BMC Bioinformatics

RESEARCH Open Access

The application of compressed sensing on tumor mutation burden calculation from overlapped pooling sequencing data

Yue Cui[1], Yi Qiao[1], Rongming An[1,2], Xuan Pan[3*] and Jing Tu[1*]

*Correspondence:
panxuan0214@163.com;
jtu@seu.edu.cn

[1] State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China
[2] Monash University-Southeast University Joint Research Institute, Suzhou 215123, China
[3] Department of Medical Oncology, Jiangsu Cancer Hospital, Jiangsu Institute of Cancer Research, The Affiliated Cancer Hospital of Nanjing Medical University, Nanjing 210009, China

Abstract

Background: Tumor Mutation Burden (TMB) is commonly characterized as the number of non-synonymous somatic SNVs per megabase within the gene region identified through whole exon sequencing or targeted sequencing in a tumor sample. It has been statistically demonstrated that TMB was related to the ability of neoantigen production and used to predict the efficacy of immunotherapy for various types of cancers. However, screening for TMB in patients poses challenges due to the extensive labor and financial resources required for the preparation of large quantities of parallel sequencing libraries.

Results: In this study, we employed compressed sensing (CS) to calculate TMB from overlapped pooling sequencing data, aiming to reduce the sequencing cost by minimizing the number of library builds. Over 90% SNPs could still be detected without a significant loss of mutation information even when the data is pooled from ten different samples. Based on this, the orthogonal matching pursuit (OMP) algorithm and the basic pursuit (BP) algorithm were used to reconstruct TMB from pooling sequencing data. The performance of these two algorithms was evaluated. The BP algorithm consistently performed well across all cases, albeit necessitating extended computational time. The OMP algorithm has been proved to be suitable for scenarios where the original matrix was sparse but it showed low overall performance. Based on an accurate calculation of TMB, we determined that the number of sequencing runs could be reduced to 0.6 times the total number of samples, resulting in a 40% reduction in sequencing cost.

Conclusions: In conclusion, we calculated TMB from overlapped pooling sequencing data utilizing compressed sensing strategy to reduce sequencing cost. Our findings confirm that the SNP calling from ten samples' pooling sequencing data is feasible. Additionally, we performed an assessment of the reconstruction efficiency of both the BP model and the OMP model.

Keywords: Tumor mutation burden, Compressed sensing, Overlapped pooling, DNA sequencing

Introduction

Tumor mutation burden (TMB) refers to the measurement of somatic nonsynonymous mutations of the coding region in a tumor tissue genome [1]. These non-synonymous mutations may generate neoantigens which can stimulate an immune response by activating T cells [2]. Higher TMB implies a higher likelihood of neoantigen generation, potentially leading to increased recognition and targeting of cancer cells by the immune system [3]. Immune checkpoint therapy (ICT) is a type of cancer immunotherapy that provide durable clinical responses and improve overall survival but only subsets of patients benefit from it [4]. In addition to PD-L1 expression and microsatellite instability (MSI), TMB has been widely used as biomarker to identify patients that are potentially sensitive to ICT [5].

Since TMB was initially identified as a possible biomarker for ICT in melanoma[6], numerous studies have demonstrated the link between TMB and clinical efficacy of ICT [7–9]. TMB is counted after removing the germline mutations, which could be obtained from either normal tissue or public database [10]. With the progress of high-throughput sequencing, TMB can be evaluated by whole exome sequencing (WES), or targeted panel sequencing [11]. Although WES has become the "gold standard" for measuring TMB, it is not currently feasible in clinical practice with large samples because of its high cost [12]. The high cost majorly comes from the library preparation for large samples prior to sequencing since the cost of sequencing itself is decreasing [13]. If the TMB of each sample can be retrieved with fewer sequencing tries, the cost will be significantly decreased.

Compressed sensing (CS) is a novel signal sampling and processing theory for sparse or compressible signals, which enables the reconstruction of a signal from a small number of measurements [14]. It provides a way to capture and represent signals efficiently from fewer measurements compared to traditional sampling methods [15]. Taking advantage of this, compressed sensing has been applied in the field of high throughput single-cell transcriptomic sequencing, where it provides an alternative method to infer single-cell expression profiles with fewer sequencing times [16–18]. In spite of the similar amount of total sequencing data, CS decreases the overall cost by significantly reducing the number of sequencing libraries. As similar as single cell expression data, non-synonymous somatic mutation profile in different tumor tissues also own sparse characteristics, indicating that it may be worthwhile to apply compressed sensing theory on calculating TMB to save cost.

In this paper, we investigated the feasibility and efficacy of employing compressed sensing as a cost saving method to determine TMB from overlapped pooling sequencing data. We assessed the precision of SNP detection from overlapped pooling sequencing data and developed a comprehensive pipeline for calculating TMB from such data. The efficacy of two reconstruction algorithms was examined by analyzing the reconstruction outcomes of 6669 samples from 30 different types of cancers, across varied compression levels. Our findings indicated that the number of sequencing times can be diminished to 0.6 times the number of samples, resulting in a 40% reduction in the number of library preparation, without compromising the accurate calculation of the TMB for each sample.

Method

Data collection

The WES data from 10 patients with nasopharyngeal carcinoma (NPC) were downloaded in FASTQ format from National Center for Biotechnology Information (NCBI) under sequence read archive (SRA) accessions SRA291701 [19]. Details of the WES data for these ten samples can be found in the supplementary material (Supplementary Table S1).

The mutation annotation format (MAF) files were downloaded from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). The MAF files were generated utilizing the Mutect2, with subsequent filtration of germline mutations through the analysis of matched normal tissue sequencing data.We chose the available MAF files obtained from the WES data and filtered out 6669 MAF files. We first downloaded a manifest file that contains the necessary information for the selected MAF files and later used the GDC Data Transfer Tool, a command-line tool provided by the Genomic Data Commons (GDC) to batch download them. Details of the MAF files for these 6669 samples can be found in the supplementary material (Supplementary Table S2).

