A Genomic Analysis Pipeline and Its Application to Pediatric Cancers

Michael Zeller, Christophe N. Magnan, Vishal R. Patel, Paul Rigor, Leonard Sender, and Pierre Baldi

Abstract—We present a cancer genomic analysis pipeline which takes as input sequencing reads for both germline and tumor genomes and outputs filtered lists of all genetic mutations in the form of short ranked list of the most affected genes in the tumor, using either the Complete Genomics or Illumina platforms. A novel reporting and ranking system has been developed that makes use of publicly available datasets and literature specific to each patient, including new methods for using publicly available expression data in the absence of proper control data. Previously implicated small and large variations (including gene fusions) are reported in addition to probable driver mutations. Relationships between cancer and the sequenced tumor genome are highlighted using a network-based approach that integrates known and predicted protein-protein, protein-TF, and protein-drug interaction data. By using an integrative approach, effects of genetic variations on gene expression are used to provide further evidence of driver mutations. This pipeline has been developed with the aim to be used in assisting in the analysis of pediatric tumors, as an unbiased and automated method for interpreting sequencing results along with identifying potentially therapeutic drugs and their targets. We present results that agree with previous literature and highlight specific findings in a few patients.

Index Terms—Pediatric cancer, genome analysis, next generation sequencing

1 Introduction

At the most fundamental level, cancer is a disease of the DNA, in which changes to the DNA sequence and the molecules that interact with it ultimately lead to uncontrolled cell proliferation. Thus high-throughput sequencing technologies capable of identifying not only the DNA sequence (DNA-seq) but, for instance, also epigenetic states (e.g. Methyl-seq) and gene expression levels (RNA-seq), hold the promise to help better understand cancer in all its various forms. Indeed large-scale cancer sequencing projects, such as the Cancer Genome Atlas [1], have already started and produced volumes of data that are already well beyond what can be transferred over the Internet. However, these projects are still at a relatively early stage of development and are fraught with numerous challenges associated with the complexity of the sequencing technology, the lack of standardization, the sheer volume of data, the heterogeneity of cancers, the complexity of cancer biology, and the problem of obtaining proper control samples, to name only a few. Although incomplete by necessity, problems, solutions, and results from these projects ought to be shared periodically in order to move the field towards more standardized solutions and accelerate the pace of discovery. Here we describe the ongoing development of a computational pipeline for the analysis of high-throughput sequencing cancer data that is currently being applied to pediatric cancer data that is regularly being sequenced, and further resequenced on recurrence, as a result of a collaboration between the University of California, Irvine (UCI) and the Children Hospital of Orange County (CHOC).

Worldwide, it is estimated that childhood cancer has an incidence of more than 175,000 per year, and a mortality rate of approximately 96,000 per year. In the United States, cancer is the second most common cause of death among children between the ages of 1 and 14 years, exceeded only by accidents, with an incidence of about 12,000 of newly diagnosed cases per year and 1,300 deaths. The most common cancers in children are (childhood) leukemia (34 percent), brain tumors (23 percent), and lymphomas (12 percent). Other, less common childhood cancer types are: Neuroblastoma (7 percent), Wilms tumor (5 percent), NonHodgkin lymphoma (4 percent), Rhabdomyosarcoma (3 percent), Retinoblastoma (3 percent), Osteosarcoma (3 percent), Ewing sarcoma (1 percent), Germ cell tumors, Pleuropulmonary blastoma, Hepatoblastoma, and hepatocellular carcinoma. White and Hispanic children are more likely than children from any other racial or ethnic group to develop cancer. The causes of most childhood cancers are unknown. The CHOC receives on the order of 100 new cases per year, and a project was started in 2012 to sequence the genome from healthy and cancer tissues of a subset of newly diagnosed cases—and therefore with no emphasis on particular tumors or tissue types—together with high-throughput gene expression measurements from cancer cells using RNA-seq.

Our goal has been to develop an analysis pipeline comprising a combination of in-house and third party software to manage and analyze the raw data produced by these experiments, in a timely manner after they become available, including the identification and ranking of affected genes containing both small and large variants, and their

- *M. Zeller, C.N. Magnan, V.R. Patel, P. Rigor, and P. Baldi are with the Department of Computer Science, University of California, Irvine, CA 92617. E-mail: {zellerm, cmagnan, vishalrp, prigor}@uci.edu, pfbaldi@ics.uci.edu.*
- *L. Sender is with the Children Hospital of Orange County, Orange, CA 92868. E-mail: lsender@uci.edu.*

Manuscript received 6 Jan. 2014; revised 8 May 2014; accepted 28 May 2014. Date of publication 11 June 2014; date of current version 2 Oct. 2014.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2330616

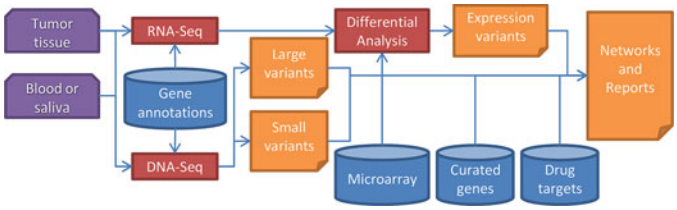

Fig. 1. Overview of the genomics analysis pipeline which starts from raw sequencing reads derived from two biological samples per patient and results in a HTML report with ranked genes and pathways.

integrative systems biology analyses against the large background of omic, literature, and other data available to us in order to derive inferences of clinical relevance specific to the cancer types of the patients sequenced.

2 Methods

2.1 Sequence Processing

2.1.1 Sequencing Data Generated for Each Patient

An overview of our pipeline is presented in Fig. 1. It begins by collecting two different samples for each patient participating in the CHOC pediatric cancers project. The first sample is collected in the tissue affected by the cancer and the second sample is collected either from blood or saliva depending on the patient to be used as a control sample during the analysis.

Both samples are then provided to either Complete Genomics (Mountain View, CA) for the *Cancer Sequencing Service* offered by the company or to Illumina, Inc. (San Diego, CA) for a rapid sequencing of both control and cancer genomes using the *RapidTrack WGS Service* offered by the company. When the sample extracted at the tumor tissue is not exhausted by the DNA extraction, RNA sequencing is also performed using an Illumina HiSeq 2500 instrument either by the Scripps Research Next Generation Sequencing Core Facility (San Diego, CA) or by the Genomics High-Throughput Facility of the University of California, Irvine (Irvine, CA). Sequencing platforms, data vendors, patient description, and data obtained for each patient are reported in Supplementary Table 1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2014.2330616.

2.1.2 Quality Controls and Data Filtering

The sequencing data quality is assessed based on the standard PHRED quality scores predicted during the base calling step on each sequencing platform and on the base call distribution for each sequencing cycle. Sequencing reads for all the datasets generated on an Illumina instrument are paired 100 bp reads. Sequencing reads provided by Complete Genomics are paired 35 bp reads where each 35 bp read is made of four shorter reads close but not necessarily contiguous on the sequenced genome. An overview of the sequencing data quality is provided in Supplementary Figs. 1 and 2, available online. The RNA sequencing data is subject to the same quality controls and is pre-filtered to remove common contaminants in RNA-seq libraries. Reads mapping to the mitochondrial or nuclear ribosomal RNA genes and PhiX control reads are removed from the original datasets.

2.1.3 Alignment to Reference Genome hg19/GRCh37

Both Complete Genomics and Illumina deliver the short-read alignment results as part of their sequencing service. Alignments delivered by Complete Genomics are performed using the CGA tools developed by the same company to handle the specific structure of the short-reads. Alignments delivered by Illumina are performed using their short-read aligner Eland v2e. The RNA sequencing data is aligned to the reference genome hg19 together with the known splice junction sequences extracted from the RefSeq database using Eland v2e.

