Genomic and machine-learning analysis of germline variants in cancer

Chioma Madubata

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Genomic and machine-learning analysis of germline variants in cancer

Chioma Madubata

Cancer often develops from specific DNA alterations, and these cancer-associated mutations influence precision cancer treatment. These alterations can be specific to the tumor DNA (somatic mutations) or they can be heritable and present in normal and tumor DNA (germline mutations). Germline variants can affect how patients respond to therapy and can influence clinical surveillance of patients and their families. While identifying cancer-associated germline variants traditionally required studying families with inherited cancer predispositions, large-scale cancer sequencing cohorts enable alternative analysis of germline variants.

In this dissertation, we develop and apply multiple strategies for analyzing germline DNA from cancer sequencing cohorts. First, we develop the Tumor-Only Boosting Identification framework (TOBI) to learn biological features of true somatic mutations and generate a classification model that identifies DNA variants with somatic characteristics. TOBI has high sensitivity in identifying true somatic variants across several cancer types, particularly in known driver genes. After predicting somatic variants with TOBI, we assess the identified somatic-like germline variants for known oncogenic germline variants and enrichment in biological pathways. We find germline and somatic variants inactivating the Fanconi anemia pathway in 11% of patients with bladder cancer. Finally, we investigate germline, diagnosis, and relapse variants in a large cohort of patients with pediatric acute lymphoblastic leukemia (ALL). Our somatic analysis captures known ALL

driver genes, and we describe the sequential order of diagnosis and relapse mutations, including late events in *NT5C2.* We apply both the TOBI framework and guidelines American College of Medical Genetics and Genomics to identify potentially cancer-associated germline variants, and nominate nonsynonymous variants in *TERT* and *ATM*.

Contents

List of Figures

Acknowledgements

I thank my family for supporting me during this journey. My friends have been essential in navigating this part of my research career. Particular thanks goes to my favorite girls and the Quincy house crew, the Wacky Wednesday gang, d.li, the med school crew, and d@6. I also have to thank my prior research mentors, who continue to help me grow as a scientist, and my current collaborators at Columbia and elsewhere. Finally, a big thanks to the current and prior Rabadan Lab members, particularly Hossein, Jiguang, Daniel, Tim, Junfei, Rachel, Albert, and Raul.

To my mom, Juliet, and my dad, Christian

Introduction

Germline variants in cancer

Cancer reflects a group of diseases characterized by unconstrained cell replication and growth, evasion of cell death, and tissue invasion and metastasis [1]. Tumor cells have numerous genomic alterations, including whole chromosomal gains or losses, copy number variations, and single nucleotide variants (SNVs). These alterations in DNA (deoxyribonucleic acid) can be somatic (unique to the tumor), or germline (found in both tumor and matched normal DNA from the same patient, and transmissible from parent to offspring).

The relative contribution of heritable cancer risk varies depending on cancer type. Analysis of 44,788 pairs of twins from Swedish, Danish, and Finnish twin registries found that while monozygotic concordance was typically less than 0.1, heritable factors significantly contributed to prostate cancer (42% of risk explained; 95 % confidence interval 29-50%%), colorectal cancer (35 %; 95 % confidence interval 10-48 %), and breast cancer (27 %; 95 % confidence interval 4-41 %)[2]. Rare Mendelian cancer syndromes caused by alterations in specific genes do contribute to this heritability, but do not fully explain it. For example, Mendelian disorders associated with colorectal cancer (Familial

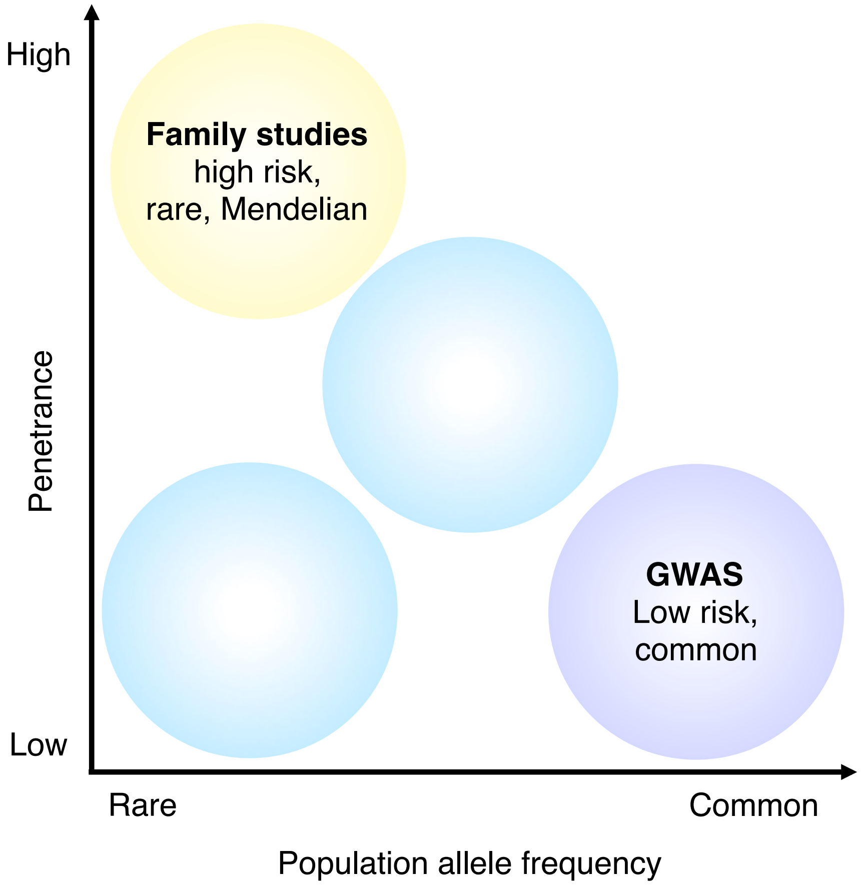


Figure 0.1: **Population allele frequency and penetrance of germline variants in cancer**. Blue circles indicate germline variants that have not traditionally been assessed in cancer. Modified from [4], Figure 2.

Adenomatous Polyposis, Lynch syndrome, *MUTYH*-associated polyposis, and rarer polyposis syndromes) only explain 5% of colorectal cancer [3], while ref. [2] estimated 35 % heritability. Common genetic variants as assayed by genome-wide association study (GWAS) and many case control studies explain only ~10% of the familial relative risk of cancer, suggesting that rare variants, potentially with greater effect sizes, explain some of the heredity[3]. Figure 0.1 depictrs the population frequency and penetrance of these germline variants.

Common variants

A GWAS analyzes whether common variants (minor allele present in greater than 5% of the population) are associated with a common trait [5]. Common loci are assayed throughout the genome to provide a relatively unbiased assessment of genomic contribution to a trait. Since the 2000s, GWAS studies have uncovered genetic associations with cancer predisposition and attributes. Multiple studies focused on pediatric acute lymphoblastic leukemia (ALL) identified common variants associated with ALL risk and treatment response[6]. For example, 6-mercaptopurine is a thiopurine and a standard component of pediatric ALL therapy, but has a narrow therapeutic index. One GWAS identified a coding and non-coding variant associated with patient drug tolerance in the gene *NUDT15* [7], which encodes the Nudix Hydrolase 15 that negatively regulates of thiopurine activation. While significant GWAS variants can be biologically or clinically informative, they also typically have small effect sizes indicating minimal contribution to disease risk.

Rare coding variants from familial studies

While common variants typically have a small effect size, rare variants with highly penetrant effects can cause inherited cancer predisposition. In 1913, Warthin described the heredity nature of carcinoma based on family case studies [8]; before that, the concept of familial cancer was controversial. Specific familial cancer predispositions include Li and Fraumeni's 1969 report of a high-frequency cancer predisposition syndrome (Li-Fraumeni syndrome) [9, 10] and Knudson's 1971 confirmation that certain retinoblastoma cases were inherited (familial retinoblastoma)[11]. Patients with these inherited cancer syndromes developed tumors at earlier ages than patients with sporadic tumors of the same type [11]. This earlier age of tumor development due to germline mutations suggests that many pediatric tumors are caused by germline driver variants, either inherited or de novo.

While clinical characteristics such as family history and early age of onset allowed for the identification of inherited cancers in the 1900s, beginning in the 1980s, improved molecular biology techniques allowed scientists to identify the genetic loci associated with these syndromes. Both tumor suppressors genes (TSG) and oncogenes cause inherited cancers. Germline variants in TSGs may be inherited in a single allele, with the TSG locus undergoing loss of heterozygosity (LOH) of the wild type allele in tumor tissue[11, 12]. LOH occurs when a formerly heterozygous locus loses a functional wildtype allele, possibly through deletion of the wildtype allele, mitotic recombination leading to two retained copies of the variant allele, or other mechanisms[13]. While some inherited oncogenes such as *MET*[14] (which causes renal cell carcinoma) and *RET*[15] (which causes Multiple Endocrine Neoplasia, types 2A/3) also exhibit LOH or copy number gain, these gains are rare compared to protein changing mutations. Other oncogenes such as *ALK* (neuroblastoma) have not shown LOH in inherited tumors[16]. The germline alterations causing these cancer-predisposition syndromes are rare in populations and highly penetrant, compared to more common alleles identified via GWAS[13]. Many causal genes in familial cancer syndromes are recurrently altered across cancer types, including *RB1* (the driver of inherited retinoblastoma[12]), *NF1* (neurofibromatosis, type 1[17]), and *TP53* (the cause of Li-Fraumeni syndrome [18]). Of note, certain germline variants that cause familial tumors occur as somatic driver mutations [19].

At least 100 genes have been associated to Mendelian cancer predisposition syndromes[3, 20], and genes associated with cancer predisposition continue to be identified as DNA sequencing technology improves. For example, recent studies using whole exome sequencing identified inherited predisposition to acute lymphoblastic leukemia via mutations in *PAX5*[21], *ETV6* [22–25], *IKZF1* [26], and *SH2B3* [27].

Rare coding variants from sporadic cancer cohorts

As DNA sequencing technology has improved, cohorts of sporadic cancers have been analyzed for rare germline variants[28]. For example, whole exome sequencing (WES) and whole genome sequencing (WGS) of hypodiploid ALL found that 43.3% of low-hypodiploid ALL with *TP53* mutations had germline *TP53* mutations, including mutations previously associated with Li-Fraumeni syndrome [28]. One patient in this cohort also had a germline *NRAS* p.Gly12Ser substitution. Sequencing of over 100 cases of early T-cell precursor acute lymphoblastic leukemia found recurrent somatic and germline mutations in the gene *ECT2L*[29]. Certain adult cancer studies have also assessed rare germline variants. One study of 429 ovarian carcinoma cases and 557 controls found germline truncations and deletions in *BRCA2* and other Fanconi pathway genes in 20% of cases[30].

Pan-cancer investigations of germline variants have found genes with pathogenic variants affected multiple cancer types. A study of 1,120 pediatric patients with different cancers found potentially pathogenic germline variants in 8.5% of patients[31], with recurrent variants in *TP53*, *APC*, *BRCA2*, *NF1*, *PMS2*, *RB1*, and *RUNX1*. This percentage only reflects variants in 60 genes known to cause familial cancers with autosomal dominant inheritance, and may underestimate the percentage of pediatric cancer patients with cancer-associated germline variants. Pan-cancer analysis of 4,034 cases cancer representing 12 cancer types from The Cancer Genome Atlas (TCGA) found that across 624 cancer-associated genes, the fraction of cases with cancer-predisposition variants in a cancer type ranged from 4% (acute myeloid leukemia) to 19% (ovarian), with frequent mutations in *ATM*, *BRCA1*, *BRCA2*, *BRIP1*, and *PALB2* [32]. Case-control study genotyping of 10 rare germline mutations in *PALB2*, *CHEK2* and *ATM* in patients with breast cancer (42,671 cases and 42,164 controls), prostate cancer (22,301 cases and 22,320 controls) and ovarian cancer (14,542 cases and 23 491 controls) found breast cancer risk associated with certain *PALB2*, *CHEK2* and *ATM* variants [33]. Specific *CHEK2* variants also associated with prostate cancer risk[33]. A recent study of 10,389 adult cases representing 33 cancer types found 8% of cases had a pathogenic variant in any of 152 cancer disposition genes, including variants with evidence of tumor LOH affecting genes *ATM*, *BRCA1*, and *NF1*[34]. While these cancer cohort studies have revealed that approximately 8% of patients with cancer have a pathogenic variant in a gene related to a cancer predisposition, an approach focused on previously identifed cancer-associated genes limits opportunities to identify new genes with cancer-associated germline variants.

Strategies to assess germline variant pathogenicity

An important issue in germline variant analysis is assessing the potential pathogenicity of a variant. Variants observed with high penetrance in a familial cancer syndrome or observed at high frequency in affected cases compared to controls have fairly high evidence of pathogenicity. Another strategy to assess a variant's association with cancer is experimental validation of aberrant function or a cancer phenotype. Unfortunately, many germline variants of interest are found in sporadic cases with no family history, and experimental validation of all germline variants remains unfeasible.

In silico or curated assessments of pathogenicity are increasingly used to prioritize variants in research and clinical care. Early curated assessments focused only on protein changing variants, particularly truncating variants, in known cancer genes. However, many germline variants in cancer genes are benign as seen from a study of 681 individuals without cancer who all exhibited protein changing variants in cancer genes [35]. Additionally, focusing on truncating variants in cancer genes does not improve identification of truly deleterious variants [36]. More sophisticated *in silico assessments* come from software that predicts variant effects based on amino acid conservation, protein structure, nucleotide structure, and other factors[37]. These methods for assessing variant significance often are not cancer-specific, instead scoring functional impact[38] or deleteriousness[39] of variants from any study.

To better standardize interpretations of pathogenicity from germline variants, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology recently released guidelines for classifying sequence variants in genes with known roles in inherited disorders as "benign", "likely benign", "uncertain significance", "likely pathogenic", or "pathogenic" [37]. Software predictions are included in the algorithm, with increased evidence for pathogenicity if multiple software programs predict a deleterious affect. However, these software programs must have different underlying bases for prediction (e.g. amino acid conservation versus nucleotide conservation).

Additionally, the ACMG cautions against applying their criteria to candidate genes, since their guidelines are not meant to identify new genes in disease[37]. Finally, discordance remains even when using the ACMG guidelines. When nine different molecular diagnostic laboratories applied ACMG guidelines to a standardized variant set, the laboratories' ACMG classifications had 34% initial concordance, with an increase to 71% after detailed review of the ACMG criteria and consensus discussions[40].

Integration of large scale clinical and genomic information to nominate cancer driver genes based on comorbidity with Mendelian disorders

Integration of germline and somatic data remains uncommon, despite biological similarities between germline and somatic alterations[19, 30] and possible relationships between germline and somatic mutations such as LOH. Further biological interactions between germline and somatic variants are suggested by the genetic similarity between cancers and comorbid Mendelian diseases [41]. A Mendelian disease is caused by a specific genetic alteration, while a complex disease such as breast cancer has multiple genetic and environmental causes. Recent study of diagnosis from electronic medical records representing 110 million patients found that each complex disease had a unique set of Mendelian disease associations, a "Mendelian code'' of disease comorbidity [42]. These unique codes included novel comorbidities and known cancer comorbidities, such as the increased risk for breast cancer in patients with the Mendelian disease ataxia telangiectasia.

Our lab further assessed whether comorbid cancers and Mendelian disorders had genetic similarity [41]. By comparing the genes somatically mutated in a cancer type with

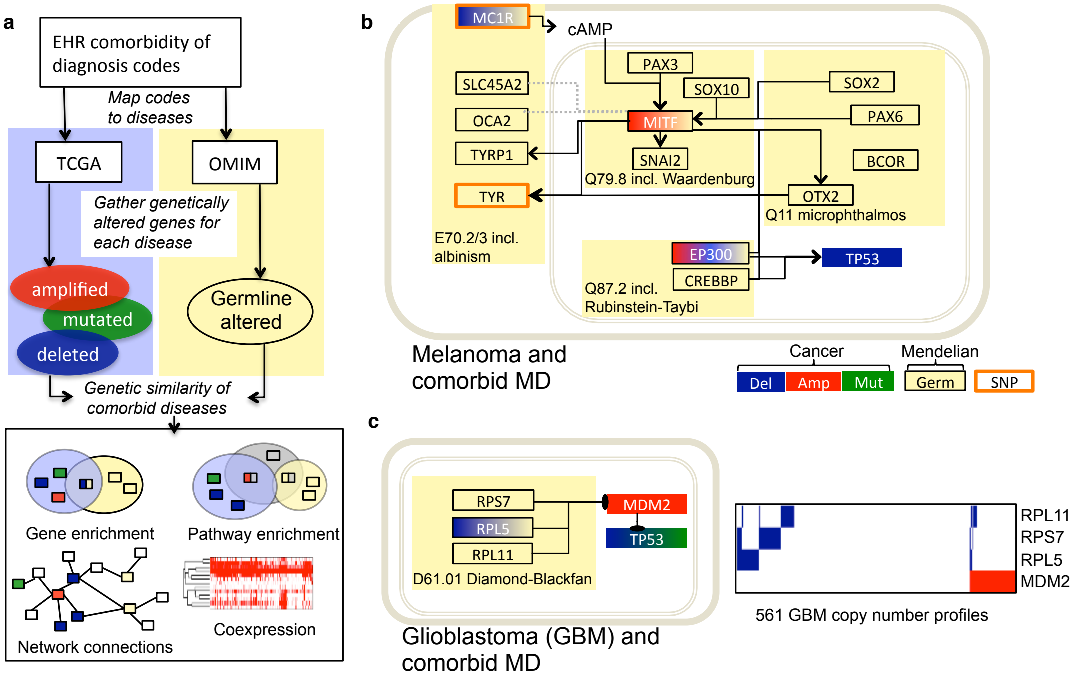


Figure 0.2: **Genetic similarity of comorbid cancers and Mendelian diseases** (a) Data integration and overview of genetic similarity metrics. (b) Diseases comorbid with skin melanoma, with yellow boxes containing Mendelian disease diagnosis codes and associated genes, and arrows indicating cellular pathways in melanocyte development and cancer-related processes. Genes that are recurrently somatically mutated in melanoma are highlighted. Solid edges represent interactions from the literature, while the dashed edges represent significant co-expression. (c) Interaction of genes altered in glioblastoma with Diamond–Blackfan anemia genes. Right shows cellular pathways; left summarizes of copy-number changes to *MDM2* and Diamond–Blackfan associated ribosomal proteins known to suppress the action of MDM2. RPL5 is recurrently and focally deleted in glioblastoma, and is mutual exclusivity with MDM2 amplification (one-tailed Fisher's exact test, P=0.033). RPL11 deletion also has mutual exclusivity with MDM2 amplification (P=0.042).amp, amplification; del, deletion; germ, germline; incl., including; mut, mutation; SNP, single-nucleotide polymorphism. Modified from ref. [41], Figures 1 and 3.

the genes causing comorbid Mendelian diseases, we found a cancer type and comorbid Mendelian diseases had significant functional similarity across multiple metrics, including significant cell-specific co-expression (Figure 0.2). This genetic similarity indicates that germline DNA alterations that lead to Mendelian diseases affect pathways that are also somatically deregulated in comorbid cancers.

Based on this significant genetic similarity, we nominated potential cancer driver genes.For example, melanoma was comorbid with multiple Mendelian disorders associated with *MITF*, a melanocyte cell fate transcription factor, suggesting that MITF is a melanoma driver gene amplified in 26% of TCGA melanomas. Our results also suggest alterations in *RPL5* may drive glioblastoma based on the comorbidity between glioblastoma and Diamond–Blackfan anemia (caused by inherited mutations in *RPL5* and other genes encoding ribosomal proteins), particularly because *RPL5* is significantly deleted in 8% of TCGA glioblastoma cases and suppresses *MDM2*, an oncogene amplified in 15% of TCGA glioblastoma. Our analysis suggests that certain Mendelian variants promote the development of comorbid cancer, and that the genes deregulated by these variants may be altered somatically because they influence cancer progression. This research also illustrates that integrated germline and somatic analysis can further inform our understanding of cancer biology.

Statement of problem and organization of the Thesis

In this thesis, we investigate how to identify germline variants that contribute to cancer development. We also develop strategies to integrate germline, somatic, and relapse genomic information when studying sporadic cancer cohorts. Traditional strategies to assess germline variants in cancer include GWAS for common variants, mapping highly penetrant variants in familial cancer predisposition syndromes, and expert curation of potentially pathogenic variants in cancer-associated genes from sporadic tumor cohorts. Given the low effect size of common variants, the infrequent nature of cancer predispositions syndromes, and the ambiguity in suggesting novel cancer-associated variants from manual analysis, we aim to learn the biological features of cancer-associated variants. Specifically, we will build a machine learning classifier that learns the features of somatic mutations and use that classifier to identify germline variants with somatic features. By using supervised learning on all somatic variants in our training set, our framework will learn biological features of variants across the exome, instead of relying on curated cancer-gene lists and expert manual curation. Our machine learning framework will also support integrated analysis of germline and somatic variants. Finally, we use this framework and other strategies to assess germline, diagnosis, and relapse mutations in pediatric leukemia.

This thesis is organized into three chapters and a conclusion. Chapter 1 focuses on the development of the Tumor-Only Boosting Identification (TOBI) framework that uses machine learning to identify somatic variants from tumor-only data or to identify somatic-like germline variants using matched germline DNA. We implement the TOBI framework and assess performance on true somatic variants, including somatic variants in driver genes. Using data from 1,769 patients from seven cancer types (bladder, glioblastoma, low-grade glioma, lung, melanoma, stomach, and pediatric glioma), we show that our framework has high sensitivity in identifying nonsynonymous somatic mutations.

In Chapter 2, we apply TOBI to integrated somatic and germline analysis. By assessing which variants are present in germline DNA and have a "somatic'' classification from TOBI, we identify "somatic-like'' germline variants. These somatic-like germline variants include known *TP53* germline variants and are enriched for genes associated with autosomal dominant cancer-predisposition syndromes. By assessing both somatic-like germline variants and true somatic variants in bladder cancer, we find that 5% of cases have germline inactivating mutation in the Fanconi anemia pathway that inactivating mutations in this pathway associate with a somatic signature of DNA repair deficiency.

Chapter 3 combines TOBI, germline variant classification, and somatic analysis to assess genomic alterations in a cohort of 627 pediatric patients with ALL. We capture known ALL driver genes in our somatic analysis, and describe a sequential ordering of variants in patients with relapsed ALL, including early mutations in *PHF6* and late mutations in *NT5C2*. We also assess TOBI's performance on the ALL cohort and nominate potentially pathogenic germline variants.

Acknowledgements

The section "Integration of large scale clinical and genomic information to nominate cancer driver genes based on comorbidity with Mendelian disorders" is a reproduction, in part, of a publication in Nature Communications by Melamed, R. D., Emmett, K. J., Madubata, C., Rzhetsky, A., and Rabadan, R. [41]. We recognize the contribution of all specimen donors. We thank Stephanie Siegmund for her editorial suggestions.

Chapter 1

Machine-learning to identify variants with somatic-features from tumor-only samples

1.1 Introduction

Cancer often results from specific DNA alterations, and identification of cancer-causing mutations underlies genome-based precision cancer treatment[43]. Somatic mutations can be identified by sequencing matched tumor and normal DNA[44], where normal samples can come from blood or any other non-tumor tissue, and then removing any shared variants (germline variants). This paired tumor-normal analysis has identified oncogenic somatic mutations in multiple cancer types, including cohorts originally analyzed by The Cancer Genome Atlas (TCGA)[45–50]. Germline DNA alterations can also be oncogenic[51].

A standardized framework for unified analysis of germline and somatic variants could reveal key oncogenic pathways. Recent analysis of sporadic ovarian cancer found significantly enriched germline and somatic alterations in the Fanconi anemia and MAPK pathways[30]. However, unified germline and somatic analysis is historically focused on cancers with known familial predispositions (e.g. ovarian and breast cancer) and often focuses on known cancer predisposition genes[31, 32, 34].

Given that certain oncogenic germline variants share biological features with known somatic variants, such as affecting the same amino acid[31], we hypothesize that learning the features of somatic mutations could promote integrated germline and somatic analysis. Specifically, we hypothesize that a machine learning framework built upon biological features of somatic variants would be able identify germline variants with somatic features that might influence tumor development. To learn the features of somatic mutations, our classifier would first learn to differentiate somatic and germline variants from a tumor sample. In the next section, we describe current techniques to classify somatic versus germline variants.

Review of strategies to distinguish somatic and germline variants

One biological feature of many recurrent somatic variants is their low population frequency. Accordingly, attempts to identify somatic variants from tumor-only whole exome sequencing (WES) data often involve removing common population variants found described in dbSNP[52]. Analysis then focuses on genes in the Catalogue Of Somatic Mutations In Cancer (COSMIC)[53]. However, this strategy fails to recognize private polymorphisms that are not annotated in public repositories and preclude the discovery of novel oncogenic events.

A limited number of computational strategies exist to identify somatic variants from tumor-only WES data. Certain strategies rely on a single patient's sequence alignment information, either predicting somatic deletions based on read-pair alignments and read depth[54] or predicting somatic single nucleotide variants (SNV) using base quality, variant allele frequency (VAF), and sequencing error[55]. Other strategies use population allele frequency tabulated from a cohort of normal genomes to remove potential germline SNPs[56]. None of these techniques integrate information from both the individual patient sequence and the total patient cohort. These techniques also fail to leverage valuable databases of somatic mutations or predicted mutation effects.

While we focused on tumor-only WES above, we note that there are strategies for identifying somatic mutations from tumor-only samples sequenced with high-depth gene panels[57]. However, this method again only assesses variants on an individual patient basis, and is constrained to a set of known cancer genes.

Our approach to learn the biological features of somatic variants involves integrating information from individual patients, patient cohorts, and curated databases. This approach would require a patient cohort with some matched tumor-normal cases and some tumor-only cases. The tumor-normal cases would form a training set for identifying true somatic mutations, and the biological features of these confirmed somatic variants would be used to classify variants from the remaining tumor-only samples. Prior studies of mixed tumor-normal and tumor-only cohorts used manual recurrence analysis of specific genes to reveal altered genes in lymphoma[58, 59], relapsed pediatric ALL[60], and pediatric glioma[61], but the focus on gene identity had decreased power to identify oncogenic variants. In contrast, we suggest using machine learning instead of manual analysis, and making predictions across the whole exome instead of focusing on specific genes.

Machine learning to distinguish somatic and germline variants

Supervised machine learning algorithms generate a model that predicts a response variable based on descriptive features in a training set of data. For example, an algorithm could learn from a training set of DNA variants with descriptive features such as variant allele frequency, and a response variable stating whether the variant was somatic or germline in patients. The algorithm could then develop a model that predicts somatic or germline status of DNA variants in an outside test set.

In a recent comparison of supervised learning methods, gradient boosted trees had the highest performance across multiple metrics compared to other methods such as random forests, bagged trees, neural nets, and calibrated support vector machines [62]. Gradient boosting generates an ensemble classifier by iteratively applying weak classifiers and learning from training set observations misclassified in prior iterations [63]. The typical base classifier used in gradient boosting is a short, low-depth decision tree. A single decision trees is a classifier that divides data into regions using covariates (e.g. variant allele frequency) and assigns all observations in a region to a response variable category. Decisions trees are biologically interpretable, but a low depth tree would predict only slightly better than random guessing, and higher depth trees can suffer from overfitting. However, boosting generates an additive ensemble of shallow decision trees that is generalizable, minimizes overfitting, and has lower error than single trees [63]. In biology research, models generated using gradient boosting have classified variants of unknown significance in Mendelian disorders [64] and identified biologically functional gene fusions in cancer [65].

Stochastic gradient boosting is one version of gradient tree boosting that improves model performance by introducing randomness into model generation [66]. In this chapter, we use stochastic gradient boosting to generate a model that distinguishes somatic and germline variants from tumor-only samples. Specifically, for a training set with N variants, with $(y_i, \vec{x_i})$ as the ith variant for $i = 1, 2, ..., N$, where y_i is our response variable of somatic or germline status, and $\vec{x_i}$ is a vector of biological covariates for sample i, we generate a model $\hat{F}(\vec{x})$ that maps $\vec{x}$ to y and minimizes the expected loss function $\Psi(y, F(x))$ We create our final additive model using the stochastic gradient boosting strategy described by Friedman [66]. Our final model, $\hat{F}(\vec{x}) = \sum_{m=0}^{M} \beta_m h(\vec{x}; \vec{a_m})$, summarizes M models, where $h(\vec{x}; \vec{a_m})$ is a simple base learner with parameters $\vec{a_m}$, and β_m are expansion coefficients. Each model $\hat{F}_m(\vec{x}) = \beta_m h(\vec{x}; \vec{a_m})$ trains on the pseudo-residuals from the prior model:

$$\tilde{y}_{im} = -\Big[\frac{\partial \Psi(y_i, F(\vec{x_i}))}{\partial F(\vec{x_i})}\Big]_{F(\vec{x_i})=F_{m-1}(\vec{x_i})}$$

Friedman [66] showed that by defining $h(\vec{x}; \vec{a_m})$ as an L-terminal node tree with L-disjoint regions $\{R_{lm}\}_1^L$, such that

$$h(\vec{x}; \{R_{lm}\}_1^L) = \sum_{l=1}^{L} \bar{y}_{im} \mathbb{1}(\vec{x} \in \{R_{lm}\})$$

the base classifier $h(\vec{x}; \vec{a_m})$ and coefficients β_m reduce to

$$\gamma_{lm} = \underset{\gamma}{\operatorname{argmin}} \sum_{\vec{x} \in \{R_{lm}\}} \Psi(y, F_{m-1}(\vec{x_i}) + \gamma).$$

To introduce randomness, we use random subsets of the training data to build each tree $h(\vec{x}; \{R_{lm}\}_1^L)$. We outline Friedman's stochastic gradient boosting algorithm in Algorithm 1.

```
Algorithm 1: Stochastic Gradient Boosting
  Result: $\hat{F}(\vec{x})$
  Initialize $F_0(\vec{x}) = \text{argmin}_\gamma \Psi(y_i, \gamma)$;
  for $m \leftarrow 0$ to $M$ trees do
      $\{\pi(i)\}_1^N = rand\_perm(\{i\}_1^N)$ for $\tilde{N} < N$ draws without replacement;
      generate random subsample $\{y_{\pi(i)}, x_{\pi(i)}\}_1^{\tilde{N}}$ ;
      for $n \leftarrow 1$ to $\tilde{N}$ observations do
          calculate pseudo-residual
              $$\tilde{y}_{\pi(i)m} = -\Big[\frac{\partial\Psi(y_\pi(i), F(x_\pi\vec{(i)}))}{\partial F(x_\pi\vec{(i)})}\Big]_{F(\vec{x})=F_{m-1}(\vec{x})}$$
      end
      Fit a regression tree to the targets $\tilde{y}_{\pi(i)m}$ giving terminal regions $\{R_{lm}\}_1^L$;
      for $l \leftarrow 1$ to $L$ regions do
          $\gamma_{lm} = \text{argmin}_\gamma \sum_{x_{\vec{\pi(i)}} \in \{R_{lm}\}} \Psi(y_{\pi(i)}, F_{m-1}(x_{\vec{\pi(i)}}) + \gamma)$;
      end
      update $F_m(\vec{x}) = F_{m-1}(\vec{x}) + v\cdot \sum_{l=1}^L \gamma_{lm} \mathbb{1}(\vec{x} \in \{R_{lm}\})$
  end
```

In this chapter, we develop our Tumor-Only Boosting Identification (TOBI) framework, using WES data from 1,769 patients across seven cancer types. We then assess TOBI's performance on predicting true somatic variants given tumor-only DNA, and compare TOBI to other software for tumor-only analysis. We find that TOBI has high true positive rates, particularly in nonsynonymous somatic variants in cancer driver genes. These results show that TOBI identifies true somatic variants using a machine-learning classifier built from somatic variant features.

1.2 Results

Framework for predicting somatic, germline and "somatic-like" germline variants

Our framework consists of four main steps: steps I-III accommodate tumor WES data at different stages of analysis, and step IV incorporates germline variant allele frequency (VAF) when available (Figure 1.1). Step I receives aligned WES files (.bam files), calls variants against a human reference genome, and annotates variants (full details in online Methods). These variant calls (.vcf files) are the input for Step II, allowing users to jump to Step II if they have previous annotated variants from tumor-only samples. Step II filters variants using biological and technical criteria described in the Online Methods, retaining high quality variants that are rare in the population (population minor allele frequency less than 1% in the 1000 Genomes Project[67]).