WES data analysis

The WES data were processed according to GATK Best Practices recommendations [20]. The sequence reads were aligned to the reference human genome (hg19) using Burrows-Wheeler Aligner [21]. Following mapping, the sam files are processed with the help of the Picard tools to sort output bam files and mark duplicates. GATK was applied for base quality score recalibration, variant discovery and variant filtering. Ultimately, VCF files storing the SNPs called from sequencing data was obtained for each sample and use Bcftools to count the number of SNPs in each vcf file.

Firstly, we counted the proportion of missing and false SNPs obtained from two samples' sequencing data before implementing SNP quality control measures. We merged raw sequencing data of two samples into a single fastq file and repeated the bioinformatics pipeline described above to get raw SNPs. Using Bcftools, the vcf file obtained from pooling data was subjected to comparison with each of the two vcf files derived from the original WES data. The objective was to generate two vcf files containing information about missing SNPs from two samples. By tallying the count of SNPs within these two vcf files, without eliminating any duplicate SNPs, the overall number of missing SNPs was determined. Subsequently, dividing the count of lost SNPs by the total count of SNPs present in the two samples enabled the calculation of the percentage of SNPs that were lost. SNPs that were exclusively called in the pooling sequencing data but not in the original WES data of two samples were designated as false-positive SNPs. The count of such SNPs was determined using Bcftools and divided by the total number of SNPs detected from the pooling sequencing data to calculate the proportion of false SNPs. Secondly, we applied quality control procedures to the obtained SNPs. Specifically, for the SNPs called from the original WES data, we excluded those with sequencing coverage depth below 5×. For SNPs obtained from pooling data, we applied various sequencing coverage depths to filter

the SNPs and assessed the proportions of missing SNPs and false SNPs across different filtering conditions.

We then merged raw sequencing data of ten samples and followed the same processing procedure employed for the two samples.

MAF files analysis

Generation of mutation matrix: The mutation matrix X is a binary matrix with dimensions of *sample number*×*SNP number*. We removed duplicate non-synonymous mutations contained in MAF files and found all SNPs. If the sample i contains the SNP j then $M(i,j)=1$, else $M(i,j)=0$. For different cancers we respectively generated their mutation matrices. Thirty mutation matrices were derived by processing MAF files obtained for thirty distinct types of cancer.

Generation of measurement matrix: Measurement matrices M are randomly generated Bernoulli matrices with dimensions of *pool number*×*sample number*, where the probabilities of 1 to appear in a matrix are set as an adjustable parameter p. We ensured that at most 10 samples are mixed in each pool by setting the parameter p. When the sample i appears in the pool j then $M(i,j)=1$ else $M(i,j)=0$. For different cancers we used the same method to generate measurement matrix as above.

Reconstruction algorithm: The compressive measurements Y were built up by multiplying M with X ($Y=M\times X$). We introduced another parameter k, defined as the ratio of the total number of samples to the number of pools. We set k from 0.1 to 0.6 (interval = 0.1) and randomly generated M. We solved $\widehat{X}$ by two reconstruction algorithms respectively: Basic Pursuit (BP) and Orthogonal Matching Pursuit (OMP).

In order to turn $\widehat{X}$ into a binary matrix containing only 0 and 1, we set the threshold as the mean of the $\widehat{X}$ plus three times the standard deviation of the $\widehat{X}$, as follows:

$$threshold = \widehat{x} + 3 \times \sigma$$

We used Pearson Correlation Coefficient to compare reconstructed $\widehat{X}$ with the original X. Though X is solved column by column, we care more about reconstructed performance of each sample. So, we calculated the Pearson Correlation Coefficient by comparing inferential mutation matrix of each sample to its original data (ρ_{sample}), and we considered mean of Pearson Correlation Coefficient of all samples as the detection consistency of different reconstruction algorithms as follows:

$$\rho_{sample} = \frac{\sum_{i=1}^{n}(x_{c,i}-\overline{x})(\widehat{x}_{c,i}-\overline{\widehat{x}})}{\sqrt{\sum_{i=1}^{n}(x_{c,i}-\overline{x})^2\sum_{j=1}^{n}(\overline{x}_{c,j}-\overline{\widehat{x}})^2}}$$

$$\rho = \frac{\sum_{sample=1}^{m}\rho_{sample}}{m}$$

We also calculated the accuracy, measuring the number of correctly reconstructed SNPs in each sample. We considered mean of accuracy of all samples as the detection accuracy of different reconstruction algorithms as follows:

$$Accuracy_{sample} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{\sum_{sample=1}^{m} Accuracy_{sample}}{m}$$

Furthermore, we computed the recall rate measuring the number of correctly reconstructed SNPs that are truly mutated in each sample. We considered mean of recall of all samples as the detection sensitivity of different reconstruction algorithms as follows:

$$Recall_{sample} = \frac{TP}{TP + FN}$$

$$Recall = \frac{\sum_{sample=1}^{m} Recall_{sample}}{m}$$

The Mann–Whitney test is employed in order to assess the presence of statistically significant distinction between the two models across various levels of compression.

TMB calculation: After filtering out non-synonymous mutations from MAF files, we divided the total number of non-synonymous mutations by the whole exome region size, as follows:

$$TMB = \frac{Numberofnon - synonymousmutationsinthemaffile}{Exonregionlength}$$

After reconstruction, we got inferential TMB of each sample based on $\widehat{X}$ by adding up the amount of 1 s in each row.

