OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae677

https://doi.org/10.1093/bib/bbae677

Problem Solving Protocol

Precise identification of somatic and germline variants in the absence of matched normal samples

Hui Li[1,‡], Lu Meng[2,‡], Hongke Wang[2,‡], Liang Cui[2], Heyu Sheng[2], Peiyan Zhao[1], Shuo Hong[2], Xinhua Du[2], Shi Yan[1], Yun Xing[2], Shicheng Feng[2], Yan Zhang[2], Huan Fang[2], Jing Bai[2,3], Yan Liu[1], Shaowei Lan[1], Tao Liu[2], Yanfang Guan[2], Xuefeng Xia[2], Xin Yi[2,*], Ying Cheng[4,*]

[1]The Medical Oncology Translational Research Laboratory, Jilin Provincial Key Laboratory of Molecular Diagnostics for Lung Cancer, Jilin Cancer Hospital, No. 1066, Jinhu Road, Changchun, 130012, China
[2]Geneplus-Beijing Institute, 9th Floor, No. 6 Building, Peking University Medical Industrial Park, Zhongguancun Life Science Park, Beijing, 102206, China
[3]College of Future Technology, Peking University, No. 5 Yiheyuan Road, Beijing, 100871, China
[4]The Department of Medical Oncology, Jilin Cancer Hospital, No. 1066, Jinhu Road, Changchun, 130012, China

*Corresponding authors. Xin Yi, Geneplus-Beijing Institute, China. E-mail: yix@geneplus.org.cn; Ying Cheng, The Department of Medical Oncology, Jilin Cancer Hospital, China. E-mail: jl.cheng@163.com

‡Hui Li, Lu Meng and Hongke Wang contributed equally to this study.

Abstract

Somatic variants play a crucial role in the occurrence and progression of cancer. However, in the absence of matched normal controls, distinguishing between germline and somatic variants becomes challenging in tumor samples. The existing tumor-only genomic analysis methods either suffer from limited performance or insufficient interpretability due to an excess of features. Therefore, there is an urgent need for an alternative approach that can address these issues and have practical implications. Here, we presented OncoTOP, a computational method for genomic analysis without matched normal samples, which can accurately distinguish somatic mutations from germline variants. Reference sample analysis revealed a 0% false positive rate and 99.7% reproducibility for variant calling. Assessing 2864 tumor samples across 18 cancer types yielded a 99.8% overall positive percent agreement and a 99.9% positive predictive value. OncoTOP can also accurately detect clinically actionable variants and subclonal mutations associated with drug resistance. For the prediction of mutation origins, the positive percent agreement stood at 97.4% for predicting somatic mutations and 95.7% for germline mutations. High consistency of tumor mutational burden (TMB) was observed between the results generated by OncoTOP and tumor-normal paired analysis. In a cohort of 97 lung cancer patients treated with immunotherapy, TMB-high patients had prolonged PFS ($P = .02$), proving the reliability of our approach in estimating TMB to predict therapy response. Furthermore, microsatellite instability status showed a strong concordance (97%) with polymerase chain reaction results, and leukocyte antigens class I subtypes and homozygosity achieved an impressive concordance rate of 99.3% and 99.9% respectively, compared to its tumor-normal paired analysis. Thus, OncoTOP exhibited high reliability in variant calling, mutation origin prediction, and biomarker estimation. Its application will promise substantial advantages for clinical genomic testing.

Keywords: OncoTOP; bioinformatics; DNA sequencing; cancer genomics; biomarker

Introduction

In recent years, massively parallel sequencing has emerged as a valuable tool in clinical settings for the characterization of tumor tissues [1, 2]. This technology enables the generation of comprehensive genomic data, which can be utilized to identify actionable alterations that inform treatment decisions [3, 4]. For this purpose, distinguishing somatic mutations from inherited germline variants is regarded as a critical step. However, detecting somatic mutations can be challenging due to heterogeneity and genomic instability in cancer specimens. This is typically solved by sequencing the tumor specimen with the matched normal tissue from the same patient and followed by comparison: variants present in both of the paired samples are determined to be germline variants shared across all cells within an individual, while those detected in tumor specimen only are classified as somatic variants [5, 6]. Owing to its precise determination of somatic mutations that occurred during cancer progression, this approach has been applied in several projects studying genomic features across cancer types, including Pan-Cancer Analysis of Whole Genomes and The Cancer Genome Atlas [7, 8]. Furthermore, genomic sequencing data can provide several informative biomarkers, such as tumor mutational burden (TMB), microsatellite instability (MSI), and human lymphocyte antigen (HLA) subtype, which have been shown to be potential predictors of response to immunotherapies in clinical trials [9–12].

Collecting a matched normal sample, however, is not a standard practice in clinical oncology, making it unconventional in the clinical field. This limitation greatly restricts the explorable research of tumor samples collected in the clinics and hinders the ability to estimate biomarkers for treatment efficacy. Although

Received: June 18, 2024. **Revised:** November 8, 2024. **Accepted:** December 16, 2024
© The Author(s) 2024. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

recent algorithmic advancements have enabled the classification of somatic and germline mutations and the estimation of complex biomarkers in tumors without a matched normal sample, several limitations still persist. These include lengthy runtimes, restrictions on tissue types, and the accuracy of variant calling, which relies on tumor purity and sequencing depth [13, 14]. The inference of tumor-only sequencing data remains to be controversial [15, 16], thus highlighting the need for a highly reliable bioinformatics tool with a comprehensive workflow.

To address these issues, we present OncoTOP (Oncologic Tumor-Only Profiling), a method for analyzing tumor samples in the absence of a matched normal counterpart. It enables (i) the identification of single-nucleotide variants (SNVs), insertions and deletions (InDels) within clinically relevant regions associated with tumor development or targeted therapeutics; (ii) determination of the somatic or germline origins of detected variants; (iii) the evaluation of complex biomarkers in tumor specimens. In this study, we provide a comprehensive and rigorous validation of OncoTOP, showcasing its reliability and robustness.

Results

Detection of genomic variants

The workflow of OncoTOP is illustrated in Fig. 1. Detailed description can be found in Methods. We first determined the limit of detection (LoD) value of OncoTOP for variant calling using contrived samples with controlled variant allele frequency (VAF). As depicted in Fig. 2A, the LoD value for hotspot SNVs and InDels was set as 2% VAF with a variant calling rate of 100% for both SNVs and InDels, while the LoD for nonhotspot SNVs and InDels was set at 5% VAF with a variant calling rate of 99.1% for SNVs and 100% for InDels. These results were validated with 82 tissue samples and their corresponding tumor-normal paired analysis results (Fig. 2B, Table S1). Remarkably, we observed high positive percent agreement (PPA) for both hotspot and nonhotspot SNVs and InDels at the determined LoD values: 99.2% and 98.3% for hotspot SNVs and InDels, and 99.8% and 100% for nonhotspot SNVs and InDels, respectively. Next, we investigated the effect of tumor purity on variant detection using four standard samples spanning four tumor lineages at four different tumor purity levels, with the results from samples at 40% tumor purity regarded as a reference (Table S2). At the tumor purity of 5%, two genomic alternations in sample B failed to be called, resulting in an overall concordance of 90% (18/20). However, at a tumor purity of 10%, the accuracy of mutation detection exceeded 95% for each sample, and the overall variant detection rate reached 98.4% (62/63). Thus, we established the LoD for tumor purity as 10%. In the limit of blank (LoB) study, we evaluated 2015 hotspot and 128 663 non-hotspot mutation sites in 49 normal cell samples. The LoB was determined to be 0.00% for both the hotspot and non-hotspot mutations, as no mutations were detected in these samples (Table S3).

We then conducted a comprehensive evaluation of OncoTOP's ability to detect SNVs and InDels using a dataset of 2864 samples encompassing 18 tumor lineages (Table S4). The detected mutations by OncoD served as the reference for true positive mutations, and only mutations with a VAF of 1% or higher were considered for evaluation in OncoTOP (Table 1). OncoD is our previously developed paired tumor-normal analysis method that has been widely used in genomic variant analysis [17–19]. The results demonstrated a high level of accuracy, with a PPA of 99.8% (99.1%–100%) and a positive predictive value (PPV) of 99.9% (99.7%–100%) for detecting genomic alterations. Furthermore, the overall accuracy remained consistently high across the majority of the 18 tumor lineages, reaching 99.7% (99.2%–100%) (Table 1). To evaluate the precision of variant calling, we utilized six formalin-fixed paraffin-embedded (FFPE) samples with 128 known mutations. Each sample was tested five times. Only mutations with a VAF higher than the determined LoD values were considered for evaluation. The overall variant calling rate was 99.7%, with a coefficient of variation (C.V.) below 20% for nearly all evaluated mutations (Table S5), demonstrating the high precision of OncoTOP in detecting SNVs and InDels.

