PLOS Computational Biology

METHODS

MRDtarget: A heuristic Gaussian approach for optimizing targeted capture regions to enhance Minimal Residual Disease detection

Xuwen Wang[1,2,3], Yanfang Guan[2,3,4], Wei Gao[4], Xin Lai[2,3], Wuqiang Cao[4], Xiaoyan Zhu[2,3], Xiaoling Zeng[4], Yuqian Liu[2,3], Shenjie Wang[1,2,3], Ruoyu Liu[1,2,3], Xin Yi[4], Shuanying Yang[1]*, Jiayin Wang[2,3]*

1 Department of Respiratory Medicine, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, China, **2** School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China, **3** Shaanxi Engineering Research Center of Medical and Health Big Data, Xi'an Jiaotong University, Xi'an, China, **4** Geneplus-Beijing, Beijing, China

* yangshuanying@xjtu.edu.cn (SY); wangjiayin@mail.xjtu.edu.cn (JW)

OPEN ACCESS

Citation: Wang X, Guan Y, Gao W, Lai X, Cao W, Zhu X, et al. (2025) MRDtarget: A heuristic Gaussian approach for optimizing targeted capture regions to enhance Minimal Residual Disease detection. PLoS Comput Biol 21(9): e1013443. https://doi.org/10.1371/journal.pcbi.1013443

Editor: Jinyan Li, Shenzhen University of Advanced Technology, CHINA

Received: February 20, 2025

Accepted: August 16, 2025

Published: September 17, 2025

Data availability statement: A Python-based software implementation of MRDtarget is made freely available in the GitHub repository (https://github.com/Sherwin-xjtu/MRDtarget).

Funding: This work was supported by the National Natural Science Foundation of China

Abstract

Molecular residual disease (MRD) detection, initially developed for hematologic malignancies, has become a critical biomarker for monitoring solid tumors. MRD detection primarily relies on circulating tumor DNA (ctDNA) analysis using next-generation sequencing, offering high sensitivity and broad genomic coverage. However, challenges remain in designing cost-effective panels that maximize mutation detection while maintaining biological relevance. Fixed panels often lack sufficient patient-specific mutation coverage, while WES-based personalized MRD assays, despite their high sensitivity, are costly and less accessible. We developed a tumor comprehensive genomic profiling (CGP)-informed personalized MRD assay to detect tumor-derived mutations, which allowed us to design patient-specific personalized panels and meanwhile, provide a cost-effective alternative to whole exome sequencing (WES). To address these limitations, we developed MRDtarget, a heuristic multivariate Gaussian model-based targeted capture region selection method. By expanding beyond traditional hotspot regions, MRDtarget optimizes variant tracking for MRD detection, significantly improving sensitivity. Using a Bayesian inference-based heuristic approach, MRDtarget integrates multi-feature informativeness rates to identify optimal genomic regions for capture. Experimental results demonstrate that MRDtarget enables the detection of more variants per patient. This study underscores the importance of rational panel design to improve MRD sensitivity and provides a novel approach to enhance precision diagnostics and treatment for solid tumor patients.

(grant number 92046009 to JW) and the Natural Science Basic Research Program of Shaanxi Province, China (grant number 2020JC-01 to JW). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: Authors Y.G., W.G., W.C., X.Z., and X.Y. are employees of Geneplus-Beijing 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.

Author summary

Minimal residual disease (MRD) detection plays a critical role in cancer prognosis and treatment monitoring, especially for solid tumors. However, existing sequencing panels often fail to provide sufficient coverage of tumor-specific mutations in every patient, limiting the clinical sensitivity of MRD detection. To address this gap, we developed MRDtarget, a computational tool that designs personalized targeted sequencing panels by optimizing the selection of genomic regions likely to contain informative mutations. Our method goes beyond conventional hotspot-based panels by incorporating multi-dimensional mutation features, including recurrence, clonality, and functional relevance, into a probabilistic model that prioritizes regions most informative for MRD detection. We evaluated MRDtarget using both clinical and public datasets and found that it consistently outperforms traditional approaches in mutation capture efficiency and patient coverage. Notably, it raises the proportion of patients with four or more trackable mutations—considered the minimum threshold for reliable MRD monitoring. MRDtarget also demonstrates robust performance across different cancer types and sequencing conditions. This approach enables a cost-effective and personalized solution for improving MRD detection, with broad implications for early relapse prediction and treatment guidance in precision oncology.

1. Introduction

Molecular residual disease (MRD), initially defined in hematologic malignancies, has been extended to solid tumors, where it serves as a valuable biomarker for recurrence risk and prognosis [1,2]. MRD detection primarily targets circulating tumor DNA (ctDNA) [3,4] or circulating tumor cells (CTCs) [5], helping clinicians identify high-risk patients early and guiding personalized treatment decisions [5,6]. Among detection methods, ctDNA mutation analysis using next-generation sequencing (NGS) has become the standard due to its high sensitivity, broad genomic coverage, and throughput. NGS-based ctDNA MRD detection is divided into fixed and personalized panel assays. Fixed panels, like CAPP-seq [7], are cost-effective and quickly deployable but have limited mutation coverage, particularly for rare cancers. WES-based personalized MRD assays, such as Signatera [8], use patient-specific tumor mutation profiles to achieve greater sensitivity and specificity with ultra-deep sequencing but are expensive and require tumor tissue sequencing. Tumor-informed strategies, which incorporate patient-specific mutation data, outperform tumor-agnostic approaches by reducing noise and improving detection. However, developing affordable, comprehensive genomic profiling (CGP) panels with sensitivity comparable to WES-MRD [9] remains a key challenge in advancing solid tumor MRD detection (detailed in the S1 Text).

The limit of detection (LOD) for ctDNA mutation-based MRD detection using NGS depends on factors such as cell-free DNA (cfDNA) input, sequencing depth, and the

number of tracked mutations. Clinical MRD detection typically involves monitoring multiple specific mutations, significantly enhancing sensitivity and reliability. Increasing the number of tracked mutations under fixed cfDNA input and sequencing depth further reduces the LOD [10]. We developed a probability model based on binomial distribution theory, demonstrating that with 30 ng and 60 ng cfDNA input and four tracked mutations, the LOD reaches 0.02% (S1 Fig) and 0.01% (S2 Fig), which is consistent with the performance of the WES-customized MRD product Signatera that also achieves an LOD of 0.01%. However, traditional multi-gene panels often focus on a limited number of critical driver genes, limiting their sensitivity due to tumor heterogeneity [11]. Effective tissue-based panel designs must maximize detectable mutations in each patient's tumor while maintaining the biological significance of mutation sites. This requires consistently detecting ctDNA at a frequency of ≥0.01% (with 60 ng cfDNA input) and achieving at least 95% patient coverage with four or more trackable mutation sites [8,12,13]. To meet these demands, novel algorithms are needed to enhance mutation detection capabilities, incorporate patient-specific mutation burdens, and maximize trackable mutations. These advancements will optimize MRD detection's sensitivity and stability, addressing the challenges posed by tumor heterogeneity and detection complexity.

