Journal of Bioinformatics and Computational Biology
Vol. 22, No. 5 (2024) 2450022 (23 pages)
© World Scientific Publishing Europe Ltd.
DOI: 10.1142/S0219720024500227

SAKit: An all-in-one analysis pipeline for identifying novel proteins resulting from variant events at both large and small scales

Yan Li[*,¶], Boran Wang[†,‖], Zengding Wu[§,**], Shiliang Ji[‡,††], Shi Xu[§,‡‡] and Caiyi Fei[§,§§]

[*]*Department of Breast Surgery*
Peking Union Medical College Hospital, Peking Union Medical College
Chinese Academy of Medical Sciences, Beijing 100730, P. R. China

[†]*Beijing Tiantan Hospital, Capital Medical University*
Beijing 100070, P. R. China

[‡]*State Key Laboratory of Pharmaceutical Biotechnology*
School of Life Sciences, Nanjing University, Nanjing 210023, P. R. China

[§]*Department of AI and Bioinformatics*
Nanjing Chengshi Biopharmaceutical (TheraRNA) Co., Ltd.
Nanjing, P. R. China
[¶]*liyan0551@163.com*
[‖]*ttboran@163.com*
[**]*wuzengding@therarna.cn*
[††]*dg21300014@smail.nju.edu.cn*
[‡‡]*xushi@therarna.cn*
[§§]*feicaiyi@therarna.cn*

Received 11 September 2023
Revised 6 August 2024
Accepted 14 August 2024
Published 19 November 2024

Background: Genetic mutations that cause the inactivation or aberrant activation of essential proteins may trigger alterations or even dysfunctions in cellular signaling pathways, culminating in the development of precancerous lesions and cancer. Mutations and such dysfunctions can result in the generation of "novel proteins" that are not part of the conventional human proteome. Identification of these proteins carries a profound potential for unraveling promising drug targets and designing innovative therapeutic models. Despite the emergence of diverse tools for detecting DNA or RNA variants, facilitated by the widespread adoption of nucleotide sequencing technology, these methods primarily target point mutations and exhibit suboptimal performance in detecting large-scale and combinatorial mutations. Additionally, the outcomes of these tools are confined to the genome and transcriptome levels, and do not provide the corresponding protein information resulting from genetic alterations. *Results:* We present the development of Sequencing Analysis Kit (SAKit), a bioinformatics pipeline for hybrid sequencing analysis integrating long-read and short-read RNA sequencing data. Long reads are utilized for detecting large-scale variations such as gene fusions, exon skipping, intron retention, and aberrant expression in non-coding regions, owing to their excellent coverage capabilities.

[¶]Corresponding author.

Short reads serve to validate these findings at breakpoints and splice junctions. Conversely, short reads are employed for identifying small-scale variations, including single nucleotide variants, deletions, and insertions, due to their superior sequencing depth, with long reads providing additional validation. SAKit is designed to perform analyses using inter-species configuration files comprising genome references and annotation data, making it applicable to both human and mouse studies. Furthermore, SAKit implements a hierarchical filtering approach to eliminate low-confidence variants and employs open reading frame (ORF) analysis to translate identified variants into protein sequences. *Conclusion:* SAKit is a robust and versatile bioinformatics tool designed for the comprehensive identification of both large-scale and small-scale variants from RNA-seq data, facilitating the discovery of novel proteins. This pipeline integrates analysis of long-read and short-read sequencing data, offering a powerful solution for researchers in genomics and transcriptomics. SAKit is freely accessible and open-source, available through GitHub (https://github.com/therarna/SAKit) and as a Docker image https://hub.docker.com/repository/docker/therarna). Implemented primarily within a Snakemake framework using Python, SAKit ensures reproducibility, scalability, and ease of use for the scientific community.

Keywords: Genetic modifications; novel proteins; sequencing technology; large-scale variants; bioinformatic pipeline; SAKit.

1. Introduction

Numerous fundamental inquiries in the field of human disease, including cancer, are closely linked to the emergence of novel proteins resulting from base substitutions, deletions, insertions, frameshifts, intron retention, alternative splicing, and the translation of novel unannotated open reading frames (nuORFs) at the genetic and transcript levels.[1,2] Because novel proteins are not included in normal cells, they would not have immune tolerance, and could be easily recognized as foreign antigens. Novel proteins could be digested into short peptides, which are represented as T cell antigens by the host system, leading to consequential stimulation of cellular immune response and eradication of the cells expressing novel proteins.[3] Based on this principle, neoantigen-based immunotherapy has been developed by pharmaceutical companies or research institutions as a promising cancer treatment.[4–6] Additionally, novel proteins are also associated with other complex diseases, for instance, Alzheimer's,[7] type II diabetes,[8] liver disease,[9–14] kidney disease,[15–20] obesity,[21–25] and cardiovascular disease.[26] Therefore, effective identification of novel proteins is imperative for understanding human diseases and developing corresponding therapies.

The traditional methods of identifying and characterizing protein are the combination of tandem mass spectrometry (MS) with subsequent database searching. However, if MS relies on reference databases containing regular sequences in the canonical proteome, it would be almost impossible to identify proteins outside the canonical proteome, especially those produced by patient-specific individualized variants. Moreover, by benchmarking with known protein repertoire, Fricker[27] proposed that MS-based peptidomic approaches have other technical shortcomings: significantly lower detection rate in Cys-rich peptides, low detection sensitivity for peptides at low molecular weight (< 500 Da) or high molecular weight (> 3 kDa). Obviously, the bias derived from technical issues limits the application of MS in

identifying novel proteins. Therefore, new bioinformatic pipelines would be necessary for identifying novel proteins for immunotherapy.

Over the past decade, Next-Generation Sequencing (NGS) platforms have made significant advancements, enabling affordable sample-specific whole genome sequencing or total RNA sequencing. The technological improvements have accelerated the identification of the vast complexity of the human proteome. In 2012, Wang *et al.*[28] proposed a protein identification workflow that constructs hypothetical protein sequence reference databases derived from RNA-seq data to facilitate deciphering mass spectrometry (MS) data. Results from this study demonstrated that such sample-specific reference databases could significantly increase peptide identification sensitivity, reduce protein assembly ambiguity, and further enable the detection of novel peptides caused by highly variant regions. This NGS-driven proteogenomic strategy has been increasingly applied to characterize the protein repertoire in basic research or disease biology. However, despite these advancements, this method combining NGS and MS still faces several deficiencies.:

(1) NGS has inherent technical defects in detecting large-scale variation due to the short read-length (generally no more than 150 bp);
(2) The identification protocol for canonical proteins is often redundant, requiring detection by both mass spectrometry (MS) at the protein level and NGS at the RNA level. While this redundancy can benefit the characterization of smaller proteins,[29] it imposes a high sample input requirement to meet both detection protocols. This can pose a significant obstacle to translational medical research, particularly in cases where regenerative clinical samples are scarce and their failure to process may result in unacceptable consequences.
(3) In terms of the nucleotide sequencing-based bioinformatic analysis workflow, most of them only analyze conventional transcripts in general or mainly focus on single or few variation categories; in other words, by design principle they cannot cover all sample-specific variants. Besides the limitation of short read-lengths in NGS platform, several other factors make it challenging to develop an analysis workflow covering all types of variants relevant to novel proteins. We would explain these factors below:

(a) Various categories of variants could all contribute to the production of novel proteins, e.g., base substitution, frame-shift, chromosomal instability, etc. The detection of each category of variants requires a suitable analysis tool and a suitable dataset customized for specific detection purposes. These tools are usually interdependent or intertwined, thus a complete analysis process involving distinct analysis on various variants would be invariably time-consuming and labor-intensive;
(b) Calling and filtering that cover all variants often require massive computing power, especially when multiple samples are analyzed in parallel, and managing such time-consuming analysis steps is also a challenge;

(c) The calling results of each variant category are extremely redundant and glutted with noise, implying high proportions of false positive results. Therefore, further fine filtering and de-redundancy with proper parameters would be necessary for achieving accurate enrichment of true positive results.