Step III receives the remaining training set variants and uses the gradient boosting machine learning algorithm to generate the somatic classification model. Gradient boosting generates a classifier from an ensemble of decision trees, where each subsequent tree learns from the previously misclassified training set observations.[66] For example, some features of previously described highly-recurrent variants will easily classify hotspot variants, while other features will be more relevant for classifying rarer mutations in subsequent trees. We optimized the gradient boosting parameters using systematic grid search. Each variant in the training set represents an observation for machine learning. Ten biological features were used for gradient boosting (full features in Appendix text 1.A); fea-

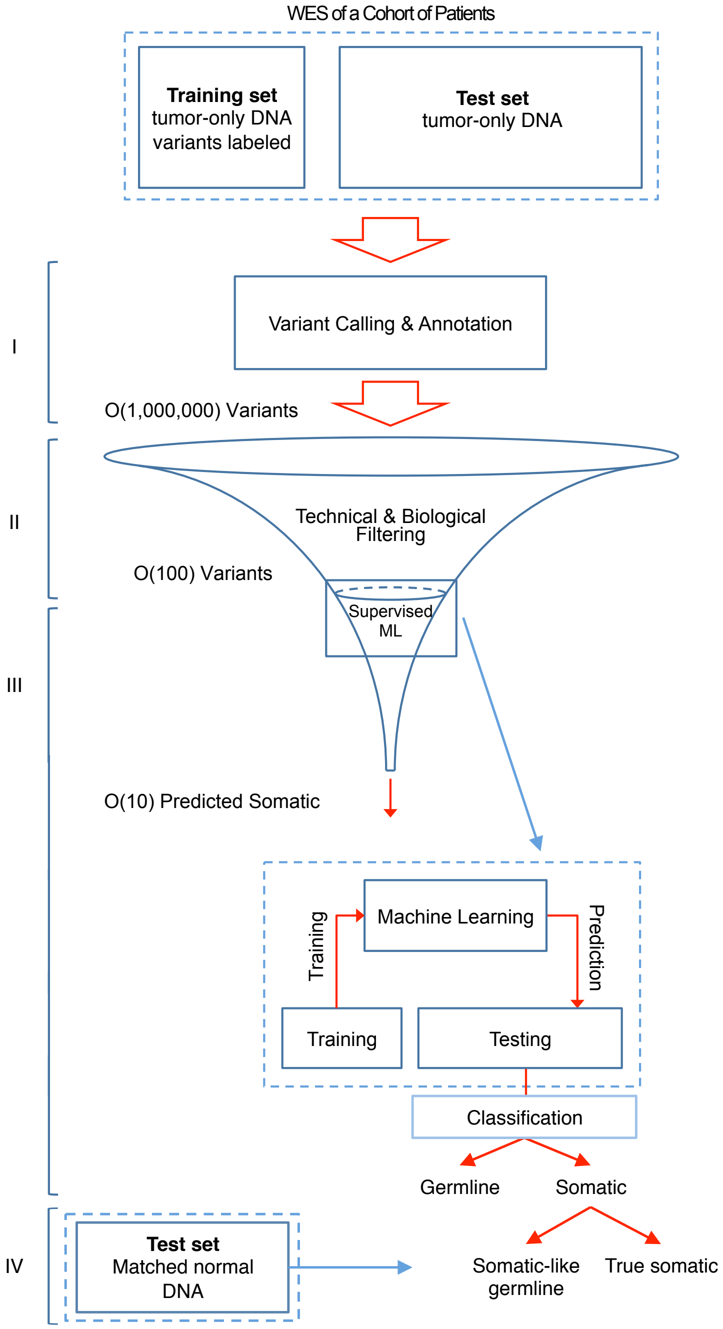


Figure 1.1: **Outline for predicting somatic variants with TOBI**. TOBI accepts tumor-only DNA, separated into a training set of cases with prior tumor-normal somatic analysis available and a test set. The steps of TOBI analysis are (I) variant calling and annotation, (II) filtering, (III) machine learning to classify "somatic" and "germline" variants, and (IV) identification of somatic-like germline variants. Step III predictions result in tens of predicted somatic variants per case.

tures include database-derived features from COSMIC, cohort-associated features such as “Variants per Gene”, and individual sequence features such as tumor VAF. Model generation requires training set variants annotated with true somatic status, defined by a user-generated list of somatic variants output from separate somatic variant calling pipelines (e.g. MuTect[68], SAVI[44]). Step III ends by applying the final somatic classification model to the test set variants.

Finally, Step IV occurs only if normal WES DNA is available for test set samples, and distinguishes somatic variants from somatic-like germline variants.

TOBI training and test sets

We developed TOBI using glioblastoma multiforme (GBM) cases from TCGA[45], and assessed TOBI on six adult cancer types from TCGA: bladder urothelial carcinoma (BLCA)[46], brain lower grade glioma (LGG)[47], lung adenocarcinoma (LUAD)[48], skin cutaneous melanoma (SKCM)[49], and stomach adenocarcinoma (STAD)[50]. We used TCGA’s previously published somatic calls as the “true somatic” calls for labeling training set variants. To assess TOBI’s performance on pediatric tumors, we analyzed pediatric glioma cases (Ped.Glioma), including cases with published tumor-normal analysis[69, 70] and tumor-only cases[61, 70, 71]. The number of cases per cancer type, and the number of cases used in each figure, is in Table 1.A.1.

Since cancer-sequencing studies have variable numbers of paired tumor-normal samples[54, 70, 71], we assessed the number of training cases required for model generation (Fig. 1.2a). Increasing the number of training set tumor samples from one to fifty samples

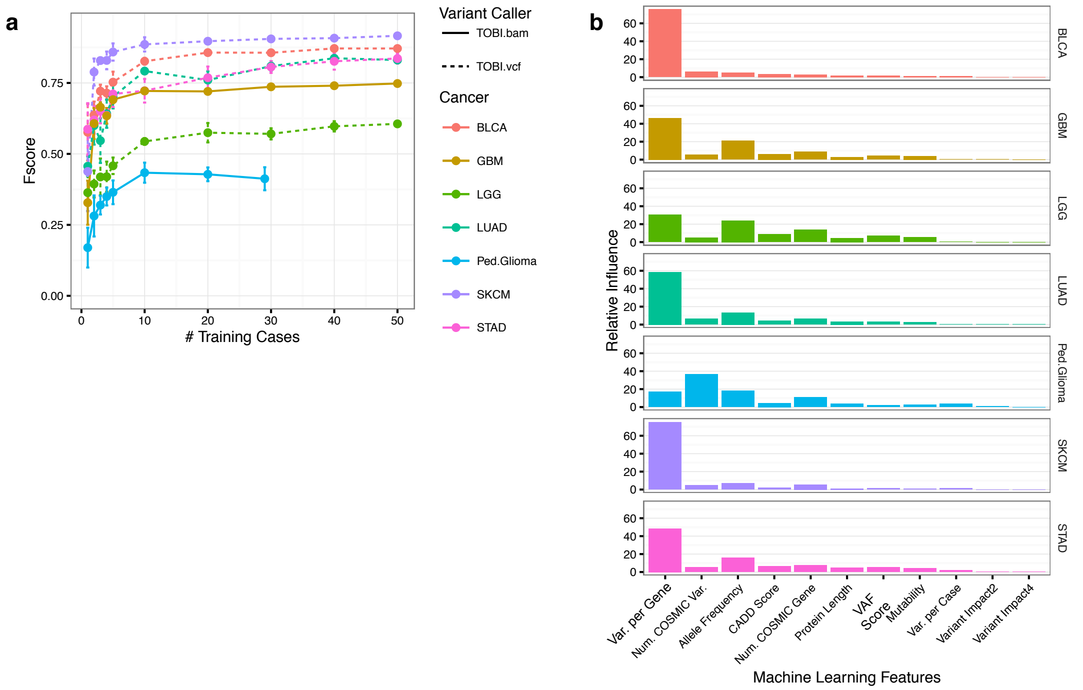


Figure 1.2: **TOBI training set size and relative importance of features**. (a) Average F-score for increasing numbers of cases in the training set in seven cancer types. Number of samples in the training set equals number in testing set. Points represent average predictions from five runs with randomly selected training and testing sets cases; error bars represent +/- s.e.m. TOBI.bam indicates samples were analyzed from aligned sequence files (.bam) using TOBI steps I-III; TOBI.vcf indicates samples were analyzed from variant call files (.vcf) using TOBI steps II-III. (b) Relative importance of features in gradient boosting classification model generated from a training set with twenty cases in each individual cancer.

improved performance, with F-scores plateauing between 20 and 50 training cases in the six adult cancers. Twenty training cases produced an average F-score within 10% of the F-score at the maximum training set size. Thus, in the remainder of our analysis, we used 20 random cases as the training set size and all remaining cases as the test set to reflect a WES scenario where the majority of patient samples are tumor-only.

Historical tumor-only samples may be formalin-fixed and paraffin-embedded (FFPE), which introduces sequencing artifacts. We applied TOBI's LUAD classification model to FFPE LUAD cases (Figure 1.A.1), and observed a slightly decreased F-score for FPPE (0.68) vs. frozen samples (0.81). FFPE samples had similar sensitivity and specificity (0.94, 0.97) compared to frozen samples (0.87, 0.96).

Next, we assessed how differences in patient ancestry, sequencing institution, or hypermutator status within a cohort might affect TOBI performance. Stratifying on patient's reported race, TOBI had decreased mean F-scores when the training and test set differed by race in almost all cancers. (Figure 1.A.2). Differing sequencing institutions between the training and test set also generated lower mean F-scores in almost all cross-institutional predictions (TCGA GBM with a cohort of 80 additional non-TCGA cases[72] in Figure 1.A.3; Ped.Glioma analysis in Figure 1.A.4). Finally, using hypermutator status from the STAD publication[50], we found no significant effect on TOBI's performance when analyzing a non-hypermutator population or mixed population (61 hypermutator, 219 non-hypermutator; figure 1.A.5). Thus, TOBI's performance might improve with features denoting racial or institutional differences, but performance appears robust to hypermutator samples.

TOBI features

We assessed the importance of our ten biological features to a cancer type's final classification model using relative influence[73], a measure of how frequently one feature is used in the decision trees within the final classification model (Figure 1.2b). In all adult cancers, the feature with greatest relative influence was "Variants in Gene", the total number of variants per gene normalized by cohort size. In pediatric glioma, the feature with greatest relative influence was "Num. COSMIC Var.", representing the number of cases

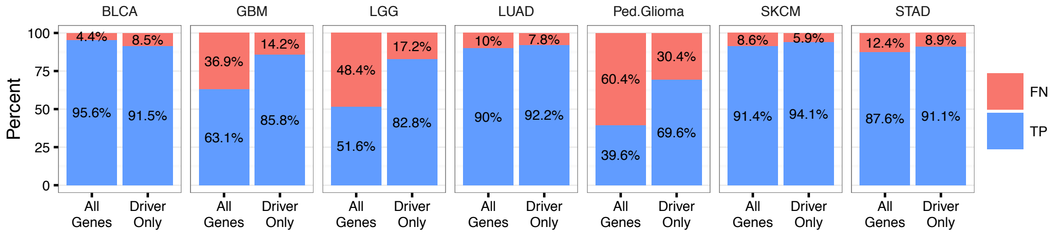


Figure 1.3: **True positive rate of TOBI somatic predictions in nonsynonymous variants.** For each indicated cancer type, percentages of true positive (TP) or false negative (FN) TOBI somatic predictions in nonsynonymous variants across all genes or only driver genes.

in COSMIC with a specific variant; this may reflect both the lower mutation burden in pediatric glioma and the prevalence of hotspot mutations in *H3F3A*. As expected, removal of these top features from the classification model caused a slight drop in F-score, while removal of other individual features or both COSMIC-derived features minimally affected performance (Figure 1.A.6).

High performance somatic variant identification

We compared TOBI's somatic classifications to published somatic calls from tumor-normal analysis of test set cases[45–50, 69, 70]. Across all variants, TOBI had a sensitivity of 86.6%; for nonsynonymous variants, TOBI had a sensitivity of 87.2%. Additional performance metrics are in figure 1.A.7. TOBI also has high sensitivity for variants with tumor VAF as low as 5% (Figure 1.A.8). Per gene, the number of cases with nonsynonymous variants predicted as somatic closely matches published somatic analysis (Fig. 1.3, 1.4). TOBI's sensitivity in a cancer type positively correlates with the median somatic SNV per megabase (Mb) across all cases of that cancer (Spearman rho 0.964, p-value < 0.003 for both all gene and driver only sensitivity, Figure 1.A.9).

While TOBI identifies variants with somatic characteristics, an important challenge in precision medicine involves finding genes that promote tumor development ("driver genes"). Thus, we assessed whether TOBI's predictions were enriched for driver genes in each tumor type, defining driver genes as those with evidence of positive selection in somatic mutation patterns as published by the Intogen group[74]. In six cancers, TOBI has a higher true positive rate of nonsynonymous variants in driver genes compared to all genes (Fig. 1.4). Such enrichment occurred despite training sets retaining synonymous variants and probable passenger variants. This driver gene enrichment did not solely arise from predicting highly recurrent genes, as suggested by TOBI's similar performance in high, medium, and low recurrence genes in most cancers (Fig. 1.A.10).

Finally, to demonstrate analysis of a truly tumor-only data set, we applied the pediatric glioma classification model to 68 tumor-only cases (Fig. 1.5), identifying known driver genes in pediatric glioma (*TP53*, *H3F3A*, *PIK3CA*). All predicted BRAF and IDH1 variants occurred at known somatic hotspots (*BRAF* V600E, *IDH1* R132H).

TOBI outperforms other tumor-only analysis tools

Using six GBM and six Ped.Glioma cases, we compared TOBI's results to those from other software for tumor-only WES somatic variant analysis: Virtual Normal Correction (VNC)[56] and SomVarIUS[55]. Compared to VNC, TOBI has higher F-scores (0.48 for Ped.Glioma and 0.22 for GBM; VNC F score less than 0.0002 for both Ped.Glioma and GBM; Table 1.A.2). SomVarIUS did not identify any true somatic mutations in Ped.Glioma. TOBI also predicts orders of magnitude fewer somatic variants per case compared to VNC and

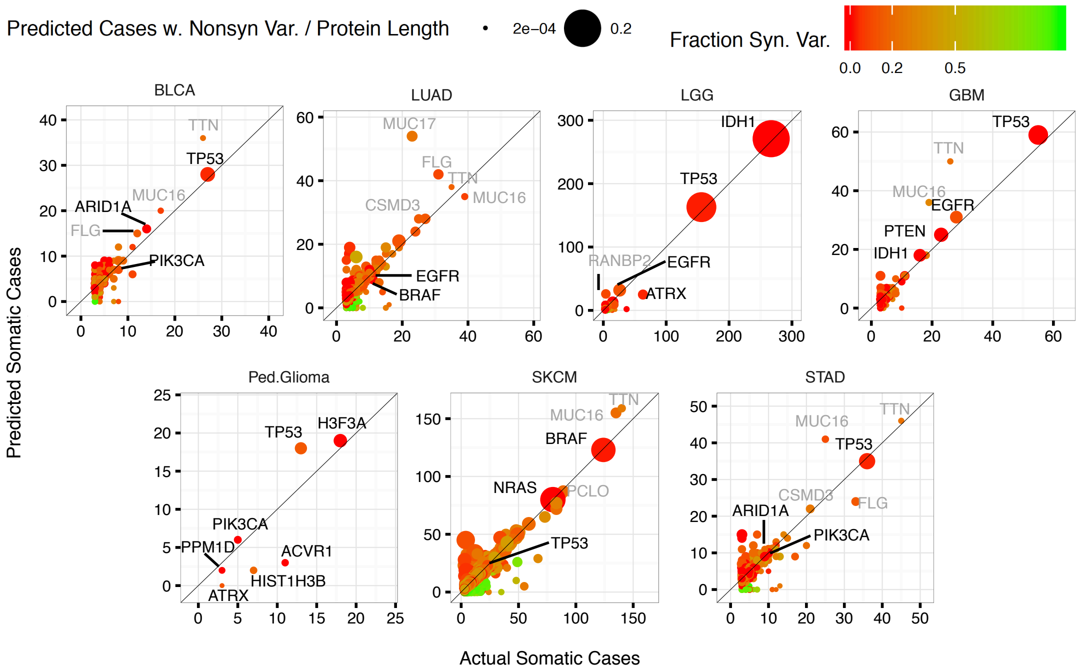


Figure 1.4: **Comparison of actual versus predicted cases with somatic, nonsynonymous variants.** Dot color corresponds to the fraction of synonymous variants out of all variants remaining after TOBI filtering (Step II); dot size corresponds to number of predicted cases over protein length in amino acids. Driver genes labeled in black; other genes in the top five most predicted cases labeled in grey. For clarity, genes with less than three previously published somatic variants are not shown.

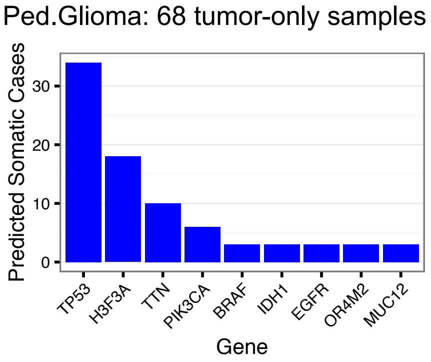


Figure 1.5: **TOBI predictions on tumor-only pediatric glioma cohort.** Number of cases with predicted somatic variants when pediatric glioma classification model is applied to 68 tumor-only samples; genes predicted in at least 3 cases shown. For all cancers, twenty randomly selected tumor-normal cases comprised training set; remaining paired tumor-normal samples formed testing set.

SomVarIUS (TOBI: ~5-50; VNC: ~300,000; SomVarIUS: ~100-3,000). TOBI's higher F-scores and biologically appropriate number of somatic variants indicates that TOBI outperforms these methods.

We also compared TOBI to methods that assess a variant's disease potential[38, 39, 75, 76] since these methods have been used to assess effects of somatic variants. Using published somatic variants from tumor-normal analysis as the gold standard, TOBI consistently had the highest AUC (Figure 1.A.11).

1.3 Discussion

In this chapter, we describe Tumor-Only Boosting Identification framework, or TOBI, a new unifying framework that uses the gradient boosting machine learning algorithm to identify somatic variants from tumor-only data or identify somatic-like germline variants in patients with tumor-normal DNA available.

In tumor-only analysis, TOBI successfully identified 87% of nonsynonymous somatic variants. Higher true positive rates in driver genes suggest that TOBI enriches for cancer-causing variants. TOBI's similar performance on frozen and FFPE samples suggests that TOBI filters certain FFPE artifacts. A TOBI modification trained on FFPE artifacts could potentially remove more FFPE sequencing artifacts, although this modification would need testing. TOBI also outperforms other methods designed for somatic variant identification from tumor-only samples. This higher performance likely reflects two fundamental differences between alternative methods and TOBI. First, alternative techniques use a single information source, but TOBI integrates biological features from individual variants, patient cohorts, and curated databases. Second, TOBI uses the powerful gradient boosting algorithm to classify variants, allowing TOBI to learn features important to specific tumor types (Fig. 1.2).

We recognize several limitations for the TOBI framework. First, TOBI's biological features include some that depend on outside databases (COSMIC variants), and future versions of these databases could affect TOBI predictions. Moreover, we only assessed a subset of biological features; alternative features could lead to improved TOBI performance. Second, FFPE status, patient ancestry, and sequencing institution do affect TOBI's performance, suggesting that TOBI will perform best on relatively homogeneous cancer cohorts. Third, TOBI's sensitivity positively correlates with the median somatic SNV rate per cancers, possibly due to the increased fraction of somatic mutations in the training set of melanoma and other cancers with high mutation rates. This suggests that TOBI will be most sensitive in cancers with high somatic mutation rates. In sum, we propose a framework that analyzes either tumor-only samples or samples with matched tumor-normal DNA for variants with somatic features. In tumor-only samples, the framework (1) promotes the study of previously collected tumor samples without matched normal DNA, potentially unlocking a vast repository of tumor-only samples without sequencing of matched normal DNA, and (2) prioritizes exome alterations in a particular patient by focusing on variants with somatic characteristics. The results of this chapter focused on developing TOBI, assessing TOBI performance on true somatic mutations, and analyzing tumor-only samples. However, we hypothesize that TOBI will identify certain true germline variants as having somatic features. We further hypothesize that these somatic-like germline variants will include true oncogenic germline variants that are biologically similar to somatic variants.

In the next chapter, we investigate whether TOBI identifies somatic-like germline variants. We assess whether known oncogenic germline variants are identified, and calculate enrichment for somatic-like germline variants in biological pathways.

1.4 Methods

Sequence access and retrieval of clinical and somatic data

We obtained approval from the database of Genotypes and Phenotypes (dbGaP) to access exome sequences and germline variant calls from TCGA (accession number phs000178.v9.p8). We downloaded WES files (.bam files) for 104 randomly selected tumor-normal GBM cases from TCGA. For the remaining five TCGA cancers (BLCA, LGG, LUAD, SKCM, STAD), we downloaded Protected Mutation vcf files with somatic and germline variants for entry into the TOBI.vcf pathway indicated in Figure 1.2a. We downloaded and analyzed all TCGA Data Matrix cases with Broad Institute-generated Protected Mutation vcf files between July 28, 2015 and September 1, 2015, as well as 226 additional LGG cases downloaded between September 1, 2016 and September 4, 2016. For STAD, 282 cases had available vcf files; 63 cases classified as "hyper-mutated" in TCGA clinical data were excluded from the main analysis. For all six TCGA cancers, clinical data was retrieved from cBioPortal[77] and publication MAFs from the TCGA Data Matrix provided true somatic variant calls.

We analyzed the WES files (.bam files) for the 92 GBM cases analyzed in Wang et al.

2016. Published somatic calls were used to label true somatic variants.

For pediatric glioma WES sequence files, we obtained approval from the appropriate Data Access Committees (DAC) and downloaded all available sequence files from EGA. Bam files were available for datasets EGAD00001000807[70] (St. Jude Children's Research Hospital-Washington University Pediatric Cancer Genome Project Steering Committee) and EGAD00001000706[69] (ICR DIPG Data Access Committee). Fastq files were available for EGAD00001000792[71] and EGAD00001000791[61] (McGill-DKFZ Pediatric Brain Tumour Consortium); samples were mapped to GRCh37.71 using BWA 0.7.12[78] before variant calling. Published somatic variant calls were used to label true somatic variants for the 74 paired samples; only experimentally validated somatic mutations from [70] were included.

Clinical data was retrieved from supplementary tables for Ped.Glioma patients[61, 69–71] and using the R cgdsr package for TCGA. To standardize nomenclature for reported race across studies, we removed samples with missing or mixed classification (“Asian & White”, “Multiple (NOS)”, “Mixed”, “”, “.”, “N/A”, “Other”, “[Not Evaluated]”, “[Unknown]”), and standardized “BLACK OR AFRICAN AMERICAN” to “black”. Patient counts after standardizing nomenclature are in Table 1.A.3.

For 1000 Genomes Project[67] samples, phase 3 bam files were downloaded from the public FTP site for the first 99 “mapped” samples listed in `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/alignment_indices/20130502.exome.alignment.index`, as well as sample NA11994, which was previously reported to have a germline variant in *TP53* (R273H)[31].

All GBM, pediatric glioma, and 1000 Genomes Project bam files went through the

TOBI.bam pathway indicated in Figure 1.2a.

Variant calling and annotation

Bam files were analyzed with Samtools and Bcftools[79] to call variants, excluding variants with mapping quality lower than 10.

Variants were annotated using SnpEff[80] and SnpSift with dbSNP build 144, Cosmic v74, and dbNSFP v2.4 databases[81]. We also annotated the variants with an in-house database of common mutations in 219 normal WES cases ("Meganormal" database).

Filtering

Filters thresholds were selected based on preliminary analysis of GBM samples. We applied two main filters on the variants: 1) Technical filter and 2) Biological filter. The technical filter retained all variants with either a quality score from Bcftools greater than 60 or variant depth higher than 10 on both strands. These filters retained a high fraction of true somatic mutations in known driver genes (e.g. *EGFR*, which had good depth but a QUAL score <=60) while removing many low quality variants. Variants with sample VAF (the number of sequencing reads supporting a variant nucleotide divided by the total number of sequencing reads at that genomic position) less than 1% were removed. We also removed the variants that had low mapping quality (mq < 40), and had strand bias, map quality bias, and tail distance bias with the p-values below 0.01. In the biological filter, we removed common SNPs (population allele frequency greater than 1% in the 1000 Genome Project populations), as well as variants that were present in our Meganormal database.

We also removed the SNPs that were in the dbSNP database, but were not in COSMIC. Variants in intragenic, non-coding exon, and splice-site regions were also filtered. We applied these filters to GBM and pediatric glioma variants.

The TCGA variants in the TOBI.vcf pathway did not have reported per strand depth, mapping quality, and technical biases; thus, we used a modified Technical filter to remove variants with total depth <10 and QUAL score <=60. Biological filters were the same across all samples.

Machine learning

We selected the gradient-boosting algorithm for machine learning given its excellent performance on diverse binary classification problems compared to other supervised learning methods[62]. This algorithm generates a classification model using an ensemble of decision trees that iteratively learn from the previously misclassified training set observations. Gradient boosting returns a probability that a variant is somatic, which TOBI converts into a binary decision using an optimized probability threshold. TOBI does not use the default threshold probability of 0.5 because that would favor the majority class (in our case, non-somatic mutations), resulting in low sensitivity[82]. Instead, TOBI selects a probability threshold that maximizes classification performance; the threshold's potential range is 0.05 to 0.95 in increments of 0.0375.

For each cancer, TOBI generates an optimum classification model by running a systematic grid search through gradient boosting's three parameters: number of trees (100, 150, 200), interaction depth (3-7 splits), and shrinkage (constant at 0.1). For each possible combination of these three parameters, TOBI performs five repeats of 5-fold cross-validation on the training set in order to avoid over-fitting to the training set. The large number of training set variants compared to features also avoids overfitting. TOBI finally selects the parameter combination that maximizes average performance across the five repeats as the final classification model.

To select the best model despite the class imbalance, we used the F-score as the model performance metric:

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{1.1}$$

where TP, FP, and FN stand for true positive, false positive, and false negative. Maximizing F-score results in maximizing TP while minimizing FP and FN. We also assessed performance by calculating sensitivity, specificity, positive predictive value, negative predictive value, prevalence, accuracy, false positive rate (FPR), false discovery rate (FDR), and AUC. For these calculations, true negatives were those variants that passed all TOBI quality filters, were not published as somatic in source publications, and were not predicted as somatic by TOBI.

Here, we describe the software implementation of gradient boosting. For each cancer, cases were randomly assigned to the training or test set using the sample() function without replacement in R. TOBI then calculated cohort-specific annotations separately for the training and test set (see Appendix text 1.A for features). Somatic status of training set variants was annotated using a user-supplied list of somatic variants, defined by affected case, genomic position, and variant nucleotide. Next, TOBI used the Caret and gbm packages in R[82] to perform gradient boosting and generate a classification model. To assess feature importance, relative influence of features was automatically calculated during model generation. Relative influence is a measure of how many times a feature is selected for splitting in all trees in the gradient boosting model, weighted and scaled so that the sum of relative influence of all features equals one hundred.

We defined drivers in Figure 1.3 using the list of driver genes provided by the Intogen group[74].

The rate of somatic SNVs per Mb for each case was calculated using the number of published somatic SNVs, after converting di-nucleotide mutations into single nucleotide components and removing indels. This number was divided by the total megabases covered in Agilent SureSelect Human All Exon 50 Mb regions.bed file.

Our framework is available online for non-commercial use `https://github.com/RabadanLab/TOBI`.

1.5 Acknowledgements

This chapter is a reproduction, in part, of a publication in npj Genomic Medicine by Madubata, C.J., Roshan-Ghias, A., Chu, T., Resnick, S., Zhao, J., Arnes, L., Wang, J. and Rabadan, R [83]. This work was funded in part by R01 CA185486-01, R01 CA179044-01A1, and NIH U54 CA193313; J.W. is supported by Precision Medicine Fellowship (UL1 TR000040); A.R. was supported by the Swiss National Science Foundation Postdoctoral Fellowship; C.J.M. is supported by the NIH Medical Scientist Training Program (5T32 GM07367) and the National Research Service Award Individual Predoctoral Fellowship

(F31CA210607). This study makes use of data generated by the St. Jude Children's Research Hospital—Washington University Pediatric Cancer Genome Project (http://explore.pediatriccancergenomeproject.org/). We also acknowledge The Institute of Cancer Research, McGill University, the German Cancer Research Center (DKFZ), and The Cancer Genome Atlas (managed by the NCI and NHGRI, cancergenome.nih.gov) as sources of primary data. We recognize the contribution of all specimen donors. We thank Hossein Khiabanian for helpful discussions during project design.

1.A Appendix

Explanation of TOBI features

- "Var. per Gene" is the total number of variants per gene normalized by the number of patients in the cohort, calculated separately for training and testing datasets.
- "Num. COSMIC Var." is the total number of samples in COSMIC[53] with this specific nucleotide variant ("CNT" in COSMIC v74 vcf).
- "Allele Frequency" is the variant allele frequency (VAF) in the tumor sample.
- "CADD Score" is the Combined Annotation–Dependent Depletion Score, a score of variant deleteriousness integrated from multiple genome annotations. For more details, please see [39].
- "Num. COSMIC Gene" is the total number of COSMIC mutations in a gene.
- "Protein Length" is the length of the protein in amino acids.
- "VAF score" is the probability of a mutation to be a germline mutation with VAF = 50%. It's calculated using the binomial distribution:

  $$VAFScore = Binom\left(dp_{var}, dp_{tot}, 0.5\right) \tag{1.2}$$

  where dp_{var} and dp_{tot} are variant depth and total depth, respectively. The justification is that assuming no copy number variation (CNV), the VAF of germline mutations should be either 50% or 100%. This can be seen in Fig. 1.A.6 where a local minimum ratio of somatic to non-somatic mutations occurs around VAF = 50%. This feature helps identify mutations with a high probability of being germline in cases without CNV.

- “Mutability” indicates if a gene is prone to mutation in a normal, non-tumor cohort. Per gene calculation involved counting the total number of mutations per gene in a cohort of 219 normal samples, and dividing by the amino acid length.
- “Var. per Case” for a particular gene and a particular sample represents the number of variants in that sample divided by the number of patients in the cohort; cohort indicates either training or testing set.
- “Variant Impact” is the predicted effect impact from SnpEff, and it can be “High”, “Moderate”, “Low”, and “Modifier”. For more details, please see [80].

Additional features used in preliminary GBM analysis

- “Recurrent” is the number of recurrent mutations for the specific variant in the cohort.
- “ID” is 1 if the mutation is in no database, 2 if it is both in dbSNP and COSMIC, and 3 if it is only in COSMIC. Mutations that are only in dbSNP where filtered out.
- “MutationAssessor is a functional impact of amino-acid substitutions in proteins, such as mutations discovered in cancer or missense polymorphisms. The functional impact is assessed based on evolutionary conservation of the affected amino acid in protein homologs”[38].
- “MutationTaster” is a composition of different scores including evolutionary conservation, splice-site changes, and loss of protein features[76].
- “Polyphen” is a score that predicts possible impact of a mutation on the struc-

ture and function of a human protein using physical and comparative considerations[84].

- “SIFT” is a score of the effect of a mutation on the protein structure[75].

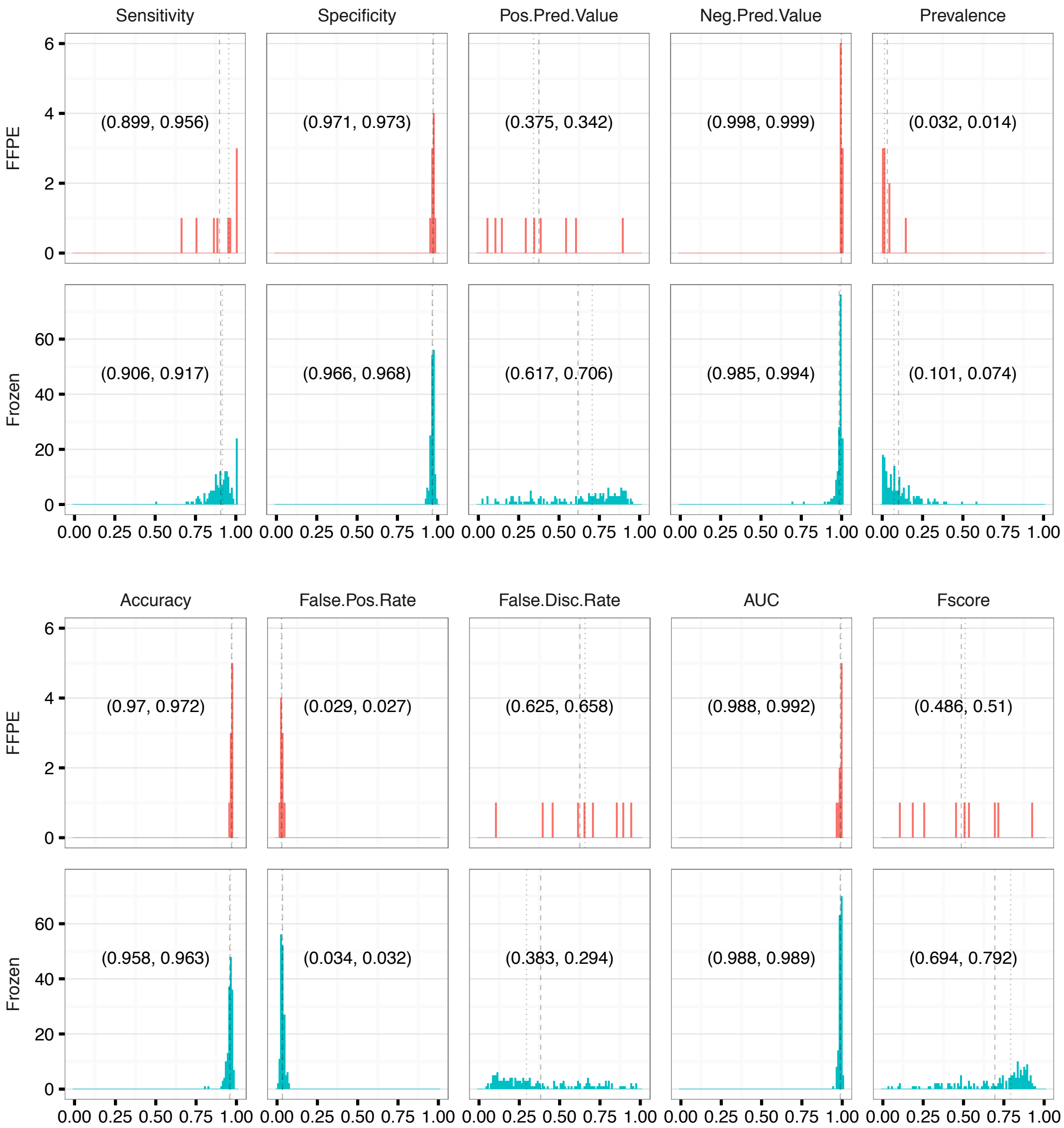


Figure 1.A.1: **Performance metrics in 9 FFPE cases and 161 frozen cases from LUAD cohort**. Metric listed on top of box; for each metric, top figure represents FFPE samples, bottom frozen samples. Y-axis of case counts, x-axis represents 0 to 1 range of metrics. In each box, ordered pair represents “(mean, median)” of metric for that patient cohort; dashed line=mean, dotted line= median.