To enhance the sensitivity of MRD detection, we focus on panel design by optimizing targeted capture sequencing regions to increase the number of mutations tracked for MRD detection, thereby improving analytical performance. To achieve this, we developed a heuristic multivariate Gaussian model-based targeted capture region selection method, MRDtarget. The core functionality of MRDtarget lies in overcoming the limitations of traditional hotspot detection regions by screening exonic regions and expanding suitable genomic regions for detection. The process begins with manually selecting features of potential expansion regions based on expert knowledge. These features are then transformed into informativeness rates using a Poisson-Binomial model. Next, multi-feature informativeness rates are integrated to construct a Multivariate Gaussian Distribution model. Finally, an optimal set of expansion regions is identified using a heuristic approach based on Bayesian inference. Experimental results demonstrate that MRDtarget, compared to traditional multi-gene panel capture region selection methods, can more precisely and efficiently increase the number of mutations detected.

2. Materials and methods

We aim to enhance the detection rate of gene mutations by optimizing targeted capture sequencing regions, thereby improving the accuracy of MRD detection. This study focuses on two key aspects: first, optimizing mainstream hotspot detection region selection strategies; and second, proposing the screening of additional exonic regions beyond hotspot target regions. To achieve the expansion of detection regions beyond hotspots, this study introduces a heuristic algorithm based on Bayesian inference. The algorithm combines the probability density of a multivariate Gaussian distribution with Mahalanobis distance to identify and expand suitable genomic regions for detection. This ensures the scientific rigor and practicality of the expanded targeted capture sequencing regions (Fig 1).

2.1 Optimization of hotspot detection region selection

Currently, mainstream hotspot region selection strategies rely primarily on records from public databases, supplemented by manual review, to define targeted capture regions based on drug guidelines, clinical recommendations, and known hotspot regions. Building on this approach, this study further optimizes and expands the capture regions with a focus on improving the detection of exonic regions. The optimization process incorporates two key aspects: selecting exonic regions with high detection rates identified through Whole Exome Sequencing (WES) and targeting exonic regions with high detection rates within hotspot genes identified by WES.

1] Identifying Seed Regions: Seed genes are selected based on their relevance to prediction, prognosis, inheritance, diagnosis, or immunotherapy efficacy in the target cancer type. The selection prioritizes key cancer-related genes, including those approved by the NCCN, FDA, or NMPA, potential targets from literature reviews, and experimental targets currently in clinical trials listed on ClinicalTrials.gov (https://clinicaltrials.gov/).

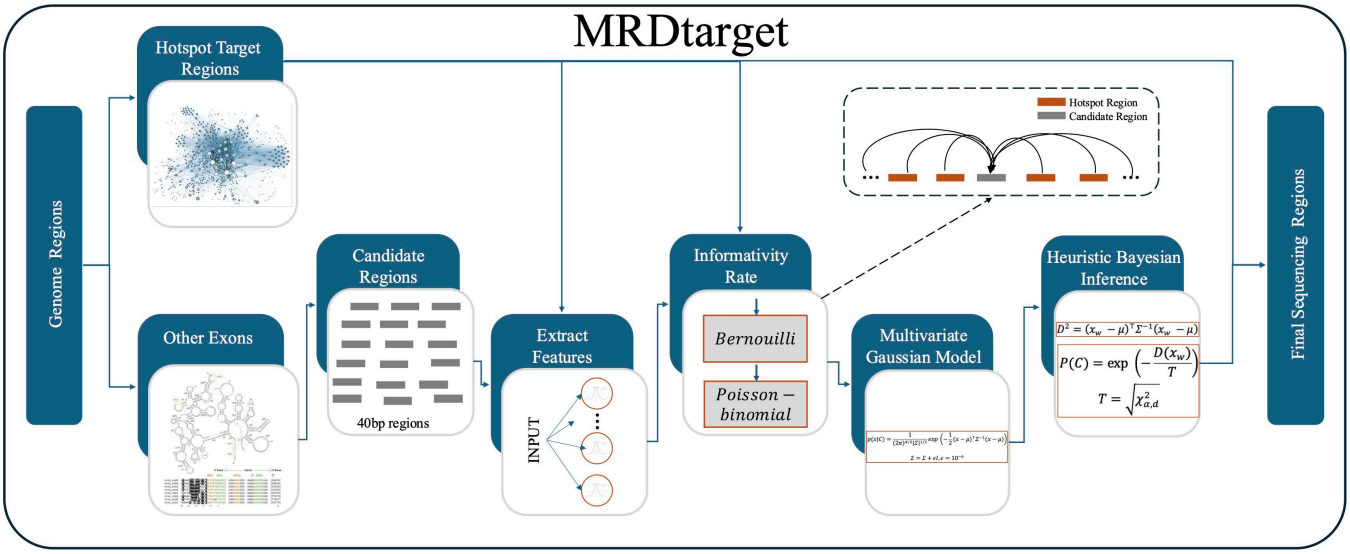

Fig 1. The workflow of MRDtarget.

https://doi.org/10.1371/journal.pcbi.1013443.g001

2] Standardizing WES Gene Exon Intervals: A unique transcript is designated for each gene to ensure analytical consistency. Commonly referenced transcripts are chosen for documented genes, while the longest transcript in NCBI is used for others. Exon region coordinates are extracted from annotation files to form the WES gene exon set, which is refined by intersecting probe-covered regions from the training dataset with exon regions.

3] Feature Extraction Based on Exons: Patient data and corresponding WES variant sets are analyzed, sourced from clinical samples, published literature, and the TCGA database (https://portal.gdc.cancer.gov/). Metrics such as the number of patients with variants, total variants, and exon length are recorded. The Recurrence Index (**RI**) is calculated for each exon to prioritize target regions, representing the average number of variants per kilobase, as defined by Eq. (1).

$$RI = 1000 * \frac{n_{wp}}{n_l * n_p} \tag{1}$$

n_p represents the number of patients in whom variants were detected in the exon region, i.e., the number of covered patients, **L** denotes the length of the exon detection region, and n is the total number of samples in the dataset. When counting the number of patients covered by the exon detection region, splice mutations (5' or 3' ends of exons) are included if the detection region contains the exon.

4) Selection of High Detection Rate Exon Regions: It involves sorting the Recurrence Index (**RI**) values in descending order to identify the most significant exon regions.

2.2 Screening of candidate regions for targeted sequencing based on a heuristic multivariate Gaussian model

In Section 2.1, we identified hotspot regions as the foundation for targeted capture region selection. Building on this, our study introduces a Bayesian inference-based heuristic algorithm to screen additional exon detection regions beyond the hotspots, expanding gene coverage and enhancing mutation detection. The algorithm (Table 1) integrates the probability density of a multivariate Gaussian distribution with Mahalanobis distance to identify high-priority exon regions for targeted sequencing. Core steps include feature extraction, multivariate Gaussian modeling, and heuristic screening via Bayesian inference.

Table 1. The algorithm of screening extended regions in targeted sequencing.

Algorithm: Method for Screening Extended Regions in Targeted Sequencing

Input:
- Known target capture sequencing regions $X \in \mathbb{R}^{n \times d}$: where n represents the number of data points and d represents the feature dimensions.
- Candidate target capture sequencing regions to be evaluated $X_t \in \mathbb{R}^{m \times d}$: where m is the number of points to be evaluated.
- Significance level: $\alpha = 0.05$.
- Regularization term: $\in = 10^{-6}$.

Output: Target capture sequencing expanded regions $R \in X_t$.