Result

Framework of TMB calculation using compressed sensing strategy

The conceptual definition of TMB refers to the total number of mutations identified within a tumor specimen. However, the specific types of genetic alterations included in the assessment of TMB have differed based on the methodologies employed [12]. Some studies take into account both somatic SNVs and indels when calculating TMBs through WES [22–24], while other studies focus exclusively on somatic SNVs [25–27]. In general, synonymous and germline variants are excluded from the calculation of TMB, as it is presumed that these variants are unlikely to play a direct role in the generation of neoantigens and somatic mutations are more likely to serve as a contributing factor to the onset of tumor. In order to effectively eliminate germline variations, it is preferable to sequence a matched non-tumor sample from each patient. However, such matched samples may not be available in clinical practice and numerous targeted panel methodologies do not incorporate matched normal samples [3]. In the context of tumor-only sequencing, it is possible to eliminate germline variants by utilizing extensive, publicly accessible germline variant databases [28].

In order to achieve compressed sampling and minimize library usage, we explored the idea of using the compressed sensing (CS) method. The basic idea is to subsample different samples into overlapped pools, where a sample can appear in multiple groups and

each group includes multiple samples. Then construct libraries for pools (less than the number of samples) to get composite sequencing data and call single nucleotide polymorphisms (SNPs) from pooling sequencing data. Finally certain CS algorithms are used to reconstruct SNPs present in each sample and calculate TMB for them. By leveraging the shared information of pools, this overlapping methodology can effectively decrease the number of tests needed through cross-information.

Equation (1) is a representation of the theoretical model for our method (Fig. 1A). Among them, M stands for the measurement matrix, a known matrix randomly generated by computer. M is a binary $p \times s$ matrix, where p indicates the number of pools, and s represents the number of samples. According to matrix M, the samples were partitioned into pools, where $M(i,j)=1$ indicates that the i^{th} pool contains the j^{th} sample. X is a binary $s \times m$ matrix, where m represents SNP. When i^{th} sample contains j^{th} SNP, $X(i,j)=1$ (Fig. 1B). Y represents the observation data, which is obtained by multiplying matrix M and matrix X. We utilized matrix Y and matrix M to reconstruct $\widehat{X}$ using reconstruction algorithms and then compared $\widehat{X}$ with X.

$$Y_{p\times m} = M_{p\times s} \times X_{s\times m} \tag{1}$$

The process of reconstruction can be likened to the endeavor of solving an ill-conditioned linear equation, which has an infinite number of solutions. Compressed sensing can help to solve this equation when X is sparse. Reconstructing the sparse signal or solving the sparse solution of the ill-conditioned linear equation is a minimum l_0 norm optimization problem. However, this problem is classified as NP-C, making it challenging to solve directly. Typically, it is converted to the minimum l_1 norm to solve or it is solved through greedy algorithm. In this paper, we applied Basis Pursuit model (BP, l_1 norm) and Orthogonal Matching Pursuit (OMP, greedy algorithm) for reconstruction.

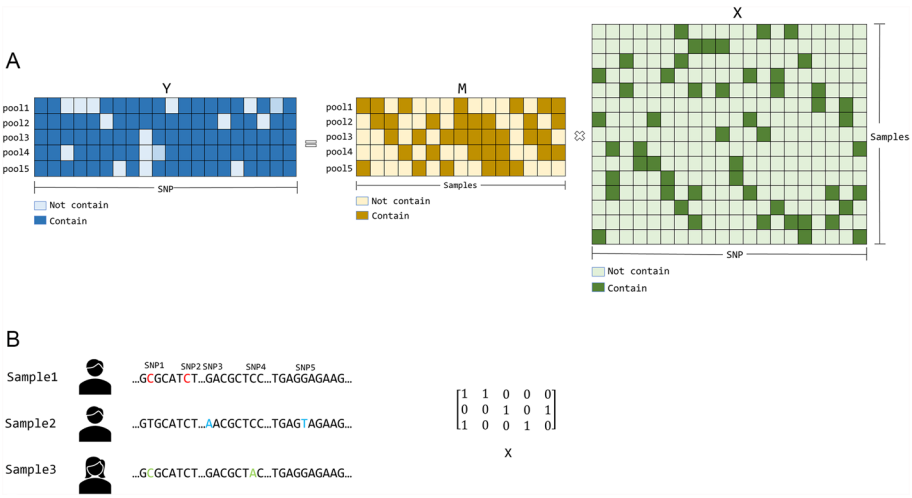

Fig. 1 Objection overview illustration. **A** Illustration of pool strategy. The *M* matrix is pool-sample matrix, which is overlapped. In this example, 20 samples are designated into 5 pools. M(1,1) is painted means pool1 contains sample1. Y is pool-SNP matrix, which is obtained by multiplying M and X in computational simulation. In actual experiment Y is obtained from sequencing data. X is sample-SNP matrix. **B** Illustration of mutation matrix generation. X(1,1) means sample1 contains SNP1. X is generated from MAF files in computational simulation. In actual experiment, we want to infer X from Y and M

The BP model aims to recover a sparse signal by finding the minimum l_1 norm solution that satisfies a given set of linear equations. While the OMP model selects atoms (basis functions) from a dictionary in an iterative manner to approximate a signal using a restricted number of non-zero coefficients.

Two issues should be validated for the utilization of CS on TMB calculation. Firstly, the utilization of the overlapped pooling method may pose challenges in accurately calling SNPs by reducing the depth of coverage for some SNPs. This can lead to these sites being mistakenly identified as sequencing errors. Secondly, it is necessary to confirm whether the original mutation matrixes show sparsity. Because the underlying principle of compressed sensing theory is that the original signal exhibits sparsity. In practical applications, if the original mutation matrixes approximately satisfy the condition of sparsity, wherein a majority of the values tend towards zero, it can be considered compressible and subjected to subsampling.