Fortunately, the third-generation sequencing technology of the Pacbio platform, known as HiFi sequencing, encompasses the advantages of long read lengths and high accuracy. These attributes render it especially well-suited for identifying the large-scale variants. Leveraging the power of Pacbio HiFi reads, Miller *et al.*[30] developed a bioinformatics tool named PB_FLIP, tailored to identifying expressed fusion and long isoform variants.

While Pacbio long reads are adept at uncovering large-scare variants, the same cannot be said for small-scale variants such as single nucleotide variants, which tend to be susceptible to disruption due to random errors within a small length-based span. Therefore, it is recommended to synergize HIFI sequencing with NGS technology to improve detection accuracy, particularly in the context of single nucleotide variation (SNV).

To enhance the identification and discovery of novel proteins generated by genomic and transcriptomic variants, we have developed an innovative bioinformatics workflow called Sequencing Analysis Kit (SAKit). SAKit integrates High-Fidelity (HiFi) RNA long-read sequencing data with NGS RNA short-read data to identify and generate sample-specific hypothetical novel proteins. This hybrid approach leverages the strengths of both sequencing technologies, enabling a more comprehensive and accurate protein prediction process. The sequences comprising the hypothetical novel proteins are ultimately converted into FASTA format. We have validated SAKit using two different samples: a human brain reference RNA sample spiked with 2% SIRV-Set 4, and an allograft of MC38 cells subcutaneously injected into C57BL/6 mice. Our automated protein identification pipeline is user-friendly and demonstrates a robust capability to detect novel proteins arising from various genomic and transcriptomic variants, as evidenced by our results.

2. Methods

This section presents a comprehensive overview of the SAKit framework and its underlying logic, along with the inter-dependence among its various modules. We also provide details on the implementation of hierarchical filtering strategies in the different modules to effectively eliminate false hits. Finally, the study demonstrates the application of SAKit by analyzing samples from both human and mice, and presents the resulting analysis.

2.1. *Pipeline overview*

SAKit incorporates Snakemake as its workflow management system, enabling controlled and scalable pipeline execution in high-performance computing environments.

Snakemake, an efficient and versatile workflow management tool, offers numerous benefits for scientific data analysis, including superior memory management, robust portability, modular design, and reproducible results.

As a comprehensive bioinformatics pipeline, SAKit is compatible with various data types, including NGS and HiFi RNA-Sequencing data. It offers a range of analysis steps including quality control, variants calling, annotation and hierarchical filtering, and novel protein sequence generation. SAKit provides an extensive array of variants calling categories, such as insertion, deletion, frameshift, gene fusion, retained intron, and alternative splicing, to enable a comprehensive and detailed analysis of transcriptomic data. Those variant categories contribute to the identification of novel transcripts or proteins, which we refer to as Novel Variant Isoform (NVI) herein.

To utilize the SAKit pipeline, a subreads.bam file of the raw data generated by HiFi sequencing is required, along with a configuration file containing optional parameters and the necessary paths of reference and annotation files. If available, raw data generated by the NGS platform may also be included. Users can either use the default parameters or adjust them to suit their specific needs before the initial test run. After local deployment, users only need to specify the sample-specific raw data path for subsequent analyses. The SAKit pipeline encompasses over two dozen rules, which can be categorically classified into four distinct parts (Fig. 1):

(1) quality control and pre-processing of raw data,
(2) calling of NVI,
(3) hierarchical filtering and annotation for calling results,
(4) generating protein sequences form filtered NVI in fasta format and identifying novel proteins that are not present in the canonical proteome.

2.2. *Quality control and pre-processing of raw data*

We employed open-source software provided by PacBio for the pre-processing of our data. Specifically, we utilized the Circular Consensus Sequencing (CCS) software[31] to obtain high-accuracy single-molecule consensus HiFi reads. This process involves cutting "polymerase reads" that alternate between adapter and insert sequences to remove the adapters. As a results, we obtain multiple insert sequences referred to as "subreads". These subread sequences are generated through circular sequencing from the insert sequences. Subreads originating from the same source insert are largely identical, but sequencing errors introduce random errors among these subreads derived from identical insert sequences. To address this, these subreads are aligned with each other and subjected to a polishingstep. In general, as long as the circular sequencing is performed more than three times, meaning that you obtain three or more subread sequences, CCS polishing can result in HIFI reads with an accuracy of up to 99.9%. It is worth noting that the CCS step is the most time-consuming step of the entire pipeline.

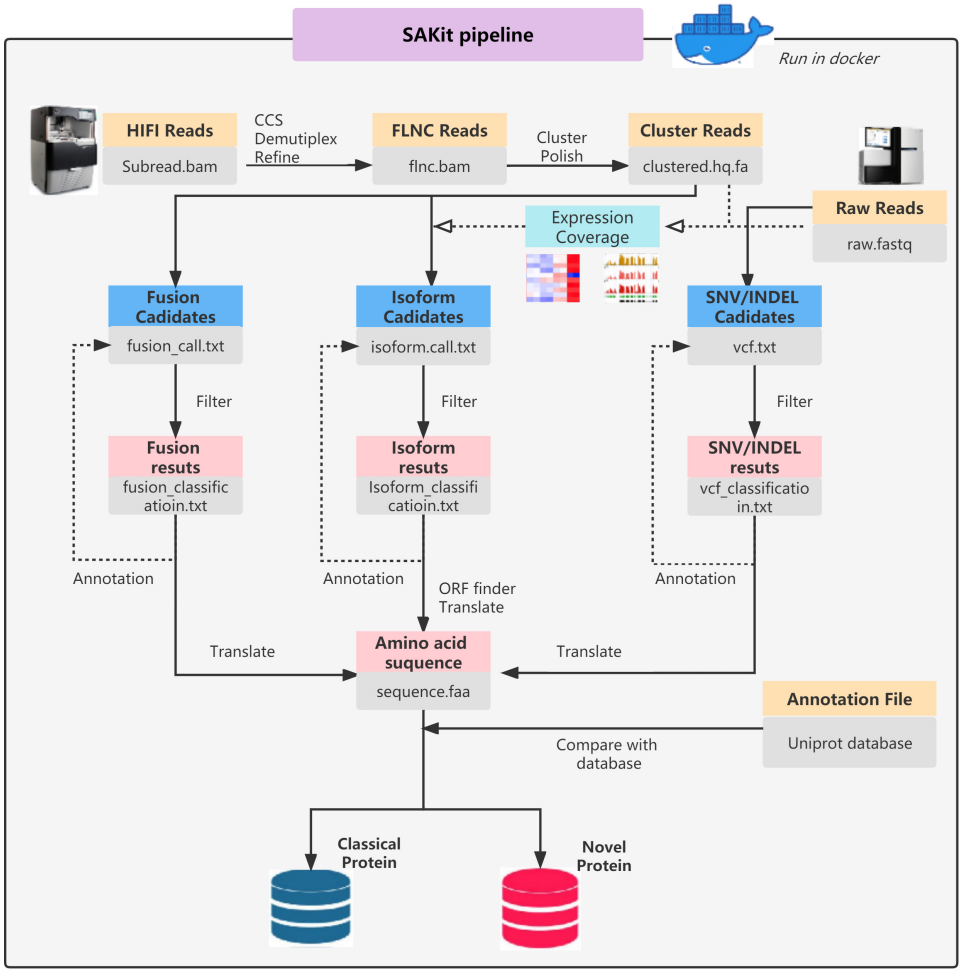


Fig. 1. Workflow of SAKit for all-in-one calling all novel variants and isoforms containing SNV, Insertion–deletion mutations (INDELs), gene fusion (FUSION), ISOFORM of intron-retention and alternative splicing simultaneously by comprehensive analysis of NGS and HiFi sequencing data. The intermediate steps of the analysis include multiple strict filtrations, and finally the result of low background noise is obtained. Finally, the converted amino acid sequences were compared with the protein database to select Novel or Classic protein.