Step 1: Compute Mean Vector and Covariance Matrix
- Compute the mean vector μ for the known target capture sequencing regions:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Compute the covariance matrix Σ for the known regions:

$$\Sigma = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^{\top}$$

- To ensure the invertibility of Σ, add a regularization term $\in$ to its diagonal:

$$\Sigma = \Sigma + \in I$$

Step 2: Compute the Critical Value of the Chi-Squared Distribution
- Based on the feature dimensions d, calculate the critical value $\chi^2_{\text{threshold}}$:

$$\chi^2_{threshold} = \chi^2_{\alpha,d}$$

Step 3: Compute Mahalanobis Distance
- For each candidate region $x_w \in X_t$, calculate the squared Mahalanobis distance:

$$D^2(x_w) = (x_w - \mu)^{\top} \Sigma^{-1} (x_w - \mu)$$

Step 4: Heuristic Bayesian Decision and Selection
- Initialize the target capture expanded regions set $R = \phi$
- For each candidate point $x_w \in X_t$:
  - if $D(x_w)^2 \leq \chi^2_{\text{threshold}}$, add x_w to R:

$$R = R \cup \{x_w\};$$

  - Otherwise, skip this point.

Step 5: Output
- Return the final set of expanded target capture sequencing regions R.

https://doi.org/10.1371/journal.pcbi.1013443.t001

2.2.1 Feature extraction. MRDtarget begins by segmenting genomic regions to define the characteristics of exon detection regions. Specifically, for each exon, mutations within a 40 bp range are aggregated into a mutation cluster. The 40 bp window size was chosen based on typical probe design constraints in targeted sequencing panels, balancing region compactness with mutation clustering frequency. This window size is consistent with prior panel designs such as CAPP-Seq [7], where similar parameters are used to accommodate hybridization efficiency and regional mutation density. The region is then defined from the start position of the first mutation in the cluster to the end position of the last mutation. As a result, the length distribution of the regions includes mutation cluster regions of 40 bp or less. Based on this segmentation, the study extracts the following four key features from each 40 bp segment within the exon detection regions:

1) The WES Recurrence Index (**RI**): This metric evaluates mutation coverage in patients for each candidate detection region, representing the average number of variants per kilobase within the whole exome sequencing sample set. It is calculated using Eq. (1), with L = 40 as specified in Eq. (2).

$$RI_e = 1000 * \frac{n_{wp}}{40 * n_p} \tag{2}$$

RI_e represents the WES sample detection rate, n_{wp} denotes the number of patients covered by the 40 bp region, and n_p is the total number of samples.

2) Recurrence Improvement Index (RII): This metric evaluates the improvement in detecting additional variants for each candidate region, assessing its potential value in enhancing sample mutation coverage. When the total number of mutations covered by the existing regions is N_{cover}. Upon including the candidate region, the additional number of mutations covered for the sample is $N_{40bp-only}$. Assuming the designed target capture region ultimately aims to achieve N mutations covered per sample, the increase in the number of mutations detected in samples is defined as follows Eq. (3):

$$N_{add} = \begin{cases} 0\ ,\ N_{cover} \geq N \\ N - N_{cover}\ ,\ N_{40bp-only} + N_{cover} \geq N \\ N_{40bp-only}\ ,\ N_{40bp-only} + N_{cover} < N \end{cases} \tag{3}$$

Recurrence Improvement Index (RII) is defined as Eq. (4):

$$RII = \frac{\sum_{i=1}^{n_p} N_{add}}{l_e} \tag{4}$$

RII represents the additional mutation detection contribution rate, N_{add} denotes the number of additionally covered mutations, and l_e represents the length of the additional probe region.

3) Clonal Recurrence Improvement Index (**Clonal RII**) measures the number of additional samples with clonal mutations detected by each candidate region, evaluating the potential value of the region in improving clonal mutation coverage. When the number of clonal mutations covered by the existing regions for a sample is $N_{cover-clonal}$, Upon including the candidate region, the additional number of clonal mutations covered for the sample is $N_{40bp-only-clonal}$. Assuming the designed target capture region ultimately aims to achieve **N** mutations covered per sample, additional Covered Clonal Mutation Count is defined as Eq. (5):

$$N_{add-clonal} = \begin{cases} 0\ ,\ N_{cover-clonal} \geq N \\ N - N_{cover-clonal}\ ,\ N_{40bp-only-clonal} + N_{cover-clonal} \geq N \\ N_{40bp-only-clonal}\ ,\ N_{40bp-only-clonal} + N_{cover-clonal} < N \end{cases} \tag{5}$$

Clonal Recurrence Improvement Index (Clonal RII, RIIC) is defined as Eq. (6). Clonal mutations were defined based on cellular prevalence estimates using PyClone-VI. Specifically, variants in the cluster with maximum cellular prevalence were classified as clonal mutations. This threshold was selected to ensure robust representation of dominant tumor clones in downstream MRD analysis.

$$RII_C = \frac{\sum_{i=1}^{n_p} N_{add-clonal}}{l_e} \tag{6}$$

4) The Functional Recurrence Improvement Index (Functional RII) was defined to evaluate the additional coverage contribution of candidate regions for functional (non-benign) mutations. When the number of non-benign (*nb*) mutations covered by the existing regions is $N_{cover-nb}$. Upon including the candidate region, the additional number of non-benign mutations covered for the sample is $N_{40bp-only-nb}$. Assuming the designed target capture region ultimately aims to

achieve **N** mutations covered per sample, the increase in the number of functional (non-benign) mutations detected in samples is defined as follows Eq. (7):

$$N_{add-nb} = \begin{cases} 0\ ,\ N_{cover-nb} \geq N \\ N - N_{cover-nb}\ ,\ N_{40bp-only-nb} + N_{cover-nb} \geq N \\ N_{40bp-only-nb}\ ,\ N_{40bp-only-nb} + N_{cover-nb} < N \end{cases} \tag{7}$$

Functional Recurrence Improvement Index (Functional RII, RII_f) is defined as Eq. (8):

$$RII_f = \frac{\sum_{i=1}^{n_p} N_{add-nb}}{l_e} \tag{8}$$

2.2.2 Feature informativity rate. To facilitate the analysis, we introduce the concept of Feature Informativity Rate (**FIR**), which measures the performance of extracted features in candidate regions for targeted sequencing. Assume the extracted feature set is f, where f_i represents the i-th feature, and the number of instances corresponding to this feature is n (n_i represents the number of instances for the i-th feature). The Informativeness is defined as I_{ij}, representing the informativeness of the j-th instance of the i-th feature. Specifically: $I_{ij} = 0$: Non-informative instance, $I_{ij} = 1$: Informative instance. I_{ij} is assumed to be independent and follows a Bernoulli distribution (Eq. (9)), which varies across features. The number of informative instances surrounding f_i is denoted as n_{is}. Notably, all hotspot detection regions mentioned in Section 2.1 are treated as informative instances. This definition of **FIR** helps quantify the contribution of different features in candidate regions to mutation detection, aiding in the systematic evaluation and optimization of targeted sequencing regions.

$$I_{ij} \sim Bernouilli\,(p_{ij}) \tag{9}$$

$$P\,(I_{ij} = 1) = p_{ij}, j = 1, \ldots, n_{is} + 1 \tag{10}$$

p_{ij} represents the informativity rate of the j^{th} instance of the i-th feature. In the given feature set f_i, the total number of informative instances, N_i, is defined as Eq. (11):

$$N_i = \sum_{j=1}^{n_{is}+1} I_{ij}, j = 1, \ldots, n_{is} + 1 \tag{11}$$

The total number of informative instances, N_i, in feature f_i represents the sum of all informative instances. Since the distribution of I_{ij} varies, N_i follows a Poisson-Binomial distribution. The cumulative distribution function (CDF) of N_i is defined as Eq. (12):

$$F_{N_i}(l) = P(N_i \leq l), l = 0, \ldots, n_{is} + 1 \tag{12}$$

We focus on the probability of having at least l informative instances in the feature set f_i. Specifically, the formula for calculating the feature informativity rate $f_i(l)$ is as follows Eq. (13):

$$f_i(l) = 1 - F_{N_i}(l-1) \tag{13}$$

where, $f_i(l)$ represents the probability that the feature set f_i contains at least l informative instances; $F_{N_i}(l-1))$ is the cumulative distribution function (CDF) of N_i, the total number of informative instances in f_i. The default value for l is set to n_{is}+1, where n_{is}denotes the number of informative instances in the surrounding region.

