frontiers | Frontiers in Immunology

TYPE Hypothesis and Theory
PUBLISHED 25 May 2023
DOI 10.3389/fimmu.2023.1151224

OPEN ACCESS

EDITED BY
Jinghua Pan,
Jinan University, China

REVIEWED BY
Kun Chen,
University of Connecticut, United States
Hongmin Cai,
South China University of Technology, China
Meng Wu,
Harvard Medical School, United States

*CORRESPONDENCE
Jiayin Wang
✉ wangjiayin@mail.xjtu.edu.cn
Xuqi Li
✉ lixuqi@163.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 25 January 2023
ACCEPTED 15 May 2023
PUBLISHED 25 May 2023

CITATION
Liu Y, Wang S, Wang Y, Li Y, Zhu X, Lai X, Zhang X, Li X, Xiao X and Wang J (2023) What makes TMB an ambivalent biomarker for immunotherapy? A subtle mismatch between the sample-based design of variant callers and real clinical cohort. *Front. Immunol.* 14:1151224. doi: 10.3389/fimmu.2023.1151224

COPYRIGHT
© 2023 Liu, Wang, Wang, Li, Zhu, Lai, Zhang, Li, Xiao and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

What makes TMB an ambivalent biomarker for immunotherapy? A subtle mismatch between the sample-based design of variant callers and real clinical cohort

Yuqian Liu[1,2†], Shenjie Wang[1,2†], Yixuan Wang[3†], Yifei Li[1], Xiaoyan Zhu[1,2], Xin Lai[1,2], Xuanping Zhang[1,2], Xuqi Li[4*], Xiao Xiao[2,5] and Jiayin Wang[1,2*]

[1]School of Computer Science and Technology, Faculty of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China, [2]Shaanxi Engineering Research Center of Medical and Health Big Data, Xi'an Jiaotong University, Xi'an, Shaanxi, China, [3]Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China, [4]Department of General Surgery, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, China, [5]Geneplus Shenzhen, Shenzhen, China

Tumor mutation burden (TMB) is a widely recognized biomarker for predicting the efficacy of immunotherapy. However, its use still remains highly controversial. In this study, we examine the underlying causes of this controversy based on clinical needs. By tracing the source of the TMB errors and analyzing the design philosophy behind variant callers, we identify the conflict between the incompleteness of biostatistics rules and the variety of clinical samples as the critical issue that renders TMB an ambivalent biomarker. A series of experiments were conducted to illustrate the challenges of mutation detection in clinical practice. Additionally, we also discuss potential strategies for overcoming these conflict issues to enable the application of TMB in guiding decision-making in real clinical settings.

KEYWORDS

clinical immunology, tumor mutation burden, categorization thresholds, sequencing data analysis, bioinformatics tool, measurement error

1 Introduction

Immunotherapy has altered cancer treatment paradigms as a result of substantial advancements in immune checkpoint blocking (1–3). Increasing numbers of advanced cancer patients benefit from immune-checkpoint inhibitor (ICI) therapy (4). Tumor mutation burden (TMB) has been intensively studied as the promising immunotherapy biomarker for patient selection (5, 6). TMB refers to the number of somatic mutations per megabase (7). Clinical studies have noted that patients with high TMB tend to benefit more from immunotherapy (8). The association of high TMB with improved patient responses and survival benefits after immunotherapy has been observed in urothelial cancer (9), small cell lung cancer (10), non-small-cell lung cancer (11), among others. The US FDA has also prioritized TMB as the recommended test for cancer patients (12).

In clinical practice, it is only practical for physicians when TMB levels can effectively categorize patients into different risk groups with varying therapeutic benefits. However, here, TMB remains highly controversial. On one hand, TMB has been approved as a companion diagnostic biomarker, and multiple clinical trials have demonstrated its relevance to immunotherapy efficacy (13–15). Multiple studies presented at the 2020 ASCO meeting confirmed the predictive value of TMB in immunization or combination therapy, including KEYNOTE-061 study (16, 17), CONDOR study (18), EAGLE study (19), and EPOC1704 study (20), consolidating TMB as an independent predictor. On the other hand, several investigators have noted that TMB is not a perfect predictor of response to anti-PD-1/PD-L1 therapy, such as in KEYNOTE-158 study (21) and RATIONALE-304 study (22). Clinical studies with RCC (23–25), HPV-positive HNSCC (26), and melanoma receiving anti-PD-1 after recurrence (27) showed that TMB alone neither distinguished responders nor accurately predicted overall survival.

A popular opinion believes that this dispute is mainly caused by the inappropriate thresholds. The quantile-based cutoffs (e.g., median, quartiles) do not accurately reflect the underlying biology of TMB and fail to distinguish patients with their prospective clinical benefits (23, 24, 27, 28). Other conventional categorization methods, which establish a generic TMB threshold based on a single endpoint, reveal only partial therapeutic benefits. A single endpoint cannot fully represent the complexity and efficacy of a disease. Since different single endpoints were used, even on the same cohort of patients, the statistical studies gave inconsistent TMB thresholds, making it difficult for clinicians to make a decision (29). Moreover, the relationship between TMB and ICI benefits may not be uniformly distributed and may also differ across cancer types and corresponding regimens (30–33). Therefore, incorporating multiple efficacy endpoints into multiple categorizations of TMB for various cancers would be more effective in resolving the dispute (34).

Why does the argument still exist when TMB thresholds seem optimal? In data management, we often hear of the “trash in, trash out” principle. Thus, the imprecision of TMB measurements (23, 35, 36) is another crucial or even more dominant fact causing such controversy. Regardless of the various TMB calculation methodologies, none of the mutation callers claim to reach 100% accuracy. They each have their own unique advantages for mutation detection; thus, the errors in TMB measurement cannot be eliminated (37, 38). To avoid the trash-in-trash-out results, it is reasonable to consider TMB errors in threshold optimization, particularly for decision models. Based on the aforementioned multiple-endpoint framework, some study have proposed fault-tolerant statistical models (36) that reduce the instability and bias caused by TMB errors in patient categorization and resulting in improved performance. Although the mutation detection accuracy on each sample may be arbitrarily improved, regardless of the cost, by combining various sequencing technologies, deepening the sequencing depth, etc., it is still hard for the errors to meet the statistical correction assumption of the proposed models. Hence, merely introducing error control or fault tolerance into the decision model is insufficient. The critical error issues from bioinformatics tools that preclude the TMB from being employed in clinical use have not been addressed yet. We are trying to discuss how the issue of errors issue made TMB an ambivalent biomarker, and propose avenues for future research to resolve these tensions.

