present (30-50%) consisting mostly of mature mononuclear cells with some plasma cells and scattered eosinophils.

S6 Fig.



Hydra *enrich* analysis identifies correlation between expression subtypes and checkpoint blockade markers in *MYCN*-NA neuroblastoma.

S7 Fig.



Hydra analysis identified *JAK1* expression clusters that correlate with mast cell expression signature in Ewing sarcoma. A: *JAK1* expression distribution for Ewing sarcoma cohort (top) and the JAK1 expression distributions for cluster 1 (green), 2 (orange), and 3 (blue). B: Boxplot showing the xCell mast cell enrichment score for the three clusters associated with *JAK1* expression.

S1 File. TARGET *MYCN*-NA neuroblastoma supplementary data.

S2 File. Hydra method documentation.

Acknowledgments

We would like to thank the patients and families who participated in translational genomics research. We would also like to thank the St. Baldrick's Foundation, the California Initiative to Advance Precision Medicine, and the National Human Genome Research Institute of the National Institutes of Health for funding support.

References

1. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil Enables Reproducible, Open Source, Big Biomedical Data Analyses;35(4):314–316. doi:10.1038/nbt.3772.

2. Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, et al. The Genetic Landscape of High-Risk Neuroblastoma;45(3):279–284. doi:10.1038/ng.2529.

3. Goldman M, Craft B, Kamath A, Brooks A, Zhu J, Haussler D. The UCSC Xena Platform for Cancer Genomics Data Visualization and Interpretation; p. 326470. doi:10.1101/326470.

4. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project;45:1113–1120. doi:10.1038/ng.2764.

5. Newton Y, Novak AM, Swatloski T, McColl DC, Chopra S, Graim K, et al. TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal;77(21):e111–e114. doi:10.1158/0008-5472.CAN-17-0580.

6. Vaske OM, Bjork I, Salama SR, Beale H, Shah AT, Sanders L, et al. Comparative Tumor RNA Sequencing Analysis for Difficult-to-Treat Pediatric and Young Adult Patients With Cancer;2(10):e1913968–e1913968. doi:10.1001/jamanetworkopen.2019.13968.

7. Joyce JA, Fearon DT. T Cell Exclusion, Immune Privilege, and the Tumor Microenvironment;348(6230):74–80. doi:10.1126/science.aaa6204.

8. Chen DS, Mellman I. Elements of Cancer Immunity and the Cancer–Immune Set Point;541(7637):321–330. doi:10.1038/nature21349.

9. Mellman I, Coukos G, Dranoff G. Cancer Immunotherapy Comes of Age;480(7378):480–489. doi:10.1038/nature10673.

10. Page DB, Postow MA, Callahan MK, Allison JP, Wolchok JD. Immune Modulation in Cancer with Antibodies;65:185–202. doi:10.1146/annurev-med-092012-112807.

11. Majzner RG, Heitzeneder S, Mackall CL. Harnessing the Immunotherapy Revolution for the Treatment of Childhood Cancers;31(4):476–485.

12. Zamora AE, Crawford JC, Allen EK, Guo XzJ, Bakke J, Carter RA, et al. Pediatric Patients with Acute Lymphoblastic Leukemia Generate Abundant and Functional Neoantigen-Specific CD8+ T Cell Responses;11(498). doi:10.1126/scitranslmed.aat8549.

13. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-Based Differential Expression Analysis of RNA Sequencing Data Using R and Bioconductor;8(9):1765–1786. doi:10.1038/nprot.2013.099.

14. Anders S, Huber W. Differential Expression Analysis for Sequence Count Data;11(10):R106. doi:10.1186/gb-2010-11-10-r106.

15. Soneson C, Delorenzi M. A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data;14(1):91. doi:10.1186/1471-2105-14-91.

16. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles;102(43):15545–15550. doi:10.1073/pnas.0506580102.

17. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes;34(3):267–273. doi:10.1038/ng1180.

18. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular Signatures Database (MSigDB) 3.0;27(12):1739–1740. doi:10.1093/bioinformatics/btr260.

19. Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, et al. Clustering Algorithms: Their Application to Gene Expression Data;10:237–253. doi:10.4137/BBI.S38316.

20. John CR, Watson D, Russ D, Goldmann K, Ehrenstein M, Lewis M, et al. M3C: A Monte Carlo Reference-Based Consensus Clustering Algorithm; p. 377002.

21. Wilkerson MD, Hayes DN. ConsensusClusterPlus: A Class Discovery Tool with Confidence Assessments and Item Tracking;26(12):1572–1573. doi:10.1093/bioinformatics/btq170.

22. Lenz M, Müller FJ, Zenke M, Schuppert A. Principal Components Analysis and the Reported Low Intrinsic Dimensionality of Gene Expression Microarray Data;6(1):1–11. doi:10.1038/srep25696.

23. Ghosh D. Mixture Models for Assessing Differential Expression in Complex Tissues Using Microarray Data;20(11):1663–1669. doi:10.1093/bioinformatics/bth139.

24. Dahl DB, Vannucci M. In: Do KA, Muller P, editors. Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. Cambridge University Press;. p. 201–218. Available from: `https://www.cambridge.org/core/product/identifier/CBO9780511584589A070/type/book_part`.

25. Kim S, Tadesse MG, Vannucci M. Variable Selection in Clustering via Dirichlet Process Mixture Models;93(4):877–893. doi:10.1093/biomet/93.4.877.

26. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB, et al. Bayesian Data Analysis. Chapman and Hall/CRC;. Available from: `https://www.taylorfrancis.com/books/9780429113079`.

27. Thall PF, Mueller P, Xu Y, Guindani M. Bayesian Nonparametric Statistics: A New Toolkit for Discovery in Cancer Research;16(6):414–423. doi:10.1002/pst.1819.

28. Teh YW. Dirichlet Process; p. 280–287.

29. Antoniak CE. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems;2(6):1152–1174. doi:10.1214/aos/1176342871.

30. Ferguson TS. A Bayesian Analysis of Some Nonparametric Problems;1(2):209–230. doi:10.1214/aos/1176342360.

31. Müller P, Quintana FA. Nonparametric Bayesian Data Analysis; p. 95–110.

32. Görür D, Edward Rasmussen C. Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution;25(4):653–664. doi:10.1007/s11390-010-9355-8.

33. Hughes MC, Sudderth E. Memoized Online Variational Inference for Dirichlet Process Mixture Models. In: Advances in Neural Information Processing Systems;. p. 1133–1141.

34. Müller P, Quintana FA, Jara A, Hanson T. Bayesian Nonparametric Data Analysis. Springer Series in Statistics. Springer International Publishing;. Available from: `https://www.springer.com/gp/book/9783319189673`.

35. Phadia EG. Prior Processes and Their Applications. Springer;.

36. Hughes MC, Sudderth EB. Bnpy : Reliable and Scalable Variational Inference for Bayesian Nonparametric Models; p. 4.

37. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: Tool for the Unification of Biology;25(1):25–29. doi:10.1038/75556.

38. Consortium GO. The Gene Ontology Resource: 20 Years and Still GOing Strong;47(D1):D330–D338.

39. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation;5(11):e13984.

40. Yu G, Wang LG, Han Y, He QY. clusterProfiler: An R Package for Comparing Biological Themes among Gene Clusters;16(5):284–287. doi:10.1089/omi.2011.0118.

41. Cai T, Liu W, Luo X. A Constrained $\ell 1$ Minimization Approach to Sparse Precision Matrix Estimation;106(494):594–607. doi:10.1198/jasa.2011.tm10155.

42. Tamayo P, Steinhardt G, Liberzon A, Mesirov JP. The Limitations of Simple Gene Set Enrichment Analysis Assuming Gene Independence;25(1):472–487. doi:10.1177/0962280212460441.

43. Korotkevich G, Sukhov V, Sergushichev A. Fast Gene Set Enrichment Analysis; p. 060012. doi:10.1101/060012.

44. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA Interference Reveals That Oncogenic KRAS-Driven Cancers Require TBK1;462(7269):108–112. doi:10.1038/nature08460.

45. Hänzelmann S, Castelo R, Guinney J. GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data;14(1):7. doi:10.1186/1471-2105-14-7.

46. Tarca AL, Bhatti G, Romero R. A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity;8(11):e79217. doi:10.1371/journal.pone.0079217.

47. Zwiener I, Frisch B, Binder H. Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures;9(1):e85150. doi:10.1371/journal.pone.0085150.

48. Lagarde P, Przybyl J, Brulard C, Pérot G, Pierron G, Delattre O, et al. Chromosome Instability Accounts for Reverse Metastatic Outcomes of Pediatric and Adult Synovial Sarcomas;31(5):608–615. doi:10.1200/JCO.2012.46.0147.

49. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks. In: Third International AAAI Conference on Weblogs and Social Media;.

50. Aran D, Hu Z, Butte AJ. xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape;18(1):220. doi:10.1186/s13059-017-1349-1.

51. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust Enumeration of Cell Subsets from Tissue Expression Profiles;12(5):453–457. doi:10.1038/nmeth.3337.

52. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data;4:2612. doi:10.1038/ncomms3612.

53. Tibshirani R, Walther G, Hastie T. Estimating the Number of Clusters in a Data Set via the Gap Statistic;63(2):411–423.

54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-Learn: Machine Learning in Python;12:2825–2830.

55. Jones E, Oliphant T, Peterson P, et al. SciPy: Open Source Scientific Tools for Python;.

56. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python; p. arXiv:1907.10121.

57. Terpilowski M. Scikit-Posthocs: Pairwise Multiple Comparison Tests in Python;4(36):1169. doi:10.21105/joss.01169.

58. Kassambara A, Kosinski M, Biecek P. Survminer: Drawing Survival Curves Using 'Ggplot2';. Available from: `https://CRAN.R-project.org/package=survminer`.

59. Morgenstern DA, Bagatell R, Cohn SL, Hogarty MD, Maris JM, Moreno L, et al. The Challenge of Defining "Ultra-High-Risk" Neuroblastoma;66(4):e27556. doi:10.1002/pbc.27556.

60. Cotto KC, Wagner AH, Feng YY, Kiwala S, Coffman AC, Spies G, et al. DGIdb 3.0: A Redesign and Expansion of the Drug–Gene Interaction Database;46(D1):D1068–D1073.

61. Foster DS, Jones RE, Ransom RC, Longaker MT, Norton JA. The Evolving Relationship of Wound Healing and Tumor Stroma;3(18). doi:10.1172/jci.insight.99911.

62. Bourgon R, Gentleman R, Huber W. Independent Filtering Increases Detection Power for High-Throughput Experiments;107(21):9546–9551. doi:10.1073/pnas.0914005107.

63. Tritchler D, Parkhomenko E, Beyene J. Filtering Genes for Cluster and Network Analysis;10(1):193. doi:10.1186/1471-2105-10-193.

64. Carcamo-Orive I, Hoffman GE, Cundiff P, Beckmann ND, D'Souza SL, Knowles JW, et al. Analysis of Transcriptional Variability in a Large Human iPSC Library Reveals Genetic and Non-Genetic Determinants of Heterogeneity;20(4):518–532.e9. doi:10.1016/j.stem.2016.11.005.

65. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K, et al. Cluster: Cluster Analysis Basics and Extensions;1(2):56.

66. Reeves E, James E. Antigen Processing and Immune Regulation in the Response to Tumours;150(1):16–24. doi:10.1111/imm.12675.

67. Hermans MAW, Schrijver B, van Holten-Neelen CCPA, Gerth van Wijk R, van Hagen PM, van Daele PLA, et al. The JAK1/JAK2- Inhibitor Ruxolitinib Inhibits Mast Cell Degranulation and Cytokine Release;48(11):1412–1420. doi:10.1111/cea.13217.

68. Maris JM, Morton CL, Gorlick R, Kolb EA, Lock R, Carol H, et al. Initial Testing of the Aurora Kinase a Inhibitor MLN8237 by the Pediatric Preclinical Testing Program (PPTP);55(1):26–34. doi:10.1002/pbc.22430.

69. Gautschi O, Heighway J, Mack PC, Purnell PR, Lara PN, Gandara DR. Aurora Kinases as Anticancer Drug Targets;14(6):1639–1648. doi:10.1158/1078-0432.CCR-07-2179.

70. Ries LaG, Smith MA, Gurney JG, Linet M, Tamra T, Young JL, et al. Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995.;.

71. Ring EK, Markert JM, Gillespie GY, Friedman GK. Checkpoint Proteins in Pediatric Brain and Extracranial Solid Tumors: Opportunities for Immunotherapy;23(2):342–350. doi:10.1158/1078-0432.CCR-16-1829.

72. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer;173(2):291–304.e6. doi:10.1016/j.cell.2018.03.022.

73. Cunanan KM, Iasonos A, Shen R, Begg CB, Gönen M. An Efficient Basket Trial Design;36(10):1568–1579. doi:10.1002/sim.7227.

74. Rhee JK, Jung YC, Kim KR, Yoo J, Kim J, Lee YJ, et al. Impact of Tumor Purity on Immune Gene Expression and Clustering Analyses across Multiple Cancer Types;6(1):87–97. doi:10.1158/2326-6066.CIR-17-0201.

75. Raphael BJ, Hruban RH, Aguirre AJ, Moffitt RA, Yeh JJ, Stewart C, et al. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma;32(2):185–203.

Part III

Novel Immunotherapy Targets from the Dark Matter of the Genome

Introduction

The tumor microenvironment (TME) is the cellular matrix consisting of cancer, stromal, and immune cells. The TME plays an important role in providing nutrients and oxygen to the tumor, shielding cancer cells from the immune system, and providing growth factors that promote cancer growth and resistance to therapies. Cancer cells co-opt the host's wound healing response to promote angiogenesis and proliferative signaling. Epithelial cells create blood vessels that bring nutrients, oxygen, and growth factors to the tumor. Stromal and immune cells exclude cytotoxic immune cells and release signaling molecules that suppress cytotoxic functions of the immune system. The tumor becomes a wound that cannot heal and the host's wound healing program contributes to cancer progression [13].

Stromal cells, such as fibroblasts, make up connective tissues that support other tissues and organs. Fibroblasts are usually quiescent, but become activated during the wound healing response. Fibroblasts are resilient and can shield tumors from radiation and chemotherapy. Fibroblasts increase the interstitial pressure within the tumor microenvironment, which excludes some immune cells from infiltrating. Cancer-associated fibroblasts (CAFs) are identified by over-expression of α-smooth muscle actin [21]. The wound healing program promotes angiogenesis and extracellular matrix remodeling [13]. Fibroblasts also secrete mitogenic signaling molecules that can promote resistance to targeted inhibitors by activating an alternative proliferative signaling pathway. Examples of mitogenic factors include hepatocyte growth factor, epithelial growth factors, insulin-like growth factors, and fibroblast growth factors. Targeted therapies inhibit the function of these growth factors, but compensatory expression of another

growth factor molecule can lead to drug resistance [13].

The cells of the TME also recruit immunosuppressive immune cells that protect cancer cells from the immune system. Infiltrating immune cells include regulatory T-cells, macrophages, neutrophils, dendritic cells, and mast cells. The TME becomes immune privileged, which prevents the immune system and immunotherapies from eliminating cancer cell populations [20]. The immune component also provides important pro-survival signals that allow cancer cells to evade apoptosis. The immune cells secrete epidermal growth factor, transforming growth factor-β, tumor necrosis factor-α, and fibroblast growth factors [13, 35].

Tumors are complex systems of interacting cells, and characterizing the individual cells in the tumor may lead to novel therapeutic directions. For my third aim, I propose a single-cell sequencing approach to describe the cell populations within a patient tumor and to identify drug resistance markers that could influence a patient's response to therapy. Treehouse targeted therapies are susceptible to drug resistance, so identifying markers of resistance before the drugs are administered can allow for early intervention of cancer resistance mechanisms. Single-cell RNA sequencing has been used to characterize tumor heterogeneity. A recent study into the tumor heterogeneity of glioblastoma tumors identified variable expression of receptor tyrosine kinases (RTKs) [33]. RTKs are important molecular targets for therapy and identification of mosiac RTK expression could contraindicate application of an RTK targeted inhibitor. Single-cell RNA sequencing is a powerful approach to identify markers of drug resistance than can be used to prioritize targeted therapies.

There has recently been increased focus on targeting the tumor microenvironment for therapeutic gain. Immunotherapies modulate the immune effector cells to target cancer cells

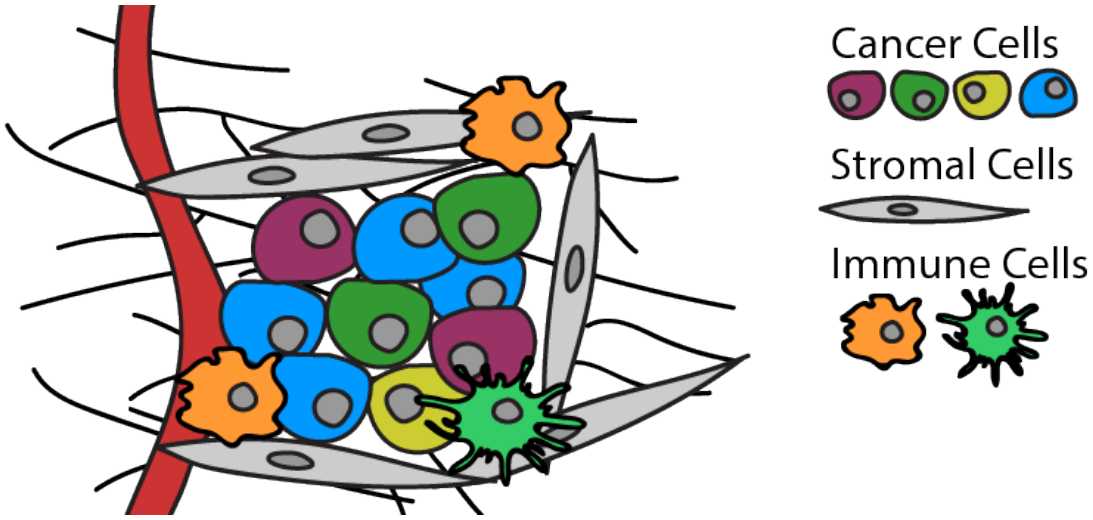


Figure 4.1: The tumor micro-environment is made up of extracellular matrix, cancer, stromal, and immune cells. The tumor microenvironment facilitates tumor growth and survival. Molecularly targeting the tumor microenvironment may yield improved therapeutic responses.

and have become a front line defense against some cancers, including melanoma. While immunotherapies, including checkpoint blockade therapy, have become widely used, the response rates remain low ($< 40\%$), so new strategies are needed to increase the response rate.

Chapter 5

vaccinaTE: A precision immuno-oncology toolkit for identifying transposable element vaccine targets

Introduction

Cancer vaccine technology today is focused on neoantigens derived from somatic mutations, but there has been limited success in bringing this approach into the clinic. One reason for this is the inability to scale this approach to the entire healthcare system. I took a different approach by investigating mobile genetic sequences that are shared across individuals but strongly repressed in healthy cells. Transposable elements make up 40% of the human genome and become overexpressed in cancer cells. This makes transposable elements attractive targets for cancer vaccine development. I developed the vaccinaTE software to identify shared transposable element (TE) epitopes across individuals. I also show that expression of TEs correlates

with a survival benefit in triple negative breast cancer and complete response to checkpoint blockade therapy in melanoma.

vaccinaTE: A precision immuno-oncology toolkit for identifying transposable element vaccine targets

Jacob Pfeil, Jason Fernandes, Lauren Sanders, Alana Weinstein, Geoff Lyle, Holly Beale, Olena Morozova, Sofie Salama, David Haussler

Abstract

Cancer immunotherapy harnesses the power of the immune system to attack cancer and has led to durable responses in advanced disease. However, only a subset of patients respond, so innovative therapeutic strategies, such as combination therapies, are needed to increase the number of patients who benefit. Cancer vaccine in combination with checkpoint blockade therapy is a promising approach to increasing the antitumor immune response, but current limitations in cancer vaccine development may prevent this approach from being widely used. For example, cancer vaccines based on private mutations may be prohibitively expensive and inhibit widespread adoption of this approach. We propose a novel strategy for personalized cancer vaccines that use public antigens that are shared across individuals.

Genomewide dysregulation of transcription and translation leads to overexpression of non-canonical protein coding genes, including transposable elements (TEs). TEs are strongly repressed in healthy cells to prevent genomic instability but become dysregulated in cancer. We developed a computational framework for identifying potential cancer vaccine targets within transposable elements using RNA-seq or mass spectrometry data. We focus on the youngest and most highly conserved transposable element in the human genome, L1HS. We applied our approach to triple negative breast cancer (TNBC) and melanoma and found that L1HS epitope kmers correlate with better survival in TNBC and complete response to checkpoint blockade therapy in melanoma. This suggests that these elements correlate with better survival, presumably through activation of the host immune system. Further activation through vaccination may lead to even stronger antitumor immune responses, which may work synergistically with checkpoint blockade therapy.

Introduction

Cancer is the second leading cause of death in the United States [1], and while there have been significant medical advances in treating this disease, the standard of care has not changed significantly over the past few decades. Chemotherapy, radiation, and surgery have been the frontline defense against cancer progression, but new therapeutic strategies are being developed that personalize the therapy to individuals. For example, targeted therapies are small-molecule drugs designed to inhibit specific molecular alterations, such as an activating kinase mutation. These therapies have generated complete responses in late-stage disease, but resistance often emerges and the cancer relapses. Targeted therapies are routinely used against recurrent activating mutations, including BRAF V600E in melanoma, but most patients do not have an actionable variant and do not benefit from these approaches. Furthermore, targeted therapies do not yield durable responses, since the cancer eventually relapses, and incur significant cost to the healthcare system [2].

Another approach for treating cancer is to amplify the antitumor immune response. This approach has achieved remarkable responses while induces minimal toxic side-effects. The

discovery that the immune system can recognize and destroy cancer cells has opened the door to an entirely new therapeutic approach. Genomewide dysregulation of transcription and translation leads to the presentation of tumor-specific antigens by major histocompatibility complex molecules. Cytotoxic cells recognize tumor-specific antigens and induce immune-mediated cell death.

Unfortunately, this process can select for cancer cells that evade immune recognition, which leads to an immunosuppressive tumor microenvironment that is able to coexist with the host's immune system [3]. For example, some cancer cells adopt immunosuppressive cell-surface markers to curb the antitumor immune response. These include the immune checkpoint molecules CTLA4 and PD1. Identification of immune checkpoint expression in cancer has led to the development of antibody therapies that block the immunosuppressive signal allowing cytotoxic T-cells to continue the antitumor attack. Checkpoint blockade therapy uses the immune system to achieve durable responses with relatively minor toxic side-effects [4–7].

The anti-CTLA4 antibody, ipilimumab, was the first checkpoint blockade therapy to achieve FDA approval [6,8]. CTLA4 has a stronger binding affinity to CD80 and CD86 than the costimulatory CD28 molecules, leading to inhibition of T-cell activation [3]. CTLA4 normally becomes expressed after T-cell activation in order to prevent off-target autoimmunity, cancer cells may express CTLA4 to prevent cytotoxic T-cell activation [4–6]. The anti-PD1 antibody pembrolizumab came later and was found to be more efficacious and have fewer side-effects [9]. PD1 is a cell-surface receptor expressed after T-cell activation. Activation of the PD1 receptor by its ligand PDL1 leads to interference of downstream signaling from the T-cell receptor which suppresses the T-cell response [7,8].

The extraordinary responses to checkpoint blockade therapy has led to this therapy becoming widely used and at increasingly earlier stages in cancer treatment [7]. Using checkpoint blockade as a monotherapy achieves a response rate between 20 and 40% for melanoma [4,9]. Current biomarkers for response include PDL1 expression, T-cell infiltration, tumor bulk, mutation burden, crippled DNA repair machinery, and microsatellite instability. In addition to identifying predictive biomarkers of response, combination immune checkpoint therapies are being investigated. Administering anti-CTLA4 and anti-PD1 therapies increases the response rate (> 40%), but at the cost of increasing the number of adverse events, including fatal pulmonary toxicity [9].

The increased response rate with combination immunotherapy shows that further activation of the immune system correlates with increased antitumor effects. The additional toxic side-effects limit this approaches utility, so new approaches are needed to similarly activate the antitumor immune response while avoiding toxic side-effects. Checkpoint blockade therapy allows infiltrating T-cells to continue their cytotoxic functions, but does not influence the T-cell clones that travel to the tumor. Therapies that expand T-cell clones that are able to recognize cancer cells may work synergistically with checkpoint blockade therapy to tip the balance in favor of immune-mediated destruction of tumors [10].

During a normal infection, antigen presenting cells enter peripheral lymph nodes to excite T-cell that recognize the antigen into rapidly expanding and circulating throughout the body in search of the antigen. Another strategy for improving response to checkpoint blockade

therapy may be to increase the number of circulating T-cells able to recognize cancer cells using a cancer vaccine approach. Cancer vaccines expand the T-cells able to recognize cancer cells and increase the number of T-cells infiltrating the tumor [11].

Despite extensive research into cancer vaccines, the clinical response to cancer vaccine monotherapy has been modest [12,13]. Sipuleucel-T is the only FDA-approved cancer vaccine that stimulates the immune response against a tumor-specific antigen [14]. This suggests that expanding the number of antitumor T-cells is not sufficient, so checkpoint blockade therapy may be required to overcome the inhibitory mechanisms within the tumor microenvironment. Recent studies have shown that vaccines work synergistically with checkpoint blockade therapy to increase response rates [10,11].

Sipuleucel-T does not target a mutated protein, but instead targets a shared antigen that is overexpressed in prostate cancer cells but not in healthy somatic cells. Being shared across patients has facilitated the development of Sipuleucel-T. The alternative cancer strategy being investigated is to identify private mutations within each tumor and synthesize a unique set of peptide vaccines based on that individuals cancer mutations. The private mutation approach does not scale well since it requires DNA sequencing, alignment, variant calling, MHC binding prediction, peptide synthesis, quality control and safety validation for each individual patient. Ideally, it would be possible to identify a set of protein-coding genes within the genome that are uniquely expressed in cancer cells but are also shared across individuals. However, this approach may also need to be personalized to the individual since the immunopeptidome reflects that patient's particular HLA genotype.

The one FDA-approved cancer vaccine targets a non-mutated gene that is overexpressed in cancer cells and not normal cells. This is an attractive model because cancer cells typically overexpress a large number of genes not usually expressed in healthy cells. Dysregulation of transcription and translation is a hallmark of cancer and causes many non-canonical genes to be expressed in tumor cells. Recent research into potential cancer neoantigens has found that over-expression of non-canonical genes, including genes from endogenous retroviruses and transposable elements, is a major source of tumor specific antigens [14,15].

Epigenetic dysregulation is a hallmark of cancer. Cancer cells take on a stem-cell-like state, with the genome taking on a more euchromatic structure. This, in combination with widespread DNA hypomethylation, allows genes that are normally silenced to become expressed. Notably, 40% of the genome is composed of self-propagating DNA elements known as transposable elements (TEs). TEs encode viral-like genes that facilitate reintegration of their sequences throughout the genome. These elements are normally repressed to prevent genomic instability, but have been identified in specific tissues and developmental stages. For example, transposable elements are under selective pressure to retrotranspose in germline cells in order to propagate across generations. There have also been reports of higher expression in brain tissue and stem cells [16–24].

Transposable elements can be subdivided into DNA transposons and retrotransposons. DNA transposons replicate with a DNA intermediate and retrotransposons replicate with an RNA intermediate coupled with a reverse transcription. There are two major classes of retrotransposon: long terminal repeat (LTR) and non-LTR elements [16]. LTR elements are

related to retroviruses. The non-LTR elements contain two subclasses, the short interspersed nuclear elements (SINEs) and the long interspersed nuclear elements (LINEs). LINEs are the only class of TE that contain the necessary protein machinery to retrotransposon. Moreover, autonomous LINEs are required for other TEs, including *Alu* SINEs, to retrotranspose. For this reason, the LINEs are strongly repressed in somatic tissues to prevent genomic instability caused from widespread retrotransposition.

The youngest sub-class of LINEs are the human specific L1HS. These elements are the youngest in the genome and their protein coding sequences are the most strongly conserved. L1HS vaccines have been developed to treat HIV patients because HIV infected cells also over-express transposable elements. The L1HS vaccines were tested in pre-clinical models, including primates, and found to be immunogenic and safe [25]. However, immunization against these elements did not have an effect in protecting macaques from SIV infection, but the vaccines were based on the consensus sequence of transposable elements and endogenous retroelements, which may not capture loci variation required for response [26].

Methods for quantifying TE expression are currently being developed, but these methods are not designed for precision immuno-oncology applications. TE expression methods quantify expression at the class level using a consensus sequence or an average across loci [15,27]. However, this approach is agostic of the targetable vaccine sequence and how it can be present at multiple loci or unique to a specific locus. We developed a novel TE epitope quantification approach to identify unique TE sequences for precision cancer vaccine development. Furthermore, DNA and RNA-level analysis of TE expression assume these sequences are translated, processed, and presented on the MHC, but this assumption is too strong. In response to this, we also developed a mass spectrometry approach that identifies MHC bound peptides. This approach confirms that TE peptides are presented on MHCs and ca be targeted using a cancer vaccine therapy.

We discuss a novel approach based on expression of unique L1HS epitope kmers and peptides in RNA-seq and mass spectrometry data. Our method prioritizes L1HS epitopes that can be uniquely identified to facilitate the identification of vaccine targets. We have developed a novel process for identifying tumor-specific epitopes that are shared among individuals, allowing for a panel of vaccine targets to be synthesized, validated, and distributed across healthcare centers and matched to patient tumors. We quantified normal expression of potential epitopes in several human tissue samples and across developmental stages. We show that L1HS peptides are processed and presented on triple negative breast cancer (TNBC) tumors but not matched normal tissue. Finally, L1HS epitope expression correlates with better survival in TNBC and complete response to checkpoint blockade therapy in melanoma.

Methods and Materials

Implementation of the vaccinaTE software

The vaccinaTE toolkit was developed to facilitate the identification of vaccine targets in cancer populations. There are three main functionalities within the toolkit. The first function is to generate necessary reference files for building a database of unique transposable element (TE) kmers and peptides. The second function is to quantify unique kmers in RNA-seq data. The last function is to generate *in silico* kmers to detect APOBEC expression related to activation of

antiviral response within the cell [28–32]. The vaccinaTE software is written in C++ to scale to genome-wide analysis of transposable element vaccine targets. We also provide several Python routines for preprocessing and analyzing the output of vaccinaTE. The vaccinaTE software tools are available to academic researchers under an Apache 2.0 license.

We automated the identification of transposable element immunotherapy targets using the vaccinaTE toolkit. The underlying database of TE vaccine targets is based on TE annotations from a human reference genome sequence (Figure 1A). The first step of the pipeline identifies unique open reading frames (ORFs) across all TEs (Fig 1B). The generateORFs command takes a genome sequence file and a transposable element annotation file and generates the transcripts and predicted protein sequences for downstream analysis. The ORFs are then used in the findBinders tool to generate a database of all peptides (typically 8, 9, 10, and 11mers) predicted to bind to HLA genotypes of interest. We used netMHCpan-4.0 [33] to predict MHCI binding, which is software available for academic researchers, but if this tool is not available for some users, we also provide support for MHCflurry, which is available under an Apache 2.0 license [34].

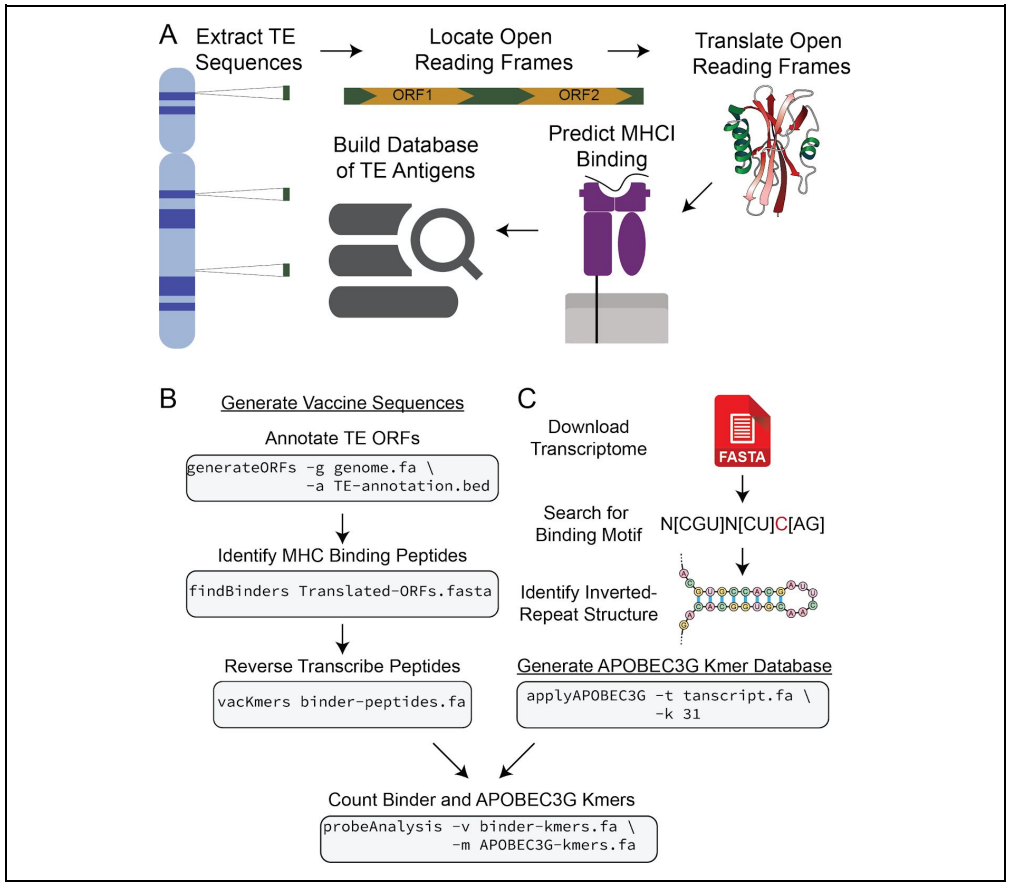


Fig 1. Overview of vaccinaTE tools for quantifying TE epitope kmers and APOBEC mutated kmers. (A) High-level overview of basic approach for developing probes for TE vaccine development. (B) Outline of computational tools available for developing TE vaccine database.

The next step is to identify the peptides within the protein sequences that bind to the HLA genotypes in the patient population. The findBinders script runs netMHCpan-4.0 or MHCflurry, whichever tool is available, to generate a database of potential TE vaccine targets. The database is used to quantify HLA-peptide kmer expression in RNA-seq data. We host several TE databases of interest to the cancer research community on the UCSC Xenahub [35]. The predicted TE epitopes are reverse transcribed back into DNA using the genomic annotation of the TE elements. Unique and multimapping DNA kmers are used for quantifying expression in RNA-seq data. The vacKmer tool reverse transcribes the peptides and matches them to the transposable element loci that could have generated this sequence. This creates the FASTA database that can be used for quantifying druggable transposable element expression in RNA-seq data.

Activation of the APOBEC antiviral response within cells is a hallmark of cancer [28,32]. The APOBEC family of proteins is also involved in repressing transposable elements through several mechanisms, including random mutagenesis of single-stranded RNA and DNA. To provide additional support to transposable element signal, we also generate a random mutagenesis database using published APOBEC mutagenesis motifs [29,30,36]. The APOBEC mutation database along with the MHC bound TE peptides is used for a complete analysis of druggable expression signatures using the probeAnalysis tool. The probeAnalysis tool generates a ranked list of MHC bound peptides and APOBEC kmers for each sample. We provide Python analysis routines for annotating these lists for precision medicine applications.

Generation of LINE-1 Epitope Database

L1HS is the youngest transposable element in the human genome and is one of the few classes of TEs that is autonomous. We hypothesized that L1HS would be strongly repressed in somatic tissue and thus would be an ideal target for developing antitumor vaccine therapies. As the youngest class of TE, L1HS is the most potent at becoming activated in the dysregulated state with cancer cells since these elements have conserved regulatory sequences and coding regions. Despite the strong conservation, there is sufficient variation for L1HS elements to show differential expression across individuals due to differences in transcriptional regulation at different loci, which makes it necessary to personalize vaccines to each tumor knowing that many of these peptides will be shared across individuals.

Of the thousands of L1HS loci, the majority have become degraded and may not generate sufficient protein for vaccine development. We used the L1base2 database to prioritize full-length L1HS elements and L1HS loci with intact ORF2 sequences [37]. We used the hg38 genome annotation for generating L1HS ORFS. The generateORFs tool was used to identify protein-coding regions within L1HS elements. We then investigated the protein domains within ORFs using the Pfam tool [38]. The netMHCpan-4.0 software was applied to the translated

L1HS ORFs for 2427 HLA genotypes. We investigated 8, 9, 10, and 11mers predicted to bind to at least one HLA allele with a minimum percentile rank of 2%. We then mapped these peptides back to the transcript kmers to create a database of corresponding probes which were used in downstream analyses.