Actionable or drug-resistant-related variants

Next, we sought to evaluate the performance of OncoTOP in detecting clinically actionable variants, which hold great significance in the clinical setting. To this end, we curated a collection of commonly encountered actionable genetic variants from the OncoKB database [20]. These included well-known oncogenic mutations such as *PIK3CA* p.E542K, p.E545K, p.H1047R, and p.R88Q in breast cancer; *BRAF* p.V600E in colorectal cancer, melanoma, and thyroid cancer; and *EGFR* p.L858R, 19del, 20ins, and p.G12C in nonsmall cell lung cancer (NSCLC). By comparing the results obtained from OncoTOP with those from tumor-normal matched analysis using OncoD, we found that OncoTOP impeccably detected these clinically actionable variants, achieving a remarkable accuracy, PPA, and PPV of 100% for all variants (Table 2). To further assess the performance of OncoTOP in detecting subclonal mutations, we examined the detection capability of *EGFR* p.T790M in NSCLC, a mutation commonly associated with subclones resistant to tyrosine kinase inhibitors [21]. Remarkably, we observed that all 40 cases of *EGFR* p.T790M detected by OncoD were accurately identified by OncoTOP (Table 2). These findings suggest that OncoTOP, solely relying on tumor genomic data, provides performance comparable to tumor-normal paired testing in detecting actionable variants and subclonal mutations associated with drug resistance.

Prediction of somatic/germline origin

In addition to detecting mutations, distinguishing between germline and somatic mutations is also a crucial issue. The accuracy of predicting the origin of mutations was evaluated by comparing the results generated by OncoTOP with those obtained from OncoD. The PPA for predicting somatic mutations was determined to be 97.4% (67 145/68 946), while for predicting germline mutations, it was 95.7% (49 768/52 018) (Fig. 3). The PPV for predicting germline mutations was 96.5%, while for somatic mutations, it was 96.8% (Table S6). We further investigated the influence of tumor lineage on the accuracy of prediction. Among SNVs and InDels, the PPA for classifying somatic mutations exceeded 95% in 16 tumor lineages, with the exception of small cell lung cancer (88.9%) and urothelial carcinoma (93.4%). As for germline mutations, it was higher than 95% in 12 tumor lineages, except for small cell lung cancer (92.2%), ovarian cancer (94.1%), urothelial carcinoma (94.3%), diffuse large B-cell lymphoma (94.1%), esophageal cancer (94.9%), and melanoma (94.4%) (Table S6). These findings highlight the impressive predictive performance of OncoTOP in determining the origin of mutations across a diverse range of cancer types.

OncoTOP employs a decision tree model to discern germline/somatic origins of variants, incorporating three key features: germP, PAD_count, and caseAF (for details, see Methods). To clarify the contributions of each of those features to the model performance, we assessed the performance of individual features

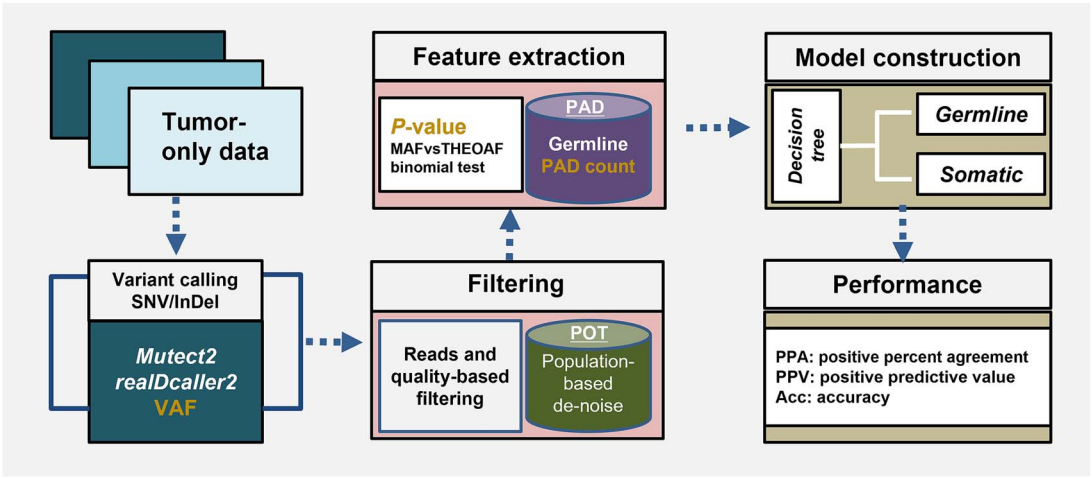

Figure 1. Schematic of OncoTOP analysis workflow. Tumor-only genomic data were obtained using our CGP assay. Genetic variants, including SNV and InDel were detected using realDcaller2 and Mutect2, and filtered with a population database (POT) to minimize background noise. Three important features, such as VAF, *P*-value, and PAD_count were further analyzed using a decision tree model to determine the germline or somatic origin of the detected variants. CGP: comprehensive genomic profiling; SNV: single-nucleotide variant; InDel: insertion/deletion; VAF: variant allele frequency; PAD_count: mutation count in the PAD database. germP: *P*-value of two-tailed binomial test between the MAFs of variants and their THEOAFs. See details in Methods.

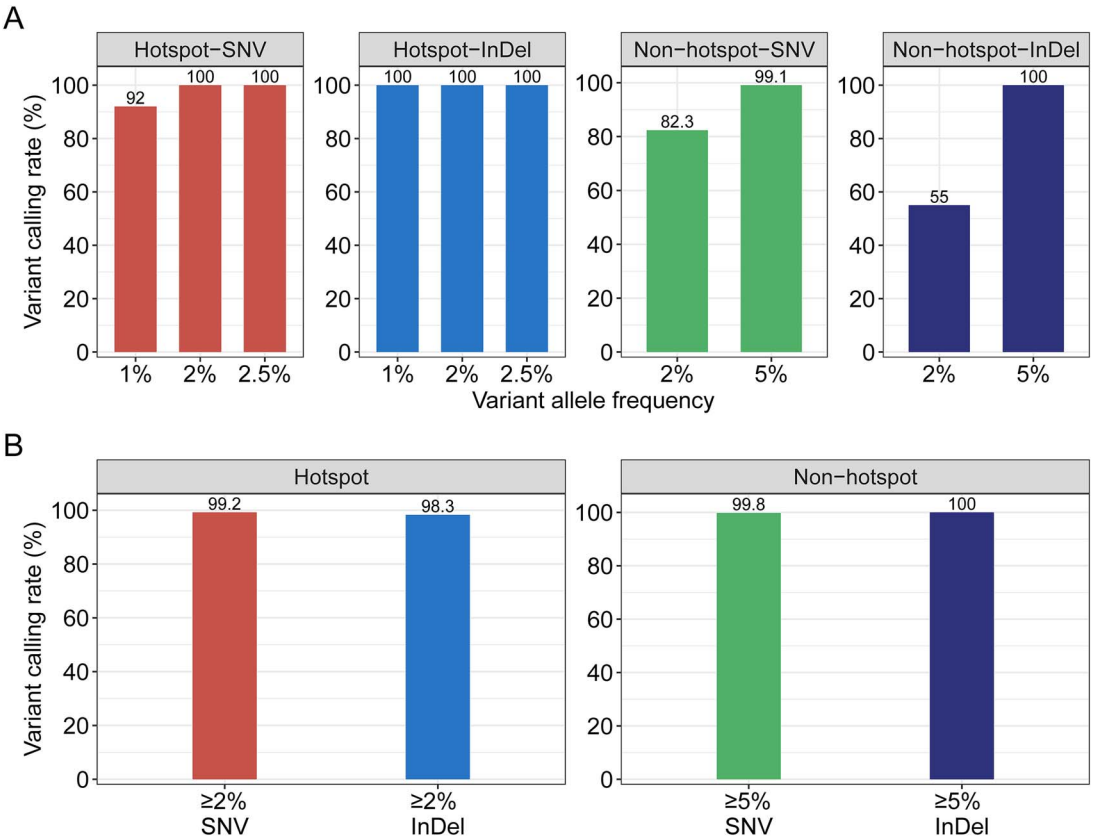

Figure 2. Estimation and validation of LoD values for VAF. (A) The estimation of LoD values for efficient variant calling based on contrived samples. The LoD value for hotspot SNVs and InDels is 2.00% VAF, while the value for nonhotspot SNVs and InDels is 5.00% VAF. (B) The validation result of the variant calling rate at the LoD values determined in (A). A total of 82 samples from 14 cancer types were employed in this analysis, as shown in Table S1. The comparison of OncoTOP versus OncoD is displayed. The X-axis shows the evaluated mutation type and VAF criteria, and the Y-axis shows the variant calling rate.

in distinguishing between germline and somatic mutations (Fig. S1A). We found that when using germP, PAD_count, and caseAF separately, the PPA for predicting germline mutations was 94.1%, 77.6%, and 58.3%, respectively, while the PPA for predicting somatic mutations was 85.6%, 93.9%, and 97.4%. These results clearly demonstrated that no single feature performs as well as the combined use of all three. Additionally, we provided the feature importance scores from the decision tree model (Fig. S1B), showing that the importance values for germP, PAD_count, and caseAF were 0.713, 0.241, and 0.046, respectively. It appears that

Table 1. Accuracy of OncoTOP for variant calling across 18 tumor lineages.

