Identification of potentially oncogenic alterations from tumor-only samples reveals Fanconi anemia pathway mutations in bladder carcinomas

Chioma J Madubata1,2, Alireza Roshan-Ghias1,3, Timothy Chu1,2, Samuel Resnick1,2, Junfei Zhao1,2, Luis Arnes1,2, Jiguang Wang1,2,3 and Raul Rabadan1,2

Cancer is caused by germline and somatic mutations, which can share biological features such as amino acid change. However, integrated germline and somatic analysis remains uncommon. We present a framework that uses machine learning to learn features of recurrent somatic mutations to (1) predict somatic variants from tumor-only samples and (2) identify somatic-like germline variants for integrated analysis of tumor-normal DNA. Using data from 1769 patients from seven cancer types (bladder, glioblastoma, low-grade glioma, lung, melanoma, stomach, and pediatric glioma), we show that “somatic-like” germline variants are enriched for autosomal-dominant cancer-predisposition genes ($p < 4.35 \times 10^{-15}$), including TP53. Our framework identifies germline and somatic nonsense variants in BRCA2 and other Fanconi anemia genes in $11\%$ (11/100) of bladder cancer cases, suggesting a potential genetic predisposition in these patients. The bladder carcinoma patients with Fanconi anemia nonsense variants display a BRCA-deficiency somatic mutation signature, suggesting treatment targeted to DNA repair.

ARTICLE

INTRODUCTION

Cancer often results from specific DNA alterations, and identification of cancer-causing mutations underlies genome-based precision cancer treatment. Somatic mutations can be identified by sequencing matched tumor and normal DNA, 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). Despite the value in sequencing matched normal DNA to truly differentiate germline and somatic variants, the historically high cost of sequencing led to tumor-only sequencing in many research projects and clinical settings. Tumor-only sequencing contains both germline and somatic alterations, but differentiating the 10–100s of somatic mutations from tens of thousands of germline variants remains challenging. Common attempts to identify somatic variants from tumor-only WES data involve removing dbSNP mutations common in the general population and focusing on genes in the Catalogue Of Somatic Mutations In Cancer (COSMIC). These strategies fail 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 or predicting somatic single nucleotide variants (SNV) using base quality, variant allele frequency (VAF), and sequencing error. Other strategies use population allele frequency tabulated from a cohort of normal genomes to remove potential germline SNPs. 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.

Integrated information from individual patients, patient cohorts, and databases can inform an alternative approach that learns biological features from known somatic variants in order to predict somatic variants from tumor-only samples. 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 test 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, relapsed pediatric acute lymphoblastic leukemia, and pediatric glioma, but the focus on gene identity had decreased power to identify oncogenic variants. In contrast, the approach we suggest would use machine learning instead of manual analysis, make predictions across the whole exome instead of focusing on specific genes, and use multiple biological features to increase power to predict somatic variants.

While tumor-only analysis remains common and somatic mutations associate with cancer development, germline DNA alterations can also be oncogenic. 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 (FA) and MAPK pathways. Furthermore, certain oncogenic germline variants share biological features with known somatic variants, such as affecting the same amino acid. A machine learning framework built upon biological features of somatic variants would have high power to identify germline variants with somatic features that might influence tumor development.

Thus, we present a framework to address the separate but related challenges of tumor-only somatic analysis and integrated germline–somatic analysis. Our Tumor-Only Boosting Identification framework (TOBI) learns from a small training set of tumor-normal pairs to generate a classification model that identifies variants with somatic characteristics from tumor-only samples. If normal DNA is available, we can assess whether TOBI predicted certain germline variants as somatic; we refer to these variants as “somatic-like” germline variants. Somatic-like germline variants complement the somatic landscape, promoting integrated analysis of oncogenic processes. TOBI uses gradient boosting, a machine learning algorithm with consistently superior performance in diverse classification tasks. Using 1769 patients across seven tumor types, we developed TOBI, evaluated TOBI’s ability to identify somatic variants, and identified somatic-like germline variants (SLG variants), including variants with known or possible oncogenic potential.

RESULTS
Framework for predicting somatic, germline and SLG 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 VAF when available (Fig. 1a). Step I receives aligned WES files (bam files), calls variants against a human reference genome, and annotates variants (full details in 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 Methods, retaining high quality variants that are rare in the population (population minor allele frequency less than 1% in the 1000 Genomes Project).