Genome and transcriptome coverage corresponding to the DNA and RNA sequencing data is reported in Supplementary Fig. 3, available online. The mean relative chromosome coverage broken by gender and sequencing platform is reported in Supplementary Figs. 4 and 5, available online, indicating a significant coverage bias for several chromosomes on the Complete Genomics sequencing platform.

2.2 Genome Assembly

2.2.1 Assembly Provided by the Sequencing Companies

Control and cancer genomes are assembled separately by both companies using a diploid model. The methods used by each company to generate a consensus assembly from the short-read alignment results have significant differences (not discussed here), which results in assemblies with heterogeneous characteristics. For instance, alleles are tracked individually by Complete Genomics whenever possible, allowing to call a position on one allele only while leaving the second allele uncalled. On the other hand, Illumina calls are based on a consensus of the aligned reads for each position resulting in either a position fully called or not called. Another significant difference is in the strategy used to call the small variations in the genome assemblies. While Complete Genomics provides a final decision for each allele and each position in the genome (including positions not called, partially called, reference calls, snps, short indels, and substitutions), the strategy followed by Illumina consists in separating what is observed at each position in the alignment results, the snp calling analysis and results, and the indel calling analysis and results, leading to a situation where conflicts between the different calls may occur. Annotations provided with the predicted variations also differ significantly.

2.2.2 Standardized Multi-Genome Assemblies

Rapidly evolving technologies and softwares, lack of standardization, and large volumes of heterogeneous data are common issues in genomic analyses and are paired here with the necessity to compare several assemblies for the same individuals to extract relevant differences potentially correlated with the corresponding disease. Our strategy to limit the effects of these problems and provide uniform downstream analyses for all the patients is to adopt a fixed

TABLE 1
Typical Distribution of the Small Variations

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

Mixed variations include cases where different small variants are called between the two alleles of a genome or between the two genomes.

representation of genome assemblies consisting in three major components: 1) the features needed for the downstream analysis; 2) a fixed level of description and annotation; and 3) a standardized scoring system and data format allowing multiple genome assemblies and RNA-seq experiments to be rapidly combined with each other and compared.

Features selected to describe each assembly include a unique call for each allele, the called allele sequence, zygosity, ploidy, call confidence level, read counts (coverage), and genomic annotations corresponding to each position. Annotations include promoter regions, untranslated regions, transcription start site (TSS), start and stop codons, donor and acceptor sites, exons, introns, coding regions, and transcription factor binding sites (TFBS) predicted by our in-house software MotifMap [2], [3].

Generating the assemblies following this model is a relatively straightforward process starting from the assemblies provided by Complete Genomics, due to the selected features being a subset of the provided ones. On the other hand, the assemblies provided by Illumina are completed using in-house software to reach the same level of description: conflicting snp and indel calls are resolved based on the PHRED score associated with each variant call, detailed genomic locations are added based on the RefSeq annotation database, and COSMIC annotations [4] are added to the small variation calls whenever available.

2.2.3 Control and Cancer Genomes Comparison

The two assembled genomes for each patient can directly be compared from the assemblies described in the previous section. The comparative analysis is described in the next sections but a few comments are given here as they apply to all the comparisons we made on the assemblies and more generally to the complexity of comparing genome assemblies. First, positions not fully called on both genomes and both alleles have been excluded from the rest of the analysis as they do not allow a reliable comparison between the two genomes. Around 95 percent of the known positions in the reference sequences are fully called on both genomes regardless of the sequencing platform. Also, small variations are extracted for the entire genome but only the ones located in genic regions (including promoter regions and putative TFBS) are further studied during the next steps of the pipeline. The numerous small variations located in the intergenic regions fall into a more general problem out of the scope of this study—that of identifying the consequences these variations may have on the organism. They are thus counted for information but not further analyzed.

2.3 Small Variations

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

2.3.1 Comparative Analysis

The small variations called during the genome assembly (see Section 2.2) are classified into four categories: SNPs, insertions, deletions, and substitutions. Between 4 and 4.5 million such variations with the reference sequences are called for each assembled genome in this project. The sequencing platform used and the software developed by both data vendors to call the small variations do not affect significantly this number. The reliability of many called small variations could easily be discussed as it is to be expected that alignment algorithms, small variation calling methods, and the natural complexity of predicting these variations in many regions of the genome probably result in many false calls and systematic biases. However, in our case, the two samples for each patient in the CHOC pediatric cancers project are sequenced on the same platform and the genomes are assembled using the same methods and software, hence a significant part of the biases and false calls is thus likely to be repeated on each assembly. By comparing both genomes and extracting only the differences between the cancer genome and the control genome, we can reasonably assume these issues to affect the resulting set of variations significantly less.

Variations observed between the cancer genome and the control genome only represent a very small fraction of the variations called on both genomes, less than 0.01 percent in most cases (example provided for one patient in Table 1), reducing drastically the number of variations to further analyze for each patient. Variations observed on both genomes are not studied further regardless of the effect these variations may have on proteins when they occur in gene regions.

2.3.2 Variant Location and Effect

Small variations observed only in the cancer genome and in genic regions can be further analyzed as their effect on the resulting proteins can be directly deduced from their

TABLE 2
Mean Size and Standard Deviation of the Gene Lists Extracted During the Various Stages of the Analysis for Each Patient

Small Variations			Large Variations			Gene Expression			Curated Gene Lists		
Gene List	Mean	StdDev	Gene List	Mean	StdDev	Gene List	Mean	StdDev	Gene List	Mean	StdDev
Missense	589.9	437.1	Deletion	406.2	555.1	Under expr. HIGH	52.0	67.1	Entrez	154.6	134.5
Nonsense	26.2	27.8	Inversion	234.3	325.1	Under expr. MED	433.2	172.3	MEDLINE	70.8	69.5
Nonstop	2.0	3.0	Tandem-Dup	122.8	188.1	Under expr. LOW	14.2	21.7	GeneRIF	63.7	74.1
Misstart	1.2	2.0	Distal-Dup	8.3	27.6	Over expr. HIGH	24.8	42.7			
Splicing	31.4	28.0	Inter-Chr	51.0	51.2	Over expr. MED	438.2	99.6			
Frameshift	113.3	161.2	Gene Fusion	24.4	21.7	Over expr. LOW	4.9	11.2			
Inframe	61.4	51.0	Higher CNV	782.2	1221.8	All tumors HIGH	54.9	71.1			
LOH	179.6	321.5	Lower CNV	488.4	1423.3	All tumors MED	457.0	147.7			
						All tumors LOW	74.0	82.1			
						Tumor sig.	1430.4	489.8			
						Control sig.	1087.2	1208.5			
						Contrast sig.	2455.7	3011.0			
UNION	781.1	563.6	UNION	2030.3	1774.7	UNION	4897.9	2570.6	UNION	243.3	196.2

These lists are used in computing a ranking score for each gene in the final reports.

location in many cases. Eight distinct types of disruptions or changes in the proteins are considered in our pipeline. Seven of them are inspired by the classification of the variant effects performed by the CGA tools (Complete Genomics) and we added the loss of heterozygosity between the germline and cancer genomes, frequently reported in cancer cases [5], resulting in the following classification:

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

Variations not matching with any of these eight categories either correspond to silent variations (no effect on the protein) or to variations with unknown effect on the protein (e.g., located in intronic regions). Variations with unknown effect are not considered in the network analysis described in Section 2.6 due to the lack of information about the severity of the effect on the corresponding product. These variations are therefore kept aside for cases where a gene is predicted to be disrupted by the differential expression analysis but does not have variations included in one of the eight categories above or in the large variations described in Section 2.4 explaining this disruption.