Generation of APOBEC kmer Database

We next investigated the ability to quantify *APOBEC* associated RNA editing/DNA mutations using RNA-seq data as input. This is a novel approach that uses *in silico mutated* transcriptome kmers to detect heightened APOBEC activity, a sign of viral infection and TE element expression and an independent predictor of response to checkpoint blockade therapy [39,40]. APOBEC3A is believed to be the main enzyme responsible for the cancer APOBEC signature [28,31,36,41]. These enzymes are typically studied for their DNA mutagenesis signature, but APOBEC3A and 3G were recently found to have an RNA signature that is more specific than the C>T DNA mutagenesis signature. These APOBEC enzymes bind to a specific RNA secondary structure that can be computationally modeled to detect APOBEC activity from RNA-seq data. We investigated the ability to exploit this biological signature to identify additional patients who may benefit from checkpoint blockade therapy.

APOBEC3A is the most active APOBEC in cancer and is involved in repressing viral and retroelement reintegration events in the human genome. APOBEC3A causes a C>T substitution across the genome at the DNA-level, but Sharma et al. (2016) identified a secondary structure preference and a [CT][CT][ATC][TC]C[GA] binding motif preference. Similarly, APOBEC3G was recently found to preferentially bind to a N[CGT]N[CT])C motif. Sharma et al. (2016) found that an inverted repeat was found in 98% of confirmed APOBEC3G mRNA edits due to a hairpin structure that facilitates APOBEC3G binding to RNA. Using the Gencode V32 transcriptome reference [42], we synthetically mutated kmers containing this motif, filtering out kmers that match kmers in the normal transcriptome database as well as kmers related to common polymorphisms in the human population using the dbSNP resource [43]. We then used the kmerCounter script to quantify the number of mutated and normal kmers in RNA-seq samples.

Mass Spectrometry Approach for Identifying Targetable TE Peptides

Current mass spectrometric approaches rely on protein databases for identifying peptides. One of the limitations of this approach is that peptides that are not present in the search database are not identified. Since the focus in the field has been on the identification of canonical proteins, there has been limited attention paid to potential targets from non-canonical protein coding genes, including genes within transposable elements. We have developed a novel approach for identifying potential vaccine targets by first precomputing a database of transposable element epitopes using the vaccinaTE software. We generated a mass spec peptide search database from the Immune Epitope Database (IEDB) [44] of known MHC bound peptides and the predicted L1HS peptides. We then used the MaxQuant software [45,46] to identify these peptides in publically available MHC peptide profiling data for a cohort of triple negative breast cancer patients (PRIDE accession: PXD009738).

Results

Creation of LINE-1 peptide kmer database and APOBEC Signature

In order to quickly quantify the expression of targetable L1HS and APOBEC signatures in healthy and cancer tissue samples, we generated a database of kmers using the Gencode V32 genome and transcriptome reference files and the L1base2.0 annotation for full length L1HS elements and L1HS elements with intact ORF2 sequences [37,42]. This resulted in the generation of 38 unique ORF1 sequences at 56 unique ORF2 sequences. We then analyzed these ORF sequences for conserved protein domains using the Pfam software [38]. We found that ORF1 contained conserved LINE-1 domains, including the L1 RNA Binding Domain (RBD)-Like domain, the double stranded RBD-like domain, and the L1 trimerization domain (SFig 1). ORF2 contained the endonuclease domain, the reverse transcriptase domain, and the domain of unknown function.

We then generated all unique 8, 9, 10, and 11mer peptide sequences using the kmerTools *generate* function. This analysis yield 22,358 unique L1HS peptide kmers. We then used the netMHCpan-4.0 tool predict which of these peptides are likely to bind to at least one of the 2,427 available HLA genotypes. We identified 8,405 unique L1HS peptides predicted to bind to at least one HLA. We applied an additional filter to remove peptides that mapped to canonical proteins and translated open reading frames from the RepeatMasker database which resulted in a final set of 2,316 L1HS epitopes. Filtering for predicted MHC binders generated a preference towards 9mer epitopes (Fig 2A). There were 2,069 kmers that mapped to a single L1HS loci and 247 kmers that mapped to more than 1 loci (SFile 1). The average overlap across HLA alleles was 12% with a single peptide predicted to bind to 407 different HLA alleles. Clustering HLA genotypes using the Gephi force model [47] found that most of the HLA genotypes clustered in a central mass with a small number of HLA types having a significant differences and clustering outside of the main cluster. For example, HLA-A03*02, HLA-A03:01, and HLA-A11:01 clustered separately from the majority of the HLA genotypes due to a small amount of overlap with all other HLAs.

We next investigated hotspots within the L1HS ORFs for generating MHCI binding peptides (Fig 2C & D). The average coverage across the ORF1 and ORF2 sequences was 16 and 11 kmers, respectively. There were hotspots at the junction between the trimerization and RBD-like domain, at the junction between the RBD-like domain and the dsRBD-like domain, and across the sdRBD-like domain in ORF1 (Fig 2C). The similarity across ORF1 sequences was fairly constant across the length of the ORF. The endonuclease domain and the region between the reverse transcriptase and DUF domains were the most highly covered. Surprisingly, we found below average coverage for the reverse transcriptase domain (Fig 2D). The similarity across ORF2 sequences was high across the necessary endonuclease and reverse transcriptase domains, but dropped sharply towards the 3’ end of the element.

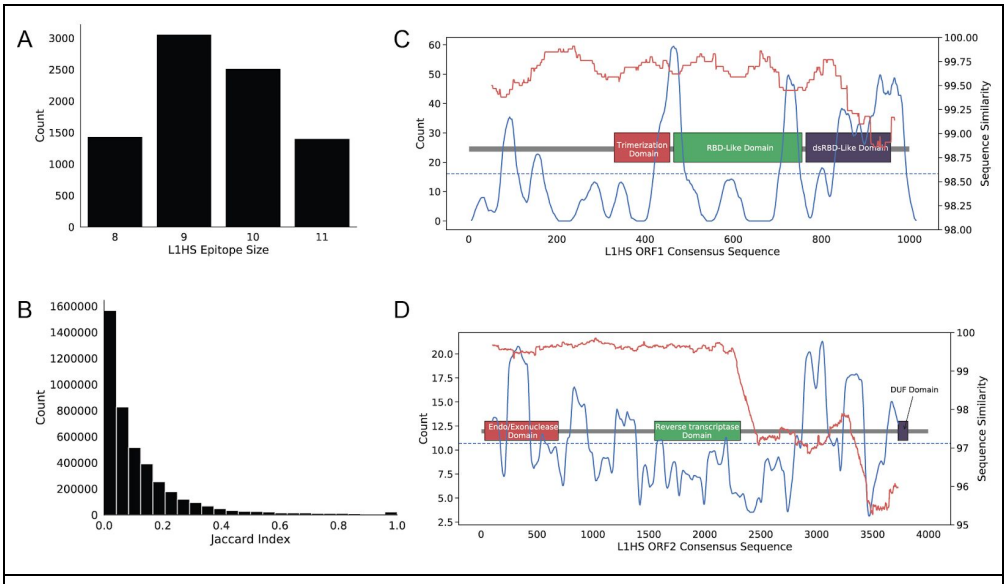


Fig 2. MHCI binding prediction identifies large number of potential L1HS cancer targets. (A) Barplot of netMHCpan-4.0 predicted L1HS epitope lengths. (B) Histogram of Jaccard similarity index values for pairwise comparison of 2,427 HLA alleles. (C & D) Coverage plots for predicted MHCI binders across consensus L1HS ORF1 and ORF2 sequences. Protein domains were annotated using Pfam software. Red line corresponds to the sequence similarity across L1HS ORF multiple sequence alignment.

MHC kmers expressed across developmental stages

The strength of this approach relies on the ability to identify L1HS peptides that are almost never expressed in healthy tissue. This is a challenge to identify because access to healthy tissue is limited, but fortunately a database of healthy human tissue was recently published (N=310) [48]. The mammalian expression database is particularly useful because it includes 7 human tissue types sampled across 23 developmental timepoints.

Transposable expression is expected to be higher during embryonal human developmental stages because regions of the genome that are not usually expressed become activated to support early human development [21]. We identified 1,649 L1HS epitope kmers with a count of at least 2 reads. There were 667 L1HS epitopes that were never detected across all 311 RNA-seq samples. We found 11 L1HS epitopes with decreasing expression across developmental stages and 36 kmers with increasing expression (Kruskal test: adjusted p-value < 0.05).

Overall, we found consistently low expression of L1HS epitope kmers across developmental stages and tissue types (Fig 3). As expected, we found constant expression of L1HS epitope sequences in brain tissues across developmental stages [24]. Similarly, we found constant expression across developmental stages in germline testis tissue [49], but we also

found constant expression in liver tissue (Kruskal test: p-value > 0.05). We found that extracranial tissue including heart and kidney had high levels of expression in the embryo, but significantly lower expression in postnatal samples (Kruskal test: p-value < 0.05).

We found differential expression of several APOPBEC genes with the highest expression at embryonic stages (SFig 2). Interestingly, we observed a spike in L1HS expression and APOBEC expression in the school-age children samples. We found a similar expression pattern in synthetically mutated APOBEC3C kmers where embryonic tissue had the highest number of mutated kmers and later stages had lower expression.

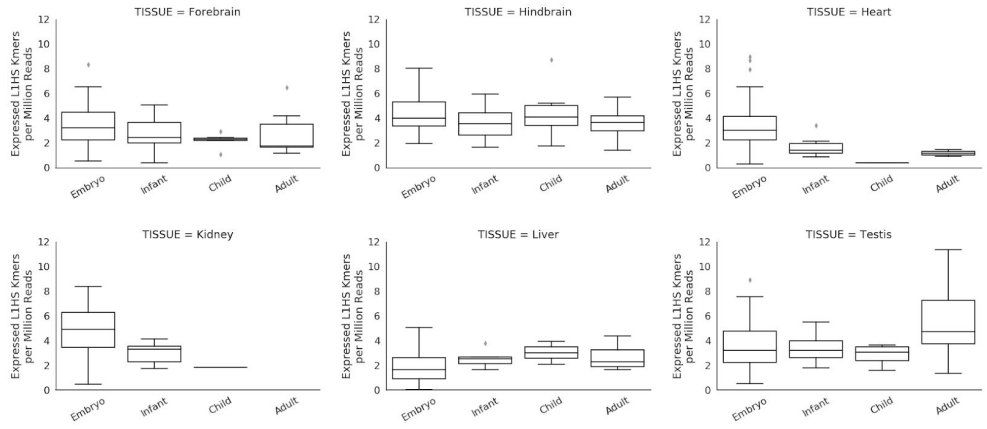


Fig 3. L1HS expression varies based on tissue and developmental stage. Box plots of the number of expressed L1HS epitope kmers per million RNA-seq reads across 6 tissue type and 4 developmental stages.

L1HS peptides are presented on triple negative breast cancer cells but not matched normal cells

Triple negative breast cancer (TNBC) is an aggressive disease that is resistant to multimodal therapy. Immunotherapy has recently been approved as a first-line treatment for TNBC, but response rates remain low and additional strategies are needed to improve durable response rates [50]. Our analysis of RNA-seq identifies epitopes that are likely to be presented on MHC molecules, but there are additional regulatory mechanisms that may prevent some of these peptides from being efficiently processed and presented on the MHC. Recent improvements in the resolution of mass spectrometry equipment has allowed for the identification of short peptides, including MHC-bound peptides [51,52]. Isolation of MHC peptides followed by high-resolution mass spectrometry identifies potential vaccine targets for TNBC.

While it is known that TEs are overexpressed in cancer cells, there has been limited data presented to show that TE peptides are presented by cancer cell MHCs. We used our L1HS epitope database, Immune Epitope Database (IEDB), and a publicly available

immunopeptidome dataset for a cohort of TNBC tumor and matched normal samples investigate whether shared vaccine targets were presented on cancer samples but not matched normal samples (Table 1). Using the MaxQuant search algorithm for mass spectrum matching, we identified three L1HS peptides presented on 5 different patient tumor samples (Table 1). Two of the peptides were shared across different TNBC samples, suggesting that public antigens are similarly processed and presented across individuals with likely different HLA genotypes. This is the first evidence that L1HS peptides are identifiable in patient tumor samples using mass spectrometry analysis and further supports these molecules as viable vaccine targets for combination immunotherapy. Furthermore, we did not detect any L1HS peptides on matched normal tissue samples that were similarly analyzed by MHC peptidome profiling.

Sample	Peptide	L1HS ORF	Protein Domain
Tumor 1	KIKGWRKI	ORF2	Endonuclease domain
Tumor 2	IKRNEQSL	ORF1	Trimerization domain
Tumor 3	IKRNEQSL	ORF1	Trimerization domain
Tumor 4	SFYEASIIL	ORF2	Reverse transcriptase domain
Tumor 5	SFYEASIIL	ORF2	Reverse transcriptase domain

Table 1. MHC-bound L1HS peptides on triple negative breast cancer tumor samples

We then investigated L1HS epitope expression in the TCGA TNBC cohort (N=190). We found 1,428 L1HS epitope kmers with a count of at least 2 reads. There were 162 L1HS epitope sequences that were never detected in the healthy tissue compendium. The average number of expressed kmers per sample 72 and the average number of expressed kmers predicted to bind to one of the patient's HLA alleles was 22. The average overlap in kmers across unrelated TNBC tumor samples with nonzero L1HS kmer expression was 6%. We then correlated the number of expressed HLA-matched L1HS epitope binders with the TNBC patient's overall survival data and discovered a 58% decrease in the Cox proportional hazard ratio (95% CI: 0.19-0.97, $p < 0.05$). The expression of L1HS epitopes may provide a survival benefit because these cells are more easily recognized by the host immune system, which limits tumor growth and extends survival. Further amplification of the anti-L1HS immune response may increase the anti-tumor attack and lead to further reduction in tumor growth and potentially immune-mediated destruction of the tumor.

Shared L1HS epitope expression occurs across TCGA cancer types but not normal samples
We then investigated whether the expressed L1HS epitopes were specific to cancer types or whether there were shared epitopes across diseases (Fig 4). L1HS epitopes were expressed higher in cancer tissue samples than the matched set of postnatal healthy control samples (Fig 4A). We found that most of the epitopes were disease specific (Fig 4B), which is consistent with previous studies in cell-specific expression of permissive loci [49]. There were 9 L1HS epitopes that were expressed in all four TCGA cancer types but not in the healthy control data set.

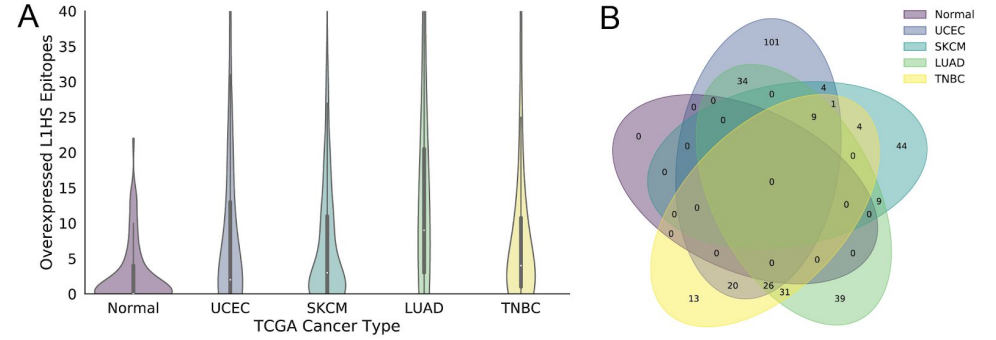


Fig 4. TCGA cancers express L1HS epitope sequences that are not expressed in healthy postnatal human samples. (A) Violin plot of overexpressed L1HS epitope sequences in postnatal healthy samples and four TCGA cancer types. (B) Venn diagram of recurrent (n > 5) L1HS epitopes across cancer types and healthy controls. UCEC: uterine corpus endometrial carcinoma, SKCM: skin cutaneous melanoma, LUAD: lung adenocarcinoma, TNBC: triple negative breast cancer

L1HS Kmers that Correlate with Checkpoint Blockade Response

We propose using TE vaccine therapies in combination with checkpoint blockade therapy. To investigate the clinical efficacy of this approach, we correlated the number of predicted L1HS epitopes with response to checkpoint blockade therapy in a set of 129 melanoma tumor samples. We found that patients with a complete response to checkpoint blockade therapy had higher predicted MHC-bound LINE-1 peptides compared to samples with progressive disease or stable disease (Mann-Whitney U-test p-value < 0.05, Fig 4). Patients with a partial response had the second highest abundance of L1HS epitopes. Amplifying the immune response against these epitopes may increase the response rate to checkpoint blockade therapy.

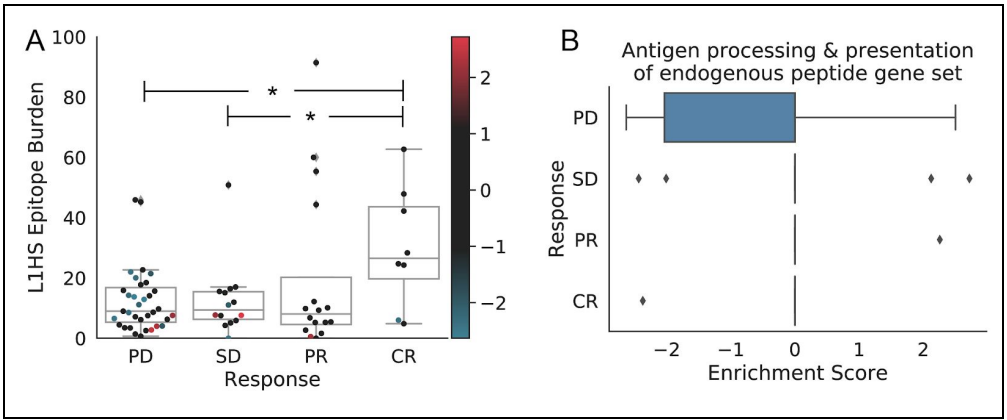


Fig 5. MHC bound peptide burden correlates with complete response to checkpoint blockade therapy. (A) Box plot of the total L1HS epitope expression across melanoma checkpoint blockade response groups (N=73). (B) Gene set enrichment analysis of the Gene Ontology antigen processing and presentation of endogenous peptide gene set.

Discussion

Checkpoint blockade therapy has generated remarkable responses in a subset of cancer patients, but further research into combination therapies is needed to increase the number of patients who benefit [4,10,53]. We have developed a computational framework for prioritizing transposable element (TE) epitopes for personalized cancer vaccine therapies. We hypothesize that combination TE vaccine immunization and checkpoint blockade therapy may tip the balance in favor of immune-mediated destruction of the tumor. A combination vaccine and checkpoint blockade therapy was used recently to treat glioblastoma and this study found that these therapies work synergistically [10]. The power of the immune system to destroy cancer at a cellular level, throughout the body, and to maintain a memory against recurrence allows for this therapeutic approach to achieve durable response and potentially cure patients of their cancer.

For this approach to be successful, we need to identify peptides that are expressed in cancer cells but not healthy cells. To address this concern, we applied our approach to a large cohort of 311 healthy RNA-seq datasets across 23 developmental stages and 7 tissue types. While we detected L1HS expression in these samples, we found that cancer cells express additional L1HS peptides that were never detected in the healthy control cohort. This suggests that it is possible to identify a subset of L1HS peptides that are only expressed in cancer cells, so amplification of an immune response against these peptides may not generate off-target effects that may be toxic to the patient.

Much of the data on TE expression in the literature is based on RNA-seq data, but whether these elements generate peptides that are presented on human cancer cell MHCs has not been sufficiently investigated. We provide evidence that indeed L1HS peptides are presented by cancer cells in triple negative breast cancer tumors but not matched normal tissue samples. This shows that not only are these elements aberrantly expressed in cancer cells, but these the TE transcripts are translated into proteins and these proteins are properly processed and presented by MHC molecules. This further underscores that druggability of these vaccine targets. Moreover, we found that expression of predicted MHC bound TE peptides lead to a 58% reduction in the Cox proportional hazards ratio for the TCGA TNBC cohort. This underscores the benefit of these molecules for treating cancer, since the expression of these molecules correlates with better patient outcomes, presumably since these molecules may induce immune responses that limit tumor growth.

Lastly, we correlated L1HS epitopes expression with response to checkpoint blockade therapy in melanoma [54,55]. Surprisingly, we found that the expression of L1HS epitopes correlated with the complete response group of melanoma patients. This suggests that these patients by chance had higher L1HS epitope expression and were naturally immunizing their

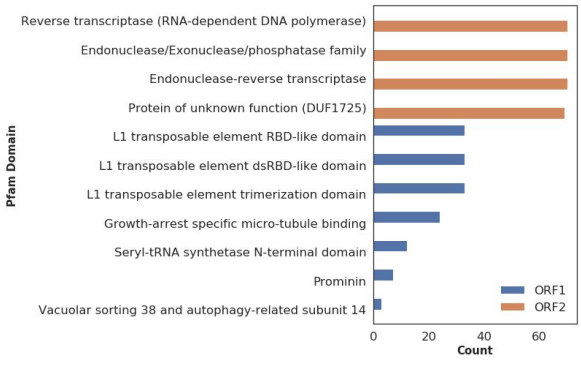
immune system against the cancer cells. Introduction of checkpoint blockade therapy may have then removed the immunosuppressive effect allowing cytotoxic T-cells to eradicate the tumor. Notably, the expression of these peptides were not zero in many of the non-responders or partial responders, but the balance between expression of these targets and the circulating T-cells able to recognize the cancer cells may not have been in these patients favor and thus there was no response or a limited response that the cancer cells quickly rebounded from.

These results provide hope that further expansion of T-cells that are able to recognize cancer cells through identification of tumor-specific TE expression analysis may increase the number of patients that experience durable responses. One of the many strengths of this approach is that these peptides are shared across individuals. We propose a novel therapeutic paradigm for matching tumors to a repository of validated cancer vaccines for efficient distribution and administration of therapy. This includes the screening of large cancer RNA-seq data sets for the most commonly overexpressed epitopes, prioritizing epitopes that correlate with patient benefit. We then propose synthesizing, quality control, and validation of these peptides before mass production and distribution to treat cancer at scale.

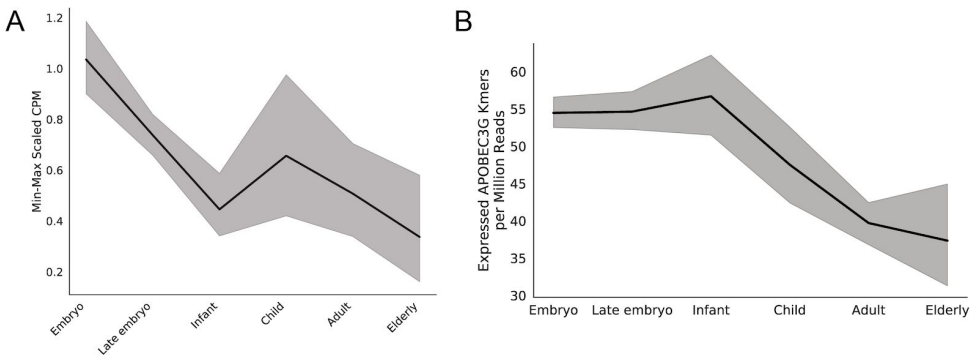
Conclusion

Transposable elements make up ~40% of the human genome, encode viral like proteins, and are strongly repressed in somatic cells. This makes them attractive targets for cancer vaccine development, but the sequence similarity and complexity of the genome makes it difficult to identify which peptides to prioritize. We developed an exciting new computational framework based on unique expression of MHC bound peptide kmers. This approach was able to identify expression of druggable L1HS epitopes that correlated with better survival outcomes and complete response to checkpoint blockade therapy. Future research investigating whether expansion of the T-cell response to these peptides in cancer patients generates stronger antitumor responses.

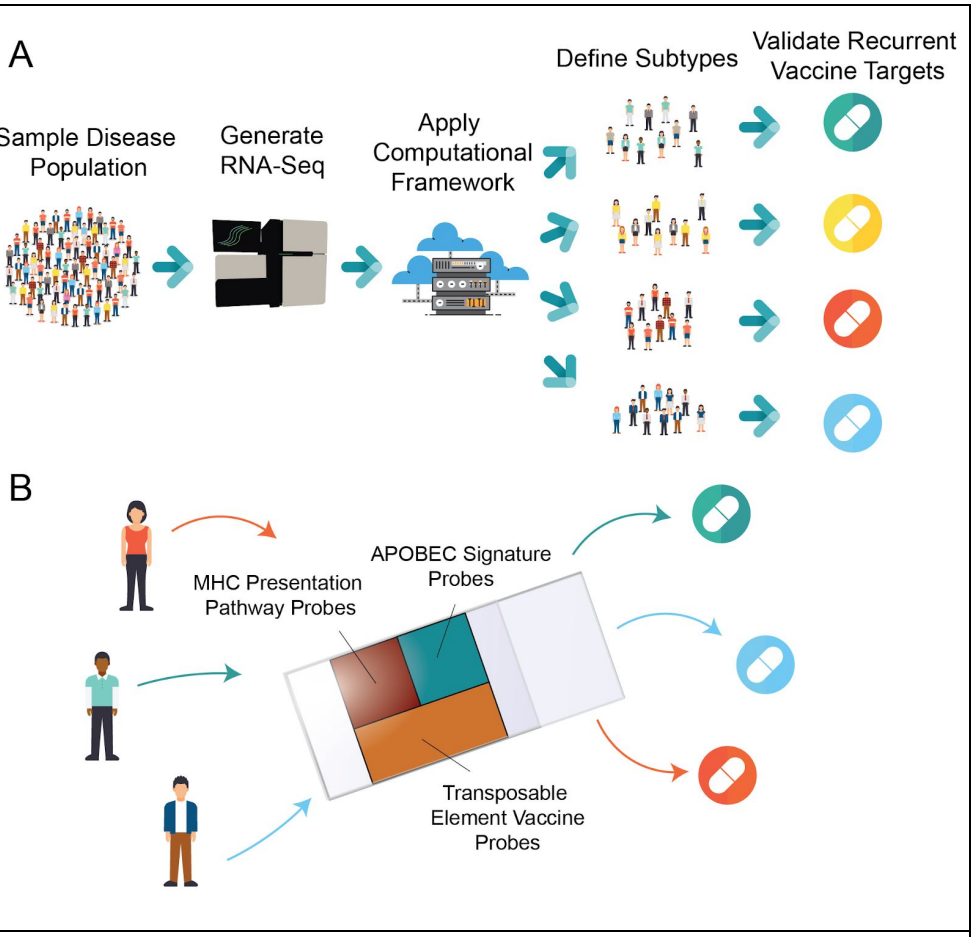
Supplement



SFig1. Predicted L1HS open reading frames contain expected protein coding domains



SFig 2. APOBEC3C expression is highest in embryonic tissue. (A) Average APOBEC3 expression across 7 tissue types and 23 developmental stages. (B) Average APOBEC3G kmer expression across the same cohort.



SFig 3. Process for prioritizing shared TE vaccine targets and matching patient tumor samples to repository of validated vaccine therapies. (A) Process for screening cancer RNA-seq data and defining subtype groups based on shared TE epitope expression. (B) Microarray diagnostic matches patient tumor samples to available vaccine therapies while also correlating TE expression with MHC presentation and APOBEC expression signatures.

References

1. Murphy SL. Mortality in the United States, 2017. 2018; 8.
2. Wilking N, Karolinska Institutet, Solna, Sweden, Lopes G, Sylvester Comprehensive Cancer Center, University of Miami, FL, US, Meier K, HKK Soltau, Lower Saxony & Heidekreis-Klinikum GmbH, Soltau, Germany, et al. Can we Continue to Afford Access to Cancer Treatment? Eur Oncol Haematol. 2017;13: 114. doi:10.17925/EOH.2017.13.02.114
3. CD28 and CTLA-4 have opposing effects on the response of T cells to stimulation. J Exp Med. 1995;182: 459–465.
4. Zappasodi R, Merghoub T, Wolchok JD. Emerging Concepts for Immune Checkpoint Blockade-Based Combination Therapies. Cancer Cell. 2018;33: 581–598. doi:10.1016/j.ccell.2018.03.005
5. Ribas A, Wolchok JD. Cancer immunotherapy using checkpoint blockade. Science. 2018;359: 1350–1355. doi:10.1126/science.aar4060
6. Pitt JM, Vétizou M, Daillère R, Roberti MP, Yamazaki T, Routy B, et al. Resistance Mechanisms to Immune-Checkpoint Blockade in Cancer: Tumor-Intrinsic and -Extrinsic Factors. 2016;44. doi:10.1016/j.immuni.2016.06.001
7. Simon S, Labarriere N. PD-1 expression on tumor-specific T cells: Friend or foe for immunotherapy? Oncoimmunology. 2017;7. doi:10.1080/2162402X.2017.1364828
8. Seidel JA, Otsuka A, Kabashima K. Anti-PD-1 and Anti-CTLA-4 Therapies in Cancer: Mechanisms of Action, Efficacy, and Limitations. Front Oncol. 2018;8. doi:10.3389/fonc.2018.00086
9. Khair DO, Bax HJ, Mele S, Crescioli S, Pellizzari G, Khiabany A, et al. Combining Immune Checkpoint Inhibitors: Established and Emerging Targets and Strategies to Improve Outcomes in Melanoma. Front Immunol. 2019;10. doi:10.3389/fimmu.2019.00453
10. Liu C, Schaettler M, Bowman-Kirigin J, Kobayashi D, Miller C, Johanns T, et al. IMMU-09. COMBINATION IMMUNE TREATMENT OF A HIGHLY AGGRESSIVE ORTHOTOPIC MURINE GLIOBLASTOMA WITH CHECKPOINT BLOCKADE AND MULTI-VALENT NEOANTIGEN VACCINATION. Neuro-Oncol. 2019;21: vi120–vi121. doi:10.1093/neuonc/noz175.503
11. Lee KL, Benz SC, Hicks KC, Nguyen A, Gameiro SR, Palena C, et al. Efficient Tumor Clearance and Diversified Immunity through Neoepitope Vaccines and Combinatorial Immunotherapy. Cancer Immunol Res. 2019;7: 1359–1370. doi:10.1158/2326-6066.CIR-18-0620
12. Burg SH van der, Arens R, Ossendorp F, Hall T van, Melief CJM. Vaccines for established cancer: overcoming the challenges posed by immune evasion. Nat Rev Cancer. 2016;16: 219–233. doi:10.1038/nrc.2016.16
13. Banchereau J, Palucka K. Cancer vaccines on the move. Nat Rev Clin Oncol. 2018;15: 9–10. doi:10.1038/nrclinonc.2017.149
14. Laumont CM, Vincent K, Hesnard L, Audemard É, Bonneil É, Laverdure J-P, et al. Noncoding regions are the main source of targetable tumor-specific antigens. Sci Transl Med. 2018;10. doi:10.1126/scitranslmed.aau5516
15. Kong Y, Rose CM, Cass AA, Williams AG, Darwish M, Lianoglou S, et al. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. Nat Commun. 2019;10: 1–14. doi:10.1038/s41467-019-13035-2
16. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19: 199. doi:10.1186/s13059-018-1577-z

17. Finnegan DJ. Transposable elements: How non-LTR retrotransposons do it. Curr Biol. 1997;7: R245–R248. doi:10.1016/S0960-9822(06)00112-6
18. Kassiotis G, Stoye JP. Immune responses to endogenous retroelements: taking the bad with the good. Nat Rev Immunol. 2016;16: 207–219. doi:10.1038/nri.2016.27
19. Burns KH. Our Conflict with Transposable Elements and Its Implications for Human Disease. Annu Rev Pathol Mech Dis. 2020;15: 51–70. doi:10.1146/annurev-pathmechdis-012419-032633
20. De Cecco M, Criscione SW, Peterson AL, Neretti N, Sedivy JM, Kreiling JA. Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. Aging. 2013;5: 867–883.
21. Gerdes P, Richardson SR, Mager DL, Faulkner GJ. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. Genome Biol. 2016;17: 100. doi:10.1186/s13059-016-0965-5
22. Chung N, Jonaid GM, Quinton S, Ross A, Sexton CE, Alberto A, et al. Transcriptome analyses of tumor-adjacent somatic tissues reveal genes co-expressed with transposable elements. Mob DNA. 2019;10: 39. doi:10.1186/s13100-019-0180-5
23. Saleh A, Macia A, Muotri AR. Transposable Elements, Inflammation, and Neurological Disease. Front Neurol. 2019;10. doi:10.3389/fneur.2019.00894
24. Terry DM, Devine SE. Aberrantly High Levels of Somatic LINE-1 Expression and Retrotransposition in Human Neurological Disorders. Front Genet. 2020;10. doi:10.3389/fgene.2019.01244
25. Sacha JB, Kim I-J, Chen L, Ullah JH, Goodwin DA, Simmons HA, et al. Vaccination with Cancer- and HIV Infection-Associated Endogenous Retrotransposable Elements Is Safe and Immunogenic. J Immunol. 2012;189: 1467–1479. doi:10.4049/jimmunol.1200079
26. Sheppard NC, Jones RB, Burwitz BJ, Nimityongskul FA, Newman LP, Buechler MB, et al. Vaccination against Endogenous Retrotransposable Element Consensus Sequences Does Not Protect Rhesus Macaques from SIVsmE660 Infection and Replication. PLOS ONE. 2014;9: e92012. doi:10.1371/journal.pone.0092012
27. Jeong H-H, Yalamanchili HK, Guo C, Shulman JM, Liu Z. An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. Pac Symp Biocomput Pac Symp Biocomput. 2018;23: 168–179.
28. Swanton C, McGranahan N, Starrett GJ, Harris RS. APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. Cancer Discov. 2015;5: 704–712. doi:10.1158/2159-8290.CD-15-0344
29. Sharma S, Baysal BE. Stem-loop structure preference for site-specific RNA editing by APOBEC3A and APOBEC3G. PeerJ. 2017;5: e4136. doi:10.7717/peerj.4136
30. Sharma S, Patnaik SK, Taggart RT, Baysal BE. The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. Sci Rep. 2016;6: 1–12. doi:10.1038/srep39100
31. Refsland EW, Harris RS. The APOBEC3 Family of Retroelement Restriction Factors. In: Cullen BR, editor. Intrinsic Immunity. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. pp. 1–27. doi:10.1007/978-3-642-37765-5_1
32. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013;45: 970.
33. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol. 2017;199: 3360–3368. doi:10.4049/jimmunol.1700893

34. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell Syst. 2018;7: 129-132.e4. doi:10.1016/j.cels.2018.05.014
35. Goldman M, Craft B, Kamath A, Brooks A, Zhu J, Haussler D. The UCSC Xena Platform for cancer genomics data visualization and interpretation. bioRxiv. 2018; 326470. doi:10.1101/326470
36. Sharma S, Patnaik SK, Taggart RT, Kannisto ED, Enriquez SM, Gollnick P, et al. APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. Nat Commun. 2015;6: 1–15. doi:10.1038/ncomms7881
37. Penzkofer T, Jäger M, Figlerowicz M, Badge R, Mundlos S, Robinson PN, et al. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. Nucleic Acids Res. 2017;45: D68–D73. doi:10.1093/nar/gkw925
38. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47: D427–D432. doi:10.1093/nar/gky995
39. Boichard A, Pham TV, Yeerna H, Goodman A, Tamayo P, Lippman S, et al. APOBEC-related mutagenesis and neo-peptide hydrophobicity: implications for response to immunotherapy. Oncoimmunology. 2018;8. doi:10.1080/2162402X.2018.1550341
40. Wang S, Jia M, He Z, Liu X-S. APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer. Oncogene. 2018;37: 3924–3936. doi:10.1038/s41388-018-0245-9
41. Burgess DJ. Switching APOBEC mutation signatures. Nat Rev Genet. 2019;20: 253–253. doi:10.1038/s41576-019-0116-4
42. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22: 1760–1774.
43. Kitts A, Sherry S. The single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation. NCBI Handb McEntyre J Ostell J Eds Bethesda MD US Natl Cent Biotechnol Inf. 2002.
44. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2015;43: D405–D412.
45. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26: 1367–1372.
46. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. J Proteome Res. 2011;10: 1794–1805. doi:10.1021/pr101065j
47. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. Third international AAAI conference on weblogs and social media. 2009.
48. Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen C, Shao Y, et al. Gene expression across mammalian organ development. Nature. 2019;571: 505–509. doi:10.1038/s41586-019-1338-5
49. Philippe C, Vargas-Landin DB, Doucet AJ, van Essen D, Vera-Otarola J, Kuciak M, et al. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. Burns K, editor. eLife. 2016;5: e13926. doi:10.7554/eLife.13926
50. Marra A, Viale G, Curigliano G. Recent advances in triple negative breast cancer: the immunotherapy era. BMC Med. 2019;17: 90. doi:10.1186/s12916-019-1326-5

51. Bassani-Sternberg M. Mass Spectrometry Based Immunopeptidomics for the Discovery of Cancer Neoantigens. Methods Mol Biol Clifton NJ. 2018;1719: 209–221. doi:10.1007/978-1-4939-7537-2_14
52. Purcell AW, Ramarathinam SH, Ternette N. Mass spectrometry–based identification of MHC-bound peptides for immunopeptidomics. Nat Protoc. 2019;14: 1687–1707. doi:10.1038/s41596-019-0133-y
53. Minn AJ, Wherry EJ. Combination Cancer Therapies with Immune Checkpoint Blockade: Convergence on Interferon Signaling. Cell. 2016;165: 272–275. doi:10.1016/j.cell.2016.03.031
54. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, et al. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell. 2016;165: 35–44. doi:10.1016/j.cell.2016.02.065
55. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, et al. Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. Cell. 2017;171: 934-949.e16. doi:10.1016/j.cell.2017.09.028

Part IV

Evaluating Preclinical Models to Accelerate Development of Targeted Therapies for Pediatric Cancers

Introduction

The development of novel therapies depends on the availability of preclinical models of human disease. As it is unethical to test unproven therapies on patients, preclinical models like cell-lines and mouse models are used to validate novel therapies. One of the biggest challenges in current drug development efforts is that most drugs that go into clinical trials fail despite showing efficacy in preclinical models. This suggests that preclinical models do not accurately reflect human diseases. To address this problem, I have developed a collaboration with Alejandro Sweet-Cordero at UCSF, who is a leading clinical oncologist and expert in patient-derived xenograft models (PDX). PDX models are generated by implanting human tumor tissue into an immunosuppressed mouse. While it is thought that PDX models better reflect human disease because they consist of human cancer tissue, it is unclear what changes occur in the mouse that may influence tumor biology. I have designed and implemented a Bayesian hierarchical model to robustly learn the evolution of PDX-specific expression. This analysis found that most genes (¿ 90%) are conserved in the PDX. The genes that are differentially expressed are associated with expected changes in the PDX, including immune and stromal expression markers. We have then used these genes to identify pathways that are differentially expressed and will share these pathways with the PDX modeling community. The goal of this analysis is to accelerate drug development by identifying pathways that are conserved in PDX models and better reflect human disease.