2 Discussion

2.1 What are the TMB measurement errors?

Traditionally, when evaluating a bioinformatics tool, researchers use the following performance metrics, including precision, recall, and F1-score, on the average of samples. The goal is to accurately detect mutations of target genes, with a focus on identifying the mutation commonalities among the genome data of patients and maintaining strict control over false positives(FPs), thereby avoiding giving the wrong medicine in targeted therapy. To ensure the detection of gene mutations, bioinformatics has developed numerous variant callers that are sensitive (39) and employs various filters to control FPs. It inevitably results in a large number of false negative(FN) errors while lowering the FP error rate in the final report (40).

In immunotherapy practice, the essence of TMB lies in the total number of mutations rather than a single or multiple targets of interest, regardless of the TMB calculation approaches used. FP and FN errors are equally important in TMB assessment and contribute to the aggregate TMB measurement errors. In TMB errors, the false-positive rate (FPR) is defined as the ratio of the number of FPs to the total number of mutations, whereas the false-negative rate (FNR) is defined as the ratio of the number of FNs to the total number of mutations. FPR and FNR may each obey a non-parametric distribution. They might be a layer-by-layer conditional probability that depends on the types and number of mutations in the sample, the mutation density and composition of the mutations in the particular segments, the design philosophy of the selected caller, the sampling quality, etc. Together, these two complex errors add up to a more complicated and unpredictable TMB measurement error.

2.2 What are the effects of complicated TMB errors on the threshold?

Existing TMB thresholds are typically obtained from retrospective investigations of specific immunotherapy patient cohorts. The disadvantage is that the optimized TMB thresholds are frequently less appropriate for broader patient populations, leading to limited generalizability results from sampling bias and measurement inaccuracy within the TMB. Generally, a particular cohort is a small group of patients sampled from a large population based on certain conditional criteria, such as cancer subtypes and enrollment requirements (41), resulting in substantial sampling bias. Due to sampling that violates the principle of randomization, patient cohorts in standard clinical trials are only partially representative of the distribution features of the entire population, resulting in TMB thresholds that are cohort-specific and scalable under extremely demanding conditions. In addition, the risk of measurement error carried by the TMB metric itself influences the transferability of the assigned threshold, even if the sampling population is regularly extended in clinical trials with the expectation that the analytic cohort would precisely characterize the entire distribution. TMB measurement errors can introduce bias in statistical inference (42), which in turn affects decision-making and hinders the effectiveness of therapeutic grouping effects. Here, we use the maximum likelihood estimation (MLE), which is the most popular in statistical inference, as an example to analyze the bias imposed by TMB measurement error on parameter estimation.

The MLE of a parameter θ is generally obtained by solving for the zero solution of a score function (the first-order derivative of the likelihood), i.e., $\Psi(\theta) = \frac{\partial \ell(\theta)}{\partial \theta^T} = 0$. The basic condition that guarantees the MLE is an unbiased estimator is the expectation unbiasedness of the score function $\Psi(\cdot)$. Nonetheless, if the TMB observations contain additive error components e, the expectation of the score function must be nonzero since the score function cannot be axisymmetric around the origin.

$$\boldsymbol{E}\left\{\Psi(TMB^{\star};\boldsymbol{\Theta})\right\} = \boldsymbol{E}\{\Psi(TMB + e;\boldsymbol{\Theta})\}$$
$$= \int_{-\infty}^{+\infty} \Psi(TMB + e;\boldsymbol{\Theta})p(e)de \neq 0 \qquad (1)$$

Further, if the error term e is assumed to obey a normal distribution with mean 0 and variance Σ_e. Z_i denotes a vector of covariates, e.g., age, gender, treatment indicator, cancer stage, we take Weibull–Cox PH model as an example, the instantaneous risk for an event depends on Z_i and TMB is defined as follow,

$$h(t|\boldsymbol{Z},\ TMB^{\star};\theta) = \lambda t^{\lambda-1}\exp(\beta_z^T\boldsymbol{Z} + \beta_m TMB^{\star}) \qquad (2)$$

Here, the expectation of the score function in Eq (1). can be expressed as

$$\boldsymbol{E}\left\{\Delta\boldsymbol{Z} - T^{\lambda}\cdot\exp(\beta_z^T\boldsymbol{Z} + \beta_m TMB^{\star})\boldsymbol{Z}\right\}$$
$$= \Delta\boldsymbol{Z} - T^{\lambda}\cdot\boldsymbol{E}\left\{\exp(\beta_z^T\boldsymbol{Z} + \beta_m TMB^{\star} + \beta_m e)\boldsymbol{Z}\right\}$$
$$= \Delta\boldsymbol{Z} - T^{\lambda}\exp(\beta_z^T\boldsymbol{Z} + \beta_m TMB^{\star})\boldsymbol{Z}\cdot\boldsymbol{E}\{\exp(\beta_m e)\} \neq 0 \qquad (3)$$

where Δ is an event indicator, T denotes the observed event time (such as tumor relapses, progression, death, etc.). The additional term $\boldsymbol{E}\{exp(\beta_m e)\}$ on the scoring function is caused by the measurement error, leading the naïve estimator to be biased apparently. If the variance fluctuation Σ_e can be controlled to approximate zero, the expectation of the score function will converge to zero.

$$\boldsymbol{E}\{\exp(\beta_m e)\}$$
$$= \int_{-\infty}^{+\infty} \exp(\beta_m e)\frac{1}{\Sigma_e\sqrt{2\pi}}exp(-\frac{e^2}{2\Sigma_e^2})de = exp(-\frac{\Sigma_e^2\beta_m^2}{2}) \approx 1 \qquad (4)$$

$$\boldsymbol{E}\left\{\Delta\boldsymbol{Z} - T^{\lambda}\cdot\exp(\beta_z^T\boldsymbol{Z} + \beta_m TMB^{\star})\boldsymbol{Z}\right\}$$
$$\approx \boldsymbol{E}\left\{\Delta\boldsymbol{Z} - T^{\lambda}\cdot\exp(\beta_z^T\boldsymbol{Z} + \beta_m TMB)\boldsymbol{Z}\right\} \approx 0 \qquad (5)$$

However, existing approaches barely achieve the variance control. The presence of the unavoidable error term destroys the impartiality nature of the score expectation, resulting in a considerably biased naive MLE estimator, which further affects the downstream TMB threshold determination. The threshold thus obtained is difficult to apply to clinical practice or other historical cohort data.

