Computational Molecular Biology 2025, Vol.15, No.4, 208-217
http://bioscipublisher.com/index.php/cmb

Research Insight Open Access

Standardizing Bioinformatics Pipelines for Clinical Genomics

Yuhong Huang, Yufen Wang, Guangman Xu
Traditional Chinese Medicine Research Center, Cuixi Academy of Biotechnology, Zhuji, 311800, China
Corresponding author: guangman.xu@cuixi.org

Computational Molecular Biology, 2025, Vol.15, No.4 doi: 10.5376/cmb.2025.15.0020
Received: 18 Jun., 2025
Accepted: 29 Jul., 2025
Published: 21 Aug., 2025

Preferred citation for this article:
Huang Y.H., Wang Y.F., and Xu G.M., 2025, Standardizing bioinformatics pipelines for clinical genomics, Computational Molecular Biology, 15(4): 208-217 (doi: 10.5376/cmb.2025.15.0020)

Abstract High-throughput sequencing technology has been widely adopted in clinical genomics for the diagnosis of genetic diseases and personalized treatment of tumors. However, the differences in bioinformatics analysis processes among various laboratories may lead to inconsistent variant detection results, affecting clinical interpretation and data sharing. Based on the research on the standardization of bioinformatics processes, this article analyzes the common data analysis processes in clinical genomics, the key steps and tools involved in each link, and clarifies the necessity and challenges of process standardization. We further explored the technical strategies for achieving standardization, including the adoption of workflow management systems, containerization technologies, unified reference standards, and quality control verification schemes, and introduced relevant domestic and international standards, norms, and application practices. The results show that standardized bioinformatics processes help improve the accuracy and repeatability of variant detection, ensure the comparability of results from different laboratories, and meet clinical diagnostic norms and regulatory requirements. This work provides a reference for the standardization of the clinical genomics student information analysis process and can promote the reliable application of sequencing data in clinical practice.

Keywords Clinical; Genomics; Bioinformatics analysis process; Standardization repeatability; Quality control

1 Introduction

The rise of high-throughput sequencing (NGS) technology has brought the diagnosis and treatment of genetic diseases into a new stage. Nowadays, whole exome sequencing and whole genome sequencing have almost become standard equipment in clinical testing. People can quickly identify potential pathogenic mutations through them, providing important decision-making basis for doctors. To standardize the interpretation of variations, there have long been mature guidelines internationally, such as the variation classification standard of ACMG. These guidelines themselves do not directly determine the results but rely on the list of candidate variations screened out by the bioinformatics analysis process - once this list is inaccurate, the subsequent interpretation will lose its basis (Lavrichenko et al., 2025).

The problem lies in the fact that the analysis processes in different laboratories vary greatly. Roy et al. (2018) pointed out that there is no unified standard in this regard in the industry at present. Even when processing the same batch of data, the consistency rate of detection results among different pipelines is not high. The coincidence rate of single nucleotide variations is approximately 60%, while the consistency rate of insertion and deletion variations is less than 30%. Researchers further found that approximately 16.5% of clinically relevant variations were identified by only one algorithm, which implies that some key pathogenic mutations might have been overlooked by other processes. It can be seen from this that if the analysis process is not standardized or not fully verified, it may eventually lead to incorrect diagnostic conclusions and have a negative impact on the treatment decisions of patients (Weißbach et al., 2021).

From a broader perspective, the standardization of bioinformatics processes is not merely a technical issue; it determines the credibility of the entire clinical genomics outcome. Errors in a laboratory may be magnified in multi-center studies or variant sharing, affecting the comparability of data. Therefore, promoting the standardization of the analysis process has become a key task at present. This study focuses on this issue, systematically analyzes the structure and deficiencies of the existing clinical sequencing process, discusses the

key technologies and practical strategies in the standardization process, and combines the latest guidelines and consensuses at home and abroad to propose feasible optimization ideas, hoping to provide reference for improving the consistency and reliability of clinical sequencing analysis.

2 An Overview of the Clinical Genomics Bioinformatics Analysis Process

2.1 Quality control and preprocessing of raw data

The first step of bioinformatics analysis usually begins with the FASTQ files output by the sequencer, which record the base sequence and quality value of each read length. The original data may seem complete, but not all of it can be directly used for downstream analysis. During sequencing, contamination from adapters, base-calling errors, or "noise" reads are inevitable; if left unfiltered, these artifacts may distort the subsequent analyses (Hao et al., 2022; Chen, 2023). Therefore, researchers usually perform quality control first: removing reads with excessive "N" bases, reads with a high proportion of low-quality bases, or directly trimming contaminated sections. Tools like FastQC can quickly visualize quality distributions and trigger filtering/trimming actions (Hao et al., 2022). These cleaned high-quality reads serve as the foundation for reliable variant detection downstream.