Chapter 6

Bayesian hierarchical modeling framework for accelerating drug development using pediatric patient derived xenografts

Introduction

Large patient-derived xenograft repositories provide a great resource for the preclinical research community. However, there has been limited investigation into the biological features of these models with respect to molecular data, including whole transcriptome sequencing analysis. A Bayesian hierarcical modeling framework provides a method of analyzing data with small sample sizes and biological similarities across related diseases.

Bayesian statistics is well-suited to building hierarchical models. The goal of Bayesian inference is to learn the probability of the parameters given the data. To calculate the probability of the parameters requires Bayes theorem

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

where θ represents the parameters of your model and x represents the data. The prior distribution $P(\theta)$ expresses your belief in the model before any data is observed. The likelihood $P(x|\theta)$ expresses the probability of observing the data given your model. Bayes theorem updates your belief in the system using the prior and likelihood distributions to generate the posterior distribution $P(\theta|x)$. The posterior distribution can now be used as the prior when a new set of data is generated. The marginal probability of the data $P(x)$ is a normalizing constant, which does not influence inference and so the posterior is often represented in an unnormalized form.

$$P(\theta|x) \propto P(x|\theta)P(\theta)$$

In a hierarchical model, the different levels of the model are encoded in the prior distribution. Repeat application of conditional probability relates each level of the hierarchy [10, 16].

$$\begin{aligned} P(\theta,\phi|x) &\propto P(x,\theta|\phi)P(\phi) \\ &= P(x|\theta,\phi)P(\theta|\phi)P(\phi) \\ &= P(x|\theta)P(\theta|\phi)P(\phi) \end{aligned}$$

The prior for one level of the model is the likelihood for the next level in the hierarchy and so on until the top of the model. The last equation is simplified because the likelihood of

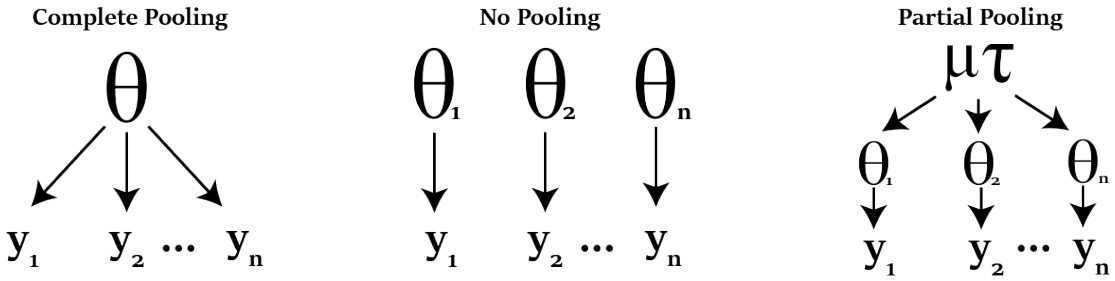


Figure 6.1: Models for the Treehouse analysis. Treehouse pan-cancer analysis is an example of the complete pooling model. In a complete pooling model, distinct groups of data are not modeled individually. Pan-cancer analysis does not account for different data features like the age, cancer type, and gender. Pan-disease analysis is a form of no-pooling model where each disease is modeled separately without considering information learned from other cancer types. A hierarchical models is a compromise between the complete and no pooling model. In a hierarchical model, separate parameters are learned for each data group while also sharing information through prior distributions on the group specific parameters.

the data does not depend on the prior on θ. Here, I use weakly informative prior distributions including the normal and half Cauchy distributions in order to improve compuational efficiency for learning the model parameters [10].

One of the benefits of hierarchical modeling is shrinkage. To correct for sampling errors that arise from using a small sample size, the hierarchical model pulls the data cluster's distribution closer to the population mean. Parameters shrink towards the population mean because the prior distribution is stronger than the likelihood. This is a valuable feature for the Treehouse compendium because it helps control for erroneous inferences when the number of samples is limited. Pediatric gene expression profiles are limited in the compendium, so data shrinkage can be used to help control estimates for rare pediatric cancer. The shrinkage features comes from adaptive regularization. The prior distributions learn the expected distribution of parameters and samples that do not conform to the population level distribution are corrected.

Therefore, the model does not exaggerate effects that result from small sample sizes. Regularizing priors introduce skepticism into the model so the model does overestimate when it observes surprising data that may be due to errors in measurement [30]. As the number of pediatric samples increases, the Treehouse hierarchical model becomes more confident in estimating pediatric gene expression differences and the shrinkage effect relaxes.

6.0.0.1 Varying intercept and slope models to predict pediatric gene expression

Genes are expressed at different levels for different tissues. In addition to tissue specific expression, there are also biological features that influence gene expression across individuals. For example, age and gender are correlated with expression of some genes. A varying effects model where the mean and the effect of biological features change depending on the tissue can be used to make better predictions of gene expression. For example, a hierarchical model can identify sex-linked expression, but the current pan-cancer and pan-disease analyses are not able to detect sex-linked expression. An example of sex-linked expression that has been associated with cancer is the XIST gene [45]. XIST controls X-chromosome silencing in females and is not usually expressed in males (Figure **??**). This is a clear example where assuming male and female gene expression comes from the same distribution leads to an exaggerated estimation of the outlier threshold. It is therefore difficult to identify potential cases where under-expression of XIST in females may contribute to their cancer. While the incidence of cancer is equal across boys and girls, boys tend to respond worse to therapy. An investigation into sex-linked gene expression may yield insights into the differences in response to cancer therapies for boys and girls.

By estimating disease and tissue specific parameters using biological features as predictors, I learned the expected gene expression as well as the influence these parameters have on gene expression. The varying effects model can be used to identify gene expression outliers by generating the posterior predictive distribution for that patient and determining if that patient's gene expression is an outlier. Alternatively, the hierarchical model can be used to infer latent variables associated with cancer gene expression and classify patients into normal and abnormal gene expression categories. The first normal linear model will be explored first and then a hierarchical mixture model will be developed to better resolve cancer-associated expression.

The initial hierarchical model will be a normal linear models. There are many ways to represent a linear model, but I prefer to use a representation that describes the sampling process. Here, the data is sampled from a normal distribution and the mean of the normal distribution is calculated as a linear combination of an intercept and slope term.

$$y_i \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu = \alpha + \beta x_i$$

$$\alpha \sim \mathcal{N}(0, 100)$$

$$\beta \sim \mathcal{N}(0, 100)$$

Here, the variance is known and we are trying to learn the mean μ and the parameters for describing μ which are α and β. There are prior distributions on α and β that do not provide any strong information, so the model will learn these values from observing the data. As written, this model describes a complete pooling linear model. An alternative model is the varying slopes

and intercepts model which learns separate α and β parameters for each data cluster.

Varying intercepts allow different data clusters to have different mean values. For the Treehouse compendium, each disease cluster can have its own mean expression level. The varying slopes allow predictors to influence the expected gene expression differently for different data clusters. For instance, the effect of being pediatric may be stronger for some tissues than others. Varying intercept and slope models can be used to learn how biological and clinical features influence expression for different disease classes.

Even relatively simple Bayesian models require solving complex integrals. This is one reason Bayesian models have not been more widely adopted. Complex integrals can be approximated using Markov Chain Monte Carlo sampling methods. Probabilistic programming closes the gap between statistical modeling and computer programming. Now, the computational framework for inference is expressed in a form that is close to the mathematical representation. The probabilistic programming environment handles MCMC sampling from the posterior distribution. The two most popular probabilistic programming libraries are STAN and PyMC3. STAN developed its own probabilistic programming language, but has interfaces in common programming languages like R and Python. PyMC3 is another probabilistic programming language that uses the widely adopted Python language and optimizes gradient calculations using the Theano library.

As a proof of concept experiment, I developed a CDK4 varying intercept model for the Treehouse compendium using the PyMC3 library (Model 6.1). Each cancer type in the Treehouse compendium gets its own intercept in the hierarchical model. The prior for the intercept is shared across all cancer types, so cancer types with a limited number of samples shrink to-

wards the pan-cancer mean. This model is equivalent to the no-pooling pan-disease model, but there is shrinkage for diseases that have a low number of samples. This varying intercept model is the first level of a more complex hierarchical model. For instance, the posterior distributions for this model will be used to model clinical features at the next level.

$$
\begin{aligned}
y_{disease} &\sim \mathcal{N}(\mu_{disease}, \sigma^2) \\
\mu_{disease} &\sim \mathcal{N}(\mu_{gene}, \sigma^2_{gene}) \\
\mu_{gene} &\sim \mathcal{N}(0, 100) \\
\sigma_{gene} &\sim \text{HalfCauchy}(5) \\
\sigma &\sim \text{HalfCauchy}(5)
\end{aligned}
\tag{6.1}
$$

In this manuscript, I describe a Bayesian hierarchical model I developed to learn which genes are differentially expressed between pediatric sarcoma PDXs and matched patient tumors. This was the first study of matched pediatric PDX tumors, and the results showed that PDXs capture the tissue-of-origin signal better than cell lines. I also proposed a framework for developing PDX models that better reflect patient tumors using the tumor microenvironment signal to prioritize tumors.

Bayesian hierarchical modeling framework for accelerating drug development using pediatric patient derived xenografts

Jacob Pfeil, Leanne Sayles, Alex Lee, Geoff Lyles, Sofie Salama, David Haussler, Olena Vaske, Alejandro Sweet-Cordero

Abstract

Molecularly targeted therapies inhibit specific cancer pathways and have fewer harmful side effects than broadly toxic chemotherapies. However, the development of targeted therapies for childhood cancers has lagged behind that of adult cancers. One factor influencing this is the lack of accurate preclinical models for validating novel drug targets. Cancer cell lines are the first line of validation studies, but cancer cell lines do not capture the full complexity of a human tumor. Patient derived xenografts (PDXs) are presumed to be more accurate models, but there has not been sufficient research into how well these models reflect human cancer, particularly with respect to differences in gene expression. We have developed a novel gene expression framework for evaluating PDX models. We show that PDX gene expression better reflects patient disease populations than cancer cell lines using TumorMap analysis. We then apply a Bayesian hierarchical model to a cohort of pediatric sarcoma PDXs to infer consistently differentially expressed genes between PDXs and matched patient samples. We found that the majority of genes are not differentially expressed (>90%) and that removing differentially expressed genes from analysis causes osteosarcoma PDXs and unmatched osteosarcoma samples to cluster, suggesting that we have identified the genes that differentiate osteosarcoma PDXs from patient samples. Lastly, we provide two examples for how this database can be used to accelerate the development of novel therapies for pediatric cancers.

Introduction

While pediatric cancers generally have high survival rates, patients who relapse have few treatment options and a low rate of survival [1,2]. The development of novel therapies depends on the availability of accurate preclinical models [3]. Before a new drug is tested in humans, the drug is first introduced to a preclinical model, including cancer cell lines and mouse models. Preclinical models are used as a surrogate for human subjects and the results of preclinical experiments are used as preliminary data for the investigational new drug application, which is a necessary step towards opening clinical trials in humans.

Preclinical models are used to test for toxicity and efficacy, but it has become clear that the results of these experiments can be misleading [3]. Only 5% of new drugs finish phase III clinical trials, despite showing efficacy in preclinical models [4–6]. The current drug development paradigm wastes time and money, and is a major factor contributing to the high cost of drugs [7]. The identification of preclinical models that better reflect the patients entering clinical trials may accelerate the development of effective cancer therapies since the efficacy in preclinical models will more likely correlate with efficacy in human subjects.

One of the most widely used preclinical models for testing anticancer therapies is the cancer cell line [8]. Cell lines reduce the complexity of the cancer system by decreasing heterogeneity as well as the effect of the host immune and stromal cells [9]. Cell lines are easy to distribute across laboratories which allows for cross-institutional analysis. Cancer cell lines also grow quickly and can be expanded to support a large number of experiments. Despite the experimental conveniences of cancer cell lines, there are several challenges that make these models less accurate for modeling cancer.

Cancer cell lines are challenging to create for individual patients because it requires the cancer cells to be able to grow in culture, which is a significantly different environment compared to the tumor microenvironment. The success rate for generating cell lines is around ~20% [8], so most tumors will not generate a cancer cell line. There may also be selective pressure for particular tumor subtypes, which biases downstream validation experiments towards particular tumor subtypes. The tumor subtypes that are more likely to generate cancer cell lines may be relatively rare in the disease population, leading to low success rates in clinical trials [5,10]. Cancer cell lines adapt to the growing in culture and may lose important genetic and transcriptomic features of the original tumor sample [11]. For example, cancer cell lines are suspended in medium and thus lose cell-cell interactions that are known to play important roles in cancer [12].

Patient-derived xenografts (PDXs) are an alternative preclinical model that is thought to more accurately reflect human tumors. PDXs are created by transplanting human cancer tissue into an immunosuppressed mouse model. The PDX supports the human cancer cell growth and allows for PDX tumors to be passaged to additional mice to maintain the original tumor. Contrary to traditional mouse models where mutations are engineered into the mouse line to induce a specific cancer phenotype, PDXs use human cancer cells in a controlled tumor microenvironment and may better reflect human cancer [13].

While PDXs are presumed to be more accurate, there has been limited investigation into the accuracy of these models. PDX tumors are comprised of human cancer cells, but these cells grow in a significantly different microenvironment than the original human tumor sample. Ben-David et al. (2017) discovered mouse-specific evolution of copy number alterations in the PDX that correlated with response to targeted therapies [14]. In addition to copy number changes, it is known that human immune and stromal cells cannot proliferate in the mouse model, so these cells are quickly replaced with mouse counterparts [15,16].

How the exchange of immune and stromal cells influence PDX tumors is currently unknown, but the lack of selective pressure imposed on the cancer cells by the human immune system may accelerate the accumulation of mutations and the downregulation of regulatory immune mechanisms. Tumor heterogeneity can also dramatically change in response to the changing microenvironment such that some cancer cell clones are lost and others become more abundant [16]. While these changes are likely associated with differences in selective pressure within the PDX, it is unclear if there are patient tumor subtypes that may be more accurately modeled in the PDX system. Identification of these subtypes would facilitate the validation of specific therapies for cancer subtypes, which is becoming a widely adopted strategy for treating cancer [13].

The tumor microenvironment plays an essential role in tumor biology and is a feature that has been overlooked in previous PDX credentialing studies. PDX mice do not have a fully functional immune system, which allows the human cancer cells to grow unchecked. This, however, may influence the selective pressure placed on PDX tumors and lead to mouse specific evolution. Comparing PDXs in the context of the tumor microenvironment is a novel approach for evaluating the accuracy of PDX models and may yield novel insights into how to best generate and interpret results from PDX models. Here, we describe a novel framework for comparing gene expression between matched PDX and patient tumor samples. Our analysis identifies the known differences between PDXs, while also proposing a novel preclinical modeling strategy that uses the database of differentially expressed genes to prioritize patients and models for pediatric drug development.

Materials & Methods

PDX Generation

We initially implant tumor fragments in the subrenal capsule to establish PDXs, followed by orthotopic implantation. All mice are monitored for 1 year to determine if the PDX was successful. Mice carrying primary PDXs will be sacrificed, PDXs will be removed and ½ will be used to FACs sort tumor cells and separate them from the mouse stroma in preparation for sequencing.

Pediatric Preclinical Testing Consortium and UCSC Treehouse Gene Expression Data

We downloaded the publicly available TARGET and Pediatric Preclinical Testing Consortium (PPTC) gene expression data from the pediatric cBioportal website [17,18]. We also downloaded PDX and matched patient samples available through the UCSC Treehouse gene expression compendium [19] published on the UCSC Xena browser [20,21]. Gene expression transcript per million mapped read values (TPM) were normalized using a log2(TPM + 1) transformation.

Clustering Analysis

TumorMap allows interactive exploration of large cancer datasets and the visualization of individual tumors in the context of other cancers [22]. We used TumorMap analysis to identify similarities between large cohorts of PDXs and related human tumor samples. We included all genes with a mean expression greater than 1 log2(TPM + 1).

Gene Expression Analysis

Complete pooling maximally underfits and no-pooling maximally overfits data, but hierarchical modeling strikes a balance between the two [23]. In a hierarchical model, each data cluster is modeled separately, but information is shared across levels of the hierarchy. We developed a Bayesian hierarchical model to learn statistically significant differences between PDXs and matched patient samples. The genewise differences for each gene was modeled as a normal distribution, with a prior over the global difference between PDX and human tumor samples. We performed a power analysis to determine the number of samples needed to identify 80% of differentially expressed genes with a mean difference of 1 log2(TPM + 1).

We performed gene set enrichment analysis (GSEA) [24,25] using the estimated differences between inferred from the hierarchical model. This created a database of differentially expressed pathways, which we visualized using the EnrichmentMap software [26].

Clusters of related gene sets were manually annotated to highlight biological features that are differentially expressed between PDX and human tumors.

We developed a novel GSEA approach that uses a mixture model to infer reference distributions for each osteosarcoma gene. We then normalize expression to this reference distribution to amplify the detection of subtype expression. We applied this method to the osteosarcoma PDX and patient samples. The recent tumor microenvironment subtypes identified by the hydra method where used to compare osteosarcoma patient tumor samples and their matched PDXs.

Results

TumorMap analysis shows that PDXs cluster near to patient tumors with same diagnosis

To investigate differences in gene expression between PDXs and their corresponding tumor samples, we first use the genomic dimensionality reduction tool known as TumorMap to reduce the feature space and identify relationships across samples. We applied the TumorMap algorithm to the PPTC and TARGET gene expression data to assess whether PDXs cluster with patients with the same diagnosis (Fig 1). We found that in general PDXs cluster near to related disease cohorts, but none of the well-represented PDXs (N > 10) actually merged into the patient cluster. We found that the TumorMap algorithm will cluster small clusters, but once the PDX cohort becomes sufficiently large, this cluster will separate from the patient cluster (SFig 1).

T-cell ALL was the only PDX that did not cluster near to the related patient tumors. As a comparison, we also did the same analysis on unmatched cancer cell lines and patient samples found that the cell lines clustered very far away from patient samples. We then compared this clustering pattern to a similar analysis using the Cancer Cell Line Encyclopedia (CCLE) and The Cancer Genome Atlas (TCGA) data, which was available on the UCSC Xena browser. We found that the cell lines clustered separately from the patient tumors, suggesting that cell lines capture less of the original tissue of origin signal than PDXs (SFig 2). The significant difference in expression may be one of many factors leading to the low attrition rate in drug development despite ample evidence in cell line models. PDXs did not cluster with patient samples, but the TumorMap algorithm found enough similarity to link them with the appropriate diseases. This may suggest that PDXs are a more accurate model than cell lines and perhaps can be improved to make PDXs even more accurate model of human cancer.

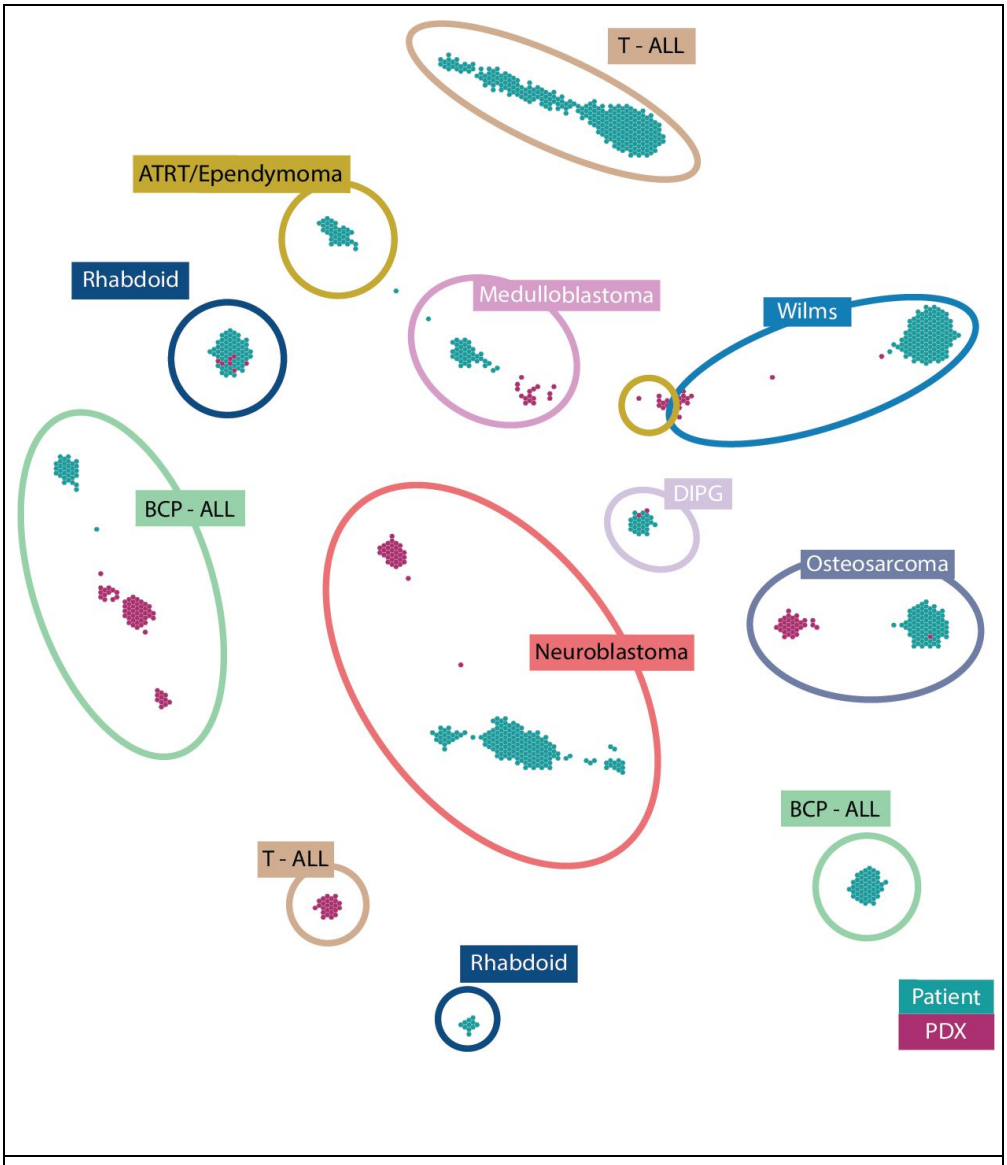


Figure 1: PDX tumors cluster near to human tumors of the same diagnosis. TumorMap visualization of PPTC PDX and related TARGET gene expression profiles. Circles group PDX and human gene expression clusters with related diagnoses.

Multilevel Differential Expression Analysis

We hypothesized that credentialing of PDX models could be improved if systematic changes that occur in all PDXs were identified. The high dimensionality of genomic data and small sample sizes pose challenges for genome-wide credentialing of PDX models. Bayesian statistics incorporate prior knowledge to improve inferences and can reduce the problems introduced by small sample size [27]. We developed a Bayesian hierarchical model that propagates information across gene-level parameters to improve inferences for all genes (Fig 2). We performed a power analysis to determine the number of patient/PDX pairs needed to robustly detect differentially expressed genes. The hierarchical model becomes well-powered (>80%) to detect a mean difference of 1 log2(TPM + 1) with 8 patient/PDX pairs.

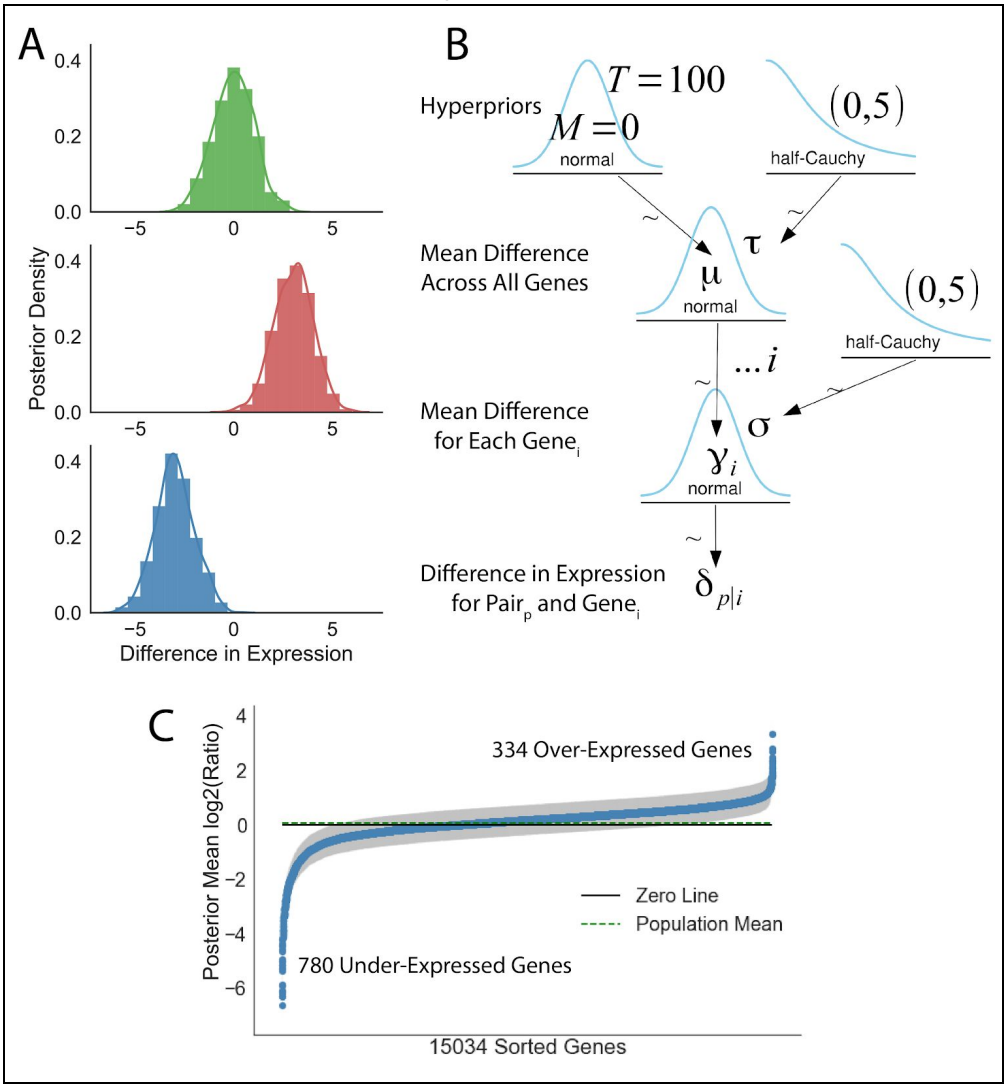


Figure 2: Hierarchical Bayesian model identifies significant expression similarity between patient tumor gene expression and matched patients.

We applied the hierarchical model to a cohort of matched pediatric PDX/primary tumor samples (8 osteosarcoma, 1 rhabdomyosarcoma, 1 Ewing sarcoma and 1 synovial sarcoma). Our model identified 334 genes with consistently higher and 780 genes with consistently lower expression in PDXs compared to matched primary tumor samples. The majority of genes were not systematically differentially expressed in PDX (~93%), suggesting that most expression effects are preserved, which is consistent with our PDX TumorMap analysis. We then investigated whether coordinated expression of biological pathways was observed in the hierarchical PDX analysis. We ranked genes by their estimated expression differences and performed gene set enrichment analysis [24,25] using the EnrichmentMap gene set database [26].

We found statistically significant upregulation of X gene sets and downregulation of Y gene sets (adjusted p-value < 0.05, SFile 1). As expected, we identified downregulation of immune and stromal pathways, but also identified upregulation of cancer-associated functions, including spliceosome, cell cycle, and transcriptional regulation pathways (Fig 3). We generated an EnrichmentMap to visualize higher-level relationships across enriched gene sets [26]. We found a large network of related immune gene sets influencing innate and adaptive immune expression gene sets. We also found downregulation of MHC presentation and extracellular organization pathways, which may be associated with the lack of selective pressure from the host immune system and thus there is no longer a survival benefit to expressing MHC genes.

While most of the gene sets were associated with downregulation, a small number of gene sets were upregulated. These gene sets were associated with expected biological functions, including upregulation of cell cycle expression, which is likely associated with the enrichment for cancer cells in PDX tumors. We were surprised to see that TNF-alpha signalling gene sets were expressed higher in PDXs, but it is unclear how this pathway may be functioning differently in the PDX. We speculate that the TNF-alpha expression may be associated with a wound healing response in the PDX that may be reflected by the cancer cells [28].

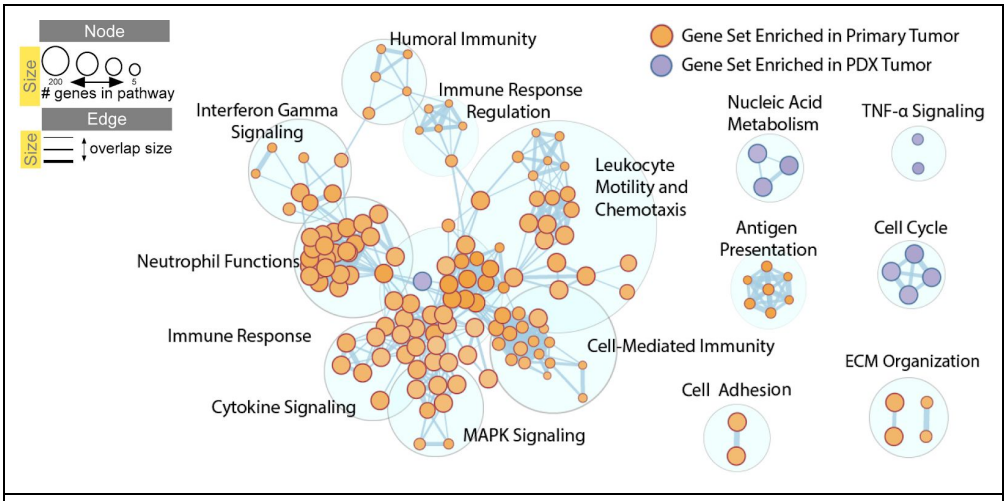


Figure 4: Differentially expressed genes in PDX are associated with immune and cell cycle regulation.

Removing sarcoma expression differences causes clustering of osteosarcoma PDX and patients samples

We clustered data from the TARGET project and the Pediatric Preclinical Testing Consortium (PPTC) and used TumorMap to visualize relationships across this large multivariate datasets. Initially, the osteosarcoma PDXs clustered separately from the TARGET osteosarcoma patient samples (Fig 1), but removing PDX-specific expression identified through the Bayesian hierarchical approach described above improved the rank correlation and led to the osteosarcoma PDX clusters merging with the TARGET osteosarcoma cluster (Fig. 3). Thus, osteosarcoma expression features were preserved in the PDX after accounting for global PDX-specific expression differences. This correction had the greatest impact on osteosarcoma, suggesting disease-specific differences between PDXs and primary tumors. We have found that having matched PDX and patient samples is essential for isolating the PDX-specific differences, since heterogeneity across cancer types may not be properly balanced in the cohort and differences between subtypes may confound differences between PDX and patient tumors.

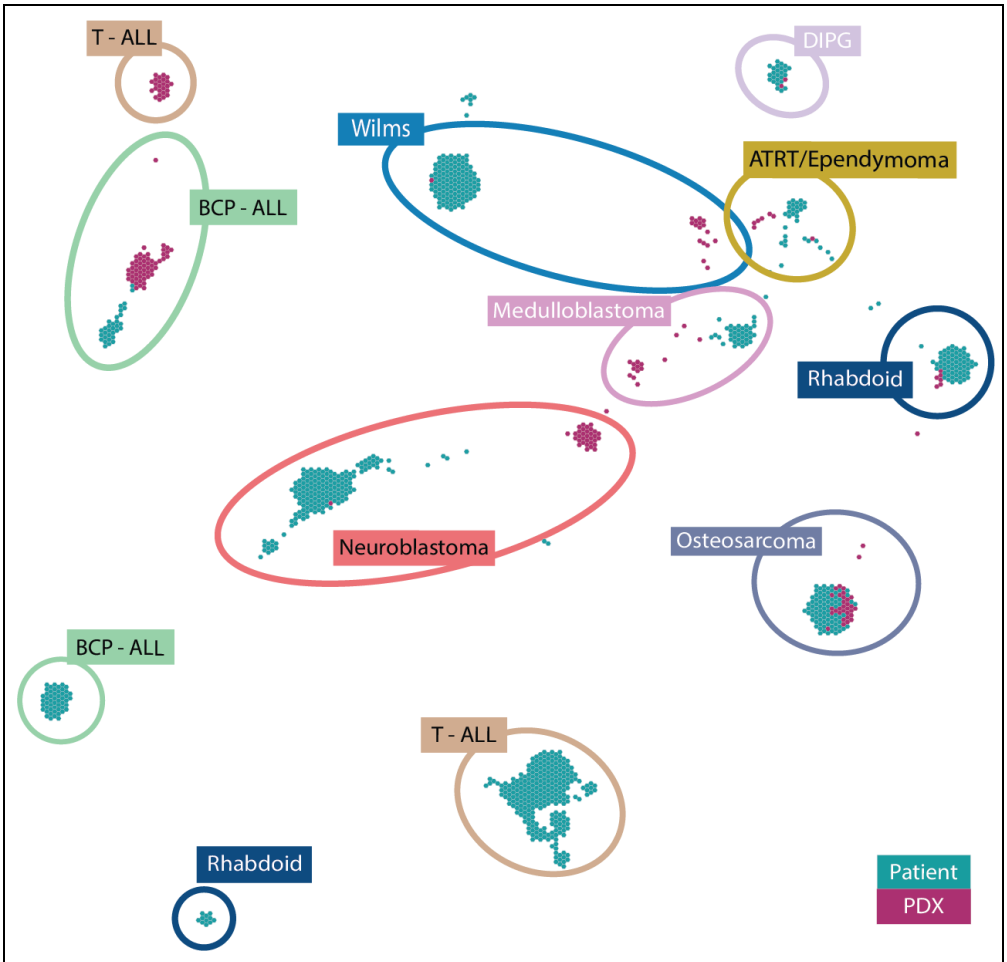


Figure 3: Excluding osteosarcoma PDX expression differences rescues clustering with osteosarcoma patient tumors but does not rescue clustering for other pediatric cancer types.

Mixture model approach for matching gene expression subtypes in patient and PDX tumors
After identifying global differences between PDXs and matched osteosarcoma samples, we wanted to develop a framework for subtyping patient samples and PDXs that may facilitate the development of new therapies. We assume that PDX models that share gene set enrichment signatures may be better able to reflect the druggable features of patient tumors. We first developed a mixture model approach for identifying differential expression within the TARGET osteosarcoma cohort. This analysis transformed all genes from a multimodal distribution to a univariate that can be used to amplify the subtype expression that may be used for subtyping

tumors (SFig 2). We then applied this approach to matched PDX and patient tumor samples. We found that PDX tumors that were derived from patient tumors with low immune and stromal expression had better conserved gene expression enrichment than PDX tumors derived from immune-active tumors (Fig 6).