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. 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 (Methods). Each variant in the training set represents an observation for machine learning. Ten biological features were used for gradient boosting (full features in Supplementary Text); features 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, SAVE). 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, and assessed TOBI on five adult cancer types from TCGA: bladder urothelial carcinoma (BLCA), brain lower grade glioma (LGG), lung adenocarcinoma (LUAD), skin cutaneous melanoma (SKCM), and stomach adenocarcinoma (STAD). 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, and tumor-only cases. The number of cases per cancer type, and the number of cases used in each figure, is in Supplementary Table 1a.

Since cancer-sequencing studies have variable numbers of paired tumor-normal samples, we assessed the number of training cases required for model generation (Fig. 1b). Increasing the number of training set tumor samples from 1 to 50 samples 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 (Supplementary Table 1b). 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 (Supplementary Fig. 7, Supplementary Table 7), and observed a slightly decreased F-score for FFPE (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 a patient’s reported race, TOBI had decreased mean F-scores when the training and test set differed by race in almost all cancers (Supplementary Fig. 2, Supplementary Table 2c). 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 in Supplementary Fig. 3a–c and Supplementary Table 3; Ped.Glioma analysis in Supplementary Fig. 3d). Finally, using hypermutator status from the STAD publication, we found no significant effect on TOBI’s performance when analyzing a non-hypermutator population or mixed population (61 hypermutator, 219 non-hypermutator) (Supplementary Fig. 4, Supplementary Table 4). Thus, TOBI’s performance might improve with features denoting patient race 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, a measure of how frequently one feature is used in the decision trees within the final classification model (Fig. 1c). 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 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 (Supplementary Fig. 1).
High performance somatic variant identification

We compared TOBI’s somatic classifications to published somatic calls from tumor-normal analysis of test set cases. 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 Supplementary Table 5 and Supplementary Figure 5a. TOBI also has high sensitivity for variants with tumor VAF as low as 5% (Supplementary figure 5b,c). Per gene, the number of cases with nonsynonymous variants predicted as somatic closely matched published somatic analysis (Figs. 2a, b). 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, Supplementary Figure 13). TOBI predictions on previously published somatic mutations are in Supplementary Table 6.

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. In six cancers, TOBI has a higher true positive rate of nonsynonymous variants in driver genes.
compared to all genes (Fig. 2b). Such enrichment occurred despite training sets containing 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 (Supplementary Fig. 6).

Finally, to demonstrate analysis of a truly tumor-only data set, we applied the pediatric glioma classification model to 68 tumor-only samples; genes predicted in at least three cases shown. For all cancers, twenty randomly selected tumor-normal cases comprised training set; remaining paired tumor-normal samples formed testing set.
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) and SomVarIUS. 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). 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 SomVarIUS (TOBI: ~5–50; VNC: ~300,000; SomVarIUS: ~100–3000). 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 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.

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, FP variants might be tumor-specific variants that were not previously published due to variability in somatic variant analysis.

First, we assessed TOBI’s overall false positive rate (FPR) in the cancer test sets. Since FP variants may include SLG variants, we also calculated the FPR from applying the Ped.Glioma classification model to a set of 100 non-tumor exomes from individuals without cancer sequenced by the 1000 Genomes Project. 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 (Supplementary Fig. 9). The higher FPR from tumor cohorts suggests that some FP calls represent somatic-like germline variants.

To identify SLG variants, we analyzed germline VAF from 1327 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 FP variant needed a germline VAF of at least 30% to decrease the probability that the germline variant represented tumor contamination or artifacts. Since certain germline variants highly increase predisposition to cancer, we analyzed SLG variants for enrichment in 60 genes associated with autosomal dominant cancer-predisposition syndromes, or “AD genes” (listed in Supplementary Table 9), and found significant enrichment of AD genes in nonsynonymous SLG variants (p < 1.33 x 10^-10; Fig. 3a). SLG nonsynonymous variants in TP53 occurred in seven cases (Fig. 3c). Certain inactivating mutations in tumor suppressors are heterozygous germline variants, but show loss of heterozygosity in the tumor. Five of TP53 SLG variants exhibit evidence of loss of heterozygosity, with germline VAFs below 45% and tumor VAFs above 70%.