SNP calling from pooling sequencing data

Pooled sequencing approaches present challenges and potential inaccuracies in SNP calling. When a SNP is shared by multiple samples, there is a possibility of allelic deletions, where one allele may be preferentially detected. This can distort the observed allele frequencies, resulting in inaccurate SNP detection. Additionally, if pooled samples have high genetic heterogeneity such as tumor samples, it becomes difficult to accurately distinguish true variants from background noise. The presence of genetic heterogeneity increases the complexity and affects the accuracy of SNP detection. In order to determine the impact of pooled sequencing on SNP calling, we mixed the WES data for 2 samples and 10 samples separately and conducted a comprehensive bioinformatics analysis. We examined characteristics of the lost and false SNPs based on the variant information recorded in the vcf files. We found that sequencing coverage depth hold great significance in SNP quality control (Supplementary Figure S1, S2). In order to determine the most effective filtering conditions, we applied various filtering schemes and compared the extent of similarity between the SNPs identified in the mixed data and those identified in the original WES data. For the SNPs called from the original WES data, we excluded those with sequencing coverage depth below 5×. Given that the majority of SNP loci in the pooling sequencing data exhibited an upward trend in coverage depth, we explored various sequencing depth thresholds for the SNPs obtained from the pooling sequencing data (Supplementary Figure S3).

In the case of both two and ten samples, the proportion of missing SNPs rises while the proportion of false SNPs declines when employing a greater depth to filter the SNPs obtained from the pooling sequencing data (Fig. 2A, B). This phenomenon occurs due to the implementation of a higher depth as threshold, which eliminates a number of SNPs called from the pooling sequencing data. Consequently, this process leads to an increased loss of SNPs and a reduction in the occurrence of false SNPs. Given that we need to compute the TMB in the subsequent analysis, it is imperative to choose the case in which the difference between the proportion of missing SNPs and the proportion of false SNPs is minimized and approaches zero. We determined that the most optimal choices for filtering SNPs from two samples' pooling sequencing data is to set a

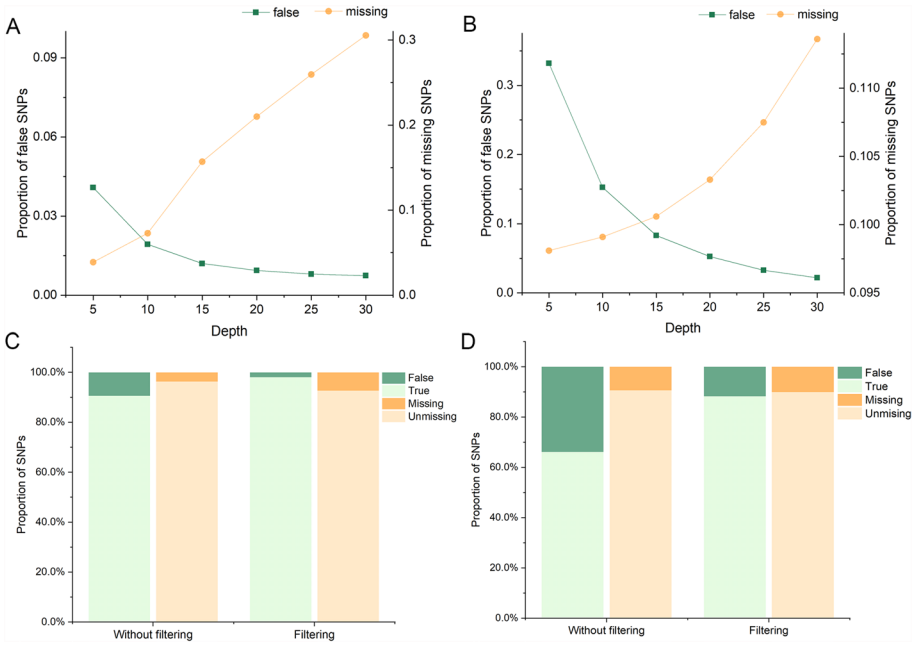

Fig. 2 Percentage of missing and false SNPs in various situations. **A** Percentage of missing and false SNPs when using various sequencing coverage depths to filter SNPs called from two samples pooling sequencing data. **B** Percentage of missing and false SNPs when using various sequencing coverage depths to filter SNPs called from ten samples pooling sequencing data. **C** Comparison of the percentage of missing and false positive SNPs before and after the application of filters on the two samples' sequencing data. **D** Comparison of the percentage of missing and false positive SNPs before and after the application of filters on the ten samples' sequencing data

threshold of 10×, and for filtering SNPs from ten samples' pooling sequencing data is to set a threshold of 15×.

For two samples' pooling sequencing data, we found that 3.68% of SNPs were lost and 9.39% of SNPs were false positives without any filtering. However, when the SNPs obtained from pooling sequencing data were filtered with a sequencing coverage depth of 10×, 7.29% of the SNPs were lost, and the percentage of false SNPs decreased to 1.93% (Fig. 2C). For the ten samples' pooling sequencing data, the percentage of missing SNPs was 9.36% and the percentage of false positive SNPs was 33.86%. After applying depth of 15× as a threshold to the SNPs called from pooling sequencing data, we observed a loss of 10.06% of the SNPs and the percentage of false SNPs decreased to 8.83% (Fig. 2D).

Over 90% of the SNPs called from the pooling sequencing data, following proper SNP quality control procedures, are usable for further analysis. Additionally, over 90% of the SNPs obtained from the original WES data can be identified through pooling sequencing. Overall, the inaccuracy of SNP detection due to pooled sequencing is deemed acceptable based on our findings.