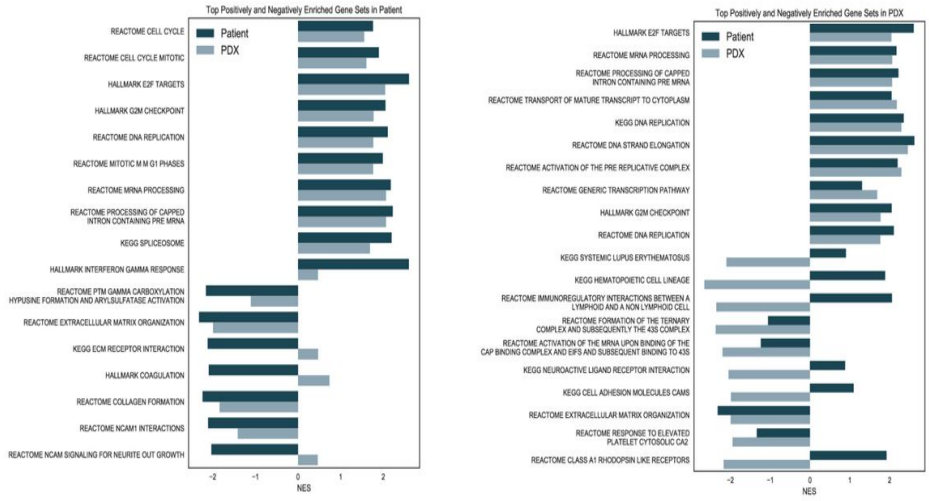


Figure 6: Top and bottom 10 enriched gene sets for a matched patient and PDX pair clustering with the proliferative osteosarcoma subtype. NES: normalized enrichment score

Differentially Expressed Genes Correlate with Response to Targeted Therapies in PDX

We then applied UCSC Treehouse outlier analysis to a cohort of PDX samples. We found druggable genes in each of the samples and the gene expression leads correlated with the CNV-based predictions of Sayles et al. (2019) in 7 out of the 9 PDXs with drug response data [29]. The PDXs that did not correlate were not tested with the gene expression lead, so it is unknown whether the PDX would have responded.

An additional PDX was identified for which CNV analysis did not identify an actionable lead, but gene expression analysis found over-expression of the FGFR1 gene. The RTK inhibitor pazopanib led to a reduction in tumor growth compared to a vehicle control (Fig 7A). We then looked in our database for evidence that the FGFR1 pathway is differentially expressed in the PDX and found that the FGFR1 expression pathway is preserved in the sarcoma PDXs. We similarly analyzed a Ewing sarcoma and found overexpression of JAK1, but the JAK1 inhibitor momelotinib did not cause a difference in tumor growth (Fig 7B). We then investigated the differentially expressed pathways in pediatric sarcoma PDXs and found the JAK signaling

pathways were significantly downregulated in PDX models, which may confound preclinical validation studies that attempt to target this pathway.

PDX	Expression Target	CNV Target	Drug Target	TGI
PSS085	MYC	MYC	CDK9	104.6
PSS089	MYC	MYC	CDK9	83.9
PSS112	CDK7	CDK2	CDK2	57%
PSS004	CDK4	CDK4	CDK4	82.7
PSS018	CDK4	CDK4	CDK4	66
PSS077	FGFR1/VEGF	VEGF	VEGF	76.5
PSS050	AKT1	AKT1	AKT1	66
PSS008	DNMT1	PTEN	AKT1	60.8
PSS078	FGFR1	FOXM1	CDK4	112.4

Table 1. Gene expression and copy number analysis correlate with drug response for 7 out of 9 PDXs tested with a CNV identified lead. The tumor growth index (TGI) is a measure of the tumor growth relative to a control experiment. A larger TGI value signifies a better response to the targeted inhibitor. A value greater than 60 is considered significant.

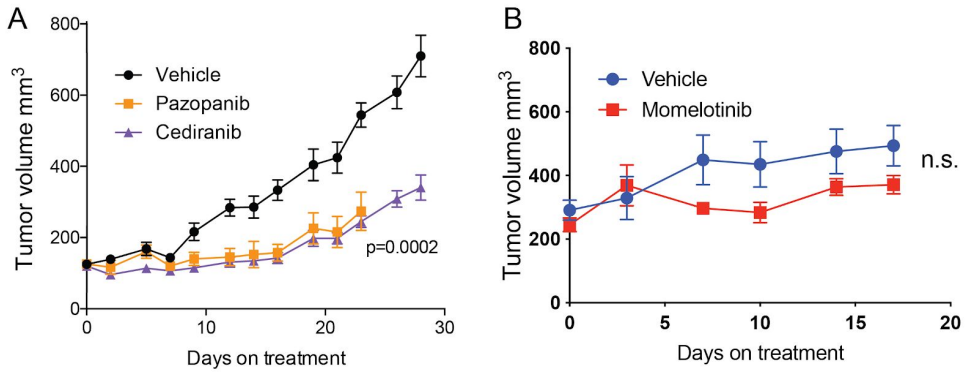


Fig 7. Gene expression outlier analysis predicts response to RTK but not JAK inhibitors.

Pathway	NES	Adjusted P-value
HALLMARK IL6 JAK STAT3 SIGNALING	-3.22	0.003
REGULATION OF JAK-STAT CASCADE	-2.62	0.003
POSITIVE REGULATION OF JAK-STAT CASCADE	-2.57	0.003

Table 2. JAK1 signaling is downregulated in PDX.

Discussion

Patient arrived xenografts (PDXs) are important preclinical models for the drug development industry. It is presumed that PDXs are more accurate models of human cancer, but there has been limited research into the molecular evolution that occurs when human cancer tissue is implanted into an immunosuppressed laboratory mouse. To address this challenge, we developed a computational framework for inferring consistent differences in gene expression and identifying coordinated expression of differentially expressed genes that participated in known biological pathways. Significant differences in gene expression may confound validation experiments that rely on consistent expression of particular biological pathways. We have provided a database of differentially expressed genes and pathways for researchers to reference when designing drug validation assays. By creating a knowledge base for specific diseases, we will facilitate the development of therapies that better reflect the human disease population. Significant differences in gene expression may lead to differences in the response to particular therapies.

The small sample sizes and statistical noise associated with cancer gene expression data led us to develop a novel Bayesian hierarchical model to infer differential gene expression [30]. The patient-to-patient heterogeneity within cancer types makes inferring differentially expressed genes between unmatched samples of patients and PDXs complicates statistical analysis. We studied differential expression across matched patient and PDX tumor gene expression to emphasize differences in gene expression associated with the PDX system. This model was well powered to detect significant differences in gene expression while also introducing statistical shrinkage to decrease the detection of false positive differences [27].

Our analysis of PDX gene expression differences found that most genes are not differentially expressed in the PDX model. Initial TumorMap analysis showed that PDXs share enough similarity in gene expression that the TumorMap algorithm was able to group related diseases close to each other but there were sufficient differences such that none of the PDX clusters merged with the patient tumor clusters. For comparison, we performed the same analysis for CCLE and TCGA gene expression and found that cell lines cluster separately from patient tumor samples, which suggests that they share fewer similarities with related patient tumors.

The inferred differences in PDX expression were associated with expected differences in the PDX models, including loss of immune and stromal expression. PDXs showed higher expression of TNF-alpha signaling, which has not been previously reported, but may be associated with the engraftment of the human tumor tissue in the mouse [31,32]. We also found enrichment of cell cycle pathway genes, which may have resulted from enrichment for cancer cells within the PDX tumor. We found that a significant number of differentially expressed genes were also among the known druggable genes. Knowledge of these differences may help in prioritizing drugs that may better reflect responses in human tumors.

Removing differentially expressed genes in osteosarcoma PDXs from the PPTC and TARGET gene expression compendia rescued clustering of osteosarcoma PDXs and human tumors, but did not improve clustering of other cancer types. This suggests there are disease specific differences in PDXs that need to be accounted for in future PDX credentialing experiments. Furthermore, we recommend that future PDX model generation attempt to generate matched RNA-seq data from the original tumor sample. This is challenging since tumor tissue is limited and a larger tissue section has a better chance of creating a viable PDX, but this work is necessary to create useful PDX data.

We provide a database of differentially expressed genes pathways which includes many of the druggable genes investigated for precision medicine applications. Our hope is that this set of pathways and genes can facilitate validation of therapies in the PDX. There may be genes and pathways that behave differently in the PDX and thus the results of the validation study is inconclusive. As an example, we showed that overexpression of FGFR1 and FGFR1 signaling pathways predicted response in a PDX and that these results may be more accurate since the FGFR1 pathway is not differentially expressed in the PDX.

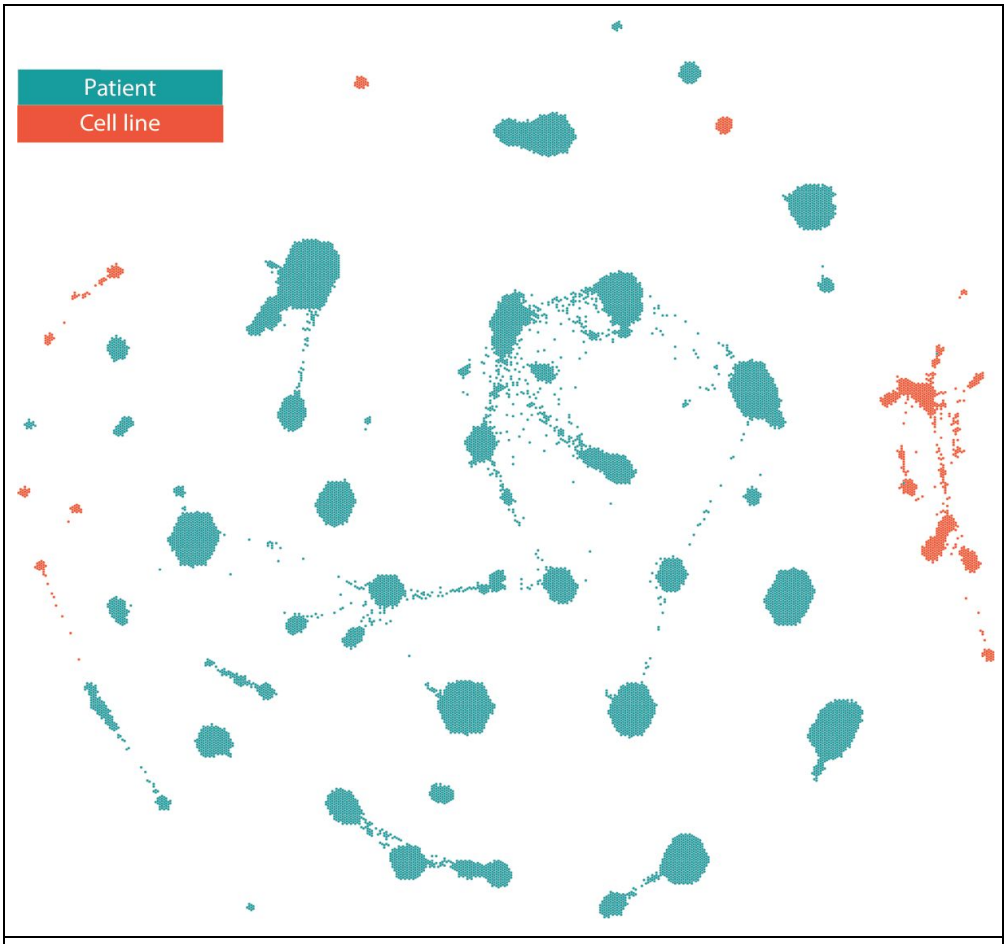
The JAK1 signaling pathway was differentially expressed in the PDX, so an experiment that targets the JAK1 pathway in PDXs may have inaccurate conclusions since many genes involved in this pathway show differences in the PDX. We tested this hypothesis by targeting overexpression of JAK1 in the PDX despite the JAK1 signaling pathway being strongly downregulated in the PDX. A JAK1 inhibitor assay resulted in no significant difference in tumor growth, but it is not clear if this is due to JAK1 not being a good target or if the JAK1 pathway behaves differently in the PDX.

It is becoming increasingly clear that the tumor microenvironment contributes significantly to cancer biology and gene expression measurements [12,33–35]. We used the recently published *hydra* gene expression subtypes of osteosarcoma to better understand PDX evolution. We found that despite the original patient tumors showing significant heterogeneity in tumor microenvironment expression patterns, the PDX tumors all showed expression signals associated with the tumor microenvironment associated with low immune infiltrate and stromal expression. There has been significant development into humanized PDX models with active human immune components, but these models need further investigation using a method similar to the framework presented here. Another approach is to use the many PDX models that have already been developed and to match the results of these experiments to patients with a similar tumor microenvironment state (SFig 3).

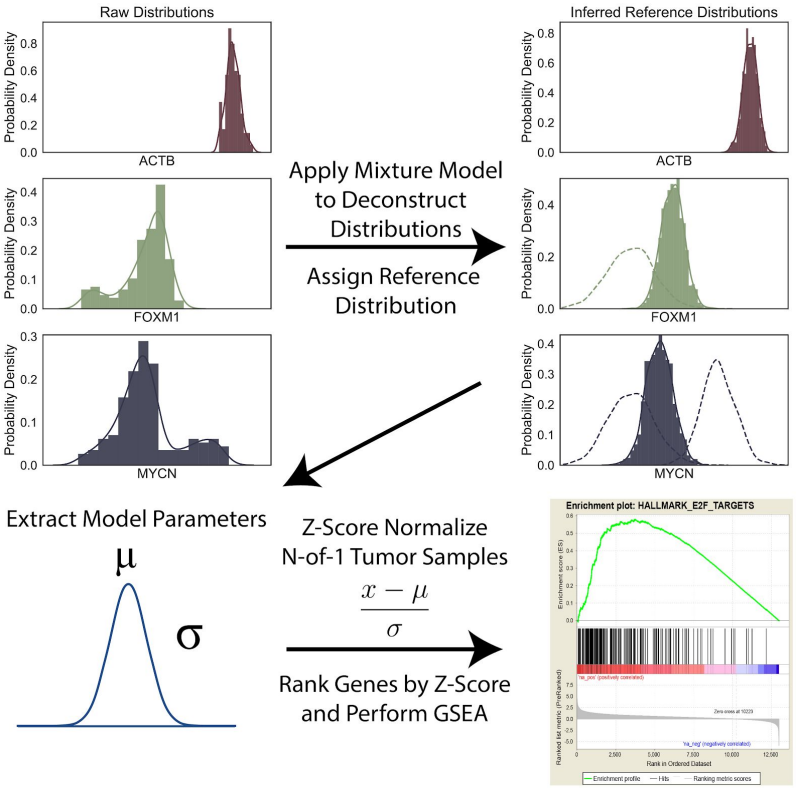
Conclusion

There is increased focus on developing more accurate preclinical models, but there has been limited research into using the models we already have more effectively. Here, we describe a framework of credentialing available PDX models to identify the most accurate PDX and patient tumor pairs for developing novel therapies. The influence of the host immune system on tumor evolution is lost in the PDX, so we have proposed prioritizing PDXs derived from immune-silent tumors since these tumors have strongly correlated gene enrichment.

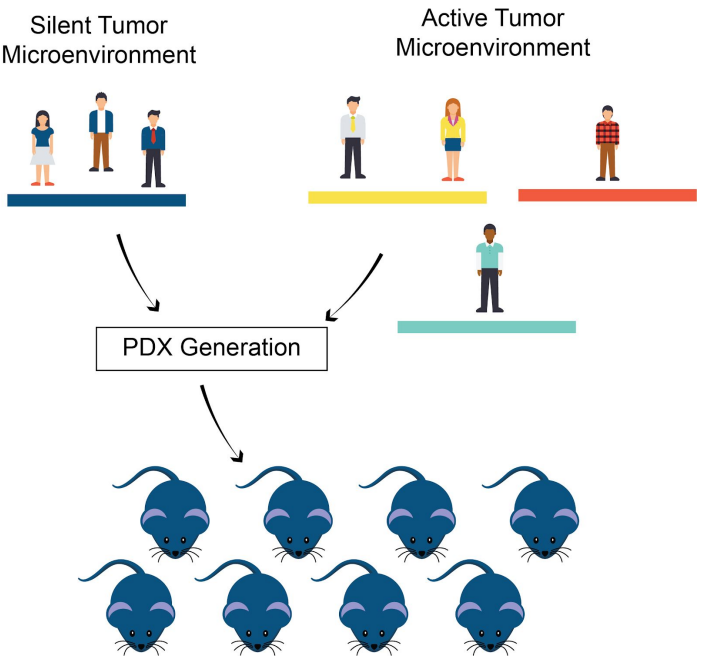
Supplement



SFig 1. Cancer Cell Line Encyclopedia and The Cancer Genome Atlas Gene Expression TumorMap.



SFig 2. Mixture model approach for identifying cancer subtype expression pathways.



SFig 3. Patient tumor microenvironment effects are lost in PDX, but may support more accurate modeling of immune-silent patient tumors.

References

1. Oberlin O, Rey A, Lyden E, Bisogno G, Stevens MCG, Meyer WH, et al. Prognostic Factors in Metastatic Rhabdomyosarcomas: Results of a Pooled Analysis From United States and European Cooperative Groups. J Clin Oncol. 2008;26: 2384–2389. doi:10.1200/JCO.2007.14.7207
2. London WB, Castel V, Monclair T, Ambros PF, Pearson ADJ, Cohn SL, et al. Clinical and Biologic Features Predictive of Survival After Relapse of Neuroblastoma: A Report From the International Neuroblastoma Risk Group Project. J Clin Oncol. 2011;29: 3286–3292. doi:10.1200/JCO.2010.34.3392
3. Suggitt M, Bibby MC. 50 years of preclinical anticancer drug screening: empirical to target-driven approaches. Clin Cancer Res Off J Am Assoc Cancer Res. 2005;11: 971–981.
4. Hutchinson L, Kirk R. High drug attrition rates—where are we going wrong? Nat Rev Clin Oncol. 2011;8: 189–190. doi:10.1038/nrclinonc.2011.34
5. Wilding JL, Bodmer WF. Cancer Cell Lines for Drug Discovery and Development. Cancer Res. 2014;74: 2377–2384. doi:10.1158/0008-5472.CAN-13-2971
6. Mullard A. Can you trust your cancer cell lines? Nat Rev Drug Discov. 2018;17: 613–613.

doi:10.1038/nrd.2018.154
7. Wilking N, Karolinska Institutet, Solna, Sweden, Lopes G, Sylvester Comprehensive Cancer Center, University of Miami, FL, US, Meier K, HKK Soltau, Lower Saxony & Heidekreis-Klinikum GmbH, Soltau, Germany, et al. Can we Continue to Afford Access to Cancer Treatment? Eur Oncol Haematol. 2017;13: 114. doi:10.17925/EOH.2017.13.02.114
8. Kodack DP, Farago AF, Dastur A, Held MA, Dardaei L, Friboulet L, et al. Primary Patient-Derived Cancer Cells and Their Potential for Personalized Cancer Patient Care. Cell Rep. 2017;21: 3298–3309. doi:10.1016/j.celrep.2017.11.051
9. Ertel A, Verghese A, Byers SW, Ochs M, Tozeren A. Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. Mol Cancer. 2006;5: 55. doi:10.1186/1476-4598-5-55
10. Gillet J-P, Varma S, Gottesman MM. The Clinical Relevance of Cancer Cell Lines. JNCI J Natl Cancer Inst. 2013;105: 452–458. doi:10.1093/jnci/djt007
11. Ben-David U, Siranosian B, Ha G, Tang H, Oren Y, Hinohara K, et al. Genetic and transcriptional evolution alters cancer cell line drug response. Nature. 2018;560: 325–330. doi:10.1038/s41586-018-0409-3
12. Nishida-Aoki N, Gujral TS. Emerging approaches to study cell-cell interactions in tumor microenvironment. Oncotarget. 2019;10: 785–797. doi:10.18632/oncotarget.26585
13. Williams JA. Using PDX for Preclinical Cancer Drug Discovery: The Evolving Field. J Clin Med. 2018;7. doi:10.3390/jcm7030041
14. Patient-derived xenografts undergo mouse-specific tumor evolution | Nature Genetics. [cited 2 Feb 2020]. Available: https://www.nature.com/articles/ng.3967
15. Choi SYC, Lin D, Gout PW, Collins CC, Xu Y, Wang Y. Lessons from patient-derived xenografts for better in vitro modeling of human cancer. Adv Drug Deliv Rev. 2014;79–80: 222–237. doi:10.1016/j.addr.2014.09.009
16. Siolas D, Hannon GJ. Patient-derived tumor xenografts: transforming clinical samples into mouse models. Cancer Res. 2013;73: 5315–9. doi:10.1158/0008-5472.CAN-13-1069
17. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6: pl1. doi:10.1126/scisignal.2004088
18. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discov. 2012;2: 401–404. doi:10.1158/2159-8290.CD-12-0095
19. Vaske OM, Bjork I, Salama SR, Beale H, Shah AT, Sanders L, et al. Comparative Tumor RNA Sequencing Analysis for Difficult-to-Treat Pediatric and Young Adult Patients With Cancer. JAMA Netw Open. 2019;2: e1913968–e1913968. doi:10.1001/jamanetworkopen.2019.13968
20. Goldman M, Craft B, Kamath A, Brooks A, Zhu J, Haussler D. The UCSC Xena Platform for cancer genomics data visualization and interpretation. bioRxiv. 2018; 326470. doi:10.1101/326470
21. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol. 2017;35: 314–316. doi:10.1038/nbt.3772
22. Newton Y, Novak AM, Swatloski T, McColl DC, Chopra S, Graim K, et al. TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal. Cancer Res. 2017;77: e111–e114. doi:10.1158/0008-5472.CAN-17-0580
23. McElreath R. Statistical rethinking: A Bayesian course with examples in R and Stan.

Chapman and Hall/CRC; 2018.
24. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102: 15545–15550. doi:10.1073/pnas.0506580102
25. Korotkevich G, Sukhov V, Sergushichev A. Fast gene set enrichment analysis. bioRxiv. 2019; 060012. doi:10.1101/060012
26. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PloS One. 2010;5: e13984.
27. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB, et al. Bayesian Data Analysis. Chapman and Hall/CRC; 2013. doi:10.1201/b16018
28. Ritsu M, Kawakami K, Kanno E, Tanno H, Ishii K, Imai Y, et al. Critical role of tumor necrosis factor-α in the early process of wound healing in skin. J Dermatol Dermatol Surg. 2017;21: 14–19. doi:10.1016/j.jdds.2016.09.001
29. Sayles LC, Breese MR, Koehne AL, Leung SG, Lee AG, Liu H-Y, et al. Genome-Informed Targeted Therapy for Osteosarcoma. Cancer Discov. 2019;9: 46–63. doi:10.1158/2159-8290.CD-17-1152
30. Ji H, Liu XS. Analyzing 'omics data using hierarchical models. Nat Biotechnol. 2010;28: 337–340. doi:10.1038/nbt.1619
31. Ziaei R, Ayatollahi M, Yaghobi R, Sahraeian Z, Zarghami N. Involvement of TNF-α in differential gene expression pattern of CXCR4 on human marrow-derived mesenchymal stem cells. Mol Biol Rep. 2014;41: 1059–1066. doi:10.1007/s11033-013-2951-2
32. Gálvez BG, Martín NS, Rodríguez C. TNF-alpha Is Required for the Attraction of Mesenchymal Precursors to White Adipose Tissue in Ob/ob Mice. PLOS ONE. 2009;4: e4444. doi:10.1371/journal.pone.0004444
33. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018;173: 291-304.e6. doi:10.1016/j.cell.2018.03.022
34. Mao Y, Feng Q, Zheng P, Yang L, Liu T, Xu Y, et al. Low tumor purity is associated with poor prognosis, heavy mutation burden, and intense immune phenotype in colon cancer. Cancer Manag Res. 2018;10: 3569–3577. doi:10.2147/CMAR.S171855
35. Hanahan D, Coussens LM. Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. Cancer Cell. 2012;21: 309–322. doi:10.1016/j.ccr.2012.02.022

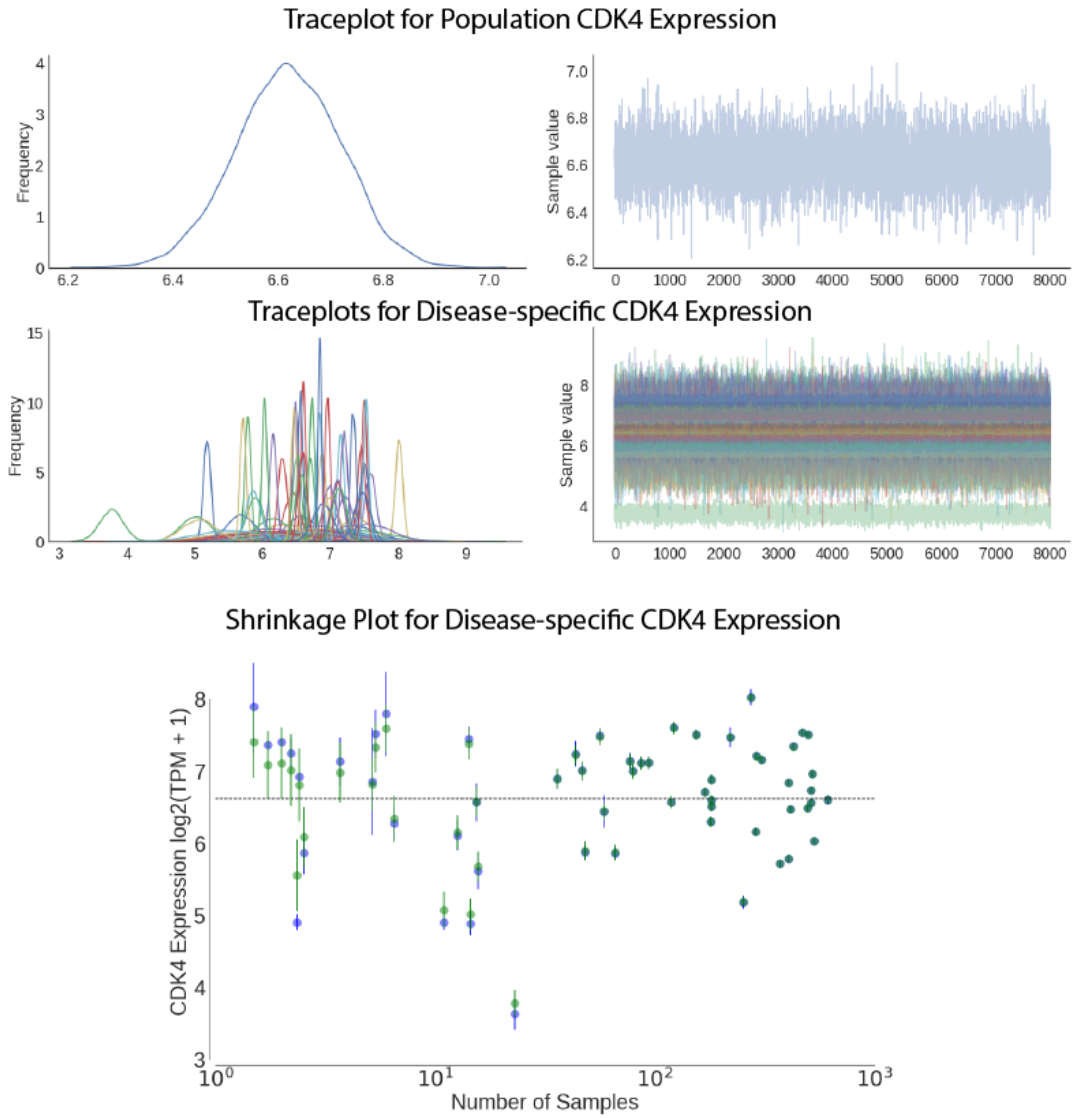


Figure 6.2: Trace and scatter plots for preliminary partial pooling model. The top trace plot shows the posterior distribution for global CDK4 expression and the lower trace plot shows the disease specific posterior distributions for mean CDK4 expression. The bottom scatter plot shows the no-pooling CDK4 model in blue and the partial pooling model in green. Note that the partial pooling model shrinks towards the population mean value.

Chapter 7

Genomic Profiling of Childhood Tumor Patient-Derived Xenograft Models to Enable Rational Clinical Trial Design

Introduction

For the Pediatric Preclinical Testing Consortium (PPTC), I developed a hierarchical model that groups cancers by the tissue of origin (Model **??**). This model will facilitate learning pediatric cancer effects in situations where children and adults do not develop the same kind of cancer. For example, bone cancer is much more common in children than adults, but linking these cancers through a shared bone-specific prior distribution will better model the biological effects of pediatric bone cancer.

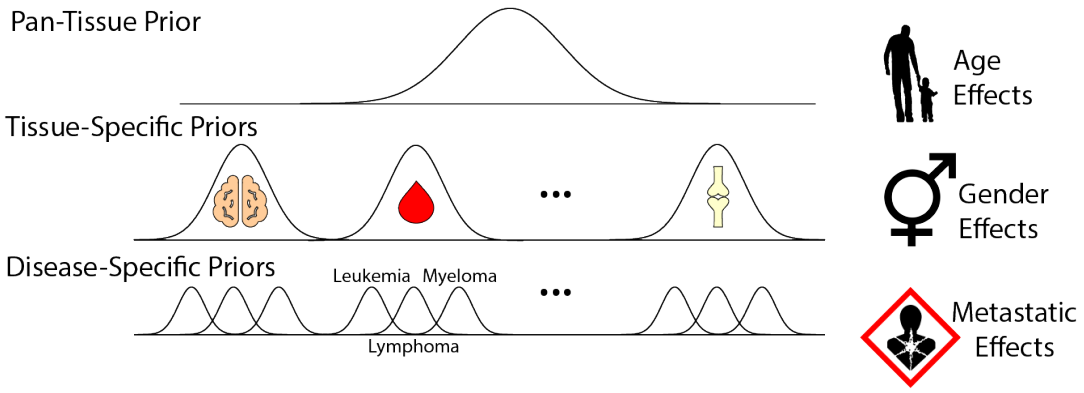


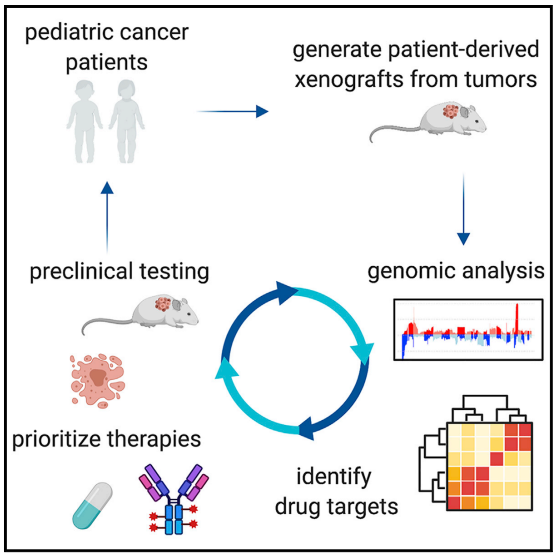
Figure 7.1: Hierarchical model for Treehouse compendium. Each tissue is modeled separately using a pan-tissue prior distribution. Cancer types are then associated with the tissue of origin. This hierarchical model takes advantage of similar expression patterns between cancers of the same tissue type. Grouping related data decreases the amount of variation and uncertainty in the model. Predictions from the hierarchical model can be used to identify abnormal expression for new patients. The model also learns varying effects on expression related to age, gender, and metastatic tissue samples that could influence gene expression.

Cell Reports

Resource

Genomic Profiling of Childhood Tumor Patient-Derived Xenograft Models to Enable Rational Clinical Trial Design

Graphical Abstract



Authors

Jo Lynne Rokita, Komal S. Rathi, Maria F. Cardenas, ..., Pichai Raman, David A. Wheeler, John M. Maris

Correspondence

maris@email.chop.edu

In Brief

Rokita et. al provide an extensively annotated genomic dataset of somatic oncogenic regulation across 37 distinct pediatric malignancies. The 261 patient-derived xenograft models are available to the scientific community, and the genomic annotations will enable rational preclinical agent prioritization and acceleration of therapeutic targets for early-phase pediatric oncology clinical trials.

Highlights

- Multiplatform analysis facilitates genomic resource of 261 pediatric cancer PDX models
- PPTC PDX models are reflective of high-risk and chemotherapy resistant disease
- Inferred TP53 pathway inactivation correlates with pediatric cancer copy number burden
- Pediatric cancer PDX models will be useful for drug development prioritization

Rokita et al., 2019, Cell Reports *29*, 1675–1689
November 5, 2019
https://doi.org/10.1016/j.celrep.2019.09.071

CellPress

Cell Reports
Resource

Genomic Profiling of Childhood Tumor Patient-Derived Xenograft Models to Enable Rational Clinical Trial Design

Jo Lynne Rokita,[1,2,3] Komal S. Rathi,[2,3] Maria F. Cardenas,[4] Kristen A. Upton,[1] Joy Jayaseelan,[4] Katherine L. Cross,[5] Jacob Pfeil,[6] Laura E. Egolf,[1,7] Gregory P. Way,[8] Alvin Farrel,[2] Nathan M. Kendsersky,[1,9] Khushbu Patel,[2] Krutika S. Gaonkar,[2,3] Apexa Modi,[1,8] Esther R. Berko,[1] Gonzalo Lopez,[1,2] Zalman Vaksman,[2] Chelsea Mayoh,[10] Jonas Nance,[11] Kristyn McCoy,[11] Michelle Haber,[10] Kathryn Evans,[10] Hannah McCalmont,[10] Katerina Bendak,[10] Julia W. Böhm,[10] Glenn M. Marshall,[10,12] Vanessa Tyrrell,[13] Karthik Kalletla,[2,3] Frank K. Braun,[14] Lin Qi,[15,16] Yunchen Du,[15,16] Huiyuan Zhang,[15,16] Holly B. Lindsay,[15,16] Sibo Zhao,[15,16] Jack Shu,[15,16] Patricia Baxter,[15,16] Christopher Morton,[17] Dias Kurmashev,[18] Siyuan Zheng,[18] Yidong Chen,[18] Jay Bowen,[19] Anthony C. Bryan,[19] Kristen M. Leraas,[19] Sara E. Coppens,[19] HarshaVardhan Doddapaneni,[4] Zeineen Momin,[4]

(Author list continued on next page)

[1]Division of Oncology, Children's Hospital of Philadelphia, and Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104-4318, USA
[2]Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[3]Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[4]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA
[5]Guardian Forensic Sciences, Abington, PA 19001, USA
[6]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[7]Cell and Molecular Biology Graduate Group, University of Pennsylvania, Philadelphia, PA 19104, USA
[8]Genomics and Computational Biology Graduate Group, University of Pennsylvania, Philadelphia, PA 19104, USA
[9]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA 19104, USA
[10]Children's Cancer Institute, School of Women's and Children's Health, UNSW Sydney, Sydney, NSW, Australia
[11]Cancer Center, Texas Tech University Health Sciences Center School of Medicine, Lubbock, TX 79430, USA
[12]Sydney Children's Hospital, Sydney, NSW, Australia
[13]Children's Cancer Institute, Kensington, NSW, Australia
[14]Texas Children's Cancer and Hematology Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA
[15]Preclinical Neurooncology Research Program, Texas Children's Cancer Research Center, Texas Children's Hospital, Houston, TX 77030, USA
[16]Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA
[17]Department of Surgery, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
[18]Greehey Children's Cancer Research Institute, University of Texas Health Science Center, San Antonio, TX 78229, USA
[19]The Research Institute at Nationwide Children's Hospital, Columbus, OH 43205, USA
[20]Division of Pediatrics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[21]Department of Global Health Technologies, RTI International, Research Triangle Park, NC 27709, USA
[22]Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Hanover, NH 03755, USA
[23]Norris Cotton Cancer Center, Lebanon, NH 03766, USA
[24]Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, PA 19102, USA

(Affiliations continued on next page)

SUMMARY

Accelerating cures for children with cancer remains an immediate challenge as a result of extensive oncogenic heterogeneity between and within histologies, distinct molecular mechanisms evolving between diagnosis and relapsed disease, and limited therapeutic options. To systematically prioritize and rationally test novel agents in preclinical murine models, researchers within the Pediatric Preclinical Testing Consortium are continuously developing patient-derived xenografts (PDXs)—many of which are refractory to current standard-of-care treatments—from high-risk childhood cancers. Here, we genomically characterize 261 PDX models from 37 unique pediatric cancers; demonstrate faithful recapitulation of histologies and subtypes; and refine our understanding of relapsed disease. In addition, we use expression signatures to classify tumors for *TP53* and *NF1* pathway inactivation. We anticipate that these data will serve as a resource for pediatric oncology drug development and will guide rational clinical trial design for children with cancer.