After the generation of HiFi reads, we used the Lima software[32] to trim barcodes so that multiple samples constructed in the same library pool could be distinguished. This step can be omitted if barcodes are not employed when a library consists of only a single sample. Subsequently, we used the refine module of the Iso-Seq3 software[33] to trim polyA tails and adapter sequences that were wrapped on both sides of HiFi reads. This allowed us to obtain full-length non-chimeric (FLNC) sequences, while also removing residual concatemers. We aligned the obtained FLNC sequences to a genome reference, generating a BAM file. Simultaneously, the NGS data were also

aligned to the same genome reference, generating another BAM file. Both BAM files will be further utilized for small-scale variants analysis. Additionally, we employed the software of Iso-Seq3 to cluster FLNC reads, employing hierarchical n*log(n) alignment and iterative cluster merging. This enabled us to identify other larger-scale variants.

The pre-processing of data involves multiple processing steps and analytical modules, which may introduce abnormalities that propagate to subsequent analysis. Therefore, it is crucial to implement multidimensional quality control at each step to ensure downstream analysis and final results regarding novel proteins. For example, the number of pores used for single-molecule sequencing, referred to as Zero-Mode Waveguide (ZMWs), represents the data volume. Furthermore, the polished HIFI reads must meet the required accuracy, and the remaining FLNC reads after polyA tail removal. Additionally, the number of "uniq_mapped_reads" aligned to a unique position in the reference genome, "NonUniq_mapped_reads" aligning to multiple positions, and the count of "Unmapped_reads" that couldn't be aligned. Those quality control metrices can be employed for the analysis at each step and provide feedback on the analysis issues. If the amount of ZMWs is too low, it suggests that there wasn't enough input library for sequencing, prompting researchers to increase sample inputs to boost data amount.

In terms of library quality, we calculate the number and proportion of RIBOSOMAL_BASES, INTRONIC_BASES, INTERGENIC_BASES, and UTR_BASES using BAM files. A high proportion of INTERGENIC_BASES may indicate DNA contamination, while a high proportion of RIBOSOMAL_BASES suggests issues with ribosomal RNA removal, leading to reduced detection sensitivity. We also use the geneBody_coverage.py module in RSeQC software[34] to analyze the sequencing depth coverage at the 5′ and 3′ ends of the gene, obtaining MEDIAN_5PRIME_BIAS and MEDIAN_3PRIME_BIAS. A low MEDIAN_5PRIME_BIAS value indicates a significantly low coverage depth at the 5′ end of the transcript, potentially indicating severe library degradation. When combined with the RIN value, which assesses RNA quality, this analysis helps comprehensively determine whether re-sequencing is necessary.

Furthermore, we employ the infer_experiment.py module in RSeQC software to determine the proportions of reads aligned to the "+" and "−" strands in reference. A balanced proportion suggests non-strand-specific library construction, while a significantly imbalanced proportion indicates strand-specific library construction. Strand-specific libraries are useful for analyzing gene expression and isoform calling of two genes with overlap.

2.3. *Calling of novel variant isoform (NVI)*

In the NVI calling step following the pre-processing step described above, we classified all genetic variants that may result in novel proteins into two categories: small-scale variants (such as short insertions, deletions, duplications, substitutions, and small-scale gene fusions) and large-scale variants (such as large gene fusions, exon skipping of

transcripts, intron retention, RNA variable splicing, and abnormal expression of noncoding regions, as well as long insertions, deletions, and duplications) (Fig. 2).

To enable simultaneous analysis of these various scale variants, we integrated various variant analysis software into SAKit to meet the analysis requirements of various types of variants and enhance overall analysis performance. We utilized the characteristics of the data to optimize the analysis approach. HiFi reads were mainly used to analyze large-scale variants due to the advantage of their long-read sequence, while NGS reads with higher accuracy at the single-base level and easier acquisition of high-depth sequencing data were mainly used to analyze small-scale variants.

For HiFi reads, we employed the collapse_isoforms_by_sam.py module of cDNA_cupcake[35] to collapse identical transcript isoforms into a unique transcript isoform. The number of full-length reads corresponding to different unique isoforms were also quantified, which served as important supporting evidence for the called isoforms and were used as a parameter in the subsequent filter step. Gene fusions are also considered another isoform and important source of novel proteins or chimeric protein, and were called by utilizing fusion_finder.py of cDNA_cupcake. Notably, HiFi reads were also used to analyze small-scale variants. For instance, we employed PBsv software[36] to analyze various structural variants, including deletions of approximately 10 bp, which may also be present in the results of NGS variant analysis.

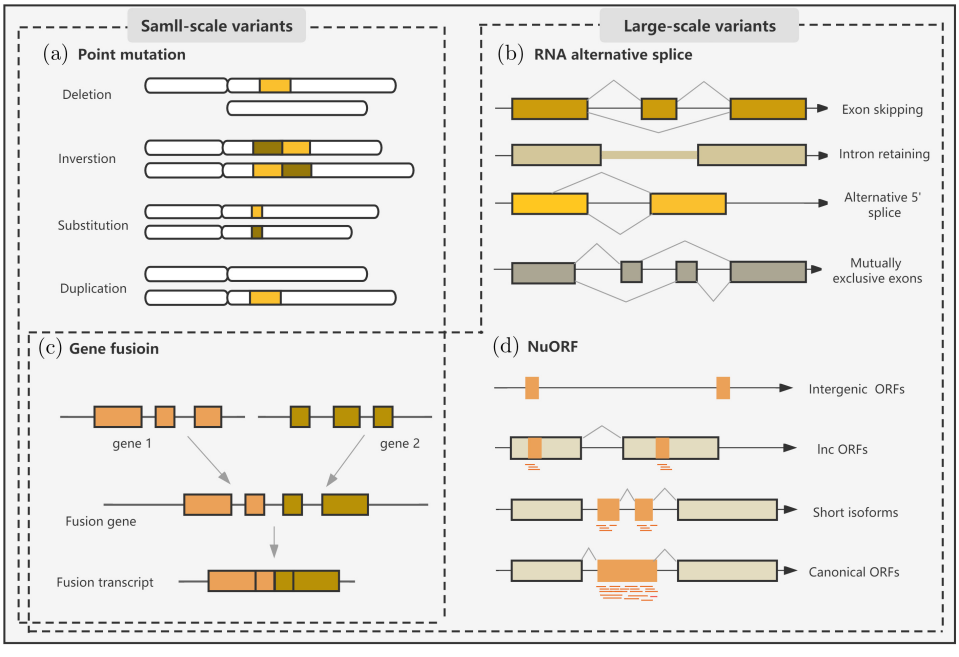


Fig. 2. Categories of NVI Calling. Small-scale variants, including (a) point mutations and (c) gene fusions, are analyzed using short-read sequencing reads. Large-scale variants, including (b) RNA alternative splicing, (c) gene fusions, and (d) NuORFs, are analyzed using long-read sequencing data. RNA fusions are analyzed using both long-read and short-read data.

Although this may appear redundant, it does not pose a significant threat to the analysis time. We removed redundancy from these variant results obtained from multiple software, which in turn increases the credibility of the results.

For NGS reads, we mainly used Vardict software[37] to generate VCF files that store all possible variants. Although Vardict software claims to call "everything," we mainly used it to analyze small-scale variants. We filtered out variants that do not produce novel proteins, such as single-base substitutions in non-coding regions or synonymous, nonsense, or silent mutations, although they may also be responsible for the occurrence of the disease. Large-scale variants analyzed based on NGS data were compared with the results of HiFi data analysis and integrated to further increase the reliability of the results.

2.4. *Hierarchical filtering*

2.4.1. *Filtering of isoform/fusion*

Since FUSION is also regarded as a specific type of isoform, its filtering process and steps are essentially the same as isoforms. To filter out called isoforms, there are typically four steps involved (Fig. 4).