For each of the eight variant effects listed above, we use the hg19 gene coordinates to extract the list of genes overlapping the called small variations. The confidence for a gene to be actually affected by such variation is directly given by the confidence of the small variation call. The sizes of the gene lists for each patient are summarized in Table 2 and range from a few genes for the most deleterious variations to a few hundred genes for missense mutations, which are more common and less likely to be deleterious than other variations.

2.3.3 Small Variations in Transcripts

RNA sequencing data is available for a large portion of the patients in this project. Small variations can therefore also be called for the transcripts based on the alignment results (Section 2.1.3). We use the software developed by Illumina, Casava variant detection and counting (VDC), to extract SNPs and indels following the same protocol as the one used by Illumina for their DNA sequencing service. The resulting variations are used as an additional control for the clinically relevant results emerging from the final network analysis.

2.3.4 Flagged Small Variations

Putative small variations, specifically single nucleotide polymorphisms (SNPs), have been categorized into three subsets—unique, common, or flagged—with respect to the latest dbSNPs (version 137) tracks from the UCSC Genome Browser [6]. Specifically, when creating these subsets we used the curated subsets of dbSNPs referred to as Common SNPs and Flagged SNPs.

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

Further, for the Flagged SNPs, only SNPs flagged as clinically associated by dbSNP, that map to a single location in the reference genome assembly, and not known to have a minor allele frequency of at least 1 percent, are included. SNPs that do not fit into either the common or flagged SNPs subsets are categorized as unique SNPs, specific to the patient.

2.3.5 Protein Domains

Besides the commonly characterized effect of small variations on protein coding sequence detailed in Section 2.3.2, we characterized the location of small variants based on predicted secondary structure and solvent accessibility using the SCRATCH software suite [7], [8]. In addition, protein domain families predicted by Pfam [9] are used to identify if the small variation affects a protein family domain, which in many cases can identify important functional portions of a protein such as protein binding domains. This

information is incorporated into our final report in order to manually investigate the consequences of small variations.

2.3.6 Variant Transcription Factor Binding Sites

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

2.4 Large Variations

Large variations, also refered to as structural variations, are the large-scale chromosomal variations or rearrangements leading to a significant change in the classical organization of the DNA in the genome. There is no software specifically recognized to perform significantly better than the others in this area and the accuracy of the predicted large variations is still very unclear in most cases. Moreover, the consistency between the different types of large variations predicted by separate tools and with the genome assembly (Section 2.2) is not checked or resolved by such tools. Such a task can rapidly become complicated as structural variations can be very complex to track in some cases, notably using only short reads. The different types of large variations considered in our pipeline are described below.

- *Novel junctions* are observed junctions between distant parts of the genome (intra- or inter-chromosomal). The position of each partner in the junction and the direction of the sequences observed on the sequencing reads allow us to predict the corresponding large variation event: inversion, deletion, tandem or distal duplication, or inter-chromosomal rearrangement. These large variations are provided by each of the vendors.
- *Gene fusions* are a special case of novel junctions leading to the fusion of two distantly located genes resulting in a new, functionally different protein product. Detecting putative gene fusions can be performed by analyzing each partner position in the detected novel junctions. This analysis is performed by Complete Genomics and performed in-house for the data delivered by Illumina.
- *Copy number variations (CNVs)* are relatively large deletion or duplication events leading to a different number of copies observed for specific regions of the genome. They can, for instance, be detected by comparing the coverage in each region of the genome with baseline distributions computed on control populations.
- *Chromosome duplications* or deletions are a particular case of CNVs where an entire chromosome is either missing or observed with more than two copies. Their detection is very similar to CNVs but requires an overall coverage bias. These variations are detected using in-house software and further validated based on the expression results obtained following the protocol described in Section 2.5.

Comparing the exact positions or sequences of the predicted large variations in both genomes results in most cases in classifying all of the large variations predicted in the cancer genome as being specific to that genome. We thus implemented a case-by-case set of rules based on the overlap length between the large variations predicted for the baseline genome and the ones predicted for the cancer genome (not detailed here) to decide if a large variation is likely to be unique to the cancer genome or not. Similarly to the small variations, we list all of the genes affected by large variations in the cancer genome considering the eight following categories:

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

The sizes of each of these gene lists are summarized in Table 2.

2.4.1 Mitelman Fusions

The Mitelman database [10] contains 3,752 entries corresponding to gene fusions implicated in different types of cancer. To identify and prioritize these gene fusions in our patients, we cross this database with all of the gene fusions found for each patient to identify high-priority fusions and to present the relevant literature in our final reports that are of clinical relevance. Three of the patients in our study contained fusions previously described. These fusions were originally identified in the same tumor type as each of the patients. All identified Mitelman fusions are listed in Supplementary Table 4, available online.

2.5 Expression Analysis

Gene expression levels in the tumor samples are computed directly from the read alignment results. Standard RPKM values (reads per kilobase of exon model per million mapped reads) are computed for each exon, splice junction, and isoform covered by the sequencing data. No RNA sequencing data is available for the baseline genome samples, or any other control tissue samples, and therefore standard differential analysis of the gene expression levels cannot be performed for each patient. Various approaches are considered in our study to predict abnormal gene regulation and are described in the next sections.

2.5.1 RNA-seq Differential Analysis

In the absence of tissue-matched control RNA-seq samples for each patient—which in many cases is not feasible to obtain—each patient's RPKM values are compared to a pooled sample created by combining the other patients'

RPKM values. Differential analysis of RPKM-normalized read counts is performed using CyberT [11] which was recently upgraded to handle both DNA microarray and RNA-seq data [12]. A confidence in the Bayesian prior of 3 is used instead of the default of 10 within CyberT to estimate the variance in gene expression. Rather than use strict p-value cutoffs, the top 5 percent most significantly over- or under-expressed genes, as well as the top 5 percent least significantly changing genes, are retained for down-stream analysis. The sizes of each of these gene lists are summarized under the gene expression column in Table 2.

2.5.2 Variant Transcription Factors

Transcription factors have been shown to have a large role in tumor progression, as evidenced by a large number of transcription factors that are known tumor suppressors. We identify potentially important affected transcription factors by making use of the predicted TFBS described in Section 2.3.6. For each transcription factor, we determine the number of binding sites predicted within 3 kb upstream and 1 kb downstream of the transcription start site of all transcripts in the human genome. We compare these counts to those within the same distance to genes in each of the following three lists:

1) The top 5 percent under-expressed genes in the patient vs. other patient RNA-seq differential analysis
2) The top 5 percent over-expressed genes in the patient vs. other patient RNA-seq differential analysis
3) The top 5 percent differential genes in the control vs. tumor microarray data obtained for this patient's cancer type, as described in Section 2.7.3.

We use a Fisher's Exact test to determine significance of the number of binding sites within the above lists, as compared to the 36,742 transcripts annotated in the human genome, and subsequently rank transcription factors by p-value. For each enriched transcription factor with p-value less than 0.05, we determine if the protein for that transcription factor is affected by any small or large variations or has abnormal gene expression for that patient. This results in lists of approximate 0-20 variant transcription factors per patient. In conjunction with the expression of the putative targets of these factors, we can identify what are likely causal relationships between over- or under-expression of certain factors and subsequent over- or under-expression of their targets.