INTRODUCTION

An estimated 15,780 children and adolescents (<20 years) are diagnosed with cancer in the United States each year, and these

OPEN ACCESS
CellPress

Wendong Zhang,[20] Gregory I. Sacks,[1] Lori S. Hart,[1] Kateryna Krytska,[1] Yael P. Mosse,[1] Gregory J. Gatto,[21] Yolanda Sanchez,[22,23] Casey S. Greene,[24,9] Sharon J. Diskin,[1,2] Olena Morozova Vaske,[25,6] David Haussler,[6,26] Julie M. Gastier-Foster,[19,27] E. Anders Kolb,[28,29] Richard Gorlick,[20] Xiao-Nan Li,[15,16,30,31] C. Patrick Reynolds,[11] Raushan T. Kurmasheva,[18] Peter J. Houghton,[18] Malcolm A. Smith,[32] Richard B. Lock,[13] Pichai Raman,[2,3] David A. Wheeler,[4] and John M. Maris[1,33,*]

[25]Department of Molecular, Cell and Developmental Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[26]Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[27]The Ohio State University College of Medicine, Departments of Pathology and Pediatrics, Columbus, OH 43210, USA
[28]Department of Pediatrics, Sidney Kimmel Medical College at Thomas Jefferson University, Philadelphia, PA 19107, USA
[29]Nemours Center for Cancer and Blood Disorders, Nemours/Alfred I. duPont Hospital for Children, Wilmington, DE 19803, USA
[30]Division of Hematology, Oncology, Neuro-oncology and Stem Cell Transplant, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL 60611, USA
[31]Department of Pediatrics, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA
[32]National Cancer Institute, NIH, Bethesda, MD 20814, USA
[33]Lead Contact
*Correspondence: maris@email.chop.edu
https://doi.org/10.1016/j.celrep.2019.09.071

diverse entities are the leading cause of disease-related deaths in children (American Childhood Cancer Organization, 2014). Despite five-year survival rates for pediatric cancers now exceeding 80%, survivors frequently have lifelong side effects from cytotoxic therapy, and survival outcomes for children with certain types of tumors remain dismal. The relative rarity of pediatric cancers, molecular and mechanistic heterogeneity of subtypes within and across histologies, genetic and molecular distinction from adult malignancies, tumor evolution in the face of cytotoxic standard therapies, and lack of targeted therapeutic agents all pose major challenges to improving outcomes for children with cancer. Indeed, there are very few drugs with specific labeled indications for pediatric malignancies, and most standard therapies are largely empiric.

Preclinical testing of new therapeutic anti-cancer agents is essential in the field of pediatric oncology due to the relative rarity of the condition and the need to prioritize agents for early-phase clinical trials. Over the past 15 years, the Pediatric Preclinical Testing Consortium (PPTC), previously known as the Pediatric Preclinical Testing Program (Houghton et al., 2002, 2007), has developed over 370 patient-derived xenograft (PDX) models from high-risk childhood cancers. In collaboration with pharmaceutical and academic partners, the PPTC systematically screens novel therapeutic agents for anti-tumor efficacy in order to help prioritize those that will move to the clinic. Previous studies have characterized subsets of pediatric xenograft models, often with limited numbers of specific histologies and/or genomic assays (Brabetz et al., 2018; El-Hoss et al., 2016; Stewart et al., 2017; Townsend et al., 2016; Whiteford et al., 2007). Here, we present a comprehensive genomic characterization of 261 models from 29 unique pediatric cancer malignancies.

RESULTS

Genomic Analysis Workflow and Histological Summary of Pediatric PDX Tumors

Figure 1 depicts the overall workflow of our study, including model histologies, site of tumor specimen, phase of therapy, and molecular assays performed. The PDX generation methods are described in the STAR Methods. We performed whole-exome sequencing (WES) on 240 childhood cancer PDX models, whole-transcriptome sequencing (RNA sequencing [RNA-seq]) on 244 models, and SNP microarrays on 252 models (Figures 1 and S1; Table S1), and we performed short tandem repeat (STR) profiling on all 261 models (Table S2). Of the 261 models profiled, 82 had available references that are also included in Table S2.

Figure S1 describes the analysis workflow (see STAR Methods for details). Of the 240 models on which WES was performed, 69 models were previously sequenced through efforts of the PPTP (dbGAP: phs000469.v17.p7), and we harmonized these data. For WES (Figure S1C) and RNA-seq (Figure S1D), we performed competitive mapping to a hybrid human-mouse reference (hg19-mm10) and used human-specific BAM files as input for downstream analyses. We validated this biochemically with qPCR by calculating the ratio of human:mouse DNA in a subset of 35 PDX tumors. We found a significant correlation between the percent of human reads following WES hybrid mapping and the percent of human DNA in the tumor extract (Figure S1B; Pearson correlation $R = 0.943$, $F = 272.5$, $df = 34$, p value $< 2.2e{-}16$). A mutation annotation format (MAF) file of common germline variation was created if a variant was present in more than five normal samples from The Cancer Genome Atlas (TCGA) patients ($n = 809$). The remaining variants, comprising both somatic and rare germline alterations, were collated into the "somatic" MAF file. Artifactual sequencing variants were removed as described in the STAR Methods. Common germline SNP distributions (allele frequency > 0.005 in any one of the three databases: Exome Aggregation Consortium, 1,000 genomes, or the NHBLI Exome Sequencing Project) were plotted for each model and visually inspected for a negatively skewed distribution to assess DNA cross-contamination in WES data. To identify potential misidentification, RNA variant calling was performed, and variant allele frequencies were correlated between WES and RNA. Models whose variants did not correlate were deemed misidentified and removed (STAR Methods). Within this cohort, five pairs of models were derived from tissue at the phase of therapy (Table S1). Thus, as additional quality control (QC), we correlated somatic mutation allele frequencies between each pair and found a high concordance of mutation frequencies (data on Figshare;

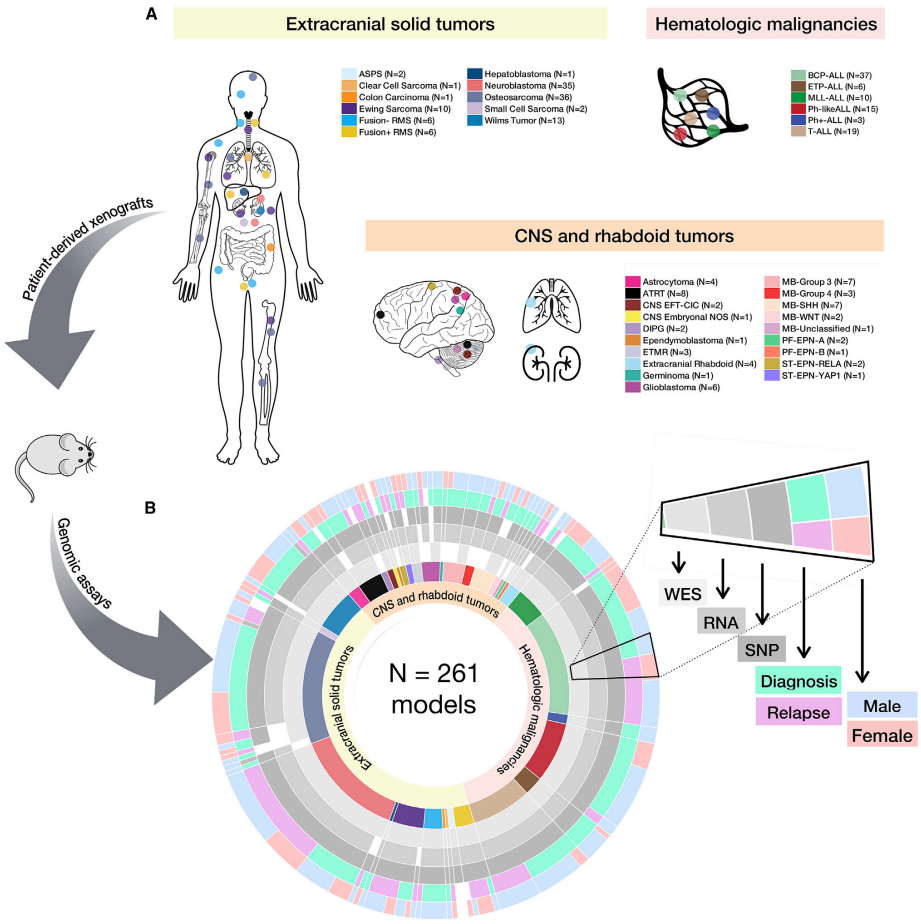


Figure 1. Study and Sample Overview

(A and B) Diversity of the 261 childhood tumors collected (A) and demographics and genomic assays performed by histology (B). Assays performed were whole-exome sequencing (n = 240), whole transcriptome (n = 244), and SNP array copy number analysis (n = 252). Each genomic assay was performed once per biological tumor sample.

See Figure S1 for analysis pipelines, Table S1 for model metadata, and Table S2 for STR profiles.

STAR Methods), confirming the biological reproducibility of creating PDX models within a center. Mutation variation is summarized per model in Table S3.

SNP arrays were processed for segmentation, focal copy number, and ethnicity inference (STAR Methods; Figures S1A and S2). As reported ethnicities were only available for a small proportion of the models, we used SNP array genotypes to infer approximate ethnicities using HapMap genotype frequencies. We assigned models to African, East Asian, European, and South Asian/Hispanic ethnicities (Figure S2; Table S1). Overall, 71% of models are of predicted European descent, 11.5% South Asian/Hispanic, 9.1% African, 5.5% mixed or unknown ethnicity, and 2.4% East Asian.

Following rigorous assessment for contamination, misidentification, and sample mislabeling, 26 full models were excluded, and 3 RNA samples were excluded. The remaining 261 models used herein were shown to be free of detectable levels of DNA contamination (STAR Methods).

PDX Models Recapitulate the Mutation and Copy-Number Landscape of Childhood Cancers

We highlight hallmark alterations in key pediatric tumor driver genes (Behjati et al., 2017; Eleveld et al., 2015; Gröbner et al., 2018; Liu et al., 2017; Ma et al., 2018; Pugh et al., 2013; Shern et al., 2014; Zhang et al., 2012) in Figure 2 and demonstrate faithful disease recapitulation across PDX models.

Acute Lymphoblastic Leukemias

Figure 2A depicts oncoprints for 90 acute lymphoblastic leukemia models.

BCP ALLs

A total of 45%–48% of B cell precursor acute lymphoblastic leukemia (BCP-ALL) PDX models contain canonical focal deletions of the tumor suppressors on chromosome 9p, *CDKN2A* or *CDKN2B* (Figures 2A and S4B), the majority of which are homozygous. The BCP-ALL models were enriched for alterations in the RAS pathway (*KRAS* mutated in 30%, *NRAS* mutated in 18%) and the JAK-STAT pathway (*JAK2/3* altered in 15%), and 15% have altered *KMT2D*. These pathways, along with *PI3K/AKT*, *TNF*α, and *TP53* signaling, were all significantly enriched in gene expression data (Figure 5B). Finally, we detected fusion transcripts in 78% of BCP-ALL models (25/33), many of which contain *ETV6* (27%; 88% of these partner with *RUNX1*), *PAX5* (18%), and *CRLF2* (6%) (Table S5).

ETP and T-ALLs

Early T cell precursor-ALL (ETP-ALL) and T cell-ALL (T-ALL) models are predominantly characterized by *CDKN2A/B* focal deletions (72%–76%; Figure S4B) and/or a *NOTCH1* mutation (68%). Genes within the JAK-STAT pathway are also frequently altered with concurrent pathway enrichment (Figure 5B). *JAK1 or JAK2* lesions were observed in 24% of the models, and 4% of the models contain lesions in *STAT5B*. We detected oncogenic fusion transcripts in nearly half (48%) of these models, many partnering with the following genes: *TRBC2* (16%), *TRBC1* (12%), *ABL1* (8%), *IGH* (8%), *LMAN2* (4%), *LMO1* (4%), *LMO2* (4%), and *ETV6* (4%).

Ph-like and Ph+ ALLs

We confirmed the presence of a *BCR-ABL1* fusion in all three Ph+-ALL models (ALL-04, ALL-55, and ALL-56). Eight Ph-like ALL models (42%; 10/19) contain a canonical *CRLF2* fusion; seven partner with *P2RY8* and one with *IGHM*. Additional frequently rearranged genes include *JAK2* (55%; 12/22) and *PAX5* (23%; 5/22). In both Ph+ and Ph-like ALL models, focal deletions of CDKN2A/B (45%, 10/22; Figure S4B) are predominant. Frequently altered pathways include Ras and JAK-STAT (Figures 2A and 5B).

MLL-ALLs

All mixed lineage leukemia-ALL (MLL-ALL) models contain a canonical *KMT2A* fusion and have relatively silent genomes with minimal copy number alterations (Figure S4B). The majority of these models were derived from children <1 year of age (Table S1).

Molecular Subtyping and Genomic Landscape of CNS Tumors

Models derived from CNS and extracranial rhabdoid tumors were further molecularly classified using pathology reports or genomic features from WES, RNA, and SNP arrays (Figure 1; Table S1). Atypical teratoid rhabdoid tumor (ATRT) models represented both Sonic hedgehog (SHH; n = 3) and MYC (n = 3) subgroups, with two models unclassified. To classify medulloblastoma models, we developed and applied a classifier for RNA-seq data (STAR Methods). The 20 medulloblastoma models in this cohort span all broad subtypes: SHH (n = 7), WNT (n = 2), group 3 (n = 7), and group 4 (n = 3), and one model without RNA-seq remained unclassified. Other CNS embryonal tumors were classified into embryonal tumor with multi-layer rosettes (ETMR; n = 3), CNS Ewing with *CIC* alteration (CNS EFT-CIC; n = 2), ependymoblastoma (n = 1), or CNS embryonal not otherwise specified (CNS embryonal not otherwise specified [NOS]; n = 1). Astrocytoma models comprised pleomorphic xanthroastrocytomas (PXA; n = 2), MYCN subtype (n = 2), glioblastomas (IDH-wild-type; n = 5), histone H3-wild-type diffuse intrinsic pontine glioma (DIPG; n = 2), and a histone H3-wild-type astrocytic tumor (n = 1). Ependymal tumors were classified into supratentorial *RELA* positive (ST-EPN-RELA; n = 2), supratentorial *YAP1* positive (ST-EPN-YAP; n = 2), posterior fossa type A (PF-EPN-A; n = 1), or posterior fossa type B (PF-EPN-B; n = 1), and one remained unclassified.

All ATRT and extracranial rhabdoid models harbor inactivating alterations (focal deletion, frameshift deletion, or nonsense mutation) in the hallmark tumor suppressor, *SMARCB1*, and/or *SMARCA4*. Hedgehog, *TNFa*, and p53 signaling were enriched in these models (Figure 5B). Interestingly, three astrocytic tumors harbored *SMARCB1* hemizygous deletions, which have not been reported but are present in multiple pediatric high-grade glioma cohorts (Mackay et al., 2017: 6.7%, n = 834; Ijaz et al., 2019: 7.5%, n = 93) and may warrant further investigation. One astrocytic model, IC-1621GBM, was generated from a patient with DNA mismatch repair deficiency syndrome and showed 124 somatic mutations per medulloblastoma (MB) (Table S3). We confirmed multiple mutations in mismatch repair genes *PMS1*, *MSH2*, *MSH5*, and *POLE* (non-exonuclease domain mutation). The likely oncogenic drivers are the nonsense mutations in *PMS1* (Q316*) and *MSH2* (G721*), which disrupt the DNA mismatch repair protein domain and the MutS domain, respectively (Figure S3C). NCH-MN-1 was derived from a patient diagnosed with an anaplastic rhabdoid meningioma with the clinical suspicion of an ATRT; however, this model had no evidence of an inactivating *SMARCB1* alteration. Rather, it harbors a *BRAF* V600E mutation and focal *CDKN2A/B* deletion, classifying this model as a high-grade glioma, herein denoted as an astrocytoma. Not surprisingly, astrocytoma and glioblastoma models had similar pathway enrichment: estrogen response, hedgehog signaling, protein secretion, *TNFa*, and p53 pathway (Figure 5B).

IC-2664PNET was derived from a patient diagnosed with a primitive neuroectodermal tumor (PNET) but was further molecularly classified as a MYCN-subtype high-grade glioma. IC-2664PNET has a focal amplification of *MYCN* and a hemizygous *SMARCB1* deletion, but it retains mRNA expression of

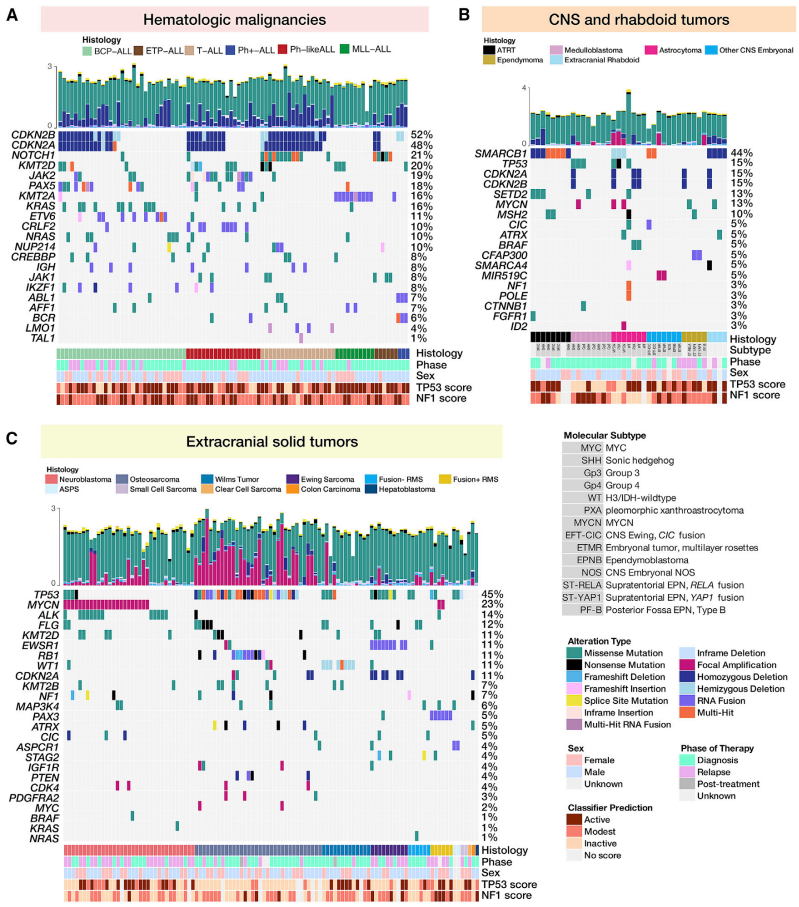


Figure 2. PDX Models Recapitulate the Mutational Landscape of Childhood Cancers

(A–C) Oncoprints of somatic alterations (homozygous deletions, amplifications, SNVs, and fusions) in hallmark driver genes for PDX models for which exome sequencing was performed (n = 240, top 20 genes per histology shown). Oncoprints are grouped by acute lymphoblastic leukemias (A), CNS and rhabdoid tumors (B), and extracranial solid tumors (C).

(A) From left to right are B cell precursor ALLs (n = 33), T cell ALLs (n = 25), Philadelphia chromosome positive (Ph+) ALLs (n = 3), mixed lineage leukemias (MLL, n = 10), early T cell precursor (ETP) ALLs (n = 6), and Philadelphia chromosome-like (Ph-like) ALLs (n = 19).

(B) From left to right are atypical teratoid rhabdoid tumors (ATRTs; n = 8), medulloblastomas (MBs; n = 8), astrocytomas (n = 7), non-MB/non-ATRT CNS embryonal tumors (n = 7), ependymomas (n = 5), and extracranial rhabdoid tumors (n = 4).

(C) From left to right are neuroblastomas (n = 35), osteosarcomas (n = 34), Wilms tumors (n = 13), Ewing sarcomas (n = 10), fusion negative rhabdomyosarcomas (n = 6), fusion positive rhabdomyosarcomas (n = 6), and rare solid tumors (n = 7). Clinical annotations for all models include histology, patient phase of therapy from which PDX was derived, and sex. CNS tumors were also annotated with molecular subtype. Hemizygous deletions in *TP53* are annotated for osteosarcoma models, in *CDKN2A* for leukemia models, and in *WT1* for Wilms tumor models. Focal homozygous deletions correspond to loss of expression (FPKM < 1) in models for which RNA was available. For fusions, only the 5′ partner is shown. Total mutations (log10) per model are plotted above each oncoprint and colored by mutation type. Each genomic assay was performed once per biological tumor sample.

SMARCB1. ICb-S1129MB, ICb-1343ENB, and IBs-2373PNET were classified as ETMRs due to an amplification of C19MC, an overexpression of *LIN28A*, and/or *TTYH1* fusions. ICb-9850PNET and IC-22909PNET-rIII, a diagnosis-relapse pair, were genetically classified as CNS EFT-CIC, as the diagnostic tumor contains a *CIC-DUX4* fusion. The two DIPG models were profiled with RNA-seq and SNP arrays and thus are not shown in the oncoprint. We confirmed both IBs-P1215DIPG and IBs-W0128DIPG have high expressions of *H3F3A* and *H3F3B* (FPKM > 50), genes encoding the histone H3.3 variant, and lack expressions of *HIST1H3B* or *HIST1H3C*, genes encoding the histone H3.1 variant. While we did not detect H3.1 or H3.3 histone mutations in these models, RNA variant calling revealed IBs-W0128DIPG contained predicted damaging (PolyPhen) missense mutations in *NRAS* (p.G13R, 0.41), *CIC* (p.C102Y, 0.44), and *KMT2C* (p.C988F, 0.45). We did not detect any hallmark damaging mutations in IBs-P1215DIPG.

Extracranial Solid Tumors

Neuroblastomas

Amplification of the *MYCN* oncogene was the most frequent alteration observed across all models (66%) and, as expected, was largely mutually exclusive of 11q deletion. Gene set enrichment analysis (GSEA) confirmed the enrichment of MYC targets in these models (Figure 5B). A majority (77%) of models had 1p deletion and 17q gain (97%; collapsed profiles are shown in Figure S4A). Consistent with previous reports (Pugh et al., 2013), we find *ALK* to be the most frequently mutated gene (37% of all models contain hotspot mutations) with additional, less frequent alterations in hallmark genes such as *TP53* (11%), *PTPN11* (9%), *NF1* (9%), *BRAF* (3%), *CIC* (3%), and *KRAS* (3%). The nonsense and frameshift deletions in *NF1* correspond with ablated expressions in COG-N-590x and NB-1771, respectively, but NB-1643 retains expression.

Osteosarcomas

The hallmark of osteosarcomas is *TP53* inactivation, and using a classifier trained on RNA expression data from TCGA, we found all osteosarcoma models with available RNA-seq data (n = 32) were predicted to have non-functional *TP53* (described below). Thus, as expected, *TP53* was the most commonly altered gene (82%) in osteosarcoma PDX models (Figure 2C), which also demonstrate global copy number changes, consistent with the high prevalence of complex genomic rearrangements found in this tumor type (Figure S4).

Ewing Sarcomas

The canonical *EWSR1-FLI1* fusion was found in all Ewing sarcoma models profiled with RNA-seq (NCH-EWS-1 was not profiled), and CHLA-258 contained an additional *FLI1* fusion partner: *RP11-9L18.2* (Table S5; Figure 2C). *TP53* mutations are present in seven cases (70%), with six showing allele frequencies at or near 1.0 due to copy-neutral loss of heterozygosity (cnLOH, ES-6, EW-8, and SK-NEP-1) or loss of heterozygosity (LOH) from a chromosomal arm deletion (EW-5, ES-8, and TC-71). Homozygous *CDKN2A/B* loss (60%) was mutually exclusive to *STAG2* mutations (20%), as expected (Tirode et al., 2014). We observe canonical (Tirode et al., 2014) broad gain of whole chromosomes 8 and 12, as well as focal 1q gain and 16q loss, in Ewing sarcomas (Figure S4A).

Wilms Tumors

The mutational and copy number landscapes of Wilms tumor (n = 13) PDX models are depicted in Figures 2C and S4A. The *WT1* gene located at 11p13 was mutated in one PDX model (NCH−WT−6−S13−1506), but we observed hemizygous deletions of *WT1* in 61% of Wilms models, many of which had LOHs of the entire 11p13 region. The 11p15.5 region, which contains imprint control regions (ICRs) 1 and 2, often undergoes loss of imprinting (LOI) either due to maternal DNA methylation or maternal LOH/paternal uniparental disomy (pUPD) in a Wilms tumor. The 11p15.5 region harbored LOHs in 69% (9/13) of Wilms tumors, consistent with previous reports (Scott et al., 2012). Two models (15%) harbored hemizygous deletions of *AMER1* (formerly known as *WTX* and/or *FAM123B*). KT-9 is the only model annotated as coming from a patient with bilateral disease, and although it does not harbor a *WT1* mutation, interestingly, it has two hits in *TP53*: a *TP53-FXR2* fusion and a partial homozygous deletion. The Wilms models (15%; KT-6 and NCH-WT-6-S13-1506) with *CTNNB1* mutations were mutually exclusive to *WTX* alterations, consistent with previous reports (Scott et al., 2012). Gains of the 1q arm, 1p LOH, and 16q LOH—adverse prognostic biomarkers for Wilms tumors (Pan et al., 2017; Segers et al., 2013; Spreafico et al., 2013)—were observed in 31% (4/13), 8% (1/13), and 23% (3/13) of models, respectively (Figure S4).

Rhabdomyosarcomas

All Fusion+ rhabdomyosarcoma (RMS) models harbored a hallmark *PAX3-FOXO1* fusion (Figure 2C; Table S1), and the median patient age of Fusion+ RMS patients (16 years) was higher than that of Fusion− RMS patients (5 years) (Table S3). As expected, we also observed focal amplifications of *MYCN* and *CDK4*. Interestingly, the amplification of *CDK4* was not retained in Rh-30R (relapse tumor paired with Rh-30; SNPs and STRs confirm identity). Ras pathway mutations (*NRAS*, *HRAS*, *KRAS*, and *NF1*) are typically observed in one-third of Fusion− RMS cases and here, Ras mutations were observed in 3/6 models (Rh-12 with *NF1 T2335fs*, NCH−ERMS−1−NCH−RMS−1 with *NRAS Q61K* mutation, and Rh-36 with *HRAS* Q61K). Of note, all models except for IRS-68 overexpress the common rhabdomyosarcoma biomarker, *MYOD1*.

Rare Histologies

Seven PDX models were derived from rare tumor types and are depicted in Figure 2C. Three models (43%) contained alterations in *TP53*; of note, an in-frame hemizygous deletion of *TP53* evolved at the relapse in NCH-CA-2 (not present in diagnostic model, NCH-CA-1). The canonical *ASPSCR1-TFE3* fusion was detected in both alveolar soft part sarcoma (ASPS) models. NCH-CA-1 and NCH-CA-2 harbored deleterious *SMARCA4* mutations, and NCH-CA-3 harbored a deleterious *NF1* nonsense mutation; each had a concurrent loss of mRNA expression and, as such, these may be potential drivers of oncogenesis in these tumors. NCH-HEP1 contained a likely oncogenic WNT pathway mutation (*CTNNB1* p.D32G).

Breakpoint Density

We calculated the total number of breakpoints per sample and breakpoint density within chromosomes, the latter as a surrogate measure of putative chromothripsis events

(STAR Methods). Consistent with pediatric cancer genomics literature, we observed very few breakpoints per sample in hematologic malignancies, compared to those in solid tumors (median = 3 breakpoints per sample in CNS embryonal NOS to median = 154.5 breakpoints per sample in osteosarcoma; Figure S4C; Table S3). We found 25% (64/252) of models profiled have a high breakpoint density (HBD) across one or more chromosomes (Figure S4D; Table S3), consistent with a recent pan-cancer chromothripsis report (Cortes-Ciriano et al., 2018). Specifically, 97% (33/34) of osteosarcomas had HBDs; 30% (10/33) of these contained HBDs on four or fewer chromosomes indicative of localized chromothripsis events, while the remaining 70% (23/33) contained HBDs on five or more chromosomes, supporting the globally rearranged genomes prevalent in this tumor type (Lorenz et al., 2016). In neuroblastoma samples, 17% of models contained HBDs on chromosomes 2, 5, 16, 17, and 19. Chromothripsis events on chromosomes 2, 5, and 17 in neuroblastoma tumors have been previously reported to be associated with *MYCN* amplification, *TERT* rearrangements, and 17q gain, respectively (Molenaar et al., 2012; Boeva et al., 2013). Recurrent loci with HBDs in medulloblastoma were chromosomes 2, 8, 14, and 17, consistent with recent reports (Rausch et al., 2012). In summary, PDX models faithfully recapitulate important prognostic copy number alterations of pediatric tumors.

Mutational Landscape of Models Derived from Tumors at Relapse

The majority of the PDX models were established at diagnosis (63%), but 6% were derived from surgical resection specimens after neoadjuvant therapy, 27% were from a relapsed specimen (14% of those were neuroblastomas from a large volume blood draw obtained immediately after death from disease progression), and 4% did not have the phase of therapy annotated. In addition, 12 pediatric cancer patients had either two or three models created across the spectrum of their therapy (Table S1). Here, we compare mutation frequencies and tumor mutation burdens (TMBs) for histologies with paired diagnosis-relapse cohorts with group $N \geq 6$: BCP-ALL ($N_{diagnosis}$ = 19/$n_{relapse}$ = 14), T-ALL ($n_{diagnosis}$ = 11/$n_{relapse}$ = 8), osteosarcoma ($n_{diagnosis}$ = 25/$n_{relapse}$ = 6), and neuroblastoma ($n_{diagnosis}$ = 12/$n_{relapse}$ = 23). Across all four histologies, there is an increased frequency of key hallmark gene alterations in relapsed disease, as indicated by the oncoprint frequencies (Figure 3A). Using somatic missense and nonsense mutations, we calculated the TMB for each PDX model (STAR Methods). The median TMB across all models was 2.66 somatic mutations per megabase (Mut/Mb; Figure S3B; Table S3). The TMBs across this cohort of PDX models are likely higher than those in previous reports for two main reasons. First, 37% of the PDX models were derived from a patient tumor at a phase of therapy other than diagnosis, and it is now known that tumors acquire significantly more somatic mutations post-therapy and following a relapse (Eleveld et al., 2015; Ma et al., 2015; Padovan-Merhar et al., 2016; Schleiermacher et al., 2014; Schramm et al., 2015). Second, without paired normal samples, rare germline and private variants could not be reliably removed from the ''somatic'' MAF. Thus, the TMB reported here is likely inflated, but the trends across histologies and phase of therapy should accurately reflect TMBs determined with a paired germline sample. In fact, we observe an overall significantly higher TMB in PDX models derived from relapse tissue (3.08 Mut/Mb) compared to those derived at diagnosis (2.57 Mut/Mb, Wilcoxon p = 2.2e−5; Figure 3B). When compared to diagnostic tumors within a histology, the TMB was higher at relapse in BCP-ALL (Wilcoxon p = 0.054) and significantly higher at relapse in neuroblastoma (Wilcoxon p = 0.016) and T-ALL (Wilcoxon p = 0.0081), but it was not different between the diagnosis and relapse for osteosarcoma (Wilcoxon p = 0.42). Finally, we compared TMBs between paired diagnosis-relapse models and found a significantly higher TMB in models derived from relapse tumors (Figure 3B; median of 98.0 versus 27.5 mutations; Wilcoxon p = 0.0083). This PDX cohort recapitulates relapsed disease and provides a model for further studying tumor progression and therapeutic resistance.

Expression Signatures Classify Pediatric PDX Models for TP53 and NF1 Inactivation

A recent study used TCGA data to classify tumors for *TP53* inactivation status and found that alterations in multiple genes phenocopy *TP53* inactivation, indicating that *TP53* mutation status alone is not necessary to infer the inactivation of the pathway (Knijnenburg et al., 2018). We applied a machine learning algorithm to infer *TP53* inactivation, *NF1* inactivation, and Ras pathway activation using PDX tumor transcriptomes. These classifiers were previously trained using gene expression data from TCGA PanCanAtlas (STAR Methods) (Knijnenburg et al., 2018; Way et al., 2017, 2018). The *TP53* (area under the receiver operator characteristic [AUROC] = 0.89) and *NF1* (AUROC = 0.77) classifiers are both accurate compared to a shuffled gene expression baseline, but performance of the Ras classifier (AUROC = 0.55) was relatively poor (Figure 4A), which may be attributed to differences in Ras pathway signatures in pediatric compared to adult tumors. Classifier scores >0.5 predict the inactivation of *TP53* or *NF1* (Table S5), and *TP53* scores are significantly higher (Wilcoxon p < 2.2e−16) in models with a *TP53* alteration (mean score = 0.790) compared to those without alterations (mean score = 0.419) (Figure 4B). Many models annotated as wild-type *TP53* have high *TP53* inactivation scores (Figure 4B). We found models with alterations in genes such as *MDM2* and *RB1* also have high *TP53* inactivation scores. These alterations may phenocopy *TP53* alterations (Figure 4C; genes chosen as primary or secondary interactors of *TP53* defined by the TP53 KEGG signaling pathway). In Figure 4D, we plot alterations for each gene by variant classification. Notably, all types of alterations within *TP53* were associated with high classifier scores, while the scores for other genes varied by type of alteration.

As *TP53* inactivation is a hallmark of osteosarcoma, we focused on these models as a proof of concept. The classifier predicted that all models profiled with RNA-seq except OS-55-SBX had *TP53* pathway inactivation. Many had a genetic alteration in a *TP53* pathway gene as supporting evidence (Figure 4E; Table S4). However, the mechanisms of *TP53* inactivation in OS-34-SJ, OS-43-TPMX, and OS-51-CHLX are still unknown and may require whole-genome sequencing to detect. To ensure

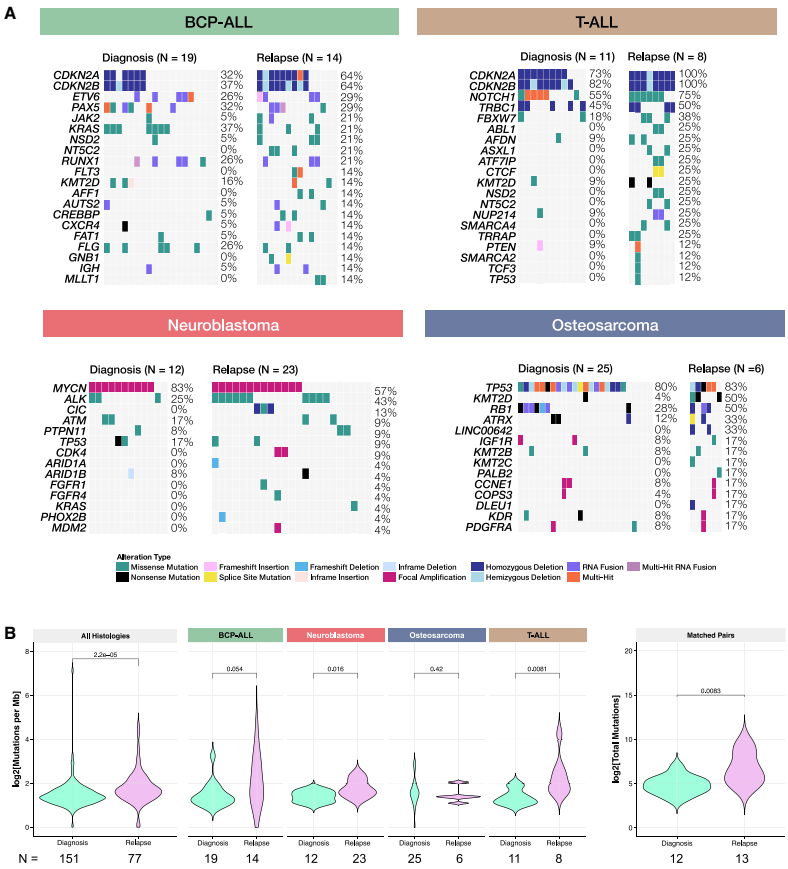


Figure 3. Mutational Landscape of Models Derived from Tumors at Relapse

(A) For BCP-ALL, T-ALL, neuroblastoma, and osteosarcoma (histologies with N ≥ 6 models and multiple phases of therapy), oncoprints comparing hallmark alterations in models derived from diagnosis tumors to models derived from relapse tumors.