Based on the above definition, a feature informativity rate vector f_i can be computed for each candidate region for targeted sequencing Eq. (16):

$$f_i = \{f_{i1}, f_{i2}, ..., f_{im}\} \tag{14}$$

where: m represents the total number of features in the candidate region for targeted sequencing. It should be noted that the Poisson-Binomial model assumes conditional independence among feature instances. While features such as RI and RIIC may exhibit mild correlations, we adopt this assumption for tractability and interpretability. Future extensions may consider dependency-aware models.

2.2.3 Multivariate Gaussian modeling. Based on the analysis in Section 2.2.2, feature informativity rates for each candidate region in targeted sequencing were obtained. According to probabilistic statistical theory, the feature informativity rates for each dimension approximately follow a Gaussian distribution [14]. Therefore, the overall feature informativity rates can be modeled using a Multivariate Gaussian Distribution, with the probability density function defined as Eq. (15):

$$p(x|C) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right) \tag{15}$$

where:

- d refers to the number of features used to represent each candidate region in the informativity vector x
- $x \in \mathbb{R}^d$: A d-dimensional feature vector representing a candidate region to be evaluated.
- $\mu \in \mathbb{R}^d$: The mean vector, representing the center of the known capture sequencing region distribution C.
- $\Sigma \in \mathbb{R}^{d\times d}$: The covariance matrix, representing the relationships between features.
- $|\Sigma|$: The determinant of the covariance matrix.
- Σ^{-1}: The inverse of the covariance matrix.

To ensure the numerical stability and invertibility of the covariance matrix Σ, a small regularization term $\in$ (default value $\in = 10^{-6}$) is added to its diagonal elements. The definition is Eq. (16):

$$\Sigma = \Sigma + \in I, \in = 10^{-6} \tag{16}$$

$\boldsymbol{I}$ is the identity matrix. Regularization techniques are widely used to address potential singularity issues in covariance matrices and are broadly applied in multivariate Gaussian modeling [15].

The multivariate Gaussian distribution is uniquely determined by the mean vector μ and the covariance matrix Σ. μ is to describe the central location of the distribution, reflecting the average feature informativity rates of known capture sequencing regions. Σ is to capture the correlations between features, helping to comprehensively characterize the statistical properties of candidate regions. For a candidate region x_w, its conformity to the distribution characteristics of known capture sequencing regions can be assessed by computing its probability density $p(x_w)$ under the distribution **C**. If $p(x_w)$ is high, the candidate region is close to the central characteristics of the known distribution and can be considered a potential expansion region for targeted sequencing. If $p(x_w)$ is low, the candidate region significantly deviates from the known distribution and can be regarded as an outlier, unsuitable for expansion.

2.2.4 Heuristic screening based on bayesian inference. To estimate the prior probability of a candidate exon region x_w, we adopt a Bayesian-inspired heuristic approach that leverages the Mahalanobis distance from x_w to the

centroid of known capture regions. This distance is transformed into a probability score using an exponential decay function, where smaller distances yield higher probabilities, indicating closer similarity to the known distribution. Although this transformation deviates from classical Bayesian priors, it effectively captures the centrality of data points within a multivariate Gaussian framework. Similar strategies have been applied in functional data classification (e.g., Galeano et al. [16]), and anomaly detection using local Mahalanobis distances (e.g., Yang et al. [17]) and weighted Mahalanobis models (e.g., Wen et al. [18]). The squared Mahalanobis distance D^2 is calculated as Eq. (17):

$$D^2 = (x_w - \mu)^\top \Sigma^{-1} (x_w - \mu) \tag{17}$$

Here, μ is the mean vector of known target regions, and Σ is the covariance matrix. A small D^2 indicates that x_w is close to the distribution center, suggesting suitability as an expanded capture region. Conversely, a large D^2 implies outlier status.

The prior probability $P(C)$ for a candidate region is computed via Eq. (18):

$$P(C) = exp\left(-\frac{D(x_w)}{T}\right) \tag{18}$$

Where T is a critical value derived from the Chi-Squared distribution, used for normalizing the distance and ensuring the prior probability lies within the range [0, 1]. And $\boldsymbol{T}$ is calculated based on the significance level α and degrees of freedom d (number of features) using Eq. (19). The use of exponential decay to transform Mahalanobis distance into a prior probability is inspired by its effectiveness in probabilistic outlier detection. This transformation allows soft penalization of outlier regions while preserving a continuous scoring spectrum. Similar approaches have been used in functional data classification (e.g., Galeano et al.) [16], where distance-based scores are mapped into probability estimates to reflect centrality in the feature space.

$$T = \sqrt{\chi^2_{\alpha,d}} \tag{19}$$

Under the assumption of multivariate normality, the squared Mahalanobis distance $D(x_w)^2$ follows a Chi-Squared distribution with degrees of freedom d. Therefore, outliers can be determined using the critical value of the Chi-Squared distribution. Given a significance level α (set to 0.05 by default) and the number of features d, the critical value is calculated using Eq. (20):

$$\chi^2_{threshold} = ChiSquared^{-1}(1-\alpha, d) \tag{20}$$

Here, ChiSquared^{-1} is the inverse cumulative distribution function of the Chi-Squared distribution. When $D(x_w)^2 \leq \chi^2_{\text{threshold}}$, x_w belongs to the known distribution and can be considered as a potential extension region for targeted sequencing. Conversely, when $D(x_w)^2 \leq \chi^2_{\text{threshold}}$, x_w is considered an outlier and unsuitable as an extension region.

3. Results

3.1. Sample preparation

To validate the effectiveness of MRDtarget, its performance was analyzed and evaluated using both public database samples and clinical samples. The validation compared WES results with targeted capture sequencing results obtained using a conventional pre-optimized panel, hereafter referred to as Tdesigner, and the optimized panel generated by our proposed algorithm, MRDtarget. Tdesigner was constructed based on standard hotspot region selection strategies, relying on public database records and clinical guidelines without incorporating recurrence-based metrics or statistical screening.

In contrast, MRDtarget integrates WES-derived recurrence features and a Bayesian heuristic model to expand the capture region beyond known hotspots. Key performance metrics, including variant density and clonal population clustering, were used to comprehensively assess the performance improvements achieved by MRDtarget over Tdesigner.