2.2 Sequence alignment

Next comes sequence alignment. The quality-controlled reads must find their best match positions in the reference genome. While it may sound straightforward, alignment is one of the most critical and algorithmically challenging steps in the whole pipeline. Well-known aligners such as BWA-MEM, Bowtie2 and NOVOAlign can all perform the task, and among them BWA-MEM remains widespread due to its balance of speed and accuracy (Alganmi et al., 2023). It uses the Burrows-Wheeler transform to enable efficient mapping. The alignment output is typically stored in SAM/BAM format, detailing each read's genomic location and match status. To reduce systematic bias between batches, many clinical labs adopt the same version of the reference genome and fix alignment parameters.

2.3 Variant detection

Only after alignment are researchers in a position to detect variants. Variant callers infer differences between the sample genome and the reference by analyzing mismatches, insertions, and deletions in the alignment data. Tools like GATK HaplotypeCaller, FreeBayes and SAMtools/BCFtools are commonly used; among them, GATK's algorithm-employing local re-assembly and Bayesian modeling-is considered a "gold standard" for small variants (Wilton and Szalay, 2023). However, for more complex structural variants (SV) or copy number variants (CNV), these tools alone are insufficient: specialized callers such as Manta, CNVnator or BreakDancer are often required (Minoche et al., 2021). Different algorithms have varying sensitivity, so in practice multiple tools are often used in cross-validation. Ultimately, variant information is saved in a VCF (Variant Call Format) file-listing each variant's coordinate, type, allele frequency and more-a format widely supported across tools and databases.

3 The Necessity of Standardizing Bioinformatics Processes

3.1 Enhance the accuracy and repeatability of the results

The reason why standardization was first proposed in clinical genomic analysis is actually to make the results more accurate and repeatable. False negatives or false positives brought by NGS analysis, once they occur in the diagnostic stage, may directly affect the treatment choices of patients (Roy et al., 2018).

However, in reality, the analytical habits of different laboratories vary greatly. The software versions and parameter settings are all different, and the results of the same sample data running in different places may not be consistent. Studies have pointed out that in multi-laboratory comparisons, the SNV detection results of the same sample by different pipelines are only 50% to 60% consistent, and the consistency rate of Indel is even lower (Samarakoon et al., 2025). This difference is not only caused by the algorithm itself, but also by multiple details such as alignment methods, filtering thresholds, and judgment criteria. If the entire industry can adopt a unified "best practice" process, these human differences will be significantly reduced, so that the conclusions of the same data can remain consistent regardless of which laboratory is analyzed (Koboldt, 2020). On the other hand, standardization can also enhance the stability within the laboratory. Once a validated process is fixed, the results

should remain highly consistent even if the analysts are changed or the same batch of data is re-analyzed at intervals. This consistency is particularly important for clinical cases that require long-term follow-up or multiple re-examinations.

3.2 Promote the comparability of results and data sharing among laboratories

Today, with the rapid development of precision medicine, data sharing has become an important means to promote new discoveries. However, if different laboratories each use different analytical processes, it is like speaking different "languages", and deviations are very likely to occur when directly communicating results. Researchers found that patient variant sets generated by different institutions are often not directly compatible. Some mutations that are reported as pathogenic in one laboratory may be completely undetectable by changed processes. This situation can seriously hinder cross-institutional data integration. To solve this problem, standardization becomes particularly crucial. Unified reference versions, quality control standards and variation interpretation rules can make results among different laboratories comparable (Brancato et al., 2024). This consistency not only helps to establish a large clinical variant database, but also facilitates international cooperation. Only when the way data is produced is consistent can the frequency and effect of variants in different populations or diseases be truly compared. In addition, the standardized process also provides a foundation for external quality assessment (EQA). Many quality assessment projects require laboratories to analyze the uniformly provided data according to the same standard in order to determine the source of differences. If the processes are different, quality evaluation loses its meaning (Cherney et al., 2024).

3.3 Comply with clinical regulations and certification requirements

At the regulatory level, the standardization of bioinformatics processes is also an unavoidable requirement. Clinical genomics testing falls within the scope of medical diagnosis, and regulatory authorities have clear requirements for its accuracy and consistency. Medical laboratory quality standards such as ISO 15189 have put forward specific norms for high-throughput sequencing processes including bioinformatics analysis (Haanpää et al., 2025). These specifications require laboratories to establish validated standard operating procedures (SOPs) that are traceable and controllable from sample processing to data analysis. In the field of molecular diagnostics, societies such as AMP and CAP have also issued relevant guidelines, emphasizing that laboratories should validate the performance of NGS analysis pipelines, evaluate the sensitivity, specificity and repeatability of the tests, and re-validate after software updates or parameter adjustments (Jennings et al., 2017; Samarakoon et al., 2025). Only laboratories that meet these requirements can obtain certification, and their test reports will be accepted by clinicians. At the same time, standardization also enables regulatory authorities to more easily formulate unified checklists, promoting the orderly development of the entire industry. Therefore, whether from the perspectives of quality management, data mutual recognition or regulatory compliance, promoting the standardization of bioinformatics processes has become an inevitable trend and is a necessary prerequisite for the long-term healthy development of clinical genomics.