TMB reconstruction across varied compression levels

In the 30 mutation matrices we generated, it was observed that the percentage of elements with a value of 0 ranged from 98 to 99%, while the percentage of elements with a value of 1 ranged from 1 to 2%. The sparsity profile of each mutation matrix can be accessed in the supplementary material (Supplementary TableS2). Consequently,

the sparse nature of the mutation matrix renders it suitable for the theoretical framework of compressed sensing. Then we pooled samples of 30 different types of cancer separately in a mixed manner of ten samples per pool. We compared the reconstruction results of BP and OMP at different compression levels with the original mutation matrix to evaluate the performance of these two algorithms (Supplementary Figure S4).

As expected, the correlation observed between the reconstructed outcome and the original data rises with the increment of the pool number due to the increase in sampling times (Fig. 3A). High correlation was already achieved when the pool number was 0.6 of the sample number, although the correlation of 0.1 pool number was as high as ~ 0.8. The standard deviation drops significantly as the pool number increased. The correlation derived from the BP model reconstruction result consistently exhibit greater magnitudes compared to the correlation observed between the OMP model reconstruction result and the original data. It can be deduced that the reconstruction efficiency of the BP model is notably superior to that of the OMP model in cases where the number of pools is limited. The accuracy of the models also improves as the number of pools increases; however, the accuracy of both models remains comparable (Fig. 3B). It should be noted that accuracy is an inadequate metric for evaluating

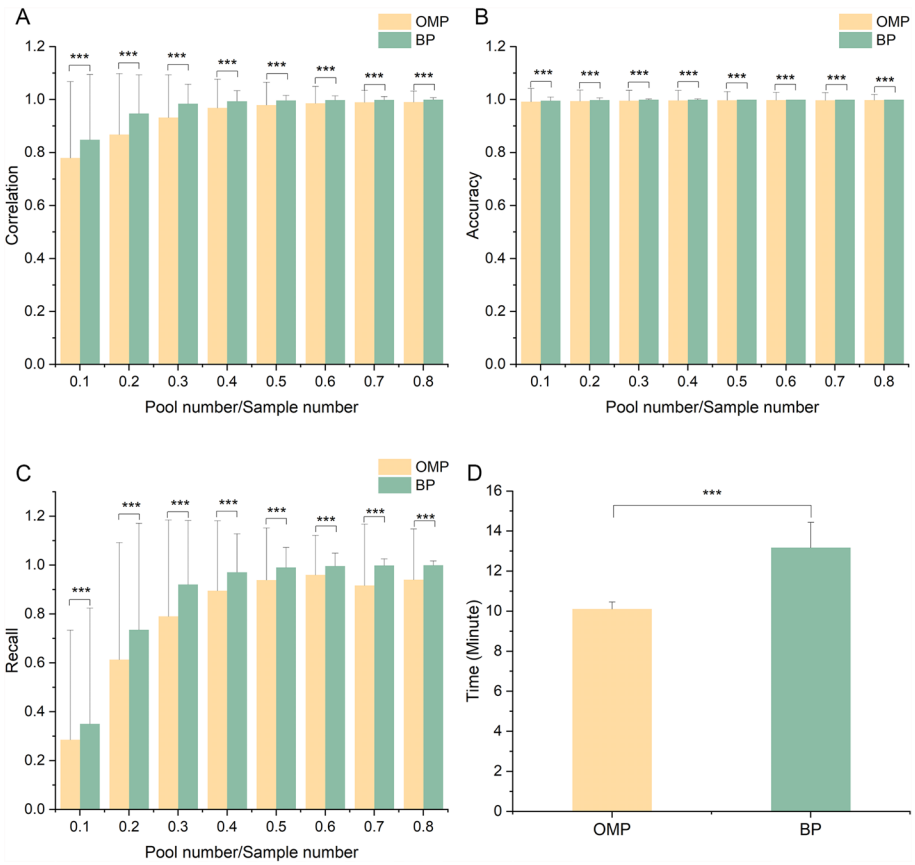

Fig. 3 Performance of BP model and OMP model at different levels of compression (**A**) Mean Pearson correlation observed between the reconstructed outcome and the original data. **B** The mean accuracy of two models. **C** The mean recall rate of two models. **D** The mean time required for reconstruction

model performance as our original matrixes contain a significantly larger number of 0 compared to 1 and the model may achieve a high accuracy by incorrectly classifying all the 1 as 0.

In order to conduct a more comprehensive evaluation of the BP model and OMP model, we proceeded to assess the recall of both models. Recall is a metric that measures the model's capacity to accurately detect and classify positive instances, thereby indicating its proficiency in minimizing false negatives. When predicting TMB, the presence of false-negative SNPs is deemed more problematic than false-positive SNPs. A high false negative rate implies that the TMB value predicted by the model will be considerably lower than the actual TMB value, consequently resulting in the failure to identify individuals who may benefit from immunotherapy. This failure can potentially lead to misdiagnosis or delayed treatment. The performance of both models, in terms of recall, improves as the number of pools increases (Fig. 3C). This suggests that with an increasing number of pools, the two models become more adept at recognizing positive SNPs, thereby reducing the occurrence of false negatives SNPs in the reconstructed results. Notably, the recall rate of the BP model consistently and significantly surpasses that of the OMP model. This observation indicates that the BP model is more effective in accurately identifying and capturing positive SNPs, thereby minimizing the number of missed or disregarded SNPs classified as negative. In addition, we recorded the time taken by both models across various levels of compression. It was observed that the OMP model exhibited significantly shorter running times compared to the BP model (Fig. 3D).

The distribution of the results for each parameter derived from BP model and OMP model across various compression levels is presented in the Supplementary Material. The violin plot provides a clearer visualization of the significant differences between the two models, thereby facilitating a more effective comparison of their performance across various compression levels (Figure S6).