The first step is to deplete isoforms below the pre-set threshold (default 2) by abundance as a high-quality isoform should be supported by at least two FLNC reads. Normally, the abundance of each called isoform could be directly quantified from PacBio HiFi sequencing data. Nonetheless if additional NGS sequencing data could be obtained from another aliquot of the same sample, it would provide an even more accurate quantification of isoforms' abundance.

The second step is to deplete the subset of isoforms (Fig. 3) that have the same exons at 3′ end as the longest one but are missing some of the 5′ exons, and those 5′ truncated transcripts are likely due to degradation of the 5′ end of RNA during library construction.[38] Moreover, the more degraded the library, the greater the number of transcripts filtered out due to 5′ truncation.

The third step is to deplete the "artificial" isoforms introduced by Chimerism in PCR, restriction enzyme ligation or sequencing error, etc. Such artificial isoforms are real signals called from the sequencing data, but they are noise to the samples and must be eradicated. Because a single feature could hardly identify random interference causing artificial isoforms, we employed the SQANTI3 tools[38] constructed using a machine learning approach based on random forests to achieve this goal. SQANTI3 extracts up to 47 features including coverage, exon coordinate, supported reads number, etc., and effectively implements filtering based on these features.

The last step is the annotation with the database. This step can identify meaningful isoforms/fusions associated with clinical diseases, as well as potential therapeutic targets. The detail and accuracy of gene annotation results are highly dependent on the quality and quantity of the annotation database. For human fusion gene annotations, we take the database of chimerDB4.0,[39] which summarizes nearly 100 k fusion candidates from patients in the TCGA project including tens of

thousands fusion genes with experimental evidence support. Conversely, the availability of suitable isoform annotation databases is still limited. Although there are some transcript-level isoform databases being developed, they are not comprehensive and are not user-friendly. It is anticipated that the advent of high-quality transcript annotation databases will enhance the clinical applicability of the identified isoforms.

2.4.2. *Filtering of SNV/INDEL*

There are also four steps to filter SNV/INDEL (Fig. 3).

The first step is to deplete the variant with sequencing depth below the confidence threshold. We calculate depth threshold using the formula below:

$$(\mathrm{VD})\mathrm{VariantDepth} = \frac{\mathrm{TPM} * \mathrm{total\_RPK}}{1000,000 * 100} * \mathrm{mean\_read\_length}.$$

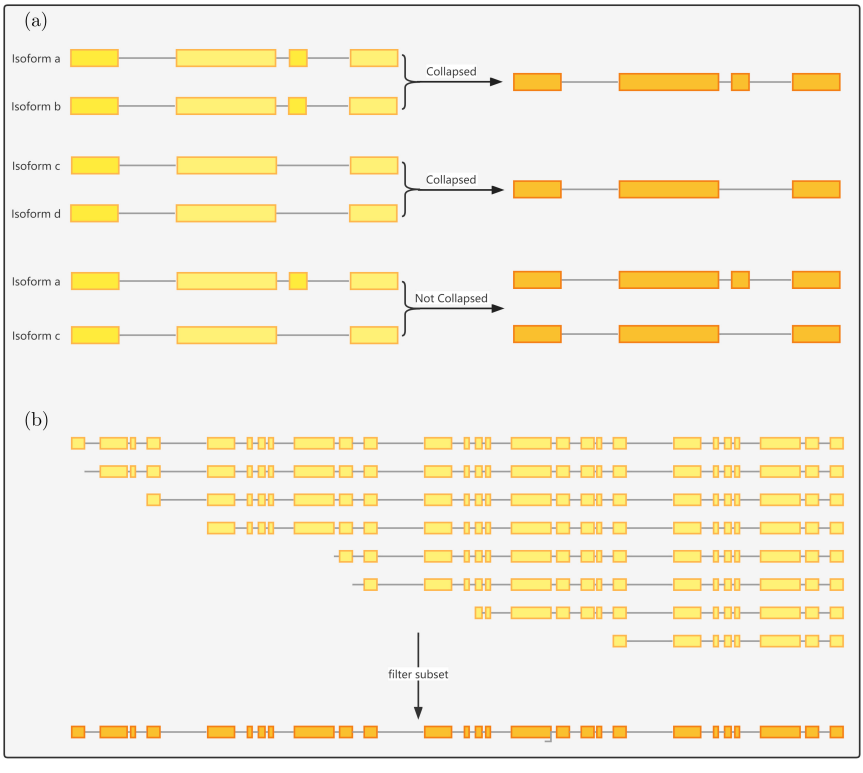


Fig. 3. (a) illustrates the process of collapsing the long read sequence. Because sequence a and b, and sequence c and d have exactly the same introns, they, respectively, collapse to get the only potential transcript candidates, while sequence a and c cannot further collapse because of their differences in introns. (b) indicates that many sequences cannot be aligned at the 5′ end, but are neat at the 3′ end. Such sequences are biologically likely to be of the same original transcript with the 5′ degradation, so they are collapsed into a unique transcript isoform.

RPK in the formula refers to Reads Per Kilobase, calculated by dividing the read counts by the length of each gene in kilobases.

TPM refers to Transcripts Per Kilobase Million, calculated by dividing the PRK values by the "per million" scaling factor.

Considering that the peptide translated by variant transcripts with a TPM higher than 35 would be considered to have immunogenicity,[40] we set the TPM to 35, and the threshold of Variant Depth can be calculated using the formula.

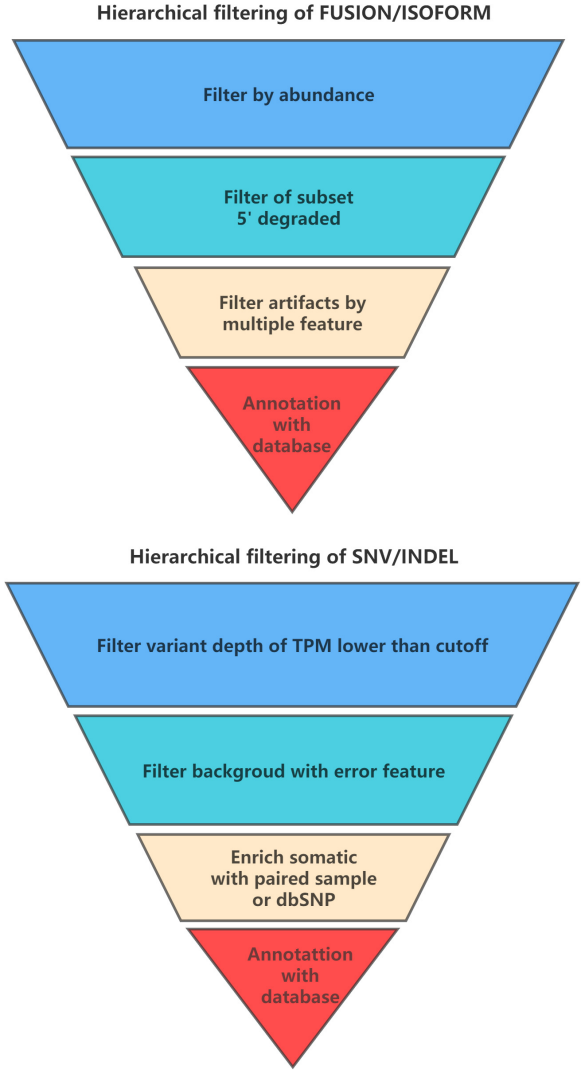


Fig. 4. Hierarchical filtering for preliminary calling results of all NVI. The inverted triangle represents the logic of filtering. Each layer represents one step filtering operation. The values on the left and right sides are the representative number of results obtained by a clinical sample after each filtering step.

The second step is to remove background noise. For this purpose, we mainly use the column of FILTER in the VCF file to remove variants with flags such as P8, Q10 and bias.

The third step is to deplete variants derived from single nucleotide polymorphism (SNP) or germline mutations by comprehensively analyzing the dbSNP annotation information and the allele frequency (AF).