4 Key Technologies and Strategies for Standardizing Bioinformatics Processes

4.1 Workflow management system

In modern bioinformatics analysis, if one wants to make the process truly repeatable and traceable, the Workflow Management System (WMS) has almost become an essential tool. It is not a simple "automated script", but a system that can clearly describe complex analysis steps and dependencies. Nextflow, Snakemake, Cromwell (supporting WDL language), and Galaxy are several common solutions at present. Among them, Nextflow and Snakemake are the most frequently used in both scientific research and clinical practice. Nextflow writes processes in DSL language and can flexibly adapt to various computing platforms; Snakemake is more like Python, with intuitive rule definitions and can also be extended for use in cluster environments (Köster and Rahmann, 2018). With the help of these systems, many analysis steps that originally needed to be carried out manually can now be automated. After defining the process, no matter who runs it, as long as the input is the same, the output result will remain consistent. This not only reduces the differences in human operation but also makes the results easier to trace. Most WMS come with a built-in logging function, which automatically records the software version, parameter configuration and operation information of each step. For clinical laboratories,

transforming the sequencing data analysis process into a fixed and reusable standard pipeline not only saves manpower but also makes process upgrades traceable. More importantly, when multiple centers use the same set of processes, the consistency of the analysis results can also be guaranteed (Jackson et al., 2021; Baykal et al., 2024).

4.2 Software containerization and environment standardization

Another major reason for inconsistent process results is actually the difference in software environments . Version differences of the same software, mismatched dependency packages, and even changes in the operating system may all cause minor deviations in the analysis results. To avoid these problems, containerization technology is widely adopted. Docker and Singularity are the two most commonly used container tools. The former is mostly used in server or cloud environments, while the latter is specifically optimized for high-performance computing platforms. The function of a container lies in that it can package analysis software, dependent environments, and system configurations all into a single image file. In this way, no matter which machine is running, the environment can be guaranteed to be consistent. In other words, as long as the container images are the same, the analysis results should also be exactly the same. Containerization also makes the deployment and maintenance of processes easy. Upgrading a certain software version only requires replacing the image and then distributing it uniformly. Workflow systems such as Nextflow themselves also support direct invocation of container images, making standardization of the entire environment simpler. Nowadays, some clinical institutions have even packaged the entire NGS analysis process into containers for the rapid deployment of the same system across different hospitals (Kadri et al., 2022; Florek et al., 2025). With the support of containerization, the portability and repeatability of bioinformatics processes have been significantly enhanced, making it easier for laboratories to meet the environmental consistency requirements stipulated by regulations.

4.3 Analysis process verification and quality control

Before applying bioinformatics processes to clinical practice, strict performance validation must be conducted to confirm their detection ability and stability for target variations (Jennings et al., 2017). Process validation typically involves the assessment of sensitivity, specificity, accuracy and repeatability (Roy et al., 2018). For this reason, the industry recommends using standard reference samples and known true value datasets for testing. For example, the high-confidence variant set of human Genome standard samples (such as NA12878) provided by the Genome in a Bottle (GIAB) Project is widely used to evaluate the detection rate of SNV/Indel in the process. During verification, the data of the standard sample was input into the process to be tested, and the conformity of the output variation with the authoritative true value set was compared to calculate the sensitivity and false positive rate. For the detection of somatic variations such as tumors, alliances such as SEQC2 have also released standard datasets and reference variation sets, which can be used to verify the detection ability of the process at different variation frequencies. In addition to using standard samples, the laboratory should also design repeat tests to evaluate the repeatability of the process, that is, whether the same results are obtained from multiple independent runs (Baykal et al., 2024). After verification, a written report should be formed to record the process version, test data, performance indicators, etc., for regulatory review (Jennings et al., 2017).

In actual operation, quality control measures should also be introduced, such as adding control samples with known variations in each batch of analysis to monitor process performance. If control variations are not detected or the results are abnormal, problems in the analysis process can be investigated and analyzed in a timely manner (Haanpää et al., 2025). In addition, version control and change management are also important parts of process quality control. When bioinformatics processes or the software tools therein are updated, the differences in results between the old and new versions must be evaluated to ensure that the improvements do not reduce the detection sensitivity. Many laboratories use version control systems such as Git to manage process code and establish automated test pipelines. They set threshold monitoring for key indicators (such as detection rate and accuracy rate), and refuse to release a new version once its performance falls below the set standards (Baykal et al., 2024; Haanpää et al., 2025). Through continuous quality monitoring and version management, the performance of bioinformatics processes can remain stable and steadily improve with technological progress.