Focusing on nonsynonymous FP variants in AD genes, we found 15 cases with TP53 mutations and at least seven cases with mutations in CDH1, RB1, RET or TSC2 (Fig. 3b). In three Ped.Glioma cases, TOBI 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, illustrating that TOBI can identify somatic variants that may be inconsistently called.

Certain germline variants in cancer-associated genes correlate with earlier age of diagnosis, so we analyzed whether presence of nonsynonymous SLG variants in 565 cancer-associated genes (list in Supplementary Table 9) associated with earlier age of diagnosis in any cancer type. Supplementary table 10 provides the number of cases with SLG variants in these cancer-associated genes for each cancer type. In LGG, patients with cancer-associated SLG variants had significantly earlier age at diagnosis (median 37 years vs. 41 years, p = 0.0013; Fig. 4a; Supplementary Fig. 10). 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 [COSM2573612]).

Unified somatic-germline study with tumor-only data

CJ Madubata et al.
variants within the FA pathway for this BLCA cohort. Somatic copy number alterations and predicted nonsynonymous SLG nonsense variants in the FA pathway based on pathway type. Bladder carcinoma cases showed significant enrichment of SLG nonsense variants in low-grade glioma associated with earlier age of diagnosis.

Bladder cancer cases with inactivating mutations in FA pathway display somatic signature of BRCA-deficiency. Truncating germline alterations in cancer predisposition genes have been reported in 4–19% of cancer types. Accordingly, we examined the exome-wide SLG nonsense variants in each cancer type. Bladder carcinoma cases showed significant enrichment of SLG nonsense variants in the FA pathway based on pathway assessment with g:Profiler (49 genes with SLG variants, 54 genes in FA pathway, 3 overlapping genes; p-value of 0.029 after multiple testing correction; Supplementary Fig. 11). The FA pathway normally performs DNA repair of interstrand crosslinks, which requires homologous recombination.

We then assessed the overall occurrence of germline and somatic nonsense mutations in the FA pathway predicted by TOBI. 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, affecting 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 (Fig. 5b: FANCM R1931*, BRCA2 Y3308*). Of note, BRCA2 variant Y3308* has been associated with hereditary colorectal and breast cancer. Mice ES cells with BRCA2 Y3308* mutations showed hypersensitivity to ionizing radiation and crosslinking agents, as well as decreased homologous recombination efficiency. Additionally, FANCM R1931* was associated with increased breast cancer risk and deficient DNA repair. Fig. 5c and Supplementary Table 11 describe published somatic copy number alterations and predicted nonsynonymous variants within 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 including our 100 test cases, we generated trinucleotide mutational spectra that decomposed into four somatic signatures (Supplementary Fig. 12a,b). Cases with FA nonsense mutations were only enriched in the fourth signature (Fig. 4c), a somatic signature similar to the BRCA1/2-deficiency signature from a pan-cancer analysis (signature 3 in the referenced publication). Enrichment of this somatic mutation signature in bladder cancer cases with nonsense FA variants suggests that these FA nonsense variants, whether somatic or germline, affect the bladder cancer somatic mutation landscape.

DISCUSSION

In this report, we present 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. Our framework is available online for non-commercial use (https://github.com/RabadanLab/TOBI).

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 capture complex, non-linear dependencies between biological features.
boosting algorithm to classify variants, allowing TOBI to learn features important to specific tumor types (Fig. 1c).

When germline VAF information is available, TOBI can identify "somatic-like" germline variants. 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. SLG variants in cancer genes also associated with earlier age of diagnosis in patients with low-grade glioma (Fig. 4a), suggesting that TOBI's SLG variants are enriched for cancer-associated variants.

Analysis of bladder carcinoma cases using TOBI revealed largely unreported germline inactivating mutations in the FA pathway, suggesting a potential genetic predisposition in 5% of patients. Additionally, recent research 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. 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.

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, so PARP inhibitors may also show increased effectiveness in bladder tumors with BRCA2 or other FA mutations. Additionally, recent research found that the presence of tumor DNA alterations in FANCC predicted beneficial response to cisplatin neoadjuvant chemotherapy. 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.

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...
anxiety, 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. Fourth, for germline variant analysis, 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 FA variants in bladder cancer requires experimental validation.

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, 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. In cases with matched normal DNA, this framework identifies germline variants that present 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 FA pathway in bladder cancer.