Furthermore, the mathematical modeling of TMB measurement error is extremely complex. It is related to a number of factors mentioned before, which are interdependent. There exists a complex logical transfer that constitutes nonparametric probability distributions on a layer-by-layer basis. Describing TMB error as a simple Gaussian noise within the conventional decision-making model lacks mathematical rigor, and definitely causes significant confusion in decision-making.

2.3 Why is this issue amplified in cancer sequencing data?

Why does this error rate issue seem not to appear in previous sequencing data analysis, especially in a similar genomics problem named population-based data analysis? This is due to the fact that, i) TMB assessment needs to count the total number of detection results, while other application scenarios only need to detect mutations of interest. The switch of needs increases the impact of error rates; ii) the types of mutations in general population are very limited, hence the accuracy of mutation detection software is much higher than that in cancer patients. For example, complex indels only exist in cancer sequencing data. It is a unique form of somatic mutation in tumor samples rarely seen in normal samples; and iii) if the accuracy of a mutation calling tool is sufficiently high, it would be capable of handling the detection and counting tasks very well. For example, the detection accuracy of the 1000 Genomes can easily reach 95% and up. This slight error rate would not affect the counting task since mutations are almost all detected. However, when it comes to complex cancer sequencing data with much lower accuracy, the impact of error rates is further deepened in the counting task. Hence, due to these facts, the original error rate issue has been noticed in cancer sequencing data.

Meanwhile, the clinical need of immunotherapy lies in the ability of variant callers to provide the total mutation count with steady state error rates on a cohort to calculate a fair TMB value for categorization. The variance control of TMB measurement becomes the focus. Despite the factors we discussed in Section 2.1, we focused on the TMB errors from the calling analysis. Bioinformatics software is comparable to a ruler in that it measures the level of patients on a specific dimension related to their immunotherapy prognosis. Just like a ruler, the measurement region should be uniform for all patients. Specifically, the variant callers must have steady performance by maintaining a stable/constant FPR and FNR across patients, as errors are inevitable. In that case, physicians will be able to categorize TMB as a baseline to separate patients into distinct risk groups with varying therapeutic benefits for subsequent clinical decision-making.

Unfortunately, the existing variant callers are unable to ensure consistent performance across samples, thereby failing to provide a fair TMB for clinical usage. We simulated a data set with 10 samples in which 500 variants, including single-nucleotide variants (SNVs), insertions, and deletions, were randomly planted in a template derived from the reference genome (hg.19). Variant calling was performed using Samtools and Bcftools. As shown in Figure 1, the performance of the caller exhibited significant fluctuations in FPR and FNR values across different samples. The coefficient of variation for the FPR and FNR values was 87.90% and 58.61%, respectively, demonstrating that the performance of callers fluctuated significantly as the sample (e.g., the proportion of different variants) varied. Further details of the experiment are presented in Supplementary 1.

As compared to the targeted therapy, the differences in the design philosophy make sophisticated bioinformatics tools unable to provide results with steady error rates and minimize TMB errors, thus performing inadequately in immunotherapy guidance. Hence, error control in bioinformatics tools becomes particularly important when using TMB to identify individuals likely to benefit from ICI treatment in a reliable and reproducible manner.

2.4 Why does bioinformatics software perform unsteadily across samples?

Existing bioinformatics software detects mutations from sequencing data based on rules, which are the mapping relationships between features of the sequencing data (e.g., split reads, abnormal read pairs, sequencing depth, etc.) and outcomes (mutation types), as summarized in Figure 2. Taking the deletion variant as an example, in which the sample is missing a fragment relative to the reference genome, there are three types of features when compared to the reference genome: 1) the read depth would be significantly reduced within this deletion region; 2) the insert size of the read pair, which is the spatial distance of the fragment generated by sequencing on both sides of the variant, would be significantly larger; and 3) a read in the sequenced fragment would be split into two fragments with the same alignment direction. Software sets the rules so that a region with these features would be reported as having a deletion variant. These rules are either summarized by researchers *via* experience and observation (43–46) or automatically learned by machine learning algorithms (47–50) based on commonalities among patients. The program reports the detection of a mutation in any genome region whose features fit the predefined rules. Therefore, the accuracy of detection in a certain region relies on the degree of matching between the preset rules of callers and the characteristics of the sequencing data. The mutation types in different samples may not differ greatly, but the proportion of each mutation type may vary significantly. The amount and proportion of mutations whose features do not match preset rules are also different across samples. Using the software with limited predetermined rules to analyze these samples will result in a significant variation in accuracy, as shown in Figure 1. The mismatch, caused by the variety of samples and the incompleteness of preset rules, is the fundamental reason for the

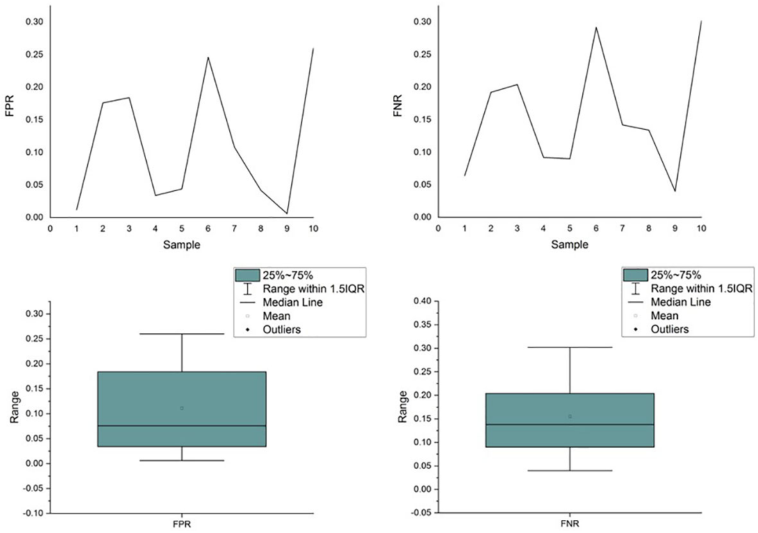

FIGURE 1
Performance comparison of the caller across simulated samples.