5 Standard Norms and Practical Applications

5.1 International guidelines and standards

As sequencing technology is increasingly used in clinical diagnosis, international norms regarding bioinformatics processes have gradually taken shape. In 2018, Association for Molecular Pathology (AMP) and College of American Pathologists (CAP) jointly proposed 17 consensus recommendations for the validation of clinical NGS analysis processes (Roy et al., 2018) which can be regarded as a relatively systematic guideline in the industry. It not only talks about process design and development, but also involves aspects such as verification and quality control. For example, it is suggested that laboratories use reference materials to evaluate the detection rate of different variant types, establish error-monitoring mechanisms, and have professionals with bioinformatics backgrounds be responsible for process management (Roy et al., 2018). In the same year, the American College of Pathologists also included bioinformatics analysis in the laboratory checklist, requiring clinical laboratories to validate the analysis software, record upgrades, and conduct regular performance evaluations. An earlier step in the field of tumor sequencing - AMP released guidelines on targeted panel sequencing in 2017, clarifying the specific standards that the bioinformatics pipeline should achieve in terms of minimum variant frequency detection, background noise control, etc. (Jennings et al., 2017; Klee et al., 2023). EMQN in Europe has also repeatedly emphasized in its quality assessment guidelines that unifying bioinformatics processes is an important prerequisite for laboratory consistency. In addition to industry associations, the International Organization for Standardization (ISO) released the technical standard ISO/TS 23357:2023 (ISO 2023) specifically for genomic informatics in 2023, which put forward unified requirements for the analysis and sharing of clinical genomic data, with detailed norms covering data formats, variant naming, and quality reporting (Haanpää et al., 2025). It can be said that these international standards provide a clear reference framework for laboratories. If anyone wants their results to be recognized across institutions, following these requirements is basically an inevitable path.

5.2 Domestic norms and practices

The standardization construction in China started a little later, but the progress has been very fast. With the entry of high-throughput sequencing into the fields of genetic disease and tumor diagnosis, the National Health Commission of the People's Republic of China has successively issued multiple technical documents, requiring that the analysis processes in laboratories must be "accurate, controllable and traceable". Meanwhile, domestic experts have also organized multiple consensus discussions. The "Consensus on the Standardization of the Entire Process of Clinical Testing for Next-Generation Sequencing of Genetic Diseases" is one of the more representative ones. It not only offers suggestions for each analysis stage but also lists commonly used software and recommended parameters, with the aim of promoting the unification of quality control, comparison, variant screening and reporting in the industry. Some large centers have taken the lead by establishing standardized analysis platforms, encapsulating modules such as quality control, comparison, variation identification, annotation, and reporting, and using automated assembly lines to reduce human differences (Chen et al., 2024). Some laboratories have even developed "one-click" systems. By simply uploading data, the system can automatically generate standard reports that meet clinical diagnostic requirements. In projects such as thalassemia and deafness gene screening, such standardized procedures have significantly enhanced the consistency of results among different batches and individuals, and also ensured the positive detection rate and the accuracy of reports. These experiences show that for international guidelines to be truly implemented, they must be optimized in light of domestic realities rather than simply copied.

5.3 Benefits of standardized processes

From the perspective of practical effects, the benefits brought by standardized processes have become quite obvious. First of all, the results are more reliable. The false alarm and missed detection rates of laboratories that adopt a unified process have significantly decreased. For example, in laboratories using the GATK best practice processes, the sensitivity performance of single-gene genetic disease detection reaches over 99%, while laboratories using different pipelines in a decentralized manner often fail to reach this level. Secondly, efficiency has been significantly enhanced. Process standardization is often accompanied by automation. Many hospitals have compressed the reporting cycle of whole exome sequencing from the original two weeks to several days,

thus winning precious time for the diagnosis of neonatal genetic diseases. Another often overlooked advantage is scalability. The standardized process structure is clear. When developing new testing items or switching to tumor panel analysis, only minor adjustments need to be made to the original process (Roy et al., 2018). Meanwhile, fixed SOPs also facilitate personnel training, enabling novices to get started more quickly (Whiffin et al., 2016). Supervision and inspection are also more efficient because there is a unified operational basis. Overall, an increasing amount of evidence indicates that the standardization of bioinformatics processes not only enhances the quality and efficiency of testing but also lays the foundation for data inter-communication and resource sharing across the country. This trend has almost become a consensus in the industry.