2.5.3 Tissue Specific Expression

The Human U133A Gene Atlas data set [13] is obtained from BioGPS [14] to be used as a measure of normal tissue expression for the tissues most similar to the tumor sample obtained in each patient. This determines a baseline gene expression profile in healthy tissue to be used as a control. This dataset contains GCRMA values as a result of normalizating the microarray samples obtained from 79 human tissues. Combining these with the RPKM values from the RNA-seq analysis, we generate profiles of gene expression in (1) all patient tumor tissue samples and (2) all of the matched normal tissue samples, in order to identify abnormal patterns of expression in patients, i.e., those that would not be expected due to normal differences between tissues from which tumors were obtained.

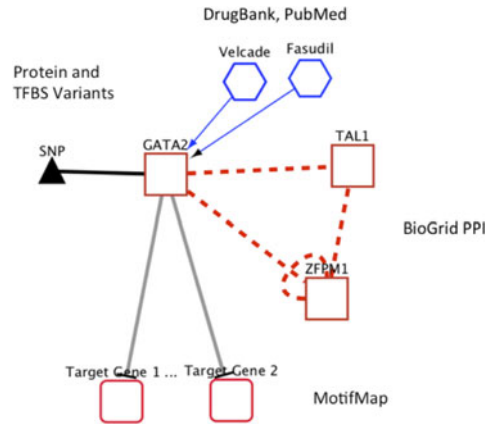

Fig. 2. Example of a network with drug, interaction, and transcription databases used to relate transcripts to each other and to potential drugs.

2.6 Network Analysis

It has been shown that cancer cells share in common multiple acquired capabilities that enable the cell to proliferate uncontrollably. These hallmarks of cancer have been highlighted previously [15], [16] and show a wide range of known pathways to be affected across different types of cancer. To visualize the connections between affected genes for each patient within known pathways—as well their connections to unaffected proteins—networks are created using in-house software which are then rendered in a web browser using CytoscapeWeb [17].

In order to initialize networks with proteins related to specific pathways, 478 known pathways are downloaded from KEGG Pathways [18] and the NCI Pathway Interaction Database [19]. Subsequently, transcription factor (TF)-DNA, TF-TF, protein-protein edges are added to the network based on the publicly available datasets from MotifMap [2], [3] and BioGRID [20], respectively. Variants on proteins, as well as the proteins identified as differential in the microarray and RNA-seq analyses, are used to highlight portions of the network and help visually interpret the biological role of the mutations. Variant TFBS are visualized by highlighting the edges between transcription factors and the genes that contain a site for that factor within its promoter. Further, drug-protein interactions are added to the network, as described in the next section. Taken together, this network approach assists in investigating potential driver mutations with a focus on identifying potential drug candidates and their targets. A simplified example of such a network is shown in Fig. 2.

2.6.1 Drug Targets

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

- DrugBank [21], [22], [23]

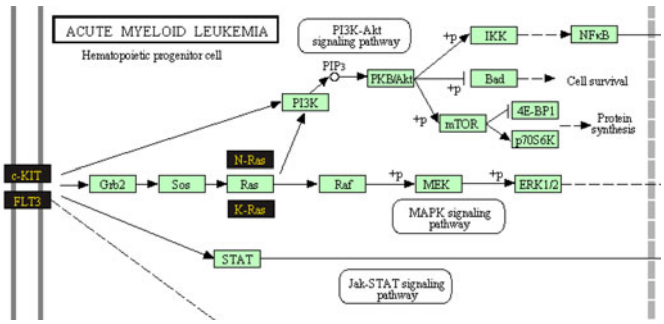

Fig. 3. Portion of KEGG AML pathway for patient CHOC36. Dark boxes indicate proteins with clinically associated variations.

- BindingDB [24]
- PharmGKB [25].

Each database provides an orthogonal set of annotations from which one can infer potential attenuation of known drug-effect, or perhaps novel drug interaction. Additional drug and drug-target information were also incorporated using semantic web resources for open drug data. These include Bio2RDF [26], [27], Chem2Bio2RDF [28], and Linked-Open Drug Data (LODD) [29]. Fig. 3 shows the original AML pathway from the KEGG database. Fig. 4 shows the corresponding auto-generated drug-target network used in exploring potential therapeutic targets.

To assist in identifying potential therapeutic drugs, we use a network-based approach which leverages the auto-generated networks for all pathways. For any gene target, we identify which KEGG or NCI pathways it is present in, and perform a breadth-first search starting at the gene target until we find a drug with an affected gene target. Additionally, if multiple such drug-targets exist at the same distance from our initial target, we choose the drug that targets the most genes, with preference given to drugs with a greater number of affected targets. Fig. 4 shows the set of drugs reached by this method via a search originating from each of the affected genes in the AML pathway for one patient. Using the pathway ranking method described in Section 2.8, we additionally prefer drugs obtained from the top ranked pathways that contain our gene target.

2.7 Cancer Specific Analysis

Without context, calling variants within a patient's genome is not enough to identify the most relevant mutations in a patient. Variants provided by commercial solutions such as Complete Genomics or Illumina, Inc., or by open-source pipelines such as VarScan2 [30], do not solve the problem of ranking the most important genic mutations. Rather, they rank the most confident of such mutations, and in most cases an overwhelming number of potential driver mutations are identified. Solving this problem requires us to perform a few steps in our pipeline that are specific to each type of cancer, in order to provide a context in which to identify the most affected genes for each patient.

2.7.1 Curated Gene Lists

To narrow down our variants to those contained within genes known to be involved in a certain type of cancer, gene lists are automatically curated from three primary sources.

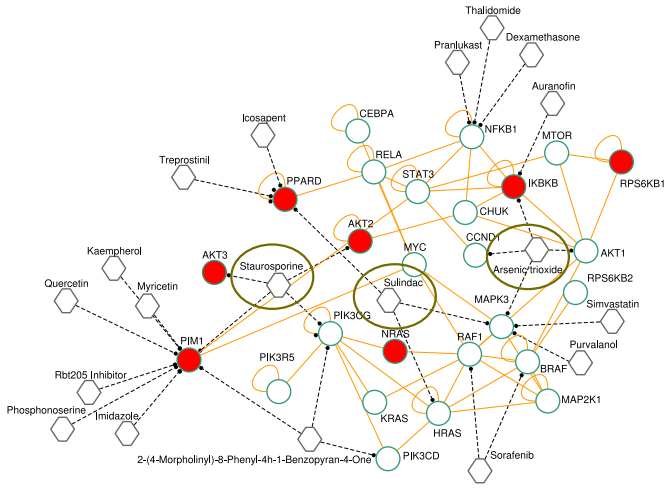

Fig. 4. CHOC36 drug-target edges in AML pathway (limited to variants). Circles denote proteins and hexagons denote drugs. Filled circles denote affected proteins, with identified potentially therapeutic drugs circled.

These three sources are (1) NCBI MEDLINE abstract and titles, (2) NCBI GeneRIF [31], and (3) NCBI Entrez queries.

For the first source, we perform text pattern matching on the corpus of abstracts and titles from NCBI Medline, using the UCSC hg19 genome annotation tables for a list of all known gene symbols in our search. Using the PubMed API, we retrieve a list of articles matching a specific type of cancer and extract all gene symbols in the titles and abstracts of these articles. Secondly, we cross reference the same articles with NCBI GeneRIF in order to find all genes that have been manually annotated for these articles. NCBI GeneRIF contains 800,000 gene symbols annotated to MEDLINE articles, 477,417 of which are for Homo sapiens. Thirdly, the NCBI Entrez web API is used to return a list of genes for any query related to each type of cancer. The final sizes of each of our curated gene list for each type of cancer are shown in Supplementary Table 3, available online.