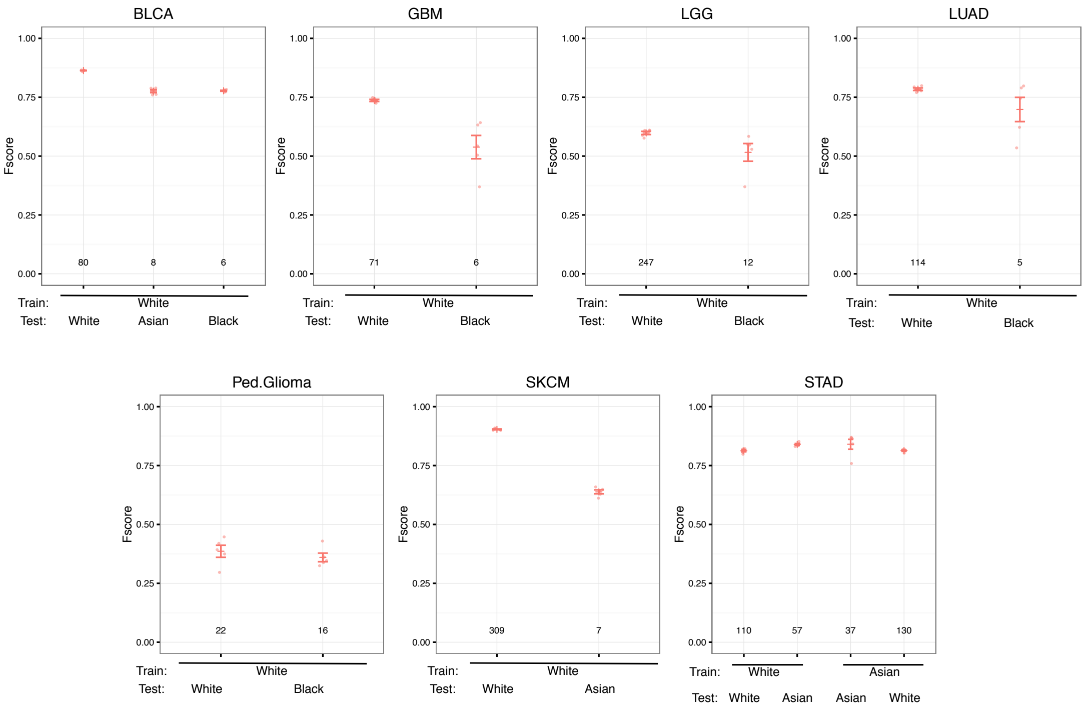


Figure 1.A.2: **TOBI performance when training and testing set are stratified by patients' self-reported race**. Each box corresponds to one cancer type. Y-axis shows F-score, x-axis shows reported race of training set used to generate model (20 randomly selected patients) above reported race of test set; number of cases in the race-stratified test set shown within plot area. Self-reported race categories required greater than 20 patients for inclusion as a training set, and a minimum of 5 patients for inclusion as a test set. Points represent F-score for five runs with randomly selected training and testing sets from specified race; error bars represent mean +/- s.e.m.

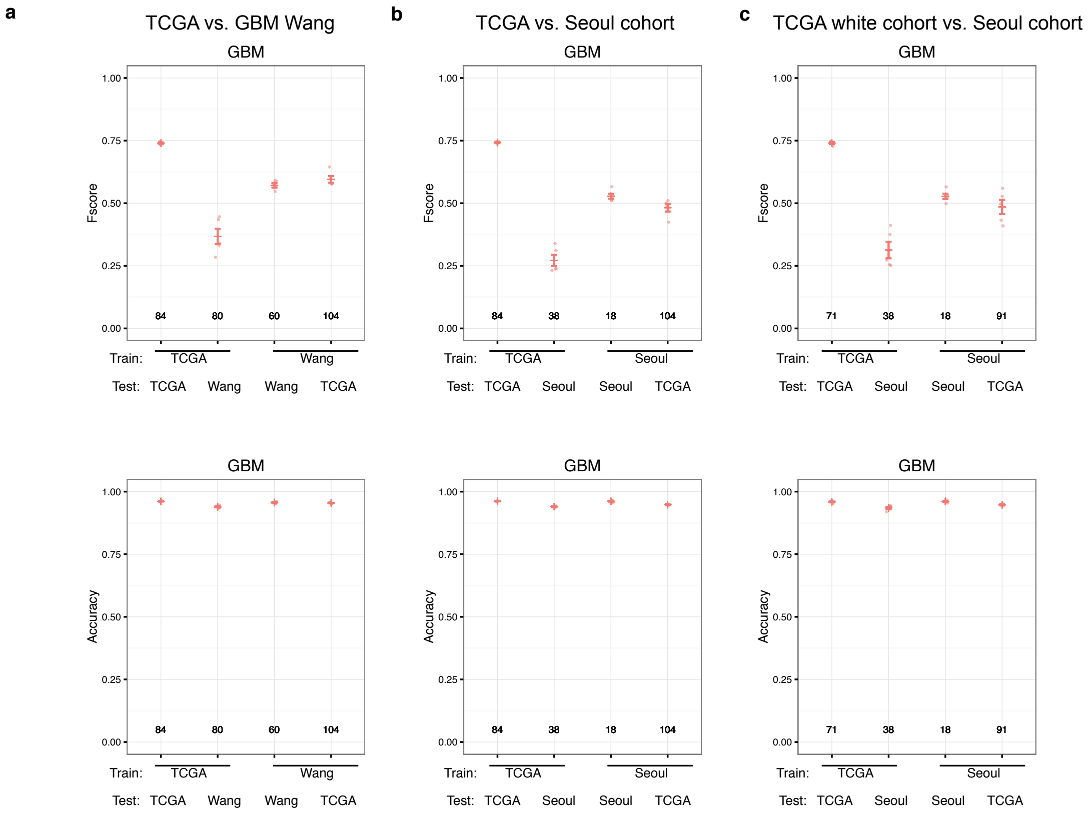


Figure 1.A.3: **TOBI performance on GBM training and testing sets stratified by institution**. Each box corresponds to one cancer type. Y-axis shows F-score or accuracy, x-axis shows reported race of training set used to generate model (20 randomly selected patients) above institution of test set; number of cases in the test set shown within plot area. An institution required greater than 20 patients for inclusion as a training set, and a minimum of 5 patients for inclusion as a test set. Points represent performance metric for five runs with randomly selected training and testing sets from specified race; error bars represent mean +/- s.e.m. (a) Stratifying GBM cases analyzed by TCGA or within Wang et al., 2016., excluding TCGA cases within Wang analysis, (b) by TCGA cases versus Wang cases collected and analyzed in Seoul, and (c) by TCGA cases identified as “white” versus Seoul cases.

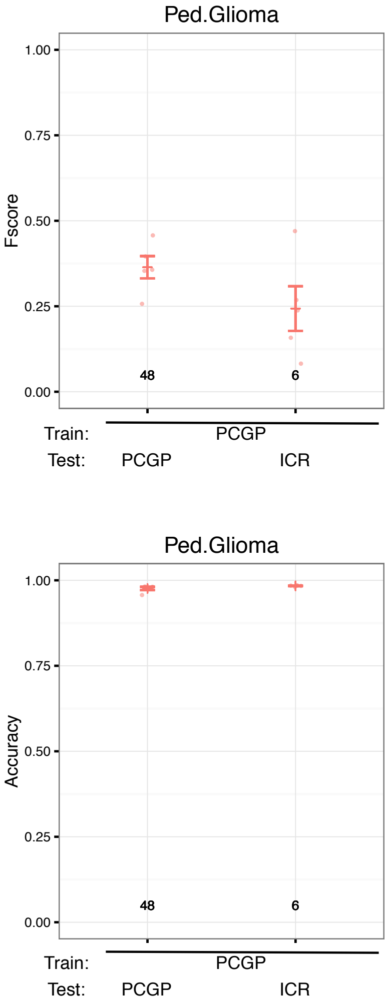


Figure 1.A.4: **TOBI performance on Ped.Glioma training and testing sets stratified by institution**. Stratifying Ped.Glioma cases analyzed by Pediatric Cancer Genome Project or The Institute of Cancer Research, London.

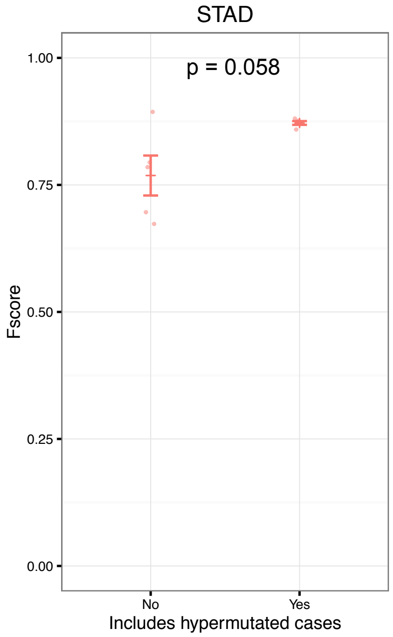


Figure 1.A.5: **Inclusion of 61 STAD cases with hypermutation phenotype does not significantly alter TOBI performance.** p-value from Welch's Two Sample t-test.

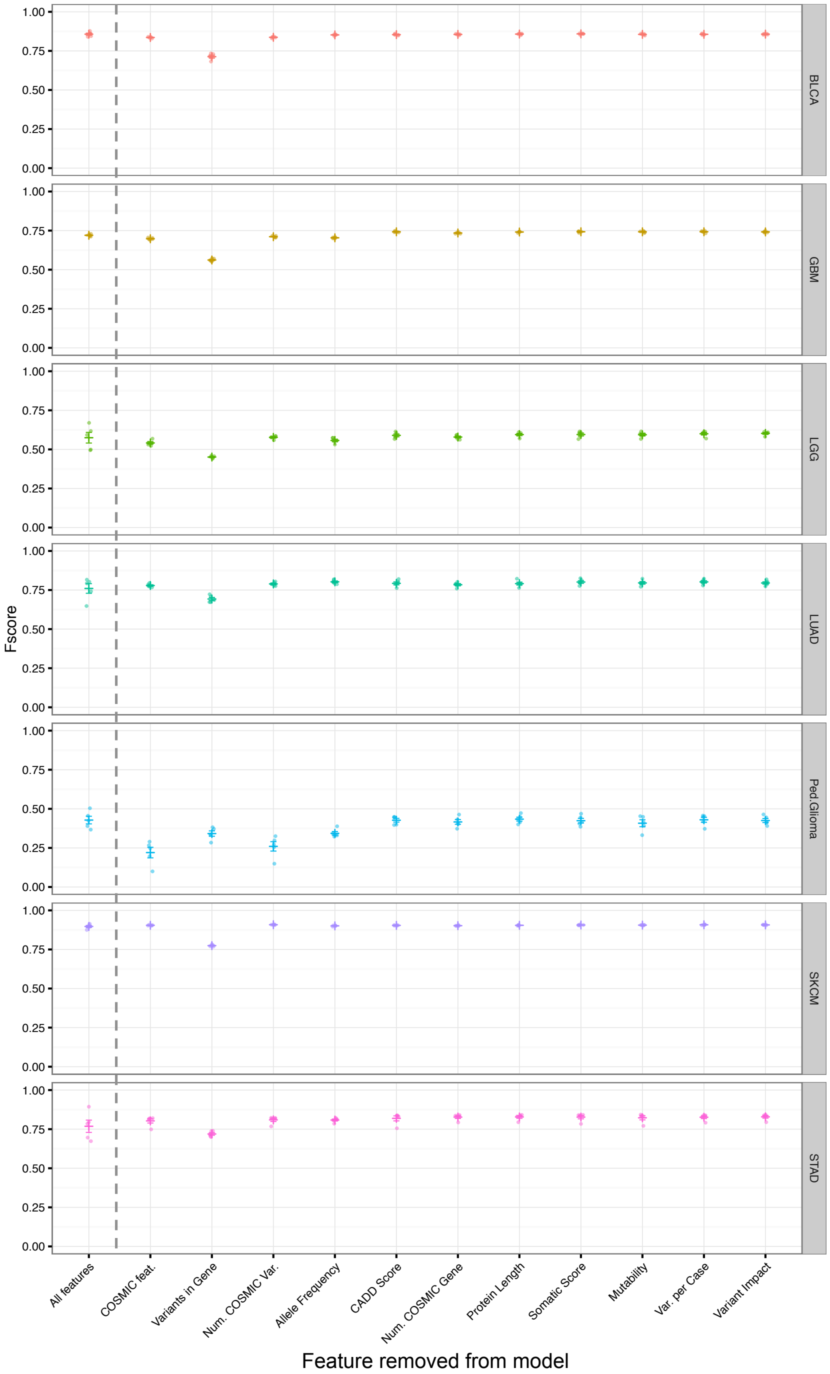


Figure 1.A.6: **Effect of removing individual features or all COSMIC associated features from the TOBI model.** Each box indicates a cancer type. Left of the dashed line indicates performance using the standard TOBI model with all features included; to the right, F-scores after the specified feature is removed from the model. Points represent F-score for five runs with randomly selected training and testing sets from specified race; error bars represent mean +/- s.e.m.

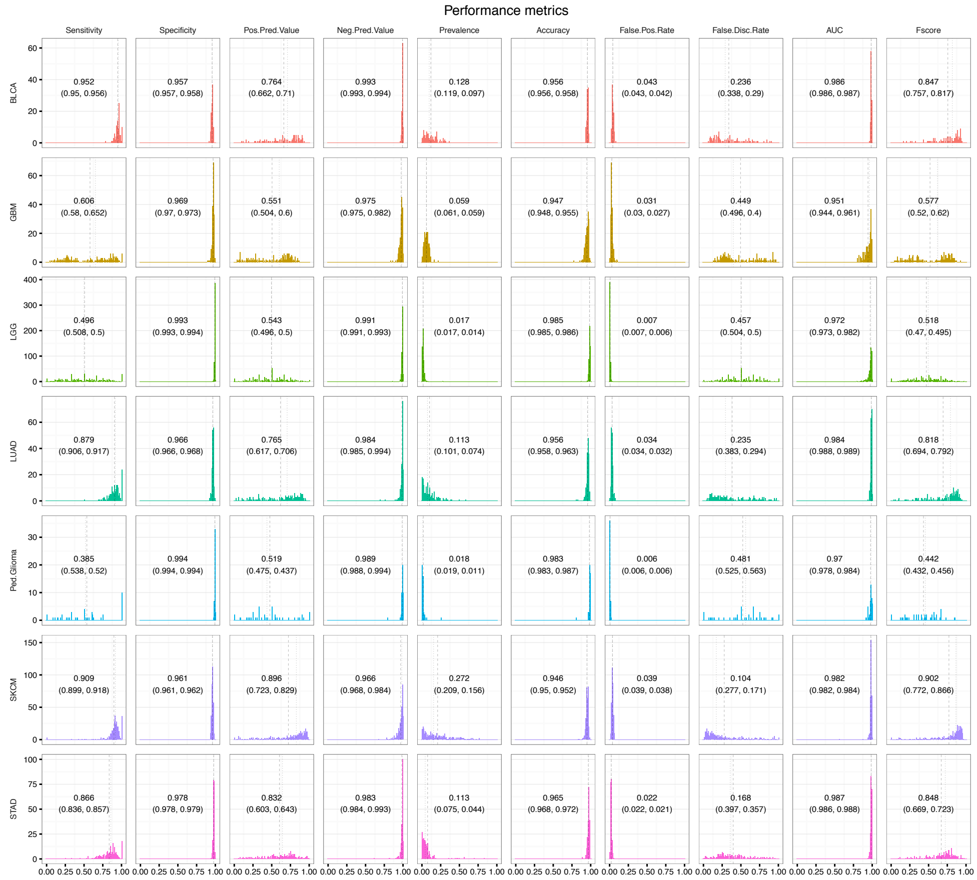


Figure 1.A.7: **Performance metrics in cancers analyzed by TOBI**. Histogram of performance across all variants in each case. Metric on top of column; each row is a cancer type. Y-axis: case counts, x-axis: 0 to 1 range of metrics. In each box, the upper number represents that performance metric across all samples and variants in that cancer subtype; the ordered pair represents "(mean, median)" of metric for that patient cohort; dashed line=mean, dotted line= median.

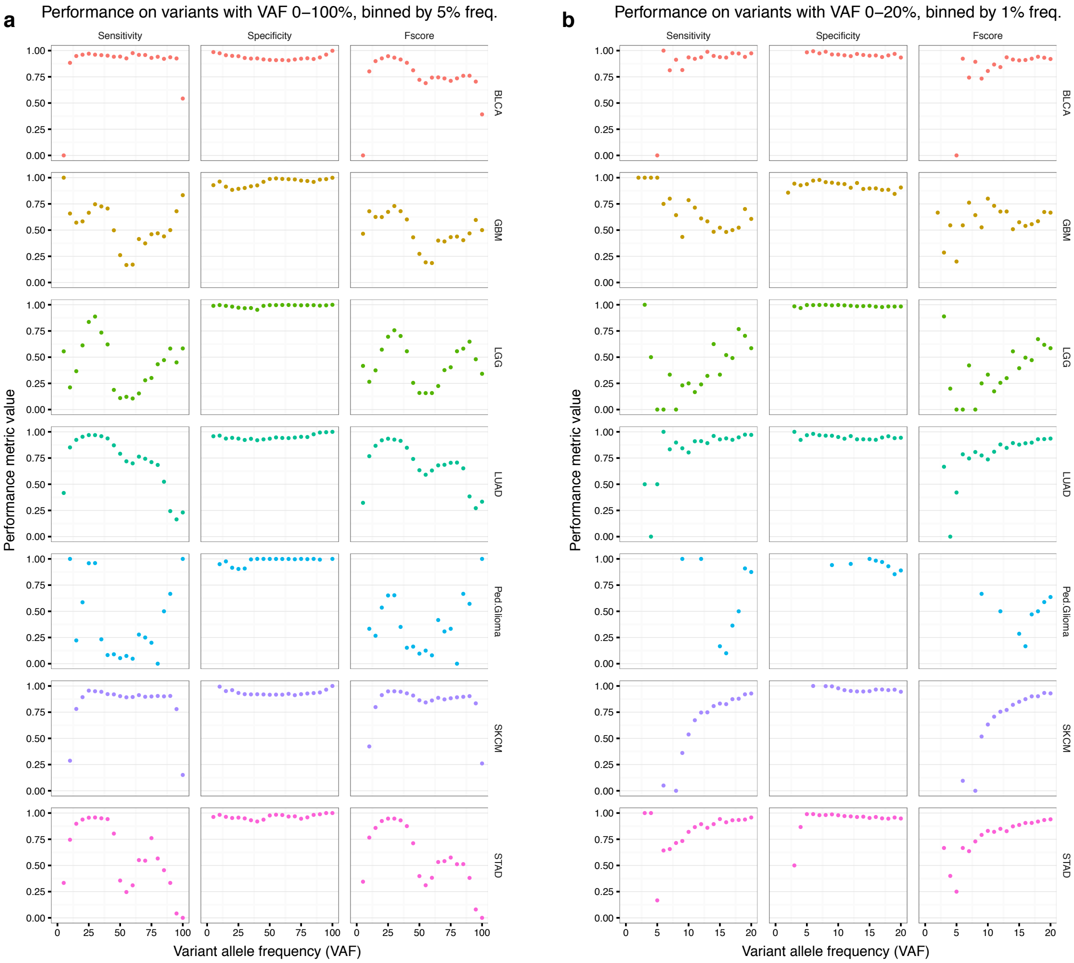


Figure 1.A.8: **Performance metrics at different variant allele frequencies**. Sensitivity, specificity, and F-score of variants (a) with VAF 0-100% binned by 5% or (b) VAF 0-20% binned by 1%.

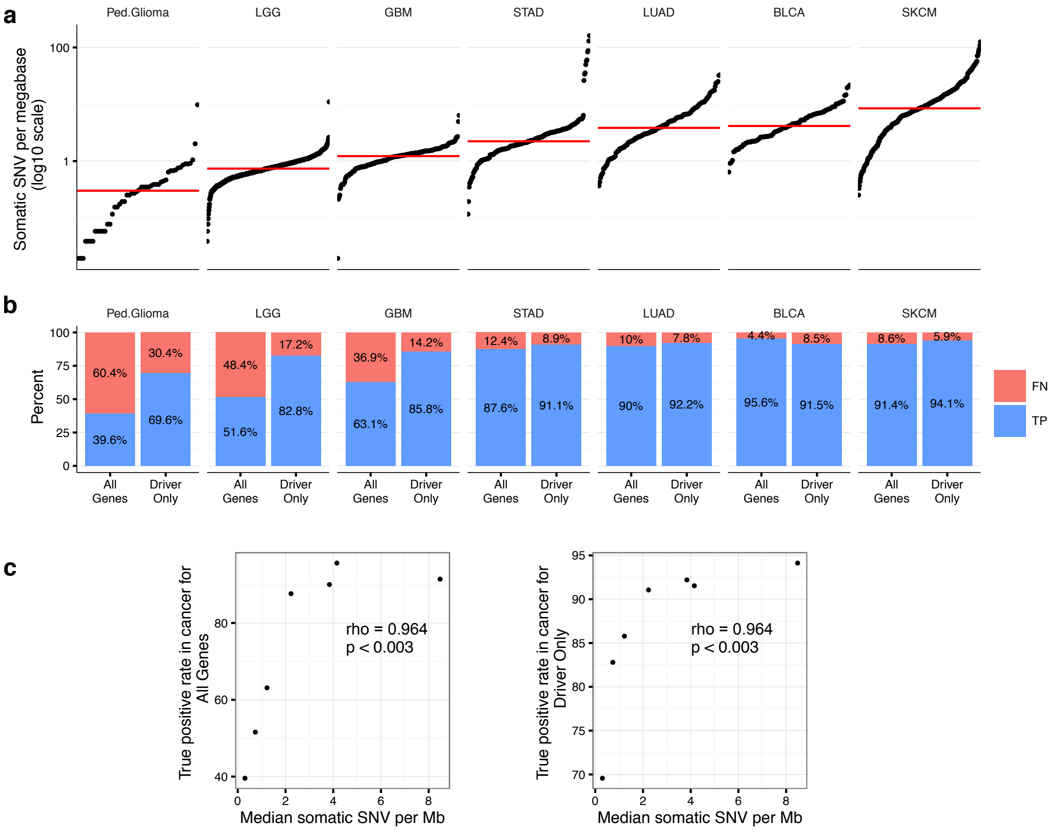


Figure 1.A.9: **TOBI sensitivity correlates with somatic SNV rate.** (a) Somatic SNV per megabase (Mb) for each cancer type. Vertical axis shows the number of somatic SNV per megabase on a log10 scale. Each point represents a tumor sample, red horizontal lines indicate median value for cancer; cancers ordered by increasing median number of somatic mutations. (b) Same as Figure 2a but cancers are ordered by increasing median number of somatic mutations. (c) Scatterplot of median somatic SNV per Mb versus true positive rate of nonsynonymous variants. Each point is a cancer type. Left panel uses true positive rate from all genes, right panel for driver genes only. P-value for Spearman correlation.

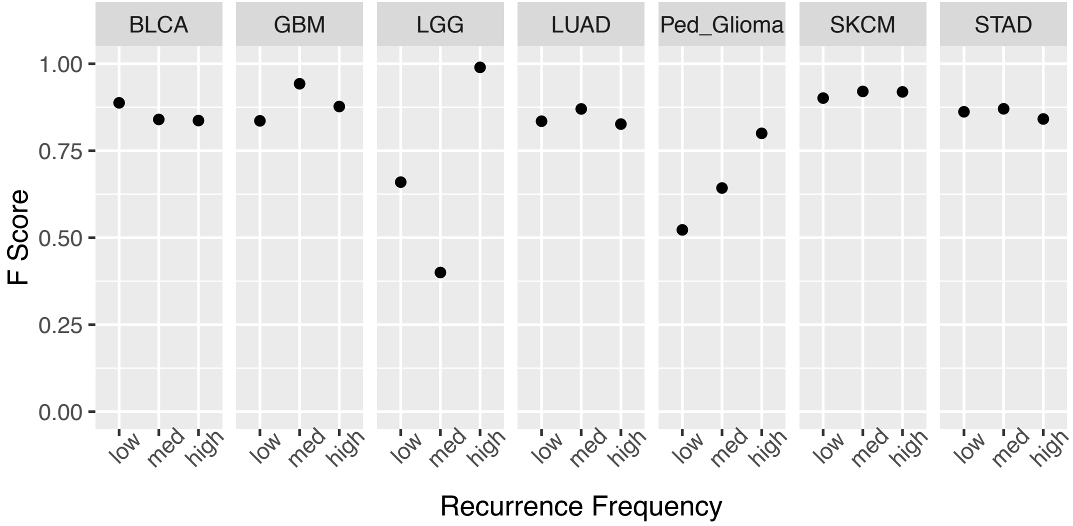


Figure 1.A.10: **F-score of TOBI prediction on genes binned by recurrence of true somatic mutations**. Recurrence bins defined as high (>20% of tumors), middle (10-20%) genes, and low (<10%)

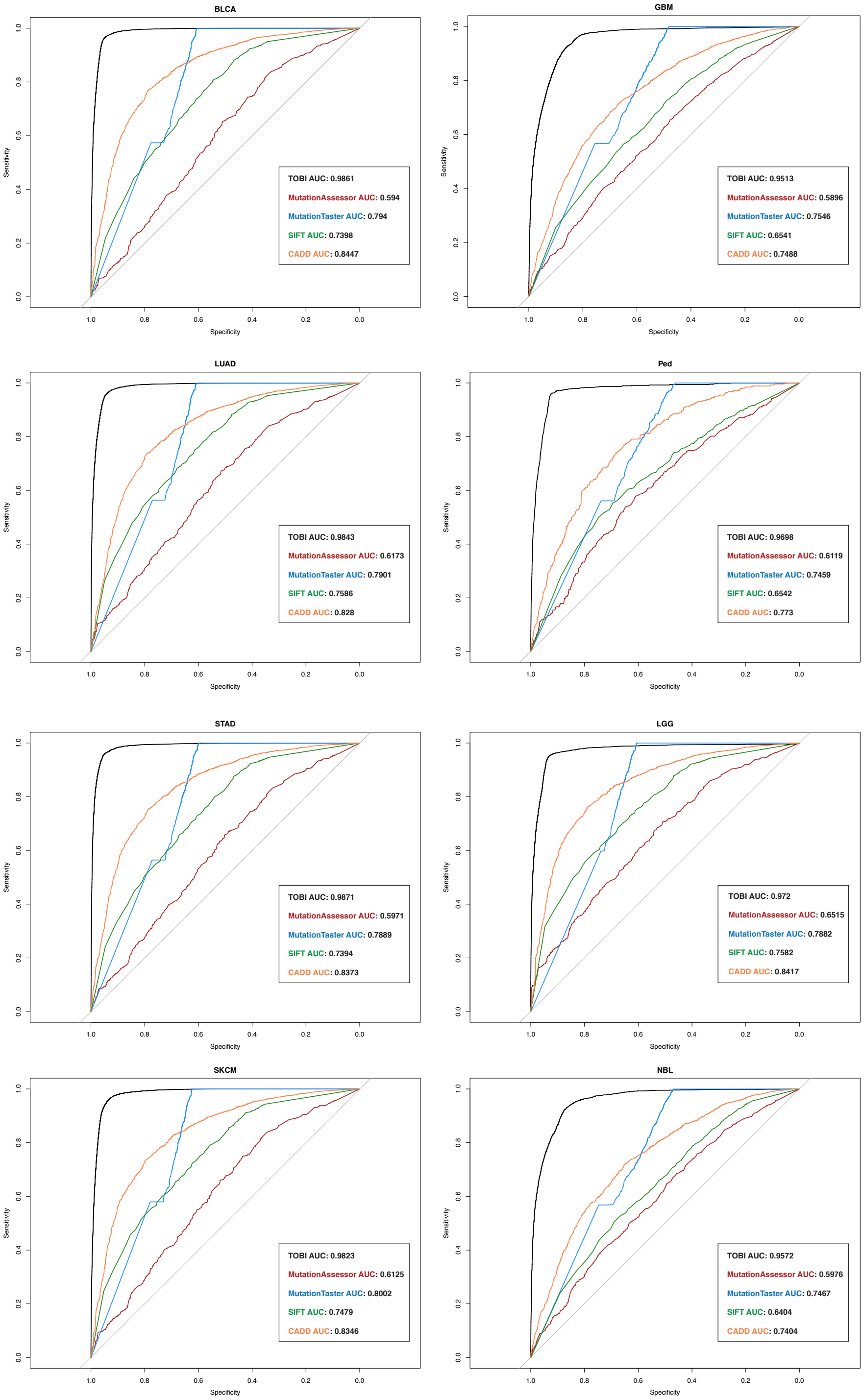


Figure 1.A.11: **TOBI somatic variant prediction outperforms other methods**. ROC curves comparing somatic variant prediction (synonymous and nonsynonymous) based on TOBI, CADD score, Mutation Assessor, SIFT and MutationTaster.

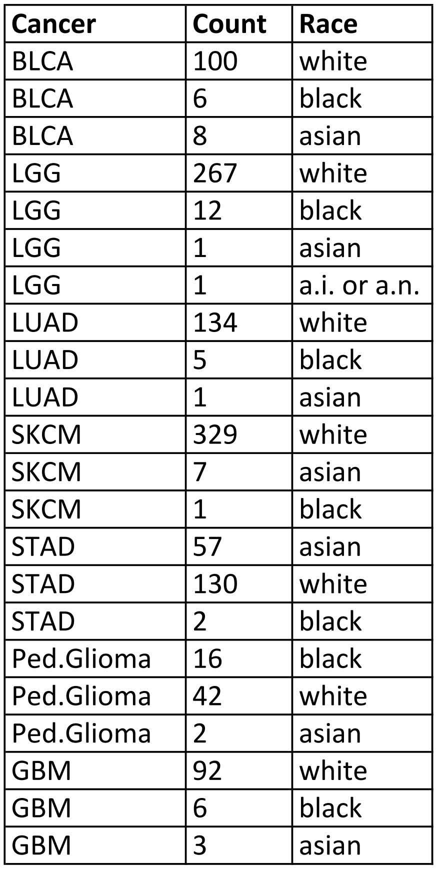
Cancer	Abbreviation	Total cases	Paired	Paired frozen	Paired FFPE	Hyper-mutator	Un-paired	Fig 1.2	Fig 1.3
Bladder urothelial carcinoma	BLCA	120	120	120	0	0	0	120	100
Brain lower grade glioma	LGG	512	512	512	0	0	0	286	492
Lung adenocarcinoma	LUAD	194	194	185	9	0	0	185	165
Skin cutaneous melanoma	SKCM	337	337	337	0	0	0	337	317
Stomach adenocarcinoma	STAD	280	280	219	0	61	0	219	199
Glioblastoma multiforme	GBM	184	184	184	0	0	0	104	176
Pediatric glioma	Ped.Glioma	142	74	74	0	0	68	74	54
	Subtotal		**1701**		**9**	**61**	**68**		
	Total samples	**1769**							

*20 cases in training set for each cancer

Table 1.A.1: Total cases per cancer. Also, the number paired frozen, paired FFPE, hypermutator, or unpaired cases. Number of cases used in figure 1 and figure 2 onward also indicated

Cancer	Method	Sensitivity	Specificity	Pos.Pred. Value	Neg.Pred. Value	Prevalence	Accuracy	False.Pos. Rate	False.Disc. Rate	AUC	Fscore	TP	TN	FP	FN	Total
Ped.Glioma	TOBI	0.382353	0.996529	0.65	0.989658	0.016581	0.986345	0.003471	0.35	0.967883	0.481481	26	4019	14	42	4101
GBM	TOBI	0.283019	0.945253	0.178571	0.969089	0.040355	0.918528	0.054747	0.821429	0.846481	0.218978	45	3574	207	114	3940
Ped.Glioma	VN	0.962963	0.144359	6.05E-05	0.999986	5.37E-05	0.144403	0.855641	0.99994	.	0.000121	78	217599	1289751	3	1507431
GBM	VN	0.989247	0.13374	7.29E-05	0.999995	6.39E-05	0.133794	0.86626	0.999927	.	0.000146	184	389444	2522512	2	2912142
Ped.Glioma	SomVarIUS	.	.	0	.	.	.	.	1	.	.	0	.	693	.	693
GBM	SomVarIUS	.	.	0.004222	.	.	.	.	0.995778	.	.	98	.	23115	.	23213

Table 1.A.2: Comparison of TOBI, SomVarIUS, and Virtual Normal Correction on six GBM and six Ped.Glioma samples.



Cancer	Count	Race
BLCA	100	white
BLCA	6	black
BLCA	8	asian
LGG	267	white
LGG	12	black
LGG	1	asian
LGG	1	a.i. or a.n.
LUAD	134	white
LUAD	5	black
LUAD	1	asian
SKCM	329	white
SKCM	7	asian
SKCM	1	black
STAD	57	asian
STAD	130	white
STAD	2	black
Ped.Glioma	16	black
Ped.Glioma	42	white
Ped.Glioma	2	asian
GBM	92	white
GBM	6	black
GBM	3	asian

Table 1.A.3: Patient counts after standardizing nomenclature

Chapter 2

Identification of germline variants with somatic variant features

2.1 Introduction

Germline DNA alterations can be oncogenic[51], with rare predisposing variants occurring in over a hundred genes[20]. Many predisposition genes are tumor suppressors. Oncogenic germline variants in tumor suppressor genes exhibit some common biological features, such as loss of heterozygosity and selection for the variant allele when comparing germline and tumor DNA[11]. There are fewer known gain-of-function predisposition variants, but kinases such as *RET* and *ALK* have oncogenic germline variants that typically occur in protein hotspots. Forty percent of these germline predisposition genes also undergo somatic mutation in cancer[20].

Review of recent literature

While predisposition genes were initially identified via studying families with cancer predisposition syndromes, recent research has focused on assessing germline variants in sporadic cancer cohorts. Analyses of both single cancer[30] and pan-cancer cohorts[31, 32, 34] have found that approximately 10% of patients with sporadic cancer have a potentially pathogenic germline variant in a known cancer predisposition gene. However, these cohort studies focus only on known cancer predisposition genes, precluding identification of oncogenic germline variants in less studied genes. Additionally, nomination of potentially oncogenic variants often involves clinical guidelines that do not account for tumor genomic information[37].

Here, we assess whether our TOBI framework predicts certain germline variants to have somatic features. Since TOBI learns biological features of somatic mutations throughout the exome, TOBI's predictions are exome-wide, rather than constrained to a known set of genes. We find that TOBI does classify certain germline variants as somatic, and refer to these germline variants as "somatic-like" germline variants (SLG variants) because SLG variants share biological features with true somatic mutations. SLG variants predicted by TOBI are enriched for known cancer predisposition genes. Finally, we find that in bladder carcinoma, TOBI predicts recurrent germline alterations in the Fanconi anemia pathway. We show that cases with germline or somatic alterations in the Fanconi anemia pathway are enriched for a DNA-repair deficiency somatic signature. These results suggest a potential inherited predisposition to bladder cancer and potential treatment with PARP inhibitors.

2.2 Results

Identification of “somatic-like” germline variants

Having established TOBI’s ability to identify somatic variants from tumor-only samples, we next assessed whether TOBI was capturing germline variants with somatic features. TOBI’s false positive (FP) variants could include germline variants that share features with true somatic variants, making them “somatic-like” germline (SLG) variants. SLG variants could be benign or oncogenic. Alternatively, false positive variants might be tumor-specific variants that were not previously published due to variability in somatic variant analysis[85].