6 Case Study: Construction of Standardized Analysis Process for Tumor Genomics

6.1 Research background and objectives

The analysis of tumor genomics has always been complex, not only due to the huge volume of data, but also because of the diverse types of results and the high requirements for interpretation. A case often involves multiple aspects such as somatic cell SNV/Indel, copy number changes, gene fusions, tumor mutational burden (TMB), and microsatellite instability (MSI). In the past, different laboratories adopted various analytical methods. Some pursued sensitivity, while others emphasized speed, and the results were often difficult to compare (Feng et al., 2023). We aim to integrate the analysis methods, output formats and report structures by establishing a standardized bioinformatics process for precise tumor diagnosis. From the reception of sequencing data to the interpretation of results, the entire process will be automated and highly consistent, truly achieving an analysis system that is “repeatable, traceable and clinically applicable”. The goal of the process is not only to increase the detection rate, but also to ensure the stability of the report, making the results of different batches and different analysts as consistent as possible. Ultimately, we want to see if such a standardized process can stand the test of time in real clinical scenarios, and also use this to summarise experiences and improve the future optimisation direction (Figure 1) (Nasra et al., 2024; Nguyen et al., 2025).

6.2 Process construction and implementation steps

The construction of the process starts from the characteristics of tumor samples and adopts a dual-channel analysis design of DNA and RNA. DNA sequencing was mainly whole-exome sequencing (WES) that matched tumor-normal controls, with average depths of 400×and 180×, respectively. RNA sequencing is the whole transcriptome data. The sequencing platform is Illumina, with 150 bp at both ends and a total read length of approximately 100 million pairs. The samples should undergo pathological assessment before being put on the machine, and the tumor content should exceed 20%. After the DNA/RNA extraction is completed, the library is constructed and sequenced. After the sequencing is finished, the LIMS system automatically records the sample information and initiates the analysis process (managed by Nextflow). The first step of the process is data quality control and comparison. DNA sequences were evaluated for quality using FastQC, and Trimmomatic was used to remove linkers and low-quality fragments. The reference genome from BWA-MEM to hg38 was compared, and then repeated labelling and base mass recalibration were performed. RNA data is monitored using FastQC/MultiQC indicators, such as the rRNA ratio, Q30 ratio, and alignment rate, and is cleaned up when necessary. Subsequently, the RNA-seq data were aligned to the hg38 transcriptome using STAR to support the identification of splicing sites. All quality control results will be automatically summarised into the log, including indicators such as sequencing depth, coverage rate, and Q30 distribution (Cabello-Aguilar et al., 2023; Zerdes et al., 2025).

In the mutation detection section, the process adopts a dual-algorithm strategy. Somatic cell SNV/Indel was jointly detected by GATK Mutect2 and Strelka2. Mutect2 first generates a candidate variant set, which then filters out sequencing noise and false positives through FilterMutectCalls. Strelka2 is more sensitive to low-frequency variations. After taking the union of the two results, the internal assessment showed that the detection sensitivity increased from approximately 85% to over 95%. The analysis of CNV and LOH was performed by CNVkit, and structural variations (SV) were identified by Manta for abnormal pairing and split reads. Fusion genes were detected at the RNA level using FusionCatcher, and the high-confidence results were then manually rechecked.

All modules are executed in parallel in Nextflow to fully utilise computing resources. The final output results include the SNV/Indel list, CNV and copy number table, fusion gene list, etc.