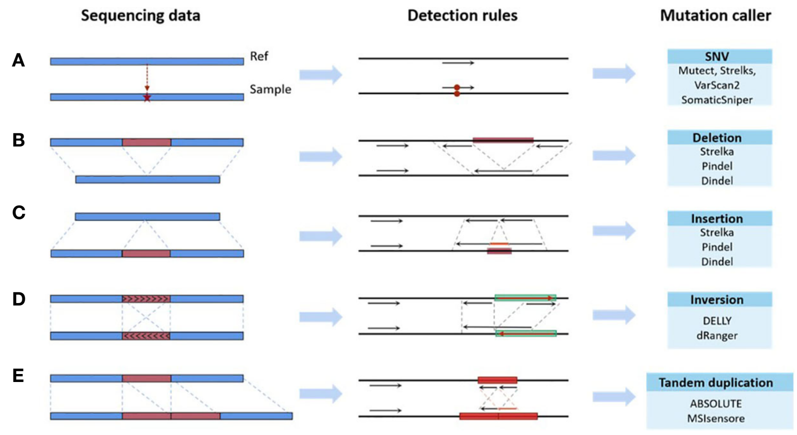

FIGURE 2
Mapping relationships between features of the sequencing data and mutation types. **(A)** SNV. **(B)** Deletion. **(C)** Insertion. **(D)** Inversion. **(E)** Tandem duplication.

fluctuation of error rates. It may also help explain why the performance of bioinformatics software differs significantly across populations and even races.

The mutation detection problem for a sample with multiple mutation types may be a non-deterministic polynomial-time hardness problem. That is, when a caller tries to combine all rules to cope with multiple mutation types, it is hard to find a solution to the problem in polynomial-time. As the number of mutation types increases, the number of rules may expand exponentially. Even with the help of machine learning, based on the probabilistic approximately correct (PAC) theory (51), it is only possible to establish an approximately correct set of rules to reduce the generalization error to an acceptable level. Thus, it is not feasible for a variant caller to establish a complete set of recognition rules that encompasses all mappings. Moreover, a general idea of the proof is given below.

Denote the reals by R, the accuracy control variable by ϵ, the confidence degree control variable by δ, the target concept by H, the possible hypothesis by h and the structural variation by SV. In Figure 3, falling within H indicates the reference SV set, while falling within h indicates the call SV set.

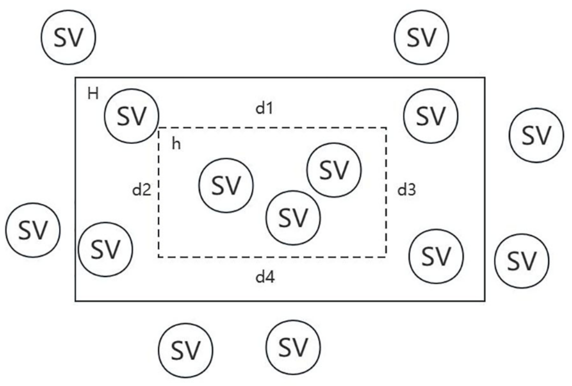

FIGURE 3
Schematic diagram of target concept and possible hypothesis.

For each $\epsilon \in R$, if $P[H] \leq \epsilon$, then $R[C_S] \leq P[H] \leq \epsilon$ is constant. If $P[H] > \epsilon$, Let's separate $d_1, d_2,\ d_3,\ d_4$ from H, and suppose that their areas are all $\frac{\epsilon}{4}$, then $S_{d_1+d_2+\ d_3+\ d_4} \leq \epsilon$. Thus, we can obtain the proposition that if $c = H_S$ intersects both $d_1, d_2,\ d_3,\ d_4$, then $H(c) \leq \epsilon$. Its converse proposition is that if $H(c)>\varepsilon$, then H_S does not intersect with at least one of $d_1, d_2,\ d_3,\ d_4$. Therefore, we can get

$$\begin{aligned}
&\mathbb{P}_{S \sim D^m}[H(H_S) > \epsilon] = \bigcap_{i=1}^{4}\left\{H_S \bigcap d_i = \phi\right\} \\
&\leq \bigcup_{i=1}^{4}\left\{H_S \bigcap d_i = \phi\right\} \\
&\leq \sum_{i=1}^{4}\mathbb{P}_{S \sim D^m}\left\{H_S \bigcap d_i = \phi\right\} \\
&\leq (1-\tfrac{\epsilon}{4})^m + (1-\tfrac{\epsilon}{4})^m + (1-\tfrac{\epsilon}{4})^m + (1-\tfrac{\epsilon}{4})^m \\
&= 4e^{\log(1-\frac{\epsilon}{4})^m} \\
&= 4e^{m\log(1-\frac{\epsilon}{4})} \\
&\leq 4e^{-\frac{\epsilon}{4}m} \\
&\leq \delta
\end{aligned} \tag{6}$$

In order to ensure that

$$Pd_{S \sim D^m}[H(C_S) \leq \epsilon] \geq 1-\delta \Leftrightarrow Pd_{S \sim D^m}[H(C_S) > \epsilon] \leq \delta \tag{7}$$

Then,

$$4e^{-\frac{\epsilon}{4}m} \leq \delta \Leftrightarrow m \geq \frac{4}{\epsilon}\log\frac{4}{\delta} \tag{8}$$

Hence, for each $\epsilon > 0$, $\delta > 0$, if

$$m \geq \frac{4}{\epsilon}\log\frac{4}{\delta} \tag{9}$$

Then,

$$Pd_{S \sim D^m}[H(C_S) \leq \epsilon] \geq 1-\delta \tag{10}$$

Thus, this concept class is PAC-learnable. Because of the correlation between some features of the sequencing data (e.g., sequencing depth), an SV can be expressed as an r-term DNF. Applying the result of Pitt and Valiant (52), that r-term DNF are not learnable using r-term DNF as hypotheses in polynomial time unless RP = NP, will complete the proof.