TMB reconstruction across varied sample characteristics

When the pool is adjusted to 0.6 times the number of samples, the Pearson correlation can reach a high value of 0.998 and the recall can also reach a high value of 0.996(Basis Pursuit). Both of these parameters surpass the threshold of 0.98, indicating a strong association and accurate retrieval of relevant information. In general, when the pool size is 0.6 times the sample size, it becomes feasible to obtain a substantial quantity of information that exhibits a high level of agreement with the original, as indicated by the visualization of the histogram (Fig. 4A) and scatter plot (Supplementary Figure S5A). This implies that the TMB for 30 different types of cancers can be accurately calculated using the reconstruction results obtained from the BP model ($p > 0.05$). This results in a cost reduction of 40% for library construction. A disparity persists between the effectiveness of the OMP and BP when the number of pools is equivalent to 0.6 times the number of samples. When the number of pools is 0.6 times the number of samples, the OMP model exhibits a correlation coefficient of 0.986 between the reconstructed results and the original data. Additionally, the OMP model demonstrates a recall of 0.960, which is comparatively lower than the BP model for both parameters. The accurate calculation of TMB for six different types of cancers cannot be achieved based on the results obtained

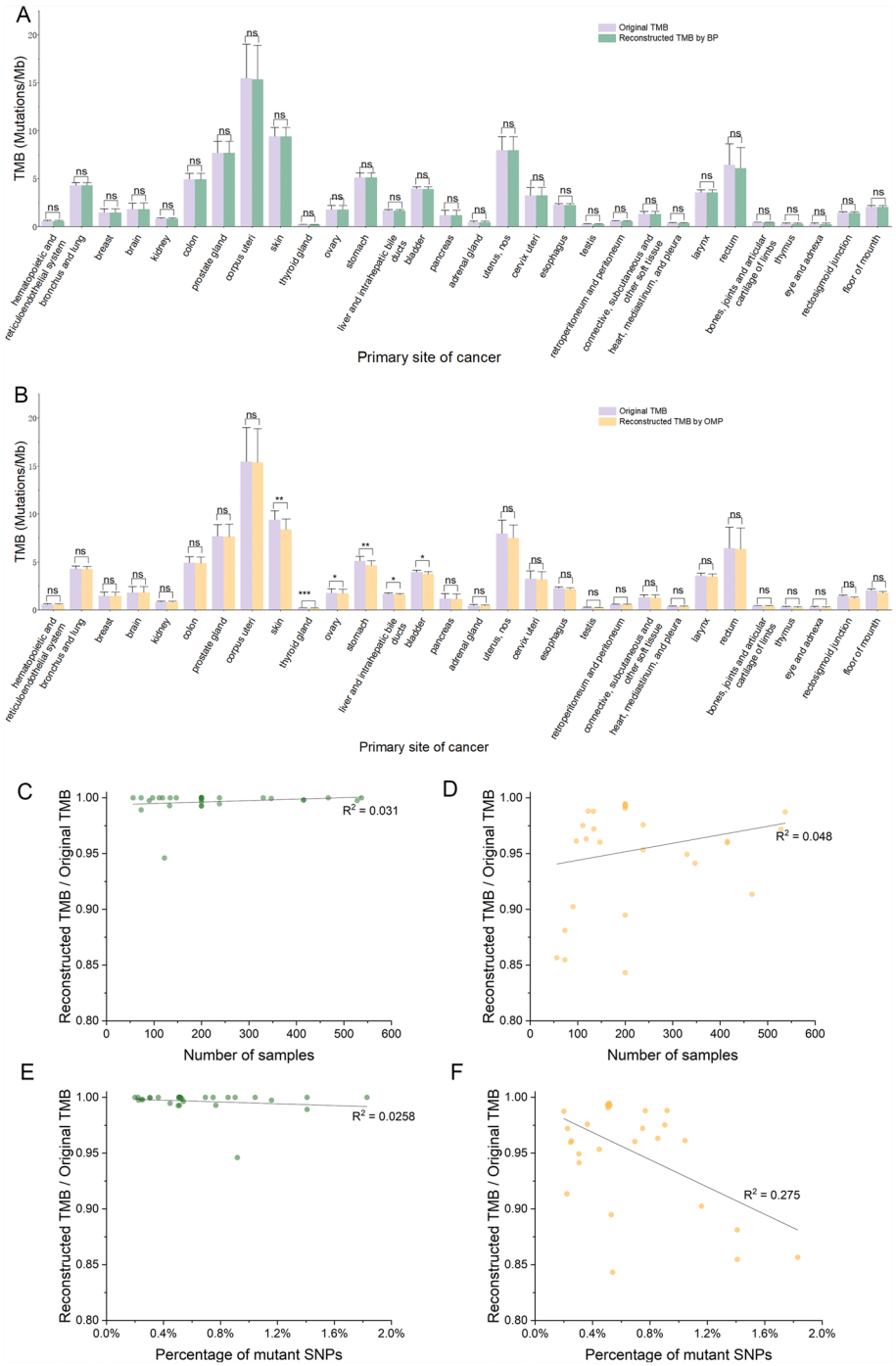

Fig. 4 The two models' accuracy in calculating TMB across different sample characteristics (pool number / sample number = 0.6). **A** Calculation the TMB using the BP model for each type of cancer. **B** Calculation the TMB using the OMP model for each type of cancer. **C** The relationship between the accuracy of TMB calculation using BP model and the number of samples. **D** The relationship between the accuracy of TMB calculation using OMP model and the number of samples. **E** The relationship between the accuracy of TMB calculation using BP model and the percentage of mutated SNPs. **F** The relationship between the accuracy of TMB calculation using OMP model and the percentage of mutated SNPs

from the OMP model (Fig. 4B, Supplementary Figure S5B). This suggests that the OMP model is less effective than the BP model in some cases.