(B) Tumor mutation burden (TMB) is significantly (or near significantly) higher in relapse models, compared to models established at diagnosis for all histologies collapsed (n_{dx} = 151, n_{rel} = 77, Wilcoxon p = 2.2e−5), BCP-ALL (n_{dx} = 19, n_{rel} = 14, Wilcoxon p = 0.051), neuroblastoma (n_{dx} = 12, n_{rel} = 23, Wilcoxon p = 0.016), and T-ALL (n_{dx} = 11, n_{rel} = 8, Wilcoxon p = 0.0081). There was no difference between osteosarcoma models established at diagnosis and relapse (n_{dx} = 25, n_{rel} = 6, Wilcoxon p = 0.42). For patients in which models were established at both diagnosis and relapse, there was a significant increase in mutational burden upon relapse (n_{dx} = 12, n_{rel} = 13, p = 0.0083). All n's denote biological replicates.

osteosarcoma models were not driving the observed association with *TP53* scores, we removed the osteosarcoma models and reanalyzed the data. We found a significantly higher *TP53* classifier score (Wilcoxon p = 1.0e−11) in models with alterations in *TP53* pathway genes (Figures S5A and S5B). We then evaluated which types of variants were associated with high *TP53* classification scores and observed that models containing fusions had highest classifier scores compared to wild types, followed by models with single nucleotide variants (SNVs) and copy number variants (CNVs). (Figure S5C; Kruskal-Wallis p = 9.8e−11). These are broken down by gene in Figure S5D. Outside of osteosarcomas, only one model contained a fusion in the *TP53* pathway: Wilms model KT-9 contained a *TP53-FXR2* fusion. We found the overall copy number burden (number of breakpoints calculated

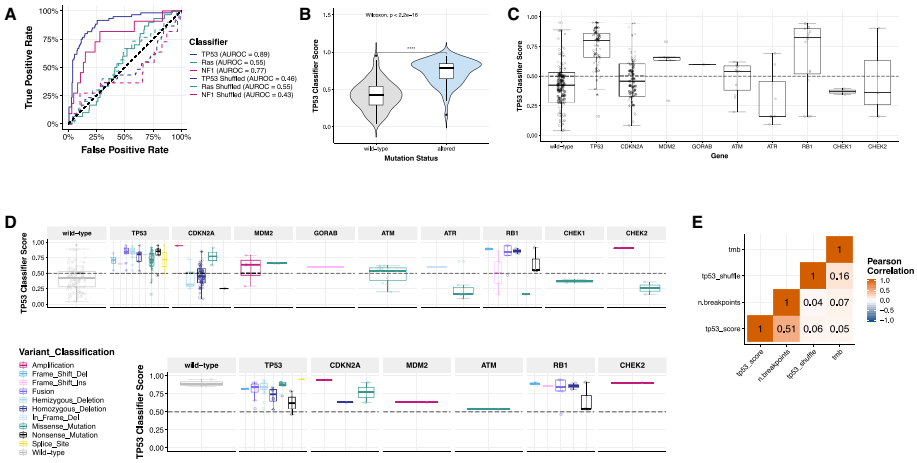


Figure 4. Expression and Mutational Signatures Classify Pediatric PDX Models for *TP53* and *NF1* Inactivation

(A) Only *TP53* and *NF1* classifiers performed well in our dataset ($AUROC_{TP53}$ = 0.89, $AUROC_{NF1}$ = 0.77, $AUROC_{Ras}$ = 0.55). Solid lines represent real scores, and dotted lines represent shuffled scores. For the samples measured (n = 244), 60 had *TP53* alterations (24.6%); 30 had *KRAS*, *HRAS*, or *NRAS* alterations (12.3%); and 11 had *NF1* alterations (4.5%).

(B) *TP53* scores are significantly higher (n_{WT} = 120, n_{ALT} = 124, Wilcoxon p < 2.2 e−16) in models with genetic aberrations in *TP53* (mean score = 0.790) compared to those without alterations (mean score = 0.419).

(C) Classifier scores are plotted based on the *TP53* pathway gene alteration present (n_{WT} = 120, n_{TP53} = 72, n_{CDKN2A} = 63, n_{MDM2} = 5, n_{GORAB} = 1, n_{ATM} = 11, n_{ATR} = 7, n_{RB1} = 16, n_{CHEK1} = 2, n_{CHEK2} = 3) or variant classification (n = 244 total samples).

(D) *TP53* classifier scores across all histologies broken down by *TP53* pathway gene (n = 240).

(E) In osteosarcoma models (n = 30), all scores, regardless of variant type or gene, were high and predicted pathway inactivation. Overall copy number burden (number of breakpoints calculated from SNP array data; STAR Methods) correlates significantly with *TP53* classifier score (R = 0.51, p = 1.8e−17, n = 239). All n's denote biological replicates.

from SNP array data; STAR Methods), but not the TMB or shuffle score, correlates significantly with the *TP53* classifier score (Figure 4E; R = 0.51, p = 1.8e−17), supporting recent published observations (Knijnenburg et al., 2018). Genetic alterations rendering *TP53* inactive may contribute to copy number instability in these models. The use of gene expression classifiers can guide preclinical studies: for example, therapeutically targeting the *TP53* pathway in tumors with high *TP53* inactivation scores rather than those with altered *TP53*.

Expression Profiles of PDX Models Cluster by Tissue of Origin and Contain Driver Fusions

We used the UCSC TumorMap (Newton et al., 2017) to visualize clusters of expression profiles across PDX histologies (Figure 5A). We observed a clear separation among unrelated histologies and an overlapping clustering among related histologies. For example, T-ALL and ETP-ALL cluster together as expected, but distinctly from other ALL histologies. The leukemias clustered by subtype and distinctly from solid tumors. Ewing sarcoma, neuroblastoma, Wilms, and medulloblastoma form distinct clusters. Osteosarcomas cluster with two ASPS models. Fusion+ and Fusion- RMS cluster near each other but distinctly. Brain tumor histologies cluster near each other with the exception of ATRTs, some of which cluster with extracranial rhabdoid tumors near sarcoma samples. We identified histology-specific expression differences using a Bayesian hierarchical model (Gelman, 2006), grouped related histologies under the same prior distribution, ranked gene expression differences for each histology, and performed GSEA. This demonstrated tissue-specific enrichment within each histology, using GSEA and the Tissue-Specific Gene Database in Cancer (TissGDB; Kim et al., 2018a) and Tissue-Specific Gene Expression and Regulation Database (TiGER; Liu et al., 2008) gene sets (Figure S5F). To investigate pathway enrichment within histologies, we ran GSEA using the MSigDB curated (C2) gene sets and plotted the normalized enrichment scores (NESs) for the Hallmark pathway gene sets in Figure 5B.

Next, we created a high-confidence fusion annotation pipeline (Figure S1; STAR Methods) using four algorithms: defuse, FusionCatcher, STARFusion, and SOAPFuse. A total of 50,796 unique fusions were called, and we defined 925 unique high-confidence fusions and 92 unique known oncogenic driver fusions defined by cytogenetics and literature (Figure 5C; Table S5). Fusions were annotated for their frame and for whether a gene partner is a known oncogene, kinase, or transcription factor to identify oncogenic potential and functional relevance. We found that PPTC PDX models largely maintain known oncogenic

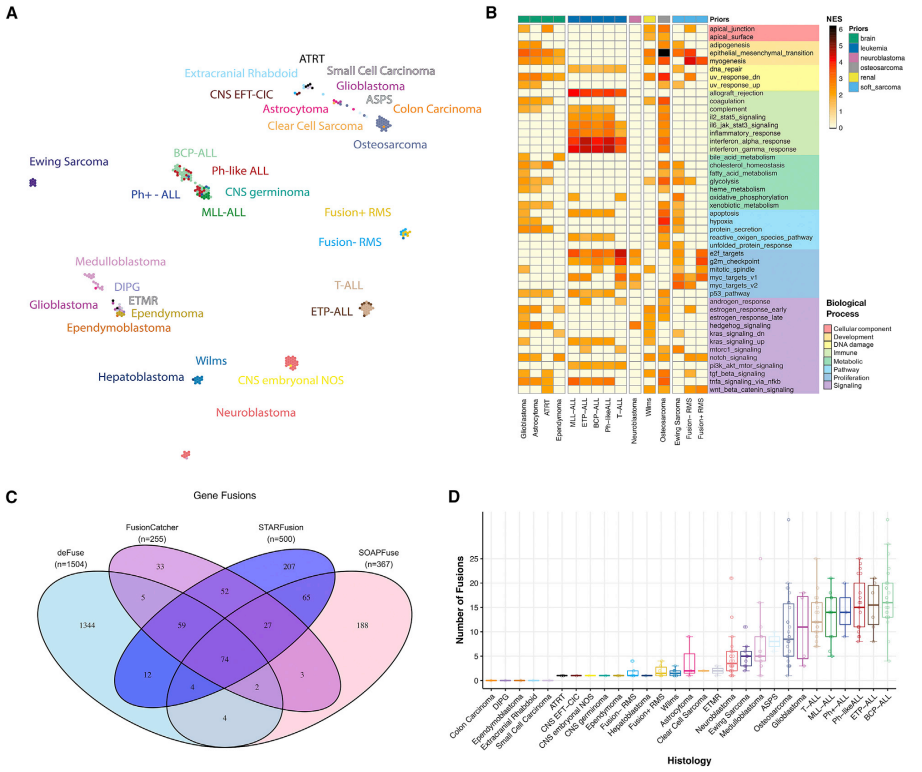


Figure 5. Expression Profiles of PDX Models Cluster by Histology and Contain Driver Fusions

(A) TumorMap rendition of PDX RNA-seq expression matrices by histology.

(B) Gene set enrichment analysis for Hallmark pathways for histologies with n ≥ 4 samples demonstrates histology-specific biologic processes significantly altered (adjusted $p < 0.05$ and NES > 2.0, N = 221). Samples were grouped by prior before GSEA ($n_{bone\ sarcoma}$ = 10, n_{brain} = 58, $n_{leukemia}$ = 90, $n_{neuroblastoma}$ = 35, $n_{osteosarcoma}$ = 36, n_{renal} = 14, $n_{soft\ sarcoma}$ = 18).

(C and D) Venn diagram of RNA fusion overlap among four algorithms (C) and high-confidence fusion totals (D) demonstrates a higher overall number of fusions in hematologic malignancies (boxplots are graphed as medians with box edges as first and third quartiles; detailed Ns in Table S3). n = 244 RNA samples used as input, and all n's represent biological replicates.

driver fusions specific to their histologies: all alveolar rhabdomyosarcoma models harbored *PAX3-FOXO1* fusions, all Ewing sarcoma samples with RNA-seq data showed *EWSR1-FLI1* fusions, all Ph+ ALL tumors contained *BCR-ABL1* fusions, and *KMT2A* (*MLL*) fusions were detected in all MLL-ALL models (Table S5). Osteosarcomas harbored *TP53* fusions, and breakpoints reside within intron one of the *TP53* gene, a mechanism of TP53 inactivation previously reported in osteosarcoma (Ribi et al., 2015). In five diagnosis-relapse pairs, we detected four fusions in the diagnostic PDX (*PAX5-RP11-465M18.1*, *IGH-MYC*, *CIC-DUX4*, and *TP53-TNR*) that were undetected in their paired relapse model, suggesting these specific gene fusions may have been acquired after an alternative initiating event that was retained.

DISCUSSION

Here, we used whole-exome, whole-transcriptome, SNP genotyping arrays, and STR profiling to characterize 261 pediatric PDX models across 37 unique molecular subtypes. We used a competitive mapping approach to remove mouse reads from DNA or RNA-seq data and demonstrated high concordance

between these pipelines and the orthogonal measurement of human:mouse DNA ratios. We showed a faithful recapitulation of primary and relapsed disease within tumor of origin type through analysis of somatic mutations, copy number alterations, RNA expression, gene fusions, and oncogenic pathways. It is clear that the models here are biased toward the most highly aggressive pediatric cancers, which is reflective of the typical pediatric phase 1 patient populations.

The data presented herein have immediate applications to the prioritization of experimental agents for testing in pediatric preclinical models, leading to eventual clinical testing. For example, there are reports identifying specific genomic alterations as predicting sensitivity to ATR inhibitors, including ATM loss, ARID1A mutation, defective homologous recombination, and ATRX mutation associated with alternative lengthening of telomeres (ALTs) (Lecona and Fernandez-Capetillo, 2018). Querying the PPTC data at PedcBioPortal can quickly identify models with these characteristics, and the models can then be used to test whether *in vivo* responsiveness to ATR inhibitors is predicted by one or more of the molecular characteristics. Similarly, PPTC RNA-seq data can be used to identify models that show elevated gene expression for the targets of immunotherapeutics such as antibody-drug conjugates and T-cell engagers. As examples, in the PPTC dataset, GPC2 and ALK are dramatically overexpressed neuroblastoma models, as previously published (Bosse et al., 2017; Sano et al., 2019), but also in multiple subsets of additional pediatric cancer histotypes, allowing for a basket trial design for preclinical testing. The PPTC RNA-seq dataset was also used to identify T-ALL as a target histology for an agent activated by the aldo-keto reductase AKR1C3 (R.B. Lock et al., 2018, Mol. Cancer Ther., abstract) and to identify ASPS xenografts as intrinsically overexpressing CD274 (PD-L1), making ASPS a target histology for the evaluation of checkpoint inhibition (C.G. O'Sullivan et al., 2018, Connective Tissue Oncology Society Annual Meeting, conference).

Further, we performed machine learning to classify tumors into TP53 and NF1 active or inactive, and we suggest that these scores might be future biomarkers for drug response. These classifiers have been used to identify tumors that may respond to novel agents, including those that target tumors driven by NF1 loss (Way et al., 2017). Although these machine learning algorithms are not ready for the clinic, the next logical step is to use PDX models to test the predictive nature of classifiers so that in the future, interdisciplinary teams can identify tumors driven by TP53 and/or NF1 loss, evaluate, and compare multiple therapies in real time.

Our study also highlights additional opportunities for pan-pediatric genomic characterization. We did not have available models for acute myelogenous leukemia, juvenile myelomonocytic leukemia, lymphomas, retinoblastoma, melanoma, thyroid malignancies, or histone mutant midline gliomas. Additionally, although we covered 37 molecular subtypes, many of the rare tumors had low numbers of models and could benefit from the creation and sequencing of additional PDXs, and we seek to generate these data and/or hope to merge our data with future pediatric cancer PDX sequencing projects. Finally, WES likely missed several pathogenic lesions, and DNA methylation profiling is particularly relevant for pediatric brain tumors. Future studies, perhaps in collaboration with ongoing similar efforts by international colleagues, could address these gaps.

We performed this project to provide a resource to the pediatric cancer research community. To date, the pediatric cancer genomic literature largely focuses on diagnostic samples, and this study includes a large number of PDXs derived during or after intensive chemoradiotherapy. Thus, the frequency of many genomic alterations is higher in these models compared to the literature. By having a large number of PDXs obtained from samples at relapse or at autopsy, we can provide models that more closely recapitulate the patients being enrolled in early-phase clinical trials after extensive chemoradiotherapy. All models and data are freely available for the cancer research community, as described in the STAR Methods.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Patient-Derived Xenograft Generation and Harvesting
- METHOD DETAILS
  - Nucleic Acid Extractions and Quality Control
  - Short Tandem Repeat (STR) Profiling
  - Biochemical Measurement of Human DNA Content in PDX Tumors
  - Additional Quality Control for Cross-Contamination and Mis-Identification
  - Whole Exome Sequencing
  - SNP Array Assay
  - Whole Transcriptome Sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Mouse Read Subtraction from WES Sequencing Data
  - Whole Exome Mutation Analysis
  - Tumor Mutation Burden Analysis
  - ATRX Deletion Analysis
  - Mutational Signatures Analysis
  - Classifier Analysis
  - mRNA Gene Expression Analysis
  - mRNA Variant Calling, Filtering, and Comparison to DNA Variants
  - Copy Number Analysis
  - Breakpoint Analysis
  - Ethnicity Inference
  - Fusion Transcript Analysis
  - RNA Expression Clustering and Pathway Analyses
  - Pediatric cBioPortal Data Processing
- DATA AND CODE AVAILABILITY
  - Raw Data Availability
- INTERMEDIATE PROCESSED DATA AVAILABILITY
  - Processed Data Availability
  - Code Created or Modified for Analysis in This Paper Have Been Deposited in GitHub

OPEN ACCESS
CellPress

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2019.09.071.

ACKNOWLEDGMENTS

The authors would like to thank the Alex's Lemonade Stand Foundation for providing the funding for this project. This work was also supported by NIH grants U01 CA199287 (J.M.M.), U01 CA199000 (R.B.L.), U01 CA199288 (X.-N.L.), U01 CA199221 (R.G.), U01 CA199297 (P.J.H.), U01 CA199222 (G.J.G.), R35 CA220500 (J.M.M.), and R01 CA221957 (C.P.R.), and NINDS R01 NS095411-01A1 (Y.S.); the Giulio D'Angio Endowed Chair (J.M.M.); the National Health and Medical Research Council of Australia (NHMRC fellowships APP1059804 and APP1157871 to R.B.L. and NHMRC program grant APP1091261); the Cancer Council New South Wales (PG 16-01); and Australian Federal Government Department of Health funding awarded to Zero Childhood Cancer, a joint initiative of Children's Cancer Institute Australia (affiliated with UNSW Sydney) and The Kids Cancer Centre, Sydney Children's Hospitals Network. We also acknowledge the NHLBI GO Exome Sequencing Project and its ongoing studies, which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926), and the Heart GO Sequencing Project (HL-103010).

We acknowledge sequencing and molecular characterization efforts of the PPTP, with which our data were harmonized (Brian Geier, D.K.). We also thank the following researchers for PDX model establishment and maintenance: Edward Favours, Doris Phelps, Alessia D'Aulerio, Colleen Larmour, and Matthew Tsang. The authors also thank Dr. David T. Teachey (Children's Hospital of Philadelphia, Philadelphia, PA), Professor Charles G. Mullighan (St Jude Children's Research Hospital, Memphis, TN), and the Children's Oncology group for providing material from which the ETP-ALL and Ph-like ALL xenografts were established and the Children's Cancer Institute Tumour Bank for providing samples and related clinical information for this study. We thank Chia Chin Wu and Jianhua Zhang for helpful discussions throughout the study and Bobby Moulder for assistance with Figure 1 graphics.

Finally, the authors gratefully acknowledge the patients and their families for donating samples for PDX establishment.

AUTHOR CONTRIBUTIONS

Conceptualization, J.M.M., D.A.W., J.L.R., M.A.S., P.J.H., R.T.K., X.-N.L., R.B.L., R.G., and M.H.; Methodology, D.A.W., J.L.R., J.P., G.P.W., and K.S.R.; Software, K.S.R., M.F.C., A.F., J.L.R., G.P.W., K.P., and J.P.; Validation, C.P.R., J.L.R., K.P., K.L.C., K.A.U., K.M., J.N., F.K.B., and K.S.G.; Formal Analysis, D.A.W., J.L.R., K.S.R., M.F.C., J.P., L.E.E., N.M.K., G.P.W., C. Mayoh, K.S.G., and P.R.; Investigation, J.L.R., J.P., K.S.R., K.A.U., K.L.C., G.P.W., K.L.C., and J.N.; Resources, J.M.G.-F., J.B., K.M.L., S.E.C., A.C.B., J.M.M., K. Krytska, Y.P.M., R.B.L., C.S.G., P.J.H., X.-N.L., R.T.K., H.V.D., Z.M., J.J., K.M., J.N., C. Morton, D.K., K.E., H.M., J.W.B., K.B., F.K.B., L.Q., Y.D., H.Z., H.B.L., S.Z., J.S., and P.B.; Data Curation, D.A.W., M.F.C., J.L.R., K.S.R., K.A.U., K. Kalletla, G.I.S., C. Mayoh, K.E., H.M., K.B., J.W.B., and E.R.B.; Writing - Original Draft, J.L.R., J.M.M., D.A.W., J.J., K.S.R., G.P.W., J.P., N.M.K., L.E.E., and K.A.U.; Writing - Review & Editing, J.L.R., M.A.S., J.M.M., C.S.G., C. Mayoh, R.B.L., Y.S., S.Z., and K. Krytska; Visualization, J.L.R., K.S.R., J.P., N.M.K., K.P., L.E.E., G.L., and A.M.; Supervision, J.M.M., D.A.W., J.L.R., M.A.S., D.H., C.P.R., S.J.D., O.M.V., and Z.V.; Project Administration, J.M.M., J.M.G.-F., J.B., K.M.L., M.A.S., and G.J.G.; Funding Acquisition, J.M.M., J.M.G.-F., D.A.W., C.P.R., R.B.L., J.L.R., M.H., G.M.M.,V.T., Y.S., R.G., P.J.H., and G.J.G.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 31, 2019
Revised: July 10, 2019
Accepted: September 24, 2019
Published: November 5, 2019

REFERENCES

Alcoser, S.Y., Kimmel, D.J., Borgel, S.D., Carter, J.P., Dougherty, K.M., and Hollingshead, M.G. (2011). Real-time PCR-based assay to quantify the relative amount of human and mouse tissue present in tumor xenografts. BMC Biotechnol. *11*, 124.

American Childhood Cancer Organization (2014). Special Section: Cancer in Children & Adolescents. In Cancer Facts and Figures 2014 (The American Cancer Society), pp. 25–42.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Bainbridge, M.N., Wang, M., Wu, Y., Newsham, I., Muzny, D.M., Jefferies, J.L., Albert, T.J., Burgess, D.L., and Gibbs, R.A. (2011). Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol. *12*, R68.

Baker, S.C., Bauer, S.R., Beyer, R.P., Brenton, J.D., Bromley, B., Burrill, J., Causton, H., Conley, M.P., Elespuru, R., Fero, M., et al.; External RNA Controls Consortium (2005). The External RNA Controls Consortium: a progress report. Nat. Methods *2*, 731–734.

Behjati, S., Tarpey, P.S., Haase, K., Ye, H., Young, M.D., Alexandrov, L.B., Farndon, S.J., Collord, G., Wedge, D.C., Martincorena, I., et al. (2017). Recurrent mutation of IGF signalling genes and distinct patterns of genomic rearrangement in osteosarcoma. Nat. Commun. *8*, 15936.

Birney, E., and Soranzo, N. (2015). Human genomics: The end of the start for population sequencing. Nature *526*, 52–53.

Boeva, V., Jouannet, S., Daveau, R., Combaret, V., Pierre-Eugène, C., Cazes, A., Louis-Brennetot, C., Schleiermacher, G., Ferrand, S., Pierron, G., et al. (2013). Breakpoint features of genomic rearrangements in neuroblastoma with unbalanced translocations and chromothripsis. PLoS One *8*, e72182.

Bosse, K.R., Raman, P., Zhu, Z., Lane, M., Martinez, D., Heitzeneder, S., Rathi, K.S., Kendsersky, N.M., Randall, M., Donovan, L., et al. (2017). Identification of GPC2 as an Oncoprotein and Candidate Immunotherapeutic Target in High-Risk Neuroblastoma. Cancer Cell *32*, 295–309.e12.

Brabetz, S., Leary, S.E.S., Gröbner, S.N., Nakamoto, M.W., Şeker-Cin, H., Girard, E.J., Cole, B., Strand, A.D., Bloom, K.L., Hovestadt, V., et al. (2018). A biobank of patient-derived pediatric brain tumor models. Nat. Med. *24*, 1752–1761.

Cancer Genome Atlas Research Network; Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. J. Stat. Softw. *76*, 1–32.

Challis, D., Yu, J., Evani, U.S., Jackson, A.R., Paithankar, S., Coarfa, C., Milosavljevic, A., Gibbs, R.A., and Yu, F. (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics *13*, 8.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Cortes-Ciriano, I., Lee, J.K., Xi, R., Jain, D., Jung, Y.L., Yang, L., Gordenin, D., Klimczak, L.J., Zhang, C.-Z., Pellman, D.S., et al. (2018). Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. bioRxiv. https://doi.org/10.1101/333617.

DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W., and Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics *28*, 1530–1532.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

El-Hoss, J., Jing, D., Evans, K., Toscan, C., Xie, J., Lee, H., Taylor, R.A., Lawrence, M.G., Risbridger, G.P., MacKenzie, K.L., et al. (2016). A single nucleotide polymorphism genotyping platform for the authentication of patient derived xenografts. Oncotarget *7*, 60475–60490.

Eleveld, T.F., Oldridge, D.A., Bernard, V., Koster, J., Colmet Daage, L., Diskin, S.J., Schild, L., Bentahar, N.B., Bellini, A., Chicard, M., et al. (2015). Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. Nat. Genet. *47*, 864–871.

Gelman, A. (2006). Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. Technometrics *48*, 432–435.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Gröbner, S.N., Worst, B.C., Weischenfeldt, J., Buchhalter, I., Kleinheinz, K., Rudneva, V.A., Johann, P.D., Balasubramanian, G.P., Segura-Wang, M., Brabetz, S., et al. (2018). The landscape of genomic alterations across childhood cancers. Nature *555*, 321–327.

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849.

Haas, B.J., Dobin, A., Stransky, N., Li, B., Yang, X., Tickle, T., Bankapur, A., Ganote, C., Doak, T.G., Pochet, N., et al. (2017). STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. bioRxiv. https://doi.org/10.1101/120295.

Houghton, P.J., Adamson, P.C., Blaney, S., Fine, H.A., Gorlick, R., Haber, M., Helman, L., Hirschfeld, S., Hollingshead, M.G., Israel, M.A., et al. (2002). Testing of new agents in childhood cancer preclinical models: meeting summary. Clin. Cancer Res. *8*, 3646–3657.

Houghton, P.J., Morton, C.L., Tucker, C., Payne, D., Favours, E., Cole, C., Gorlick, R., Kolb, E.A., Zhang, W., Lock, R., et al. (2007). The pediatric preclinical testing program: description of models and early testing results. Pediatr. Blood Cancer *49*, 928–940.

Ijaz, H., Koptra, M., Gaonkar, K.S., Rokita, J., Baubet, V.P., Yauhid, L., Zhu, Y., Brown, M., Lopez, G., Zhang, B., et al. (2019). Pediatric High Grade Glioma Resources From The Children's Brain Tumor Tissue Consortium (CBTTC) And Pediatric Brain Tumor Atlas (PBTA). bioRxiv. https://doi.org/10.1101/656587.

Ji, H., and Liu, X.S. (2010). Analyzing 'omics data using hierarchical models. Nat. Biotechnol. *28*, 337–340.

Jia, W., Qiu, K., He, M., Song, P., Zhou, Q., Zhou, F., Yu, Y., Zhu, D., Nickerson, M.L., Wan, S., et al. (2013). SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biol. *14*, R12.

Jun, G., Wing, M.K., Abecasis, G.R., and Kang, H.M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. *25*, 918–925.

Kim, P., Park, A., Han, G., Sun, H., Jia, P., and Zhao, Z. (2018). TissGDB: tissue-specific gene database in cancer. Nucleic Acids Res. *46* (*D1*), D1031–D1038.

Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., et al. (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods *15*, 591–594.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B.E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J.B., Grout, J., Corlay, S., et al. (2016). Jupyter Notebooks-a publishing format for reproducible computational workflows. In Positioning and Power in Acadmic Publishing: Players, Agents and Agendas, F. Loizides and B. Scmidt, eds. (IOS Press), pp. 87–90.

Knijnenburg, T.A., Wang, L., Zimmermann, M.T., Chambwe, N., Gao, G.F., Cherniack, A.D., Fan, H., Shen, H., Way, G.P., Greene, C.S., et al.; Cancer Genome Atlas Research Network (2018). Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. Cell Rep. *23*, 239–254.e6.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Lecona, E., and Fernandez-Capetillo, O. (2018). Targeting ATR in cancer. Nat. Rev. Cancer *18*, 586–595.

Lee, S., Lee, S., Ouellette, S., Park, W.-Y., Lee, E.A., and Park, P.J. (2017). NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types. Nucleic Acids Res. *45*, e103.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

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

Liem, N.L.M., Papa, R.A., Milross, C.G., Schmid, M.A., Tajbakhsh, M., Choi, S., Ramirez, C.D., Rice, A.M., Haber, M., Norris, M.D., et al. (2004). Characterization of childhood acute lymphoblastic leukemia xenograft models for the preclinical evaluation of new therapies. Blood *103*, 3905–3914.

Liu, X., Yu, X., Zack, D.J., Zhu, H., and Qian, J. (2008). TiGER: a database for tissue-specific gene expression and regulation. BMC Bioinformatics *9*, 271.

Liu, Y., Easton, J., Shao, Y., Maciaszek, J., Wang, Z., Wilkinson, M.R., McCastlain, K., Edmonson, M., Pounds, S.B., Shi, L., et al. (2017). The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. Nat. Genet. *49*, 1211–1218.

Lock, R.B., Liem, N., Farnsworth, M.L., Milross, C.G., Xue, C., Tajbakhsh, M., Haber, M., Norris, M.D., Marshall, G.M., and Rice, A.M. (2002). The nonobese diabetic/severe combined immunodeficient (NOD/SCID) mouse model of childhood acute lymphoblastic leukemia reveals intrinsic differences in biologic characteristics at diagnosis and relapse. Blood *99*, 4100–4108.

Lorenz, S., Barøy, T., Sun, J., Nome, T., Vodák, D., Bryne, J.-C., Håkelien, A.-M., Fernandez-Cuesta, L., Möhlendick, B., Rieder, H., et al. (2016). Unscrambling the genomic chaos of osteosarcoma reveals extensive transcript fusion, recurrent rearrangements and frequent novel TP53 aberrations. Oncotarget *7*, 5273–5288.

Ma, X., Edmonson, M., Yergeau, D., Muzny, D.M., Hampton, O.A., Rusch, M., Song, G., Easton, J., Harvey, R.C., Wheeler, D.A., et al. (2015). Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. Nat. Commun. *6*, 6604.

Ma, X., Liu, Y., Liu, Y., Alexandrov, L.B., Edmonson, M.N., Gawad, C., Zhou, X., Li, Y., Rusch, M.C., Easton, J., et al. (2018). Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature *555*, 371–376.

Mackay, A., Burford, A., Carvalho, D., Izquierdo, E., Fazal-Salom, J., Taylor, K.R., Bjerke, L., Clarke, M., Vinci, M., Nandhabalan, M., et al. (2017). Integrated Molecular Meta-Analysis of 1,000 Pediatric High-Grade and Diffuse Intrinsic Pontine Glioma. Cancer Cell *32*, 520–537.e5.

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., and Koeffler, H.P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. *28*, 1747–1756.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome

Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

McKinney, W. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (SCIPY), pp. 51–56.

McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M.G.F., Griffith, M., Heravi Moussavi, A., Senz, J., Melnyk, N., et al. (2011). deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data. PLoS Comput. Biol. *7*, e1001138.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41.

Molenaar, J.J., Koster, J., Zwijnenburg, D.A., van Sluis, P., Valentijn, L.J., van der Ploeg, I., Hamdi, M., van Nes, J., Westerman, B.A., van Arkel, J., et al. (2012). Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature *483*, 589–593.

Newton, Y., Novak, A.M., Swatloski, T., McColl, D.C., Chopra, S., Graim, K., Weinstein, A.S., Baertsch, R., Salama, S.R., Ellrott, K., et al. (2017). TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal. Cancer Res. *77*, e111–e114.

Nicorici, D., Satalan, M., Edgren, H., and Kangaspeska, S. (2014). FusionCatcher-a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv. https://doi.org/10.1101/011650.

Padovan-Merhar, O.M., Raman, P., Ostrovnaya, I., Kalletla, K., Rubnitz, K.R., Sanford, E.M., Ali, S.M., Miller, V.A., Mossé, Y.P., Granger, M.P., et al. (2016). Enrichment of Targetable Mutations in the Relapsed Neuroblastoma Genome. PLoS Genet. *12*, e1006501.

Pan, Z., He, H., Tang, L., Bu, Q., Cheng, H., Wang, A., Lyu, J., and You, H. (2017). Loss of heterozygosity on chromosome 16q increases relapse risk in Wilms' tumor: a meta-analysis. Oncotarget *8*, 66467–66475.

Peters, T.L., Kumar, V., Polikepahad, S., Lin, F.Y., Sarabia, S.F., Liang, Y., Wang, W.-L., Lazar, A.J., Doddapaneni, H., Chao, H., et al. (2015). BCOR-CCNB3 fusions are frequent in undifferentiated sarcomas of male children. Mod. Pathol. *28*, 575–586.

Pugh, T.J., Morozova, O., Attiyeh, E.F., Asgharzadeh, S., Wei, J.S., Auclair, D., Carter, S.L., Cibulskis, K., Hanna, M., Kiezun, A., et al. (2013). The genetic landscape of high-risk neuroblastoma. Nat. Genet. *45*, 279–284.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Rausch, T., Jones, D.T.W., Zapatka, M., Stütz, A.M., Zichner, T., Weischenfeldt, J., Jäger, N., Remke, M., Shih, D., Northcott, P.A., et al. (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell *148*, 59–71.

Reid, J.G., Carroll, A., Veeraraghavan, N., Dahdouli, M., Sundquist, A., English, A., Bainbridge, M., White, S., Salerno, W., Buhay, C., et al. (2014). Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. BMC Bioinformatics *15*, 30.

Ribi, S., Baumhoer, D., Lee, K., Edison, Teo, A.S., Madan, B., Zhang, K., Kohlmann, W.K., Yao, F., Lee, W.H., et al. (2015). TP53 intron 1 hotspot rearrangements are specific to sporadic osteosarcoma and can cause Li-Fraumeni syndrome. Oncotarget *6*, 7727–7740.

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., and Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. *17*, 31.

Sano, R., Krytska, K., Larmour, C.E., Raman, P., Martinez, D., Ligon, G.F., Lillquist, J.S., Cucchi, U., Orsini, P., Rizzi, S., et al. (2019). An antibody-drug conjugate directed to the ALK receptor demonstrates efficacy in preclinical models of neuroblastoma. Sci. Transl. Med. *11*, eaau9732.

Schleiermacher, G., Javanmardi, N., Bernard, V., Leroy, Q., Cappo, J., Rio Frio, T., Pierron, G., Lapouble, E., Combaret, V., Speleman, F., et al. (2014). Emergence of new ALK mutations at relapse of neuroblastoma. J. Clin. Oncol. *32*, 2727–2734.

Schramm, A., Köster, J., Assenov, Y., Althoff, K., Peifer, M., Mahlow, E., Odersky, A., Beisser, D., Ernst, C., Henssen, A.G., et al. (2015). Mutational dynamics between primary and relapse neuroblastomas. Nat. Genet. *47*, 872–877.

Scott, R.H., Murray, A., Baskcomb, L., Turnbull, C., Loveday, C., Al-Saadi, R., Williams, R., Breatnach, F., Gerrard, M., Hale, J., et al. (2012). Stratification of Wilms tumor by genetic and epigenetic analysis. Oncotarget *3*, 327–335.

Segers, H., van den Heuvel-Eibrink, M.M., Williams, R.D., van Tinteren, H., Vujanic, G., Pieters, R., Pritchard-Jones, K., and Bown, N.; Children's Cancer and Leukaemia Group and the UK Cancer Cytogenetics Group (2013). Gain of 1q is a marker of poor prognosis in Wilms' tumors. Genes Chromosomes Cancer *52*, 1065–1074.

Sergushichev, A.A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv. https://doi.org/10.1101/060012.

Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E.A., Liu, Y., Weinstock, G.M., Wheeler, D.A., Gibbs, R.A., and Yu, F. (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. *20*, 273–280.

Shern, J.F., Chen, L., Chmielecki, J., Wei, J.S., Patidar, R., Rosenberg, M., Ambrogio, L., Auclair, D., Wang, J., Song, Y.K., et al. (2014). Comprehensive genomic analysis of rhabdomyosarcoma reveals a landscape of alterations affecting a common genetic axis in fusion-positive and fusion-negative tumors. Cancer Discov. *4*, 216–231.

Spreafico, F., Gamba, B., Mariani, L., Collini, P., D'Angelo, P., Pession, A., Di Cataldo, A., Indolfi, P., Nantron, M., Terenziani, M., et al.; AIEOP Wilms Tumor Working Group (2013). Loss of heterozygosity analysis at different chromosome regions in Wilms tumor confirms 1p allelic loss as a marker of worse prognosis: a study from the Italian Association of Pediatric Hematology and Oncology. J. Urol. *189*, 260–266.

Stewart, E., Federico, S.M., Chen, X., Shelat, A.A., Bradley, C., Gordon, B., Karlstrom, A., Twarog, N.R., Clay, M.R., Bahrami, A., et al. (2017). Orthotopic patient-derived xenografts of paediatric solid tumours. Nature *549*, 96–100.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics *31*, 2032–2034.

Tirode, F., Surdez, D., Ma, X., Parker, M., Le Deley, M.C., Bahrami, A., Zhang, Z., Lapouble, E., Grossetête-Lalami, S., Rusch, M., et al.; St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project and the International Cancer Genome Consortium (2014). Genomic landscape of Ewing sarcoma defines an aggressive subtype with co-association of STAG2 and TP53 mutations. Cancer Discov. *4*, 1342–1353.

Townsend, E.C., Murakami, M.A., Christodoulou, A., Christie, A.L., Köster, J., DeSouza, T.A., Morgan, E.A., Kallgren, S.P., Liu, H., Wu, S.-C., et al. (2016). The Public Repository of Xenografts Enables Discovery and Randomized Phase II-like Trials in Mice. Cancer Cell *29*, 574–586.

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

Wang, L., Ni, X., Covington, K.R., Yang, B.Y., Shiu, J., Zhang, X., Xi, L., Meng, Q., Langridge, T., Drummond, J., et al. (2015). Genomic profiling of Sézary syndrome identifies alterations of key T cell signaling and differentiation genes. Nat. Genet. *47*, 1426–1434.

Way, G.P., Allaway, R.J., Bouley, S.J., Fadul, C.E., Sanchez, Y., and Greene, C.S. (2017). A machine learning classifier trained on cancer transcriptomes detects NF1 inactivation signal in glioblastoma. BMC Genomics *18*, 127.

Way, G.P., Sanchez-Vega, F., La, K., Armenia, J., Chatila, W.K., Luna, A., Sander, C., Cherniack, A.D., Mina, M., Ciriello, G., et al.; Cancer Genome

Atlas Research Network (2018). Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Rep. *23*, 172–180.e3.