Additionally, lists of genes which are known to be affected in or related to cancer are pulled from three public sources, in order to create a list of genes commonly implicated in a wide range of cancers. These sources, along with the number of symbols in each, are: (1) The Bushman Lab Cancer Gene List [32] (2,032), (2) The Cancer Gene Census [33] (489), and (3) Network of Cancer Genes 3.0 [34] (1,495). In total, 450 genes symbols were found in all three sources, across the 2,916 genes present in at least one source.

2.7.2 Microarray Data Sets

Microarray datasets are automatically obtained from the Gene Expression Omnibus (GEO) [35] for the cancer types of our pediatric patient samples using the cancer type as a query. Control datasets and samples for each cancer type are also obtained if available. These control datasets are used in lieu of proper control sample RNA-seq datasets for each patient, which are not realistically obtainable for the pediatric patients being sequenced. We use these control datasets in the methods below to complement the RNA-seq differential analysis in such cases. The datasets, platforms, and number of samples per type of cancer are shown in Supplementary Tables 2 and 5, available online.

2.7.3 Microarray Differential Analysis

Following standard microarray practices, each microarray dataset obtained is background normalized using the MAS5 algorithm [36]. Using platform annotations provided by GEO, probes are matched to gene symbols, and for cases where there are multiple probes per gene symbol, the probe with the maximum expression is retained. In the cases where raw expression is available, expression values are log-normalized to correct for the variance-mean bias commonly observed in microarray data. For any preprocessed datasets where raw microarray data is not available, data is log normalized if it was not already, to keep scales consistent.

After all datasets for all types of cancer present within our patients are preprocessed, gene symbols are then matched across all microarray samples. After removing samples and symbols that are missing more than 75 percent of data, 17,011 unique gene symbols remain, for which any missing data is imputed using k-nearest neighbors. Lastly, quantile normalization is used to normalize between all arrays and the distribution of expression values across tumor types is shown in Supplementary Fig. 6, available online. This preprocessing step is performed again if any additional patients with distinct types of cancer are obtained.

To test for differential transcripts, CyberT [11], [12] with a Benjamini-Hochberg multiple test corrected p-value cutoff of 0.05 is performed on a number of different contrasts utilizing the microarray data. In particular, when control samples exist for a type of cancer, CyberT is used to identify differentially expressed transcripts specific to that type of cancer which can be used to prioritize (1) variants within those transcripts or (2) those transcripts that were also identified by the RNA-seq analysis differential analysis in patients afflicted with that cancer.

CyberT is additionally used to perform the exact same analysis as is done using the pooled cancer patient RNA-seq samples described in Section 2.5.1. In lieu of the patient samples, the median expression values for each gene symbol are used across all microarray samples for that type of cancer. The median microarray sample for each patient is tested for differential expression against the set of median microarray samples derived for each other patient as was done for the RNA-seq data. Additionally, in the types of cancer where control data is available, we perform the same differential analysis for all patients with that type of cancer using the median control microarray data instead of the median tumor microarray data. Lastly, using all of the tumor microarray data for all types of cancer, we use CyberT to identify transcripts that are commonly expressed, or unexpressed, in cancer. In summary, we define the following gene lists using microarray data:

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

2.7.4 Gene List Overlaps

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

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

The significance of the last list above prompted us to priortize our RNA-seq differential transcripts using a similar gene list overlap approach, since this overlap was found to enrich for transcripts with small variants—which likely influences the expression of those affected genes. For gene lists (2) and (3) from our microarray analysis in Section 2.7.3, we can identify tissue-specific genes and genes we would expect to change, respectively, in the RNA-seq analysis against the pooled patient samples for patients with that type of cancer. These help further prioritize the differential RNA-seq transcripts into `HIGH`, `MEDIUM`, and `LOW` gene lists based on overlaps with microarray gene lists (1), (3), and (2), respectively, where the `HIGH` category corresponds to the same overlap we found a significant enrichment of small variants in. The average size of these lists are summarized in Table 2.

2.8 Reporting

A novel approach to reporting is developed in order to sift through the still many affected genes found in each patient, despite having removed variations present in the germline control samples. When looking at the overlap with the 487 KEGG and NCI pathways in our network analysis, on average 342 pathways had at least one small variation per patient. To reduce this to a more clinically relevant and manageable number of affected pathways, it is necessary to limit pathways based on their importance in a each specific type of cancer. After doing so, we build networks for the pathways most affected in each patient in order to visualize the interactions between genes with variations within the same pathway and to identify potential drug candidates.

In order to filter genes with variations down to the ones most probable to contain driver mutations, we develop a ranking method for both pathways and genes based on enrichment scores. Using the list of curated genes described previously—those specific to a type of cancer—we use a Fisher's exact test to determine the statistical significance of the overlap between the curated gene list and the list of genes in each pathway. Pathways are then ranked based on this significance value. This ranking is

specific to each type of cancer but not specific to any individual patient. Additionally, for each patient, the ranked pathways for their type of cancer are filtered for only those pathways containing at least one genetic variation (small or large) within the curated list of genes for that cancer. In most patients, this reduces the average affected pathways from 342 down to less than 50 affected pathways. Specifically, we compute the pathway enrichment p-value as the probability of observing the overlap between the pathway gene list and our curated gene lists, assuming 39,131 genes in the human genome:

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

Similarly, for each patient, we compute an enrichment score for any single gene based on the list of variants which are affecting that gene. The enrichment score of each individual type of variant listed in Table 2 (under the small variations, large variations, and gene expression columns) is determined using the overlap of variants of a particular type within the table with the list of curated genes for that patient's cancer, calculated using a Fisher's Exact test:

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

To justify this approach, assuming the curated lists reflect the genes we expect to be mutated in patients with this type of cancer, we should observe more variations within this list than in a random gene list of the same size. As remarked on in Section 2.7.4, we find this to be the case across patients. Types of genetic variations that score higher will be ones that contain a larger number of affected genes within the curated list, and therefore we might expect our driver mutations to be carried by the same categories of mutations in those genes, and others, within the same patient.

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

Lastly, using the list of 450 symbols common to most cancers as was defined in Section 2.7.1, we look for affected genes within this more general list that rank highly but are not contained within the curated list of genes. These are genes that have been implicated in any type of cancer. Therefore, any affected genes within this list warrant further consideration aside from those in our curated gene lists for each type of cancer. Our final reports are in the form of network views of the top ranked pathways, along with tables of the top ranked genes along with their associated pathways, drug candidates, and expression profiles.

TABLE 3
Top 10 Ranked Pathways for CHOC23 (AML)

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

3 Interesting Findings

3.1 Patient CHOC23 Acute Myeloid Leukemia (AML)

To demonstrate the effectiveness of our pipeline in identifying genes affected by likely driver mutations, we explore the ranking observed for one of our patients with acute myeloid leukemia, CHOC23. Our objective is to identify the affected genes within the tumor genome of CHOC23 that most directly relate to AML. We employed the ranking method described previously to rank the pathways that would be of most interest in AML, the results of which are presented in Table 3. The top three pathways for this patient were PI3K-Akt signaling pathway, Chronic myeloid leukemia, and Acute myeloid leukemia. This initial ranking of pathways is not specific to this patient and is shared with all AML patients. As we would expect, the leukemia pathways for CML and AML rank near the top.

The gene ranking method also performs well for this patient, and in contrast to the pathway ranking, is specific to this patient. As shown in Table 4, this method ranks MLL3 the highest. The score for MLL3 is calculated as follows:

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

The higher value for Score(Deletion) reflects the fact that, in this patient, deletions are more enriched within the curated list of genes for AML than any of the other variations. For the majority of patients, the genetic variations that score highest are: (1) microarray differential genes, (2) fusions, (3) deletions, and (4) genes with lower RPKM compared to other patients. The fact that this gene ranks at the top is of no small consequence, it is one of the few identified Mitelman fusions in all of the patients. The other mutations identified provide further evidence that MLL3 is significantly altered in this patient, and contribute to it being the highest ranked.