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 TOBLbcf pathway indicated in Fig. 1b. 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 cBioPortal64 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 EGAD00000100080723 (St. Jude Children’s Research Hospital—Washington University Pediatric Cancer Genome Project Steering Committee) and EGAD00000100070646 (ICR DIPG Data Access Committee). Fastq files were available for EGAD00000100079235 and EGAD0000010007971 (McCill-DKFZ Pediatric Brain Tumour Consortium); samples were mapped to GRCh37.71 using BWA 0.7.1255 before variant calling. Clinical data was retrieved from supplementary tables. Published somatic variant calls were used to label true somatic variants for the 74 paired samples; only experimentally validated somatic mutations from Wu et al. 201466 were included.

For 1000 Genomes Project 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). All GBM, pediatric glioma, and 1000 Genomes Project bam files went through the TOBLbam pathway indicated in Fig. 1b.

Variant calling and annotation

Bam files were analyzed with Samtools and Bcftools56 to call variants, excluding variants with mapping quality lower than 10.

Variants were annotated using Snpeff57 and Snpsift with dbSNP build 144, Cosmic v74, and dbNSFP v2.4 databases.58 We also annotated the variants with an in-house database of common mutations in 219 normal 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 Bcftools56 greater than 60 and variant depth greater 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 SNVs 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 TOBLbcf 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.25 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.59 Instead, TOBI selects a probability threshold that maximizes classification performance; the threshold’s potential range is 0.05–0.95 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 subsample size (constant for each cancer). 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:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

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. F-score can be calculated from sensitivity, specificity, positive predictive value, negative predictive value, prevalence, accuracy, 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 Supplementary Text 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\textsuperscript{42} 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 normalized so that the sum of relative influence of all features equals one hundred.

We defined drivers in Fig. 2 using the list of driver genes provided by the Intogen group.\textsuperscript{43}

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 MbS covered in Agilent SureSelect Human All Exon 50 Mb regions.bed file.

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.\textsuperscript{29} For enrichment of genes in 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 dataset with N variants found in G genes, as the proportion of a gene value greater than (n – 1) with lambda = \( \frac{n}{\lambda} \) using the Poisson function: \( p(n-1) \), \( g \times n / G \), \( n \times G / C0 \). Protein domain names and coordinates from PFAM.\textsuperscript{60}

Clinical data was retrieved from supplementary tables for Ped.Glioma patients and using the R \texttt{cgsr} package for TCGA. To standardize nomenclature for reported race across studies, we reviewed 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 Supplementary Table 1b. We compared the distribution of diagnosis age for cases with or without SLG variants using the Wilcoxon–Mann–Whitney test in \texttt{R}, \texttt{wilcox.test}().

gProfiler\textsuperscript{44} analysis of BLCA nonsense SLG variants was run using defaults (significant only; Hierarchical sorting. Numeric IDs treated as: WIKIGENE_ACC; Significance threshold: 0.01; Statistical domain size: Only annotated genes.) Multiple testing correction for \( p \)-values calculated using the ontology-focused correction method gSCS as previously described.\textsuperscript{43} FA pathway in Fig. 5c modified from KEGG FA pathway and Ceccaldi et al. 2016.\textsuperscript{61} CNV data was retrieved from cBioPortal.

Comparison to other techniques

In order to compare results from TOBI to other techniques, we ran six GBM samples and six Pediatric Glioma samples through SomVariUS (Smith et al., 2015) and VNC (Hiltemann et al. 2015) and compared their results to TOBI.

Code for SomVariUS was obtained through their github page (https://github.com/kylessmith/SomVariUS). To build the reference database, we supplied an hg19 dbSNP bed to generate the required pickle file. The \texttt{call\_mutations} command was then run with the following options:

\begin{verbatim}
gem\_pos All.filt.pickle --dbsnp\_bed All.filt.bed --min\_reads 10 --min\_support 4 --min\_af 0.05 --min\_pvalue 0.0001 --min\_fr 0.8 --min\_qual 25 --min\_se 0.999 --min\_hetero 0.95 --min\_mapq 55 --ref\_filter True --dbsnp\_bed All.filt.bed --min\_baseq 13 --binom False --hapmap All.filt.pickle.
\end{verbatim}