2.5 Why the ensemble strategies for bioinformatics software cannot solve this issue?

Currently, powerful toolkits often adopt ensemble strategies. Multiple mutation detection tools were ensembled, and the consensus voting strategy was used to determine the final detection output. Voting may help to improve the detection of specific candidate targets, hence reducing the risk of FPs. However, it may neglect the important true mutations found by the minority. For example, a delicately designed software detects a mutation that is ignored by all others, yet due to the voting principle, this true mutation is filtered out as a false-positive error, resulting in a false-negative error. We simulated a data set with 15 samples in which 500 variants, including SNVs, insertions and deletions, were randomly planted in a template obtained from the reference genome (hg.19). Three commonly used variant calling flows: samtools + bcftools, freebayes and GATK mutect2 were adopted for the variant calling. We calculated the positive and negative error rates of variant calling using ensemble strategy, as shown in Figure 4. Furthermore, we provided two FP and two FN error examples, respectively, caused by the ensemble strategy (Supplementary 2.3). The ensemble strategy led to non-negligible errors in both the positive and negative, and the error rate fluctuated significantly, as depicted in Figure 4. Through calculation in out experiment, the coefficient of variation of the positive and negative error rates of the ensemble strategy reached 42.79% and 30.86%, respectively, indicating that when the sample changed (e.g., the percentage of various variants changes), the ensemble strategy performance changed accordinssgly (More details are available in Supplementary 2). This is because, despite having hundreds of variant callers, their fundamental rules are limited. There are huge overlaps in the basic variant-calling components. In particular, some mutation sites in alleles with low frequency are more likely to be filtered by the voting strategy, hence increasing the risk of FNs. As previously noted, FNs and FPs are equally crucial for TMB. The ensemble strategies voting cannot, therefore, resolve this issue.

3 Potential solutions

3.1 Software recommendation with improved error variance control performance

Some empirical studies have compared the performance of various variant callers on some benchmarking datasets and demonstrated that most callers have obvious advantages in specific data. These advantages are attributed to the variant caller's own preset rules, which enable them to handle data with specific mutations. For example, in a benchmark experiment by Kosugi et al. (53), a total of 69 variant callers were tested on second-generation and third-generation sequencing data, both real and simulated. The study revealed that each caller exhibited distinct advantages for specific samples, and no single caller performed optimally across all samples.

As previously noted, based on the clinical need for immunotherapy, bioinformatics mutation detection software is required to take the "personality of samples" into consideration. Technically, the objective is to improve the matching degree between mutations in samples and detection rules. Thus, we can

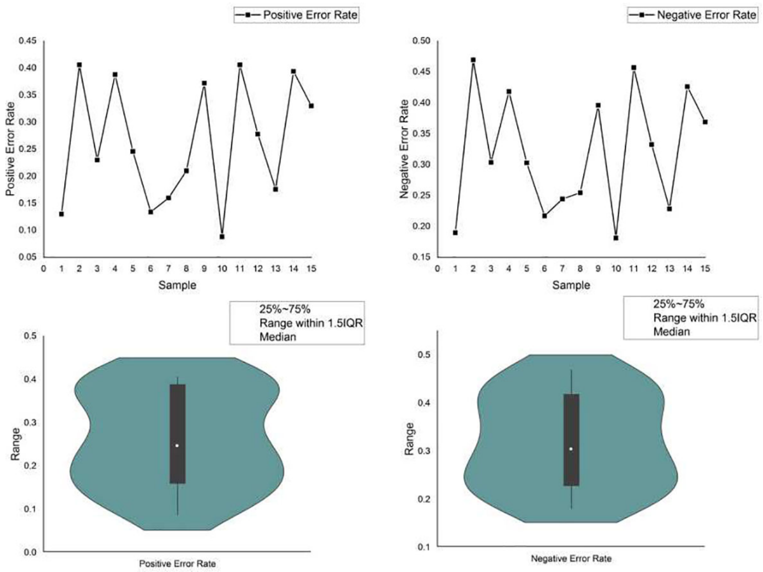

FIGURE 4
The positive and negative error rate of the ensemble strategy.

benefit from the idea of the recommender system, which provides suggestions for items that are most pertinent to a particular user. The most suitable bioinformatics software, determined by high matching degree for the samples to be analyzed, can be automatically recommended in a clinical setting. Thus, the recommended bioinformatics software ensures both effectiveness and efficiency in clinical practice, eliminating the need for additional prerequisites. Unlike the traditional recommender system in which the characteristics of users and items are apparent, the bioinformatics software recommender system suggests compatible software for samples whose characteristics are hidsssssden. In this context, we have to mine sample features and evaluate software performance in advance.

Specifically, in the design of traditional recommender systems, researcher guides the process of recommending an item to a user by establishing associations between items and the users. During this process, we can explicitly obtain the user's gender, age, educational background, and other characteristics to describe a user, as well as the item's size, color, material, and other attributes to describe an item. Motivated by this, we can achieve software recommendation by finding the characteristics that can describe the sequencing samples and software performance, and subsequently establishing the relationship between them. In this scenario, the user and item are the sample and software, respectively. It appears feasible to implement software recommendations in the same manner as traditional recommender systems. Unfortunately, it is more difficult to obtain the characteristics of a sequencing sample and software than it is to obtain the characteristics of an item or a user. We need to analyze sequencing samples to determine the characteristics that can distinguish among sample differences and software performance variations, such as read length and sequencing depth, among others. Similarly, we need to carry out a large number of experiments to test the performance of the software, in order to achieve a description of the software performance. Therefore, the user and item characteristics in the software recommendation scenario are hidden, which makes this problem challenging.

Some pioneering attempts have been made in this direction. Wang et al. (54) have presented an online SV caller recommendation tool implemented under the meta-learning framework, which automatically recommends the most compatible caller for the input sequencing data. There are some other recommender systems available, such as the SLP-based, ML-KNN-based, and collaborative filtering-based recommendation methods, among others (55). Moreover, the online caller recommendation tool documented in (54) is mainly used for SV caller recommendation. For SNV, indel, CNV, and other scenarios, meta features of the corresponding scenes can also be extracted to guide software recommendation. Therefore, we believe there will be more exciting software recommendation work, such as considering different application scenarios, personalized requirements and so on.

3.2 Developing novel bioinformatics tools with error control

Software recommendation is a good way to improve matching degree, however, its performance is still capped by the recommended candidates. Hence, it is still necessary to develop novel bioinformatics tools to achieve control over the error rate. From the perspective of error control, we believe there are several potential studies that warrant further investigation.