TABLE 4
Top 10 Curated Gene Ranking for CHOC23 (AML)

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

We make use of our automated method for drug recommendations to address the problem of a lack of directly druggable targets. For this patient, none of the top 10 ranking curated genes have any drugs that directly target them, making therapeutic recommendations for these genes problematic. To address this, we search over each of the top ranked genes and identify which of the top ranked pathways, if any, that gene is contained within. If one or more pathways exists, we perform a graph-based search for the nearest best drug candidates, as described earlier in Section 2.6.1.

For this patient, the first such targets with drug candidates are TCF7L1 and NCOR2. In the absense of directly druggable targets, we find that TCF7L1 has two potential drug candidates. The first of which, Staurosporin—a potent protein kinase inhibitor—is identified in the Prostate Cancer pathway through TCF7L1's interaction with CTNNB1, which interacts with Staurosporin's direct target GSK3B. Interestingly, we find a number of additional targets for Staurosporin in the AML pathway (shown in Fig. 4 for a different AML patient)—AKT2, AKT3, KIK3CG, and PIM1, all containing genetic variants within this patient. The second drug candidate, Vorinostat, a HDAC1 inhibitor, is identified in the Regulation of β-catenin pathway, again through TCF7L1's interaction with CTNNB1, which interacts with HDAC1.

We also identify Vorinostat as the best drug candidate for our next highest ranked gene, NCOR2, through an entirely different pathway, the Notch signalling pathway where HDAC1 and NCOR2 have a protein-protein interaction. This drug has been in Phase II clinical trials for AML patients, and while it was not shown to be effective alone, it shows promise as a potential drug for high-risk patients in conjunction with other drugs [37]. It may be the case that only a subset of patients, such as this one, would respond to this drug. In the absence of any direct drug candidates in the top ranked genes, we are able to identify reasonable drug candidates through this pathway-based approach.

3.2 Patient CHOC33 (Neuroblastoma)

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

Our second ranked gene, PARK2, contains a deletion in the first intron as well as an exonic region of higher copy number, as shown in Fig. 5, which contributes to its high rank. PARK2 is also identified as having a relatively higher expression in this patient compared to others (Fig. 6). The gene expression profile provides strong evidence that PARK2 gene expression is being altered as a result of its genetic variants. This lends credibility to the called genetic variants, as well as informing on the direction of change of PARK2 expression in this patient's tumor. Additionally, by overlapping Pfam predicted domains for PARK2, we identified a portion of the ubiqitin domain that confers PARK2 a role in the ubiqitin-ligase pathway (Fig. 5). This further suggests that PARK2 is functioning in tumor progression in this patient. Recently, PARK2 has been shown to have an emerging role in cancer [38].

3.3 Patients CHOC36 and CHOC03 (AML)

We perform a meta analysis on our two primary AML patients, CHOC36 and CHOC03, in which we attempt to find genes that had common variations: either genetic or in their expression. One such gene is EPOR, which stood out as having significantly higher expression in both patients compared to other patients' tumors (Fig. 7). Despite a known higher expression in healthy bone marrow as compared to other tissues (data not shown), the level of EPOR expression observed for these two patients is not observed

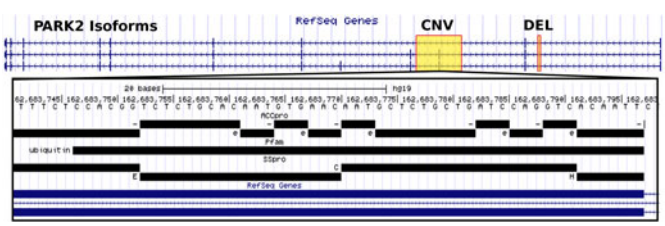

Fig. 5. UCSC genome browser displaying genetic variations in PARK2 transcripts in patient CHOC33. Zoomed in region highlights features of exon 2 within the copy number variant, which includes a functional ubiquitin domain identified by Pfam.

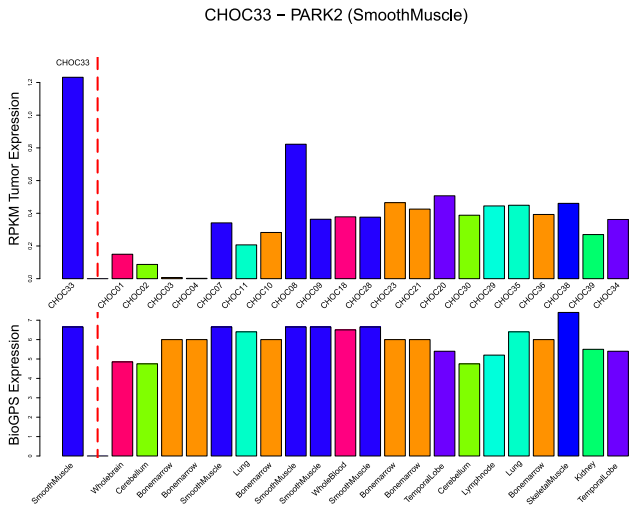

Fig. 6. PARK2 gene expression in patient tumors (top) and BioGPS normal tissue gene expression in tumor-matched tissues (bottom), where patient CHOC33 is the first bar on far left. We observe higher expression for PARK2 in CHOC33 as compared to other patients, in contrast to a relatively constant gene expression across healthy tissues.

in other patients for which the RNA-seq data was also obtained from the patient's bone marrow.

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

While not differentially expressed between our curated AML control vs. tumor microarray samples ($P = 0.6354$), it has been shown previously that in approximately 60 percent of AML patients, EPOR is unexpressed [39]. Additionally, remission times for patients with higher EPOR expression is significantly lower compared to those without EPOR expression [40] which is likely the case for these patients since all of the patients sequenced are after recurrence of the primary tumor. Additionally, in some cases patients with AML are being treated with erythropoiesis-stimulating agents, but it is believed that this could cause proliferation in a subset of AML patients with EPOR expression [39], [41], suggesting that these AML patients fall into a specific subtype of AML, differentiating them from our secondary AML patients CHOC23 and CHOC26, for which we do not observe an increase in EPOR expression.

Additionally, we investigated what transcription factors were enriched in each of our AML patients based on the location of predicted binding sites upstream of our differential

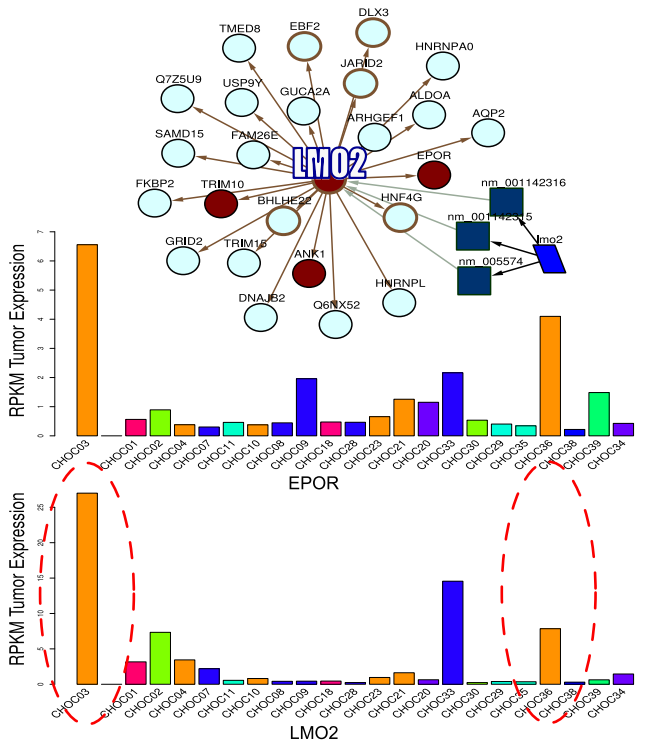

Fig. 7. Enriched transcription factor LMO2, along with three of 25 of its predicted targets (top; dark circles for genes with high expression), has high expression compared to other patients for both of our primary AML patients (bottom; primary AML patients circled). One of these targets, EPOR, was identified as being to be significantly higher in both primary AML patients and not others (middle).