Code VNC was obtained through their github page (https://github.com/Shitlemann/Virtual-Normal-Correction). To build the reference virtual normal, 433 CG-sequenced normal exomes were downloaded from 1000Genomes (https://ftp.ebi.ac.uk/pub/databases/1000genomes/vol1/ftp/historical_data/1000genomes_complete_genomes_indices/20130725.cg_data.untar.index). The virtual-normal-correction-smallvariants.sh script was run using the following commands:

\begin{verbatim}
  --threshold 1 --threshold\_highconf 3
\end{verbatim}

Data availability statement

All genomic datasets used for analysis come from publically accessible repositories after approval for controlled data access:

- The database of Genotypes and Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap): TCGA (accession number phs000178.v9.p8); Clinical information for TCGA from cBioPortal (http://www.cbioportal.org/);
- The Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra); non-TCGA GBM cases analyzed in Wang et al. 2016 (SRP674425);
- The European Genome-phenome Archive (https://www.ebi.ac.uk/ega/): St. Jude Children’s Research Hospital—Washington University Pediatric Cancer Genome Project Steering Committee (EGAD0000100080732), ICR DIPG Data Access Committee (EGAD0000100070645), McGill-DKFZ Pediatric Brain Tumour Consortium (EGAD0000100079231, EGAD0000100079122).

The 1000 Genomes Project phase 3 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/).

Publication MAFs for TCGA samples from the TCGA Data Matrix (now the Genomic Data Commons https://gdc.cancer.gov/). Publication MAFs from all other publications acquired from supplemental tables of publications.

Code availability

The TOBI framework is fully available for academic use on Github (https://github.com/RabadanLab/TOBI). This Github page describes all dependencies and versions. We also have a public Amazon Machine Image (AMI) on Amazon Web Services, that contains all the software, dependencies and reference databases used in this article, and it can be shared upon request.

ACKNOWLEDGEMENTS

This work was funded in part by R01 CA185846-01, R01 CA179044-01A1, and NIH U54 CA193313; J.W. is supported by Precision Medicine Fellowship (UL1 TR000404); A.R. was supported by the Swiss National Science Foundation Postdoctoral Fellowship; C.J.M. is supported by the NIH Medical Scientist Training Program (T32 GM07367) and the National Research Service Award Individual Predoctoral Fellowship (F31CA210607). We want to thank Hossein Khiabanian for helpful discussions during project design. This study makes use of data generated by the St. Jude Children's Research Hospital—Washington University Pediatric Cancer Genome Project (https://explore.pediatriccancergenomproject.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. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

C.J.M., A.R-G., J.W., and R.R. designed the study. C.J.M., A.R-G., T.C. and J.W. developed the pipeline. C.J.M., A.R-G., S.R., J.Z., performed somatic prediction. C.J.M. performed germline analysis. C.J.M., T.C., and J.Z. compared alternative methods. J.W. and CJM analyzed mutational spectra. C.J.M., A.R-G., L.A., J.W., and R.R. wrote the manuscript.

ADDITIONAL INFORMATION

Supplementary Information accompanies the paper on the npj Genomic Medicine website (doi:10.1038/s41525-017-0032-5).

Competing interests: The authors declare that they have no competing financial interests.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Garraway, L. A., Verweij, J. & Ballman, K. V. Precision oncology: an overview. J. Clin. Oncol. 31, 1803–1805 (2013).
2. Nowell, P. C. The clonal evolution of tumor cell populations. Science 194, 23–28 (1976).
3. Brennan, C. W. et al. The somatic genomic landscape of glioblastoma. Cell 155, 462–477 (2013).
4. Network, T. C. G. A. R. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. New Engl. J. Med. 372, 2481–2498 (2015).
5. The Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550 (2014).
6. The Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urethelial bladder carcinoma. Nature 507, 315–322 (2014).
7. Jones, S. et al. Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci. Transl. Med.* **7**, 283ra253–283ra253 (2015).

8. Fontebasso, A. M. et al. Recurrent somatic mutations in ACVR1 in pediatric midline high-grade astrocytoma. *Nat. Genet.* **46**, 462–466 (2014).

9. Kim, J., Kim, S., Nam, H., Kim, S. & Lee, D. SoloDet: a probabilistic model for detecting low-frequent somatic deletions from unmatched sequencing data. *Bioinformatics* **31**, 3105–3113 (2015).