2.5. *Novel protein generating*

2.6. *Dataset*

To demonstrate the workings of our in silico SAKit pipeline, we applied it to two distinct samples. The first sample was a human brain reference RNA sample spiked with 2% SIRV-Set 4 (Ref. 43) (https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR16762346&display=data-access), a Biologics Reference Standard that contains an artificial isoform set. The PacBio sequencing data for this sample were acquired in FLNC.read.bam format, which indicates that it has been pre-processed, from the NCBI. SIRV-Set 4 comprises 69 artificial transcript variants that simulate the splicing characteristics of seven human model gene loci to comprehensively reflect the diversity of alternative splicing, alternative transcription

start- and end-sites, overlapping genes, and antisense transcripts. Furthermore, SIRV-Set 4 includes 15 artificial transcript variants in five length categories (4 kb, 6 kb, 8 kb, 10 kb, and 12 kb, respectively) to mimic the length complexity of human transcripts.

The second sample was obtained in-house by subcutaneously injecting MC38 cell lines into C57BL/6 mice to form tumor allografts. After sacrificing the mice, the tumor allografts and paired normal small intestine tissue were collected and fragmented into small pieces of approximately 300 mg each. These tissue pieces were immediately washed with 1× PBS and placed in separate centrifuge tubes, transported at low temperature to GeneDenovo Co., Ltd. for RNA extraction, PacBio HiFi long-read sequencing, and Illumina short-read sequencing. After one month, the sequencing data were obtained and analyzed using the SAKit pipeline. Given the substantial differences in transcriptome and proteome between tumor allografts and normal tissues, we aimed to use the SAKit pipeline to identify mutations and aberrant transcript expression specific to the tumor allografts and obtain corresponding novel proteins.

3. Results and Discussion

SAKit analysis of the human brain reference RNA sample utilized PacBio HiFi long-read data, beginning with the alignment step. Prior to our analysis, these data had undergone pre-processing, specifically circular consensus sequence (CCS) polishing. Consequently, our workflow was initiated with the alignment of these pre-polished HiFi reads using MINIMAP2, thus bypassing the need for CCS polishing within our pipeline. As this sample lacked short-read data, we focused solely on analyzing large-scale variants, particularly different isoforms within the spike-in SIRV-Set 4, which are well-suited for long-read sequencing. This approach precluded the initiation of Vardict analysis or any subsequent small variant analyses that would typically require short-read data. In contrast, the MC38 allograft sample contained both PacBio HiFi long-read and Illumina short-read data, necessitating the analysis of both large-scale and small-scale mutation types. As both datasets were raw data, the analysis of MC38 allograft began with data pre-processing and quality control steps, and all analysis modules of SAKit were executed. SAKit ran on a high-performance computer cluster with an Intel(R) Xeon(R) Platinum processor (8269CY CPU, 3.10 GHz) and a Linux operating system of Ubuntu 20.04.4 LTS. Nearly one hour later, SAKit completed the analysis of the human brain reference RNA sample data, while the analysis of MC38 allograft data took nearly 5 h. By statistically analyzing the running time per module, we found that the Circular Consensus Sequencing (CCS) step was the most time-consuming. Notably, the human brain reference RNA sample, without CCS steps, saved a significant amount of time (Supplementary Figure 2).

In regard to the human brain reference RNA sample, we initially assessed the quality control metrics as presented in Supplementary Table 1. The total number of HiFi reads obtained was 1.9 million, representing almost half of the data capacity of a

SMRT Cell chip. The PCT_RIBOSOMAL_BASES metric was 0%, indicating a negligible presence of tRNA. The MEDIAN_5PRIME_BIAS and MEDIAN_3PRIME_BIAS values were 0.73 and 0.99, respectively, and these metrics, together with the gene body coverage curve (Supplementary Figure 1a), suggest that the RNA library was relatively intact and not significantly degraded. The "++−−" and "+−, −+" values were 0.859 and 0.0042, respectively, indicating that the library construction was strand-specific. These quality control metrics collectively demonstrate that the quantity and quality of the data met the required standards. Additional details of other quality control metrics can be found in Supplementary Table 1.

Since SIRV-Set 4 spiked in human brain reference RNA sample is a type of Biologics Reference Standards with the accuracy information (https://www.lexogen.com/store/sirv-set4) of isoforms in them, we confirmed the calling results of all true isoforms and calculated the recall ratio to determine whether the SAKit pipeline would erroneously miss the true positives. We achieved a 100% (12/12) recall ratio (Tables 1 and 2) for the 12 long transcripts of SIRV-set4. However, for the detection of the 69 short SIRV transcripts, we only achieved a recall ratio of 89.9% (62/69), missing seven isoforms. We investigated why those isoforms were missed case by case and found that the failure to detect these isoforms was attributed to the short length of their introns or exons, leading to a high mapping score in the global alignment and subsequent misclassification as a soft clip or deletion. Notably, statistics from the human RefSeq transcriptome indicate that the vast majority of exons and introns are longer than 20 bp. The exons and introns of the three misclassified isoforms are only 9 bp (Supplementary Figure 3a), 31 bp (Supplementary Figure 3b), and 20 bp (Supplementary Figure 3c), respectively, representing extreme cases. The absence of the other four missing isoforms in the raw data suggests they were lost during library construction or nucleotide sequencing.

For the MC38 tumor allografts, we also performed quality control evaluation first, which indicated that the data quantity and quality were sufficient for subsequent analysis. We detected numerous small-scale variants with high confidence using Vardict and PBsv on short-read data, while large-scale variants, such as intron retention, exon skipping, and abnormal expression in non-coding regions, were

Table 1. Short SIRV isoform of called and mis-called by SAKit.

Short SIRV				
Transcript category ID	Total number of transcripts	Recall number	Recall ratio	Missed
SIRV1	8	7	87.50%	SIRV105
SIRV2	6	6	100.00%	
SIRV3	11	10	90.90%	SIRV311
SIRV4	7	6	85.70%	SIRV404
SIRV5	12	10	83.30%	SIRV503, SIRV512
SIRV6	18	17	94.40%	SIRV618
SIRV7	7	6	85.70%	SIRV708
Summary	69	62	89.90%	

Table 2. Long SIRV isoform of called and mis-called by SAKit.

Long SIRV				
Transcript category ID	Total number of transcripts	Recall number	Recall ratio	Missed
SIRV10001	1	1	100.00%	/
SIRV10002	1	1	100.00%	/
SIRV10003	1	1	100.00%	/
SIRV12001	1	1	100.00%	/
SIRV12002	1	1	100.00%	/
SIRV12003	1	1	100.00%	/
SIRV4001	1	1	100.00%	/
SIRV4002	1	1	100.00%	/
SIRV4003	1	1	100.00%	/
SIRV6001	1	1	100.00%	/
SIRV6002	1	1	100.00%	/
SIRV6003	1	1	100.00%	/
SIRV8001	1	1	100.00%	/
SIRV8002	1	1	100.00%	/
SIRV8003	1	1	100.00%	/
Summary	15	15	100.00%	/

mainly detected through long-read HiFi data using cDNA_cupcake, including gene fusions identified through fusion_finder. These large-scale variants were integrated with the small-scale variants after SQANTI and hierarchical filtering and stored in a separate fasta file (Fig. 5(a)). Regarding large-scale genetic variants, we did not discern any conclusive evidence of gene fusions or structural alterations, which is in line with prior observations of the MC38 cell line. Nevertheless, we ascertained the existence of five transcript variants for a specific transcript isoform, substantiated by robust levels of supporting evidence. Four of these variants, named PB.349.2, PB.458.1, PB.1517.1, and PB.1528.1, were found and located in the Cep131, F2RL1, Gm5345, and KLHL26 gene locus in the mouse genome, respectively. All four variants expressed an additional segment as a novel exon in the intronic region (Supplementary Figure a–d). The remaining transcript variant, PB.1596.5, was located in an intergenic region of the mouse genome and was abnormally expressed as a novel transcript (Fig. 5(b)). All these variants were initially identified using HiFi long reads and subsequently validated by NGS short reads at their respective splice junctions, Importantly, none of these variants was detected in corresponding paired normal tissues, indicating their tumor-specificity. Furthermore, we generated the corresponding protein sequences (Fig. 5(c)) for the variants using GMST and performed homologous analysis with UniProt database tr_mouse_canon_isoform.fasta via BLAST, revealing that the identity value of the highest homologous protein for the first four variants, including PB.349.2, PB.458.1, PB.1517.1, and PB.1528.1, was less than 60%, indicating their novelty. However, PB.1596.5 exhibited a distinct pattern, as evidenced by the alignment result revealing that the identity value of the most homologous protein, Q1KYM0, was 99% (667/669) (Fig. 5(d)). As reported by Julien,[44] Q1KYM0 is an endogenous retrovirus (ERV) protein that is expressed in