First, we assessed TOBI’s overall false positive rate (FPR) in the cancer test sets. Since false positive variants may include SLG variants, we also calculated the FPR from applying the Ped.Glioma classification model to a set of 100 germline exomes from individuals without cancer sequenced by the 1000 Genomes Project[67]. The FPR in these 1000 Genomes individuals (median FPR 0.25%, range 0.15-1.62%) was significantly lower than the FPR in any of the cancer cohorts (Figure 2.1). The higher FPR in tumor cohorts suggests that some false positive calls represent somatic-like germline variants.

To identify SLG variants, we analyzed germline variant allele frequency (VAF) from 1,327 test cases in six cancers excluding GBM. VAF is the fraction of exome sequencing reads corresponding to the variant allele at a genomic site within a specific patient sample. To be classified as an SLG variant, a variant needed a germline VAF of at least 30% to decrease the probability that the germline variant represented tumor contamination or an artifact[31]. Since certain germline variants highly increase predisposition to cancer[31,

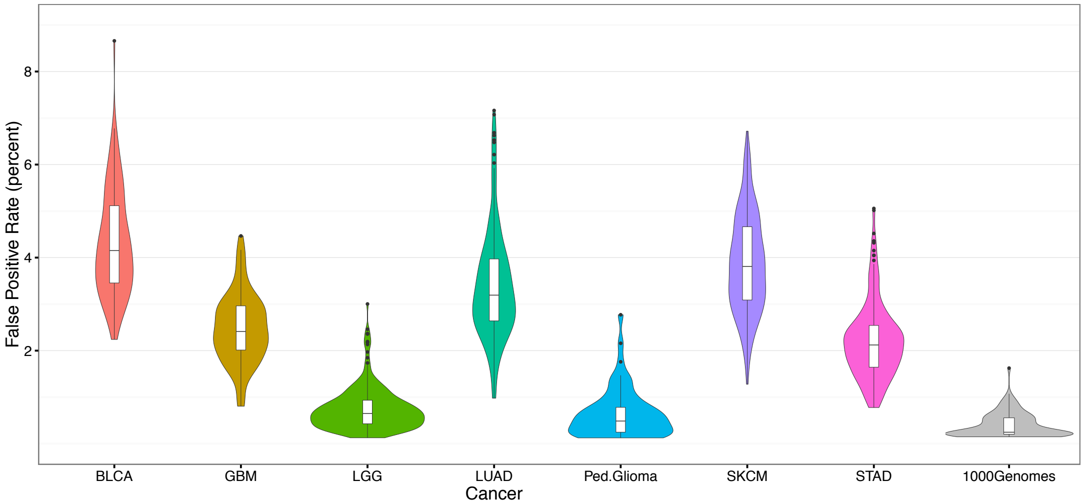


Figure 2.1: **TOBI false positive rate (FPR) in seven cancers and the 1000 Genomes project.** Distribution of FPR per case for each cancer was compared to the FPR from 100 cases from the 1000 Genomes project. For the seven cancers, FPR was calculated as the number of false positive TOBI somatic calls divided by the total number of true non-somatic variants in each case after filtering; for 1000 Genomes samples, false positives were defined as any variant predicted as somatic by TOBI, and FPR calculated by dividing the number of false positives by the total number of variants after filtering. LGG includes original 266 test cases. p-value calculated with the two-sided Wilcoxon–Mann–Whitney test.

32], we analyzed SLG variants for enrichment in a published list of 60 genes associated with autosomal dominant cancer-predisposition syndromes[31] ("AD genes", listed in Table 2.A.1), and found significant enrichment of AD genes in nonsynonymous SLG variants (p<1.53 x 10-10; Figure 2.2a).

Focusing on nonsynonymous FP variants in AD genes, we found at least seven cases with mutations in *CDH1*, *RB1*, *RET* or *TSC2*, and fifteen cases with mutations in the tumor suppressor *TP53* (Figure 2.2b). Certain inactivating mutations in tumor suppressors are heterozygous germline variants, but show loss of heterozygosity in the tumor[11]. Five of the *TP53* SLG variants exhibit evidence of loss of heterozygosity, with germline VAFs below 45% and tumor VAFs above 70% (Figure 2.3). In three Ped.Glioma cases, TOBI

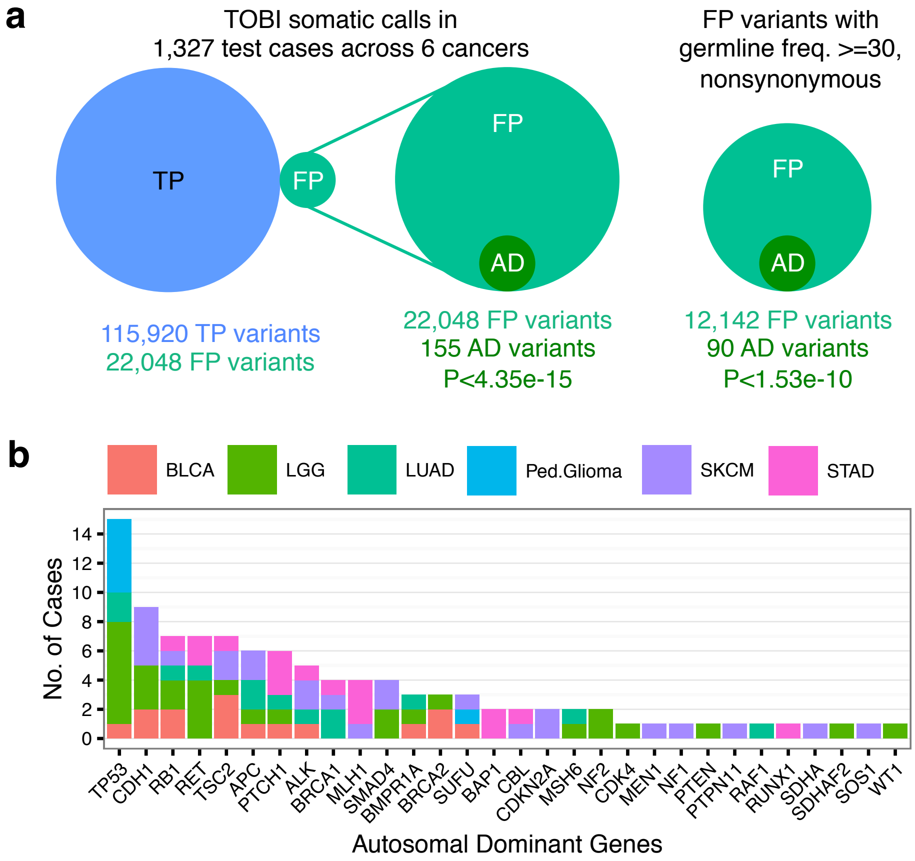


Figure 2.2: **"Somatic-like" germline (SLG) variants are enriched for genes associated with autosomal dominant cancer-predisposition syndromes (AD genes).** (a) Variants predicted as somatic by TOBI include 22,048 variants not reported as somatic in published analysis of 1,327 cases from five adult cancer types and pediatric glioma, with significant enrichment for AD genes in all FP variants and the subset of nonsynonymous variants with germline allele frequency greater than 30%. p-value from Poisson cumulative distribution. (b) Distribution of patient cases with FP variants in AD genes. Cancer abbreviations and color consistent with Figure 2.1.

predicted somatic *TP53* variants with tumor VAF greater than 65% and germline VAF of 0% (Fig. 3c; variants G105V, R175H, and R273C). Despite the high tumor VAF and low germline VAF, these variants were not published as somatic variants in outside tumor-normal analysis[70], illustrating that TOBI can identify somatic variants that may be inconsistently called.

Certain germline variants in cancer-associated genes correlate with earlier age of di-

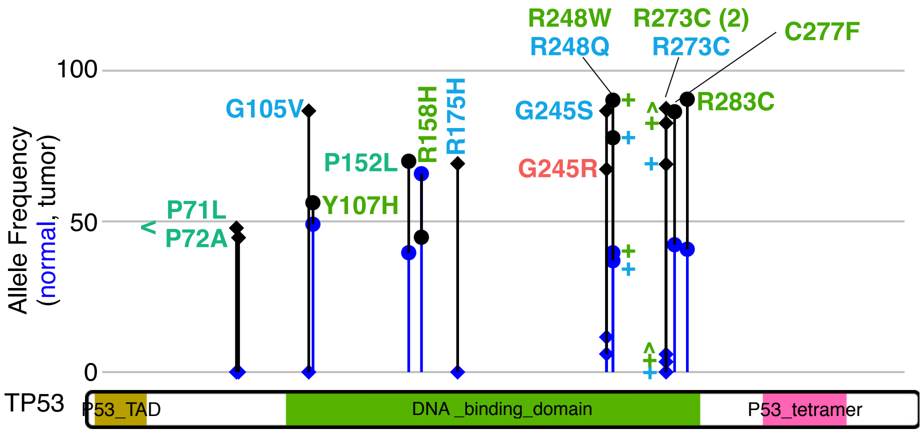


Figure 2.3: **FP variants in*TP53* domains**. Height of line represents allele frequency, with normal frequency at the blue point and tumor frequency in black. Circles indicate patients where normal frequency of variant is greater than or equal to 30%; diamonds indicate normal frequency less than 30%. Color of variant name corresponds to cancer color in Figure 2.1 and 2.2 (b). “<” indicates P71L and P72A occurred in same LUAD patient. “R273C (2)” indicates two patients with LGG had this variant. Colored “+” or “” indicate individual patient allele frequencies.

agnosis[32], so we analyzed whether the presence of nonsynonymous SLG variants in 565 cancer-associated genes associated with earlier age of diagnosis in any cancer type. In LGG, patients with cancer-associated SLG variants had significantly earlier ages at diagnosis (median 37 years vs. 41 years, p= 0.0013; Figure 2.4a; Figure 2.A.1). The most LGG cases had SLG variants in *TP53* (n=4), followed by *IDH1* (three cases: V71I [COSM96923], one case: R82K [COSM4169909]) and *RET* (Y791F [COSM1159820], I852M [COSM4573611], R982H [COSM1264016], T1038A [COSM4650197]). Many genes with SLG variants in LGG have also shown recurrent somatic mutations in prior analysis[47], including *TP53*, *IDH1*, *EGFR*, and *NF2* (Figure 2.4b).

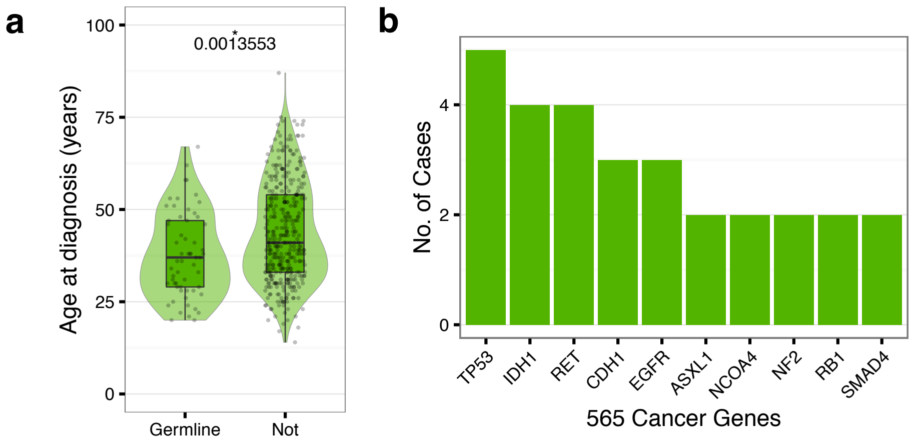


Figure 2.4: **SLG variants in low-grade glioma associated with earlier age of diagnosis.** (a) Distribution of diagnosis age in 492 LGG test set cases with or without nonsynonymous SLG variants in 565 cancer genes. For the violin plots, width of shape indicates density. In overlaid boxplots, the horizontal center line indicates the median (37 years vs. 41 years), upper and lower box edges correspond to the 25th and 75th percentiles, and the upper and lower whiskers extends from the closest box edge to the highest or lowest value within 1.5x the interquartile range, respectively. p-value calculated with two-sided Wilcoxon–Mann–Whitney test; * indicates p<0.01. (b) Cancer genes with recurrent nonsynonymous SLG in LGG.

Bladder cancer cases with inactivating mutations in Fanconi anemia pathway display somatic signature of BRCA-deficiency

Truncating germline alterations in cancer predisposition genes have been reported in 4-19% of cancer types[32]. Accordingly, we examined exome-wide SLG nonsense variants in each cancer type. Bladder carcinoma cases showed significant enrichment of SLG nonsense variants in the Fanconi anemia (FA) pathway based on pathway assessment with g:Profiler[86] (49 genes with SLG variants, 54 genes in FA pathway, 3 overlapping genes; p-value of 0.029 after multiple testing correction). The FA pathway normally performs DNA repair of interstrand crosslinks, which requires homologous recombination.[87]

We then assessed the overall occurrence of germline and somatic nonsense mutations predicted by TOBI in the FA pathway (Figure 2.5a). In bladder cancer, TOBI predicted these variants in 11% (11/100) of patients. Less than 2.5% of patients in any other cancer type had predicted nonsense FA variants. True somatic nonsense variants occurred in 6% of BLCA cases, affecting genes *BRCA2*, *FANCM*, *FANCE*, *REV3L*, and *SLX4*. Germline nonsense variants were predicted in 5% of BLCA cases, in the genes *BRCA2*, *FANCM*, and *FANCD2*. Several of these germline variants showed potential loss of heterozygosity based on increased VAF in tumor DNA compared to germline DNA (Figure 2.5b: *FANCM* R1931*, *BRCA2* Y3308*). Of note, *BRCA2* variant Y3308* has been associated with hereditary colorectal and breast cancer[88]. Mice ES cells with *BRCA2* Y3308* mutations showed hypersensitivity to ionizing radiation and crosslinking agents, as well as decreased homologous recombination efficiency[89]. Additionally, *FANCM* R1931* was associated with increased breast cancer risk and deficient DNA repair[90].

Figure 2.6 and table 2.A.2 describe TOBI's predicted nonsynonymous variants and published somatic copy number alterations affecting the FA pathway for this BLCA cohort.

Finally, we assessed whether BLCA cases with predicted FA pathway nonsense mutations had significantly different mutational signatures compared to wildtype cases. Using all somatic mutations published for 130 TCGA BLCA cases[46] including our 100 test cases, we generated trinucleotide mutational spectra that decomposed into four somatic signatures (figure 2.A.2). Cases with FA nonsense mutations were only enriched in the fourth signature (Figure 2.7), a somatic signature similar to the BRCA1/2-deficiency signature from a pan-cancer analysis (signature 3 in [91]). Enrichment of this somatic mutation signature in bladder cancer cases with nonsense FA variants suggests that these FA non-

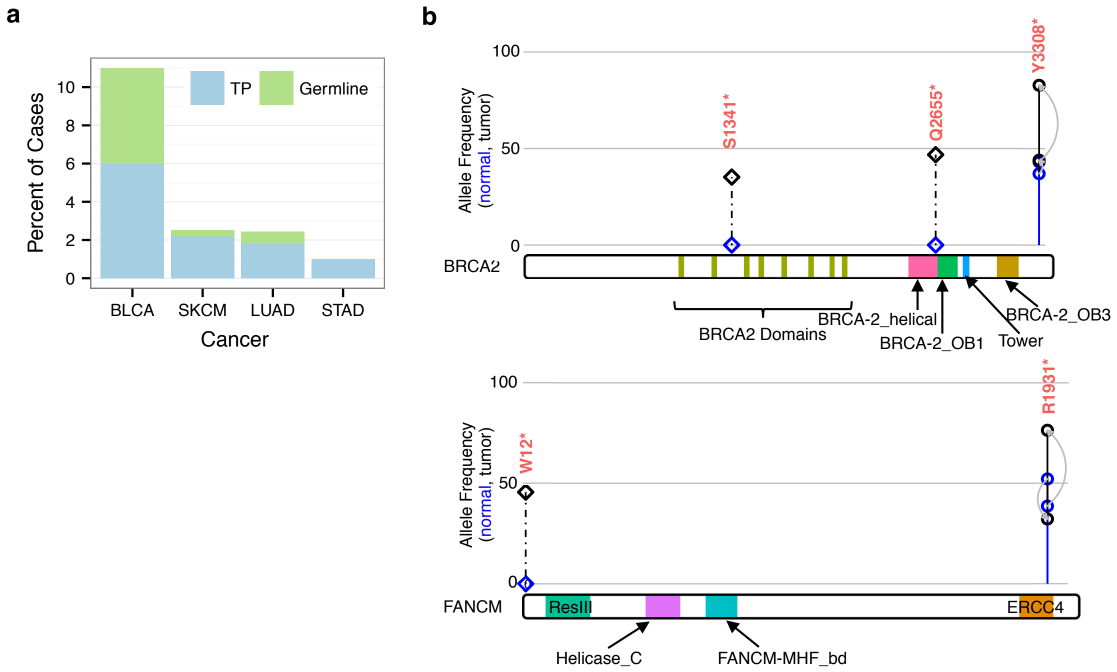


Figure 2.5: **Nonsense mutations in the Fanconi anemia pathway identified by TOBI.** (a). Percentage of test set cases with TOBI-somatic nonsense mutations; “Germline” indicates variant allele frequency (VAF) >= 30% in normal; “TP”, or true positives, were previously reported as somatic and have VAF < 30% in normal. Total number of test cases: 100 BLCA, 317 SKCM, 165 LUAD, and 199 STAD. (b) TOBI-somatic nonsense variants in BRCA2 and FANCM; diamond and dashed line indicate TP variants; solid line and circle are germline; grey arrows go from VAF in normal to tumor.

sense variants, whether somatic or germline, affect the bladder cancer somatic mutation landscape.

2.3 Discussion

In this chapter, we use TOBI for integrated germline and somatic analysis. When germline VAF information is available, TOBI can identify “somatic-like” germline variants. We assessed which somatic variants from TOBI were truly germline variants, and we found these “somatic-like” germline variants were enriched in genes associated with autoso-

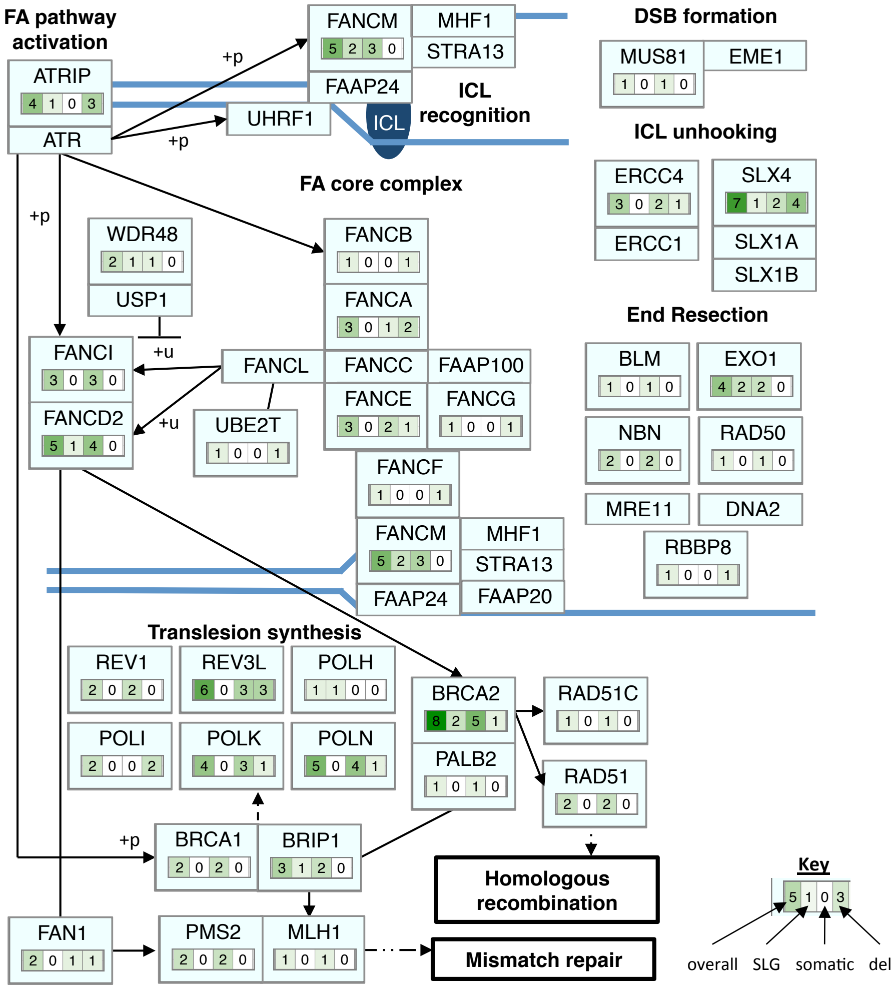


Figure 2.6: **Fanconi anemia pathway with number of altered bladder cancer cases shown per gene.**

mal dominant cancer-predisposition syndromes (Figure 2.2a). These SLG variants include oncogenic germline variants validated by outside groups, such as the *TP53* R248Q alteration confirmed as germline by tumor-normal analysis of a pediatric glioma case[70]. Additionally, SLG variants in cancer genes associated with earlier age of diagnosis in patients with low-grade glioma (Figure 2.4a), suggesting that TOBI's SLG variants are enriched for cancer-associated variants.

Analysis of bladder carcinoma cases using TOBI revealed largely unreported germline

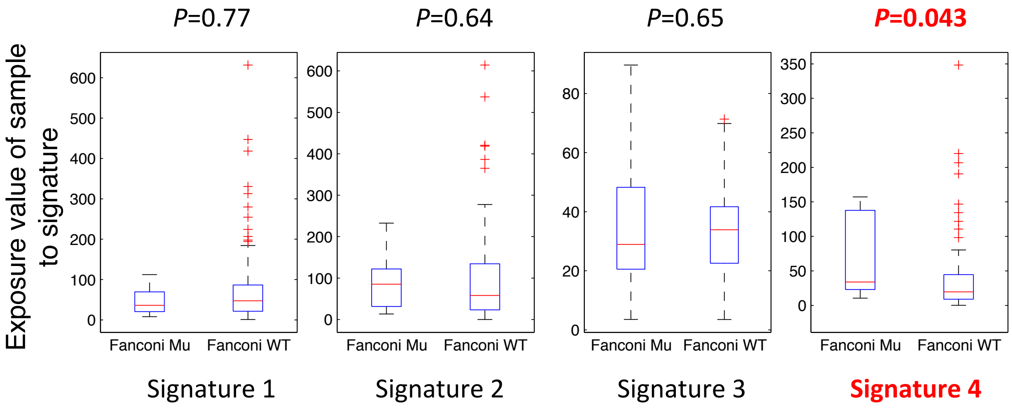


Figure 2.7: **Association of Fanconi anemia mutation status and somatic signature 4 in bladder cancer**. p-value calculated with rank sum test. Mu = mutant, WT = wildtype.

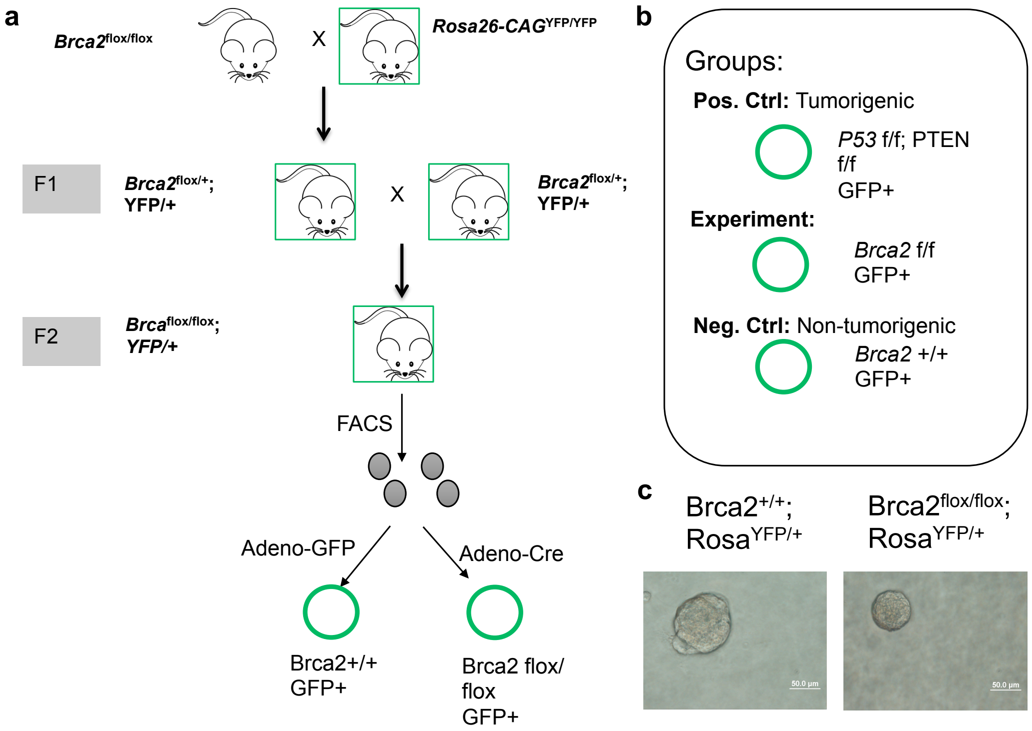


Figure 2.8: **Experimental design of murine *Brca2*-KO bladder organoids**. (a) Outline of mouse genetic crosses to generate $Brca2^{flox/flox;YFP/+}$ mice, fluorescence activated cell sorting of $Brca2^{flox/flox;YFP/+}$ mouse bladder urothelium, and delivery of adeno-Cre or adeno-GFP to organoids. (b) Genetics of experimental and control groups. (c) Light microscopy of organoids taken at 2-3 weeks. "f/f" indicates "flox/flox". Figure designed by Lijie Rong.

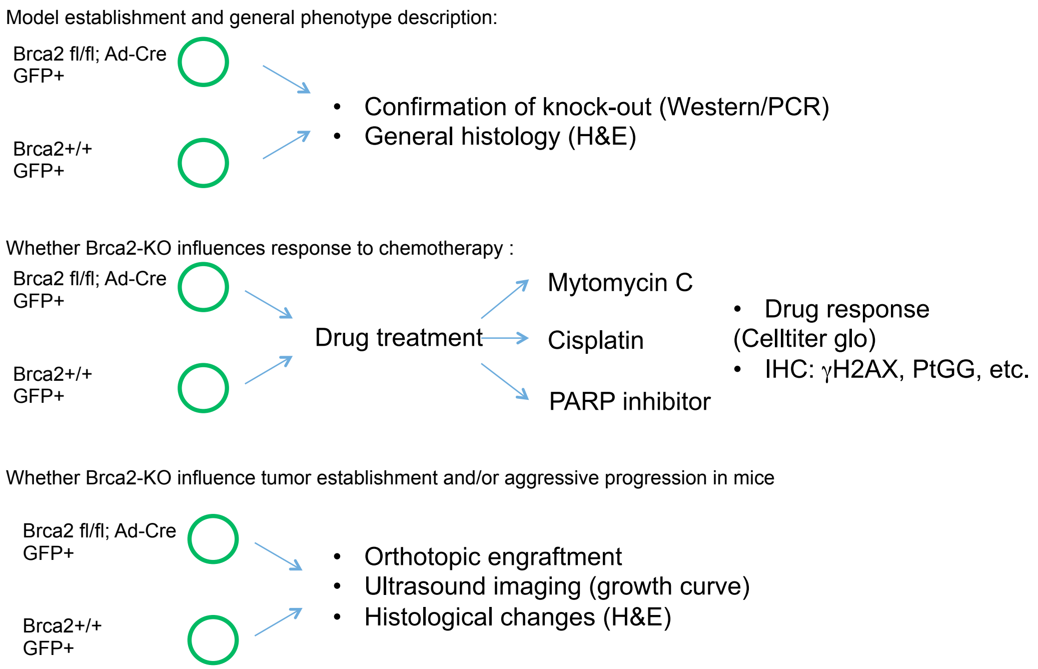

Model establishment and general phenotype description:
Brca2 fl/fl; Ad-Cre GFP+
Brca2+/+ GFP+
• Confirmation of knock-out (Western/PCR)
• General histology (H&E)
Whether Brca2-KO influences response to chemotherapy :
Brca2 fl/fl; Ad-Cre GFP+
Brca2+/+ GFP+
Drug treatment
Mytomycin C
Cisplatin
PARP inhibitor
• Drug response (Celltiter glo)
• IHC: γH2AX, PtGG, etc.
Whether Brca2-KO influence tumor establishment and/or aggressive progression in mice
Brca2 fl/fl; Ad-Cre GFP+
Brca2+/+ GFP+
• Orthotopic engraftment
• Ultrasound imaging (growth curve)
• Histological changes (H&E)

Figure 2.9: **Planned characterization of murine *Brca2*-KO bladder organoid.** Figure designed by Lijie Rong.

inactivating mutations in the Fanconi anemia pathway, suggesting a potential genetic predisposition in 5% of patients. Outside analysis of a 14-patient bladder tumor cohort[92] found a germline nonsense variant in *BRCA2*, but did not assess overall Fanconi anemia pathway mutations. Germline *BRCA2* nonsense mutations in bladder carcinoma may reflect the pan-cancer susceptibility attributed to germline *BRCA2* mutations in analysis of other adult cancers[32]. Future assessment of a larger BLCA cohort may reveal associations between germline FA mutations and clinical outcomes, similar to how an expanded cohort of prostate cancer patients revealed significantly more deleterious germline mutations in DNA repair genes in patients with metastatic versus localized prostate cancer[93].

Our integrated somatic and germline analysis identified nonsense FA pathway mutations in 11% of BLCA cases, suggesting a role for aberrant interstrand crosslink repair in bladder tumor development. Enrichment for a BRCA-deficiency somatic signature in these patients indicates similarity between FA mutant bladder cancers and BRCA-mutant breast cancers. However, further biological experiments would clarify the role of the FA mutations in bladder cancer. Treating BRCA-mutant breast cancers with PARP inhibitors improved patient outcome[94], so PARP inhibitors may also show increased effectiveness in bladder tumors with *BRCA2* or other FA mutations. Additionally, recent research in muscle-invasive bladder cancer found that the presence of tumor DNA alterations in *FANCC* (a member of the FA pathway), *ATM*, and *RB1* predicted beneficial response to cisplatin neoadjuvant chemotherapy[95]. Future research could determine whether FA nonsense mutations also predict beneficial response to Cisplatin, particularly given the beneficial response to cisplatin in patients with *BRCA1* mutant breast cancers[96]. We are collaborating on experiments with Cory Abate-Shen's laboratory to assess the role of germline Fanconi anemia pathway alterations in mouse bladder cancer organoids (experimental strategy described in Figures 2.8 and 2.9.

We recognize several limitations in germline variant analysis with TOBI. The previous chapter described several limitations for developing somatic variant classifiers with TOBI. In addition to those limitations, we recognize that TOBI's designation of SLG variants denotes "somatic-like" status, but does not differentiate oncogenic and benign germline variants. Finally, fully understanding the role of Fanconi anemia variants in bladder cancer requires experimental validation.

Overall, we believe this chapter has shown that in cases with matched normal DNA, the TOBI framework identifies germline variants that have somatic-like features and may inform tumor developments Integrated analysis of germline and somatic variants remains uncommon, making TOBI's identification of both somatic-like germline variants and somatic variants a unique strength. Applying the TOBI framework to seven cancer types illustrated that TOBI recovers known oncogenic variants of somatic and germline origin, and suggests a previously unreported role for inactivating mutations in the Fanconi anemia pathway in bladder cancer.

In the next chapter, we will apply the TOBI framework for integrated germline and somatic analysis to a cohort of patients with pediatric acute lymphoblastic leukemia.

2.4 Methods

Germline variant analysis and clinical data associations

Germline VAFs were available in Protected Mutation vcf files for five TCGA cancers (BLCA, LGG, LUAD, SKCM, STAD). For tumor-normal pediatric glioma cases, germline VAFs were determined using the SAVI variant caller[44]. We used SAVI to perform joint variant calling on normal and tumor pediatric glioma samples, then assessed whether variants predicted as somatic by TOBI in tumor-only analysis were also present in the germline SAVI calls at germline VAF > 30%. For enrichment of gene sets in false positive (FP) variants, the Poisson cumulative distribution was calculated for each gene set, with g total genes and n FP variants in those genes from a cancer cohort with N variants found in G genes, as the probability of a value greater than $(n-1)$ with $lambda = (g*N)/G$ using the R ppois function. Lists of 60 autosomal dominant genes and 565 cancer-associated genes are from [31], Supplementary Appendix 2. Protein domain names and coordinates from PFAM[97].

We compared the distribution of diagnosis age for cases with or without SLG variants using the Wilcoxon–Mann–Whitney test in R.

Fanconi anemia pathway enrichment

g:Profiler[86] analysis of BLCA nonsense SLG variants was run using defaults (Significant only; Hierarchical sorting; Numeric IDs treated as: WIKIGENE_ACC; Significance threshold: g:SCS threshold; Statistical domain size: Only annotated genes.) Multiple testing correction for p-values was calculated using the ontology-focused correction method g:SCS as previously described in the g:Profiler manuscript[86]. FA pathway in Figure 2.6 was modified from the KEGG FA pathway and [98]. Bladder carcinoma CNV data was retrieved from cBioPortal.

Mutation spectra and signatures

The non-negative matrix factorization approach developed by Alexandrov et al.[91] was applied to infer the mutational signatures of bladder cancer, using their supplied software package (`http://www.mathworks.com/matlabcentral/fileexchange/38724`). All 130 cases of bladder cancer from TCGA were used to generate signatures.

2.5 Acknowledgements

This chapter is a reproduction, in part, of a publication in npj Genomic Medicine by Madubata, C.J., Roshan-Ghias, A., Chu, T., Resnick, S., Zhao, J., Arnes, L., Wang, J. and Rabadan, R [83]. This work was funded in part by R01 CA185486-01, R01 CA179044-

01A1, and NIH U54 CA193313; J.W. is supported by Precision Medicine Fellowship (UL1 TR000040); A.R. was supported by the Swiss National Science Foundation Postdoctoral Fellowship; C.J.M. is supported by the NIH Medical Scientist Training Program (5T32 GM07367) and the National Research Service Award Individual Predoctoral Fellowship (F31CA210607). This study makes use of data generated by the St. Jude Children's Research Hospital—Washington University Pediatric Cancer Genome Project (`http://explore.pediatriccancergenomeproject.org/`). We also acknowledge The Institute of Cancer Research, McGill University, the German Cancer Research Center (DKFZ), and The Cancer Genome Atlas (managed by the NCI and NHGRI, `cancergenome.nih.gov`) as sources of primary data. We recognize the contribution of all specimen donors. We thank Hossein Khiabanian for helpful discussions during project design. Lijie Rong from Cory Abate-Shen's laboratory is performing experiments on murine *Brca2*-KO bladder organoids.