1) For some specific types of mutations, such as CNV or MSI, the mapping structure of the rules will not alter even if the variances across samples are substantial. By adjusting the parameters of the rules, it is possible to optimally match the features of different samples with the appropriate rules. Consequently, it is feasible to overcome this kind of problem by incorporating adaptation into the detection scheme. The parameters may self-adjust based on the features of the samples, thereby increasing the degree of matching in an adaptive manner. Some pioneer researches have made attempts. Taking CNV detection as an example, the size of sliding window is the crucial factor impacting the detection precision. Xuwen et al. (56) have presented a CNV caller with a dynamic sliding window that automatically adjust the window size based on the length of CNV to achieve the optimal setting. The adaptive window with self-adopted size makes it capable of handling CNVs with various lengths ranging from kb-scale to chromosome-arm level.

2) From the perspective of the control system, this problem can be viewed as a detection quality control problem. By setting the detection error as the adaptive control goal, it is feasible to introduce the adaptive control mechanism into the algorithm design to utilize the error feedback information. Hence, the detection algorithm would have the adaptability of dynamic adjustment and automatic matching of sequencing sample characteristics. However, there are two challenges that need to be tackled here. First of all, how do we establish the model for this detection quality control problem? Establishing accurate models of systems with nonlinear characteristics has been one of the most challenging and critical challenges in control science. Meanwhile, the mutation detection process is a special process of statistical analysis of static sequencing data and outputting results. The traditional modeling methods of physical systems cannot be directly applied in this context. Instead, it is necessary to study the key factors affecting quality control, such as the multi-dimensional sample characteristics and detection rule parameters and define the dynamic characteristics of the detection process. Secondly, how do we design a control strategy with adaptiveness and robustness to achieve the quality control of mutation detection? A complex nonlinear system with high uncertainties is one of the classical control problems in the field of nonlinear systems. There are numerous control approaches that have been studied and developed. However, the field currently lacks mathematical tools to uniformly deal with nonlinear systems, and a general optimal design solution is yet to be established, which needs to be analyzed and handled case-by-case. It needs to study how to introduce adaptive mechanisms into mutation detection methods, design control strategies based on the error feedback information and system model, and guide the design of adaptive detection algorithms.

3) Another one is to deal with multiple types of mutations simultaneously, each with its own set of rules. By adopting the concept of the recommender system in the selection of bioinformatics software, a sample may be viewed as a collection of different mutations with varying proportions and types. The sequencing data is divided into a series of intervals, with each

interval containing just one mutation. These intervals are then clustered. The best combination of bioinformatics software is selected by recommending the optimally matched rule sets for the clustered classes.

These studies may be further ensembled to address more complex cases, such as the overlap of different mutations within a single interval window. As an added value, this might shed light on why bioinformatics tools have inconsistent performance when detecting mutations in data from individuals of different races.

3.3 New threshold optimization methods to better consider the error

The threshold optimization method considering FN and FP errors is also a viable research topic for addressing the TMB issue. From the perspective of the errors, its fundamental concept is to correct the number of mutations detected within any specific interval of the samples in the statistical framework.

Thereby, the machine learning algorithm may be employed to learn the features of TMB that cannot be directly observed during detection, to predict the deviation risk of detected mutations, and to thereafter monitor the risk outliers among samples. Based on the predicted risk, the new threshold optimization methods would be able to assist in immune decision-making modeling, which will facilitate clinically precise diagnosis and therapy.

Compared to the existing multiple-end statistical models, which directly estimate the TMB error of each sample, this model uses machine learning to predict FPs and FNs instead, tracing the errors back one step farther. The information loss is reduced. Thus, the errors are more effectively handled across samples.

4 Conclusion

This article investigates the underlying reasons why TMB becomes an ambivalent biomarker. The definition of TMB error is given first. Then the requirements of immunotherapy for bioinformatics tools are analyzed to trace the source of the TMB error. The simulation results also demonstrate that the variant caller performs unsteadily across samples and cannot fulfill the requirements. The effects of TMB error on threshold and the reasons why error issue is amplified in cancer sequencing data are also discussed. By analyzing the design philosophy behind callers, the conflict between the incompleteness of biostatistics rules and the variety of clinical samples is the critical issue that renders TMB an ambivalent biomarker. Additionally, the article also proposes potential research topics in order to address the conflict issue.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.internationalgenome.org.

Author contributions

JW and YQL contributed to conception and design of the study; SW performed the experiments and analyzed the data; XYZ, XL, XPZ, XQL and XX helped perform the analysis with constructive discussions; YQL and JW wrote the first draft of the manuscript; YW, SW and YFL wrote some sections of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the National Natural Science Foundation of China, grant numbers 72293581, 72293580, 72274152. This work was also supported by the Natural Science Basic Research Program of Shaanxi, grant number 2020JC-01. The article processing charge was funded by the Natural Science Basic Research Program of Shaanxi, grant number 2020JC-01.

Acknowledgments

We thank all faculty members and graduate students who discussed the mathematical and statistical issues in seminars.

Conflict of interest

Author XX was employed by the company Geneplus Shenzhen Institute.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1151224/full#supplementary-material

References

1. Topalian SL, Drake CG, Pardoll DM. Immune checkpoint blockade: a common denominator approach to cancer therapy. *Cancer Cell* (2015) 27(4):450–61. doi: 10.1016/j.ccell.2015.03.001

2. Mellman I, Coukos G, Dranoff G. Cancer immunotherapy comes of age. *Nature* (2011) 480(7378):480–9. doi: 10.1038/nature10673

3. Majc B, Novak M, Kopitar-Jerala N, Jewett A, Breznik B. Immunotherapy of glioblastoma: current strategies and challenges in tumor model development. *Cells* (2021) 10(2):265. doi: 10.3390/cells10020265

4. Sharma P, Allison JP. Immune checkpoint targeting in cancer therapy: toward combination strategies with curative potential. *Cell* (2015) 161(2):205–14. doi: 10.1016/j.cell.2015.03.030

5. Hellmann MD, Ciuleanu T-E, Pluzanski A, Lee JS, Otterson GA, Audigier-Valette C, et al. Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden. *N Engl J Med* (2018) 378(22):2093–104. doi: 10.1056/NEJMoa1801946

6. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J, et al. Pan-tumor genomic biomarkers for pd-1 checkpoint blockade–based immunotherapy. *Science* (2018) 362(6411):eaar3593. doi: 10.1126/science.aar3593

7. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med* (2017) 9(1):1–14. doi: 10.1186/s13073-017-0424-2

8. Legrand FA, Gandara DR, Mariathasan S, Powles T, He X, Zhang W, et al. Association of high tissue tmb and atezolizumab efficacy across multiple tumor types. *Am Soc Clin Oncol* (2018) 36(15_suppl):15. doi: 10.1200/JCO.2018.36.15_suppl.12000