Tumor lineage	TP	FP	TN	FN	PPA	NPA	PPV	NPV	Acc
Breast cancer	9709	0	539	0	100.0%	100.0%	100.0%	100.0%	100.0%
Ovarian cancer	8049	0	25	0	100.0%	100.0%	100.0%	100.0%	100.0%
Head and neck cancer	3943	0	7	0	100.0%	100.0%	100.0%	100.0%	100.0%
Endometrial cancer	10,004	35	30	0	100.0%	46.2%	99.7%	100.0%	99.7%
Cervical cancer	3501	1	0	4	99.9%	0.0%	100.0%	0.0%	99.9%
Small cell lung cancer	4011	4	7	7	99.8%	63.6%	99.9%	50.0%	99.7%
Urothelial carcinoma	5467	2	16	9	99.8%	88.9%	100.0%	64.0%	99.8%
Diffuse large B-cell lymphoma	4148	7	2075	15	99.6%	99.7%	99.8%	99.3%	99.6%
Thyroid cancer	2286	0	7	1	100.0%	100.0%	100.0%	87.5%	100.0%
Colorectal cancer	24,171	1	272	0	100.0%	99.6%	100.0%	100.0%	100.0%
Liver cancer	3246	6	40	2	99.9%	87.0%	99.8%	95.2%	99.8%
Gastric cancer	3902	1	56	0	100.0%	98.2%	100.0%	100.0%	100.0%
Gastrointestinal stromal tumor	2816	1	6	8	99.7%	85.7%	100.0%	42.9%	99.7%
Biliary tract tumor	3139	1	33	0	100.0%	97.1%	100.0%	100.0%	100.0%
Pancreatic cancer	2621	0	8	6	99.8%	100.0%	100.0%	57.1%	99.8%
Nonsmall cell lung cancer	24,110	6	1669	208	99.1%	99.6%	100.0%	88.9%	99.2%
Esophageal cancer	3453	1	26	1	100.0%	96.3%	100.0%	96.3%	99.9%
Melanoma	2388	0	6	1	100.0%	100.0%	100.0%	85.7%	100.0%
Overall	120,964	66	4822	262	99.8%	98.6%	99.9%	94.8%	99.7%

Totally, 2864 samples across 18 cancer types were employed in the accuracy analysis of OncoTOP for variant calling. Method comparison: OncoTOP versus OncoD. Mutations detected by OncoD were regarded as reference, and validation for OncoTOP was done by evaluating alternations with a VAF no less than 1%. SNV, single-nucleotide variant; InDel, insertion and deletion; VAF, variant allele frequency; TP, true positive; FP, false positive; TN, true negative; FN, false negative; PPA, positive percent agreement; NPA, negative percent agreement; PPV, positive predictive value; NPV, negative predictive value; Acc, accuracy.

Table 2. Accuracy of detection of actionable or drug-resistant-related variants.

Cancer type	Gene	pHGVS	TP	FP	TN	FN	PPA	PPV	Acc
Breast cancer	PIK3CA	p.E542K	14	0	0	0	100.0%	100.0%	100.0%
Breast cancer	PIK3CA	p.E545K	20	0	0	0	100.0%	100.0%	100.0%
Breast cancer	PIK3CA	p.H1047R	46	0	0	0	100.0%	100.0%	100.0%
Breast cancer	PIK3CA	p.R88Q	10	0	0	0	100.0%	100.0%	100.0%
Thyroid cancer	BRAF	p.V600E	82	0	0	0	100.0%	100.0%	100.0%
Colorectal cancer	BRAF	p.V600E	14	0	0	0	100.0%	100.0%	100.0%
Melanoma	BRAF	p.V600E	12	0	0	0	100.0%	100.0%	100.0%
Nonsmall cell lung cancer	EGFR	19del	199	0	0	0	100.0%	100.0%	100.0%
Nonsmall cell lung cancer	EGFR	20ins	51	0	0	0	100.0%	100.0%	100.0%
Nonsmall cell lung cancer	EGFR	p.G719	16	0	0	0	100.0%	100.0%	100.0%
Nonsmall cell lung cancer	EGFR	p.L858R	168	0	1	0	100.0%	100.0%	100.0%
Nonsmall cell lung cancer	KRAS	p.G12C	19	0	0	0	100.0%	100.0%	100.0%
Nonsmall cell lung cancer	EGFR	p.T790M	40	0	0	0	100.0%	100.0%	100.0%

From the variant calling results of all the aforementioned 2864 samples, we extracted the corresponding detection outcomes to assess the accuracy of OncoTOP in detecting actionable or drug-resistant-related variants. TP, true positive; FP, false positive; TN, true negative; FN, false negative; PPA, positive percent agreement; NPA, negative percent agreement; PPV, positive predictive value; NPV, negative predictive value; Acc, accuracy.

germP, derived from the two-tailed binomial test comparing the observed minor allele frequencies (MAFs) of variants with their theoretical allele frequencies (THEOAFs), contributes the most to the model, followed by PAD_count, and then caseAF.

The concordance with somatic-germline-zygosity of FoundationOne CDx

We proceeded to evaluate the agreement between OncoTOP and SGZ (somatic-germline-zygosity), a method developed previously based on FoundationOne CDx for tumor-only sequencing data analysis [13]. FoundationOne CDx is a comprehensive genomic profiling assay targeting 324 genes, approved by the U.S. food and drug administration (FDA), and is widely used in clinical practice. We observed that the PPA of OncoTOP in predicting somatic variants (97.4%, 67 145/68 946) was higher than that of SGZ (95%, 312/327) (Table 3). Although the predicted germline mutations exhibited a slightly lower PPA compared to those reported by the SGZ paper [13], it is important to consider that the discrepancy may be influenced by the larger scale of our dataset. Our dataset comprises 120 964 mutations from over 2000 samples across 18 cancer types, whereas the validation dataset for SGZ consists of only 480 mutations from 30 lung and colorectal cancer samples.

To assess the accuracy of OncoTOP in detecting actionable variants, we further conducted tumor-only variant calling using both OncoTOP and FoundationOne CDx on an additional set of 22 samples. Our focus was primarily on actionable variants associated with targeted therapies. Among the 22 patients, OncoTOP identified 21 actionable variants (15 SNVs and 6 InDels), while FoundationOne CDx detected 22 actionable variants (15 SNVs and 7 InDels). Notably, all 21 variants detected by OncoTOP were concordant with FoundationOne CDx, resulting in an overall concordance of 95.4% (Table S7). For a broader range of clinically relevant variants, specifically class 1 and class 2 mutations as reported by FoundationOne CDx, the concordance between these two methods reached 98.7% (Table S7). These findings indicate that OncoTOP performs comparably to FoundationOne CDx in