To gain a deeper understanding of the factors that impact the reconstruction outcomes of the two models, we conduct an analysis considering the variables of sample size and the proportion of mutant SNPs in original matrixes. In our study, we designated the SNPs present in each sample as mutant SNPs. These SNPs were represented by the numerical value 1 in the original mutation matrixes.

In our analysis, it is evident that the accuracy of calculating TMB from OMP reconstruction result is enhanced when a larger sample size is employed (Fig. 4D). Since pool number is equal to 0.6 times the number of samples, as the number of samples increases, the number of pools will also increase. A larger pool size increases the number of equations in the ill-conditioned system of equations, thereby yielding a more precise solution and consequently improving the quality of the reconstruction outcomes. It is worth noting that cancers with a smaller number of samples can still achieve satisfactory reconstruction results. This is primarily due to the reduced total SNPs present in these cancer samples. Consequently, despite the limited number of equations, the smaller number of unknowns facilitates accurate solutions to the equations. Furthermore, it is observed that the performance of the OMP model in reconstructing the original mutation matrix deteriorates as the proportion of mutant SNPs in the matrix increases (Fig. 4F). This can be attributed to two factors. Firstly, as the number of mutant SNP increases, the computational complexity of the OMP model also increases, resulting in higher computational demands and potentially reduced efficiency. Secondly, when the number of mutant SNP is high, the OMP model is more susceptible to overfitting issues, leading to suboptimal reconstruction outcomes.

The behavior of the OMP model is not constant and can vary depending on the specific problem and data characteristics. The OMP model is particularly well-suited for scenarios where the original matrix is sparser. The OMP model exhibits a considerable sensitivity to the proportion of mutant SNPs. The proportion of mutant SNPs is indicative of the TMB in cancer patients, with a higher TMB being associated with an increased proportion of mutant SNPs. This indicates that the OMP model is more appropriate for cancer types characterized by lower TMB. On the other hand, the BP model consistently performs well across all cases (Fig. 4C, E). However, the computational complexity of the BP model is higher than that of the OMP model, resulting in longer processing times. Therefore, in situations where the original mutation matrix consists of a greater number of positive SNPs, the preference is to use the BP model.

The precise identification of somatic mutations is essential for the calculation of TMB. However, detecting somatic mutations with varying allele fractions in tumor tissues presents significant challenges, because of sample degradation, tissue contamination, and insufficient coverage. In our study, we analyzed the distributions of allele fractions for somatic mutations across thirty different types of cancer. A significant disparity exists in the frequency distributions of somatic mutant allele fraction among patients diagnosed with different types of tumor (Figure S7). An appropriately enhanced sequencing coverage facilitates the identification of SNVs with varying allele fractions. The elevated sequencing coverage of the pool utilized in our methodology may enhance the detection of somatic mutations. It is anticipated that researchers will develop additional

algorithms and statistical models to enhance the accuracy of somatic mutation detection and improve the precision of TMB assessment.

Conclusion

In this study, we calculated TMB from overlapped pooling sequencing data with compressed sensing strategy in order to save the cost of sequencing. Our findings validated that the SNP calling from ten samples' pooling sequencing data is achievable. Furthermore, we conduct an analysis and comparison of the reconstruction efficacy of the BP model and OMP model. The cost on library preparation could be saved for 40% without significant TMB difference and a potential higher compressing level is expected.

Abbreviations

TMB	Tumor mutation burden
CS	Compressed sensing
OMP	Orthogonal matching pursuit
BP	Basic pursuit
SNP	Single nucleotide polymorphism
ICT	Immune checkpoint therapy
PD-L1	Programmed death-ligand 1
MSI	Microsatellite instability
WES	Whole exome sequencing
NCBI	National center for biotechnology information
NPC	Nasopharyngeal carcinoma
MAF	Mutation annotation format
TCGA	The cancer genome atlas
GDC	Genomic data commons
BWA	Burrows-wheeler aligner
GATK	Genome Analysis Toolkit
VCF	Variant call format

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06148-7.

Additional file 1.

Additional file 2.

Additional file 3.

Additional file 4.

Additional file 5.

Additional file 6.

Acknowledgements

We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

Author contributions

J.T. and X.P. conceived and designed the experiments; Y.C., R.A. and Y.Q. performed the experiment and analysis. Y.C. and Y.Q. wrote the main manuscript text and prepared all figures. J.T. and X.P. reviesed the manuscript. All authors reviewed the manuscript.

Funding

This work was supported by the Natural Science Foundation of Jiangsu Province (BK20211513), the Fundamental Research Funds for the Central Universities (2242023K5005), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX21_0144).

Availability of data and materials

The WES data were downloaded from NCBI under SRA accessions SRA291701. Details of the WES data for these ten samples can be found in Supplementary Table S1. The MAF files were downloaded from TCGA database (https://portal.gdc.cancer.gov/). Projects involved in this work can be found in Supplementary Table S3 and MAF files' ID for these 6,669 samples can be found in Supplementary Table S4.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 10 November 2023 Accepted: 24 April 2025
Published online: 20 May 2025