9. Deleuze A, Saout J, Dugay F, Peyronnet B, Mathieu R, Verhoest G, et al. Immunotherapy in renal cell carcinoma: the future is now. *Int J Mol Sci* (2020) 21 (7):2532. doi: 10.3390/ijms21072532

10. Hellmann MD, Callahan MK, Awad MM, Calvo E, Ascierto PA, Atmaca A, et al. Tumor mutational burden and efficacy of nivolumab monotherapy and in combination with ipilimumab in small-cell lung cancer. *Cancer Cell* (2018) 33(5):853–61.e4. doi: 10.1016/j.ccell.2018.04.001

11. Hellmann MD, Nathanson T, Rizvi H, Creelan BC, Sanchez-Vega F, Ahuja A, et al. Genomic features of response to combination immunotherapy in patients with advanced non-Small-Cell lung cancer. *Cancer Cell* (2018) 33(5):843–52.e4. doi: 10.1016/j.ccell.2018.03.018

12. Subbiah V, Solit D, Chan T, Kurzrock R. The fda approval of pembrolizumab for adult and pediatric patients with tumor mutational burden (Tmb)≥ 10: a decision centered on empowering patients and their physicians. *Ann Oncol* (2020) 31(9):1115–8. doi: 10.1016/j.annonc.2020.07.002

13. Carbone DP, Reck M, Paz-Ares L, Creelan B, Horn L, Steins M, et al. First-line nivolumab in stage iv or recurrent non–Small-Cell lung cancer. *New Engl J Med* (2017) 376(25):2415–26. doi: 10.1056/NEJMoa1613493

14. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Mutational landscape determines sensitivity to pd-1 blockade in non–small cell lung cancer. *Science* (2015) 348(6230):124–8. doi: 10.1126/science.aaa1348

15. Spigel DR, Schrock AB, Fabrizio D, Frampton GM, Sun J, He J, et al. Total mutation burden (Tmb) in lung cancer (Lc) and relationship with response to pd-1/Pd-L1 targeted therapies. *Am Soc Clin Oncol* (2016) 34:9017. doi: 10.1200/JCO.2016.34.15_suppl.9017

16. Fuchs CS, Özgüroğlu M, Bang Y-J, Di Bartolomeo M, Mandalà M, Ryu M-H, et al. The association of molecular biomarkers with efficacy of pembrolizumab versus paclitaxel in patients with gastric cancer (Gc) from keynote-061. *Am Soc Clin Oncol* (2020) 38:4512. doi: 10.1200/JCO.2020.38.15_suppl.4512

17. Shitara K, Özgüroğlu M, Bang Y-J, Di Bartolomeo M, Mandalà M, Ryu M-H, et al. The association of tissue tumor mutational burden (Ttmb) using the foundation medicine genomic platform with efficacy of pembrolizumab versus paclitaxel in patients (Pts) with gastric cancer (Gc) from keynote-061. *Am Soc Clin Oncol* (2020) 38:4537. doi: 10.1200/JCO.2020.38.15_suppl.4537

18. Li W, Matakidou A, Ghazoui Z, Si H, Wildsmith S, Morsli N, et al. Molecular biomarkers to identify patients (Pts) who may benefit from durvalumab (D; anti-Pd-L1) ± tremelimumab (T; anti-Ctla-4) in Recurrent/Metastatic head and neck squamous cell carcinoma (R/M hnscc) from hawk and condor studies. *Am Soc Clin Oncol* (2020) 38:6548. doi: 10.1200/JCO.2020.38.15_suppl.6548

19. Li W, Wildsmith S, Ye J, Si H, Morsli N, He P, et al. Plasma-based tumor mutational burden (Btmb) as predictor for survival in phase iii eagle study: durvalumab (D) ± tremelimumab (T) versus chemotherapy (Ct) in Recurrent/Metastatic head and neck squamous cell carcinoma (R/M hnscc) after platinum failure. *Am Soc Clin Oncol* (2020) 38:6511. doi: 10.1200/JCO.2020.38.15_suppl.6511

20. Kawazoe A, Yamamoto N, Kotani D, Kuboki Y, Taniguchi H, Harano K, et al. Tas-116, an oral Hsp90 inhibitor, in combination with nivolumab in patients with colorectal cancer and other solid tumors: an open-label, dose-finding, and expansion phase ib trial (Epoc1704). *Am Soc Clin Oncol* (2020) 38:4044. doi: 10.1200/JCO.2020.38.15_suppl.4044

21. Marabelle A, Le DT, Ascierto PA, Di Giacomo AM, De Jesus-Acosta A, Delord J-P, et al. Efficacy of pembrolizumab in patients with noncolorectal high microsatellite Instability/Mismatch repair–deficient cancer: results from the phase ii keynote-158 study. *J Clin Oncol* (2020) 38(1):1. doi: 10.1200/JCO.19.02105

22. Lu S, Sun M, Liu Y, Hu Y, Xie Y, Wang Z, et al. Abstract Lb512: rationale-304: the association of tumor mutational burden (Tmb) with clinical outcomes of tislelizumab (Tis)+ chemotherapy (Chemo) versus chemo alone as first-line treatment for advanced non-squamous non-small cell lung cancer (Nsq-nsclc). *Cancer Res* (2022) 82(12_Supplement):LB512–LB. doi: 10.1158/1538-7445.AM2022-LB512

23. Wood MA, Weeder BR, David JK, Nellore A, Thompson RF. Burden of tumor mutations, neoepitopes, and other variants are weak predictors of cancer immunotherapy response and overall survival. *Genome Med* (2020) 12(1):1–16. doi: 10.1186/s13073-020-00729-2

24. Colli LM, Machiela MJ, Myers TA, Jessop L, Yu K, Chanock SJ. Burden of nonsynonymous mutations among tcga cancers and candidate immune checkpoint inhibitor responsespossible checkpoint inhibitor response across tcga cancers. *Cancer Res* (2016) 76(13):3767–72. doi: 10.1158/0008-5472.CAN-16-0170

25. Miao D, Margolis CA, Gao W, Voss MH, Li W, Martini DJ, et al. Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* (2018) 359(6377):801–6. doi: 10.1126/science.aan5951