3.1.1. TCGA sample cohort. This study utilized 1,024 lung cancer samples from The Cancer Genome Atlas (TCGA) database, sequenced using Whole Exome Sequencing. The cohort included 544 lung adenocarcinoma cases (TCGA Project ID: TCGA-LUAD) and 480 lung squamous cell carcinoma cases (TCGA Project ID: TCGA-LUSC), representing the pathological subtypes of lung cancer. By comparing variant detection results from WES, pre-optimized, and post-optimized targeted capture panels, the study aimed to comprehensively assess the sensitivity and practical utility of the optimized panel design.

3.1.2. Clinical sample cohort. We also included 150 clinical solid tumor tissue samples, comprising 123 lung cancer cases, 10 breast cancer cases, 8 digestive system tumor cases, and 9 cases of other solid tumors. Each sample was analyzed using both pre-optimized and post-optimized targeted capture panels, with sequencing depth exceeding 500x. Variant counts were compared between the two panels for the same samples. PyClone-VI [19] was subsequently employed to conduct clustering analysis on the detected variants, enabling the classification of clonal and sub-clonal variants.

3.1.3. Preprocessing. Raw sequencing data were preprocessed by trimming terminal adaptors and removing low-quality sequences, defined as those with more than 50% N bases or more than 50% of bases having a quality score (Q) <5. The clean reads were aligned to the reference human genome (GRCh37.p13, GCF_000001405.25) using BWA-MEM2 (version 2.2.1a) [20]. Patient-specific somatic variants were identified by analyzing sequencing data from primary tumors and matched peripheral blood lymphocyte (PBL) samples. Tumor somatic single-nucleotide variants (SNVs) and small insertions and deletions (InDels) were called using RealDcaller2 (version 2.0.9) and TNscope (v3.8.0; Sentieon Inc.), as described previously [21,22]. Structural variants (SVs) were profiled using NCsv2 (version 1.2.0), an in-house tool developed at Geneplus-Beijing [21,22].

3.2. Performance on TCGA samples

3.2.1 Variant density. Existing studies have demonstrated that increasing the number of variant monitoring sites can lower the limit of detection (LoD) and enhance the sensitivity of MRD detection [23]. In this study, variants from the TCGA dataset were filtered by excluding those with Variant_Classification values of RNA, 3'Flank, 5'Flank, or intergenic region (**IGR**), as well as mutations located outside the ± 50 bp range of intron regions. To evaluate the effectiveness of the proposed MRDtarget tool, we compared the optimized method with WES across TCGA samples. Variant density for both MRDtarget and WES was calculated, with the variant density (d), representing the average number of variants per kilobase, as defined by Eq. (21):

$$d = 1000 * \frac{n_m}{l_r} \tag{21}$$

Here, d represents the variant density, n_m is the number of variants (sourced from a publicly available variant dataset) detected within the targeted capture region, and l_r is the length of the targeted capture region. Variant density quantifies the effectiveness of targeted sequencing methods in detecting variants, with higher values indicating enhanced detection capabilities. We calculated and visualized the variant density across all chromosomes (excluding the Y chromosome) for each targeted capture region (Fig 2A). MRDtarget, representing the optimized method, showed a significant improvement over WES, achieving a 5.67-fold and 1.57-fold increase in mean and median variant density, respectively. This demonstrates the superior variant capture capability of MRDtarget, consistently outperforming WES across all autosomes and the X chromosome.

To evaluate the performance of MRDtarget against traditional panel design strategies, we compared the variant densities of MRDtarget, Tdesigner, and WES (Fig 2B). The significantly distinct variant density distributions ($p < 0.01$) among the three methods highlight their differing gene detection and capture strategies. MRDtarget outperformed Tdesigner,

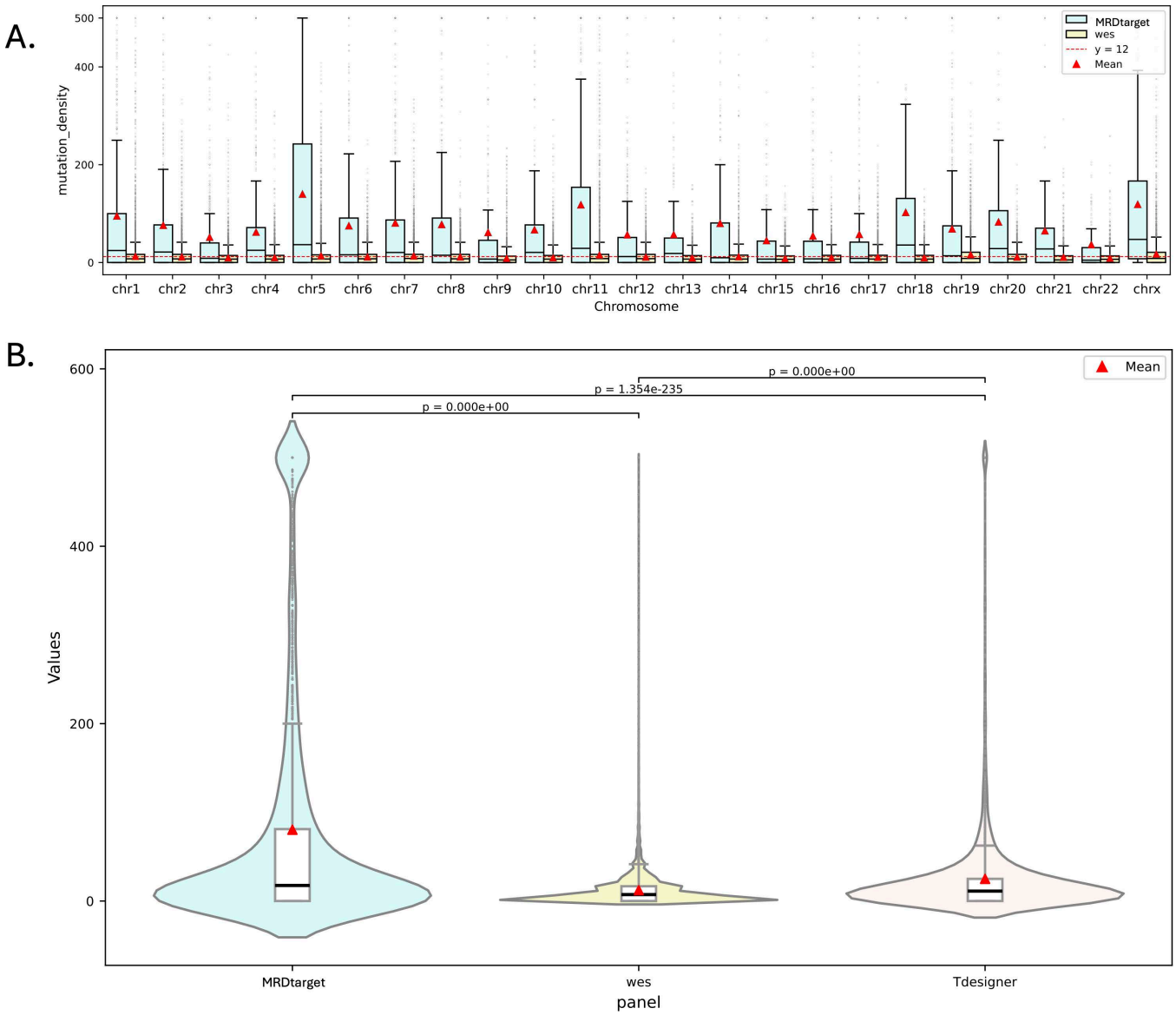

Fig 2. Comparison of Variant Density Distributions Across Methods. Variant density distributions across all chromosomes (excluding the Y chromosome) are shown for MRDtarget, Tdesigner, and WES. (A) Variant densities for each targeted capture region were calculated and visualized, highlighting the improved performance of MRDtarget compared to WES. (B) A comparison of the variant density distributions among MRDtarget, Tdesigner, and WES demonstrates their distinct gene detection strategies. MRDtarget consistently achieves higher variant densities across autosomes and the X chromosome, emphasizing its optimized capture capability.