References

1. Marabelle A, Fakih M, Lopez J, Shah M, Shapira-Frommer R, Nakagawa K, Chung HC, Kindler HL, Lopez-Martin JA, Miller WH, et al. Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. Lancet Oncol. 2020;21(10):1353–65.
2. Litchfield K, Reading JL, Puttick C, Thakkar K, Abbosh C, Bentham R, Watkins TBK, Rosenthal R, Biswas D, Rowan A, et al. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. Cell. 2021;184(3):596-614.e514.
3. Sha D, Jin ZH, Budczies J, Kluck K, Stenzinger A, Sinicrope FA. Tumor mutational burden as a predictive biomarker in solid tumors. Cancer Discov. 2020;10(12):1808–25.
4. Galluzzi L, Humeau J, Buque A, Zitvogel L, Kroemer G. Immunostimulation with chemotherapy in the era of immune checkpoint inhibitors. Nat Rev Clin Oncol. 2020;17(12):725–41.
5. Gjoerup O, Brown CA, Ross JS, Huang RSP, Schrock A, Creeden J, Fabrizio D, Tolba K. Identification and utilization of biomarkers to predict response to immune checkpoint inhibitors. Aaps J. 2020;22(6):132.
6. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, Walsh LA, Postow MA, Wong P, Ho TS, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. N Engl J Med. 2014;371(23):2189–99.
7. Barroso-Sousa R, Jain E, Cohen O, Kim D, Buendia-Buendia J, Winer E, Lin N, Tolaney SM, Wagle N. Prevalence and mutational determinants of high tumor mutation burden in breast cancer. Ann Oncol. 2020;31(3):387–94.
8. Ma K, Huang FX, Wang Y, Kang Y, Wang QL, Tang JQ, Sun PF, Lou JJ, Qiao RP, Si JM, et al. Relationship between tumor mutational burden, gene mutation status, and clinical characteristics in 340 cases of lung adenocarcinoma. Cancer Med-Us. 2022;11(22):4389–97.
9. Gabbia D, De Martin S. Tumor mutational burden for predicting prognosis and therapy outcome of hepatocellular carcinoma. Int J Mol Sci. 2023;24(4):3441.
10. Anagnostou V, Bardelli A, Chan TA, Turajlic S. The status of tumor mutational burden and immunotherapy. Nat Cancer. 2022;3(6):652–6.
11. Xu Z, Dai J, Wang D, Lu H, Dai H, Ye H, Gu J, Chen S, Huang B. Assessment of tumor mutation burden calculation from gene panel sequencing data. Onco Targets Ther. 2019;12(1178–6930):3401–9.
12. Jardim DL, Goodman A, de Melo GD, Kurzrock R. The challenges of tumor mutational burden as an immunotherapy biomarker. Cancer Cell. 2021;39(2):154–73.
13. Bayle A, Droin N, Besse B, Zou Z, Boursin Y, Rissel S, Solary E, Lacroix L, Rouleau E, Borget I, et al. Whole exome sequencing in molecular diagnostics of cancer decreases over time: evidence from a cost analysis in the French setting. Eur J Health Econ. 2021;22(6):855–64.
14. Donoho DL. Compressed sensing. IEEE Trans Inf Theory. 2006;52(4):1289–306.
15. Rani M, Dhok SB, Deshmukh RB. A systematic review of compressive sensing: concepts, implementations and applications. IEEE Access. 2018;6:4875–94.
16. Huang M, Yang Y, Wen X, Xu W, Lu N, Sun X, Tu J, Lu Z. Inferring single cell expression profiles from overlapped pooling sequencing data with compressed sensing strategy. Nucleic Acids Res. 2021;49(14):7995–8006.
17. Yu Z, Bian C, Liu G, Zhang S, Wong KC, Li X. Elucidating transcriptomic profiles from single-cell RNA sequencing data using nature-inspired compressed sensing. Brief Bioinform. 2021; 22(5).
18. Zhang S, Li X, Lin Q, Wong KC. Nature-inspired compressed sensing for transcriptomic profiling from random composite measurements. IEEE Trans Cybern. 2021;51(9):4476–87.
19. Dai W, Zheng H, Cheung AK, Tang CS, Ko JM, Wong BW, Leong MM, Sham PC, Cheung F, Kwong DL, et al. Whole-exome sequencing identifies MST1R as a genetic susceptibility gene in nasopharyngeal carcinoma. Proc Natl Acad Sci U S A. 2016;113(12):3317–22.
20. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.
21. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
22. Tang B, Zhu J, Zhao Z, Lu C, Liu S, Fang S, Zheng L, Zhang N, Chen M, Xu M, et al. Diagnosis and prognosis models for hepatocellular carcinoma patient's management based on tumor mutation burden. J Adv Res. 2021;33:153–65.
23. Bi F, Chen Y, Yang Q. Significance of tumor mutation burden combined with immune infiltrates in the progression and prognosis of ovarian cancer. Cancer Cell Int. 2020;20:373.

24. Karn T, Denkert C, Weber KE, Holtrich U, Hanusch C, Sinn BV, Higgs BW, Jank P, Sinn HP, Huober J, et al. Tumor mutational burden and immune infiltration as independent predictors of response to neoadjuvant immune checkpoint inhibition in early TNBC in GeparNuevo. Ann Oncol. 2020;31(9):1216–22.
25. Carbone DP, Reck M, Paz-Ares L, Creelan B, Horn L, Steins M, Felip E, van den Heuvel MM, Ciuleanu TE, Badin F, et al. First-line nivolumab in stage IV or recurrent non-small-cell lung cancer. N Engl J Med. 2017;376(25):2415–26.
26. Das A, Tabori U, Sambira Nahum LC, Collins NB, Deyell R, Dvir R, Faure-Conter C, Hassall TE, Minturn JE, Edwards M, et al. Efficacy of nivolumab in pediatric cancers with high mutation burden and mismatch repair deficiency. Clin Cancer Res. 2023;29(23):4770–83.
27. Yusko E, Vignali M, Wilson RK, Mardis ER, Hodi FS, Horak C, Chang H, Woods DM, Robins H, Weber J. Association of tumor microenvironment T-cell repertoire and mutational load with clinical outcome after sequential checkpoint blockade in melanoma. Cancer Immunol Res. 2019;7(3):458–65.
28. Melendez B, Van Campenhout C, Rorive S, Remmelink M, Salmon I, D'Haene N. Methods of measurement for tumor mutational burden in tumor tissue. Transl Lung Cancer Res. 2018;7(6):661–7.

Publisher's Note