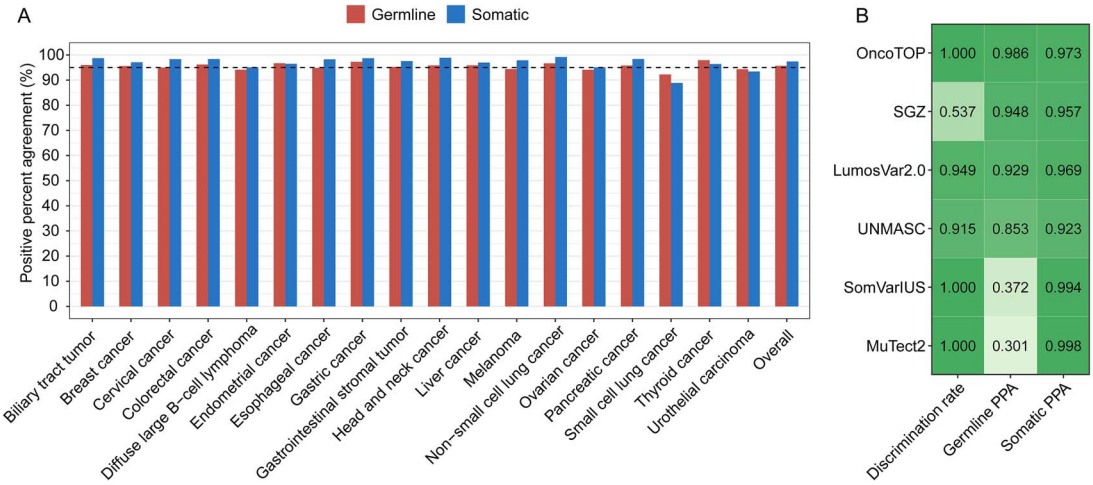

Figure 3. Accuracy of mutation status prediction across 18 tumor lineages. (A) OncoTOP shows continuously high accuracy of mutation status prediction among most of the 18 cancer types. $N = 2864$, see Table S4; method comparison: OncoTOP versus OncoD; red bars represent the accuracy of predicting germline mutations, and blue bars indicate that of predicting somatic mutations. The X-axis shows the cancer type, and the Y-axis displays the PPAs for predicting mutations as either somatic or germline. (B) Performance comparison for prediction of germline/somatic origins across OncoTOP and five other tumor-only methods. Discrimination rate: the proportion of detected mutations for which the algorithm can provide unambiguous germline or somatic prediction results; PPA: positive predictive agreement.

Table 3. Comparison of performance on variant calling and origin prediction between SGZ and OncoTOP.

Method	Sample cohort (tumor type)	Somatic variants PPA	Germline variants PPA
SGZ (FoundationOne CDx)	30 (lung & colon)	95% (312/327)	99% (151/153)
OncoTOP	2864 (Pan-cancer)	97.4% (67 145/68 946)	95.7% (49 768/52 018)

SGZ is a previously published computational method by FoundationOne CDx for analyzing tumor-only sequencing data. PPA, positive percent agreement; SGZ, somatic-germline-zygosity.

detecting clinically significant variants, providing reliable and accurate information for precision medicine applications.

Comparative analysis of OncoTOP and other tumor-only methods

To properly and fairly measure the performances of OncoTOP, we benchmarked tumor-only germline/somatic discrimination algorithms comparable to OncoTOP, including SGZ [13], LumosVar2.0 [22], UNMASC [23], SomVarIUS [24], and Mutect2 [25]. We have summarized the key features of these tumor-only methods in Table 4. Among these, Mutect2 and SomVarIUS, which rely solely on a panel of normals (PoN) and population databases like gnomAD to distinguish germline mutations, were considered as baseline methods. We randomly selected 247 samples from our training set, using tumor-normal matched results as the ground truth. Only mutations detected by each algorithm that overlapped with the true mutation sets were included to evaluate the performance of predicting germline or somatic origins. We used discrimination rate to indicate the proportion of detected mutations for which the algorithm can provide unambiguous germline or somatic prediction results, and PPA to evaluate the accuracy of the predictions against the ground truth (Fig. 3B).

OncoTOP, SomVarIUS, and Mutect2 were able to provide prediction results for all detected mutations, thus yielding a 100% discrimination rate. However, LumosVar2.0, UNMASC, and SGZ were unable to provide prediction results for some mutations, with discrimination rates of 94.9%, 91.5%, and 53.7%, respectively. Among the mutations with unambiguous predictions, OncoTOP demonstrated the best performance in differentiating germline and somatic variants, with a PPA of 98.6% for germline variants and 97.3% for somatic variants. The baseline methods, MuTect2 and SomVarIUS, relying only on population databases, were relatively poor in identifying germline mutations, with PPAs of 30.1% and 37.2%, respectively. However, for somatic mutations, high PPAs of 99.8% and 99.4% were observed. LumosVar2.0 and UNMASC performed moderately well, with PPA values of 92.9% and 85.3% for differentiating germline mutations and 96.9% and 92.3% for somatic mutations, respectively. Although SGZ had a lower discrimination rate, its ability to differentiate germline and somatic mutations had PPAs of 94.8% and 95.7%, consistent with the reported performance in the SGZ paper [13]. These results demonstrate that OncoTOP outperforms those other tumor-only methods we have benchmarked in distinguishing between germline and somatic mutations.

Table 4. A summary of representative tumor-only methods and their key features.

Method	CNV	UMN	pubSFDB	priSFDB
OncoTOP	✓	✓	✓	✓
SGZ	✓		✓	
LumosVar2.0	✓	✓	✓	
UNMASC	✓	✓	✓	
SomVarIUS			✓	
MuTect2			✓	

This table shows the key features of the six benchmarked tumor-only methods. Features include regional copy number variations (CNVs), utilization of unmatched normals (UMN), reliance on public population databases for variant sample frequencies (pubPDB), and private population databases for variant sample frequencies (priPDB).

Measurement of tumor mutational burden

Next, we determined the LoD for tumor purity in measuring TMB. Initially, four standard samples with 40% tumor purity,

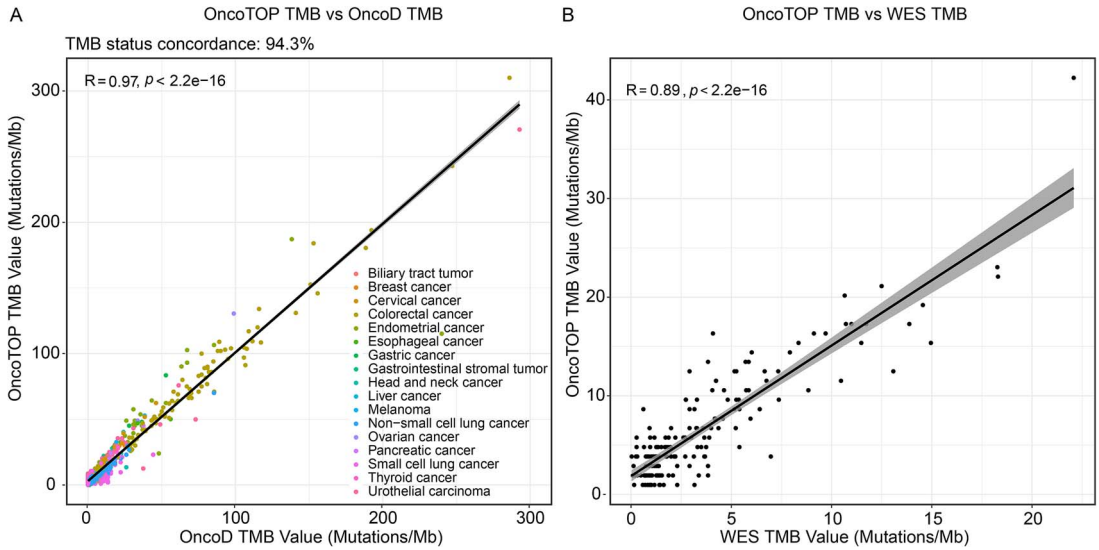

Figure 4. Comparison of TMB results yielded by OncoTOP and other methods. (A) Comparison of TMB values between OncoTOP and OncoD. A cohort consisting of 1813 tumor specimens across 17 cancer types was employed for validation. Each point represents a sample analyzed by both methods, with color-coding based on cancer type. The *X*-axis shows the TMB value (mutations/Mb) determined by OncoD, while the *Y*-axis shows the TMB value (mutations/Mb) determined by OncoTOP. The TMB status concordance is 94.3%. Pearson correlation coefficient method is applied for TMB value comparison: $R = 0.97$ with 95% CI [0.97, 0.98], *P*-value <.001, indicating a statistically significant correlation between TMB values generated by OncoTOP and OncoD. (B) Comparison of TMB values between OncoTOP and WES. $N = 167$. The *X*-axis shows the TMB value (mutations/Mb) determined by WES, while the *Y*-axis shows the TMB value (mutations/Mb) determined by OncoTOP. Pearson correlation coefficient: $R = 0.89$ with 95% CI [0.85, 0.92], *P*-value <.001, thus statistically significant.

representing various tumor lineages, including melanoma, NSCLC, bladder cancer, and colorectal cancer, were subjected to whole-exome sequencing (WES) to establish their TMB status as reference (Table S8). All samples were then analyzed by OncoTOP, and the TMB cut-off value (i.e. the upper quartile) was determined as 9 for TMB-A/B/C and 20 for TMB-D. The results showed that at tumor purities of 5%, 10%, 20%, and 40%, the corresponding concordance rates for TMB were 0% (0/4), 75% (3/4), 100% (4/4), and 100% (4/4), respectively. These findings suggest that the LoD for tumor purity in detecting TMB is 20% (Table S9).