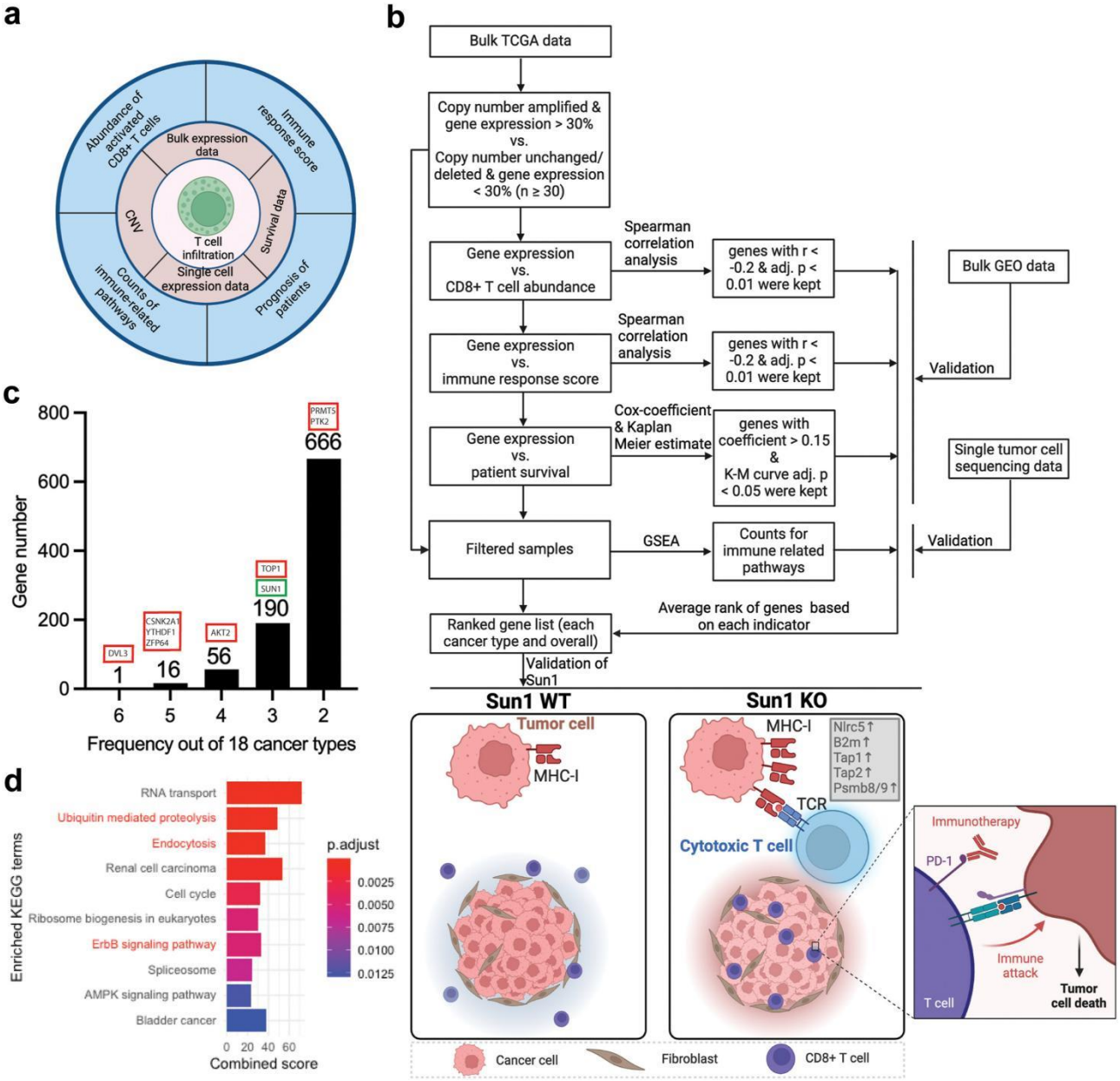

Figure 1 Identification of tumour-specific genes whose expression potentially impedes T-cell infiltration (Adopted from Nasra et al., 2024)

6.3 Implementation results and clinical application

Nearly 200 patient samples were included in the clinical application stage. Doctors generally report that the new process has significantly improved work efficiency. In the past, a senior analyst could produce at most 2 to 3 reports a day, but now the same number of people can complete 5 to 6. The report format is also clearer: the homepage centrally displays key information such as driver genes, therapeutic targets, TMB/MSI, etc., while the appendix section details references and database annotations. For instance, the report of a lung cancer patient clearly indicates the EGFR L858R mutation and amplification, along with information on drugs such as osimertinib and the level of evidence, enabling doctors to directly formulate treatment plans. Another colorectal cancer patient reported MSI-H and high TMB, and the doctor immediately decided to treat with immune checkpoint inhibitors. The optimisation of this report structure not only improves readability but also enhances the efficiency of communication between doctors and patients (Ghoreyshi et al., 2025; Nguyen et al., 2025). Patient

surveys show that standardized reports make it easier for them to understand the test results. Some patients who have received popular science education can even interpret the significance of driver mutations by themselves.

Overall, the benefits brought by this standardized process are multi-faceted. Firstly, it ensures that key variations will not be missed, guaranteeing the integrity of the diagnosis. Secondly, a unified report template reduces the differences among analysts, making the results more consistent and reliable. Thirdly, automation has significantly shortened the analysis cycle-the average time from submission for inspection to reporting has been reduced from 10 working days to 7. This improvement is particularly important for patients with advanced tumors who need to make quick decisions. The doctor satisfaction survey also shows that after using the new process, their trust in the test results and understanding of the reports have both improved. Nevertheless, there are still areas that need improvement. For instance, there are still certain limitations in the detection of low-frequency variations (<5 %). In the future, it is planned to introduce UMI markers or adopt more sensitive algorithms. Meanwhile, for the classification of rare mutations, they are currently mostly treated as “unclear meaning”, and the interpretation may be further optimised in combination with AI-assisted annotation systems.

7 Challenges and Future Prospects

Artificial intelligence (AI) has been increasingly prominent in genomic data analysis in recent years, especially in machine learning and deep learning technologies. Many of the analysis steps that originally relied on manual rule setting and parameter adjustment can now be "learned" and completed by the algorithms themselves. Take mutation detection as an example. AI models are beginning to replace traditional statistical methods. For instance, Google's DeepVariant has significantly improved the accuracy of mutation recognition by leveraging deep neural networks. It can be imagined that in the future, standardized processes are likely to be equipped with such an AI Caller, no longer merely performing command-based tests, but further reducing false positives while ensuring sensitivity. Such integration makes standardized processes smarter, moving from automation to true intelligence. However, AI is not omnipotent. Issues such as the interpretability of the model, data bias, and transparency remain thorny. Even if a model performs extremely well, if it cannot explain why a certain conclusion is given, it is still difficult to be fully trusted in clinical scenarios. Therefore, in the future, standardization organizations may need to introduce corresponding norms to ensure that AI models must undergo strict validation and clearly indicate the confidence range when outputting results.