2.A Appendix

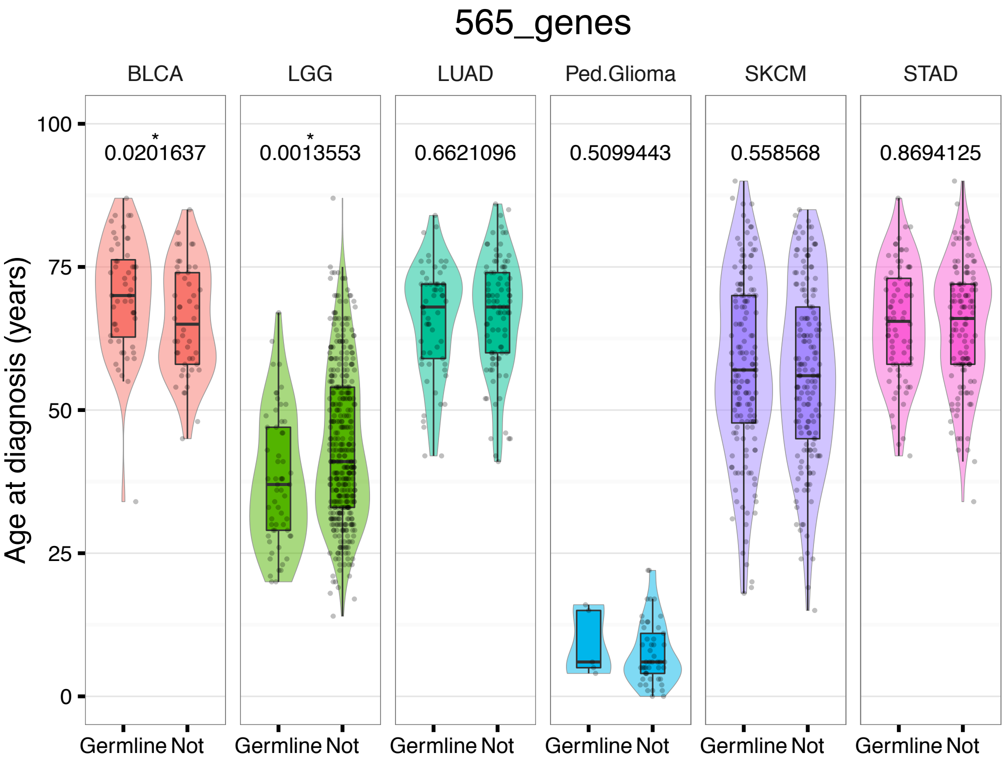


Figure 2.A.1: **Age distribution based on cancer-associated SLG in seven cancer types.** Colors consistent with Figure 2.1. p-value calculated with the two-sided Wilcoxon–Mann–Whitney test; * indicates p<0.01.

ALK
APC
BAP1
BMPR1A
BRAF
BRCA1
BRCA2
CBL
CDC73
CDH1
CDK4
CDKN1C
CDKN2A
CEBPA
DICER1
EPCAM
FH
GATA2
HRAS
KRAS
MAP2K1
MAP2K2
MAX
MEN1
MLH1
MSH2
MSH6
NF1
NF2
NRAS
PALB2
PAX5
PHOX2B
PMS2
PRKAR1A
PTCH1
PTEN
PTPN11
RAF1
RB1
RET
RUNX1
SDHA
SDHAF2
SDHB
SDHC
SDHD
SHOC2
SMAD4
SMARCA4
SMARCB1
SOS1
STK11
SUFU
TMEM127
TP53
TSC1
TSC2
VHL
WT1

Table 2.A.1: Genes associated with autosomal dominant cancer-predisposition (AD genes). List of genes from Supplementary Appendix 2 of [31]

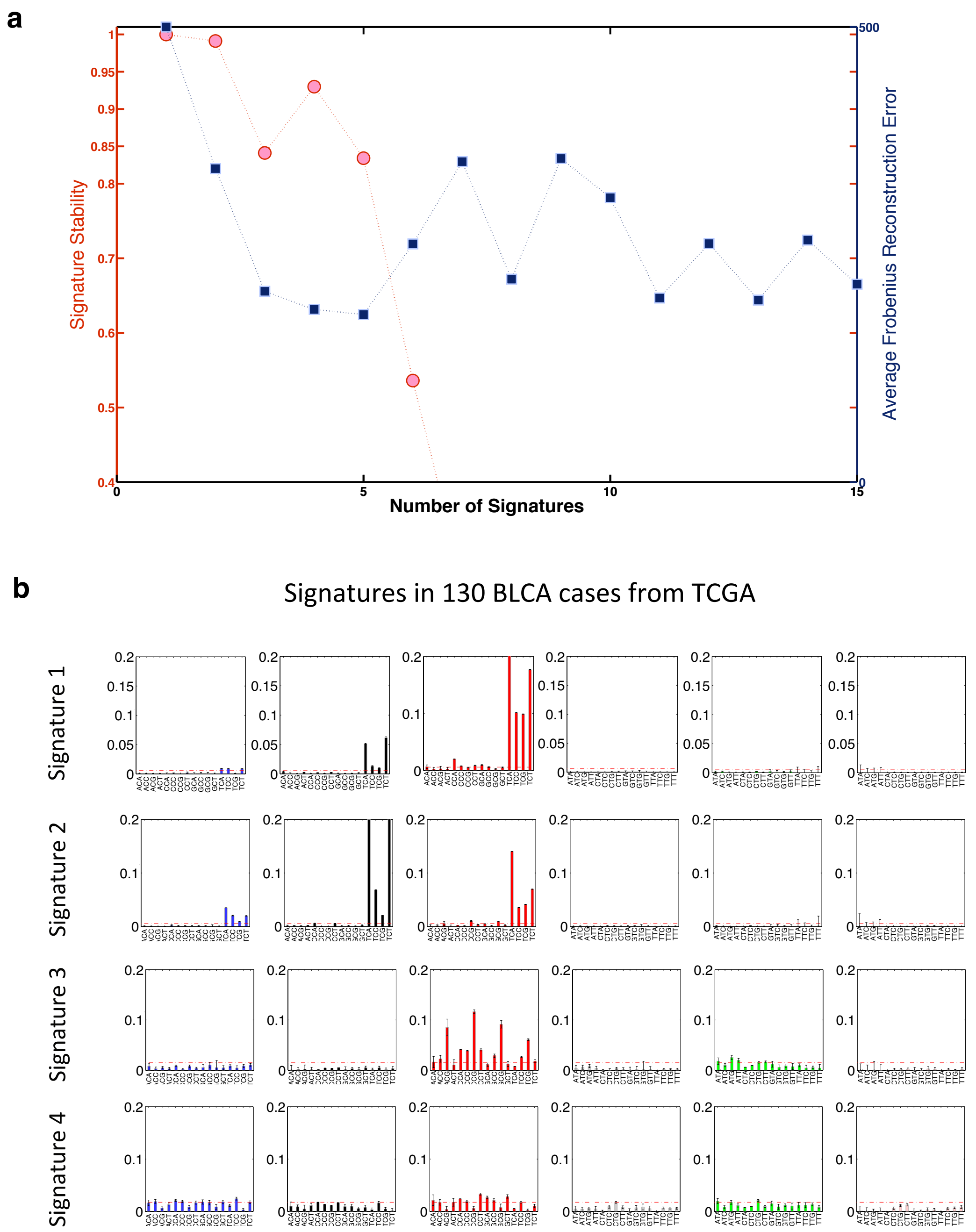
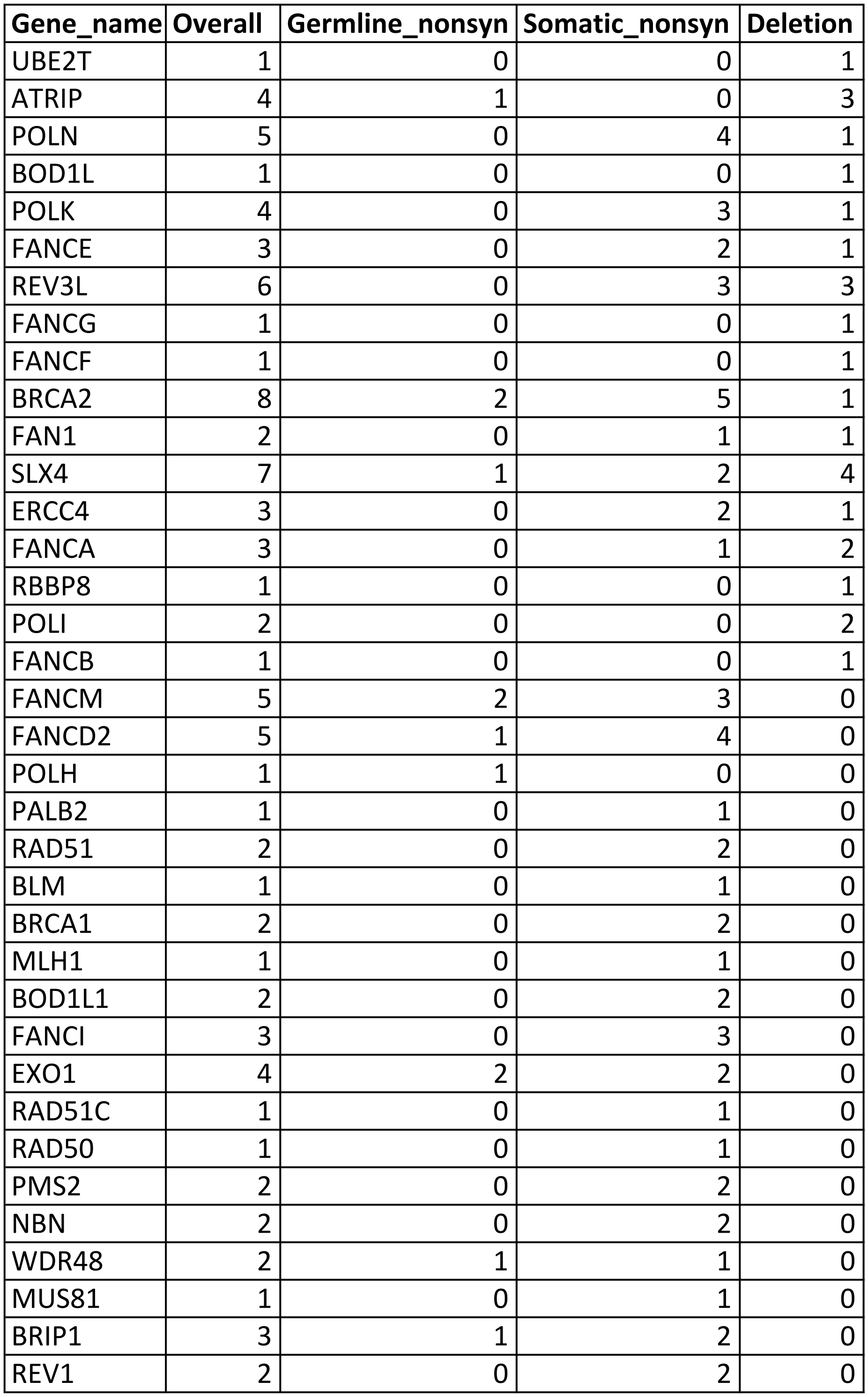


Figure 2.A.2: **Four somatic signatures for bladder carcinoma**. (a) Selection of k=four somatic signatures for BLCA maximizes stability and minimizes error. (b) Somatic signatures from TCGA BLCA cohort (TCGA, Nature 2014), generated using techniques from [91]. Signature 4 resembles the BRCA somatic signature described in [91].



Gene_name	Overall	Germline_nonsyn	Somatic_nonsyn	Deletion
UBE2T	1	0	0	1
ATRIP	4	1	0	3
POLN	5	0	4	1
BOD1L	1	0	0	1
POLK	4	0	3	1
FANCE	3	0	2	1
REV3L	6	0	3	3
FANCG	1	0	0	1
FANCF	1	0	0	1
BRCA2	8	2	5	1
FAN1	2	0	1	1
SLX4	7	1	2	4
ERCC4	3	0	2	1
FANCA	3	0	1	2
RBBP8	1	0	0	1
POLI	2	0	0	2
FANCB	1	0	0	1
FANCM	5	2	3	0
FANCD2	5	1	4	0
POLH	1	1	0	0
PALB2	1	0	1	0
RAD51	2	0	2	0
BLM	1	0	1	0
BRCA1	2	0	2	0
MLH1	1	0	1	0
BOD1L1	2	0	2	0
FANCI	3	0	3	0
EXO1	4	2	2	0
RAD51C	1	0	1	0
RAD50	1	0	1	0
PMS2	2	0	2	0
NBN	2	0	2	0
WDR48	2	1	1	0
MUS81	1	0	1	0
BRIP1	3	1	2	0
REV1	2	0	2	0

Table 2.A.2: Counts of Fanconi anemia alterations in bladder carcinoma

Chapter 3

Integrated germline and somatic analysis of a large pediatric ALL cohort

3.1 Introduction

Pediatric acute lymphoblastic leukemia epidemiology and treatment

Acute lymphoblastic leukemia (ALL) is the most common cancer in children, with ALL representing 26% of new cancer diagnosis in children 0-14 years of age and 8% of cases in adolescents ages 15-19[99]. Symptoms at diagnosis include fatigue and pallor from anemia, bleeding or bruising from low platelets, and infection due to low neutrophil counts [100]. ALL arises from two lymphocyte lineages: B cell (B-ALL) or T cell (T-ALL). B-ALL is more common, representing 85-90% of pediatric cases, while T-ALL occurs in 10-15% of cases and is more common in boys[101]. Recent ALL five-year survival rates approach 90% across both lineages, with survival steadily increasing since 1975. However, among the 10-20% of patients with ALL who relapse, there is a cure rate of less than 40%[102]. Given this low cure rate and the high number of initial ALL cases, relapsed ALL is the most common cause of cancer death in children [103].

ALL risk stratification uses patient age and WBC count at diagnosis, morphological and cytogenetic characterization of leukemia cells, and assessment of early treatment response [104]. High-risk criteria include age <1 year or >10 years, T-ALL, extreme hypodiploidy (less than 44 chromosomes), translocation t(9;22) leading to a *BCR/ABL* fusion, and induction failure[105].

Contemporary ALL therapy consists of several discrete phases: induction, consolidation, and maintenance [100]. The goal of induction therapy is remission, or the restoration of normal hematopoiesis and the removal of all clinically detectable leukemia burden [106]. Induction therapy lasts 4 to 6 weeks and typically involves daily corticosteroids, weekly administration of vincristine, and L-asparaginase, with possible intrathecal chemotherapy [107]. Patients with Philadelphia chromosome t(9;22) positive ALL have imatinib or dasatinib added to their regimen [108]. Early response to induction therapy, defined as less than 1000 blasts/μL during the first week of treatment, is a favorable prognostic factor [104].

The next stage of ALL therapy after remission is the 6-8 month consolidation phase[100]. Consolidation therapy aims to consolidate remission and prevent leukemia regrowth, especially CNS leukemia or drug resistant disease. To achieve this, patients receive drug combinations with mechanisms of action and schedules designed to minimize development of resistance [109]. Chemotherapies include methotrexate, cytarabine, and cyclophosphamide [110].

After completing the induction and consolidation phases, patients begin maintenance therapy. Maintenance therapy consists of longterm (24 to 36 months) daily 6-mercaptopurine (6-MP) and weekly methotrexate [111]. Recent trials suggest adding pulses of vincristine and steroids to maintenance therapy to increase survival[112]. Given the duration of maintenance therapy, patient adherence is variable. Patients with adherence rates of less than 90% have a 3.9-fold increased risk of relapse[113].

During and after treatment, patients are monitored for treatment side effects and disease relapse. In this monitoring, the factor most associated with prognosis is minimal residual disease (MRD, submicroscopic levels of leukemia), with higher MRD associating with leukemia relapse and lower survival[114]. Relapsed ALL is more resistant to chemotherapy, due to selection of a minor subclone present at diagnosis or from acquisition of drug resistance mutations during chemotherapy [115]. After relapse, predictors of outcome include lymphocyte phenotype, time to relapse, and site of relapse, with the Children's Oncology Group (COG) and the Berlin-Frankfurt-Munster Group (BFM) both assigning worse prognosis to early relapse, T-cell phenotype, and relapse in the bone marrow versus isolated extramedullary sites[116].

The strongest predictor of survival is time to relapse after diagnosis, where patients who relapsed less than 18 months after diagnosis had a five-year survival rate of 21% [117]. When a patient relapses, they receive salvage therapy partially based on agents used in the induction, consolidation, or maintenance phases; a common treatment strategy is alternating short-course multi-agent chemotherapy (systemic and intrathecal) and standard maintenance therapy[115]. Patients with high risk relapsed ALL additionally undergo hematopoietic stem cell transplantation (HSCT) [115]. However, the survival rate for these heterogeneous salvage therapies remains around 35%-40%[118]. Addressing the low survival rates in relapsed pediatric ALL could involve generating more targeted therapies or clinical strategies to prevent relapse. Both of these strategies can be informed by the genomic landscape of pediatric ALL at diagnosis, relapse, and within the germline genome.

Somatic and relapse mutations in pediatric ALL

Somatic mutations in diagnosis and ALL samples have been assessed through multiple next-generation sequencing studies[119–123]. In B-ALL, recurrent diagnosis mutations often occur in genes encoding transcriptional regulators of B-lineage differentiation [124], such as *PAX5* (30% of cases), *IKZF1–3*, and *EBF1*. Additional genes mutated in B-ALL include *KRAS*, *NRAS*, *FLT3*. In T-ALL, T-cell development genes are frequently mutated, particularly *NOTCH1* that is mutated in 60% of T-ALL cases[101]. Both T-ALL and B-ALL have recurrent alterations in *JAK-STAT* pathway genes and *CREBBP*. Pediatric ALL cases have a low somatic burden at diagnosis compared to adult cancers[91, 125].

Certain genomic alterations at diagnosis correlate with patient relapse, and could be used for risk-stratification or targeted therapy. Deletion of nuclear receptor subfamily 3 group C member 1 (*NR3C1*) is associated with relapse in *ETV6*-textitRUNX1 rearranged leukemia [126] In B-cell–progenitor ALL, *IKZF1* deletion or mutation associates with relapse [121, 127]. Mutations in *CREBBP* are both associated with relapse and linked to resistance to glucocorticoids[128], where glucocorticoids are a key component of ALL therapy.

Many studies have compared genomic alterations at diagnosis and relapse, and have found certain genes recurrently mutated at diagnosis are further selected for at relapse. *TP53* is acquired at relapse in many cases[129, 130], with *TP53* mutations associated with worse prognosis[131]. Relapsed clones often gain mutations in *CREBBP*[119] and other epigenetic regulators (*SETD2*, *MSH6*, *KDM6A* and *MLL2*)[132]. Mutations in the glucocorticoid-receptor gene *NR3C1* are also commonly found at relapse[133], showcasing another mechanism of resistance to combination chemotherapy.

The most prominent genetic events at relapse involve relapse-specific mutations in *NT5C2* [60, 134], a nucleotidase that metabolizes purine nucleotides. The standard purine synthesis pathway is essential for effective ALL therapy by converting the prodrugs 6-MP and 6-thioguanine into cytotoxic thioguanine nucleotides. However, *NT5C2* exports purine nucleoside monophosphates, and gain-of-function mutations at relapse result in decreased intracellular concentrations of cytotoxic therapies[60, 134, 135]. Expression of these mutations in ALL cells causes resistance to 6-MP [60, 134]. At relapse, gain-of-function mutations in *NT5C2* are present in 5% of B-ALL and 20% of T-ALL cases in one study[119] and 45% of B-ALL cases in another study [123]. NT5C2 mutations are relapse-specific, possibly due to negative selection for the purine substrate imbalance created by *NT5C2* mutations [135]. Resistance to 6-MP also arises in relapsed ALL with activating mutations in *PRPS1*, a gene encoding the enzyme that begins purine and pyrimidine synthesis; these mutations prevent 6-MP's entry into the purine salvage pathway[122].

Known germline variants associated with ALL

While diagnosis and relapse mutations are acquired during ALL development, certain germline variants present at birth also influence a patient's disease progression and response to treatment. For example, two genome-wide association studies assessing the B-ALL subtype Philadelphia chromosome–like ALL (Ph-like, or *BCR-ABL1*-like) found SNPs in *GATA3* associated with Ph-like ALL susceptibility[136, 137]. The *GATA3* locus also associated with genomic alterations at diagnosis (*CRLF2* rearrangements, *JAK* mutations, and *IKZF1* deletion). Additionally, patients with the *GATA3* variant were at higher risk of poor early treatment response and relapse[136, 137]. GWAS has identified other common variants associated with ALL prognosis, including loci associated with ALL susceptibility (in *ARID5B*, *IKZF1*, *CEBPE*, *CDKN2A*, *CDKN2B*, *PIP4K2A*, *BMI1*, *GATA3*, *TP63*), treatment toxicities (*NUDT15*, *ACP1*, *GRIA1*, *HLA-DRB1*, *ASNS*, *CBR3*, *HAS3*, *CEP72*, *SLCO1B1*), and treatment outcome (*TPMT*, *IL15*, *PYGL*, *PDE4B*, and *GATA3*)[6]. While many of these variants are noncoding, one GWAS identified a missense variant in *CDKN2A* (p.A148T) associated with ALL development, and experimental validation of this varianted revealed allele specific expression in patient tumor samples and increased leukemia cell growth *in vitro*[138].

Rare variants related to pediatric ALL or myeloid leukemia have been identified through family studies. Studying families with inherited ALL led to the identification of these ALL predisposition genes: *SH2B3* [27], *PAX5*[21], *ETV6* [22–25], and *IKZF1* [26]. A recent study of patients with multiple de novo leukemia diagnoses identified germline mutations in *TYK2*, a member of the JAK tyrosine kinase family[139]. Genes associated with inherited myeloid leukemia include *RUNX1*[140] and *CEBPA* [141]. Additionally, certain inherited Mendelian disorders confer increased risk of ALL, such as *NSD1*[142](associated with Sotos syndrome), *NF1* [143] (neurofibromatosis 1), and *TP53* (Li-Fraumeni syndrome). Assessment of *TP53* germline variants in patients with sporadic ALL and no Li-Fraumeni diagnosis reveals germline variants in a fraction of cases. Among patients with pediatric low-hypodiploid ALL and *TP53* mutations, 43.3% of these

TP53-mutant cases had germline *TP53* mutations[28]. Targeted sequencing of 3,801 children with ALL found 22 rare, putatively pathogenic germline variants in *TP53* associated with a higher risk of second cancers[144].

Plan for this chapter

Since clonal evolution of leukemia occurs within the context of each patient's germline genetics, we propose integrated analysis of potentially pathogenic germline variants and somatic variants. Here, we investigate the germline and somatic variants in a cohort of over 600 pediatric patients with ALL. First, we assess the somatic landscape of the cohort and investigate associations between diagnosis alterations and relapse status. We find that known driver genes are recurrently mutated in the cohort, with *WT1* mutations associated with relapse status. Using patients with germline, diagnosis, and relapse samples available, we describe the temporal order of somatic mutations in ALL clones. Finally, we assess two strategies to nominate candidate oncogenic germline variants. Strategy one uses the TOBI framework to predict SLG variants across the exome, and strategy two assesses rare germline variants in known leukemia and cancer predisposition genes. We show that TOBI captures ALL driver genes, but has moderate performance compared to adult cancers, and we identify potential pathogenic variants in predisposition genes in a subset of patients. These results suggest strategies for nominating candidate germline oncogenic variants.

3.2 Results

Somatic mutations in relapsed and non-relapsed ALL

To assess the landscape of somatic ALL mutations in pediatric patients with or without ALL relapse, we performed somatic variant analysis on a set of 539 patients with pediatric ALL (TARGET ALL Phase 2, dbGAP accession number phs000464[121]) and combined these results with previously published diagnosis variants from 88 patients with relapsed pediatric ALL[119, 122, 123]. In this section, somatic variants refer to mutations present in the ALL diagnosis sample, and exclude variants specific to relapse samples. Based on available clinical data, 133 patients (21.2%) in this cohort had a reported ALL relapse. This cohort contains both B-ALL (437 cases, 69.7%) and T-ALL (190 cases, 30.3%) diagnoses.

Across all 627 samples, the most recurrently altered genes based on SNVs and indels (Figure 3.1) were *NOTCH1* (17.4%, 109 of 627 cases), *KRAS* (13.6%, 85 cases), *NRAS* (12.6%, 79 cases), *FBXW7* (6.4%, 40), *PHF6* (4.9%, 31), *PAX5* (4.9%, 30), and *FLT3* (4.6%, 29 of 627 cases). The recurrence of *NOTCH1* mutations in this majority B-ALL cohort reflected the high prevalence of *NOTCH1* variants in T-ALL (106 of 190 T-ALL cases, 55.8% of T-ALL). Looking specifically at T-ALL, other known driver genes were recurrently mutated, including *FBXW7* (20.5%, 39 cases of 190 T-ALL cases), *PHF6* (15.8%, 30), *WT1* (10.5%, 20), *DNM2* (11.6%, 22), *PTEN* (10%, 19 cases), and *NRAS* (9.5%, 18 of 190 cases). Similarly, the 437 patients with B-ALL had recurrent alterations in known B-ALL drivers including *KRAS* (17.6% of B-ALL cases, 77 cases), *NRAS* (13.9%, 61), *PAX5* (6.9%, 30), *FLT3* (6.4%, 28), *PTPN11* (4.6%, 20), and *JAK2* (3.9%, 17 of 437 cases).

We assessed whether certain genomic alterations were associated with reported re-

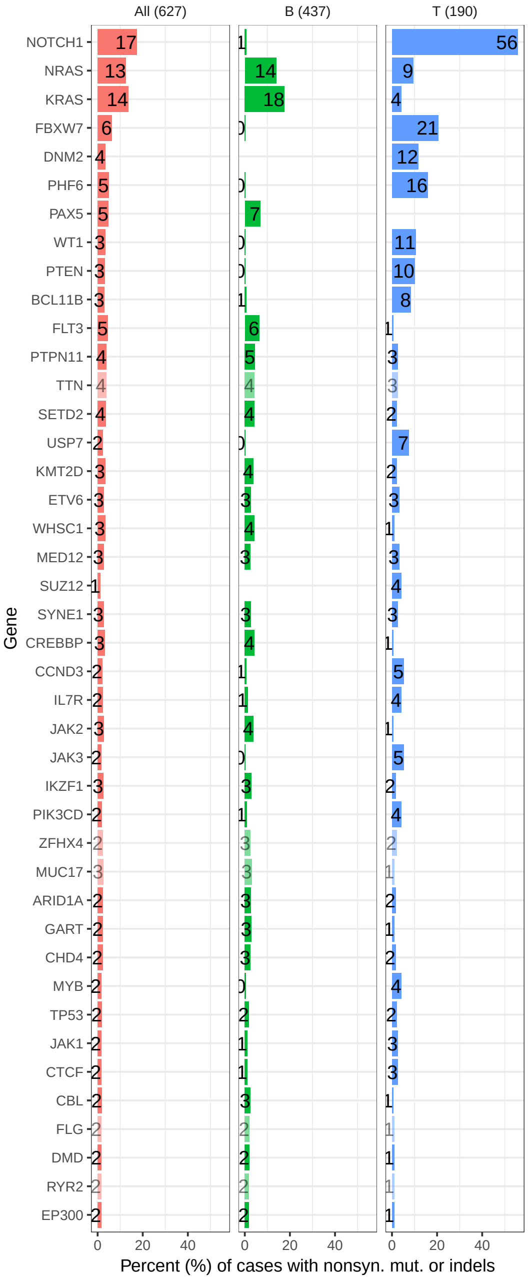


Figure 3.1: **Genes with recurrent somatic nonsynonymous SNVs or indels in pediatric ALL.** From left to right, percentage of patients with somatic mutations in a gene across all 627 patients (red), 437 patients with B-ALL (green), and 190 patients with T-ALL (blue). Panel headings indicate the patient subgroup, with the number of patients in parentheses. Genes names on the y-axis. Numbers show the percentage rounded to the nearest percent; “0” indicates percentages less than 1. Genes previously defined as drivers of pediatric ALL or relapse-associated genes are plotted with full opacity.

lapse across the entire cohort of 627 patients with pediatric ALL. For this analysis, we used only SNVs causing nonsynonymous or splice variants (SnpEff effect impact of "HIGH" or "MODERATE") because SNV calls are higher confidence compared to indel calls. We found that somatic *WT1* SNVs associated with relapse across all 627 patients ($p < 2.6e-10$) and in the 190 T-ALL cases ($p < 6e-5$, Fisher's exact uncorrected). One study of adult patients with pre-treatment T-ALL found significantly decreased relapse-free survival in *WT1*-mutant patients in the thymic T-ALL subgroup, although *WT1*-mutant status had no significant difference in relapse-free survival in the full adult T-ALL cohort [145].

Order of sequential mutations in relapsed ALL

We assessed the order of diagnosis and relapse mutations using an integrated sequential network (ISN) [146]. Mutation data from 88 pediatric cases with relapsed ALL [119, 122, 123] were used to generate ISNs. Using all relapsed cases, we found mutations in *PAX5*, *PHF6*, *DNM2*, and *KRAS* were significantly early events (Figure 3.2). While *KRAS* was an early event in our analysis, evidence for positive and negative clonal selection of *RAS* mutations has been observed in early sequencing studies of lymphoid malignancies[147] and recent next-generation sequencing studies of relapsed ALL [119, 123, 148]. Significantly late events occurred in *NT5C2* and *CREBBP* (Figure 3.2). NT5C2 is mutated specifically at relapse, with gain-of-function mutations causing resistance to 6-mercaptopurine [60, 134]. Metabolic analysis of *NT5C2*-mutant leukemia cells found decreased intracellular purine substrates and increased products of *NT5C2* activity, creating a metabolic imbalance that may be selected against in early ALL progression [135].

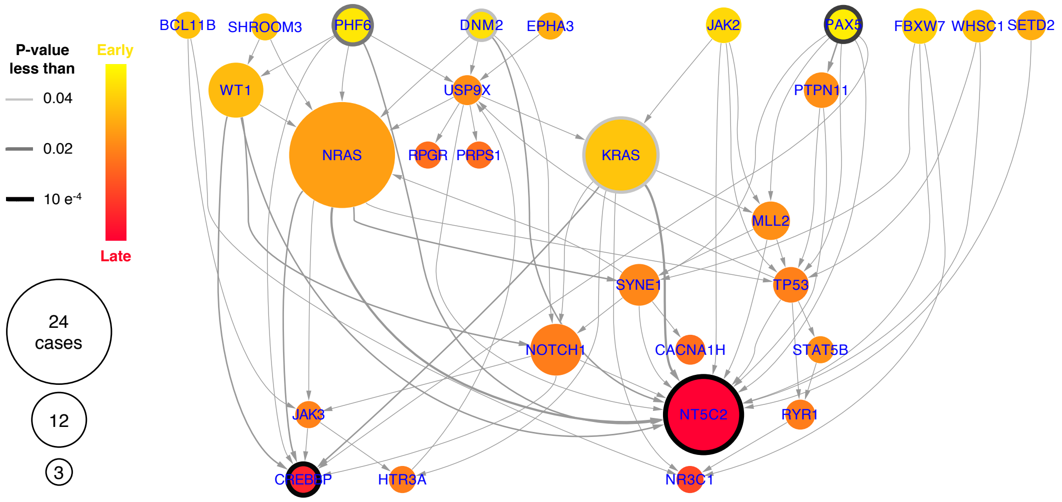


Figure 3.2: **ISN of sequential mutation order in 88 cases of relapsed ALL.** ISN illustrating the sequential order of diagnosis and relapse mutations in relapsed ALL by pooling evolutionary paths across patients. Each node represents a gene and each arrow points from a gene with an early event to a gene with a late event. To test whether a gene within the ISN was significantly early or late, we used a one-sided binomial test based on the in-degree and out-degree of each node.

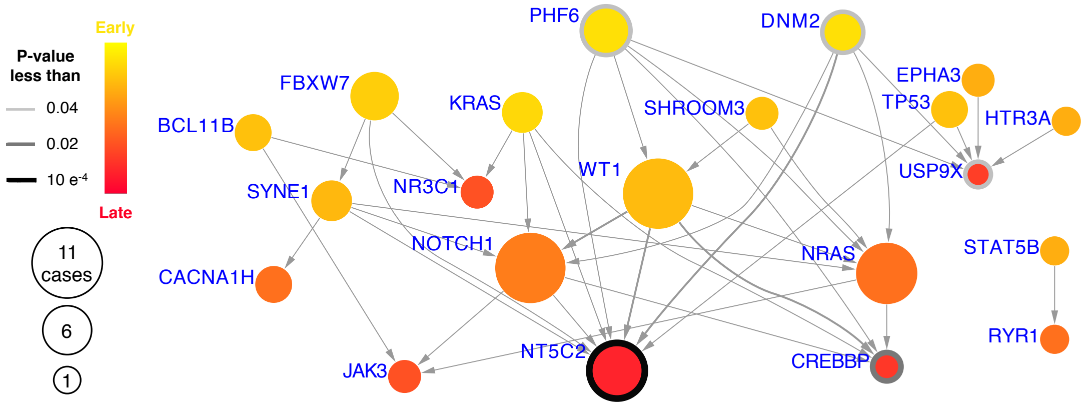


Figure 3.3: **ISN of sequential mutation order in 37 cases of relapsed T-lineage ALL.** Each node represents a gene and each arrow points from a gene with an early event to a gene with a late event. To test whether a gene was significantly early or late, we used a one-sided binomial test on each node's in-degree and out-degree.

ISN analysis of only T-ALL samples (37 patients) again found that *PHF6* and *DNM2* alterations were early events, and that *NT5C2* and *CREBBP* mutations were significantly late (Figure 3.3). Additionally, *USP9X* mutations were significantly late events.