gene lists (see Section 2.5.2). Further validating the importance of EPOR in these patients, we identify a significant enrichment for transcription factor LMO2 in our list of over-expressed transcripts (rank #3 for CHOC03; p-value = 4.5E-5). LMO2 and three out of its 25 targets predicted by MotifMap (EPOR, ANK1, and TRIM10) all have high expression in this patient (Fig. 7). LMO2 has been previously found to be involved in AML [42], and its high expression, particularly in CHOC03, compared with other patients is further evidence of a subtype of AML within our primary AML patients.

4 Discussion

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

Our pipeline adheres to the five steps of a cancer analysis pipeline outlined by Valencia and Hidalgo [43]: 1. Genome analysis: We analyze DNA-seq and RNA-seq data from commercial vendors using a uniformed format for calling variants. 2. Consequences of mutations and genomics alterations: For small variations we identify the affect on protein sequence in addition to protein secondary structure, solvent accessibility, and known protein domains. 3. Network level analysis: We make use of NCI and KEGG pathways to identify the most relevant pathways for each type of cancer. 4.

Drug: We make use of the ranked pathways and genes to identify potential drug candidates for each patient. And lastly, 5. Collaborative interfaces: We integrate multiple sources of information into a network view that includes regulatory information across all patients and tissue types for exploring the interactions among affected genes and potential drugs.

In contrast to other published pipelines [44], our pipeline successfully integrates expression data into our ranking, in addition to giving priority to mutated variant transcription factors. A recent opinion paper [45], highlights the importance of the integration of multiple-omic datasets, which we have demonstrated.

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

5 Conclusion

By combining RNA-seq, DNA-seq, and microarray data, in addition to numerous sources of annotations on the reference genome, we were able to identify likely driver mutations in pediatric cancers. We found that such an integrative approach is essential, and information from gene expression data in particular, can complement a search for genetic variants, making results more robust. Typically, we observe many mutations within the top ranked pathways, indicating that multiple genes are likely affected in a tumor cell in order to effectively knockout critical pathways, as shown to be required in cancer [15], [16].

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

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

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

Acknowledgments

Michael Zeller and Christophe N. Magnan contributed equally to this work. This work was supported by a grant from the Hyundai Foundation to L.S. and grants NIH LM01, NIH NLM T15 LM07, and US National Science Foundation (NSF) IIS-0513376 to PB. The authors acknowledge also the support of the CHOC, the UCI Institute for Genomics and Bioinformatics, the UCI Genomics High-Throughput Facility, and a hardware donation by NVIDIA. Additional support of their computational infrastructure was provided by Jordan Hayes and Yuzo Kanomata.

References

[1] (2013, Sep.). The cancer genome atlas homepage [Online]. Available: http://cancergenome.nih.gov/

[2] X. Xie, P. Rigor, and P. Baldi, "MotifMap: A human genome-wide map of candidate regulatory motif sites," *Bioinformatics*, vol. 25, no. 2, pp. 167–174, Jan. 2009.

[3] K. Daily, V. Patel, P. Rigor, X. Xie, and P. Baldi, "MotifMap: Integrative genome-wide maps of regulatory motif sites for model species," *BMC Bioinformatics*, vol. 12, no. 1, p. 495, 2011.

[4] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, and P. A. Futreal, "Cosmic: Mining complete cancer genomes in the catalogue of somatic mutations in cancer," *Nucleic Acids Res.*, vol. 39, no. suppl. 1, pp. D945–D950, 2011.

[5] A. G. Knudson. (1971, Apr.). Mutation and cancer: Statistical study of retinoblastoma. *Proc. Nat. Acad. Sci. USA* [Online]. *8(4)*, pp. 820–823. Available: http://www.pnas.org/content/68/4/820.abstract

[6] L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, B. J. Raney, A. Pohl, V. S. Malladi, C. H. Li, B. T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R. A. Harte, M. Haeussler, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, T. R. Dreszer, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent, "The UCSC Genome Browser database: extensions and updates 2013," *Nucleic Acids Res.*, vol. 41, no. database issue, pp. D64–D69, Jan. 2013.

[7] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi, "SCRATCH: A protein structure and structural feature prediction server," *Nucleic Acids Res.*, vol. 33, no. suppl. 2, pp. W72–W76, Jul. 2005.

[8] C. N. Magnan and P. Baldi, "SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity," *Bioinformatics*, 2014.

[9] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn, "The Pfam protein families database," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D290–D301, Jan. 2012.

[10] F. Mitelman, B. Johansson, and F. Mertens, Eds. (2013). Mitelman database of chromosome aberrations and gene fusions in cancer. [Online]. Available: http://cgap.nci.nih.gov/Chromosomes/Mitelman

[11] P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, no. 6, pp. 509–519, Jun. 2001.

[12] M. A. Kayala and P. Baldi, "Cyber-T web server: Differential analysis of high-throughput data," *Nucleic Acids Res.*, vol. 40, no. web server issue, pp. W553–W559, Jul. 2012.

[13] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch, "A gene atlas of the mouse and human protein-encoding transcriptomes," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 16, pp. 6062–6067, Apr. 2004.

[14] C. Wu, I. Macleod, and A. I. Su, "BioGPS and MyGene.info: Organizing online, gene-centric information," *Nucleic Acids Res.*, vol. 41, no. database issue, pp. D561–D565, Jan. 2013.

[15] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, Jan. 2000.

[16] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011.

[17] C. T. Lopes, M. Franz, F. Kazi, S. L. Donaldson, Q. Morris, and G. D. Bader, "Cytoscape web: An interactive web-based network browser," *Bioinformatics*, vol. 26, no. 18, pp. 2347–2348, Sep. 2010.

[18] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The KEGG resource for deciphering the genome," *Nucleic Acids Res.*, vol. 32, no. suppl. 1, pp. D277–D280, Jan. 2004.

[19] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow, "PID: The pathway interaction database," *Nucleic Acids Res.*, vol. 37, no. database issue, pp. D674–D679, Jan. 2009.

[20] C. Stark, B.-J. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: A general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, no. database issue, pp. D535–D539, Jan. 2006.

[21] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart, "DrugBank 3.0: A comprehensive resource for 'omics' research on drugs," *Nucleic Acids Res.*, vol. 39, no. database issue, pp. D1035–1041, Jan. 2011.

[22] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "DrugBank: A knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Res.*, vol. 36, no. database issue, pp. D901–906, Jan. 2008.

[23] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "DrugBank: A comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Res.*, vol. 34, no. database issue, pp. D668–672, Jan. 2006.

[24] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities," *Nucleic Acids Res.*, vol. 35, no. database issue, pp. 198–201, Jan. 2007.

[25] M. Whirl-Carrillo, E. M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C. F. Thorn, R. B. Altman, and T. E. Klein, "Pharmacogenomics knowledge for personalized medicine," *Clin. Pharmacol. Therapeutics*, vol. 92, no. 4, pp. 414–417, Oct. 2012.