We then evaluated the accuracy of TMB estimation by comparing the results obtained by OncoTOP with those from OncoD and WES ($N = 1813$ and $N = 167$, respectively). High concordance was observed in both analyses (Fig. 4, Pearson correlation coefficient, both $P < .001$; Fig. 4A, concordance = 94.3%, $R = 0.97$, 95% CI: [0.97, 0.98] when compared to OncoD; Fig. 4B, $R = 0.89$, 95% CI: [0.85, 0.92] when compared to WES). In addition, 12 out of 17 tumor lineages showed a TMB status concordance rate of higher than 90% when evaluating the accuracy of determining TMB status across all 1813 samples (Table S10). The precision of TMB measurement remained consistent both within and between runs, with a 100% concordance in TMB status prediction (Table S11).

To confirm the predictive value of TMB results generated by OncoTOP in determining treatment response, we performed survival analysis on an additional cohort of 97 lung cancer patients who were treated with immune checkpoint inhibitors. As anticipated, the TMB-low subgroup was significantly associated with poorer progression-free survival outcomes compared to the TMB-high subgroup (Fig. S2; 'log-rank sum test', HR = 0.58, $P = .02$), indicating a less efficient response to treatment in patients with low TMB levels. This finding aligns with previous reports [26], further confirming the reliability of OncoTOP in determining TMB levels and its ability to predict treatment response.

Evaluation of microsatellite instability

We analyzed 267 cancer samples with a tumor purity of at least 10% (191 colorectal cancers and 76 endometrial cancers) and compared their MSI status determined by OncoTOP with the gold-standard method of polymerase chain reaction (PCR). As illustrated in Fig. 5, the accuracy of OncoTOP in colorectal cancer was exceptionally high, reaching 99.9% (MSI-high sample: 100%, 139/139; microsatellite stable (MSS) sample: 96.2%, 50/52) (Fig. 5A). In the case of endometrial cancer, the accuracy rate was 92.1% (MSI-high sample: 90.3%, 56/62; MSS sample: 100%, 14/14) (Fig. 5B). Overall, irrespective of tumor type, OncoTOP exhibited an impressive overall accuracy of 97% (MSI-high sample: 97%, 195/201; MSS sample: 97%, 64/66) (Fig. 5C). Furthermore, when compared to the gold-standard results of six FFPE samples, we observed a high concordance in the predicted MSI status (100%) and consistently measured MSI scores, both within and between runs (Table S12). These results demonstrate the high accuracy and precision of OncoTOP in determining the MSI status of tumor specimens.

Subtyping of human leukocyte antigens

We further evaluated if OncoTOP could correctly determine the classical HLA class I (HLA-I) subtype of tumor specimens. By evaluating a total of 295 samples using OncoTOP and comparing the results with those generated by OncoD, we found high concordance in the classification of HLA-I subtypes and homozygosity (concordance of HLA-I subtypes: 99.3%; concordance of homozygosity: 99.9%). The performance of HLA predictions for each subtype is shown in Fig. 6. Our results illustrate that OncoTOP can be effectively applied for HLA-I subtyping and homozygosity determination of clinical tumor samples without a matched normal.

Discussion

In this study, we present OncoTOP, a reliable method for analyzing tumor-only targeted DNA sequencing data. This

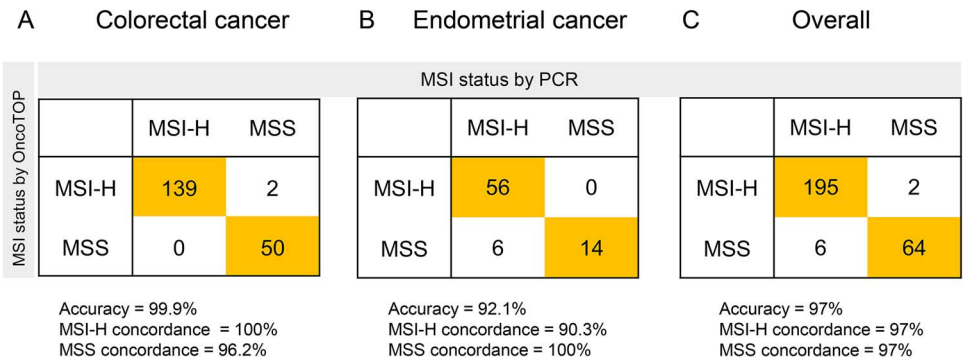

Figure 5. Validation on OncoTOP's accuracy for MSI status classification. Confusion matrix of predicted classification using OncoTOP-determined MSI results compared to actual classification determined by PCR. We utilized tumor samples of 191 colorectal cancers and 76 endometrial cancers with a tumor purity of at least 10%. PCR: polymerase chain reaction; MSS: microsatellite stable and MSI-H: microsatellite instability.

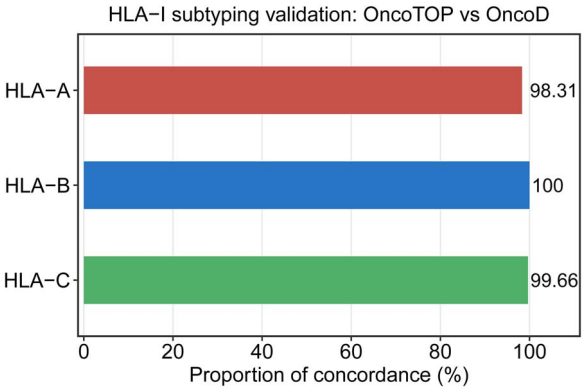

Figure 6. Concordance of HLA-I subtyping between OncoTOP and OncoD. The concordance of HLA-I subtyping between OncoTOP and OncoD was reviewed on HLA-A/B/C. The concordance was 98.31% for HLA-A, 100% for HLA-B, and 99.66% for HLA-C. $N = 295$.

innovative approach enables the identification of genomic alterations, prediction of their germline or somatic origins, and comprehensive evaluation of several clinically relevant biomarkers. Unlike previous tumor-only analysis methods that have been limited in performance validation to a few tumor lineages and specific cancer types [13, 14, 22, 27], OncoTOP was rigorously validated using a diverse set of pan-cancer clinical tumor samples and compared with WES, PCR, and its tumor-normal paired analysis mode OncoD. To the best of our knowledge, the cohort employed to validate the performance of OncoTOP represents the largest sample size and encompasses the most diverse range of cancer types compared to other validation cohorts for tumor-only analysis methods.

We comprehensively evaluated the analytical performance of OncoTOP for variant calling in 18 cancer types and across >1000 genes. The sensitivity of OncoTOP was demonstrated with an LoD of 2% VAF for hotspot SNVs and InDels and 5% VAF for nonhotspot SNVs and InDels, with a 98.4% variant detection rate at 10% tumor purity. Additionally, its false-positive variant calling rate was evaluated to be 0% for over 130 000 variants in normal cell samples. We further demonstrated its high reproducibility with 99.7% variant detection rate and <20% C.V. across 128 mutations. Its accuracy was validated using a large cohort of over 2800 samples across 18 cancer types, with an overall PPA of 99.8% and PPV of 99.9% compared to the tumor-normal paired analysis, and consistently good performance across tumor lineages. OncoTOP has also demonstrated high accuracy in detecting clinically actionable variants, such as *BRAF* p.V600E, as well as resistance-associated subclonal mutations like *EGFR* p.T790M. Moreover, its performance in detecting clinically significant variants has been comparable to that of FoundationOne CDx. This capability is particularly valuable for making clinical decisions regarding targeted therapies in situations where matched normal samples are unavailable. Together, these results demonstrate the robust performance and high reliability of OncoTOP in variant calling.

For variant calling, we employed both our in-house developed realDcaller2 and the widely utilized Mutect2, primarily for the following reasons. First, in addition to filtering low-quality bases like Mutect2 (e.g. low sequencing quality, poor alignment quality, bases near read ends), realDcaller2 employs an empirical blacklist to filter out unreliable mutations specific to our 1021 gene panel experimental system. Second, using unique molecular identifiers (UMIs) to remove duplicates preserves more useful reads and enhances the accuracy in detecting low-frequency mutations. In the preprocessing step of OncoTOP, UMIs are utilized to eliminate PCR duplicates. The realDcaller2 software can leverage information from UMI deduplication to characterize the template features of mutation-supporting reads (such as reads supporting mutations on both forward and reverse strands, i.e. duplex reads, and reads supporting mutations on a single strand, i.e. single-strand reads). Incorporating these template features is beneficial for accurately identifying false-positive mutations, whereas MuTect2 does not recognize UMI-related tags. Third, because Mutect2 uses local assembly and realignment to detect mutations, it can identify longer (>10 bp) InDels. In contrast, realDcaller2 relies on alignment alone, limiting its ability to detect long InDels. Therefore, we use both realDcaller2 and Mutect2 to call mutations, improving our detection of long InDels. Finally, since Mutect2 employs assembly algorithms and estimates mutation VAF through a statistical model, this process is somewhat opaque. In contrast, realDcaller2 calculates VAF through direct counting, offering excellent interpretability, which is crucial in clinical applications.