Whiteford, C.C., Bilke, S., Greer, B.T., Chen, Q., Braunschweig, T.A., Cenacchi, N., Wei, J.S., Smith, M.A., Houghton, P., Morton, C., et al. (2007). Credentialing preclinical pediatric xenograft models using gene expression and tissue microarray analysis. Cancer Res. *67*, 32–40.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871.

Yu, L., Baxter, P.A., Voicu, H., Gurusiddappa, S., Zhao, Y., Adesina, A., Man, T.-K., Shu, Q., Zhang, Y.-J., Zhao, X.-M., et al. (2010). A clinically relevant orthotopic xenograft model of ependymoma that maintains the genomic signature of the primary tumor and preserves cancer stem cells in vivo. Neuro-oncol. *12*, 580–594.

Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S.L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M., et al. (2012). The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature *481*, 157–163.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
KAPA HiFi DNA Polymerase	Kapa Biosystems	KK2612
Agencourt AMPure XP beads	Beckman Coulter	A63882
SeqCap EZ HGSC VCRome Kit v 2.1	Roche	06266380001
TruSeq SBS kit v3 HS	Illumina	FC-401-3001
Oligo(dT)25 Dynabeads	Life Technologies	61002
ERCC spike-in mix #1	Ambion, Life Technologies	4456740
NEBNext RNA First Strand Synthesis Module	New England Biolabs	E7525S
NEBNext Ultra Directional RNA Second Strand Synthesis Module	New England Biolabs	E7550S
Uracil-DNA Glycosylase	New England Biolabs	M0280L
Phusion High-Fidelity PCR Master Mix	New England Biolabs	M0531L
Infinium OmniExpress-24 Kit	Illumina	WG-315-1101
GenePrint24 System for STR Typing	Promega	B1870
Investigator Quantiplex Kit	QIAGEN	387018
PrimeTime Gene Expression 2x qPCR mix	IDT	1055772
Deposited Data		
WES human and mouse BAM files	This paper	dbGAP phs001437
RNA-Seq human and mouse BAM files	This paper	dbGAP phs001437
Intermediate files	This paper	https://figshare.com/projects/Genomic_landscape_of_childhood_cancer_patient-derived_xenograft_models/38147
Processed data – somatic mutations, gene expression, RNA fusions, segmentation files, focal copy number	This paper	https://pedcbioportal.org/login.jsp#summary
Processed data – SNP array-associated analyses files, FPKM matrix, WES MAF files	This paper	Figshare
HapMap 3 draft release 2	International HapMap project	ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/latest_phaseIII_ncbi_b36/plink_format/
Experimental Models: Organisms/Strains		
261 pediatric PDX models	This paper	Table S1
Oligonucleotides		
Human PTGER2 qPCR FWD primer, 5′-GCT GCTTCTCATTGTCTCGG-3′	IDT	custom
Human PTGER2 qPCR REV primer, 5′-GC CAGGAGAATGAGGTGGTC-3′	IDT	custom
Human pTGER2 qPCR probe, 5′-FAM-CAG TGTCATTCTCAACCTCATCCGCA-IOWA-BLACK-3′	IDT	custom
Mouse pTGER2 qPCR FWD primer, 5′-AC ATCAGCGTTATCCTCAACC-3′	IDT	custom
Mouse pTGER2 qPCR REV primer, 5′-GC TACTGCCAGACAATCCG-3′	IDT	custom
Mouse pTGER2 qPCR probe, 5′-TXRED-TCATTCGCATGCACCGTCGGA- IOWA-BLACK-3′	IDT	custom

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
FusionCatcher 0.99.7b	Nicorici et al., 2014	https://github.com/ndaniel/fusioncatcher
STAR-Fusion 1.1.0	Haas et al., 2017	https://github.com/STAR-Fusion
STAR 2.5.2b	Dobin et al., 2013	https://github.com/alexdobin/STAR
RSEM 1.2.28	Li and Dewey, 2011	https://github.com/deweylab/RSEM
TumorMap 1.0	Newton et al., 2017	https://tumormap.ucsc.edu/
Stan 2.16.0	Carpenter et al., 2017	https://github.com/stan-dev/cmdstan
Fgsea 1.5.1	Sergushichev, 2016	https://bioconductor.org/packages/release/bioc/html/fgsea.html
Pandas 0.23.0	McKinney, 2010	https://pandas.pydata.org/
R, various	R Core Team	http://www.R-project.org/
Python 3.6.5	Python Core Team	https://www.python.org/
Jupyter 1.0.0	Kluyver et al., 2016	https://jupyter.org/index.html
Seaborn 0.8.1	Seaborn Core Team	https://seaborn.pydata.org/
Maftools 2.0.15	Mayakonda et al., 2018	https://github.com/PoisonAlien/maftools
R 3.4.3	R Core Team	http://www.R-project.org/
ComplexHeatmap 2.1.0	Gu et al., 2016	https://www.bioconductor.org/packages/3.7/bioc/html/ComplexHeatmap.html
deconstructSigs 1.8.0	Rosenthal et al., 2016	https://github.com/raerose01/deconstructSigs
Nexus 8.0	Biodiscovery	https://www.biodiscovery.com/
GISTIC 2.0.23	Mermel et al., 2011	https://www.broadinstitute.org/node/358411
MutSigCV 1.3.01	Lawrence et al., 2013	http://software.broadinstitute.org/cancer/software/genepattern/modules/docs/MutSigCV
HGSC Mercury 3.2	Reid et al., 2014	https://www.hgsc.bcm.edu/software/mercury
BWA 0.7.17-r1188	Li and Durbin, 2010	http://bio-bwa.sourceforge.net/
GATK 3.8.1	McKenna et al., 2010	https://www.broadinstitute.org/gatk/
PLINK 1.9	Chang et al., 2015	https://www.cog-genomics.org/plink/1.9/
PLINK 1.07	Purcell et al., 2007	http://zzz.bwh.harvard.edu/plink/
Samtools 1.9	Li et al., 2009	http://samtools.sourceforge.net/
Sambamba 0.6.6	Tarasov et al., 2015	https://github.com/biod/sambamba
Picard 2.18.14-0	2018	https://github.com/broadinstitute/picard
Cufflinks 2.2.1	Trapnell et al., 2010	https://github.com/cole-trapnell-lab/cufflinks
RNA-SeQC 1.1.8	Deluca et al., 2012	https://github.com/broadinstitute/rnaseqc
AlignStats 0.3	BCM-HGSC	https://github.com/jfarek/alignstats
SOAPFuse 1.26	Jia et al., 2013	https://sourceforge.net/projects/soapfuse/
HTSeq 0.9.1	Anders et al., 2015	https://github.com/simon-anders/htseq
Pindel 0.2.5b5	Ye et al., 2009	https://github.com/genome/pindel
deFuse 0.7.0	McPherson et al., 2011	https://github.com/amcpherson/defuse
Bamutil 1.0.14	Jun et al., 2015	https://github.com/statgen/bamUtil
Trinity 2.5.1	Grabherr et al., 2011	https://github.com/trinityrnaseq/trinityrnaseq
Strelka 2.9.2	Kim et al., 2018b	https://github.com/Illumina/strelka
NGSCheckmate 1.0	Lee et al., 2017	https://github.com/parklab/NGSCheckMate
Other		
TARGET pediatric tumors RNA-sequencing dataset	The TARGET Consortium	https://ocg.cancer.gov/programs/target/data-matrix
GTEx normal tissues RNA-sequencing dataset	The GTEx Consortium. 2013	http://www.gtexportal.org/home/index.html
Exome Aggregation Consortium 0.3.1	Lek et al., 2016	http://exac.broadinstitute.org/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
The International Genome Sample Resource and 1000 genomes project	Birney and Soranzo, 2015	https://www.internationalgenome.org/
NHBLI Exome Sequencing Project (ESP)	Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [date (month, year) accessed].	http://evs.gs.washington.edu/EVS/

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to, and will be fulfilled by, the Lead Contact, John M. Maris (maris@email.chop.edu). All PDX models are available through the Pediatric Preclinical Testing Consortium with a completed Material Transfer Agreement.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Patient-Derived Xenograft Generation and Harvesting

Patient-derived xenograft models from the Pediatric Preclinical Testing Program (PPTP) were generated as described (Houghton et al., 2002, 2007; Whiteford et al., 2007). Briefly, for solid tumors, C.B-Igh-1b/IcrTac-Prkdcscid (Taconic Farms, Germantown NY), were subcutaneously flank-engrafted into male or female mice (Table S1) and passaged once tumors reached 200 mm^3. For CNS tumors, patient tumors were stereotactically-transplanted into anesthetized (50 mg/kg sodium pentobarbital) RAG2, NOD.129S7(B6)-Rag1tm1Mom/J, or RAG1tm1Mom/J mouse brains in the diagnosis-specific orthotopic locations noted in Table S1 (Yu et al., 2010). PDX tumor cells (1×10^5) were suspended in 2 ul of culture media and slowly injected through a burr hole using a 10 ul, 26 gauge syringe into the brain region of interest. Once moribund, or displaying neurological deficit symptoms, mice were euthanized and whole murine brains containing visible tumors were aseptically removed and transferred to the tissue culture laboratory. Tumors were microscopically dissected from surrounding brain tissue, mechanically dissociated into cell suspensions, and filtered. Single tumor cells were subsequently injected into the brains of SCID mice as described above. Sub-transplantation process was repeated to complete a total of five tumor passages. All animal experiments were conducted according to an Institutional Animal Care and Use Committee-approved protocol. All leukemia animal experimentation was approved by the Animal Care and Ethics Committee, UNSW Sydney (Sydney, Australia). Experiments used continuous PDXs established previously in 20-25 g female non-obese diabetic/severe combined immuno-deficient (NOD.CB17-Prkdc^{scid}/SzJ, NOD/SCID) or NOD/SCID/interleukin-2 receptor γ–negative (NOD.Cg-Prkdc^{scid} Il2rg^{tm1Wjl}/SzJ, NSG) mice. Leukemia cells were inoculated intravenously into 6-8 week-old NOD/SCID or NSG mice (Australian BioResources, Moss Vale, NSW, Australia) and leukemia burden monitored via enumeration of human CD45$^+$ (%huCD45$^+$) cells versus total CD45$^+$ leukocytes (human plus mouse) in the peripheral blood (PB) and tissues, as reported (Liem et al., 2004; Lock et al., 2002). The continuation of xenograft lines was accomplished through harvesting human leukemia cells from the spleens of the engrafted mice. Harvesting required more than 3×10^8 leukemia cells per spleen, at 85% purity. Additional details per model including sex, age, and mass are included in Table S1.

METHOD DETAILS

Nucleic Acid Extractions and Quality Control

PDX samples were submitted from Children's Cancer Institute, Children's Hospital of Philadelphia, Greehey Children's Cancer Research Institute, and Montefiore Medical Center to the Nationwide Children's Hospital Biospecimen Core Resource at −190°C using an MVE cryoshipper. Cytospins and H&E frozen sections were prepared from leukemia and solid tissue PDX specimens, respectively. Slides were assessed by board-certified pathologists to determine blast percentage in leukemia PDX samples, and percent tumor nuclei and necrosis of the solid PDX samples. DNA and RNA were co-extracted from the PDXs using a modification of the DNA/RNA AllPrep kit (QIAGEN). The flow-through from the QIAGEN DNA column was processed using a mirVana miRNA Isolation Kit (Ambion). DNA was quantified by PicoGreen assay and RNA samples were quantified by measuring Abs_{260} with a UV spectrophotometer. DNA specimens were resolved by 1% agarose gel electrophoresis to confirm high molecular weight fragments. RNA was analyzed via the RNA6000 Nano assay (Agilent) for determination of an RNA Integrity Number (RIN). The PPTC study committee reviewed the pathology and molecular QC data and selected DNA and RNA aliquots for sequencing.

Short Tandem Repeat (STR) Profiling

Each tumor DNA sample was subjected to STR profiling performed by Guardian Forensic Sciences. DNA samples were quantified using QIAGEN Investigator Quantiplex Kit (Cat# 387018) on a QIAGEN RotorGene Q instrument. The GenePrint24 System for STR

profiling (Promega, Cat#B1870) was used to amplify 0.05 ng of template DNA in a 12.5 μL volume using the following conditions: 96°C for 1 minute, 27 cycles of {94°C for 10 s, 59°C for 1 minute, 72°C for 30 s}, 60°C for 10 minutes using the RotorGene Q instrument. Samples were injected into the Applied Biosystems ABI 310 Genetic Analyzer and profiles were interpreted by forensic biologists. Only those samples deemed not misidentified and free of contamination were used in this study.

Biochemical Measurement of Human DNA Content in PDX Tumors

To determine the composition of human and mouse DNA within PDX tumors, PDX DNA samples were amplified using modified version of the published *pTGER2* (*prostaglandin E receptor 2*) qPCR assay (Alcoser et al., 2011). Depending upon sample availability, 2-20 ng of PDX tumor DNA were added to 500 nM each human- and mouse-specific forward primers, reverse primers, probes (sequences in resource document) and 1X IDT PrimeTime Gene Expression 2X Mastermix (Integrated DNA Technologies) in a total of 20 uL. Reactions were thermalcycled at 95°C for 8 min and 42 cycles of {95°C for 15 s, 64°C for 1 min}. Five-point standard curves were performed using a mixture of CHLA-90 and COG-N-603 neuroblastoma cell lines as human-specific template and pooled liver/spleen/muscle DNA from a naive NU/NU mouse as the mouse-specific template to confirm each primer efficiency was between 90%–110%. The DNA equivalent of one diploid copy of either mouse or human template was run as a reference template. Three technical replicates were performed for each standard and sample. Average C_T values of the reference DNA samples were used as "ground truth" C_T values for one DNA copy. To estimate relative copy number, $2^{-\Delta CT}$ values were calculated for each unknown for each species: $2^{-\Delta CT} = 2^{-(CT\ of\ Unknown - CT\ of\ Refererence)}$. To estimate percent human content, the following equation was used: $\%Human\ content = (Relative\ human\ genome\ copies \times 100 / Relative\ mouse\ genome\ copies)$.

Additional Quality Control for Cross-Contamination and Mis-Identification

Common germline SNP distributions (allele frequency > 0.005 in any one of the three databases: Exome Aggregation Consortium, 1000 genomes, or the NHBLI Exome Sequencing Project) were plotted for each model and visually inspected for a negatively skewed distribution to assess DNA cross-contamination in WES data. To identify potential mis-identification, RNA variant calling was performed and variant allele frequencies correlated between WES and RNA. Models whose variants did not correlate were deemed mis-identified and removed (STAR Methods). For remaining models, NGSCheckmate was performed between WES and RNA data. All models except for ICb-2002EPN had correlation values of $\geq$ 0.61 at depths of $\geq$ 10, deeming these models matched as recommended by Lee et al. (2017). ICb-2002EPN had a borderline correlation of 0.6025 at a depth of 14.51, but deemed matched from WES-RNA mutation correlations. Within this cohort, five pairs of models were derived from tissue at phase of therapy (Table S1). Thus, as additional QC, we correlated somatic mutation allele frequencies between each pair and found high concordance of mutation frequencies (data on Figshare, STAR Methods), confirming biological reproducibility of creating PDX models within a center. Mutation variation is summarized per model in Table S3.

Whole Exome Sequencing

Illumina paired-end pre-capture libraries were constructed from PDX DNA samples according to the manufacturer's protocol (Illumina Multiplexing_SamplePrep_Guide_1005361_D) modified as described in the BCM-HGSC Illumina Barcoded Paired-End Capture Library Preparation protocol. The complete protocol including oligonucleotide sequences used as adaptors and blockers are accessible from the HGSC website https://www.hgsc.bcm.edu/sites/default/files/documents/Protocol-Illumina_Whole_Exome_Sequencing_Library_Preparation-KAPA_Version_BCM-HGSC_RD_03-20-2014.pdf. The DNA sequence production is briefly described below.

Library Preparation

500 ng (or 250 ng if sample quantity was limiting) of DNA in 50ul volume were sheared into fragments to an average size of 200-300 bp in a Covaris plate with E220 system (Covaris, Inc. Woburn, MA) followed by end-repair, A-tailing and ligation of the Illumina multiplexing PE adaptors. Pre-capture Ligation Mediated-PCR (LM-PCR) was performed for 6-8 cycles using the Library Amplification Readymix containing KAPA HiFi DNA Polymerase (Kapa Biosystems, Inc.). Universal primer LM-PCR Primer 1.0 and LM-PCR Primer 2.0 were used to amplify the ligated products. Reaction products were purified using 1.8X Agencourt AMPure XP beads (Beckman Coulter) after each enzymatic reaction. Following the final 1.2X Agencourt XP beads purification, quantification and size distribution of the pre-capture LM-PCR product was determined using Fragment Analyzer capillary electrophoresis system (Advanced Analytical Technologies, Inc.).

Capture Enrichment

Four pre-capture libraries were pooled together (~750 ng/sample, 3 ug/pool) and then hybridized in solution to the HGSC VCRome 2.1 design1 (Bainbridge et al., 2011) according to the manufacturer's protocol NimbleGen SeqCap EZ Exome Library SR User's Guide (Version 2.2) with minor revisions. Probes for exome coverage across > 3,500 clinically relevant genes that are previously < 20X (~2.72Mb) is supplemented into the VCRome 2.1 probe. Human COT1 DNA was added into the hybridization to block repetitive genomic sequences. Blocking oligonucleotides from Sigma (individually sequence specifically synthesized) or xGen Universal Blocking oligonucleotides (Integrated DNA Technologies) were added into the hybridization to block the adaptor sequences. Hybridization was carried out at 560C for ~16h. Post-capture LM-PCR amplification was performed using the Library Amplification Readymix containing KAPA HiFi DNA Polymerase (Kapa Biosystems, Inc.) with 12 cycles of amplification. After the final AMPure XP bead purification, quantity and size of the capture library was analyzed using the Agilent Bioanalyzer 2100 DNA Chip 7500. The efficiency

of the capture was evaluated by performing a qPCR-based quality check on the four standard NimbleGen internal controls. Successful enrichment of the capture libraries was estimated to range from a 6 to 9 of ΔC_T value over the non-enriched samples.

DNA Sequencing

Library templates were prepared for sequencing using Illumina's cBot cluster generation system with TruSeq PE Cluster Generation Kits (Illumina) according to the manufacturer's protocol. Briefly, these libraries were denatured with sodium hydroxide and diluted to 6-9 pM in hybridization buffer in order to achieve a load density of ~800K clusters/mm^2. Each library pool was loaded in a single lane of a HiSeq flow cell, and each lane was spiked with 1% phiX control library for run quality control. The sample libraries then underwent bridge amplification to form clonal clusters, followed by hybridization with the sequencing primer. Sequencing runs were performed in paired-end mode using the Illumina HiSeq 2000 platform. Using the TruSeq SBS Kits (Illumina), sequencing-by-synthesis reactions were extended for 101 cycles from each end, with an additional 7 cycles for the index read. With sequencing yields averaging 12.1 Gb per sample, samples achieved an average of 97.64% of the targeted exome bases covered to a depth of 20X or greater.

Primary Data Analysis

Initial sequence analysis was performed using the HGSC Mercury analysis pipeline (Challis et al., 2012; Reid et al., 2014). In summary, the. bcl files produced on-instrument were first transferred into the HGSC analysis infrastructure by the HiSeq Real-time Analysis module. Mercury then ran the vendor's primary analysis software (CASAVA) to de-multiplex pooled samples and generate sequence reads and base-call confidence values (qualities), followed by the mapping of reads to the GRCh37 Human reference genome (https://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/) using the Burrows-Wheeler aligner (Li and Durbin, 2010). The resulting BAM (binary alignment/map) file underwent quality recalibration using GATK, and where necessary the merging of separate sequence-event BAMs into a single sample-level BAM. BAM sorting, duplicate read marking, and realignment to improve in/del discovery all occur at this step. Next, Atlas-SNP and Atlas-indel from the Atlas2 suite (Shen et al., 2010) were used to call variants and produce a variant call file (VCF). Finally, annotation data was added to the VCF using a suite of annotation tools "Cassandra" (https://www.hgsc.bcm.edu/software/cassandra) that brings together frequency, function, and other relevant information using AnnoVar with UCSC and RefSeq gene models, as well as a host of other internal and external data resources.

SNP Array Assay

In brief, 200 ng of genomic DNA were denatured with NaOH, followed by isothermal whole genome amplification at 37°C for 20-24 hours. The amplified DNA was enzymatically fragmented and hybridized to the BeadChip for 16-24 hours at 48°C (24 samples were processed in parallel for each BeadChip). After a series of washing steps to remove unhybridized and non-specifically hybridized DNA fragments, allele-specific single-base extension reactions were performed to incorporate labeled nucleotides into the bead-bound primers. A multi-layer staining process was conducted to amplify signals from the labeled extended primers, and then the coated beads were imaged with the Illumina iScan system.

Chip types used were humanomniexpress-24-v1-1-a.bpm and InfiniumOmniExpress-24v1-2_A1.bpm.

Whole Transcriptome Sequencing

Whole-transcriptome RNA sequencing (RNA-seq) was performed using total RNA extracted as described above. Strand-specific, poly-A+ RNA-seq libraries for sequencing on the Illumina platform were prepared using manufacturer guidelines with minor modifications described herein (Peters et al., 2015; Wang et al., 2015). RNA Integrity was confirmed (RIN > 7.0) on a Bioanalyzer (Agilent). Briefly, poly-A+ mRNA was extracted from 1 μg total RNA using Oligo(dT)25 Dynabeads (Life Technologies), to which 4 μL of 1:100 dilution of the ERCC spike-in mix 1 (Ambion, Life technologies) was already added (Baker et al., 2005). There are a total of 92 polyadenylated transcripts in this mix that are used to monitor sample and process consistency. mRNA is then fragmented by heat at 94°C for 15 minutes or less depending on sample RIN. First strand cDNA was synthesized using NEBNext RNA First Strand Synthesis Module (New England BioLabs) and during second strand cDNA synthesis, dNTP mix containing dUTP was used to introduce strand-specificity with NEBNext Ultra Directional RNA Second Strand Synthesis Module (New England BioLabs). For Illumina paired-end library construction, the resultant cDNA is processed through end-repair and A-tailing, ligated with Illumina PE adapters, and then digested with 10 units of Uracil-DNA Glycosylase (New England BioLabs). Libraries are prepared on the Beckman BioMek FXp robots and amplification of the libraries was performed for 13 PCR cycles using the Phusion High-Fidelity PCR Master Mix (New England BioLabs); 6-bp molecular barcodes that were also incorporated during this step. Libraries were purified with Agencourt AMPure XP beads (Beckman Coulter) after each enzymatic reaction, and after PCR amplification, and were quantified using Fragment Analyzer electrophoresis system. Libraries were pooled in equimolar amounts (4 libraries/pool). Library templates were prepared and sequenced exactly as described above for DNA Sequencing. Sequencing runs generated approximately 300-400 million successful reads on each lane of a flow cell, yielding 75-100M reads per sample.

QUANTIFICATION AND STATISTICAL ANALYSIS

Mouse Read Subtraction from WES Sequencing Data

Raw fastq files (n = 240) from Whole exome sequencing data were aligned to a combined hybrid genome of human hg19 and mouse mm10 genomes using the *Burrows-Wheeler transformation algorithm (BWA v0.7.17-r1188)*. Reads overlapping specifically to either the human or mouse genome were extracted and separated in corresponding human and mouse bam files using Samtools v1.9. The

mouse subtracted bam files containing reads specific to human genome were then sorted by name and only paired reads were kept using the Samtools parameter *-f 1*. Following this, duplicated reads were marked using Sambamba v0.6.6. The resulting bam files were then used as input for local realignment around indels using IndelRealigner and base quality score recalibration using BaseRecalibrator utilities from GATK v3.8.1.

Whole Exome Mutation Analysis

Many of these PDX models have been established decades ago, thus matched primary and/or normal tissue either were not collected or is not currently available. To filter common germline variation from these tumor models, we used a panel of 809 normal samples supplied from TCGA WBC tissue to generate consensus germline variant calls. Rare germline variation was retained and defined as < 0.005 minor allele frequency in any one of the three databases: Exome Aggregation Consortium (ExAC) (Lek et al., 2016), 1000 genomes, or the NHBLI Exome Sequencing Project (ESP). Filtered variants also present in COSMIC were scavenged back. We performed MutSigCV (Lawrence et al., 2013) analysis on the entire cohort to identify and remove false positive variants. With the exception of known oncogenes and tumor suppressors, novel significantly mutated genes (SMGs) common across all histologies should be rare. We manually inspected the top 100 SMGs and found that most novel genes harbored a high number of private mutations and thus were not removed. Other novel variants were false positives due to germline inclusion or sequencing/mapping errors (data on FigShare, link below). Data were thus split into germline MAF and somatic MAF files, the latter of which retained private variants.

Tumor Mutation Burden Analysis

Using the maftools R package (Mayakonda et al., 2018), total number of mutations per variant type per model were calculated. We defined tumor mutation burden using only mononucleotide substitutions resulting in amino acid changes: $(\Sigma(\textit{somatic nonsynonymous} + \textit{missense variants})/45.1\ \textit{Mb})$. The denominator was the 45.1 Mb size of the Roche Nimblegen VCRome v. 2.1 capture panel.

ATRX Deletion Analysis

The *ATRX* locus on chromosome X contains too few probes in OmniExpress arrays to accurately assess deletion, even in cases of known sex. Thus, from WES bam files, total read base counts for ATRX exons were calculated using Samtools v1.9 bedcov utility and total library size was calculated using Samtools v1.9 flagstat utility. To convert exon read counts to Fragments per kilobase per million reads (FPKM), the library sizes were first transformed to per million scaling factors. Following this, raw read counts of each exon were normalized using the per million scaling factors and the corresponding exon length.

Mutational Signatures Analysis

The deconstructSigs R package with the COSMIC 30 signature reference was used. We ran this workflow on models with $\geq$ 50 total somatic mutations. We chose a cosine similarity value cutoff at 0.1 and plotted the proportion of signatures in each model as a stacked barplot.

Classifier Analysis

We applied models derived from three supervised machine learning algorithms to all PDX models with available RNA-Seq data (n = 244). The models were previously trained on RNaseq, copy number, and mutation data across 33 different adult cancer-types from The Cancer Genome Atlas PanCanAtlas project (Cancer Genome Atlas Research Network et al., 2013). Briefly, the algorithm was an elastic net penalized logistic regression classifier that took FPKM and z-score normalized RNaseq data as input and, in three independent classifiers, was trained to predict Ras pathway activation, *NF1* inactivation, and *TP53* inactivation using mutation and copy number alteration status of corresponding samples. The Ras pathway and *NF1* classifiers and the overall method were described in more detail in Way et al. (2018). The application and validation of the *TP53* classifier was described in Knijnenburg et al. (2018).

To assess performance of the TCGA trained classifiers applied to the PDX data, we used orthogonal evidence of gene alterations in each PDX sample. Specifically, we used samples with observed missense, nonsense, frameshift, and splice site mutations in *ALK*, *BRAF*, *CIC*, *DMD*, *HRAS*, *KRAS*, *NF1*, *NRAS*, *PTPN11*, and *SOS1* as samples with possible Ras pathway activation. We used samples with only non-silent *NF1* mutations for the *NF1* classifier, and samples with deleterious *TP53* mutations, copy number deletions, and fusions for the *TP53* classifier. We assessed model performance using receiver operating characteristic (ROC) and precision recall (PR) curves using these samples as the positive set and all others as the negative set. We also applied the classifiers to shuffled PDX gene expression matrices and compared performance to the real data to assess potential model bias. The reproducible analysis pipeline can be viewed at https://github.com/marislab/pdx-classification and the software is archived on Zenodo at https://doi.org/10.5281/Zenodo.1475249.

mRNA Gene Expression Analysis

Raw fastq files (n = 244) from RNA-sequencing data were aligned to a combined hybrid genome of human hg19 and mouse mm10 genomes using the STAR aligner v2.5.3a. Reads overlapping specifically to either the human or mouse genome were extracted and

separated in corresponding human and mouse bam files using Samtools v1.9. The mouse subtracted bam files containing reads specific to human genome were then sorted by name and only paired reads were kept using the Samtools parameter *-f 1*. Following this, duplicated reads were marked using Sambamba v0.6.6. The resulting bam files were used to extract and separate reads into paired-ended fastq files using the *SamToFastq* utility of Picard v2.18.14-0. The resulting paired-ended fastq files obtained after mouse subtraction were re-aligned to human genome hg19 using STAR aligner and marked for duplicate reads using Picard *MarkDuplicates*. Gene expression was quantified in terms of Fragments Per Kilobase of transcript per Million mapped reads (FPKM) using HTSeq v0.9.1 and Cufflinks v2.2.1. We also processed RNA-sequencing patient data from TARGET (ALL, n = 533; AML, n = 364; NBL, n = 169; RT, n = 70; OS, n = 87; WT, n = 136) and PPTC PDX data (n = 244) using STAR alignment and RSEM normalization using hg38 as reference genome and Gencode v23 gene annotation to get transcript per million (TPM) expression values. For PPTC PDX data, human bam files generated from the mouse subtraction pipeline were used in order to generate input fastq files.

mRNA Variant Calling, Filtering, and Comparison to DNA Variants

Variant calling for RNA-seq samples was performed with Strelka v2.9.2 germline indels calling pipeline using hg19 primary assembly reference fasta and default parameters. VCFs were converted to MAF and variants were filtered for those that passed VEP and were non-silent (! = Silent or Intron). Variant allele frequencies for all non-silent, VEP-passed RNA variants were calculated. For each model on which both WES and RNA-Seq were performed, WES variants with RNA evidence were matched in the DNA MAF and VAF correlations were plotted and are stored in the QC folder of the FigShare project: https://figshare.com/projects/Genomic_profiling_of_childhood_tumor_patient-derived_xenograft_models_to_enable_rational_clinical_trial_design/38147.

Copy Number Analysis

SNP arrays were processed at the HGSC using the Illumina Infinium HTS Assay according to the manufacturer's guidelines. Human OmniExpress arrays (Illumina, catalog No. WG-315-1101) were used, interrogating 741 thousand SNP loci with a MAF detection limit of 5%. SNP calls were collected using Illumina's GenomeStudio software (version 1.0/2.0) in which standard SNP clustering and genotyping were performed with the default settings recommended by the manufacturer. Data from samples that met a minimum SNP call rate of 0.9 were considered passing and were included in subsequent analyses. Output files from Genome Studio containing BAF and LRR were used as input for Nexus 8.0. Quadratic systematic correction was performed using a custom file (Figshare repository, below) containing common snp probes from the two chip types. The significance threshold was reduced to 1×10^{-8} to reduce background noise. Segmentation was performed using Nexus's SNPRANK algorithm. To extract segments, gain was set to 0 and loss to -1×10^{-11}. The output table was reformatted to segmentation file format for input to GISTIC2.0, which was used to calculate broad and focal, hemizygous gene-level copy number events. Relevant arm and band level alterations were used in oncoprints. Since normal DNA was not available for paired analyses, sex chromosomes were removed. Focal homozygous deletions and amplifications were annotated using the segmentation file created post-Nexus analysis. A cutoff of LRR > = (0.538578182) was used for amplifications and > = ($-$1.739) for deletions. Cutoffs were determined by assessing histogram splits for MYCN amplification, SMARCB1 deletion, and CDKN2A/B deletions. Homozygous deletions remained only if mRNA FPKM was < 5 or if RNA-Seq for a sample was not available. Manual inspections were performed to confirm alterations for *SMARCB1, TP53, WT1, MYCN, C19MC, CDKN2A/B* and edited when necessary (see code).

Breakpoint Analysis

We defined breakpoint regions as regions with 10% copy number change between adjacent segments. These were tabulated per autosome per model and plotted by histology in Figure S4C. To defined regions of high breakpoint density (HBD) as $\geq$ 10 breakpoints per chromosome (Figure S4D; Table S3).

Ethnicity Inference

Approximate genomic ancestries for each PDX model were inferred through principal component analysis of SNP array genotypes. Illumina-designated plus-strand genotypes were exported from GenomeStudio and processed using PLINK 1.9. Sex chromosomes and SNPs with minor allele frequency < 1%, call rate < 90%, or a deviation from Hardy-Weinberg equilibrium surpassing p = 0.00005 were excluded. The PDX dataset was then merged with HapMap 3 (draft release 2), restricting to only the intersecting SNPs. This set was pruned to remove highly correlated SNPs using a window size of 50 variants, step size of 5 variants, and pairwise r^2 threshold of 0.1. The 39,544 remaining SNPs were used to calculate the top 20 principal components. Approximate ethnicities were inferred using the first two components. Individuals were classified into four broad population groups: European (including HapMap CEU and TSI population samples), African (ASW, LWK, MKK, and YRI), East Asian (CHB, CHD, and JPT), and South Asian or Hispanic (GIH and MXL).

Fusion Transcript Analysis

We used four different fusion callers: STAR-Fusion v1.1.0, FusionCatcher v0.99.7b, deFuse and SOAPFuse on RNA-sequencing data of the PDX models (n = 244). A total of 50,796 unique fusions were predicted with the following breakdown: STAR-Fusion (n = 9,496),

FusionCatcher (n = 3,822), deFuse (n = 30,393), and SOAPFuse (n = 7,085). To reduce the number of false positives, we used two parallel approaches: first to keep all fusions predicted as in-frame and second to keep all fusions where the 5′ or 3′ gene fuses promiscuously with multiple partners within the same histology. To filter out unreliable predictions, we further filtered the in-frame fusions by keeping fusions that were recurrently predicted in two or more models within the sample histology or fusions that were supported by at least two fusion callers. We removed any fusions where expression of both genes in the gene pair was found to be < 1 TPM value across all models or it was not reported by the gene quantification algorithm. We then combined the lists from the two approaches discussed above and filtered out any fusions that were predicted in more than one histology. To remove spurious fusions, we filtered all fusions annotated as ''read-through'' as a result of fusions between adjacent or neighboring genes. We further removed fusions identified in non-cancer tissues and cells as per GTEx in order to remove chimeric RNA that is normally found in healthy tissue. Next, we scavenged and annotated fusions that have been identified as ''driver'' fusions in literature and fusions that were validated using cytogenetics. Finally, we annotated the gene fusion partners with oncogenes from COSMIC, kinases from Kinase.com, and transcription factors from AnimalTFDB to identify any oncogenic potential and functional relevance.

RNA Expression Clustering and Pathway Analyses

The UCSC TumorMap analysis was used to visualize clusters of expression profiles across PDX histologies (Newton et al., 2017). The expression values were transformed into $\log_2$(TPM + 1) space. We removed genes where more than 80% of the samples had no measurable expression and we applied a variance filter to remove the 20% least varying genes. This generated a gene by sample matrix containing 28,482 genes and 244 PDX samples. The expression values and PDX annotations were uploaded to the TumorMap portal for analysis. A Bayesian hierarchical model was used to infer differences in expression across PDX histologies. We used a hierarchical modeling strategy to leverage similarities across related tissues and to improve inferences for histologies with small sample sizes (Ji and Liu, 2010). The hierarchical model was implemented using the Stan statistical programming language (Carpenter et al., 2017).

We inferred the biological function of histology-specific expression by ranking the expression differences for each histology and performing gene-set enrichment analysis (GSEA). GSEA was performed using the fgsea software (Sergushichev, 2016). Statistically significant enrichment was defined as having an adjusted p value less than 0.05 and a normalized enrichment score greater than 2.0. Statistically insignificant enrichment scores were set to zero for heatmap visualization. The normalized enrichment scores were visualized using the seaborn clustermap software for tissue database scores and R for Hallmark pathway scores.

Pediatric cBioPortal Data Processing

All processed data: RNA-sequencing expression values (FPKM and Z-score), RNA fusions, mutation calls in Mutation Annotation Format (MAF), segmentation, and focal copy number values were formatted using the current cBioPortal v1.2.2 file format documentation.

DATA AND CODE AVAILABILITY

Raw Data Availability

Mouse and human separated DNA and RNA BAM files have been deposited into dbGAP under accession number phs001437.v1.p1.