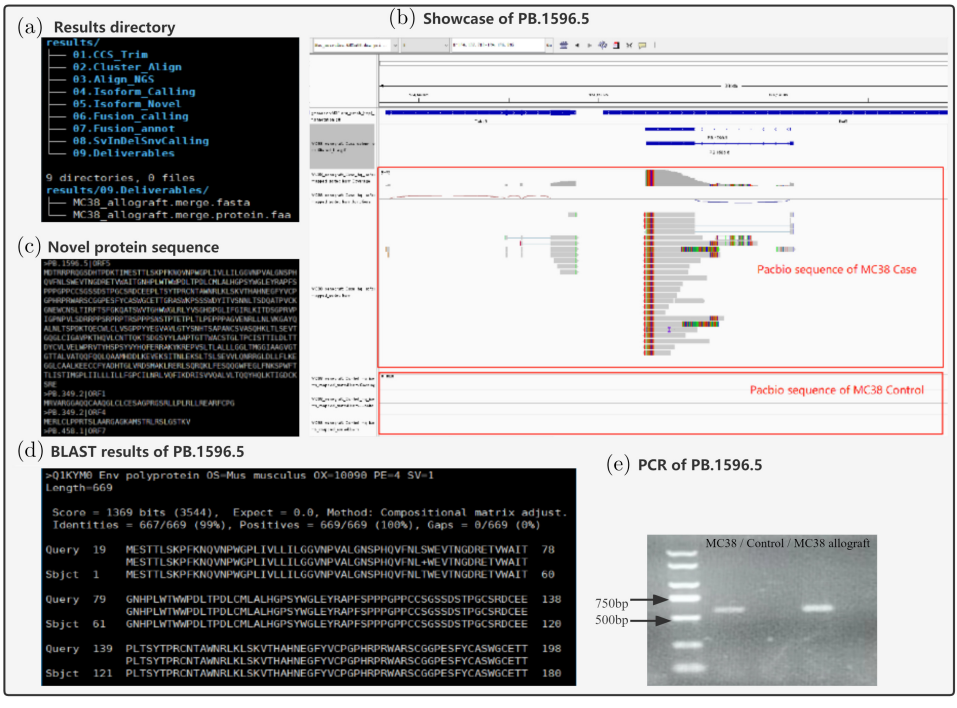


Fig. 5. Results of MC38 Analysis. (a) Analysis results displaying the nucleotide and corresponding protein sequences of the investigated variants under the Deliverables section. (b) Results of the PB1596.5 variant presented in IGV using the PacBio BAM file. (c) The final Novel protein sequence resulting from the integrated analysis. (d) Comparison of the protein sequence of the PB1596.5 variant to Q1KYM0 using BLAST. (e) Electrophoresis results of PB1596.5, arranged from left to right as Marker group, MC38 cell, Control group of normal tissue, and MC38 allograft.

Neuro-2a tumor cells and can cause immune evasion. Knocking down the gene, which expresses Q1KYM0, with siRNA can delay tumor growth and increase mouse survival. These findings strengthen the evidence supporting the tumor-specificity of PB.1596.5, and underscore the importance of combining large-scale and small-scale variant detection and subsequently novel protein analysis.

Finally, to further confirm the reliability of the identified variants results, we designed primers targeting PB.1596.5 and extracted RNA for PCR validation. The results showed significant overexpression in both MC38 cells and MC38 allografts, while the control normal tissues showed no expression (Fig. 5(e)). All results indicate that the SAKit pipeline can analyze all-in-one variants, and acquire high performance in identifying long isoforms.

4. Conclusions

The concept of "neoantigen" has been prevalent in the new era of cancer immunotherapy. Neoantigens could be defined as immunogenic proteins/epitopes derived

from aberrant genetic changes in tumor cells. Because normal cells do not have such aberrant genetic changes, neoantigens could be regarded as "absolutely" specific biomarkers of tumor cells. Therefore, in recent years, biotech companies and research institutions have been actively developing immunotherapies against neoantigens. Nowadays, how to effectively identify neoantigens from clinical samples has become a valuable question to be addressed for translational research.

The "classical" approach for identifying neoantigens is based on searching non-synonymous mutations in canonical ORFs. Such a "classical" approach is easily understandable, though its limitation is also apparent: only tumors with a high mutation burden have sufficient nonsynonymous mutations inside ORFs to induce novel proteins[45]; moreover, neoantigens with immunogenicity are only a tiny subset of novel proteins. Unfortunately, many cancer types have a low tumor mutation burden (<5 mutations/Mb), e.g., prostate cancer, glioblastoma, acute myeloid leukemia, etc. For these cancer types, it is barely possible to find sufficient neoantigens in canonical ORFs for immunotherapy.[46]

However, due to aberrant genetic changes, most types of cancer cells involve alternative RNA splicing that does not exist in normal cells.[47] RNA-seq analysis of The Cancer Genome Atlas data furnishes compelling evidence for the prevalence of tumor-specific alternative splicing events which produce neoantigens that are predicted to be more immunogenic than missense mutations,[48,49] moreover, some studies have shown that RNA alternative splicing provides another source of novel proteins, which are translated from the RNA sequences that are never translated in normal cells.[45,50,51] Therefore, identifying large-scale mutation-based novel proteins and neoantigens, such as RNA alternative splicing, would be an effective new approach to break the limitation of the "classical" approach, therefore greatly expanding the application of immunotherapies against neoantigens.

As a powerful tool for studying peptidomes or proteomes, MS theoretically could identify novel proteins derived from RNA alternative splicing. In fact, the performance of MS heavily relies on the characteristics of reference databases. As discussed in the introduction section, the canonical human proteome is a disadvantageous reference database for identifying novel proteins that do not belong to canonical human proteome at all. To solve the logical paradox above, the prevalent HiFi and NGS technology provide a scientifically sound method to establish the sample-specific reference database with hypothetical protein sequences derived from RNA translation.

Achieving comprehensive detection of all variants that have the potential to induce protein alterations is a key necessity when analyzing RNA sequencing data. In this regard, multiple mechanisms, including fusion, SNP, and INDEL, contribute to the generation of novel RNA transcripts, as well as the subsequent induction of novel proteins. Given the complexity of the task at hand, a powerful bioinformatics tool is required to process RNA sequencing data and cover all types of RNA transcript variants.

Taking this into consideration, we have developed the SAKit pipeline, which provides a one-key running solution to analyze HiFi and NGS RNA-seq data comprehensively. By far, the SAKit has the broadest coverage of all variants that can induce potential novel proteins than published workflows. Additionally, we integrated and converted all types of variants into fasta-format protein sequences to ensure convenience for users to use for subsequent analysis, such as protein structure prediction, neoantigen immune epitope analysis, among other things.

Our pipeline's efficiency and computational capabilities make it suitable for processing high-throughput data, notwithstanding limited computational power and time consumption. In conclusion, we have developed a user-friendly, comprehensive SAKit pipeline capable of detecting all variants that have the potential to induce protein alterations, thereby providing an all-in-one solution for the identification of novel proteins and neoantigens, which have clinical applications for immunotherapy purposes.