Categorizing genetic variants as germline or somatic origin is an essential step in identifying novel oncogenic targets [13, 14]. However, this has been largely limited in clinical practice due to the absence of matched normal sequencing data. Here, we demonstrated the robust performance of OncoTOP in predicting the origins of genetic variants. By comparing with the results generated by tumor-normal paired analysis, the overall PPA was 97.4% for predicted somatic mutations and 95.7% for predicted germline mutations. It also demonstrated robust performance across different types of cancers, with a PPA of over 95% for

predicting germline and somatic mutations in most of the 18 cancer types. Feature importance analysis indicates that germP and PAD_count are pivotal features enabling OncoTOP's accurate discrimination between somatic and germline mutations.

Through benchmarking against five other tumor-only methods, we further demonstrated that OncoTOP outperforms existing alternatives in distinguishing between germline and somatic mutations. As one of the most commonly utilized variant callers, Mutect2's robustness has been extensively validated. Its public accessibility has significantly facilitated broader research applications, benefiting a wide range of research communities. However, its tumor-only mode has been reported to generate a significant number of false-positive mutation sites when identifying somatic mutations [28], a common pitfall of methods solely reliant on filtering germline mutations based on population databases [29]. Our benchmark results also confirm this issue, with Mutect2 and SomVarIUS showing PPA of only 30.1% and 37.2%, respectively, in identifying germline mutations, significantly lower than OncoTOP. False-positive somatic mutations can lead to improper guidance for cancer-targeted therapies, posing risks to patient safety and increasing healthcare costs. Another commonly used method, SGZ, was developed previously based on FoundationOne CDx for tumor-only sequencing data analysis. The SGZ method employs a copy number model inferred through circular binary segmentation (CBS) and likelihood-based purity fitting to distinguish between somatic and germline mutations. Its ability to differentiate between somatic and germline mutations is heavily reliant on the copy number model. Even minor misfits in the model can result in an elevated rate of no calls, rendering it unable to make predictions. Major misfits in the model can further lead to misclassification of somatic versus germline mutations. However, this pivotal copy number model is proprietary, with its scripts lacking public accessibility. Therefore, during benchmarking analysis, we resorted to using the alternative algorithm allele-specific copy number analysis of tumors (ASCAT) [30], as recommended by SGZ, to infer allele-specific copy number variation (ASCNV). ASCAT utilizes the ASPCF segmentation algorithm and relies on the quantity and distribution of SNP in probes. In regions lacking heterozygous mutations, ASCNV results cannot be provided, impeding the differentiation of somatic and germline mutations and contributing to a lower discrimination rate (53.7%) in our benchmarking analysis. Furthermore, the inherent algorithmic logic of SGZ itself contributes to its relatively low discrimination rate, especially in scenarios where the differences in expected allele frequencies between germline and somatic mutations are minimal, making predictions challenging. Consequently, the inaccessibility of the copy number model, coupled with the low discrimination rate, raises the barrier for users and diminishes SGZ's performance in discerning the origin of mutations, thereby limiting its utility in tumor-only genomic research.

Several genomic features, such as TMB, MSI, and HLA-I subtypes, have emerged as potential biomarkers for predicting response to immunotherapy and guiding therapeutic decisions. However, prior to analyzing these biomarkers, sequencing of matched normal samples is critical for the existing approaches [6, 7]. Addressing this obstacle would be of great benefit to treatment selection for patients whenever matched normal samples are unavailable. OncoTOP provides highly accurate and precise measurement in TMB, which was demonstrated with (i) 94.3% concordance when compared to OncoD; (ii) a correlation coefficient of 0.89 compared to WES; (iii) over 90% concordance in evaluating TMB status across various tumor lineages; and (iv) 100% concordance in classifying TMB status within and between runs. Furthermore, by performing survival analysis based on TMB results yielded by OncoTOP, we also found poorer response to immune checkpoint inhibitors in lung cancer patients with low TMB compared to those with high TMB, proving that TMB results generated by OncoTOP can be inferred for predicting treatment response. Similarly, sufficient performance was observed when evaluating MSI status, with 97% concordance with PCR-validated results and 100% concordance within and between runs. When predicting HLA-I subtypes, although the underlying logic of OncoTOP differs from that of tumor-normal paired analysis, the generated results were highly concordant (>99% concordance). Collectively, these results demonstrate the satisfactory performance of OncoTOP in measuring genomic features and its potential to be applied in clinical practice.

However, OncoTOP still has some limitations that need to be addressed. First, despite the promising performance of OncoTOP, there still remains a risk of misclassification between germline and somatic variations. In certain cancer types, such as small cell lung cancer, the classification performance still needs to be further improved. Second, HLA-LOH estimation is solely based on tumor sequencing data, which can lead to inaccurate predictions of HLA-I subtypes in cases where loss of heterozygosity has occurred in the tumor.

Conclusion

OncoTOP has demonstrated sufficient performance in detecting variants, predicting mutation origin, and estimating three clinically significant biomarkers. The application of OncoTOP may shed light on analyzing tumor samples and make inferences when a matching normal sample is unavailable. It may also support clinical decision-making and aid in the discovery of novel oncogenic targets for cancer treatment.

Materials and Methods

Overview of OncoTOP

Identification of genomic alterations

OncoTOP was developed based on a comprehensive genomic profiling (CGP) assay that encompasses 1021 frequently mutated genes associated with cancer, covering a genomic region of ~1.6 megabases (MB), and the overall workflow is depicted in Fig. 1. Single-nucleotide variants (SNVs) were detected using realDcaller2 (v1.1.3 Geneplus-Beijing, in-house) specifically optimized for ultra-low frequency mutation calling and Mutect2 was used as an auxiliary tool to improve the detection of longer insertions/deletions (InDels).

Determination of somatic/germline origins for variants

Other tumor-only methods, such as SGZ [13], typically estimate tumor purity and allele-specific copy number variation (ASCNV) by utilizing logR and MAF, followed by distinguishing between germline and somatic mutations. However, considering the limitations of ASCNV calculations, such as inaccurate estimation for tumor purity below 30% and heavy reliance on a uniform distribution of SNP counts. OncoTOP does not directly rely on tumor purity and ASCNV, thus avoiding the influence of accuracy and precision of the tumor purity model. The utilization of raw signals, namely, logR and MAF, renders our method more robust. The specific workflow is as follows: After filtering out low-quality variants, we employ the CBS algorithm to partition the genome into regions of uniform copy number. Subsequently, we calculate

the MAF for each SNP within these segments, using equation (1), where AF represents the allele frequency. Clustering based on MAF values is then performed to identify potential germline and somatic variants. For each segment, we also compute the weighted average of the MAFs of variants within the segment, which serves as the theoretical minor allele frequency (THEOAF) using the equation (2), where W_i is the weight of the variant i, which was defined as equation (3). In equation (3), α represents a parameter that we have set to 0.005 based on the training results of the model. ΔMAF denotes the difference between the MAF of the variant and the upper quartile of the MAF values within the variant cluster. Next, we perform a two-tailed binomial test, comparing the observed MAF of variants with their THEOAF, and record the resulting P-value of equation (4), where a is the observed mutant allele depth and d is the observed total allele depth. Our method not only circumvents the background noise caused by outlier genomic variations but also takes into consideration the impact of mutation sequencing depth.

$$\mathrm{MAF} = \min(\mathrm{AF}, 1 - \mathrm{AF}) \tag{1}$$

$$\mathrm{THEOAF} = \sum_{i=1}^{j} W_i \times \mathrm{MAF}_i \tag{2}$$

$$W_i = \frac{\mathrm{Depth}_i \times (\alpha + \Delta\mathrm{MAF}_i)^2}{\sum_{i=1}^{j} \mathrm{Depth}_i \times (\alpha + \Delta\mathrm{MAF}_i)^2} \tag{3}$$

$$P(y|G; \mathrm{THEOAF}) = B(a, d, \mathrm{THEOAF}) \tag{4}$$

Furthermore, we have developed a baseline population database (POT) containing 2000 tumor samples, which serves as a reference for filtering out background noise. For each mutation, OncoTOP calculates its maximum VAF and occurrence in the POT database, and the VAFs of mutations detected in the analyzed sample. Different stringent thresholds are applied for hotspot mutations (recommended by consensus guidelines, FDA-approved drug targets, and clinically validated resistance-related mutations), semihotspot mutations (frequently observed inactivating mutations of tumor suppressor genes or mutations with lower drug target relevance), and nonhotspot mutations (class III variants with uncertain clinical significance) to minimize the impact of background noise. Finally, the retained mutations undergo subsequent prediction for their germline or somatic origins.