From a broader perspective, the integration of AI is merely one direction in the evolution of bioinformatics process standardization. The real challenge lies in how to keep the process at an "updated pace" in the rapidly changing technological environment. With the continuous evolution of sequencing technology, new platforms such as long-read and single-molecule sequencing have gradually entered clinical practice, and new algorithms are also emerging one after another. If standardized processes cannot be dynamically adjusted, they will soon be left behind by The Times. Therefore, it is crucial to establish a mechanism that can be regularly reviewed, evaluated and improved. Just as the medical field often refers to "continuous quality improvement", processes also require this kind of cyclical iteration - planning (Plan), doing (Do), checking (Check), and acting (Act), constantly making corrections and optimizations.

In the future, perhaps we will witness the emergence of regional or even national genomic data platforms. The variant and phenotypic data produced by different hospitals through a unified process are aggregated and shared, achieving the true meaning of "one process, national reference". Of course, beyond technology, there are also complex issues such as data mutual recognition and legal compliance, but standardized output is undoubtedly the first step taken. Overall, the integration of AI and standardization is unstoppable. It will enable bioinformatics processes not only to "run automatically" but also to "learn to run", ultimately making clinical genomics analysis more efficient and accurate.

Acknowledgments

The authors extend sincere thanks to two anonymous peer reviewers for their invaluable feedback on the manuscript.

Conflict of Interest Disclosure

The authors affirm that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Baykal P., Łabaj P., Markowetz F., Schriml L., Stekhoven D., Mangul S., and Beerenwinkel N., 2024, Genomic reproducibility in the bioinformatics era, Genome Biology, 25(1): 213.
https://doi.org/10.1186/s13059-024-03343-2

Brancato V., Esposito G., Coppola L., Cavaliere C., Mirabelli P., Scapicchio C., Borgheresi R., Neri E., Salvatore M., and Aiello M., 2024, Standardizing digital biobanks: integrating imaging, genomic, and clinical data for precision medicine, Journal of Translational Medicine, 22(1): 136.
https://doi.org/10.1186/s12967-024-04891-8

Cabello-Aguilar S., Vendrell J., and Solassol J., 2023, A bioinformatics toolkit for next-generation sequencing in clinical oncology, Current Issues in Molecular Biology, 45(12): 9737-9752.
https://doi.org/10.3390/cimb45120608

Chen S., 2023, Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp, iMeta, 2(2): e107.
https://doi.org/10.1002/imt2.107

Cherney B., Diaz A., Chavis C., Gattas C., Evans D., Arambula D., and Stang H., 2024, The next-generation sequencing quality initiative and challenges in clinical laboratories, Emerging Infectious Diseases, 31(13): 24-1175.
https://doi.org/10.3201/eid3113.241175

Feng B., Lai J., Fan X., Liu Y., Wang M., Wu P., Zhou Z., Yan Q., and Sun L., 2024, Systematic comparison of variant calling pipelines of target genome sequencing across multiple next-generation sequencers, Frontiers in Genetics, 14: 1293974.
https://doi.org/10.3389/fgene.2023.1293974

Florek K., Young E., Incekara K., Libuit K., and Kapsak C., 2025, Advantages of software containerization in public health infectious disease genomic surveillance, Emerging Infectious Diseases, 31(Suppl 1): S18.
https://doi.org/10.3201/eid3113.241363

Ghoreyshi N., Heidari R., Farhadi A., Chamanara M., Farahani N., Vahidi M., and Behroozi J., 2025, Next-generation sequencing in cancer diagnosis and treatment: clinical applications and future directions, Discover Oncology, 16: 578.
https://doi.org/10.1007/s12672-025-01816-9

Hao Z., Liang X., and Li G., 2022, Quality control and preprocessing of sequencing reads, Bio-protocol, 12: 13.
https://doi.org/10.21769/BioProtoc.4454

Jackson M., Kavoussanakis K., and Wallace E., 2021, Using prototyping to choose a bioinformatics workflow management system, PLoS Computational Biology, 17(2): e1008622.
https://doi.org/10.1371/journal.pcbi.1008622

Jennings L., Arcila M., Corless C., Kamel-Reid S., Lubin I., Pfeifer J., Temple-Smolkin R., Voelkerding K., and Nikiforova M., 2017, Guidelines for validation of next-generation sequencing-based oncology panels: A joint consensus recommendation, Journal of Molecular Diagnostics, 19(3): 341-365.
https://doi.org/10.1016/j.jmoldx.2017.01.011