Availability of Data and Materials

The human brain reference RNA sample spiked with 2% SIRV-Set 4 can be accessed at NCBI Trace (https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR16762346&display=data-access). Additionally, the MC38 tumor allografts dataset, which includes both PacBio long reads and NGS short reads, is available in the NCBI Sequence Read Archive (SRA) under project PRJAN1136964, with accession numbers SRR29872125, SRR29872126, SRR29872127, and SRR29872128.

Competing Interests

The authors declare no competing interests.

Authors' Contributions

BW as SJ performed the experiments, ZW and FC built up the pipeline. YL and SX prepared the paper, YL reviewed this paper. All authors read and approved the final paper.

Acknowledgments

This study was supported by the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2023-RW320-07) and National Natural Science Foundation of China (823B2095 (SJ)). Yan Li, Boran Wang and Zengding Wu have contributed equally to this work and share first authorship.

Supplementary Information

The Supplementary Information are available at: https://www.worldscientific.com/doi/suppl/10.1142/S0219720024500227

ORCID

Yan Li https://orcid.org/0000-0003-0316-7810
Boran Wang https://orcid.org/0009-0005-2879-7212
Zengding Wu https://orcid.org/0009-0004-7989-1164
Shiliang Ji https://orcid.org/0000-0002-0393-1324
Shi Xu https://orcid.org/0000-0002-8815-9707
Caiyi Fei https://orcid.org/0000-0001-8726-7063

References

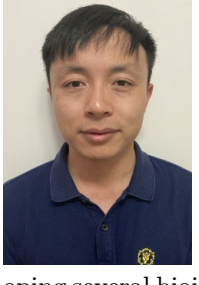
1. Walley JW, Briggs SP, Dual use of peptide mass spectra: Protein atlas and genome annotation, *Curr Plant Biol* **2**:21–24, 2015, https://doi.org/10.1016/j.cpb.2015.02.001.
2. Sheynkman GM, Shortreed MR, Cesnik AJ, Smith LM, Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation, *Annu Rev Anal Chem Palo Alto Calif* **9**:521–545, 2016, https://doi.org/10.1146/annurev-anchem-071015-041722.
3. Yarchoan M *et al.*, Targeting neoantigens to augment antitumour immunity, *Nat Rev Cancer* **17**:569, 2017, https://doi.org/10.1038/nrc.2017.74.
4. Tanyi JL *et al.*, Personalized cancer vaccine effectively mobilizes antitumor T cell immunity in ovarian cancer, *Sci Transl Med* **10**:eaao5931, 2018, https://doi.org/10.1126/scitranslmed.aao5931.
5. Sahin U *et al.*, Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer, *Nature* **547**:222–226, 2017, https://doi.org/10.1038/nature23003.
6. Hilf N *et al.*, Actively personalized vaccination trial for newly diagnosed glioblastoma, *Nature* **565**:240–245, 2019, https://doi.org/10.1038/s41586-018-0810-y.
7. Bai B *et al.*, Proteomic landscape of Alzheimer's Disease: novel insights into pathogenesis and biomarker discovery, *Mol Neurodegener* **16**:55, 2021, https://doi.org/10.1186/s13024-021-00474-z.
8. DeFronzo RA *et al.*, Type 2 diabetes mellitus, *Nat Rev Dis Primer* **1**:15019, 2015, https://doi.org/10.1038/nrdp.2015.19.
9. Shi X *et al.*, MxA is a positive regulator of type I IFN signaling in HCV infection, *J Med Virol* **89**:2173–2180, 2017, https://doi.org/10.1002/jmv.24867.
10. Jiao B *et al.*, Insulin receptor substrate-4 interacts with ubiquitin-specific protease 18 to activate the Jak/STAT signaling pathway, *Oncotarget* **8**:105923–105935, 2017, https://doi.org/10.18632/oncotarget.22510.
11. Duan X *et al.*, MicroRNA 130a regulates both hepatitis C virus and hepatitis B virus replication through a central metabolic pathway, *J Virol* **92**:e02009-17, 2018, https://doi.org/10.1128/JVI.02009-17.
12. Chen Y *et al.*, ISG12a inhibits HCV replication and potentiates the anti-HCV activity of IFN-α through activation of the Jak/STAT signaling pathway independent of autophagy and apoptosis, *Virus Res* **227**:231–239, 2017, https://doi.org/10.1016/j.virusres.2016.10.013.
13. Li Y *et al.*, Interferon-stimulated gene 15 conjugation stimulates hepatitis B virus production independent of type I interferon signaling pathway *in vitro*, *Mediators Inflamm* **2016**:7417648, 2016, https://doi.org/10.1155/2016/7417648.

14. Qu L-L *et al.*, Gastrodin ameliorates oxidative stress and proinflammatory response in nonalcoholic fatty liver disease through the AMPK/Nrf2 pathway, *Phytother Res* **30**:402–411, 2016, https://doi.org/10.1002/ptr.5541.
15. Jiao B *et al.*, Pharmacological inhibition of STAT6 ameliorates myeloid fibroblast activation and alternative macrophage polarization in renal fibrosis, *Front Immunol* **12**:735014, 2021, https://doi.org/10.3389/fimmu.2021.735014.
16. Wen J, Jiao B, Tran M, Wang Y, Pharmacological inhibition of s100a4 attenuates fibroblast activation and renal fibrosis, *Cells* **11**:2762, 2022, https://doi.org/10.3390/cells11172762.
17. An C *et al.*, Myeloid PTEN deficiency aggravates renal inflammation and fibrosis in angiotensin II-induced hypertension, *J Cell Physiol* **237**:983–991, 2022, https://doi.org/10.1002/jcp.30574.
18. Wang X *et al.*, QiDiTangShen granules activate renal nutrient-sensing associated autophagy in db/db mice, *Front Physiol* **10**:1224, 2019. https://doi.org/10.3389/fphys.2019.01224.
19. Jin X *et al.*, AMP-activated protein kinase contributes to cisplatin-induced renal epithelial cell apoptosis and acute kidney injury, *Am J Physiol Renal Physiol* **319**:F1073–F1080, 2020, https://doi.org/10.1152/ajprenal.00354.2020.
20. Jiao B *et al.*, STAT6 deficiency attenuates myeloid fibroblast activation and macrophage polarization in experimental folic acid nephropathy, *Cells* **10**:3057, 2021, https://doi.org/10.3390/cells10113057.
21. Li C *et al.*, MicroRNA regulated macrophage activation in obesity, *J Transl Intern Med* **7**:46–52, 2019, https://doi.org/10.2478/jtim-2019-0011.
22. Karlinsey K, Qu L, Matz AJ, Zhou B, A novel strategy to dissect multifaceted macrophage function in human diseases, *J Leukoc Biol* **112**:1535–1542, 2022, https://doi.org/10.1002/JLB.6MR0522-685R.
23. Qu L *et al.*, Macrophages at the crossroad of meta-inflammation and inflammaging, *Genes* **13**:2074, 2022, https://doi.org/10.3390/genes13112074.
24. Matz A, Qu L, Karlinsey K, Zhou B, Impact of microRNA regulated macrophage actions on adipose tissue function in obesity, *Cells* **11**:1336, 2022, https://doi.org/10.3390/cells11081336.
25. Matz AJ, Qu L, Karlinsey K, Zhou B, MicroRNA-regulated B cells in obesity, *Immunometabolism Cobham Surrey Engl* **4**:e00005, 2022, https://doi.org/10.1097/IN9.0000000000000005.
26. Li C *et al.*, Atherospectrum reveals novel macrophage foam cell gene signatures associated with atherosclerotic cardiovascular disease risk, *Circulation* **145**:206–218, 2022, https://doi.org/10.1161/CIRCULATIONAHA.121.054285.
27. Fricker LD, Limitations of mass spectrometry-based peptidomic approaches, *J Am Soc Mass Spectrom* **26**:1981–1991, 2015, https://doi.org/10.1007/s13361-015-1231-x.
28. Wang X *et al.*, Protein identification using customized protein sequence databases derived from RNA-Seq data, *J Proteome Res* **11**:1009–1017, 2012, https://doi.org/10.1021/pr200766z.
29. Qiu X *et al.*, Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning, *Appl Environ Microbiol* **67**:880–887, 2001, https://doi.org/10.1128/AEM.67.2.880-887.2001.
30. Miller AR *et al.*, Pacific biosciences fusion and long isoform pipeline for cancer transcriptome–based resolution of isoform complexity, *J Mol Diagn* **24**:1292–1306, 2022, https://doi.org/10.1016/j.jmoldx.2022.09.003.
31. CCS Home. in CCS Docs. Available at https://ccs.how/ [accessed on 20 March 2023].
32. Lima Home. in Lima Docs. Available at https://lima.how/ [accessed on 20 March 2023].