[26] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2rdf: Towards a mashup to build bioinformatics knowledge systems," *J. Biomed. Inf.*, vol. 41, no. 5, pp. 706–716, 2008.

[27] A. Callahan, J. Cruz-Toledo, and M. Dumontier. (2013). Ontology-based querying with bio2rdf's linked open data. *J. Biomed. Semantics* [Online]. *4 (suppl. 1)*, p. S1. Available: http://www.jbiomedsem.com/content/4/S1/S1

[28] B. Chen, X. Dong, D. Jiao, H. Wang, Q. Zhu, Y. Ding, and D. Wild,. (2010). Chem2bio2rdf: A semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics* [Online]. *11 (1)*, p. 255. Available: http://www.biomedcentral.com/1471-2105/11/255

[29] M. Samwald, A. Jentzsch, C. Bouton, C. Kallesoe, E. Willighagen, J. Hajagos, M. Marshall, E. Prud'hommeaux, O. Hassanzadeh, E. Pichler, and S. Stephens. (2011). Linked open drug data for pharmaceutical research and development. *J. Cheminformatics* [Online] *3(1)*, p. 19. Available: http://www.jcheminf.com/content/3/1/19

[30] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson, "VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing," *Genome Res.*, vol. 22, no. 3, pp. 568–576, Mar. 2012.

[31] J. A. Mitchell, A. R. Aronson, J. G. Mork, L. C. Folk, S. M. Humphrey, and J. M. Ward. (2003). Gene indexing: characterization and analysis of NLM's GeneRIFs. *Proc. AMIA Annu. Symp.*, pp. 460–464 [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480312/

[32] F. Bushman. (2013, Sep.) Bushman lab: Genelists. [Online]. Available: http://www.bushmanlab.org/links/genelists

[33] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes," *Nat. Rev. Cancer*, vol. 4, no. 3, pp. 177–183, Mar. 2004.

[34] M. D'Antonio, V. Pendino, S. Sinha, and F. D. Ciccarelli, "Network of cancer genes (NCG 3.0): Integration and analysis of genetic and network properties of cancer genes," *Nucleic Acids Res.*, vol. 40, database issue D1, pp. D978–D983, Jan. 2012.

[35] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: Archive for functional genomics data sets update," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D991–D995, Jan. 2013.

[36] S. Pepper, E. Saunders, L. Edwards, C. Wilson, and C. Miller, "The utility of MAS5 expression summary and detection call algorithms," *BMC Bioinformatics*, vol. 8, no. 1, p. 273, 2007.

[37] E. W. Schaefer, A. Loaiza-Bonilla, M. Juckett, J. F. DiPersio, V. Roy, J. Slack, W. Wu, K. Laumann, I. Espinoza-Delgado, S. D. Gore, and Mayo P2C Phase II Consortium. (2009, Oct.). A phase 2 study of vorinostat in acute myeloid leukemia. *Haematologica* [Online]. *94 (10)*, pp. 1375–1382, Oct. 2009. Available: http://view.ncbi.nlm.nih.gov/pubmed/19794082

[38] L. Xu, D.-c. Lin, D. Yin, and Koeffler,, "An emerging role of PARK2 in cancer," *J. Mol. Med.*, vol. 92, no. 1, pp. 31–42, 2014.

[39] G.-L. L. Cheng, W. Wang, H.-Y. Y. Wang, and Z.-G. G. Cui. (2011, Feb.). Expression of EPOR on acute leukemia cells and its clinical significance. *J. Exp. Hematol./Chinese Assoc. Pathophysiol.* [Online]. *19(1)*, pp. 15–18. Available: http://view.ncbi.nlm.nih.gov/pubmed/21362213

[40] A. Takeshita, K. Shinjo, K. Naito, K. Ohnishi, M. Higuchi, and R. Ohno, "Erythropoietin receptor in myelodysplastic syndrome and leukemia," *Leukemia Lymphoma*, vol. 43, no. 2, pp. 261–264, Feb. 2002.

[41] M. Feng and Y.-C. C. Li, "Expression of erythropoietin receptor in leukemia cells and relation of erythropoietin level with leukemic anemia," *J. Experimental Hematol./Chinese Assoc. Pathophysiol.*, vol. 16, no. 6, pp. 1265–1270, Dec. 2008.

[42] U. Cobanoglu, M. Sonmez, H. M. M. Ozbas, N. Erkut, and G. Can. (2010, Jun.). The expression of LMO2 protein in acute B-cell and myeloid leukemia," *Hematology* [Online]. *15(3)*, pp. 132–134. Available: http://view.ncbi.nlm.nih.gov/pubmed/20557670

[43] A. Valencia and M. Hidalgo, "Getting personalized cancer genome analysis into the clinic: The challenges in bioinformatics," *Genome Med.*, vol. 4, no. 7, p. 61, Jul. 2012.

[44] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson, and L. Ding, "MuSiC: Identifying mutational significance in cancer genomes," *Genome Res.*, vol. 22, no. 8, pp. 1589–1598, Jul. 2012.

[45] E. R. Mardis, "Genome sequencing and cancer," *Current Opinion Genetics Develop.*, vol. 22, no. 3, pp. 245–250, Jun. 2012.

Michael Zeller received the BS degrees in computer science and in biomedical engineering from Washington University in St. Louis in 2008. He is currently working toward the PhD degree at the University of California, Irvine, under Prof. Pierre Baldi in the Computer Science department. His research focuses on machine learning using gene expression data and high-throughput sequencing data analysis.

Christophe N. Magnan received the PhD degree from the University of Provence, France, in 2007. He is currently an assistant project scientist at the Institute for Genomics and Bioinformatics, University of California, Irvine. His work focuses on machine learning applications to proteomics and genomic sequencing data analysis.

Vishal R. Patel received the BTech degree in industrial biotechnology from Anna University, Chennai, India, in 2009 and the PhD degree in computer science from the University of California, Irvine, in 2014. His research focuses on data mining and large scale data analysis.

Paul Rigor received the BS degrees in neuroscience and in cognitive science with computing specialization from the University of California, Los Angeles. He is currently working toward the PhD degree under the supervision of Prof. Pierre Baldi in the Donald Bren School of Information and Computer Sciences. His research focuses on machine learning applications for high-throughput sequence analyses leveraging high-performance computing resources.

Leonard Sender received the MD degree from the University of the Witwatersrand, Johannesburg, South Africa, in 1982 followed by a pediatrics internship and residency at UC Irvine Medical Center. His pediatrics hematology/oncology sub specialty training included Childrens Hospital of Los Angeles. He is currently a clinical professor of Medicine at the UCI School of Medicine, director of the Adolescent and Young Adult (AYA) Cancer Program at the Children Hospital of Orange County (CHOC), director of Clinical Operations and Program Development and division chief, Pediatric Oncology, at the UCI Medical Centers Chao Family Comprehensive Cancer Center.

Pierre Baldi received the PhD degree from the California Institute of Technology. He is currently the chancellors professor at the Department of Computer Science, director of the Institute for Genomics and Bioinformatics, and an associate director of the Center for Machine Learning and Intelligent Systems at the University of California, Irvine. His research focuses on artificial intelligence and statistical machine learning and their applications to problems in the life sciences, particularly in chemoinformatics, proteomics, genomics, and systems biology. He is credited with pioneering the development and application of graphical models and deep neural networks in bioinformatics and chemoinformatics and has published more than 250 research articles and four books. He is a fellow of the AAAI, the AAAS, the ACM, the IEEE, and the ISCB and received the 2010 Eduardo Caianiello Prize for research in neural networks.