26. Hanna GJ, Lizotte P, Cavanaugh M, Kuo FC, Shivdasani P, Frieden A, et al. Frameshift events predict anti–Pd-1/L1 response in head and neck cancer. *JCI Insight* (2018) 3(4):3. doi: 10.1172/jci.insight.98811

27. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, et al. Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell* (2017) 171 (4):934–49.e16. doi: 10.1016/j.cell.2017.09.028

28. Miao D, Margolis CA, Vokes NI, Liu D, Taylor-Weiner A, Wankowicz SM, et al. Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. *Nat Genet* (2018) 50(9):1271–81. doi: 10.1038/s41588-018-0200-2

29. Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, et al. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancerstmb predicts response to immunotherapy in diverse cancers. *Mol Cancer Ther* (2017) 16(11):2598–608. doi: 10.1158/1535-7163.MCT-17-0386

30. Sheth M, Ko J. Exploring the relationship between overall survival (Os), progression free survival (Pfs) and objective response rate (Orr) in patients with advanced melanoma. *Cancer Treat Res Commun* (2021) 26:100272. doi: 10.1016/j.ctarc.2020.100272

31. Hashim M, Pfeiffer BM, Bartsch R, Postma M, Heeg B. Do surrogate endpoints better correlate with overall survival in studies that did not allow for crossover or reported balanced postprogression treatments? an application in advanced non–small cell lung cancer. *Value Health* (2018) 21(1):9–17. doi: 10.1016/j.jval.2017.07.011

32. Colloca GA, Venturino A, Guarneri D. Analysis of response-related endpoints in trials of first-line medical treatment of metastatic colorectal cancer. *Int J Clin Oncol* (2019) 24(11):1406–11. doi: 10.1007/s10147-019-01504-z

33. Yoshida Y, Kaneko M, Narukawa M. Magnitude of advantage in tumor response contributes to a better correlation between treatment effects on overall survival and progression-free survival: a literature-based meta-analysis of clinical trials in patients with metastatic colorectal cancer. *Int J Clin Oncol* (2020) 25(5):851–60. doi: 10.1007/s10147-020-01619-8

34. Ristl R, Urach S, Rosenkranz G, Posch M. Methods for the analysis of multiple endpoints in small populations: a review. *J biopharmaceutical Stat* (2019) 29(1):1–29. doi: 10.1080/10543406.2018.1489402

35. Samstein RM, Lee C-H, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet* (2019) 51(2):202–6. doi: 10.1038/s41588-018-0312-8

36. Wang Y, Lai X, Wang J, Xu Y, Zhang X, Zhu X, et al. Tmbcat: a multi-endpoint p-value criterion on different discrepancy metrics for superiorly inferring tumor mutation burden thresholds. *Front Immunol* (2022) 13. doi: 10.3389/fimmu.2022.995180

37. Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat Commun* (2015) 6(1):1–13. doi: 10.1038/ncomms10001

38. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* (2014) 15(1):1–10. doi: 10.1186/1471-2164-15-244

39. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput Struct Biotechnol J* (2018) 16:15–24. doi: 10.1016/j.csbj.2018.01.003

40. Kim Y-H, Song Y, Kim J-K, Kim T-M, Sim HW, Kim H-L, et al. False-negative errors in next-generation sequencing contribute substantially to inconsistency of mutation databases. *PloS One* (2019) 14(9):e0222535. doi: 10.1371/journal.pone.0222535

41. group ICH. International conference on harmonization (Ich) of technical requirements for regulations of pharmaceuticals for human use. In: *Ich tripartite guideline e-9 document, statistical principles for clinical trials* (1998). Available at: https://www.ich.org/page/efficacy-guidelines#9-1.

42. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement error in nonlinear models: a modern perspective*: Chapman and Hall/CRC. (2006). doi: 10.1201/9781420010138

43. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* (2009) 6(9):677–81. doi: 10.1038/nmeth.1363

44. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* (2012) 28(18):i333–i9. doi: 10.1093/bioinformatics/bts378

45. Layer RM, Chiang C, Quinlan AR, Hall IM. Lumpy: a probabilistic framework for structural variant discovery. *Genome Biol* (2014) 15(6):1–19. doi: 10.1186/gb-2014-15-6-r84

46. Iakovishina D, Janoueix-Lerosey I, Barillot E, Regnier M, Boeva V. Sv-bay: structural variant detection in cancer genomes using a Bayesian approach with correction for gc-content and read mappability. *Bioinformatics* (2016) 32(7):984–92. doi: 10.1093/bioinformatics/btv751

47. Spinella J-F, Mehanna P, Vidal R, Saillour V, Cassart P, Richer C, et al. Snooper: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC Genomics* (2016) 17(1):1–11. doi: 10.1186/s12864-016-3281-2

48. Ding J, Bashashati A, Roth A, Oloumi A, Tse K, Zeng T, et al. Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics* (2012) 28(2):167–75. doi: 10.1093/bioinformatics/btr629

49. Freed D, Pan R, Aldana R. Tnscope: accurate detection of somatic mutations with haplotype-based variant candidate detection and machine learning filtering. *Biorxiv* (2018), 250647. doi: 10.1101/250647

50. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal snp and small-indel variant caller using deep neural networks. *Nat Biotechnol* (2018) 36(10):983–7. doi: 10.1038/nbt.4235

51. Valiant LG. A theory of the learnable. *Commun ACM* (1984) 27(11):1134–42. doi: 10.1145/1968.1972

52. Pitt L, Valiant LG. Computational limitations on learning from examples. *J ACM (JACM)* (1988) 35(4):965–84. doi: 10.1145/48014.63140

53. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y, et al. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* (2019) 20:117. doi: 10.1186/s13059-019-1720-5

54. Wang S, Liu Y, Wang J, Zhu X, Shi Y, Wang X, et al. Is an sv caller compatible with sequencing data? an online recommendation tool to automatically recommend the optimal caller based on data features. *Front Genet* (2023) 13. doi: 10.3389/fgene.2022.1096797

55. Zhu X, Ying C, Wang J, Li J, Lai X, Wang G, et al. Ensemble of ML-KNN for classification algorithm recommendation - ScienceDirect. *Knowledge-Based Syst* (2021) 221:106933. doi: 10.1016/j.knosys.2021.106933c

56. Wang X, Xu Y, Liu R, Lai X, Liu Y, Wang S, et al. PEcnv: accurate and efficient detection of copy number variations of various lengths. *Briefings Bioinf* (2022) 23(5): bbac375. doi: 10.1093/bib/bbac375