We developed a decision tree model to discriminate between germline and somatic mutations. The model incorporated three features: mutation frequency (caseAF), which represents the detected mutation frequency in the analyzed sample; mutation count in the population allele database (PAD_count), indicating the occurrence of the mutation in the PAD database, which was established with 89 767 cancer patients who underwent genetic testing in GenePlus Co., Ltd and includes mutations identified in the normal tissue samples; and germP obtained from the two-tailed binomial test comparing the observed MAF of variants with their THEOAF. The tumor-normal paired analysis results were used as the gold standard, and the three mutation features were used to build the decision tree model using the DecisionTreeClassifier library in sklearn.tree.

The performance of the model was evaluated using the positive predictive agreement (PPA) for predicting somatic and germline mutations. A 3-fold cross-validation approach was employed, where two-thirds of the mutations were used for training, while one-third was used for validation. Finally, the GridSearchCV functionality was employed to perform a grid search on hyperparameters such as max_depth, min_samples_leaf, and min_samples_split within the DecisionTreeClassifier method. This pruning process aimed to prevent overfitting and select the model with the highest performance score.

Validation of analytical performance

Limit of detection

The limit of detection (LoD) refers to the lowest signal level at which a substance can be detected with statistical significance [31]. In this study, the statistical significance was set at a 95% confidence level. The LoD was evaluated for both tumor purity and VAF to ensure an optimal variant detection rate and for tumor purity to accurately determine TMB status.

Limit of blank

The limit of blank (LoB) illustrates the highest quantity value that is likely to be observed with a stated probability (false-positive rate) in a blank sample [32]. In this study, we employed 49 cell samples without mutations and calculated the number of false-positive variants at 2015 hotspot mutation sites and 128 663 nonhotspot mutation sites. The variant calling rate was expected to be lower than the 1% cut-off value.

Precision analysis

The aim of precision analysis is to assess the repeatability within runs and reproducibility between runs of the results produced by OncoTOP. We evaluated the precision of OncoTOP in detecting variants, determining TMB status, and predicting MSI. When assessing precision for variant calling, the C.V. was calculated with equation (5).

$$\mathrm{C.V.} = (\text{Standard deviation}) / \text{mean} \tag{5}$$

Ideally, the consistency rate for detected mutations is 95% or higher, and the C.V. is no greater than 20%. We conducted precision evaluation using six FFPE samples with known mutation information (Table S5), which were repeated five times to evaluate repeatability and reproducibility.

Accuracy analysis

To conduct accuracy analysis, we compared OncoTOP with other methods such as WES, PCR, and OncoD, following previously published validation methods [31–33]. We evaluated the concordance for SNV and InDel calls, predicted mutation origin, and estimated TMB, MSI, and HLA subtypes, respectively. When evaluating performance in variant calling, we calculated positive percent agreement (PPA), negative percent agreement (NPA), positive predictive value (PPV), negative predictive value (NPV), and accuracy (Acc) using equations (6–10) as defined in the literature reports [31–33].

$$\mathrm{PPA} = \frac{\mathrm{Count}_{\text{True positive}}}{\mathrm{Count}_{\text{True positive}} + \mathrm{Count}_{\text{False negative}}} \tag{6}$$

$$\mathrm{NPA} = \frac{\mathrm{Count}_{\text{True negative}}}{\mathrm{Count}_{\text{True negative}} + \mathrm{Count}_{\text{False positive}}} \tag{7}$$

$$\mathrm{PPV} = \frac{\mathrm{Count}_{\text{True positive}}}{\mathrm{Count}_{\text{True positive}} + \mathrm{Count}_{\text{False positive}}} \tag{8}$$

$$\mathrm{NPV} = \frac{\mathrm{Count}_{\text{True negative}}}{\mathrm{Count}_{\text{True negative}} + \mathrm{Count}_{\text{False negative}}} \tag{9}$$

$$Acc = \frac{Count_{True\ positive} + Count_{True\ negative}}{Count_{True\ positive} + Count_{True\ negative} + Count_{False\ positive} + Count_{False\ negative}} \quad (10)$$

Survival analysis

Survival analysis was conducted using the R package survival version 3.4.0 (https://cran.r-project.org/web/packages/survival/index.html, last accessed November 17, 2022), and R version 4.1.2 under the RStudio environment (https://www.r-project.org/, last accessed 17 November 2022). The analysis was based on the "survfit" function and visualized with the "ggsurvplot" function from the R package survminer version 0.4.9 (https://cran.r-project.org/web/packages/survminer/index.html, last accessed 17 November 2022).

Evaluation of clinically important biomarkers

TMB has become a promising biomarker and can be used to stratify patients who could benefit from immune checkpoint inhibitors [9, 34]. To measure the TMB value, the total number of somatic mutations within coding regions and with a VAF no less than 5% is counted, after excluding driver gene mutations that may cause the bias of dataset. Then, TMB is measured as the total number of mutations divided by the length of coding regions covered by our gene panel (1.114 MB) and reported in units of mutations per megabase (mutations/MB). TMB status is classified as either TMB-high or TMB-low based on cut-off values established for different tumor lineages (summarized in Table S13).

MSI serves as an indicator of the replication error phenotype that is caused by the dysfunction in DNA mismatch repair processes. It potentially enables the stratification of patients who may benefit from chemotherapies and immunotherapies [35]. To determine MSI status, a baseline dataset was established based on PCR-validated MSS- and MSI-high samples that met quality control criteria. Loci that are highly sensitive for characterization were selected for MSI status classification of one sample. MSI statistical magnitude is calculated as the product of proportion and entropy of loci not covered by the reference genome, and the MSI score for each targeted locus is measured by normalizing Z-score values with the application of MSS statistical magnitude distribution. Then, MSI scores are weighted and averaged to get an overall score, and the MSI status is determined according to the number of effective loci and the threshold of MSI scores. The MSI status is defined as MSI-high if the MSI score of one sample is no less than 0.135, a cut-off value previously determined.

Loss of heterozygosity (LOH) at the HLA-I locus is commonly recognized as an indicator of poor response to immune checkpoint inhibitors and immune evasion [36, 37]. When genotyping HLA loci for a paired tumor-normal analysis, the HLA subtype of the normal sample is first obtained to detect LOH in HLA genes within the tumor. However, in the case of tumor-only analysis, the HLA subtype of the control sample is not available. OncoTOP offers a direct evaluation of the HLA subtype within tumors. This is achieved by aligning the sequenced reads to known HLA allele sequences and obtaining subtyping results based on the unique alignment of reads to specific HLA alleles. And the obtained subtyping results allow us for the evaluation of homozygosity at the HLA locus.

Samples used for validation

The LoD was measured by analyzing several contrived samples, and validation was performed using a cohort of 82 previously stored tissue samples across different tumor lineages. The LoB was evaluated by employing 49 normal cell samples from patients. The accuracy of OncoTOP for variant calling and somatic or germline origin prediction was validated using 2864 FFPE samples collected by GenePlus Co., Ltd. A total of 18 types of cancer tissues were involved in the validation. OncoD, which is our previously developed tumor-normal paired analysis method widely applied in cancer research [17, 38–40], was employed for these samples and their matched normal samples for method comparison. Of this cohort, 1813 tumor specimens across 17 tumor lineages were further employed for TMB validation. A further comparison was made using a cohort of 167 clinical samples with known TMB results inferred from WES. The concordance of OncoTOP and SGZ of FoundationOne CDx was confirmed by an additional set of 22 tumor samples. The accuracy of HLA and MSI estimation was validated by analyzing additional 295 and 267 samples, respectively. Six FFPE samples were used as referenced standard for evaluating the intrarun and inter-run reproducibility and concordance of called variants, MSI/TMB values, and statuses. Ninety-seven tumor samples with both treatment and PFS information were employed for a survival analysis to demonstrate the reliability of the TMB results generated by OncoTOP in predicting response to treatment.

Key Points

- We have presented a computational method named OncoTOP that enables tumor genomic analysis when matched normal samples are unavailable.
- Analyses of 2864 samples across 18 cancer types showed an overall positive percent agreement of 99.8% and a positive predictive value of 99.9% for tumor-only variant calling with OncoTOP.
- OncoTOP can accurately detect clinically actionable variants and subclonal mutations associated with drug resistance.
- OncoTOP has an overall accuracy of 97.4% for predicting somatic mutations and 95.7% for predicting germline mutations.
- OncoTOP can be used to accurately estimate clinically important biomarkers including TMB, MSI, and HLA subtypes.