Kadri S., Sboner A., Sigaras A., and Roy S., 2022, Containers in bioinformatics: applications, practical considerations, and best practices in molecular pathology, The Journal of Molecular Diagnostics, 24(5): 442-454.
https://doi.org/10.1016/j.jmoldx.2022.01.006

Koboldt D., 2020, Best practices for variant calling in clinical sequencing, Genome Medicine, 12: 91.
https://doi.org/10.1186/s13073-020-00791-w

Köster J., and Rahmann S., 2018, Snakemake-a scalable bioinformatics workflow engine, Bioinformatics, 28(19): 2520-2522.
https://doi.org/10.1093/bioinformatics/bty350

Lavrichenko K., Engdal E., Marvig R., Jemt A., Vignes J., Almusa H., Saether K., Briem E., Caceres E., Elvarsdóttir E., Gíslason M., Haanpää M., Henmyr V., Hotakainen R., Kaasinen E., Kanninga R., Khan S., Lie-Nielsen M., Madsen M., Mähler N., Maqbool K., Neethiraj R., Nyrén K., Paavola M., Pruisscher P., Sheng Y., Singh A., Srivastava A., Stautland T., Andreasen D., de Boer E., Vang S., Wirta V., and Bagger F., 2025, Recommendations for bioinformatics in clinical practice, Genome Medicine, 17: 1-14.
https://doi.org/10.1186/s13073-025-01543-4

Martín R., Gaitán N., Jarlier F., Feuerbach L., de Soyres H., Arbonés M., Gutman T., Puiggròs M., Ferriz A., Gonzalez A., Estelles L., Gut I., Capella-Gutierrez S., Stein L., Brors B., Royo R., Hupé P., and Torrents D., 2024, ONCOLINER: a new solution for monitoring, improving, and harmonizing somatic variant calling across genomic oncology centers, Cell Genomics, 4(9): 100639.
https://doi.org/10.1016/j.xgen.2024.100639

Minoche A., Lundie B., Peters G., Ohnesorg T., Pinese M., Thomas D., Zankl A., Roscioli T., Schonrock N., Kummerfeld S., Burnett L., Dinger M., and Cowley M., 2021, ClinSV: clinical-grade structural and copy number variant detection from whole genome sequencing data, Genome Medicine, 13: 32.
https://doi.org/10.1186/s13073-021-00841-x

Nasra S., Lin L., Mohammad A., Mahroo M., Moridi A., Wang M., Zemp F., Mahoney D., and Wang E., 2024, A computational pipeline for identifying gene targets and signalling pathways in cancer cells, eBioMedicine, 104: 105167.
https://doi.org/10.1016/j.ebiom.2024.105489

Nguyen C., Nguyen T., Trivitt G., Capaldo B., Yan C., Chen Q., Renzette N., Topaloglu U., and Meerzaman D., 2025, Modular and cloud-based bioinformatics pipelines for high-confidence biomarker detection in cancer immunotherapy clinical trials, PLoS One, 20(8): e0330827.
https://doi.org/10.1371/journal.pone.0330827

Samarakoon P., Fournous G., Hansen L., Wijesiri A., Zhao S., Alex A., Nandi T., Madduri R., Rowe A., Thomssen G., Hoving E., and Razick S., 2025, Benchmarking accelerated next-generation sequencing analysis pipelines, Bioinformatics Advances, 5(1): vbaf085.
https://doi.org/10.1093/bioadv/vbaf085

Weißbach S., Sys S., Hewel C., Todorov H., Schweiger S., Winter J., Pfenninger M., Torkamani A., Evans D., Burger J., Everschor-Sitte K., May-Simera H., and Gerber S., 2021, Reliability of genomic variants across different next-generation sequencing platforms and bioinformatic processing pipelines, BMC Genomics, 22: 62.
https://doi.org/10.1186/s12864-020-07362-8

Whiffin N., Brugger K., and Ahn J., 2016, Practice guidelines for development and validation of software, with particular focus on bioinformatics pipelines for processing NGS data in clinical diagnostic laboratories, PeerJ Preprints, 5: e2996v1.
https://doi.org/10.7287/peerj.preprints.2996v1

Wilton R., and Szalay A., 2023, Short-read aligner performance in germline variant identification, Bioinformatics, 39(8): btad480.
https://doi.org/10.1093/bioinformatics/btad480

Zerdes I., Filis P., Fountoukidis G., El-Naggar A., Kalofonou F., D'Alessio A., Pouptsis A., Foukakis T., Pentheroudakis G., Ahlgren J., Smith D., and Valachis A., 2025, Comprehensive genome profiling for treatment decisions in patients with metastatic tumors: real-world evidence meta-analysis and registry data implementation, JNCI: Journal of the National Cancer Institute, 117(6): 1117-1124.
https://doi.org/10.1093/jnci/djaf015