https://doi.org/10.1371/journal.pcbi.1013443.g002

achieving a mean variant density of 80 versus 25 and a median of 18 versus 10, representing 2.20-fold and 0.8-fold improvements, respectively. Both MRDtarget and Tdesigner showed higher variant densities than WES, with MRDtarget achieving 5.67-fold and 1.57-fold increases in mean and median densities, while Tdesigner achieved 1.083-fold and 0.429-fold improvements. These results emphasize the superior performance of MRDtarget in enhancing variant detection by focusing on high-variant-density regions, providing precise genomic insights for applications such as precision oncology and MRD detection. Further optimization of Tdesigner, incorporating data-driven strategies, could enhance its ability to prioritize key genomic regions.

3.3. Performance on clinical cancer samples

We analyzed 150 clinical solid tumor tissue samples, including 123 lung cancer cases, 10 breast cancer cases, 8 digestive system tumor cases, and 9 cases of other solid tumors, using both pre-optimized (Tdesigner) and post-optimized (MRDtarget) targeted capture panels with sequencing depths exceeding 500x. Previous studies [12,13] have emphasized the need for NGS-based ctDNA multi-gene panels to comprehensively cover class I/II gene variants and reliably detect ctDNA at abundances ≥0.01% with 60 ng of cfDNA. Such panel designs must increase the number of detectable tumor tissue variants per patient while ensuring that 95% of patients have four or more monitorable variant sites, thereby enhancing MRD detection sensitivity. Variant counts between the two panels were compared, and clonal population clustering was conducted using PyClone-VI to annotate clonal and sub-clonal variants.

3.3.1 Variant counts. We first evaluated the performance of MRDtarget and Tdesigner in detecting variants across different cancer types. As shown in Fig 3A, MRDtarget demonstrated superior variant detection capabilities compared to

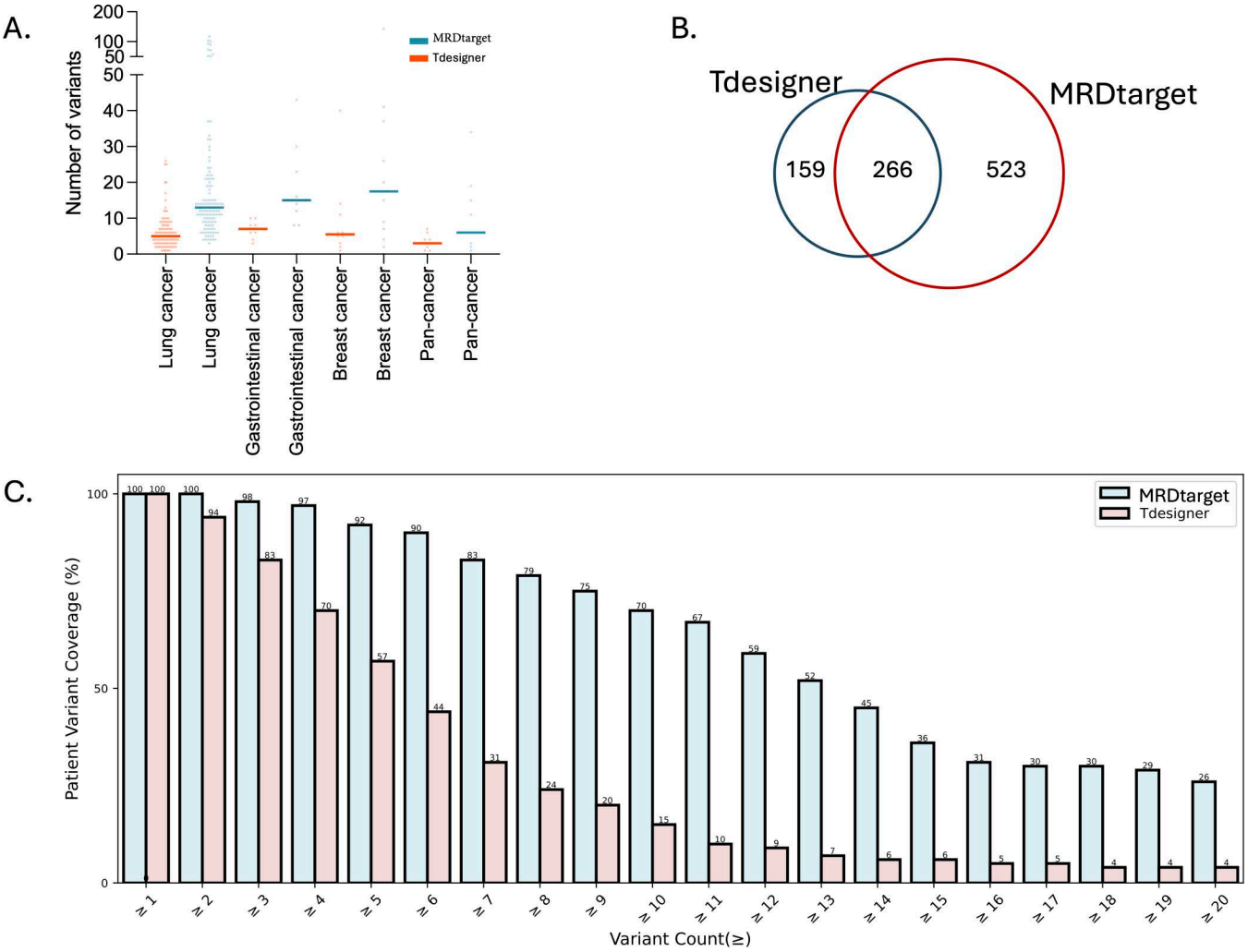

Fig 3. The performance on variant detection. (A) Median variant counts detected by MRDtarget and Tdesigner across 123 lung cancer cases, 10 breast cancer cases, 8 digestive system tumor cases, and 9 other solid tumors. (B) Comparison of clonal variants detected by MRDtarget and Tdesigner. (C) Patient coverage analysis for MRDtarget and Tdesigner.

https://doi.org/10.1371/journal.pcbi.1013443.g003

Tdesigner across all tumor types. The median variant counts detected by MRDtarget and Tdesigner in 123 lung cancer cases, 10 breast cancer cases, 8 digestive system tumor cases, and 9 other solid tumor cases were (13, 15, 17.5, 6) and (5, 7, 5.5, 3), respectively. These results highlight MRDtarget's enhanced ability to identify variants across diverse tumor types, emphasizing its robustness and reliability for broader clinical applications. Next, we compared the performance of MRDtarget and Tdesigner in detecting clonal variants. As shown in Fig 3B, Tdesigner identified 425 clonal variants, while MRDtarget detected 789. In 64.7% of the samples (97 out of 150), MRDtarget identified more clonal variants than Tdesigner, with a median increase of three variants per sample. These findings demonstrate that the optimized MRDtarget panel significantly improves the detection of clonal variants. The ability to detect more clonal variants highlights MRDtarget's improved sensitivity and efficiency, which are critical for MRD monitoring. Clonal variants are closely associated with disease progression and treatment outcomes, making MRDtarget a more effective tool for reliable MRD assessments. Finally, we analyzed the patient coverage achieved by MRDtarget and Tdesigner under a fixed number of mutations. As shown in Fig 3C, MRDtarget achieved 97% patient coverage for those with four or more trackable mutation sites, exceeding the 95% threshold required to maximize population coverage and MRD detection sensitivity [8,12,13]. In comparison, Tdesigner only covered 70% of patients meeting this criterion. From a clinical perspective, achieving four or more trackable mutations per patient is often considered a critical threshold for reliable MRD monitoring. Our optimized MRDtarget achieves this in 97% of patients, exceeding the 95% benchmark suggested in prior studies [8,12,13].