10. Wu, G. et al. The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. *Nat. Genet.* **46**, 444–450 (2014).

11. Raymond, V. M. et al. Germline mutations and cancer risk in the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

12. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer drivers. *Science* **341**, 1370–1374 (2013).

13. Li, H. A statistical framework for SNP calling, mutation discovery, association analysis and variant function prediction from short-read DNA sequence bioinformatics. *Bioinformatics* **25**, 2063–2069 (2009).

14. DePristo, M. A. et al. A framework for variation discovery and genotyping using short-read sequencing technology. *Nat. Genet.* **43**, 191–199 (2011).

15. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **30**, 1303–1311 (2002).

16. Akbani, R. et al. Genomic characterization of gastric adenocarcinoma. *Clin. Cancer Res.* **16**, 513–520 (2010).

17. Schwartzentruber, J. et al. Driver mutations in histone H3.3 and chromatin remodeling genes in paediatric glioblastoma. *Nature* **482**, 226–231 (2012).

18. Tzoneva, G. et al. Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. *Nat. Genet.* **46**, 166–170 (2014).

19. Abate, F. et al. Distinct viral and mutational spectrum of endemic Burkitt lymphoma. *PLoS Pathog.* **11**, e1005158 (2015).

20. Palomero, T. et al. Recurrent mutations in epigenetic regulators, RHOD and FYN kinase in peripheral T cell lymphomas. *Nat. Genet.* **46**, 166–170 (2014).

21. Tzonova, G. et al. Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. *Nat. Med.* **19**, 368–371 (2013).

22. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer drivers. *Science* **341**, 1370–1374 (2013).

23. Kanchi, K. L. et al. Integrated analysis of germline and somatic variants in ovarian cancer. *Nat. Commun.* **6**, 10383 (2015).

24. Zhang, J. et al. Germline mutations in predisposition genes in pediatric cancer. *N. Engl. J. Med.* **370**, 2065–2074 (2014).

25. Caruana, R. & Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms, in *Proceedings of the 23rd international conference on Machine learning*, 161–168, doi:10.1145/1143844.1143865 (ACM, 2006).

26. Consortium, T. G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

27. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).

28. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

29. Trifonov, V., Pascualucchini, L., Tiacci, E., Falini, B. & Rabadán, R. SAVI: a statistical algorithm for variant frequency identification. *BMC Syst. Biol.* **16**, E2402 (2013).

30. Akbani, R. et al. Genomic characterization of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).

31. Tzonova, G. et al. Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. *Nat. Med.* **19**, 368–371 (2013).

32. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer drivers. *Science* **341**, 1370–1374 (2013).

33. Wang, J. et al. Clonal evolution of glioblastoma in vivo. *Nature* **513**, 202–209 (2014).

34. Kanchi, K. L. et al. Integrated analysis of germline and somatic variants in ovarian cancer. *Nat. Commun.* **6**, 10383 (2015).

35. Zhang, J. et al. Germline mutations in predisposition genes in pediatric cancer. *N. Engl. J. Med.* **370**, 2065–2074 (2014).

36. Caruana, R. & Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms, in *Proceedings of the 23rd international conference on Machine learning*, 161–168, doi:10.1145/1143844.1143865 (ACM, 2006).

37. Consortium, T. G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

38. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).

39. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

40. Trifonov, V., Pascualucchini, L., Tiacci, E., Falini, B. & Rabadán, R. SAVI: a statistical algorithm for variant frequency identification. *BMC Syst. Biol.* **16**, E2402 (2013).

41. Akbani, R. et al. Genomic characterization of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).

42. The Cancer Genome Atlas Research Network Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).

43. Kathryn, R. T. et al. Recurrent activating ACVR1 mutations in diffuse intrinsic pontine glioma. *Nat. Genet.* **46**, 457–461 (2014).

44. Wang, J. et al. Clonal evolution of glioblastoma under therapy. *Nat. Genet.* **48**, 768–776 (2016).

45. Ethir, J., Leathwick, J. R. & Hastie, T. A working guide to boosted regression trees. *J. Anim. Ecol.* **77**, 802–813 (2008).

46. Rubio-Perez, C. et al. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targetting opportunities. *Cancer Cell* **27**, 382–396 (2015).

47. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

48. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).

49. Reba, V., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118–e118 (2011).

50. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7**, 575–576 (2010).