INTERMEDIATE PROCESSED DATA AVAILABILITY

Variant files, SNP array files, contamination assessment files:
https://figshare.com/projects/Genomic_landscape_of_childhood_cancer_patient-derived_xenograft_models/38147

Processed Data Availability

WES mutations, mRNA expression, RNA fusions, segmentation, and gene copy number has been deposited into the publicly-available pediatric cBioportal at: https://pedcbioportal.org/study?id=pptc#summary

Code Created or Modified for Analysis in This Paper Have Been Deposited in GitHub

PDX mouse subtraction: https://github.com/marislab/pdx-mouse-subtraction
NGSCheckmate analysis: https://github.com/d3b-center/ngs_checkmate_wf
Correlation analyses: https://github.com/marislab/create-pptc-pdx-corplots
PDX pie chart (Figure 1): https://github.com/marislab/create-pptc-pdx-pie
Oncoprint generation (Figures 2 and 3): https://github.com/marislab/create-pptc-pdx-oncoprints
Medulloblastoma classification (Figure 2): https://github.com/PichaiRaman/MedulloClassifier
Tumor mutation burden (Figures 3 and S3): https://github.com/marislab/pptc-pdx-tmb

OPEN ACCESS
CellPress

Gene classification (Figure 4): https://github.com/marislab/pdx-classification
Classifier analysis (Figures 4 and S5): https://github.com/marislab/pptc-pdx-classifier-analysis
RNA clustering and heatmaps (Figure 5): https://github.com/marislab/pptc-pdx-RNA-Seq-clustering
RNA fusion analysis (Figure 5): https://github.com/marislab/pptx-pdx-fusion-analysis
Ethnicity inference (Figure S2): https://github.com/marislab/pptc-pdx-ethnicity-inference
Mutational signatures (Figure S3): https://github.com/marislab/pptc-pdx-mut-sigs
Copy number, breakpoint, and SV (*ATRX* deletion) analysis (Figure S4): https://github.com/marislab/pptc-pdx-copy-number-and-SVs

Cell Reports, Volume *29*

Supplemental Information

Genomic Profiling of Childhood Tumor Patient-Derived Xenograft Models to Enable Rational Clinical Trial Design

Jo Lynne Rokita, Komal S. Rathi, Maria F. Cardenas, Kristen A. Upton, Joy Jayaseelan, Katherine L. Cross, Jacob Pfeil, Laura E. Egolf, Gregory P. Way, Alvin Farrel, Nathan M. Kendsersky, Khushbu Patel, Krutika S. Gaonkar, Apexa Modi, Esther R. Berko, Gonzalo Lopez, Zalman Vaksman, Chelsea Mayoh, Jonas Nance, Kristyn McCoy, Michelle Haber, Kathryn Evans, Hannah McCalmont, Katerina Bendak, Julia W. Böhm, Glenn M. Marshall, Vanessa Tyrrell, Karthik Kalletla, Frank K. Braun, Lin Qi, Yunchen Du, Huiyuan Zhang, Holly B. Lindsay, Sibo Zhao, Jack Shu, Patricia Baxter, Christopher Morton, Dias Kurmashev, Siyuan Zheng, Yidong Chen, Jay Bowen, Anthony C. Bryan, Kristen M. Leraas, Sara E. Coppens, HarshaVardhan Doddapaneni, Zeineen Momin, Wendong Zhang, Gregory I. Sacks, Lori S. Hart, Kateryna Krytska, Yael P. Mosse, Gregory J. Gatto, Yolanda Sanchez, Casey S. Greene, Sharon J. Diskin, Olena Morozova Vaske, David Haussler, Julie M. Gastier-Foster, E. Anders Kolb, Richard Gorlick, Xiao-Nan Li, C. Patrick Reynolds, Raushan T. Kurmasheva, Peter J. Houghton, Malcolm A. Smith, Richard B. Lock, Pichai Raman, David A. Wheeler, and John M. Maris

Figure S1, related to Figures 1-5

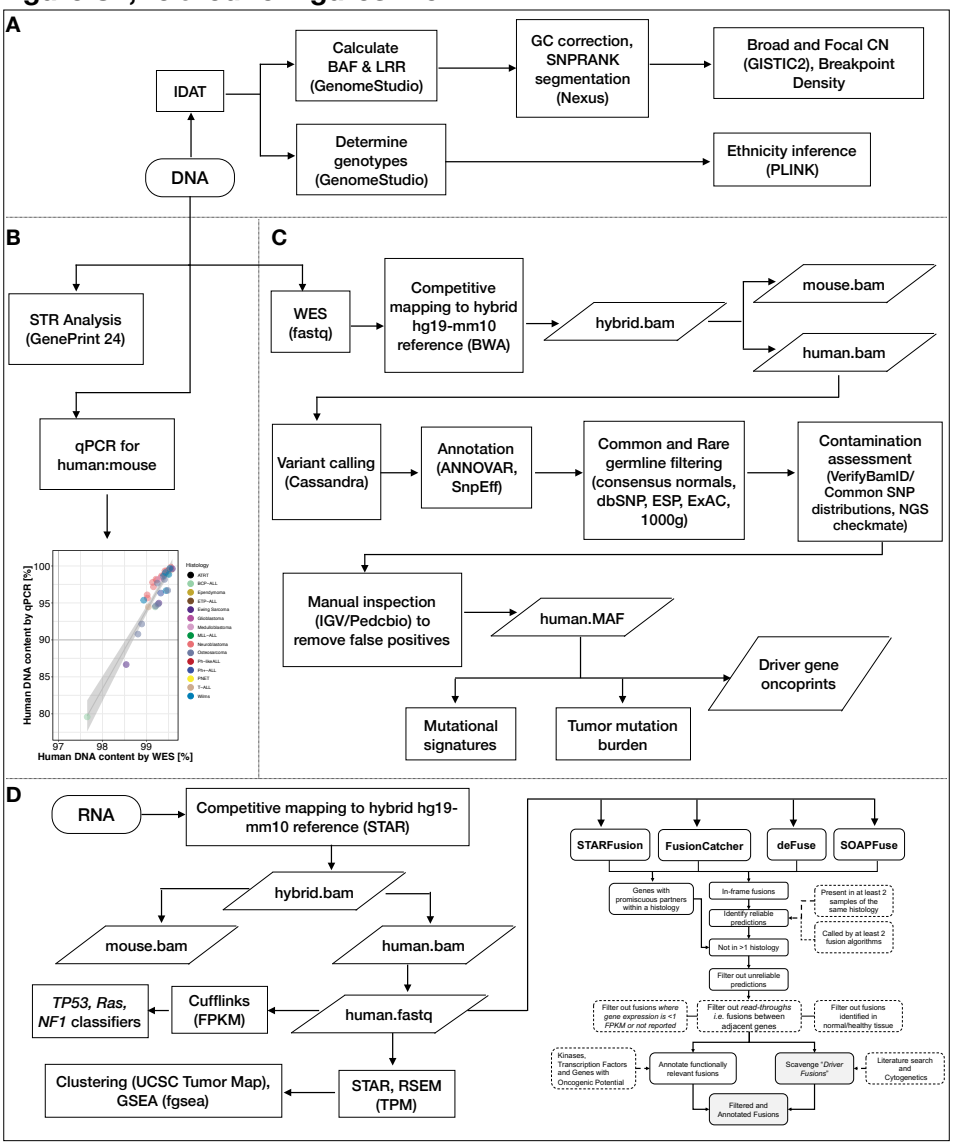
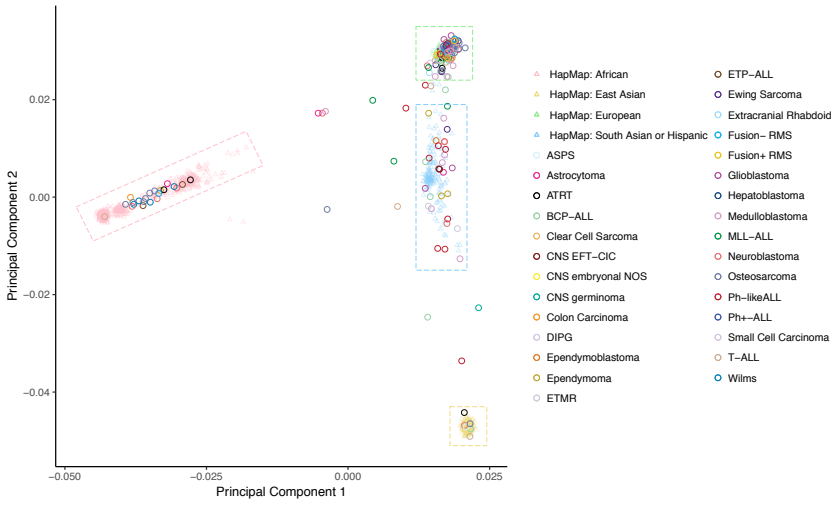


Figure S1. Analysis pipeline for somatic mutations, gene expression, RNA fusions, and copy number profiling in pediatric PDX tumors, Related to Figures 1-5. Figure S1 displays an overview of analysis methods utilized. Genomic DNA from PDX tumors was used for SNP array copy number analysis (A, N = 252), short-tandem repeat identity testing (B, N = 261), quantitative PCR to assess human:mouse DNA content (B, N = 35 samples with N = 3 technical replicates), and whole exome sequencing (C, N = 240). Total RNA from PDX tumors was used for whole transcriptome sequencing (D, N = 244). See Table S1 for Ns per assay per histology and Table S2 for STR profiles. Unless otherwise noted, Ns denote biological replicates.

Figure S2, related to Figure 1

A



B

	Number of Models	% of Total
European	181	71.8%
African	22	8.7%
East Asian	6	2.4%
South Asian or Hispanic	29	11.5%
Mixed or Unknown	14	5.6%
Total	**252**	100%

C

		Reported Ethnicity						
		African American	European	Hispanic or Latino	Mixed	Non-Hispanic	Other	Unknown
Inferred Ethnicity	African	5	1	0	0	2	0	14
	EastAsian	0	0	0	1	0	0	5
	European	3	25	3	0	10	1	139
	Mixed or Unknown	0	0	1	0	0	1	12
	SouthAsianOrHispanic	0	0	12	0	2	0	15

Figure S2. Ethnicity prediction, Related to Figure 1. Principal components analysis grouping of European, African, East Asian, and South Asian/Hispanic HapMap reference populations used to predict PDX ethnicities (A). The first two principal components calculated from SNP array genotypes for PDX models (circles, N = 252) are plotted alongside HapMap reference samples (triangles, N = 1,184). Dashed boxes represent the cutoffs used to classify PDXs into four broad population groups: European (including HapMap CEU and TSI population samples), African (ASW, LWK, MKK, and YRI), East Asian (CHB, CHD, and JPT), and South Asian or Hispanic (GIH and MXL). Tabulated counts and frequencies of ethnicities in PDX cohort (B) and a comparison table of reported versus inferred ethnicities in the PDX cohort (C). Ns represent biological replicates.

Figure S3, related to Figure 2

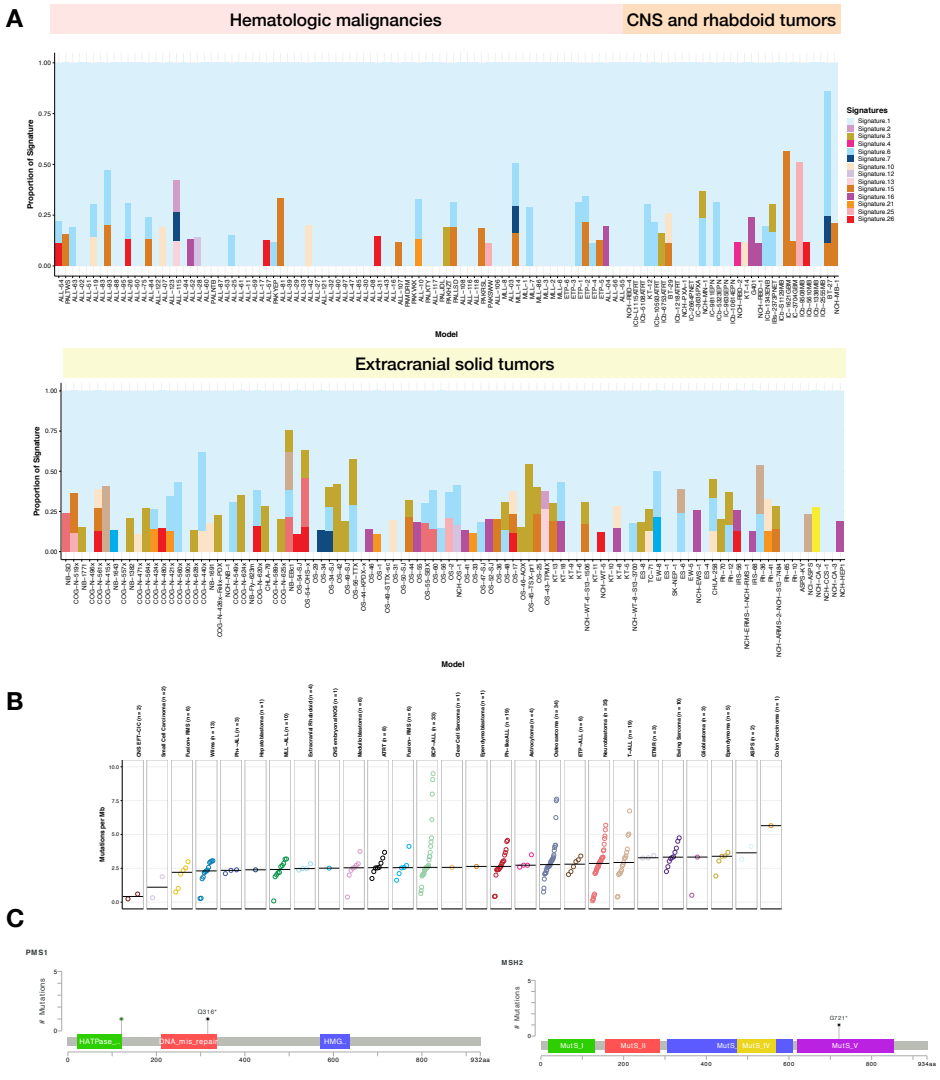


Figure S3. Mutational signatures and tumor mutational burden, Related to Figure 2. Mutational signatures per model displayed as proportion of signatures in a stacked barplot (A) Models with ≥ 50 mutations are depicted ($N_{leukemia}$ = 82, N_{brain} = 32, N_{solid} = 98). Tumor mutation burden by histology across N = 240 models on which WES was performed (B, STAR methods). Histologies are plotted in rank order by median (y-intercept) and Ns per histology are listed. Lollipop plots for oncogenic mutations in DNA repair genes, *PMS1* and *MSH2* for hypermutated model, IC-1621GBM (C). Ns represent biological replicates.

Figure S4, related to Figure 2

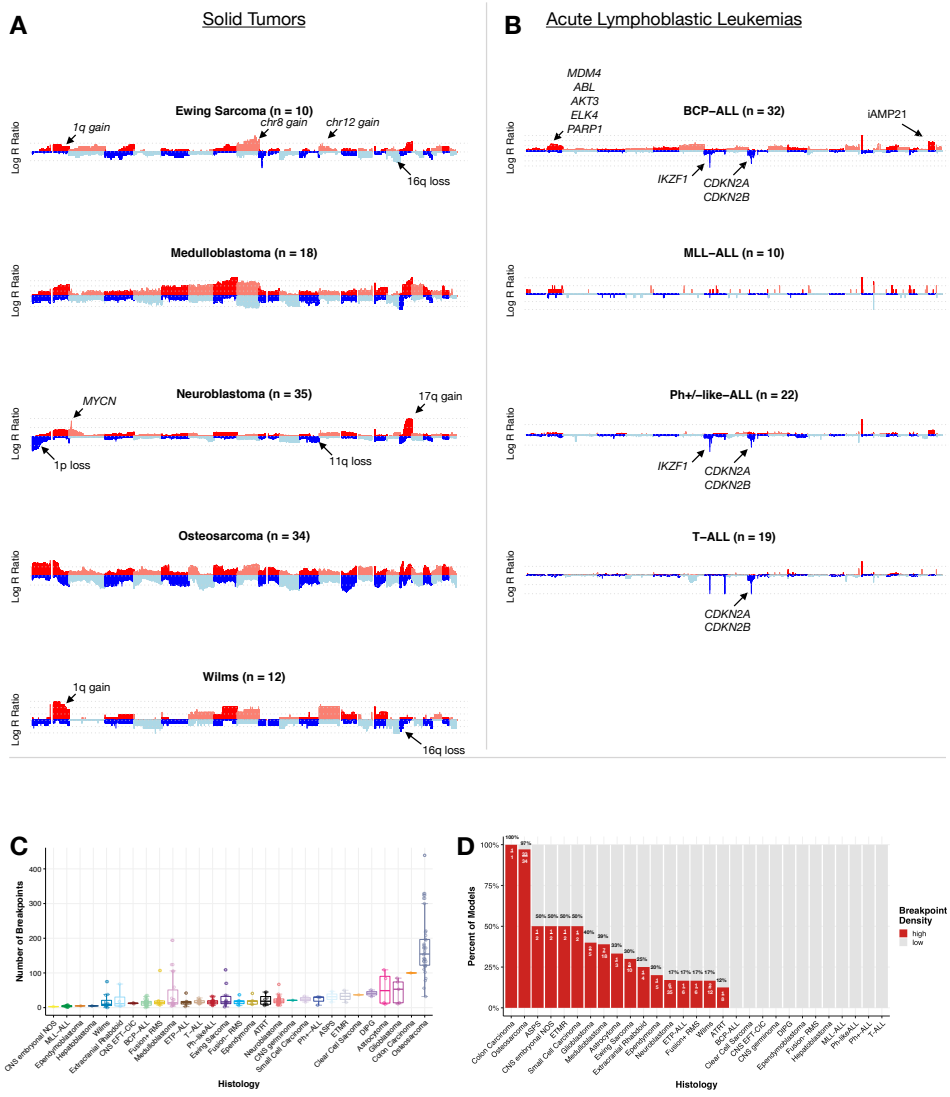


Figure S4. Copy number and breakpoint density. Related to Figure 2. Plotted are genome-wide copy-number profiles for histologies with N ≧ 10 models (Panel A, solid tumors: Ewing sarcoma, N = 10; Medulloblastoma, N = 18; Neuroblastoma, N = 35; Osteosarcoma, N = 34; Wilms, N = 12 and Panel B, leukemias: BCP-ALL, N = 32; MLL-ALL, N = 10; Ph+ or Ph-like ALL, N = 22; T-ALL, N = 19). Canonical broad and focal lesions are annotated by histology. Breakpoints per histology are plotted in C (boxplots are graphed as medians with box edges as first and third quartiles; detailed Ns in Table S4) and breakpoint density across histologies is plotted in D (displayed as % of models per histology with N/total; details in Table S4). Ns represent biological replicates.

Figure S5, related to Figures 4 and 5

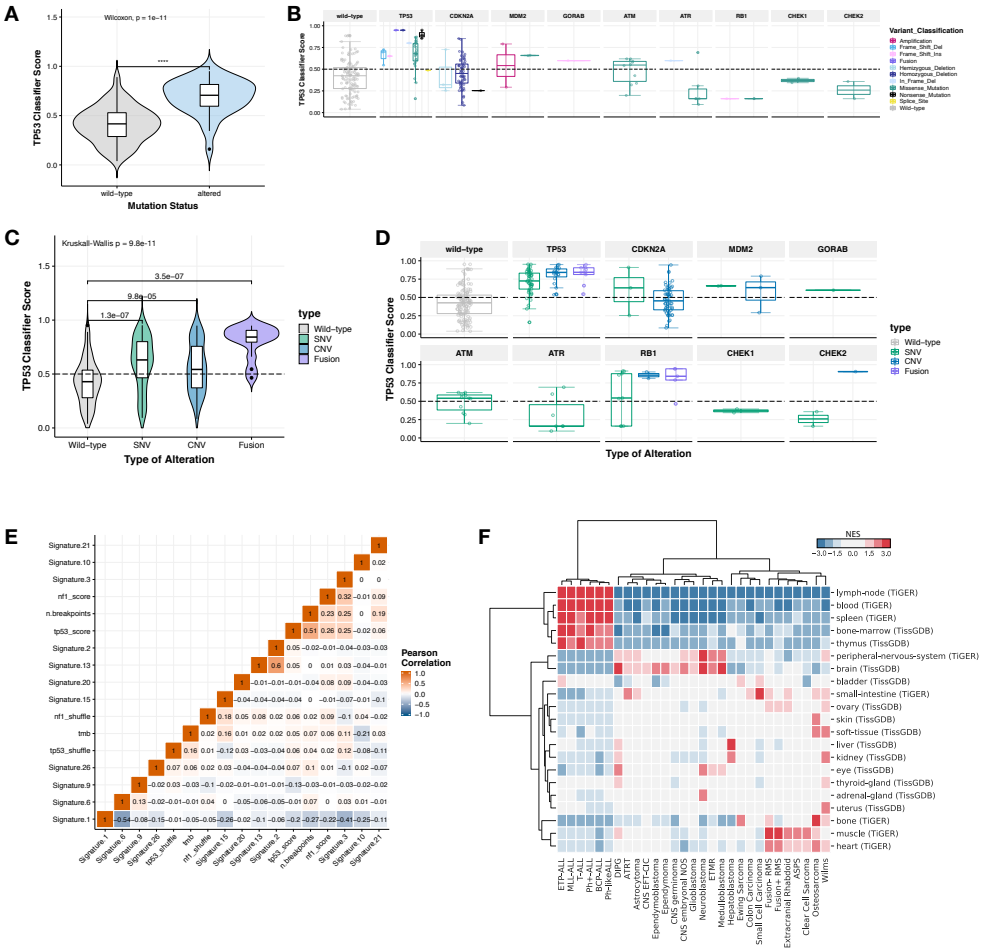


Figure S5. Classifier scores and mutational signature correlations, Related to Figure 4. With osteosarcoma models removed from analysis, *TP53* classifier scores were still significantly higher (N_{WT} = 180, N_{ALT} = 34, Wilcoxon p = 1e-11) in models with a *TP53* alteration (A), but alterations in other pathway genes don't consistently phenocopy *TP53* inactivation (B). Models containing fusions had highest classifier scores, followed by models with SNVs and CNVs, respectively (C, Kruskal-Wallis p = 9.8e-11, N_{WT} = 120, $N_{FUSIONS}$ = 14, N_{SNV} = 81, N_{CNV} = 85). *Post hoc* Wilcoxon p-values and group comparisons are displayed. Panel D breaks down the data in C by gene. Validation of mutational signatures via Pearson correlation matrix: Signatures 2 and 13 correlate strongly (R = 0.6, p= 6.5e-25, N = 260), Signature 1 is inversely correlated with impaired DNA repair mutational signatures, 3 (R = -0.41, p = 3.29e-11 , N = 260) and 6 (R = -0.54, p = 8.12e-20, N = 260) (E). Hierarchical clustering depicts tissue-specific enrichment within each histology (F, N = 244, NES = normalized enrichment score). All Ns denote biological replicates.

Part V

Conclusion

This work was motivated by the exigent need for new pediatric and adult cancer therapies. The UCSC Treehouse Childhood Cancer Initiative is an innovator in gene expression analysis and has led the way for bringing these technologies into the clinic in California. There are many facets to developing precision pediatric oncology methods. This thesis was concerned with the identification of tumor subtypes for the development of immunotherapies and the validation of therapies in preclinical models.

Unsupervised clustering of pediatric gene expression using the hydra method identified recurrent expression subtypes associated with the tumor microenvironment. The infiltration of immune cells correlated with chromatin remodeling, specifically increasing euchromatic state across normally silenced regions of the genome as a result of loss of ATRX functions. This led to the hypothesis that expression of the dark matter of the genome may be related to immune infiltration.

Progress in targeting cancer using immunotherapy has been impeded by current approaches relying on a single molecular target. This selects for subclones that do not express these targets. However, dysregulation of transcription and translation is a hallmark of cancer, so a combination cancer vacccine approach with checkpoint blockade may prevent subclones from evading the attack. We showed that the expression of the transposable element L1HS correlated with complete response to checkpoint blockade therapy in melanoma. This is evidence that our vaccination model may work since the complete responders were already predisposed to overexpressing the epitopes and their T-cell repertoire was already prepared to recognize and destroy the tumor. By preimmunizing against tumor TE epitopes, the T-cells circulating the body may be biased towards activation by cancer cells, tipping the balance in favor of response

to checkpoint blockade therapy.

These studies into cancer subtypes and potential therapeutic targets depend on the availability of preclinical models for validating these leads before testing in human subjects. The third and last main theme of this thesis was to develop analysis methods to evaluate an important preclinical model, the patient-derived xenograft, for its ability to reflect molecular features of the tumor of origin. This resulted in a new framework for designing PDX experiments that simplifies the interpretation of results and prioritizes models and tumor subtypes that are more accurately modeled in the PDX.

The scope of this thesis addresses related problems whose solutions will help facilitate the advancement of precision pediatric oncology. This work has initiated several ongoing collaborations and has impacted the trajectory of several research projects. In doing this work, I have developed essential research skills and will continue to advance my career based on the experience I had in writing this thesis. I am grateful for the students and mentors that I have worked with and look forward to fostering additional collaborations with the UCSC research community as I advance in my career in drug development.

While at UCSC, I have taken advantage of the interdisciplinary training offered by the Baskin School of Engineering. I have tailored my coursework to complement my background in biochemistry with graduate-level courses in bioinformatics, machine learning, and Bayesian statistics. I have excelled in my studies, achieving a 3.72 GPA. For my academic performance, I was awarded an NHGRI graduate training fellowship. I represented UCSC at the 2017 NHGRI training conference and was presented with an award for my poster presentation. Finally, I was award the 2019 BSOE Dissertation Fellowship which allowed me to expand the scope of my

dissertation and include several achievements that would not have been possible without this support. Training opportunities within the Baskin School of Engineering have expanded my skill set and allowed me to achieve my research goals.

My research addressed challenges that many clinical researchers are facing today. I have been invited to present research posters at the American Association for Cancer Research (AACR) Pediatric Cancer Research Conference, the TGen Precision Pediatric Oncology Conference, and the American Society of Clinical Oncology (ASCO) Conference. I was given a travel award to present my work at the TGen meeting and was one of the few researchers from UCSC to be invited to present at the high-profile ASCO meeting. My unique training at UCSC has allowed me to make progress on difficult problems in the computational analysis of cancer gene expression data.

Scientific publications has been another important component of my education at UCSC. I have contributed to several manuscripts from an early point in my graduate training. My significant contributions have been acknowledged in the Toil manuscript [41], the Treehouse gene expression outlier manuscript [40], and the ProTECT manuscript (in review). I have also contributed to manuscripts with collaborators at UCSF and the University of Pennsylvania. I contributed a neoepitope burden analysis of a patient who had an exceptional response to checkpoint blockade therapy at UCSF (in review). I also contributed to a pediatric preclinical modeling paper in collaboration with John Maris lab at the University of Pennsylvania [34]. I also have a first author publication accepted in the high-profile PloS Computational Biology journal describing the hydra computational analysis for precision oncology research.

In addition to scientific publications, I have also written significant portions of grants

that have been funded to support future research at UCSC. I collaborated with Dr. Alejandro Sweet-Cordero at UCSF to develop a novel framework for evaluating pre-clinical models. We co-wrote an National Cancer Institute grant that was recently funded and will go into effect June 2020 and will support several researchers at UCSC. The hydra method was also featured in a Treehouse grant to fund undergraduate research to improve subtyping of acute myeloid leukemia.

Lastly, I have been an effective mentor to high school and undergraduate students, and have helped my mentees achieve recognition for their research. I was invited to co-author the BD2K Summer Research Workshop and presented the workshop for two summers. The workshop focuses on computational tools for biomedical research and has prepared college students for research at UCSC. I have mentored high school students as part of the UCSC Science Internship Program. These students have been invited to present their research at the AMIA High School Scholar and Sigma Xi Student Research Conferences. I am grateful for the opportunity to mentor students and prepare them to be effective contributors in the field of biomedical research.

Bibliography

[1] Eleni Berger. Late and long-term effects of treatment of childhood leukemia. 2016.

[2] Kathrin MMD Bernt and Stephen PMD Hunger. Current concepts in pediatric Philadelphia chromosome-positive acute lymphoblastic leukemia. *Frontiers in oncology*, 4:54, 2014.

[3] Michele Ceccarelli, Floris P Barthel, Tathiane M Malta, Thais S Sabedot, Sofie R Salama, Bradley A Murray, Olena Morozova, Yulia Newton, Amie Radenbaugh, Stefano M Pagnotta, et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, 164(3):550–563, 2016.

[4] James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202, 2014.

[5] Vincent T DeVita and Edward Chu. A history of cancer chemotherapy. *Cancer research*, 68(21):8643–8653, 2008.

[6] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali

Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[7] Hariharan Easwaran, Hsing-Chen Tsai, and Stephen B Baylin. Cancer epigenetics: Tumor heterogeneity, plasticity of stem-like states, and drug resistance. *Molecular cell*, 54(5):716–727, 2014.

[8] Hannah Farmer, Nuala McCabe, Christopher J Lord, Andrew NJ Tutt, Damian A Johnson, Tobias B Richardson, Manuela Santarosa, Krystyna J Dillon, Ian Hickson, Charlotte Knights, et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, 434(7035):917–921, 2005.

[9] Jianxing Feng, Clifford A Meyer, Qian Wang, Jun S Liu, X Shirley Liu, and Yong Zhang. GFOLD: A generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics*, 28(21):2782–2788, 2012.

[10] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[11] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press, 2006.

[12] Malachi Griffith, Obi L Griffith, Adam C Coffman, James V Weible, Josh F McMichael, Nicholas C Spies, James Koval, Indraniel Das, Matthew B Callaway, James M Eldred, et al. DGIdb: Mining the druggable genome. *Nature methods*, 10(12):1209–1210, 2013.

[13] Douglas Hanahan and Lisa M Coussens. Accessories to the crime: Functions of cells recruited to the tumor microenvironment. *Cancer cell*, 21(3):309–322, 2012.

[14] David C Hoaglin, Boris Iglewicz, and John W Tukey. Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 81(396):991–999, 1986.

[15] Robert Huether, Li Dong, Xiang Chen, Gang Wu, Matthew Parker, Lei Wei, Jing Ma, Michael N Edmonson, Erin K Hedlund, Michael C Rusch, et al. The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. *Nature communications*, 5, 2014.

[16] Carolyn M. Huston. *Complex Bayesian Models: Construction, and Sampling Strategies*. PhD thesis, Simon Fraser University, 2011.

[17] Chris Jones and Suzanne J Baker. Unique genetic and epigenetic mechanisms driving paediatric diffuse high-grade glioma. *Nature Reviews Cancer*, 14(10):651–661, 2014.

[18] Steven JM Jones, Janessa Laskin, Yvonne Y Li, Obi L Griffith, Jianghong An, Mikhail Bilenky, Yaron S Butterfield, Timothee Cezard, Eric Chuah, Richard Corbett, et al. Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome biology*, 11(8):R82, 2010.

[19] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: A knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl 1):D428–D432, 2005.

[20] Johanna A Joyce and Douglas T Fearon. T cell exclusion, immune privilege, and the tumor microenvironment. *Science*, 348(6230):74–80, 2015.

[21] Raghu Kalluri. The biology and function of fibroblasts in cancer. *Nature Reviews Cancer*, 16(9):582–598, 2016.

[22] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[23] Lisa M Kopp, Puja Gupta, Luz Pelayo-Katsanis, Brenda Wittman, and Emmanuel Katsanis. Late effects in adult survivors of pediatric cancer: A guide for the primary care physician. *The American journal of medicine*, 125(7):636–641, 2012.

[24] Vishal Kothari, Iris Wei, Sunita Shankar, Shanker Kalyana-Sundaram, Lidong Wang, Linda W Ma, Pankaj Vats, Catherine S Grasso, Dan R Robinson, Yi-Mi Wu, et al. Outlier kinase expression by RNA sequencing as targets for precision therapy. *Cancer discovery*, 3(3):280–293, 2013.

[25] Bo Li and Colin N Dewey. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

[26] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[27] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo

Tamayo, and Jill P Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[28] Stephen V Liu, Vincent A Miller, Marinus W Lobbezoo, and Giuseppe Giaccone. Genomics-based early-phase clinical trials in oncology: Recommendations from the task force on Methodology for the Development of Innovative Cancer Therapies. *European Journal of Cancer*, 50(16):2747–2751, 2014.

[29] James W Macdonald and Debashis Ghosh. COPA—cancer outlier profile analysis. *Bioinformatics*, 22(23):2950–2951, 2006.

[30] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, volume 122. CRC Press, 2016.

[31] S. Mukherjee. *The Emperor of All Maladies: A Biography of Cancer*. Scribner, 2010.

[32] Robin E Norris and Peter C Adamson. Challenges and opportunities in childhood cancer drug development. *Nature Reviews. Cancer*, 12(11):776, 2012.

[33] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[34] Jo Lynne Rokita, Komal S. Rathi, Maria F. Cardenas, Kristen A. Upton, Joy Jayaseelan, Katherine L. Cross, Jacob Pfeil, Laura E. Egolf, Gregory P. Way, Alvin Farrel, Nathan M. Kendsersky, Khushbu Patel, Krutika S. Gaonkar, Apexa Modi, Esther R. Berko, Gonzalo

Lopez, Zalman Vaksman, Chelsea Mayoh, Jonas Nance, Kristyn McCoy, Michelle Haber, Kathryn Evans, Hannah McCalmont, Katerina Bendak, Julia W. Böhm, Glenn M. Marshall, Vanessa Tyrrell, Karthik Kalletla, Frank K. Braun, Lin Qi, Yunchen Du, Huiyuan Zhang, Holly B. Lindsay, Sibo Zhao, Jack Shu, Patricia Baxter, Christopher Morton, Dias Kurmashev, Siyuan Zheng, Yidong Chen, Jay Bowen, Anthony C. Bryan, Kristen M. Leraas, Sara E. Coppens, HarshaVardhan Doddapaneni, Zeineen Momin, Wendong Zhang, Gregory I. Sacks, Lori S. Hart, Kateryna Krytska, Yael P. Mosse, Gregory J. Gatto, Yolanda Sanchez, Casey S. Greene, Sharon J. Diskin, Olena Morozova Vaske, David Haussler, Julie M. Gastier-Foster, Kolb E. Anders, Richard Gorlick, Xiao-Nan Li, C. Patrick Reynolds, Raushan T. Kurmasheva, Peter J. Houghton, Malcolm A. Smith, Richard B. Lock, Pichai Raman, David A. Wheeler, and John M. Maris. Genomic Profiling of Childhood Tumor Patient-Derived Xenograft Models to Enable Rational Clinical Trial Design. *Cell reports*, 29(6):1675–1689.e9, November 2019.

[35] Mariacarmela Santarpia and Niki Karachaliou. Tumor immune microenvironment characterization and response to anti-PD-1 therapy. *Cancer biology & medicine*, 12(2):74, 2015.

[36] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2016. *CA: a cancer journal for clinicians*, 66(1):7–30, 2016.

[37] Tom AB Snijders. *Multilevel Analysis*. Springer, 2011.

[38] T. Strachan and A. P. Read. *Human Molecular Genetics*. Garland Science, Taylor & Francis Group LLC, 4th edition, 2011.

[39] The American Cancer Society. Cancers that develop in children. August 2016.

[40] Olena M. Vaske, Isabel Bjork, Sofie R. Salama, Holly Beale, Avanthi Tayi Shah, Lauren Sanders, Jacob Pfeil, Du L. Lam, Katrina Learned, Ann Durbin, Ellen T. Kephart, Rob Currie, Yulia Newton, Teresa Swatloski, Duncan McColl, John Vivian, Jingchun Zhu, Alex G. Lee, Stanley G. Leung, Aviv Spillinger, Heng-Yi Liu, Winnie S. Liang, Sara A. Byron, Michael E. Berens, Adam C. Resnick, Norman Lacayo, Sheri L. Spunt, Arun Rangaswami, Van Huynh, Lilibeth Torno, Ashley Plant, Ivan Kirov, Keri B. Zabokrtsky, S. Rod Rassekh, Rebecca J. Deyell, Janessa Laskin, Marco A. Marra, Leonard S. Sender, Sabine Mueller, E. Alejandro Sweet-Cordero, Theodore C. Goldstein, and David Haussler. Comparative Tumor RNA Sequencing Analysis for Difficult-to-Treat Pediatric and Young Adult Patients With Cancer. *JAMA Network Open*, 2(10):e1913968–e1913968, October 2019.

[41] John Vivian, Arjun Arkal Rao, Frank Austin Nothaft, Christopher Ketchum, Joel Armstrong, Adam Novak, Jacob Pfeil, Jake Narkizian, Alden D Deran, Audrey Musselman-Brown, et al. Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology*, 35(4):314–316, 2017.

[42] Chenwei Wang, Alperen Taciroglu, Stefan R Maetschke, Colleen C Nelson, Mark A Ragan, and Melissa J Davis. mCOPA: Analysis of heterogeneous features in cancer expression data. *Journal of clinical bioinformatics*, 2(1):22, 2012.

[43] Ian R Watson, Koichi Takahashi, P Andrew Futreal, and Lynda Chin. Emerging patterns of somatic mutations in cancer. *Nature reviews Genetics*, 14(10):703–718, 2013.

[44] Stephane Wong and Owen N Witte. The BCR-ABL story: Bench to bedside and back. *Annu. Rev. Immunol.*, 22:247–306, 2004.

[45] Eda Yildirim, James E Kirby, Diane E Brown, Francois E Mercier, Ruslan I Sadreyev, David T Scadden, and Jeannie T Lee. Xist RNA is a potent suppressor of hematologic cancer in mice. *Cell*, 152(4):727–742, 2013.

[46] Jinghui Zhang, Michael F Walsh, Gang Wu, Michael N Edmonson, Tanja A Gruber, John Easton, Dale Hedges, Xiaotu Ma, Xin Zhou, Donald A Yergeau, et al. Germline mutations in predisposition genes in pediatric cancer. *New England Journal of Medicine*, 373(24):2336–2346, 2015.