3.3.2 Clonal variant counts. Study [23] has shown that clonal variants hold greater prognostic value than subclonal variants in MRD monitoring. In our experiment with 150 clinical cancer samples, as illustrated in Fig 4, each bar represents a single patient. The top portion of each bar shows the number of clonal variants detected, while the bottom portion indicates the subclonal variants, using MRDtarget and Tdesigner for the same patients. A comparison demonstrates that MRDtarget consistently detects a higher total number of variants for most patients, with notable increases in both clonal and subclonal variants. The improved detection of clonal variants aligns with MRDtarget's design goal of prioritizing clinically significant regions, while the enhanced detection of subclonal variants reflects its broader sensitivity to low-abundance variants. These advancements are particularly significant

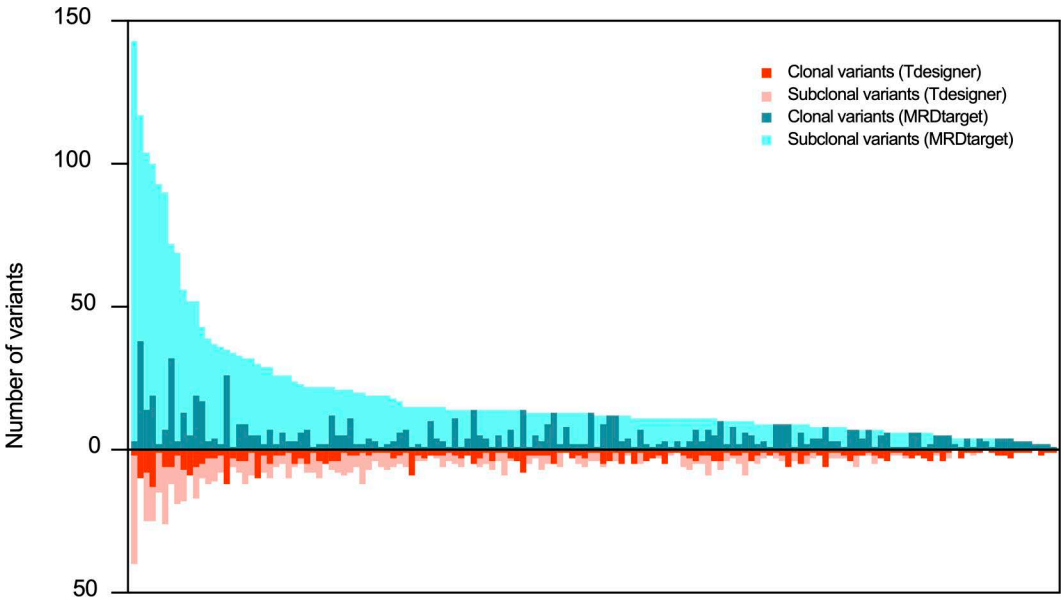

Fig 4. The performance on clonal variants. Each bar represents a single patient, with the top portion indicating the number of major clonal variants and the bottom portion representing subclonal variants detected using MRDtarget and Tdesigner.

https://doi.org/10.1371/journal.pcbi.1013443.g004

for MRD monitoring, where comprehensive variant detection is essential for accurately assessing tumor burden and disease progression. By identifying more variants across both categories, MRDtarget generates a more detailed and reliable genomic profile, enhancing patient stratification and supporting the development of personalized therapeutic strategies.

4. Discussion and conclusion

MRDtarget is an advanced tool for optimizing targeted capture sequencing regions in MRD detection. By integrating a heuristic multivariate Gaussian model, MRDtarget overcomes key limitations of traditional panel design strategies, such as limited patient-specific mutation coverage and reduced sensitivity. By expanding beyond hotspot regions and incorporating high-density exonic regions, MRDtarget enhances its utility across diverse cancer types. Compared to Tdesigner and WES, MRDtarget demonstrates superior performance, achieving significantly higher variant densities and providing better coverage of clonal and subclonal variants. Notably, it consistently outperforms Tdesigner in patient coverage, with 97% of patients having four or more trackable mutation sites, exceeding the 95% threshold for optimal MRD detection. Its ability to detect more clonal variants, which are critical for disease progression and treatment decisions, enhances the precision of MRD assessments and supports personalized therapeutic strategies. MRDtarget's tailored capture strategies address tumor heterogeneity by prioritizing regions likely to harbor key variants, enabling a more detailed genomic profile. This is essential for applications in biomarker discovery, precision oncology, and genetic research. The use of Bayesian inference and multivariate Gaussian modeling provides a reliable statistical foundation, ensuring accuracy and reproducibility. All features are currently standardized and treated equally, with no additional weighting applied. However, we would like to emphasize that in the heuristic screening stage based on Bayesian inference, the final selection of candidate regions is guided by their similarity to known hotspot regions. These hotspot regions themselves are determined using conventional panel design strategies, which often prioritize regions with high RI values. As a result, although we do not explicitly assign weights to individual features, the upstream selection of hotspot regions may implicitly introduce a bias that favors certain features—particularly RI—during the final region selection process. However, the tool's reliance on existing features for region selection highlights an area for improvement. Expanding the feature set could enhance its robustness and adaptability across diverse cancer types and genomic contexts. Future efforts will focus on expanding the feature set to further improve precision and applicability. In addition, we plan to conduct a sensitivity analysis to quantitatively assess the relationship between probe length, sequencing cost, and detection sensitivity. This will enable a more balanced design of capture panels that optimize both efficiency and clinical feasibility. Although the current analysis was based on the hs37d5 (GRCh37.p13) reference genome, chosen for its compatibility and widespread use in clinical pipelines, we acknowledge its limitations in handling sex-specific variation. Notably, Y chromosome variants were not explicitly excluded during downstream analysis. This design choice aimed to ensure broad applicability across mixed-sex populations. As a result, features from X- and Y-linked regions were calculated uniformly, without sex-specific weighting. While this may reduce precision in certain scenarios, it simplifies implementation across heterogeneous cohorts. Following best practices in sex-aware genomic analysis (Olney et al., 2020) [24], future versions of MRDtarget will adopt sex-specific reference genomes based on GRCh38.p14 and incorporate population-stratified modeling strategies to improve the accuracy, robustness, and equity of variant detection.

In conclusion, MRDtarget represents a significant advancement in MRD detection, offering an efficient and sensitive framework for targeted sequencing panel optimization. MRDtarget can push more patients above the four-variant threshold further underscores its potential to improve real-world MRD detection coverage. And its strong performance in variant detection and patient coverage underscores its potential for advancing precision oncology and improving clinical outcomes. With further refinement and validation, MRDtarget could play a central role in developing cost-effective, high-performance genomic profiling panels for cancer diagnosis and treatment.