Nomination of somatic-like germline variants in ALL using TOBI

To nominate potential oncogenic germline variants in ALL, we applied our TOBI framework to pediatric ALL samples. In addition to the 539 cases from TARGET, we included 19 cases with relapsed B-ALL from ref. [123], 55 cases with relapsed B-precursor or T-ALL from [119], and 8 cases with B-ALL from ref. [149], for a total of 621 cases (186 cases with T-ALL, 435 cases with B-ALL). All 621 cases had matched tumor and normal DNA available. We used the somatic calls from SAVI as the "true somatic" variants in the 539 cases from TARGET (SAVI calls included in Figure 3.1), and the previously published diagnosis variants as "true somatic" calls for samples from [123], [119], and [149]. TOBI was run separately for T-ALL cases and B-ALL cases given the different diagnosis mutation distributions in B-ALL and T-ALL (Figure 3.1).

Our prior assessment of seven non-ALL cancer types found that TOBI's sensitivity in a cancer type positively correlated with the median mutation rate in that cancer.Since ALL has a lower mutation rate than most adult cancers [91, 125], we assessed whether increasing the training set size from 20 cases would improve performance in B-ALL. As the number of cases in the training set increased from 20 to 200 cases, the model F-score also increased (Figure 3.4). Accordingly, we used 200 training set cases to generate the B-ALL model and 100 training set cases to generate the T-ALL model.

After comparing TOBI's somatic classifications to true somatic calls in ALL, we found that across all variants TOBI had a sensitivity of 28.2% in T-ALL and 42.6% in B-ALL. For nonsynonymous variants, TOBI had a sensitivity of 34.3% in T-ALL and 46% in B-ALL (Figure 3.5a). These sensitivities are similar to those observed in pediatric glioma, and

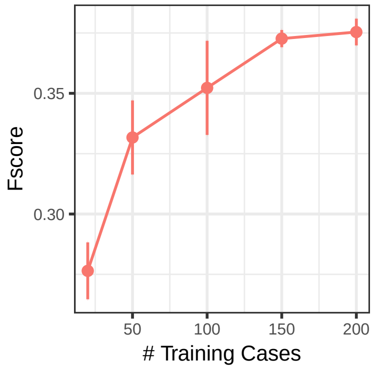


Figure 3.4: **TOBI training set size in 435 pediatric cases with B-ALL**. The average F-score for increasing numbers of cases in the training set; the number of samples in the training set equals the number in the testing set. Points represent average predictions from five runs with randomly selected training and testing sets cases; error bars represent +/- s.e.m.

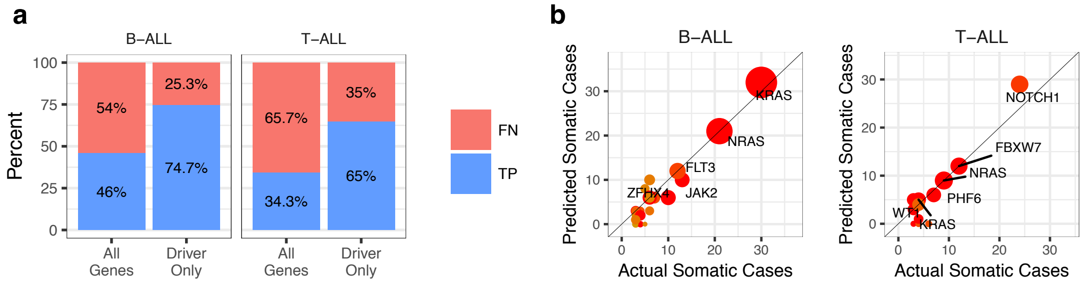


Figure 3.5: **True positive rate and actual versus predicted cases from TOBI analysis of pediatric ALL**. (a) Percentages of true positive (TP) or false negative (FN) TOBI somatic predictions in nonsynonymous variants across all genes or only driver genes. (b) Comparison of actual vs. predicted cases with somatic, nonsynonymous variants. Dot color corresponds to the fraction of synonymous variants out of all variants remaining after TOBI filtering; red indicates a lower fraction of synonymous variants (same key as Figure 1.4). Dot size reflects the number of predicted cases over the protein length in amino acids, with larger dots indicating a larger ratio. For clarity, genes with less than three previously published somatic variants are not shown.

likely reflect the positive correlation between TOBI sensitivity and the mutation rate of a cancer type. Focusing on per gene predictions, the number of cases TOBI predicted with somatic nonsynonymous variants was very similar to the published number of cases with somatic alterations (Figure 3.5b). Top predicted genes were known drivers of T-ALL and B-ALL.

Rare germline variants in cancer-susceptibility genes

An alternative strategy for investigating the potential role of germline variants in ALL development involved assessing variant pathogenicity in genes associated with leukemia predisposition or cancer predisposition. Leukemia predisposition genes were *PAX5*, *CEBPA*, *ETV6*, *RUNX1*, *NSD1*, *NF1*, and *TP53*. For cancer predisposition across multiple solid and lymphoid tumors, we used the list of genes from [34]. The union of these lists was a 153-gene list used for further analysis.

Interpreting the potential clinical consequences of germline variants remains a major challenge. The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology recently provided guidelines for interpreting sequence variants in genes with known associations to inherited disorders [37]. These guidelines classify variants as “benign”, “likely benign”, “uncertain significance”, “likely pathogenic”, or “pathogenic” using integrated knowledge of disease biology, inheritance, population genetics, and *in silico* effect predictions. A cancer-specific implementation of the ACMG guidelines called the Characterization of Germline Variants (CharGer) pipeline was recently developed to study rare germline variants in 33 adult cancers [34]. We used CharGer to nominate potential pathogenic variants in pediatric ALL using our 153-gene list.

In a set of 74 relapsed ALL cases[119, 123] with germline variants called using SAVI, CharGer nominated one pathogenic nonsynonymous SNV: *ATM*:p.Q675*, described in ClinVar (ID 231933) as “pathogenic/likely pathogenic” for ataxia-telangiectasia, an autosomal recessive disorder that increases cancer predisposition [150]. Patients with ataxia-

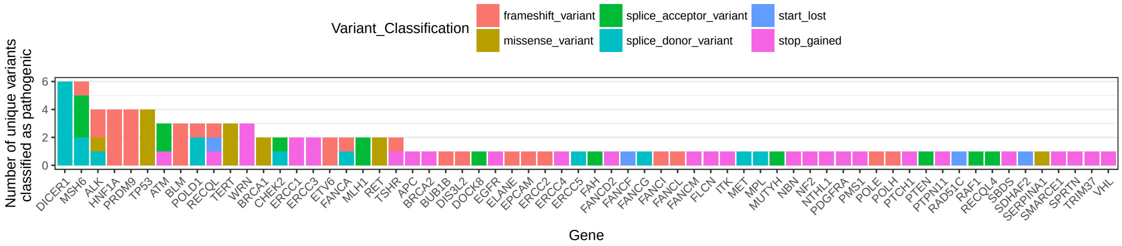


Figure 3.6: **Functional effects of predicted pathogenic variants in 539 ALL cases.** Number of unique variants (chromosome, position, reference nucleotide > alternate nucleotide) predicted as "pathogenic'' in each gene.

telangiectasia have an increased risk of developing lymphoid malignancies including ALL[151]. Recent studies have found that the initial presenting symptom of ataxia-telangiectasia in children and adolescents can be ALL [150]. CharGer also nominated three "likely pathogenic" variants: *TERT*:p.H412Y, *TRIM37*:p.Q104*, and *MSH2*:p.Y66*. The *TERT*:p.H412Y variant has been associated with dyskeratosis congenita[152, 153], a disorder that increases the risk of myeloid leukemia. The *MSH2*:p.Y66* variant (ClinVar ID 182608) and the *TRIM37*:p.Q104* variant (ClinVar ID 195377) are described as "benign/-Likely benign'' in ClinVar. The patient in our cohort with an *MSH2* variant did not display a hypermutation phenotype. CharGer also nominated multiple indels as pathogenic; however, SAVI's variant calling strategy was developed primarily using SNVs, so these indels are lower confidence.

Since the GATK variant caller included local realignment for indels, we used GATK to call germline variants in cancer predisposition genes in 539 TARGET samples. Pathogenic variants were called in 64 genes, with *DICER* and *MSH6* having the most unique "pathogenic'' variants (Figure 3.6). While CharGer and ACMG nominations of "pathogenic'' and "likely pathogenic'' narrowed the list of candidate germline variants, variant annotations and CharGer parameters may influence nominations. For example, an *ATM* splice acceptor variant at chr11:108114679 was reported as “likely pathogenic’’ in our analysis, but was reported as “pathogenic’’ in a study of rare germline variants in 33 adult cancers [34].

3.3 Discussion

In this chapter, we integrate germline and somatic analysis of a large cohort of patients with pediatric ALL. Our cohort included patients with B-precursor lineage and T-lineage ALL, as well as patients with and without reported ALL relapse. We found recurrent somatic mutations in known driver genes of ALL, including *NOTCH1*, *KRAS*, *NRAS*, and *FBXW7*. In comparing the diagnosis mutations found in patients with and without relapse, we found a significant association between having *WT1* somatic SNVs and relapse status. This association was found in the whole 627-patient cohort and within the 190-patient T-ALL cohort. Across the 190 cases with T-ALL, 13 cases (6.8%) have *WT1* SNVs and 20 cases (10.5%) have either SNVs or indels. These percentages are similar to those of previous pediatric T-ALL studies, which reported *WT1* alterations (SNV, indel, splice site alterations, and copy number variants) in 9.1%[120] and 13.2% of patients [154]. However, when our T-ALL cases are stratified by relapse status, we find *WT1* SNVs in 3 of 143 (2%) of cases with no reported relapse and 10 of 47 (21.2%) of relapsed cases. This high percentage of *WT1*-mutant relapsed cases is striking, particularly since we only included nonsynonymous SNVs in that analysis. Enrichment for *WT1* mutations at relapse has been suggested [155], and *WT1* mutations are associated with decreased relapse-free survival in adult patients with thymic T-ALL[145].

We also analyzed the order of somatic mutations across 88 cases with relapsed B-lineage or T-lineage ALL using integrated sequential network[146]. Across all 88 patients, alterations in *PAX5*, *PHF6*, *DNM2*, and *KRAS* were significantly early events, while *NT5C2* and *CREBBP* alterations were late events (Figure 3.2). PAX5 normally maintains mature B cell identity and function, and experimental findings that PAX5 deletion in mature B cells leads to dedifferentiation into pro-B cells and lymphoma development suggests that early PAX5 alterations generate a pro-B cell state that promote B-cell malignancy [21]. Late occurrence of *NT5C2* variants in ALL evolution may reflect selection against *NT5C2* variants early in leukemia development because *NT5C2* variants cause purine metabolism imbalances [135], or selection for increased purine metabolism by relapsed leukemia cells when patients are treated with the purine analogue 6-MP [60, 134]. Mutation ordering was relatively similar in T-ALL only-analysis, with the addition of *USP9X* as a late event. ISNs provide comprehensive descriptions of tumor evolution across multiple patients, summarizing and complementing analysis of clonal dynamics in individual ALL patients. The late alterations in *NT5C2* and *CREBBP* were recognized in clonal analysis of individual patients [119, 123]. The presence of significantly early and late variants across an ALL cohort may stem from the high percentage of ALL cases with branched patterns of evolution [119], where relapsed samples contain only a subset of the genetic alterations seen in the major diagnosis clone.

Diagnosis and relapse variants occur atop each patient's unique germline genetic background, so we employed multiple strategies to assess germline variants that could influence ALL. First, we applied TOBI to our pediatric ALL cohort to assess somatic-like germline variants (SLG). We previously saw that since TOBI generates models that successfully identify true somatic variants from tumor-only samples, germline variants that are classified as somatic are enriched for genes causing autosomal dominant cancer predisposition syndromes. When we assessed TOBI's ability to identify true somatic variants in our pediatric ALL cohort, we found that TOBI had a sensitivity for nonsynonymous variants of 34.3% in T-ALL and 46% in B-ALL. Although TOBI does identify true somatic variants in driver genes (Figure 3.5), this relatively low sensitivity for somatic variant classification may suggest that the TOBI model did not fully capture biological features of ALL somatic variants. Thus, SLG variants nominated by TOBI's ALL model may only partially reflect ALL biology; these SLG variants need further critical analysis and ultimately experimental validation.

Our second strategy to investigate candidate leukemia-associated germline variants was interrogation of high-quality, coding germline variants in a set of curated genes related to leukemia or cancer predisposition. From 74 relapsed ALL cases, 2 nonsynonymous variants in *TERT* and *ATM* were nominated as "pathogenic" or "likely pathogenic". In a set of 539 ALL cases, 64 of 153 cancer predisposition genes were nominated to have a pathogenic germline variant. Interpreting germline variant pathogenicity remains challenging even with ACMG guidelines, particularly given potential annotation discrepancies.

We recognize several limitations of the analysis in this chapter. Our somatic analysis focused on SNVs and indels, but translocations, copy number variants, and splice variants are common somatic alterations in ALL. Additionally, our indel calls are only from one variant caller, but calling indels with multiple variant callers and intersecting the calls may have lead to more high quality indels. TOBI captured high-quality variants in driver genes that passed quality filters. However, certain putative somatic variants called in tumor-normal analysis were removed based on TOBI filters, including many variants in *PTEN*. Future research could identify optimal methods of setting TOBI quality filters depending on the cohort being analyzed. Classifying germline variants as benign or pathogenic is complicated by differences in variant annotation software and reference transcripts. While we nominated candidate pathogenic germline variants in pediatric ALL using CharGer and ACMG criteria, validation experiments or observed inheritance of candidate variants in an affected pedigree would strengthen pathogenicity claims.

In summary, this chapter described somatic and germline variants in over 600 patients with pediatric ALL. We find that *WT1* mutations are significantly enriched in patients who have relapsed. The TOBI framework identified true somatic variants at modest sensitivity. Future assessment of SLG variants from TOBI will generate candidate germline alterations for further study. ACMG guidelines nominated germline variants in *TERT* and *ATM* as pathogenic variants. Future work will involve curating high quality germline variant sets, and assessing their associations with relapse status and the order of somatic mutations in ALL.

3.4 Methods

ISN of relapsed ALL

We illustrated the sequential order of somatic mutations in relapsed ALL using the ISN25 that pools evolutionary paths across all patients. We selected recurrently mutated genes that were previously defined as drivers of paediatric ALL[156–158] and relapse-genes[119, 122]. Only non-synonymous single nucleotide variants were used in analysis. For each patient, we generated a sequential network that defined early events as mutations observed in both the primary tumour and the relapsed tumour, whereas late events were mutations only observed in the relapsed tumour. Each node represented a gene, and each arrow pointed from a gene with an early event to a gene with a late event. The ISN then pooled sequential networks across all patients. To test whether a gene within the ISN was significantly early or late, we used the binomial test based on the in-degree and out-degree of each node. Somatic mutation data used to generate ISN were aggregated from previously published studies [119, 122, 123]. Figure 3.2 included all 88 published patients, while figure 3.3 included only 37 patients with T-ALL.

Genomic sequence access and retrieval of clinical data

We obtained approval from the database of Genotypes and Phenotypes (dbGaP) to access exome sequences from TARGET (accession number phs000218.v16.p6). Metadata from dbGaP described 539 patients with ALL and whole exome sequencing of matched tumor and normal DNA. These tumor and normal WES alignment files were downloaded using SRAToolkit.

Clinical data was downloaded from the TARGET data matrix `https://ocg.cancer.gov/programs/target/data-matrix`.

Somatic variant calling

Somatic variants and indels were called using the SAVI (Statistical Algorithm for Variant Identification) algorithm [44] on matched tumor and normal samples, with annotations from SnpEff version 4.1c [80]. SAVI is an empirical Bayesian framework that constructs empirical priors for variant allele frequency (VAF) in each sample. Specifically, high quality candidate variants are obtained by removing positions with only reference reads, low sequence depth, strand bias, or no high quality reads. For each variant position, SAVI uses the number of remaining reference and variant supporting reads to build the prior and posterior distributions of VAF. Variants were considered present with a VAF greater than 3% and a posterior probability of < $1e - 6$ for variant presence. To call somatic variants, SAVI calculates credibility intervals for VAF difference between one sample (e.g. tumor) and another sample (e.g. normal), with significant differences defined by high-credibility intervals at posterior probability less than $1e10 - 5$. These candidate somatic variants were further filtered to retain variants with normal VAF < 3%, tumor VAF >= 3%, strand bias p-value > 0.01 by Fisher's exact test, and no annotation as "COMMON'' in dbSNP build 144.

We also excluded variants found in our in-house database of common mutations in 219 normal WES cases ("Meganormal" database), variants in Ig and T-Cell variable genes annotated from SnpEff's transcript_biotype, and variants found recurrently in normal

DNA from the current ALL cohort (recurrence defined as present in normal DNA of >1% of patients).

Mutations in ALL driver genes [135] or in genes recurrently mutated in pediatric cancer [125] that were called by SAVI and subsequently filtered out were rescued for the final somatic variant list. Additional variants in these driver genes were retained in the somatic variant list if they met the following criteria: < 3 variant reads in normal sample, forward and reverse variant-supporting reads in tumor sample, and (alternative allele depth >= 5 OR tumor VAF >=15%). We excluded potential oxoG artifacts from this rescue set by removing C>A or G>T mutations with tumor VAF < 20% that were not observed in other samples after all initialy quality filters.

To standardize annotations between this cohort and previously published diagnosis and relapse variants from [119, 122, 123], published variants were annotated using SnpEff version 4.3t.

TOBI framework on ALL cases

Diagnosis (tumor) WES data from 621 patients with pediatric ALL were analyzed through the TOBI.bam pathway indicated in Figure To briefly review, variants were called from diagnosis samples using Samtools and Bcftools[79], excluding variants with mapping quality < 10. Variants were annotated with dbSNP build 144, Cosmic v74, and dbNSFP v2.4 databases[81], and our lab's "Meganormal" database, using SnpEff[80]. Technical and biological variant filters were applied to remove variants with VAF < 1%, mq < 40, p-values < 0.01 for any of bias metrics (strand bias, map quality bias, tail distance bias), common SNPs with population allele frequency > 1% in the 1000 Genome Project populations, Meganormal variants, SNVs present in only dbSNP and not in COSMIC, and SNVs in intragenic, non-coding exon, or splice-site regions. The driver genes in Section 3.4 were also used as driver genes in TOBI assessments.

Germline variant calling

For all the 539 TARGET normal exomes, germline variants were called using The Genome Analysis Toolkit 4 (GATK4) [159]. We generated gVCFs using HaplotypeCaller from GATK4 (v4.beta.5) across chromosomes 1-22, X, and Y. Joint genotyping was performed across all gVCFs using GATK4, v4.0.0.0 commands. First, gVCFs for all samples were merged using GenomicsDBImport on each chromosome interval. Next, all samples underwent joint genotyping using GenotypeGVCFs. We retained only those variants in the exome calling intervals defined in the Exome Aggregation Consortium (ExAC)[160] (`ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/exome_calling_regions.v1.interval_list`).

Germline calls had a germline VAF > 20%, a genotype that was not set to ”missing” or homozygous reference, and alternate depth >=5 reads.

SAVI was used to call germline variant calls for 74 relapsed ALL cases from [119] and [123]. Germline calls had a germline VAF > 30%.

Raw germline variant calls were filtered to remove variants with a population frequency >1% in the 1000 Genomes project or ExAC-non-TCGA subset. We used the set of 152 genes curated by [34] for cancer susceptibility (described in [34]'s supplementary table 1). Additionally, we included the gene *NSD1* since *NSD1* germline mutations cause Sotos syndrome with inherited pediatric ALL risk[142], for a total list of 153 genes. To apply ACMG pathogenicity guidelines, we used the CharGer pipeline[34] v0.5.2 and VEP[161] v.92.3 as installed with conda. First, we generated a sites-only vcf of germline variants in the 153 genes using GATK4’s MakeSitesOnlyVcf, and removed the “AF’’ INFO field (vcf-annotate -r INFO/AF). Next, we annotated using VEP: /Software/perl/src/perl-5.22.2/perl vep –everything –offline –cache –dir /home/.vep/ –assembly GRCh37 –format vcf –vcf -i input.vcf -o vep.input.vcf –force_overwrite –fasta /refs/GRCh37.71.chr.fa –fork 4 –buffer_size 2000 –merged –use_given_ref. We ran CharGer on the VEP-annotated vcf: charger –include-vcf-details -f input.vcf -o out.txt -O -D –inheritanceGeneList demo/inheritanceGeneList.txt -z pathogenic_var.vcf.gz -H demo/somaticHotspots.hotspot3d.clusters -l -x –PP2GeneList demo/inheritanceGeneList.txt. Our curated list of gene-specific pathogenic variants (pathogenic_var.vcf.gz) was generated according to the methods of [34]; specifically, we included all variants from the ASU database for *TERT* mutations (`http://telomerase.asu.edu/diseases.html`) and the ARUP MEN2 database for *RET* (`http://www.arup.utah.edu/database/MEN2/MEN2_display.php`), only variants carried by an affected proband and confirmed as germline variant in the IARC database for *TP53*[162], and variants marked as clinically important in the BIC database for *BRCA1* and *BRCA2* (`https://research.nhgri.nih.gov/bic/`).

3.5 Acknowledgements

ISN analysis of 88 pediatric cases with relapsed ALL is a reproduction, in part, of a publication in Nature by Tzoneva, G., Dieck, C.L., Oshima, K., Ambesi-Impiombato1, A., Sánchez-Martín, M., Madubata, C.J., Khiabanian, H., Yu, J., Waanders, E., Iacobucci, I., Sulis, M.L., Kato, M., Koh, K., Paganin, M., Basso, G., Gastier-Foster, J.M., Loh, M.L., Kirschner-Schwabe, R., Mullighan, C.G., Rabadan, R., and Ferrando, A.A. [135]. The results published here are in part based upon data generated by the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative managed by the NCI. The data used for this analysis are available under dbGAP accession number phs000218 for TARGET, and specifically phs000464 for TARGET ALL. Information about TARGET can be found at http://ocg.cancer.gov/programs/target. We recognize the contribution of all specimen donors. We thank Stephanie Siegmund for her editorial suggestions.

Conclusion

Motivation

In this work, our goals were to (1) identify germline variants that contribute to cancer development, and (2) integrate germline, somatic, and relapse genomic information in a systematic manner to elucidate biological aspects of cancer, particularly pediatric ALL. Certain somatic variants and germline cancer-associated variants share biological features, such as mutating a particular amino acid. We hypothesized that these shared biological features would allow us to identify potential cancer-associated germline variants.

Summary

In Chapter 1, we developed a framework, TOBI, that uses machine learning to identify somatic variants from tumor-only data or identify somatic-like germline variants. Assessing true somatic variants, we found TOBI has a high true positive rate across all somatic variants. TOBI identifid driver genes in different tumor types, and outperformed other methods of tumor-only analysis. Given TOBI's performance on true somatic variants, we hypothesized that germline variants identified by TOBI would be enriched for cancer-predisposition pathways.

In Chapter 2, we used TOBI for integrated somatic and germline analysis. We assessed which variants classified as somatic by TOBI are germline variants with somatic features, or "somatic-like'' germline variants. Somatic-like germline variants were enriched in genes associated with autosomal dominant cancer-predisposition syndromes and included known *TP53* germline variants. Using TOBI, we found that 5% of patients with bladder carcinoma had germline inactivating mutations in the Fanconi anemia pathway. Comparing the 11% of bladder carcinoma patients with somatic or germline Fanconi anemia mutations to the remaining bladder carcinoma patients, we found mutant cases were enriched for a *BRCA*-deficiency somatic signature.

In Chapter 3, we report integrated analysis of germline, somatic, and relapse variants in a cohort of patients with pediatric acute lymphoblastic leukemia. Analyzing a cohort of over 600 mixed B- and T-ALL cases, we capture known driver genes of ALL. We also report an association between *WT1* mutations and relapse. By analyzing the order of somatic mutations in relapsed ALL, we confirmed prior reports that *NT5C2* and *CREBBP* are late events in leukemia evolution, and found that *PAX5*, *PHF6*, *DNM2*, and *KRAS* are early events.

Next, we began assessing potential germline variants in ALL. We applied TOBI to ALL, and found that while driver genes are captured, TOBI had low sensitivity in pediatric ALL, decreasing our confidence in SLG variants. We also nominated pathogenic variants in a set of cancer-predisposition genes using ACMG guidelines, identifying nonsynonymous variants in *TERT* and *ATM* as potential pathogenic germline variants.

Limitations

While many of our results were promising, we recognize several limitations in this work. The choice of variant annotations can affect both our TOBI pipeline and the ACMG germline classification pipeline we used in Chapter 3. TOBI performance is very correlated to a cancer cohort's median somatic mutation rate, making TOBI more sensitive in high mutation rate cancers and less sensitive to most pediatric cancers. Heterogeneity in patient ancestry or sample sequencing protocol also affected TOBI performance. We also note that the somatic-like germline variants TOBI nominates are not necessarily oncogenic. Further assessment and experimental validation is required to fully understand the role of these variants. Our germline and somatic variant analysis included indels, which may be called unreliably without indel realignment.

Future directions

SLG variants from TOBI require further study for assessing pathogenicity. Additionally, given the depth of knowledge on aberrant diagnosis and relapse pathways in ALL, future germline analysis can focus on genes related to those pathways.

Bibliography

1. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70. ISSN: 0092-8674, 1097-4172 (Jan. 7, 2000).

2. Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A. & Hemminki, K. Environmental and Heritable Factors in the Causation of Cancer—Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. *New England journal of medicine* **343**, 78–85 (2000).

3. Bodmer, W. & Tomlinson, I. Rare Genetic Variants and the Risk of Cancer. *Current Opinion in Genetics & Development. Molecular and genetic bases of disease* **20**, 262–267. ISSN: 0959-437X (June 1, 2010).

4. Krawczyk, M., Müllenbach, R., Weber, S. N., Zimmer, V. & Lammert, F. Genome-Wide Association Studies and Genetic Risk Assessment of Liver Diseases. *Nature Reviews Gastroenterology & Hepatology* **7**, 669–681. ISSN: 1759-5053 (Dec. 2010).

5. Consortium, T. W. T. C. C. Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls. *Nature* **447**, 661–678. ISSN: 1476-4687 (June 2007).

6. Moriyama, T., Relling, M. V. & Yang, J. J. Inherited Genetic Variation in Childhood Acute Lymphoblastic Leukemia. *Blood* **125**, 3988–3995. ISSN: 0006-4971, 1528-0020 (June 25, 2015).

7. Yang, J. J., Landier, W., Yang, W., Liu, C., Hageman, L., Cheng, C., Pei, D., Chen, Y., Crews, K. R., Kornegay, N., Wong, F. L., Evans, W. E., Pui, C.-H., Bhatia, S. & Relling, M. V. Inherited NUDT15 Variant Is a Genetic Determinant of Mercaptopurine Intolerance in Children With Acute Lymphoblastic Leukemia. *Journal of Clinical Oncology* **33**, 1235–1242. ISSN: 0732-183X (Apr. 10, 2015).

8. Warthin, A. S. HEREDITY WITH REFERENCE TO CARCINOMA: AS SHOWN BY THE STUDY OF THE CASES EXAMINED IN THE PATHOLOGICAL LABORATORY OF THE UNIVERSITY OF MICHIGAN, 1895-1913. *Archives of Internal Medicine* **XII**, 546–555. ISSN: 0730-188X (Nov. 1, 1913).

9. Li, F. P. & Fraumeni, J. F. **Soft-Tissue Sarcomas, Breast Cancer, and Other Neoplasms: A Familial Syndrome?** *Annals of Internal Medicine* **71**, 747–752. ISSN: 0003-4819 (Oct. 1, 1969).

10. Li, F. P. & Fraumeni, J. F. Rhabdomyosarcoma in Children: Epidemiologic Study and Identification of a Familial Caneer Syndrome. *JNCI: Journal of the National Cancer Institute* **43**, 1365–1373. ISSN: 0027-8874 (Dec. 1, 1969).

11. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America* **68**, 820–823. ISSN: 0027-8424 (Apr. 1971).

12. Friend, S. H., Bernards, R., Rogelj, S., Weinberg, R. A., Rapaport, J. M., Albert, D. M. & Dryja, T. P. A Human DNA Segment with Properties of the Gene That Predisposes to Retinoblastoma and Osteosarcoma. *Nature* **323**, 643–646. ISSN: 0028-0836 (Oct. 16, 1986).

13. Foulkes, W. D. Inherited Susceptibility to Common Cancers. *New England Journal of Medicine* **359**, 2143–2153 (2008).

14. Zhuang, Z., Park, W.-S., Pack, S., Schmidt, L., Vortmeyer, A. O., Pak, E., Pham, T., Weil, R. J., Candidus, S., Lubensky, I. A., Linehan, W. M., Zbar, B. & Weirich, G. Trisomy 7-Harbouring Non-Random Duplication of the Mutant MET Allele in Hereditary Papillary Renal Carcinomas. *Nature Genetics* **20**, 66–69. ISSN: 1061-4036 (Sept. 1998).

15. Huang, S. C., Koch, C. A., Vortmeyer, A. O., Pack, S. D., Lichtenauer, U. D., Mannan, P., Lubensky, I. A., Chrousos, G. P., Gagel, R. F., Pacak, K. & Zhuang, Z. Duplication of the Mutant RET Allele in Trisomy 10 or Loss of the Wild-Type Allele in Multiple Endocrine Neoplasia Type 2-Associated Pheochromocytomas. *Cancer Research* **60**, 6223–6226. ISSN: 0008-5472, 1538-7445 (Nov. 15, 2000).

16. Maris, J. M. & Knudson, A. G. Revisiting Tissue Specificity of Germline Cancer Predisposing Mutations. *Nature Reviews Cancer* **15**, 65–66. ISSN: 1474-175X (Feb. 2015).

17. Wallace, M. R., Marchuk, D. A., Andersen, L. B., Letcher, R., Odeh, H. M., Saulino, A. M., Fountain, J. W., Brereton, A., Nicholson, J., Mitchell, A. L. & Al, E. Type 1 Neurofibromatosis Gene: Identification of a Large Transcript Disrupted in Three NF1 Patients. *Science* **249**, 181–186. ISSN: 0036-8075, 1095-9203 (July 13, 1990).

18. Malkin, D., Li, F. P., Strong, L. C., Fraumeni, J. F., Nelson, C. E., Kim, D. H., Kassel, J., Gryka, M. A., Bischoff, F. Z., Tainsky, M. A. & others. Germ Line P53 Mutations in a Familial Syndrome of Breast Cancer, Sarcomas, and Other Neoplasms. *Science* **250**, 1233–1238 (1990).

19. Mossé, Y. P., Laudenslager, M., Longo, L., Cole, K. A., Wood, A., Attiyeh, E. F., Laquaglia, M. J., Sennett, R., Lynch, J. E., Perri, P., Laureys, G., Speleman, F., Kim, C., Hou, C., Hakonarson, H., Torkamani, A., Schork, N. J., Brodeur, G. M., Tonini, G. P., Rappaport, E., Devoto, M. & Maris, J. M. Identification of ALK as a Major Familial Neuroblastoma Predisposition Gene. *Nature* **455,** 930–935. ISSN: 0028-0836 (Oct. 16, 2008).

20. Rahman, N. Realizing the Promise of Cancer Predisposition Genes. *Nature* **505,** 302–308. ISSN: 0028-0836 (Jan. 16, 2014).

21. Shah, S., Schrader, K. A., Waanders, E., Timms, A. E., Vijai, J., Miething, C., Wechsler, J., Yang, J., Hayes, J., Klein, R. J., Zhang, J., Wei, L., Wu, G., Rusch, M., Nagahawatte, P., Ma, J., Chen, S.-C., Song, G., Cheng, J., Meyers, P., Bhojwani, D., Jhanwar, S., Maslak, P., Fleisher, M., Littman, J., Offit, L., Rau-Murthy, R., Fleischut, M. H., Corines, M., Murali, R., Gao, X., Manschreck, C., Kitzing, T., Murty, V. V., Raimondi, S. C., Kuiper, R. P., Simons, A., Schiffman, J. D., Onel, K., Plon, S. E., Wheeler, D. A., Ritter, D., Ziegler, D. S., Tucker, K., Sutton, R., Chenevix-Trench, G., Li, J., Huntsman, D. G., Hansford, S., Senz, J., Walsh, T., Lee, M., Hahn, C. N., Roberts, K. G., King, M.-C., Lo, S. M., Levine, R. L., Viale, A., Socci, N. D., Nathanson, K. L., Scott, H. S., Daly, M., Lipkin, S. M., Lowe, S. W., Downing, J. R., Altshuler, D., Sandlund, J. T., Horwitz, M. S., Mullighan, C. G. & Offit, K. A Recurrent Germline PAX5 Mutation Confers Susceptibility to Pre-B Cell Acute Lymphoblastic Leukemia. *Nature Genetics* **45,** 1226–1231. ISSN: 1061-4036 (Oct. 2013).

22. Zhang, M. Y., Churpek, J. E., Keel, S. B., Walsh, T., Lee, M. K., Loeb, K. R., Gulsuner, S., Pritchard, C. C., Sanchez-Bonilla, M., Delrow, J. J., Basom, R. S., Forouhar, M., Gyurkocza, B., Schwartz, B. S., Neistadt, B., Marquez, R., Mariani, C. J., Coats, S. A., Hofmann, I., Lindsley, R. C., Williams, D. A., Abkowitz, J. L., Horwitz, M. S., King, M.-C., Godley, L. A. & Shimamura, A. Germline ETV6 Mutations in Familial Thrombocytopenia and Hematologic Malignancy. *Nature Genetics* **47,** 180–185. ISSN: 1061-4036 (Feb. 2015).

23. Noetzli, L., Lo, R. W., Lee-Sherick, A. B., Callaghan, M., Noris, P., Savoia, A., Rajpurkar, M., Jones, K., Gowan, K., Balduini, C. L., Pecci, A., Gnan, C., De Rocco, D., Doubek, M., Li, L., Lu, L., Leung, R., Landolt-Marticorena, C., Hunger, S., Heller, P., Gutierrez-Hartmann, A., Xiayuan, L., Pluthero, F. G., Rowley, J. W., Weyrich, A. S., Kahr, W. H. A., Porter, C. C. & Di Paola, J. Germline Mutations in ETV6 Are Associated with Thrombocytopenia, Red Cell Macrocytosis and Predisposition to Lymphoblastic Leukemia. *Nature Genetics* **advance online publication.** ISSN: 1061-4036. doi:`10.1038/ng.3253`. <`http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.3253.html`> (visited on 04/10/2015) (Mar. 25, 2015).