33. PacificBiosciences/IsoSeq, 2023.
34. Wang L, Wang S, Li W, RSeQC: quality control of RNA-seq experiments, *Bioinforma Oxf Engl* **28**:2184–2185, 2012, https://doi.org/10.1093/bioinformatics/bts356.
35. Tseng E, cDNA_Cupcake, 2023.
36. PacificBiosciences/pbsv, 2023.
37. Lai Z *et al.*, VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research, *Nucl Acids Res* **44**:e108, 2016, https://doi.org/10.1093/nar/gkw227.
38. Tardaguila M *et al.*, SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification, *Genome Res* **28**:396–411, 2018, https://doi.org/10.1101/gr.222976.117.
39. Jang YE *et al.*, ChimerDB 4.0: an updated and expanded database of fusion genes, *Nucl. Acids Res.* **48**:D817–D824, 2020, https://doi.org/10.1093/nar/gkz1013.
40. Wells DK *et al.*, Key Parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction, *Cell* **183**:818–834.e13, 2020, https://doi.org/10.1016/j.cell.2020.09.015.
41. Lomsadze A, Gemayel K, Tang S, Borodovsky M, Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes, *Genome Res.* **28**:1079–1089, 2018, https://doi.org/10.1101/gr.230615.117.
42. McGinnis S, Madden TL, BLAST: At the core of a powerful and diverse set of sequence analysis tools, *Nucl. Acids Res.* **32**:W20–W25, 2004, https://doi.org/10.1093/nar/gkh435.
43. Paul L *et al.*, SIRVs: Spike-in RNA variants as external isoform controls in RNA-sequencing, 080747, 2016.
44. Pothlichet J, Heidmann T, Mangeney M, A recombinant endogenous retrovirus amplified in a mouse neuroblastoma is involved in tumor growth in vivo, *Int J Cancer* **119**:815–822, 2006, https://doi.org/10.1002/ijc.21935.
45. Laumont CM *et al.*, Noncoding regions are the main source of targetable tumor-specific antigens, *Sci Transl Med* **10**:eaau5516, 2018, https://doi.org/10.1126/scitranslmed.aau5516.
46. Capietto A-H, Hoshyar R, Delamarre L, Sources of cancer neoantigens beyond single-nucleotide variants, *Int J Mol Sci* **23**:10131, 2022, https://doi.org/10.3390/ijms231710131.
47. Zhang Y, Qian J, Gu C, Yang Y, Alternative splicing and cancer: a systematic review, *Signal Transduct Target Ther* **6**:78, 2021, https://doi.org/10.1038/s41392-021-00486-7.
48. Jayasinghe RG *et al.*, Systematic analysis of splice-site-creating mutations in cancer, *Cell Rep* **23**:270–281.e3, 2018, https://doi.org/10.1016/j.celrep.2018.03.052.
49. Kahles A *et al.*, Comprehensive analysis of alternative splicing across tumors from 8,705 patients, *Cancer Cell* **34**:211-224.e6, 2018, https://doi.org/10.1016/j.ccell.2018.07.001.
50. Hoyos LE, Abdel-Wahab O, Cancer-specific splicing changes and the potential for splicing-derived neoantigens, *Cancer Cell* **34**:181–183, 2018, https://doi.org/10.1016/j.ccell.2018.07.008.
51. Smart AC *et al.*, Intron retention is a source of neoepitopes in cancer, *Nat. Biotechnol.* **36**:1056–1058, 2018, https://doi.org/10.1038/nbt.4239.

Yan Li received his doctor's degree in Clinical Medicine at Peking Union Medical College. Currently, he is working as a breast surgeon/oncologist at the Department of Breast Surgery, Peking Union Medical College Hospital, Beijing, China. His research interests mainly focus on breast cancer, genetic diagnosis, genetic treatment, and artificial intelligence.

Boran Wang received his bachelor's degree in Clinical Medicine and his master's degree in Epidemiology and Health Statistics, with a research focus on health management. Currently, he is pursuing his PhD degree in Computer Science and Technology at the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China. He is currently employed at the Beijing Tiantan Hospital affiliated with the Capital Medical University, Beijing, China, working in the field of medical management. His research interests mainly focus on artificial intelligence (natural language processing, smart healthcare, and auxiliary diagnosis of nervous system diseases). He also serves as Executive Member of the Adult Education Committee of the China Medical Education Association (CMEA).



Zengding Wu received his B.S. degree in Biotechnology from the National Huaqiao University in 2011 and his M.S. degree in Biochemistry and Molecular Biology from the Kunming University of Science and Technology in 2014. During his internship at BGI, he gained valuable experience in cancer genomics data analysis, where he honed his bioinformatics skills. After completing his graduate studies, Zengding Wu worked as a bioinformatician at a gene sequencing company, where he played a key role in developing several bioinformatics pipelines and software for cancer mutation detection. He is currently the Deputy Director of the Intelligent Bioinformatics Department at TheraRNA Co., Ltd., where his work focuses on leveraging bioinformatics to support drug development and regulatory science.

Shi-Liang Ji received his B.S. degree from the Xinxiang Medical University at Xinxiang City, an M.S. degree from the Soochow University, and an M.D. degree from the Nanjing University. He began his academic career through the Licensed Pharmacist at the Suzhou Hospital, Affiliated Hospital of Medical School, Nanjing University. His research focuses on the preparation and application of personalized tumor mRNA vaccines. His research shows that tumor vaccines offer a gentler, personalized treatment for cancer patients, with the promise of a better quality of life for patients.

Shi Xu received his Ph.D in Pharmaceutical Sciences and M.S in Regulatory Science both from the University of Southern California in 2013. Xu is the founder and CEO of TheraRNA Co., Ltd., where he spearheads the company's innovative drug discovery initiatives. His expertise covers the development of advanced mRNA therapeutics and delivery systems, especially lipid nanoparticles. Xu has been listed as the author of many scientific publications, and he is also the inventor of several patents for his pioneering contributions. His research interests are deeply rooted in pharmaceutical sciences and immunology, with an emphasis on enhancing efficacy and safety of mRNA vaccines methodologies. Under his leadership, TheraRNA continues to advance the development of cutting-edge therapeutic solutions, at the forefront of pharmaceutical innovation.

Caiyi Fei received his B.Eng. degree in Thermal Engineering from the Nanjing University of Science and Technology in 2006 and his Ph.D. in Bioinformatics from the University of Science and Technology of China in 2014. Fei is Co-Founder and Vice President of TheraRNA Co., Ltd., where he leads the AI & Bioinformatics Department. His research is centered on the integration and analysis of high-throughput cancer "omics" data using computational and machine learning methods. Fei's work focuses on understanding and overcoming resistance mechanisms in cancer treatment, with a particular emphasis on developing multimodal deep learning models for HLA antigen presentation and antigen immunogenicity prediction systems. His research has led to multiple patents and publications in international journals. Fei's current research interests lie in bioinformatics and cancer systems biology, where he strives to seamlessly integrate computational and statistical methods to advance experimental and clinical cancer research.