Data availability

The original code of OncoTOP and raw DNA sequencing data are available from the corresponding authors upon formal and reasonable request.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Funding

This study was supported by the Science and Technology Planning Project of Jilin Province [grant numbers 20210303002SF, 20210204031YY, YDZJ202202CXJD009].

Author contributions

Y.C. and X.Y. supervised the study. H.L. and L.M. analyzed patient data. H.K.W. and L.C. drafted the paper. P.Y.Z., S.Y., Y.L., and S.W.L.

prepared the research materials. Y.X. and S.C.F. performed benchmarking. H.S., S.H., X.H.D., Y.Z., H.F., T. L., and Y.F.G. provided insight into methodological approaches and analysis. J.B., Y.C., and X.Y. revised the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The participants provided their written informed consent in this study. The present study was approved by the Medical Ethics Committee of Jilin Cancer Hospital [202306-05-01].

Consent for publication

Not applicable.

References

1. The Oslo Breast Cancer Consortium (OSBREAC), Stephens PJ. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* 2012;**486**:400–4.
2. Kanchi KL, Johnson KJ, Lu C. *et al.* Integrated analysis of germline and somatic variants in ovarian cancer. *Nat Commun* 2014;**5**:3156.
3. Zehir A, Benayed R, Shah RH. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med* 2017;**23**:703–13.
4. Chalmers ZR, Connelly CF, Fabrizio D. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med* 2017;**9**:34.
5. Li A, Liu Y, Zhao Q. *et al.* Genome-wide identification of somatic aberrations from paired normal-tumor samples. *PloS One* 2014;**9**:e87212.
6. Larson DE, Harris CC, Chen K. *et al.* SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 2012;**28**:311–7.
7. Goldman MJ, Zhang J, Fonseca NA. *et al.* A user guide for the online exploration and visualization of PCAWG data. *Nat Commun* 2020;**11**:3400.
8. Chang K, Creighton CJ, Davis C. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20.
9. Hellmann MD, Callahan MK, Awad MM. *et al.* Tumor mutational burden and efficacy of Nivolumab monotherapy and in combination with Ipilimumab in small-cell lung cancer. *Cancer Cell* 2018;**33**:853–861.e854.
10. Carbone DP, Reck M, Paz-Ares L. *et al.* First-line Nivolumab in stage IV or recurrent non-small-cell lung cancer. *New England Journal of Medicine* 2017;**376**:2415–26.
11. le DT, Uram JN, Wang H. *et al.* PD-1 blockade in Tumors with mismatch-repair deficiency. *New England Journal of Medicine* 2015;**372**:2509–20.
12. Chowell D, Morris LGT, Grigg CM. *et al.* Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science* 2018;**359**:582–7.
13. Sun JX, He Y, Sanford E. *et al.* A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLoS Comput Biol* 2018;**14**:e1005965.
14. Oh S, Geistlinger L, Ramos M. *et al.* Reliable analysis of clinical tumor-only whole-exome sequencing data. *JCO Clinical Cancer Informatics* 2020;**4**:321–35.
15. Shi W, Ng CKY, Lim RS. *et al.* Reliability of whole-exome sequencing for assessing Intratumor genetic heterogeneity. *Cell Rep* 2018;**25**:1446–57.
16. Teer JK, Zhang Y, Chen L. *et al.* Evaluating somatic tumor mutation detection without matched normal samples. *Hum Genomics* 2017;**11**:1–13.
17. Lin G, Li C, Li PS. *et al.* Genomic origin and EGFR-TKI treatments of pulmonary adenosquamous carcinoma. *Ann Oncol* 2020;**31**:517–24.
18. Qiu Y, Xuan T, Yin M. *et al.* Clinical characteristics and genetic analysis of gene mutations in a Chinese pedigree with Peutz-Jeghers syndrome. *Clin Case Rep* 2019;**7**:735–9.
19. Luo X, Chang RZ, Kuang D. *et al.* Case report: Successful conversion and salvage resection of huge hepatocellular carcinoma with portal vein tumor thrombosis and intrahepatic metastasis via sequential hepatic arterial infusion chemotherapy, lenvatinib plus PD-1 antibody followed by simultaneous transcatheter arterial chemoembolization, and portal vein embolization. *Front Immunol* 2023;**14**:1285296.
20. Chakravarty D, Gao J, Phillips S. *et al.* OncoKB: A precision oncology Knowledge Base. *JCO Precis Oncol* 2017;**1**:1–16.
21. Yun C-H, Mengwasser KE, Toms AV. *et al.* The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc Natl Acad Sci* 2008;**105**:2070–5.
22. Halperin RF, Carpten JD, Manojlovic Z. *et al.* A method to reduce ancestry related germline false positives in tumor only somatic variant calling. *BMC Med Genomics* 2017;**10**:61.
23. Little P, Jo H, Hoyle A. *et al.* UNMASC: Tumor-only variant calling with unmatched normal controls, NAR. *Cancer* 2021;**3**:zcab040.
24. Smith KS, Yadav VK, Pei S. *et al.* SomVarIUS: Somatic variant identification from unpaired tissue samples. *Bioinformatics* 2015;**32**:808–13.
25. Benjamin D, Sato T, Cibulskis K. *et al.* Calling somatic SNVs and Indels with Mutect2. *bioRxiv* 2019;861054.
26. Ricciuti B, Wang X, Alessi JV. *et al.* Association of High Tumor Mutation Burden in non-small cell lung cancers with increased immune infiltration and improved clinical outcomes of PD-L1 blockade across PD-L1 expression levels. *JAMA Oncol* 2022;**8**:1160–8.
27. Kalatskaya I, Trinh QM, Spears M. *et al.* ISOWN: Accurate somatic mutation identification in the absence of normal tissue controls. *Genome Med* 2017;**9**:1–18.
28. Ha YJ, Kang S, Kim J. *et al.* Comprehensive benchmarking and guidelines of mosaic variant calling strategies. *Nat Methods* 2023;**20**:2058–67.
29. Jones S, Anagnostou V, Lytle K. *et al.* Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci Transl Med* 2015;**7**:283ra253.
30. van Loo P, Nordgard SH, Lingjærde OC. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci USA* 2010;**107**:16910–5.
31. Jennings LJ, Arcila ME, Corless C. *et al.* Guidelines for validation of next-generation sequencing-based oncology panels: A joint consensus recommendation of the Association for Molecular Pathology and College of American pathologists. *J Mol Diagn* 2017;**19**:341–65.
32. Woodhouse R, Li M, Hughes J. *et al.* Clinical and analytical validation of FoundationOne liquid CDx, a novel 324-gene cfDNA-based comprehensive genomic profiling assay for cancers of solid tumor origin. *PloS One* 2020;**15**:e0237802.
33. Lih CJ, Harrington RD, Sims DJ. *et al.* Analytical validation of the next-generation sequencing assay for a Nationwide signal-finding clinical trial: Molecular analysis for therapy choice clinical trial. *J Mol Diagn* 2017;**19**:313–27.
34. Klempner SJ, Fabrizio D, Bane S. *et al.* Tumor mutational burden as a predictive biomarker for response to immune checkpoint

inhibitors: A review of current evidence. *Oncologist* 2020;**25**: e147–59.
35. Copija A, Waniczek D, Witkoś A. *et al.* Clinical significance and prognostic relevance of microsatellite instability in sporadic colorectal cancer patients. *Int J Mol Sci* 2017;**18**:107.
36. Montesion M, Murugesan K, Jin DX. *et al.* Somatic HLA class I loss is a widespread mechanism of immune evasion which refines the use of tumor mutational burden as a biomarker of checkpoint inhibitor response. *Cancer Discov* 2021;**11**:282–92.
37. McGranahan N, Rosenthal R, Hiley CT. *et al.* Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* 2017;**171**:e1211.
38. Chen K, Bai J, Reuben A. *et al.* Multiomics analysis reveals distinct Immunogenomic features of lung cancer with ground-glass opacity. *Am J Respir Crit Care Med* 2021;**204**: 1180–92.
39. Hu C, Zhao L, Liu W. *et al.* Genomic profiles and their associations with TMB, PD-L1 expression, and immune cell infiltration landscapes in synchronous multiple primary lung cancers. *J Immunother Cancer* 2021;**9**: e003773.
40. Wang S, du M, Zhang J. *et al.* Tumor evolutionary trajectories during the acquisition of invasiveness in early stage lung adenocarcinoma. *Nat Commun* 2020;**11**:6083.

© The Author(s) 2024. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
For commercial re-use, please contact journals.permissions@oup.com
Briefings in Bioinformatics, 2025, **26(1)**, bbae677
https://doi.org/10.1093/bib/bbae677
Problem Solving Protocol