24. Topka, S., Vijai, J., Walsh, M. F., Jacobs, L., Maria, A., Villano, D., Gaddam, P., Wu, G., McGee, R. B., Quinn, E., Inaba, H., Hartford, C., Pui, C.-h., Pappo, A., Edmonson, M., Zhang, M. Y., Stepensky, P., Steinherz, P., Schrader, K., Lincoln, A., Bussel, J., Lipkin,

S. M., Goldgur, Y., Harit, M., Stadler, Z. K., Mullighan, C., Weintraub, M., Shimamura, A., Zhang, J., Downing, J. R., Nichols, K. E. & Offit, K. Germline ETV6 Mutations Confer Susceptibility to Acute Lymphoblastic Leukemia and Thrombocytopenia. *PLoS Genet* **11**, e1005262 (June 23, 2015).

25. Moriyama, T., Metzger, M. L., Wu, G., Nishii, R., Qian, M., Devidas, M., Yang, W., Cheng, C., Cao, X., Quinn, E., Raimondi, S., Gastier-Foster, J. M., Raetz, E., Larsen, E., Martin, P. L., Bowman, W. P., Winick, N., Komada, Y., Wang, S., Edmonson, M., Xu, H., Mardis, E., Fulton, R., Pui, C.-H., Mullighan, C., Evans, W. E., Zhang, J., Hunger, S. P., Relling, M. V., Nichols, K. E., Loh, M. L. & Yang, J. J. Germline Genetic Variation in ETV6 and Risk of Childhood Acute Lymphoblastic Leukaemia: A Systematic Genetic Study. *The Lancet Oncology* **16**, 1659–1666. ISSN: 14702045 (Dec. 2015).

26. Churchman, M. L., Qian, M., te Kronnie, G., Zhang, R., Yang, W., Zhang, H., Lana, T., Tedrick, P., Baskin, R., Verbist, K., Peters, J. L., Devidas, M., Larsen, E., Moore, I. M., Gu, Z., Qu, C., Yoshihara, H., Porter, S. N., Pruett-Miller, S. M., Wu, G., Raetz, E., Martin, P. L., Bowman, W. P., Winick, N., Mardis, E., Fulton, R., Stanulla, M., Evans, W. E., Relling, M. V., Pui, C.-H., Hunger, S. P., Loh, M. L., Handgretinger, R., Nichols, K. E., Yang, J. J. & Mullighan, C. G. Germline Genetic IKZF1 Variation and Predisposition to Childhood Acute Lymphoblastic Leukemia. *Cancer Cell* **33**, 937–948.e8. ISSN: 1535-6108 (May 14, 2018).

27. Perez-Garcia, A., Ambesi-Impiombato, A., Hadler, M., Rigo, I., LeDuc, C. A., Kelly, K., Jalas, C., Paietta, E., Racevskis, J., Rowe, J. M., Tallman, M. S., Paganin, M., Basso, G., Tong, W., Chung, W. K. & Ferrando, A. A. Genetic Loss of SH2B3 in Acute Lymphoblastic Leukemia. *Blood* **122**, 2425–2432. ISSN: 0006-4971, 1528-0020 (Oct. 3, 2013).

28. Holmfeldt, L., Wei, L., Diaz-Flores, E., Walsh, M., Zhang, J., Ding, L., Payne-Turner, D., Churchman, M., Andersson, A., Chen, S.-C., McCastlain, K., Becksfort, J., Ma, J., Wu, G., Patel, S. N., Heatley, S. L., Phillips, L. A., Song, G., Easton, J., Parker, M., Chen, X., Rusch, M., Boggs, K., Vadodaria, B., Hedlund, E., Drenberg, C., Baker, S., Pei, D., Cheng, C., Huether, R., Lu, C., Fulton, R. S., Fulton, L. L., Tabib, Y., Dooling, D. J., Ochoa, K., Minden, M., Lewis, I. D., To, L. B., Marlton, P., Roberts, A. W., Raca, G., Stock, W., Neale, G., Drexler, H. G., Dickins, R. A., Ellison, D. W., Shurtleff, S. A., Pui, C.-H., Ribeiro, R. C., Devidas, M., Carroll, A. J., Heerema, N. A., Wood, B., Borowitz, M. J., Gastier-Foster, J. M., Raimondi, S. C., Mardis, E. R., Wilson, R. K., Downing, J. R., Hunger, S. P., Loh, M. L. & Mullighan, C. G. The Genomic Landscape of Hypodiploid Acute Lymphoblastic Leukemia. *Nature Genetics* **45**, 242–252. ISSN: 1061-4036 (Mar. 2013).

29. Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S. L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M., Lu, C., Chen, S.-C., Wei, L., Collins-Underwood, J. R., Ma, J., Roberts, K. G., Pounds, S. B., Ulyanov, A., Becksfort, J., Gupta, P., Huether,

R., Kriwacki, R. W., Parker, M., McGoldrick, D. J., Zhao, D., Alford, D., Espy, S., Bobba, K. C., Song, G., Pei, D., Cheng, C., Roberts, S., Barbato, M. I., Campana, D., Coustan-Smith, E., Shurtleff, S. A., Raimondi, S. C., Kleppe, M., Cools, J., Shimano, K. A., Hermiston, M. L., Doulatov, S., Eppert, K., Laurenti, E., Notta, F., Dick, J. E., Basso, G., Hunger, S. P., Loh, M. L., Devidas, M., Wood, B., Winter, S., Dunsmore, K. P., Fulton, R. S., Fulton, L. L., Hong, X., Harris, C. C., Dooling, D. J., Ochoa, K., Johnson, K. J., Obenauer, J. C., Evans, W. E., Pui, C.-H., Naeve, C. W., Ley, T. J., Mardis, E. R., Wilson, R. K., Downing, J. R. & Mullighan, C. G. The Genetic Basis of Early T-Cell Precursor Acute Lymphoblastic Leukaemia. *Nature* **481**, 157–163. ISSN: 0028-0836 (Jan. 12, 2012).

30. Kanchi, K. L., Johnson, K. J., Lu, C., McLellan, M. D., Leiserson, M. D. M., Wendl, M. C., Zhang, Q., Koboldt, D. C., Xie, M., Kandoth, C., McMichael, J. F., Wyczalkowski, M. A., Larson, D. E., Schmidt, H. K., Miller, C. A., Fulton, R. S., Spellman, P. T., Mardis, E. R., Druley, T. E., Graubert, T. A., Goodfellow, P. J., Raphael, B. J., Wilson, R. K. & Ding, L. Integrated Analysis of Germline and Somatic Variants in Ovarian Cancer. *Nature Communications* **5**. doi:10.1038/ncomms4156. <http://www.nature.com/ncomms/2014/140122/ncomms4156/full/ncomms4156.html> (visited on 05/20/2014) (Jan. 22, 2014).

31. Zhang, J., Walsh, M. F., Wu, G., Edmonson, M. N., Gruber, T. A., Easton, J., Hedges, D., Ma, X., Zhou, X., Yergeau, D. A., Wilkinson, M. R., Vadodaria, B., Chen, X., McGee, R. B., Hines-Dowell, S., Nuccio, R., Quinn, E., Shurtleff, S. A., Rusch, M., Patel, A., Becksfort, J. B., Wang, S., Weaver, M. S., Ding, L., Mardis, E. R., Wilson, R. K., Gajjar, A., Ellison, D. W., Pappo, A. S., Pui, C.-H., Nichols, K. E. & Downing, J. R. Germline Mutations in Predisposition Genes in Pediatric Cancer. *New England Journal of Medicine* **373**, 2336–2346. ISSN: 0028-4793 (Dec. 10, 2015).

32. Lu, C., Xie, M., Wendl, M. C., Wang, J., McLellan, M. D., Leiserson, M. D. M., Huang, K.-l., Wyczalkowski, M. A., Jayasinghe, R., Banerjee, T., Ning, J., Tripathi, P., Zhang, Q., Niu, B., Ye, K., Schmidt, H. K., Fulton, R. S., McMichael, J. F., Batra, P., Kandoth, C., Bharadwaj, M., Koboldt, D. C., Miller, C. A., Kanchi, K. L., Eldred, J. M., Larson, D. E., Welch, J. S., You, M., Ozenberger, B. A., Govindan, R., Walter, M. J., Ellis, M. J., Mardis, E. R., Graubert, T. A., Dipersio, J. F., Ley, T. J., Wilson, R. K., Goodfellow, P. J., Raphael, B. J., Chen, F., Johnson, K. J., Parvin, J. D. & Ding, L. Patterns and Functional Implications of Rare Germline Variants across 12 Cancer Types. *Nature Communications* **6**, 10086 (Dec. 22, 2015).

33. Southey, M. C. *et al.* PALB2, CHEK2 and ATM Rare Variants and Cancer Risk: Data from COGS. *Journal of Medical Genetics* **53**, 800–811. ISSN: 0022-2593, 1468-6244 (Dec. 1, 2016).

34. Huang, K.-l. *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**, 355–370.e14. ISSN: 0092-8674, 1097-4172 (Apr. 5, 2018).

35. Bodian, D. L., McCutcheon, J. N., Kothiyal, P., Huddleston, K. C., Iyer, R. K., Vockley, J. G. & Niederhuber, J. E. Germline Variation in Cancer-Susceptibility Genes in a Healthy, Ancestrally Diverse Cohort: Implications for Individual Genome Sequencing. *PLoS ONE* **9,** e94554 (Apr. 11, 2014).

36. Snape, K., Ruark, E., Tarpey, P., Renwick, A., Turnbull, C., Seal, S., Murray, A., Hanks, S., Douglas, J., Stratton, M. R. & Rahman, N. Predisposition Gene Identification in Common Cancers by Exome Sequencing: Insights from Familial Breast Cancer. *Breast Cancer Research and Treatment* **134,** 429–433. ISSN: 0167-6806, 1573-7217 (Apr. 18, 2012).

37. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L. & Committee, o. b. o. t. A. L. Q. A. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* **17,** 405–423. ISSN: 1098-3600 (May 2015).

38. Reva, B., Antipin, Y. & Sander, C. Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. *Nucleic Acids Research* **39,** e118–e118. ISSN: 0305-1048, 1362-4962 (Jan. 9, 2011).

39. Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. & Shendure, J. A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants. *Nature Genetics* **46,** 310–315. ISSN: 1061-4036 (Mar. 2014).

40. Amendola, L. M., Jarvik, G. P., Leo, M. C., McLaughlin, H. M., Akkari, Y., Amaral, M. D., Berg, J. S., Biswas, S., Bowling, K. M., Conlin, L. K., Cooper, G. M., Dorschner, M. O., Dulik, M. C., Ghazani, A. A., Ghosh, R., Green, R. C., Hart, R., Horton, C., Johnston, J. J., Lebo, M. S., Milosavljevic, A., Ou, J., Pak, C. M., Patel, R. Y., Punj, S., Richards, C. S., Salama, J., Strande, N. T., Yang, Y., Plon, S. E., Biesecker, L. G. & Rehm, H. L. Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *The American Journal of Human Genetics* **98,** 1067–1076. ISSN: 0002-9297 (June 2, 2016).

41. Melamed, R. D., Emmett, K. J., Madubata, C., Rzhetsky, A. & Rabadan, R. Genetic Similarity between Cancers and Comorbid Mendelian Diseases Identifies Candidate Driver Genes. *Nature communications* **6,** 7033. ISSN: 2041-1723 (Apr. 30, 2015).

42. Blair, D. R., Lyttle, C. S., Mortensen, J. M., Bearden, C. F., Jensen, A. B., Khiabanian, H., Melamed, R., Rabadan, R., Bernstam, E. V., Brunak, S., Jensen, L. J., Nicolae, D., Shah, N. H., Grossman, R. L., Cox, N. J., White, K. P. & Rzhetsky, A. A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. *Cell* **155,** 70–80. ISSN: 00928674 (Sept. 2013).

43. Garraway, L. A., Verweij, J. & Ballman, K. V. Precision Oncology: An Overview. *Journal of Clinical Oncology* **31**, 1803–1805. ISSN: 0732-183X, 1527-7755 (May 20, 2013).

44. Trifonov, V., Pasqualucci, L., Tiacci, E., Falini, B. & Rabadan, R. SAVI: A Statistical Algorithm for Variant Frequency Identification. *BMC Systems Biology* **7**, 1–11. ISSN: 1752-0509 (Oct. 1, 2013).

45. Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., Zheng, S., Chakravarty, D., Sanborn, J. Z., Berman, S. H., Beroukhim, R., Bernard, B., Wu, C.-J., Genovese, G., Shmulevich, I., Barnholtz-Sloan, J., Zou, L., Vegesna, R., Shukla, S. A., Ciriello, G., Yung, W., Zhang, W., Sougnez, C., Mikkelsen, T., Aldape, K., Bigner, D. D., Van Meir, E. G., Prados, M., Sloan, A., Black, K. L., Eschbacher, J., Finocchiaro, G., Friedman, W., Andrews, D. W., Guha, A., Iacocca, M., O'Neill, B. P., Foltz, G., Myers, J., Weisenberger, D. J., Penny, R., Kucherlapati, R., Perou, C. M., Hayes, D. N., Gibbs, R., Marra, M., Mills, G. B., Lander, E., Spellman, P., Wilson, R., Sander, C., Weinstein, J., Meyerson, M., Gabriel, S., Laird, P. W., Haussler, D., Getz, G. & Chin, L. The Somatic Genomic Landscape of Glioblastoma. *Cell* **155**, 462–477. ISSN: 00928674 (Oct. 2013).

46. The Cancer Genome Atlas Research Network. Comprehensive Molecular Characterization of Urothelial Bladder Carcinoma. *Nature* **507**, 315–322. ISSN: 0028-0836 (Mar. 20, 2014).

47. Network, T. C. G. A. R. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *New England Journal of Medicine* **372**, 2481–2498. ISSN: 0028-4793 (June 25, 2015).

48. The Cancer Genome Atlas Research Network. Comprehensive Molecular Profiling of Lung Adenocarcinoma. *Nature* **511**, 543–550. ISSN: 0028-0836 (July 31, 2014).

49. Akbani, R. *et al.* Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681–1696. ISSN: 0092-8674 (June 18, 2015).

50. The Cancer Genome Atlas Research Network. Comprehensive Molecular Characterization of Gastric Adenocarcinoma. *Nature* **513**, 202–209. ISSN: 0028-0836 (Sept. 11, 2014).

51. Nowell, P. C. The Clonal Evolution of Tumor Cell Populations. *Science* **194**, 23–28. ISSN: 0036-8075, 1095-9203 (Oct. 1, 1976).

52. Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. & Sirotkin, K. dbSNP: The NCBI Database of Genetic Variation. *Nucleic Acids Research* **29**, 308–311. ISSN: 0305-1048, 1362-4962 (Jan. 1, 2001).

53. Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U. & Campbell, P. J. COSMIC: Exploring the World's Knowledge of Somatic Mutations in Human Cancer. *Nucleic Acids Research* **43**, D805–D811. ISSN: 0305-1048, 1362-4962 (Jan. 28, 2015).

54. Kim, J., Kim, S., Nam, H., Kim, S. & Lee, D. SoloDel: A Probabilistic Model for Detecting Low-Frequent Somatic Deletions from Unmatched Sequencing Data. *Bioinformatics* **31**, 3105–3113. ISSN: 1367-4803, 1460-2059 (Jan. 10, 2015).

55. Smith, K. S., Yadav, V. K., Pei, S., Pollyea, D. A., Jordan, C. T. & De, S. SomVarIUS: Somatic Variant Identification from Unpaired Tissue Samples. *Bioinformatics,* btv685. ISSN: 1367-4803, 1460-2059 (Nov. 20, 2015).

56. Hiltemann, S., Jenster, G., Trapman, J., van der Spek, P. & Stubbs, A. Discriminating Somatic and Germline Mutations in Tumour DNA Samples without Matching Normals. *Genome Research,* gr.183053.114. ISSN: 1088-9051, 1549-5469 (July 24, 2015).

57. Sun, J. X., He, Y., Sanford, E., Montesion, M., Frampton, G. M., Vignot, S., Soria, J.-C., Ross, J. S., Miller, V. A., Stephens, P. J., Lipson, D. & Yelensky, R. A Computational Approach to Distinguish Somatic vs. Germline Origin of Genomic Alterations from Deep Sequencing of Cancer Specimens without a Matched Normal. *PLOS Computational Biology* **14** (ed Dunbrack, R. L.) e1005965. ISSN: 1553-7358 (Feb. 7, 2018).

58. Abate, F., Ambrosio, M. R., Mundo, L., Laginestra, M. A., Fuligni, F., Rossi, M., Zairis, S., Gazaneo, S., Falco, G. D., Lazzi, S., Bellan, C., Rocca, B. J., Amato, T., Marasco, E., Etebari, M., Ogwang, M., Calbi, V., Ndede, I., Patel, K., Chumba, D., Piccaluga, P. P., Pileri, S., Leoncini, L. & Rabadan, R. Distinct Viral and Mutational Spectrum of Endemic Burkitt Lymphoma. *PLOS Pathog* **11**, e1005158. ISSN: 1553-7374 (Oct. 15, 2015).

59. Palomero, T., Couronné, L., Khiabanian, H., Kim, M.-Y., Ambesi-Impiombato, A., Perez-Garcia, A., Carpenter, Z., Abate, F., Allegretta, M., Haydu, J. E., Jiang, X., Lossos, I. S., Nicolas, C., Balbin, M., Bastard, C., Bhagat, G., Piris, M. A., Campo, E., Bernard, O. A., Rabadan, R. & Ferrando, A. A. Recurrent Mutations in Epigenetic Regulators, RHOA and FYN Kinase in Peripheral T Cell Lymphomas. *Nature Genetics* **46**, 166–170. ISSN: 1061-4036 (Feb. 2014).

60. Tzoneva, G., Perez-Garcia, A., Carpenter, Z., Khiabanian, H., Tosello, V., Allegretta, M., Paietta, E., Racevskis, J., Rowe, J. M., Tallman, M. S., Paganin, M., Basso, G., Hof, J., Kirschner-Schwabe, R., Palomero, T., Rabadan, R. & Ferrando, A. Activating Mutations in the NT5C2 Nucleotidase Gene Drive Chemotherapy Resistance in Relapsed ALL. *Nature Medicine* **19**, 368–371. ISSN: 1078-8956 (Mar. 2013).

61. Schwartzentruber, J., Korshunov, A., Liu, X.-Y., Jones, D. T. W., Pfaff, E., Jacob, K., Sturm, D., Fontebasso, A. M., Quang, D.-A. K., Tönjes, M., Hovestadt, V., Albrecht, S., Kool, M., Nantel, A., Konermann, C., Lindroth, A., Jäger, N., Rausch, T., Ryzhova, M., Korbel, J. O., Hielscher, T., Hauser, P., Garami, M., Klekner, A., Bognar, L., Ebinger, M., Schuhmann, M. U., Scheurlen, W., Pekrun, A., Frühwald, M. C., Roggendorf, W., Kramm, C., Dürken, M., Atkinson, J., Lepage, P., Montpetit, A., Zakrzewska, M., Zakrzewski, K., Liberski, P. P., Dong, Z., Siegel, P., Kulozik, A. E., Zapatka, M., Guha, A., Malkin, D., Felsberg, J., Reifenberger, G., von Deimling, A., Ichimura, K., Collins, V. P., Witt, H., Milde, T., Witt, O., Zhang, C., Castelo-Branco, P., Lichter, P., Faury, D., Tabori, U., Plass, C., Majewski, J., Pfister, S. M. & Jabado, N. Driver Mutations in Histone H3.3 and Chromatin Remodelling Genes in Paediatric Glioblastoma. *Nature* **482**, 226–231. ISSN: 0028-0836 (Feb. 9, 2012).

62. Caruana, R. & Niculescu-Mizil, A. *An Empirical Comparison of Supervised Learning Algorithms* in (ACM Press, 2006), 161–168. ISBN: 978-1-59593-383-6. doi:10.1145/1143844.1143865. <http://portal.acm.org/citation.cfm?doid=1143844.1143865> (visited on 08/23/2017).

63. Friedman, J., Hastie, T. & Tibshirani, R. Additive Logistic Regression: A Statistical View of Boosting (With Discussion and a Rejoinder by the Authors). *The Annals of Statistics* **28**, 337–407. ISSN: 0090-5364 (Apr. 2000).

64. Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A. & Bejerano, G. M-CAP Eliminates a Majority of Variants of Uncertain Significance in Clinical Exomes at High Sensitivity. *Nature Genetics* **48**, 1581–1586. ISSN: 1061-4036 (Dec. 2016).

65. Abate, F., Zairis, S., Ficarra, E., Acquaviva, A., Wiggins, C. H., Frattini, V., Lasorella, A., Iavarone, A., Inghirami, G. & Rabadan, R. Pegasus: A Comprehensive Annotation and Prediction Tool for Detection of Driver Gene Fusions in Cancer. *BMC Systems Biology* **8**, 97. ISSN: 1752-0509 (Sept. 4, 2014).

66. Friedman, J. H. Stochastic Gradient Boosting. *Computational Statistics & Data Analysis. Nonlinear Methods and Data Mining* **38**, 367–378. ISSN: 0167-9473 (Feb. 28, 2002).

67. Consortium, T. . G. P. An Integrated Map of Genetic Variation from 1,092 Human Genomes. *Nature* **491**, 56–65. ISSN: 0028-0836 (Nov. 1, 2012).

68. Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S. & Getz, G. Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples. *Nature Biotechnology* **31**, 213–219. ISSN: 1087-0156 (Mar. 2013).

69. Taylor, K. R., Mackay, A., Truffaux, N., Butterfield, Y. S., Morozova, O., Philippe, C., Castel, D., Grasso, C. S., Vinci, M., Carvalho, D., Carcaboso, A. M., de Torres, C., Cruz, O., Mora, J., Entz-Werle, N., Ingram, W. J., Monje, M., Hargrave, D., Bullock, A. N., Puget, S., Yip, S., Jones, C. & Grill, J. Recurrent Activating ACVR1 Mutations in Diffuse Intrinsic Pontine Glioma. *Nature Genetics* **46**, 457–461. ISSN: 1061-4036 (May 2014).

70. Wu, G., Diaz, A. K., Paugh, B. S., Rankin, S. L., Ju, B., Li, Y., Zhu, X., Qu, C., Chen, X., Zhang, J., Easton, J., Edmonson, M., Ma, X., Lu, C., Nagahawatte, P., Hedlund, E., Rusch, M., Pounds, S., Lin, T., Onar-Thomas, A., Huether, R., Kriwacki, R., Parker, M., Gupta, P., Becksfort, J., Wei, L., Mulder, H. L., Boggs, K., Vadodaria, B., Yergeau, D., Russell, J. C., Ochoa, K., Fulton, R. S., Fulton, L. L., Jones, C., Boop, F. A., Broniscer, A., Wetmore, C., Gajjar, A., Ding, L., Mardis, E. R., Wilson, R. K., Taylor, M. R., Downing, J. R., Ellison, D. W., Zhang, J. & Baker, S. J. The Genomic Landscape of Diffuse Intrinsic Pontine Glioma and Pediatric Non-Brainstem High-Grade Glioma. *Nature Genetics* **46**, 444–450. ISSN: 1061-4036, 1546-1718 (Apr. 6, 2014).

71. Fontebasso, A. M., Papillon-Cavanagh, S., Schwartzentruber, J., Nikbakht, H., Gerges, N., Fiset, P.-O., Bechet, D., Faury, D., De Jay, N., Ramkissoon, L. A., Corcoran, A., Jones, D. T. W., Sturm, D., Johann, P., Tomita, T., Goldman, S., Nagib, M., Bendel, A., Goumnerova, L., Bowers, D. C., Leonard, J. R., Rubin, J. B., Alden, T., Browd, S., Geyer, J. R., Leary, S., Jallo, G., Cohen, K., Gupta, N., Prados, M. D., Carret, A.-S., Ellezam, B., Crevier, L., Klekner, A., Bognar, L., Hauser, P., Garami, M., Myseros, J., Dong, Z., Siegel, P. M., Malkin, H., Ligon, A. H., Albrecht, S., Pfister, S. M., Ligon, K. L., Majewski, J., Jabado, N. & Kieran, M. W. Recurrent Somatic Mutations in ACVR1 in Pediatric Midline High-Grade Astrocytoma. *Nature Genetics* **46**, 462–466. ISSN: 1061-4036 (May 2014).

72. Wang, J., Cazzato, E., Ladewig, E., Frattini, V., Rosenbloom, D. I. S., Zairis, S., Abate, F., Liu, Z., Elliott, O., Shin, Y.-J., Lee, J.-K., Lee, I.-H., Park, W.-Y., Eoli, M., Blumberg, A. J., Lasorella, A., Nam, D.-H., Finocchiaro, G., Iavarone, A. & Rabadan, R. Clonal Evolution of Glioblastoma under Therapy. *Nature Genetics* **48**, 768–776. ISSN: 1061-4036 (July 2016).

73. Elith, J., Leathwick, J. R. & Hastie, T. A Working Guide to Boosted Regression Trees. *Journal of Animal Ecology* **77**, 802–813. ISSN: 0021-8790, 1365-2656 (July 2008).

74. Rubio-Perez, C., Tamborero, D., Schroeder, M. P., Antolín, A. A., Deu-Pons, J., Perez-Llamas, C., Mestres, J., Gonzalez-Perez, A. & Lopez-Bigas, N. In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer Cell* **27**, 382–396. ISSN: 1535-6108 (Mar. 9, 2015).

75. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the Effects of Coding Non-Synonymous Variants on Protein Function Using the SIFT Algorithm. *Nature Protocols* **4**, 1073–1081. ISSN: 1754-2189 (June 2009).

76. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster Evaluates Disease-Causing Potential of Sequence Alterations. *Nature methods* **7**, 575–576 (2010).

77. Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C. & Schultz, N. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* **2**, 401–404. issn: 2159-8274, 2159-8290 (Jan. 5, 2012).

78. Li, H. & Durbin, R. Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform. *Bioinformatics* **25**, 1754–1760. issn: 1367-4803, 1460-2059 (July 15, 2009).

79. Li, H. A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data. *Bioinformatics* **27**, 2987–2993. issn: 1367-4803, 1460-2059 (Jan. 11, 2011).

80. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. & Ruden, D. M. A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3. *Fly* **6**, 80–92. issn: 1933-6934 (Apr. 1, 2012).

81. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: A Database of Human Non-Synonymous SNVs and Their Functional Predictions and Annotations. *Human Mutation* **34**, E2393–E2402. issn: 1098-1004 (2013).

82. Kuhn, M. & Johnson, K. in *Applied Predictive Modeling* 419–443 (Springer New York, 2013). isbn: 978-1-4614-6848-6 978-1-4614-6849-3. doi:`10.1007/978-1-4614-6849-3_16`. <`http://link.springer.com/chapter/10.1007/978-1-4614-6849-3_16`> (visited on 08/22/2016).

83. Madubata, C. J., Roshan-Ghias, A., Chu, T., Resnick, S., Zhao, J., Arnes, L., Wang, J. & Rabadan, R. Identification of Potentially Oncogenic Alterations from Tumor-Only Samples Reveals Fanconi Anemia Pathway Mutations in Bladder Carcinomas. *npj Genomic Medicine* **2**, 29. issn: 2056-7944 (Oct. 3, 2017).

84. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. A Method and Server for Predicting Damaging Missense Mutations. *Nature Methods* **7**, 248–249. issn: 1548-7091 (Apr. 2010).

85. Roberts, N. D., Kortschak, R. D., Parker, W. T., Schreiber, A. W., Branford, S., Scott, H. S., Glonek, G. & Adelson, D. L. A Comparative Analysis of Algorithms for Somatic SNV Detection in Cancer. *Bioinformatics* **29**, 2223–2230. issn: 1367-4803 (Sept. 15, 2013).

86. Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H. & Vilo, J. G:Profiler—a Web Server for Functional Interpretation of Gene Lists (2016 Update). *Nucleic Acids Research,* gkw199. ISSN: 0305-1048, 1362-4962 (Apr. 20, 2016).

87. Schlacher, K., Wu, H. & Jasin, M. A Distinct Replication Fork Protection Pathway Connects Fanconi Anemia Tumor Suppressors to RAD51-BRCA1/2. *Cancer Cell* **22,** 106–116. ISSN: 1535-6108, 1878-3686 (July 10, 2012).

88. Naseem, H, Boylan, J, Speake, D, Leask, K, Shenton, A, Lalloo, F, Hill, J, Trump, D & Evans, D. Inherited Association of Breast and Colorectal Cancer: Limited Role of CHEK2 Compared with High-Penetrance Genes. *Clinical Genetics* **70,** 388–395. ISSN: 1399-0004 (Nov. 1, 2006).

89. Kuznetsov, S. G., Liu, P. & Sharan, S. K. Mouse Embryonic Stem Cell–Based Functional Assay to Evaluate Mutations in BRCA2. *Nature Medicine* **14,** 875–881. ISSN: 1078-8956 (Aug. 2008).

90. Peterlongo, P., Catucci, I., Colombo, M., Caleca, L., Mucaki, E., Bogliolo, M., Marin, M., Damiola, F., Bernard, L., Pensotti, V., Volorio, S., Dall'Olio, V., Meindl, A., Bartram, C., Sutter, C., Surowy, H., Sornin, V., Dondon, M.-G., Eon-Marchais, S., Stoppa-Lyonnet, D., Andrieu, N., Sinilnikova, O. M., Mitchell, G., James, P. A., Thompson, E., Marchetti, M., Verzeroli, C., Tartari, C., Capone, G. L., Putignano, A. L., Genuardi, M., Medici, V., Marchi, I., Federico, M., Tognazzo, S., Matricardi, L., Agata, S., Dolcetti, R., Puppa, L. D., Cini, G., Gismondi, V., Viassolo, V., Perfumo, C., Mencarelli, M. A., Baldassarri, M., Peissel, B., Roversi, G., Silvestri, V., Rizzolo, P., Spina, F., Vivanet, C., Tibiletti, M. G., Caligo, M. A., Gambino, G., Tommasi, S., Pilato, B., Tondini, C., Corna, C., Bonanni, B., Barile, M., Osorio, A., Benitez, J., Balestrino, L., Ottini, L., Manoukian, S., Pierotti, M. A., Renieri, A., Varesco, L., Couch, F. J., Wang, X., Devilee, P., Hilbers, F. S., Asperen, V., J, C., Viel, A., Montagna, M., Cortesi, L., Diez, O., Balmaña, J., Hauke, J., Schmutzler, R. K., Papi, L., Pujana, M. A., Lázaro, C., Falanga, A., Offit, K., Vijai, J., Campbell, I., Burwinkel, B., Kvist, A., Ehrencrona, H., Mazoyer, S., Pizzamiglio, S., Verderio, P., Surralles, J., Rogan, P. K. & Radice, P. FANCM c.5791C>T Nonsense Mutation (Rs144567652) Induces Exon Skipping, Affects DNA Repair Activity and Is a Familial Breast Cancer Risk Factor. *Human Molecular Genetics* **24,** 5345–5355. ISSN: 0964-6906 (Sept. 15, 2015).

91. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinsk, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N.,

Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, Icgc PedBrain, Zucman-Rossi, J., Andrew Futreal, P., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J. & Stratton, M. R. Signatures of Mutational Processes in Human Cancer. *Nature* **500**, 415–421. ISSN: 0028-0836 (Aug. 22, 2013).

92. Nickerson, M. L., Dancik, G. M., Im, K. M., Edwards, M. G., Turan, S., Brown, J., Ruiz-Rodriguez, C., Owens, C., Costello, J. C., Guo, G., Tsang, S. X., Li, Y., Zhou, Q., Cai, Z., Moore, L. E., Lucia, M. S., Dean, M. & Theodorescu, D. Concurrent Alterations in TERT, KDM6A, and the BRCA Pathway in Bladder Cancer. *Clinical Cancer Research* **20**, 4935–4948. ISSN: 1078-0432, 1557-3265 (Sept. 15, 2014).

93. Pritchard, C. C., Mateo, J., Walsh, M. F., De Sarkar, N., Abida, W., Beltran, H., Garofalo, A., Gulati, R., Carreira, S., Eeles, R., Elemento, O., Rubin, M. A., Robinson, D., Lonigro, R., Hussain, M., Chinnaiyan, A., Vinson, J., Filipenko, J., Garraway, L., Taplin, M.-E., AlDubayan, S., Han, G. C., Beightol, M., Morrissey, C., Nghiem, B., Cheng, H. H., Montgomery, B., Walsh, T., Casadei, S., Berger, M., Zhang, L., Zehir, A., Vijai, J., Scher, H. I., Sawyers, C., Schultz, N., Kantoff, P. W., Solit, D., Robson, M., Van Allen, E. M., Offit, K., de Bono, J. & Nelson, P. S. Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. *New England Journal of Medicine* **0**, null. ISSN: 0028-4793 (July 6, 2016).

94. Tutt, A., Robson, M., Garber, J. E., Domchek, S. M., Audeh, M. W., Weitzel, J. N., Friedlander, M., Arun, B., Loman, N., Schmutzler, R. K., Wardley, A., Mitchell, G., Earl, H., Wickens, M. & Carmichael, J. Oral Poly(ADP-Ribose) Polymerase Inhibitor Olaparib in Patients with BRCA1 or BRCA2 Mutations and Advanced Breast Cancer: A Proof-of-Concept Trial. *The Lancet* **376**, 235–244. ISSN: 0140-6736 (July 30, 2010).

95. Plimack, E. R., Dunbrack, R. L., Brennan, T. A., Andrake, M. D., Zhou, Y., Serebriiskii, I. G., Slifker, M., Alpaugh, K., Dulaimi, E., Palma, N., Hoffman-Censits, J., Bilusic, M., Wong, Y.-N., Kutikov, A., Viterbo, R., Greenberg, R. E., Chen, D. Y. T., Lallas, C. D., Trabulsi, E. J., Yelensky, R., McConkey, D. J., Miller, V. A., Golemis, E. A. & Ross, E. A. Defects in DNA Repair Genes Predict Response to Neoadjuvant Cisplatin-Based Chemotherapy in Muscle-Invasive Bladder Cancer. *European Urology* **68**, 959–967. ISSN: 0302-2838 (Dec. 2015).

96. Byrski, T., Dent, R., Blecharz, P., Foszczynska-Kloda, M., Gronwald, J., Huzarski, T., Cybulski, C., Marczyk, E., Chrzan, R., Eisen, A., Lubinski, J. & Narod, S. A. Results of a Phase II Open-Label, Non-Randomized Trial of Cisplatin Chemotherapy in Patients with BRCA1-Positive Metastatic Breast Cancer. *Breast Cancer Research* **14**, R110. ISSN: 1465-542X (2012).

97. Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J. & Bateman, A. The Pfam Protein Families Database: Towards a More Sustainable Future. *Nucleic Acids Research* **44**, D279–D285. ISSN: 0305-1048, 1362-4962 (Jan. 4, 2016).

98. Ceccaldi, R., Sarangi, P. & D'Andrea, A. D. The Fanconi Anaemia Pathway: New Players and New Functions. *Nature Reviews Molecular Cell Biology* **17**, 337–349. ISSN: 1471-0072 (June 2016).

99. Ward, E., DeSantis, C., Robbins, A., Kohler, B. & Jemal, A. Childhood and Adolescent Cancer Statistics, 2014: Cancer in Children and Adolescents. *CA: A Cancer Journal for Clinicians* **64**, 83–103. ISSN: 00079235 (Mar. 2014).

100. Hunger, S. P. & Mullighan, C. G. Acute Lymphoblastic Leukemia in Children. *New England Journal of Medicine* **373** (ed Longo, D. L.) 1541–1552. ISSN: 0028-4793, 1533-4406 (Oct. 15, 2015).

101. Belver, L. & Ferrando, A. The Genetics and Mechanisms of T Cell Acute Lymphoblastic Leukaemia. *Nature Reviews Cancer* **16**, 494–507. ISSN: 1474-1768 (Aug. 2016).

102. Freyer, D. R., Devidas, M., La, M., Carroll, W. L., Gaynon, P., Hunger, S. P. & Seibel, N. L. Post-Relapse Survival in Childhood Acute Lymphoblastic Leukemia Is Independent of Initial Treatment Intensity: A Report from the Children's Oncology Group. *Blood,* blood–2010–07–294678. ISSN: 0006-4971, 1528-0020 (Jan. 1, 2010).

103. Roy, A., Cargill, A., Love, S., Moorman, A. V., Stoneham, S., Lim, A., Darbyshire, P. J., Lancaster, D., Hann, I., Eden, T. & Saha, V. Outcome after First Relapse in Childhood Acute Lymphoblastic Leukaemia - Lessons from the United Kingdom R2 Trial. *British Journal of Haematology* **130**, 67–75. ISSN: 0007-1048, 1365-2141 (July 2005).

104. Carroll, W. L. & Raetz, E. A. Clinical and Laboratory Biology of Childhood Acute Lymphoblastic Leukemia. *The Journal of Pediatrics* **160**, 10–18. ISSN: 0022-3476 (Jan. 1, 2012).

105. Schultz, K. R., Pullen, D. J., Sather, H. N., Shuster, J. J., Devidas, M., Borowitz, M. J., Carroll, A. J., Heerema, N. A., Rubnitz, J. E., Loh, M. L., Raetz, E. A., Winick, N. J., Hunger, S. P., Carroll, W. L., Gaynon, P. S. & Camitta, B. M. Risk- and Response-Based Classification of Childhood B-Precursor Acute Lymphoblastic Leukemia: A Combined Analysis of Prognostic Markers from the Pediatric Oncology Group (POG) and Children's Cancer Group (CCG). *Blood* **109**, 926–935. ISSN: 0006-4971, 1528-0020 (Feb. 1, 2007).

106. Pui, C.-H. & Evans, W. E. Treatment of Acute Lymphoblastic Leukemia. *New England Journal of Medicine* **354,** 166–178. ISSN: 0028-4793 (Jan. 12, 2006).

107. Pui, C.-H., Sandlund, J. T., Pei, D., Campana, D., Rivera, G. K., Ribeiro, R. C., Rubnitz, J. E., Razzouk, B. I., Howard, S. C., Hudson, M. M., Cheng, C., Kun, L. E., Raimondi, S. C., Behm, F. G., Downing, J. R., Relling, M. V. & Evans, W. E. Improved Outcome for Children with Acute Lymphoblastic Leukemia: Results of Total Therapy Study XIIIB at St Jude Children's Research Hospital. *Blood* **104,** 2690–2696. ISSN: 0006-4971, 1528-0020 (Nov. 1, 2004).

108. Rives, S., Estella, J., Gómez, P., López-Duarte, M., de Miguel, P. G., Verdeguer, A., Moreno, M. J., Vivanco, J. L., Couselo, J. M., Fernández-Delgado, R., Maldonado, M., Tasso, M., López-Ibor, B., Lendínez, F., López-Almaraz, R., Uriz, J., Melo, M., Fernández-Teijeiro, A., Rodríguez, I. & Badell, I. Intermediate Dose of Imatinib in Combination with Chemotherapy Followed by Allogeneic Stem Cell Transplantation Improves Early Outcome in Paediatric Philadelphia Chromosome-Positive Acute Lymphoblastic Leukaemia (ALL): Results of the Spanish Cooperative G: Imatinib in Paediatric Ph+ ALL. *British Journal of Haematology* **154,** 600–611. ISSN: 00071048 (Sept. 2011).

109. Lauer, S. J., Shuster, J. J., Jr, D. M., Winick, N., Toledano, S., Munoz, L., Kiefer, G., Pullen, J. D., Steuber, C. P. & Camitta, B. M. A Comparison of Early Intensive Methotrexate/Mercaptopurine with Early Intensive Alternating Combination Chemotherapy for High-Risk B-Precursor Acute Lymphoblastic Leukemia: A Pediatric Oncology Group Phase III Randomized Trial. *Leukemia* **15,** 1038–1045. ISSN: 1476-5551 (July 2001).

110. Seibel, N. L., Steinherz, P. G., Sather, H. N., Nachman, J. B., DeLaat, C., Ettinger, L. J., Freyer, D. R., Mattano, L. A., Hastings, C. A., Rubin, C. M., Bertolone, K., Franklin, J. L., Heerema, N. A., Mitchell, T. L., Pyesmany, A. F., La, M. K., Edens, C. & Gaynon, P. S. Early Postinduction Intensification Therapy Improves Survival for Children and Adolescents with High-Risk Acute Lymphoblastic Leukemia: A Report from the Children's Oncology Group. *Blood* **111,** 2548–2555. ISSN: 0006-4971, 1528-0020 (Mar. 1, 2008).

111. Möricke, A., Zimmermann, M., Reiter, A., Henze, G., Schrauder, A., Gadner, H., Ludwig, W. D., Ritter, J., Harbott, J., Mann, G., Klingebiel, T., Zintl, F., Niemeyer, C., Kremens, B., Niggli, F., Niethammer, D., Welte, K., Stanulla, M., Odenwald, E., Riehm, H. & Schrappe, M. Long-Term Results of Five Consecutive Trials in Childhood Acute Lymphoblastic Leukemia Performed by the ALL-BFM Study Group from 1981 to 2000. *Leukemia* **24,** 265–284. ISSN: 1476-5551 (Feb. 2010).

112. Moerloose, B. D., Suciu, S., Bertrand, Y., Mazingue, F., Robert, A., Uyttebroeck, A., Yakouben, K., Ferster, A., Margueritte, G., Lutz, P., Munzer, M., Sirvent, N., Norton, L., Boutard, P., Plantaz, D., Millot, F., Philippet, P., Baila, L., Benoit, Y., Otten, J., of

Cancer (EORTC), f. t. C. L. G. C. o. t. E. O. f. R. & Treatment. Improved Outcome with Pulses of Vincristine and Corticosteroids in Continuation Therapy of Children with Average Risk Acute Lymphoblastic Leukemia (ALL) and Lymphoblastic Non-Hodgkin Lymphoma (NHL): Report of the EORTC Randomized Phase 3 Trial 58951. *Blood* **116**, 36–44. ISSN: 0006-4971, 1528-0020 (July 8, 2010).

113. Bhatia, S., Landier, W., Hageman, L., Kim, H., Chen, Y., Crews, K. R., Evans, W. E., Bostrom, B., Casillas, J., Dickens, D. S., Maloney, K. W., Neglia, J. P., Ravindranath, Y., Ritchey, A. K., Wong, F. L. & Relling, M. V. 6MP Adherence in a Multiracial Cohort of Children with Acute Lymphoblastic Leukemia: A Children's Oncology Group Study. *Blood* **124**, 2345–2353. ISSN: 0006-4971, 1528-0020 (Oct. 9, 2014).

114. Pui, C.-H., Pei, D., Coustan-Smith, E., Jeha, S., Cheng, C., Bowman, W. P., Sandlund, J. T., Ribeiro, R. C., Rubnitz, J. E., Inaba, H., Bhojwani, D., Gruber, T. A., Leung, W. H., Downing, J. R., Evans, W. E., Relling, M. V. & Campana, D. Clinical Utility of Sequential Minimal Residual Disease Measurements in the Context of Risk-Based Therapy in Childhood Acute Lymphoblastic Leukaemia: A Prospective Study. *The Lancet Oncology* **16**, 465–474. ISSN: 1470-2045 (Apr. 1, 2015).

115. Locatelli, F., Schrappe, M., Bernardo, M. E. & Rutella, S. How I Treat Relapsed Childhood Acute Lymphoblastic Leukemia. *Blood* **120**, 2807–2816. ISSN: 0006-4971, 1528-0020 (Oct. 4, 2012).

116. Pierro, J., Hogan, L. E., Bhatla, T. & Carroll, W. L. New Targeted Therapies for Relapsed Pediatric Acute Lymphoblastic Leukemia. *Expert Review of Anticancer Therapy* **17**, 725–736. ISSN: 1473-7140 (Aug. 3, 2017).

117. Nguyen, K., Devidas, M., Cheng, S.-C., La, M., Raetz, E. A., Carroll, W. L., Winick, N. J., Hunger, S. P., Gaynon, P. S. & Loh, M. L. Factors Influencing Survival after Relapse from Acute Lymphoblastic Leukemia: A Children's Oncology Group Study. *Leukemia* **22**, 2142–2150. ISSN: 1476-5551 (Dec. 2008).

118. Raetz, E. A. & Bhatla, T. Where Do We Stand in the Treatment of Relapsed Acute Lymphoblastic Leukemia? *ASH Education Program Book* **2012**, 129–136. ISSN: 1520-4391, 1520-4383 (Aug. 12, 2012).

119. Oshima, K., Khiabanian, H., da Silva-Almeida, A. C., Tzoneva, G., Abate, F., Ambesi-Impiombato, A., Sanchez-Martin, M., Carpenter, Z., Penson, A., Perez-Garcia, A., Eckert, C., Nicolas, C., Balbin, M., Sulis, M. L., Kato, M., Koh, K., Paganin, M., Basso, G., Gastier-Foster, J. M., Devidas, M., Loh, M. L., Kirschner-Schwabe, R., Palomero, T., Rabadan, R. & Ferrando, A. A. Mutational Landscape, Clonal Evolution Patterns, and Role of RAS Mutations in Relapsed Acute Lymphoblastic Leukemia. *Proceedings of the National Academy of Sciences* **113**, 11306–11311. ISSN: 0027-8424, 1091-6490 (Apr. 10, 2016).

120. Liu, Y., Easton, J., Shao, Y., Maciaszek, J., Wang, Z., Wilkinson, M. R., McCastlain, K., Edmonson, M., Pounds, S. B., Shi, L., Zhou, X., Ma, X., Sioson, E., Li, Y., Rusch, M., Gupta, P., Pei, D., Cheng, C., Smith, M. A., Auvil, J. G., Gerhard, D. S., Relling, M. V., Winick, N. J., Carroll, A. J., Heerema, N. A., Raetz, E., Devidas, M., Willman, C. L., Harvey, R. C., Carroll, W. L., Dunsmore, K. P., Winter, S. S., Wood, B. L., Sorrentino, B. P., Downing, J. R., Loh, M. L., Hunger, S. P., Zhang, J. & Mullighan, C. G. The Genomic Landscape of Pediatric and Young Adult T-Lineage Acute Lymphoblastic Leukemia. *Nature Genetics* **49**, 1211–1218. ISSN: 1061-4036, 1546-1718 (July 3, 2017).

121. Mullighan, C. G., Su, X., Zhang, J., Radtke, I., Phillips, L. A., Miller, C. B., Ma, J., Liu, W., Cheng, C., Schulman, B. A., Harvey, R. C., Chen, I.-M., Clifford, R. J., Carroll, W. L., Reaman, G., Bowman, W. P., Devidas, M., Gerhard, D. S., Yang, W., Relling, M. V., Shurtleff, S. A., Campana, D., Borowitz, M. J., Pui, C.-H., Smith, M., Hunger, S. P., Willman, C. L., & Downing, J. R. Deletion of IKZF1 and Prognosis in Acute Lymphoblastic Leukemia. *New England Journal of Medicine* **360**, 470–480. ISSN: 0028-4793 (Jan. 29, 2009).

122. Li, B., Li, H., Bai, Y., Kirschner-Schwabe, R., Yang, J. J., Chen, Y., Lu, G., Tzoneva, G., Ma, X., Wu, T., Li, W., Lu, H., Ding, L., Liang, H., Huang, X., Yang, M., Jin, L., Kang, H., Chen, S., Du, A., Shen, S., Ding, J., Chen, H., Chen, J., von Stackelberg, A., Gu, L., Zhang, J., Ferrando, A., Tang, J., Wang, S. & Zhou, B.-B. S. Negative Feedback-Defective PRPS1 Mutants Drive Thiopurine Resistance in Relapsed Childhood ALL. *Nature Medicine* **21**, 563–571. ISSN: 1078-8956 (June 2015).

123. Ma, X., Edmonson, M., Yergeau, D., Muzny, D. M., Hampton, O. A., Rusch, M., Song, G., Easton, J., Harvey, R. C., Wheeler, D. A., Ma, J., Doddapaneni, H., Vadodaria, B., Wu, G., Nagahawatte, P., Carroll, W. L., Chen, I.-M., Gastier-Foster, J. M., Relling, M. V., Smith, M. A., Devidas, M., Auvil, J. M. G., Downing, J. R., Loh, M. L., Willman, C. L., Gerhard, D. S., Mullighan, C. G., Hunger, S. P. & Zhang, J. Rise and Fall of Subclones from Diagnosis to Relapse in Pediatric B-Acute Lymphoblastic Leukaemia. *Nature Communications* **6**. doi:10.1038/ncomms7604. <http://www.nature.com/ncomms/2015/150319/ncomms7604/full/ncomms7604.html> (visited on 04/07/2015) (Mar. 19, 2015).

124. Mullighan, C. G. Molecular Genetics of B-Precursor Acute Lymphoblastic Leukemia. *The Journal of Clinical Investigation* **122**, 3407–3415. ISSN: 0021-9738 (Oct. 1, 2012).

125. Ma, X., Liu, Y., Liu, Y., Alexandrov, L. B., Edmonson, M. N., Gawad, C., Zhou, X., Li, Y., Rusch, M. C., Easton, J., Huether, R., Gonzalez-Pena, V., Wilkinson, M. R., Hermida, L. C., Davis, S., Sioson, E., Pounds, S., Cao, X., Ries, R. E., Wang, Z., Chen, X., Dong, L., Diskin, S. J., Smith, M. A., Guidry Auvil, J. M., Meltzer, P. S., Lau, C. C., Perlman, E. J., Maris, J. M., Meshinchi, S., Hunger, S. P., Gerhard, D. S. & Zhang, J. Pan-Cancer Genome and Transcriptome Analyses of 1,699 Paediatric Leukaemias and Solid Tumours. *Nature* **555**, 371–376. ISSN: 1476-4687 (Mar. 2018).

126. Kuster, L., Grausenburger, R., Fuka, G., Kaindl, U., Krapf, G., Inthal, A., Mann, G., Kauer, M., Rainer, J., Kofler, R., Hall, A., Metzler, M., Meyer, L. H., Meyer, C., Harbott, J., Marschalek, R., Strehl, S., Haas, O. A. & Panzer-Grümayer, R. ETV6/RUNX1-Positive Relapses Evolve from an Ancestral Clone and Frequently Acquire Deletions of Genes Implicated in Glucocorticoid Signaling. *Blood* **117**, 2658–2667. ISSN: 0006-4971, 1528-0020 (Mar. 3, 2011).

127. Kuiper, R. P., Waanders, E., van der Velden, V. H. J., van Reijmersdal, S. V., Venkatachalam, R., Scheijen, B., Sonneveld, E., van Dongen, J. J. M., Veerman, A. J. P., van Leeuwen, F. N., van Kessel, A. G. & Hoogerbrugge, P. M. *IKZF1* Deletions Predict Relapse in Uniformly Treated Pediatric Precursor B-ALL. *Leukemia* **24**, 1258–1264. ISSN: 1476-5551 (July 2010).

128. Mullighan, C. G., Zhang, J., Kasper, L. H., Lerach, S., Payne-Turner, D., Phillips, L. A., Heatley, S. L., Holmfeldt, L., Collins-Underwood, J. R., Ma, J., Buetow, K. H., Pui, C.-H., Baker, S. D., Brindle, P. K. & Downing, J. R. *CREBBP* Mutations in Relapsed Acute Lymphoblastic Leukaemia. *Nature* **471**, 235–239. ISSN: 1476-4687 (Mar. 2011).

129. Hsiao, M. H., Yu, A. L., Yeargin, J., Ku, D. & Haas, M. Nonhereditary P53 Mutations in T-Cell Acute Lymphoblastic Leukemia Are Associated with the Relapse Phase. *Blood* **83**, 2922–2930. ISSN: 0006-4971, 1528-0020 (May 15, 1994).

130. Goker, E., Waltham, M., Kheradpour, A., Trippett, T., Mazumdar, M., Elisseyeff, Y., Schnieders, B., Steinherz, P., Tan, C. & Berman, E. Amplification of the Dihydrofolate Reductase Gene Is a Mechanism of Acquired Resistance to Methotrexate in Patients with Acute Lymphoblastic Leukemia and Is Correlated with P53 Gene Mutations. *Blood* **86**, 677–684. ISSN: 0006-4971, 1528-0020 (July 15, 1995).

131. Krentz, S., Hof, J., Mendioroz, A., Vaggopoulou, R., Dörge, P., Lottaz, C., Engelmann, J. C., Groeneveld, T. W. L., Körner, G., Seeger, K., Hagemeier, C., Henze, G., Eckert, C., von Stackelberg, A. & Kirschner-Schwabe, R. Prognostic Value of Genetic Alterations in Children with First Bone Marrow Relapse of Childhood B-Cell Precursor Acute Lymphoblastic Leukemia. *Leukemia* **27**, 295–304. ISSN: 1476-5551 (Feb. 2013).

132. Mar, B. G., Bullinger, L. B., McLean, K. M., Grauman, P. V., Harris, M. H., Stevenson, K., Neuberg, D. S., Sinha, A. U., Sallan, S. E., Silverman, L. B., Kung, A. L., Nigro, L. L., Ebert, B. L. & Armstrong, S. A. Mutations in Epigenetic Regulators Including SETD2 Are Gained during Relapse in Paediatric Acute Lymphoblastic Leukaemia. *Nature Communications* **5**, 3469. ISSN: 2041-1723 (Mar. 24, 2014).

133. Ferrando, A. A. & López-Otín, C. Clonal Evolution in Leukemia. *Nature Medicine* **23**, 1135–1145. ISSN: 1546-170X (Oct. 2017).

134. Meyer, J. A., Wang, J., Hogan, L. E., Yang, J. J., Dandekar, S., Patel, J. P., Tang, Z., Zumbo, P., Li, S., Zavadil, J., Levine, R. L., Cardozo, T., Hunger, S. P., Raetz, E. A.,

Evans, W. E., Morrison, D. J., Mason, C. E. & Carroll, W. L. Relapse-Specific Mutations in *NT5C2* in Childhood Acute Lymphoblastic Leukemia. *Nature Genetics* **45**, 290–294. ISSN: 1546-1718 (Mar. 2013).

135. Tzoneva, G., Dieck, C. L., Oshima, K., Ambesi-Impiombato, A., Sánchez-Martín, M., Madubata, C. J., Khiabanian, H., Yu, J., Waanders, E., Iacobucci, I., Sulis, M. L., Kato, M., Koh, K., Paganin, M., Basso, G., Gastier-Foster, J. M., Loh, M. L., Kirschner-Schwabe, R., Mullighan, C. G., Rabadan, R. & Ferrando, A. A. Clonal Evolution Mechanisms in NT5C2 Mutant-Relapsed Acute Lymphoblastic Leukaemia. *Nature.* ISSN: 0028-0836, 1476-4687. doi:10.1038/nature25186. <http://www.nature.com/doifinder/10.1038/nature25186> (visited on 01/26/2018) (Jan. 17, 2018).

136. Roberts, K. G., Harvey, R. C., Yang, W., Cheng, C., Pei, D., Xu, H., Gastier-Fostier, J., E, S., Lim, J. Y.-S., Chen, I.-M. L., Fan, Y., Devidas, M., Borowitz, M. J., Smith, C., Neale, G. A., Burchard, E. G., Torgerson, D. G., Antillon, F., Rolando, C., Winick, N. J., Camitta, B., Raetz, E. A., Wood, B. L., Yue, F., Carroll, W. L., Larsen, E. C., Bowman, W. P., Loh, M. L., Dean, M., Bhojwani, D., Pui, C.-H., Evans, W. E., Relling, M. V., Hunger, S. P., Willman, C. L., Mullighan, C. G., Frcpa & Yang, J. J. Inherited GATA3 Genetic Variants Are Associated With Childhood BCR-ABL1-Like Acute Lymphoblastic Leukemia and Increased Risk Of Relapse. *Blood* **122**, 617–617. ISSN: 0006-4971, 1528-0020 (Nov. 15, 2013).

137. Perez-Andreu, V., Roberts, K. G., Harvey, R. C., Yang, W., Cheng, C., Pei, D., Xu, H., Gastier-Foster, J., E, S., Lim, J. Y.-S., Chen, I.-M., Fan, Y., Devidas, M., Borowitz, M. J., Smith, C., Neale, G., Burchard, E. G., Torgerson, D. G., Klussmann, F. A., Villagran, C. R. N., Winick, N. J., Camitta, B. M., Raetz, E., Wood, B., Yue, F., Carroll, W. L., Larsen, E., Bowman, W. P., Loh, M. L., Dean, M., Bhojwani, D., Pui, C.-H., Evans, W. E., Relling, M. V., Hunger, S. P., Willman, C. L., Mullighan, C. G. & Yang, J. J. Inherited GATA3 Variants Are Associated with Ph-like Childhood Acute Lymphoblastic Leukemia and Risk of Relapse. *Nature Genetics* **45**, 1494–1498. ISSN: 1061-4036 (Dec. 2013).

138. Xu, H., Zhang, H., Yang, W., Yadav, R., Morrison, A. C., Qian, M., Devidas, M., Liu, Y., Perez-Andreu, V., Zhao, X., Gastier-Foster, J. M., Lupo, P. J., Neale, G., Raetz, E., Larsen, E., Bowman, W. P., Carroll, W. L., Winick, N., Williams, R., Hansen, T., Holm, J.-C., Mardis, E., Fulton, R., Pui, C.-H., Zhang, J., Mullighan, C. G., Evans, W. E., Hunger, S. P., Gupta, R., Schmiegelow, K., Loh, M. L., Relling, M. V. & Yang, J. J. Inherited Coding Variants at the CDKN2A Locus Influence Susceptibility to Acute Lymphoblastic Leukaemia in Children. *Nature Communications* **6**. doi:10.1038/ncomms8553. <http://www.nature.com/ncomms/2015/150624/ncomms8553/full/ncomms8553.html> (visited on 06/29/2015) (June 24, 2015).

139. Waanders, E., Scheijen, B., Jongmans, M. C. J., Venselaar, H., van Reijmersdal, S. V., van Dijk, A. H. A., Pastorczak, A., Weren, R. D. A., van der Schoot, C. E., van de Vorst, M., Sonneveld, E., Hoogerbrugge, N., van der Velden, V. H. J., Gruhn, B.,

Hoogerbrugge, P. M., van Dongen, J. J. M., van Kessel, A. G., van Leeuwen, F. N. & Kuiper, R. P. Germline Activating *TYK2* Mutations in Pediatric Patients with Two Primary Acute Lymphoblastic Leukemia Occurrences. *Leukemia* **31,** 821–828. ISSN: 1476-5551 (Apr. 2017).

140. Owen, C. J., Toze, C. L., Koochin, A., Forrest, D. L., Smith, C. A., Stevens, J. M., Jackson, S. C., Poon, M.-C., Sinclair, G. D., Leber, B., Johnson, P. R. E., Macheta, A., Yin, J. A. L., Barnett, M. J., Lister, T. A. & Fitzgibbon, J. Five New Pedigrees with Inherited RUNX1 Mutations Causing Familial Platelet Disorder with Propensity to Myeloid Malignancy. *Blood* **112,** 4639–4645. ISSN: 0006-4971, 1528-0020 (Dec. 1, 2008).

141. Smith, M. L., Cavenagh, J. D., Lister, T. A. & Fitzgibbon, J. Mutation of CEBPA in Familial Acute Myeloid Leukemia. *New England Journal of Medicine* **351,** 2403–2407. ISSN: 0028-4793 (Dec. 2, 2004).

142. Martínez-Glez, V. & Lapunzina, P. Sotos Syndrome Is Associated with Leukemia/-Lymphoma. *American Journal of Medical Genetics Part A* **143A,** 1244–1245. ISSN: 15524825, 15524833 (June 1, 2007).

143. Patil, S. & Chamberlain, R. S. Neoplasms Associated with Germline and Somatic NF1 Gene Mutations. *The Oncologist* **17,** 101–116. ISSN: 1083-7159, 1549-490X (Jan. 1, 2012).

144. Qian, M., Cao, X., Devidas, M., Yang, W., Cheng, C., Dai, Y., Carroll, A., Heerema, N. A., Zhang, H., Moriyama, T., Gastier-Foster, J. M., Xu, H., Raetz, E., Larsen, E., Winick, N., Bowman, W. P., Martin, P. L., Mardis, E. R., Fulton, R., Zambetti, G., Borowitz, M., Wood, B., Nichols, K. E., Carroll, W. L., Pui, C.-H., Mullighan, C. G., Evans, W. E., Hunger, S. P., Relling, M. V., Loh, M. L. & Yang, J. J. TP53 Germline Variations Influence the Predisposition and Prognosis of B-Cell Acute Lymphoblastic Leukemia in Children. *Journal of Clinical Oncology* **36,** 591–599. ISSN: 0732-183X (Jan. 4, 2018).

145. Heesch, S., Goekbuget, N., Stroux, A., Tanchez, J. O., Schlee, C., Burmeister, T., Schwartz, S., Blau, O., Keilholz, U., Busse, A., Hoelzer, D., Thiel, E., Hofmann, W.-K. & Baldus, C. D. Prognostic Implications of Mutations and Expression of the Wilms Tumor 1 (WT1) Gene in Adult Acute T-Lymphoblastic Leukemia. *Haematologica* **95,** 942–949. ISSN: 0390-6078, 1592-8721 (June 1, 2010).

146. Wang, J., Khiabanian, H., Rossi, D., Fabbri, G., Gattei, V., Forconi, F., Laurenti, L., Marasca, R., Poeta, G. D., Foà, R., Pasqualucci, L., Gaidano, G. & Rabadan, R. Tumor Evolutionary Directed Graphs and the History of Chronic Lymphocytic Leukemia. *eLife.* Cancer is a clonal evolutionary process, caused by successive accumulation of genetic alterations providing milestones of tumor initiation, progression, dissemination and/or resistance to certain therapeutic regimes. To unravel these milestones we propose a framework, tumor evolutionary directed graphs (TEDG), which is able

to characterize the history of genetic alterations by integrating longitudinal and cross-sectional genomic data. We applied TEDG to a chronic lymphocytic leukemia (CLL) cohort of 70 patients spanning 12 years, and show that: (a) the evolution of CLL follows a time-ordered process represented as a global flow in TEDG that proceeds from initiating events to late events; (b) there are two distinct and mutually exclusive evolutionary paths of CLL evolution; (c) higher fitness clones are present in later stages of the disease, indicating a progressive clonal replacement with more aggressive clones. Our results suggest that TEDG may constitute an effective framework to recapitulate the evolutionary history of tumors., e02869. ISSN: 2050-084X (Dec. 11, 2014).

147. Neri, A., Knowles, D. M., Greco, A., McCormick, F. & Dalla-Favera, R. Analysis of RAS Oncogene Mutations in Human Lymphoid Malignancies. *Proceedings of the National Academy of Sciences* **85**, 9268–9272. ISSN: 0027-8424, 1091-6490 (Dec. 1, 1988).

148. Malinowska-Ozdowy, K., Frech, C., Schönegger, A., Eckert, C., Cazzaniga, G., Stanulla, M., zur Stadt, U., Mecklenbräuker, A., Schuster, M., Kneidinger, D., von Stackelberg, A., Locatelli, F., Schrappe, M., Horstmann, M. A., Attarbaschi, A., Bock, C., Mann, G., Haas, O. A. & Panzer-Grümayer, R. *KRAS* and *CREBBP* Mutations: A Relapse-Linked Malicious Liaison in Childhood High Hyperdiploid Acute Lymphoblastic Leukemia. *Leukemia* **29**, 1656–1667. ISSN: 1476-5551 (Aug. 2015).

149. Messina, M., Chiaretti, S., Wang, J., Fedullo, A. L., Peragine, N., Gianfelici, V., Piciocchi, A., Brugnoletti, F., Giacomo, F. D., Pauselli, S., Holmes, A. B., Puzzolo, M. C., Ceglie, G., Apicella, V., Mancini, M., te Kronnie, G., Testi, A. M., Vitale, A., Vignetti, M., Guarini, A., Rabadan, R., Foà, R., Messina, M., Chiaretti, S., Wang, J., Fedullo, A. L., Peragine, N., Gianfelici, V., Piciocchi, A., Brugnoletti, F., Giacomo, F. D., Pauselli, S., Holmes, A. B., Puzzolo, M. C., Ceglie, G., Apicella, V., Mancini, M., te Kronnie, G., Testi, A. M., Vitale, A., Vignetti, M., Guarini, A., Rabadan, R. & Foà, R. Prognostic and Therapeutic Role of Targetable Lesions in B-Lineage Acute Lymphoblastic Leukemia without Recurrent Fusion Genes. *Oncotarget* **7**, 13886–13901. ISSN: 1949-2553 (Feb. 12, 2016).

150. Roohi, J., Crowe, J., Loredan, D., Anyane-Yeboa, K., Mansukhani, M. M., Omesi, L., Levine, J., Politi, A. R. & Zha, S. New Diagnosis of Atypical Ataxia-Telangiectasia in a 17-Year-Old Boy with T-Cell Acute Lymphoblastic Leukemia and a Novel *ATM* Mutation. *Journal of Human Genetics* **62**, 581–584. ISSN: 1435-232X (May 2017).

151. Taylor, A. M., Metcalfe, J. A., Thick, J. & Mak, Y. F. Leukemia and Lymphoma in Ataxia Telangiectasia. *Blood* **87**, 423–438. ISSN: 0006-4971, 1528-0020 (Jan. 15, 1996).

152. Yamaguchi, H., Calado, R. T., Ly, H., Kajigaya, S., Baerlocher, G. M., Chanock, S. J., Lansdorp, P. M. & Young, N. S. Mutations in *TERT,* the Gene for Telomerase

Reverse Transcriptase, in Aplastic Anemia. *New England Journal of Medicine* **352,** 1413–1424. ISSN: 0028-4793, 1533-4406 (Apr. 7, 2005).

153. Du, H.-Y., Pumbo, E., Manley, P., Field, J. J., Bayliss, S. J., Wilson, D. B., Mason, P. J. & Bessler, M. Complex Inheritance Pattern of Dyskeratosis Congenita in Two Families with 2 Different Mutations in the Telomerase Reverse Transcriptase Gene. *Blood* **111,** 1128–1130. ISSN: 0006-4971, 1528-0020 (Feb. 1, 2008).

154. Tosello, V., Mansour, M. R., Barnes, K., Paganin, M., Sulis, M. L., Jenkinson, S., Allen, C. G., Gale, R. E., Linch, D. C., Palomero, T., Real, P., Murty, V., Yao, X., Richards, S. M., Goldstone, A., Rowe, J., Basso, G., Wiernik, P. H., Paietta, E., Pieters, R., Horstmann, M., Meijerink, J. P. P. & Ferrando, A. A. WT1 Mutations in T-ALL. *Blood* **114,** 1038–1045. ISSN: 0006-4971, 1528-0020 (July 30, 2009).

155. Bordin, F., Piovan, E., Masiero, E., Ambesi-Impiombato, A., Minuzzo, S., Bertorelle, R., Sacchetto, V., Pilotto, G., Basso, G., Zanovello, P., Amadori, A. & Tosello, V. WT1 Loss Attenuates the TP53-Induced DNA Damage Response in T-Cell Acute Lymphoblastic Leukemia. *Haematologica* **103,** 266–277. ISSN: 0390-6078, 1592-8721 (Feb. 1, 2018).

156. Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. & Stratton, M. R. A Census of Human Cancer Genes. *Nature Reviews Cancer* **4,** 177–183. ISSN: 1474-175X (Mar. 2004).

157. Van Vlierberghe, P. & Ferrando, A. The Molecular Basis of T Cell Acute Lymphoblastic Leukemia. *Journal of Clinical Investigation* **122,** 3398–3406. ISSN: 0021-9738 (Oct. 1, 2012).

158. Zhang, J., Mullighan, C. G., Harvey, R. C., Wu, G., Chen, X., Edmonson, M., Buetow, K. H., Carroll, W. L., Chen, I.-M., Devidas, M., Gerhard, D. S., Loh, M. L., Reaman, G. H., Relling, M. V., Camitta, B. M., Bowman, W. P., Smith, M. A., Willman, C. L., Downing, J. R. & Hunger, S. P. Key Pathways Are Frequently Mutated in High-Risk Childhood Acute Lymphoblastic Leukemia: A Report from the Children's Oncology Group. *Blood* **118,** 3080–3087. ISSN: 0006-4971, 1528-0020 (Sept. 15, 2011).

159. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. & Daly, M. J. A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data. *Nature Genetics* **43,** 491–498. ISSN: 1061-4036 (May 2011).

160. 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., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-

Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., MacArthur, D. G. & Exome Aggregation Consortium. Analysis of Protein-Coding Genetic Variation in 60,706 Humans. *Nature* **536**, 285–291. ISSN: 0028-0836 (Aug. 18, 2016).

161. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P. & Cunningham, F. The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122. ISSN: 1474-760X (June 6, 2016).

162. Bouaoun, L., Sonkin, D., Ardin, M., Hollstein, M., Byrnes, G., Zavadil, J. & Olivier, M. *TP53* Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data: Human Mutation. *Human Mutation* **37**, 865–876. ISSN: 10597794 (Sept. 2016).