5. Key points

- Novel Heuristic Model: MRDtarget integrates a multivariate Gaussian model with Bayesian inference to optimize targeted capture region selection, achieving higher variant densities and ensuring 97% patient coverage with four or more trackable mutations.
- Expansion of Exonic Regions: Focuses on high-density exonic regions to enhance variant detection beyond traditional hotspot strategies.
- Iterative Refinement of Target Regions: Employs a systematic process of feature extraction, statistical modeling, and heuristic optimization to iteratively refine and expand capture regions for enhanced variant detection.

Supporting information

S1 Text. Supplementary text and Supplementary Figures.
(DOCX)

S1 Fig. 30ng cfDNA.
(TIF)

S2 Fig. 60ng cfDNA.
(TIF)

Acknowledgments

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

Author contributions

Conceptualization: Shuanying Yang, Jiayin Wang.

Data curation: Yanfang Guan, Wei Gao, Shenjie Wang, Ruoyu Liu.

Formal analysis: Xuwen Wang, Yanfang Guan, Wei Gao, Shenjie Wang, Ruoyu Liu.

Funding acquisition: Shuanying Yang, Jiayin Wang.

Investigation: Xuwen Wang.

Methodology: Shuanying Yang, Jiayin Wang.

Resources: Xin Lai, Wuqiang Cao, Xiaoyan Zhu, Xiaoling Zeng, Yuqian Liu, Xin Yi.

Supervision: Shuanying Yang, Jiayin Wang.

Visualization: Xuwen Wang.

Writing – original draft: Xuwen Wang.

Writing – review & editing: Yanfang Guan, Wei Gao, Xin Lai, Shuanying Yang, Jiayin Wang.

References

1. Böckmann B, Grill HJ, Giesing M. Molecular characterization of minimal residual cancer cells in patients with solid tumors. Biomol Eng. 2001;17(3):95–111. https://doi.org/10.1016/s1389-0344(00)00073-3 PMID: 11222984
2. Coakley M, Garcia-Murillas I, Turner NC. Molecular residual disease and adjuvant trial design in solid tumors. Clinical Cancer Research. 2019;25(20):6026–34.

3. Chin R-I, Chen K, Usmani A, Chua C, Harris PK, Binkley MS, et al. Detection of solid tumor Molecular Residual Disease (MRD) using Circulating Tumor DNA (ctDNA). Molecular Diagnosis & Therapy. 2019;23(3):311–31.
4. Kasi PM, Fehringer G, Taniguchi H, Starling N, Nakamura Y, Kotani D, et al. Impact of circulating tumor DNA-based detection of molecular residual disease on the conduct and design of clinical trials for solid tumors. JCO Precis Oncol. 2022;6:e2100181. https://doi.org/10.1200/PO.21.00181 PMID: 35263168
5. Pantel K, Alix-Panabières C. Minimal residual disease as a target for liquid biopsy in patients with solid tumours. Nat Rev Clin Oncol. 2025;22(1):65–77. https://doi.org/10.1038/s41571-024-00967-y PMID: 39609625
6. Pantel K, Alix-Panabières C. Liquid biopsy and minimal residual disease - latest advances and implications for cure. Nat Rev Clin Oncol. 2019;16(7):409–24. https://doi.org/10.1038/s41571-019-0187-3 PMID: 30796368
7. Newman AM, Bratman SV, To J, Wynne JF, Eclov NCW, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. 2014;20(5):548–54. https://doi.org/10.1038/nm.3519 PMID: 24705333
8. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature. 2017;545(7655):446–51. https://doi.org/10.1038/nature22364 PMID: 28445469
9. Powles T, Young A, Nimeiri H, Madison RW, Fine A, Zollinger DR, et al. Molecular residual disease detection in resected, muscle-invasive urothelial cancer with a tissue-based comprehensive genomic profiling-informed personalized monitoring assay. Front Oncol. 2023;13:1221718. https://doi.org/10.3389/fonc.2023.1221718 PMID: 37601688
10. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol. 2016;34(5):547–55. https://doi.org/10.1038/nbt.3520 PMID: 27018799
11. Song P, Wu LR, Yan YH, Zhang JX, Chu T, Kwong LN, et al. Limitations and opportunities of technologies for the analysis of cell-free DNA in cancer diagnostics. Nat Biomed Eng. 2022;6(3):232–45. https://doi.org/10.1038/s41551-021-00837-3 PMID: 35102279
12. Moding EJ, Nabet BY, Alizadeh AA, Diehn M. Detecting liquid remnants of solid tumors: circulating tumor DNA minimal residual disease. Cancer Discov. 2021;11(12):2968–86. https://doi.org/10.1158/2159-8290.CD-21-0634 PMID: 34785539
13. Reinert T, Henriksen TV, Christensen E, Sharma S, Salari R, Sethi H, et al. Analysis of plasma cell-free DNA by ultradeep sequencing in patients with stages I to III colorectal cancer. JAMA Oncol. 2019;5(8):1124–31. https://doi.org/10.1001/jamaoncol.2019.0528 PMID: 31070691
14. Duda R, Hart P, G SD. Pattern classification. 2001.
15. Bishop C. Pattern Recognition and Machine Learning: Springer; 2006.
16. Galeano P, Joseph E, Lillo RE. The mahalanobis distance for functional data with applications to classification. Technometrics. 2015;57(2):281–91.
17. Yang J, Delpha C. An incipient fault diagnosis methodology using local Mahalanobis distance: detection process based on empirical probability density estimation. Signal Processing. 2022;190:108308.
18. Wen J, Gao H. Remaining useful life prediction of the ball screw system based on weighted Mahalanobis distance and an exponential model. Journal of Vibroengineering. 2018;20:1691–707.
19. Gillis S, Roth A. PyClone-VI: scalable inference of clonal population structures using whole genome data. BMC Bioinformatics. 2020;21(1):571. https://doi.org/10.1186/s12859-020-03919-2 PMID: 33302872
20. Vasimuddin M, Misra S, Li H, Aluru S, editors. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS); 2019, 20-24 May 2019.
21. Zhang JT, Liu SY, Gao W, Liu SM, Yan HH, Ji L. Longitudinal undetectable molecular residual disease defines potentially cured population in localized non-small cell lung cancer. Cancer Discov. 2022;12(7):1690–701.
22. Pan Y, Zhang J-T, Gao X, Chen Z-Y, Yan B, Tan P-X, et al. Dynamic circulating tumor DNA during chemoradiotherapy predicts clinical outcomes for locally advanced non-small cell lung cancer patients. Cancer Cell. 2023;41(10):1763-1773.e4. https://doi.org/10.1016/j.ccell.2023.09.007 PMID: 37816331
23. Wang S, Li M, Zhang J, Xing P, Wu M, Meng F, et al. Circulating tumor DNA integrating tissue clonality detects minimal residual disease in resectable non-small-cell lung cancer. J Hematol Oncol. 2022;15(1):137. https://doi.org/10.1186/s13045-022-01355-8 PMID: 36183093
24. Olney KC, Brotman SM, Andrews JP, Valverde-Vesling VA, Wilson MA. Reference genome and transcriptome informed by the sex chromosome complement of the sample increase ability to detect sex differences in gene expression from RNA-Seq data. Biol Sex Differ. 2020;11(1):42. https://doi.org/10.1186/s13293-020-00312-9 PMID: 32693839