scientific reports

Check for updates

OPEN

RnaXtract, a tool for extracting gene expression, variants, and cell-type composition from bulk RNA sequencing

Sophiane G. Bouirdene[1,2,3], Simon Gotty[2,3], Mickaël Leclercq[2,3], Charles Joly-Beauparlant[2,3], Emeric Texeraud[2,3], Steve Bilodeau[4,5,6,7]✉ & Arnaud Droit[2,3,6,7,8]✉

RNA sequencing (RNA-seq) is a widely used method in transcriptomics research, offering insights into gene expression, variant discovery, and, when deconvoluted, the cellular composition of complex tissues. However, existing RNA-seq pipelines frequently emphasize gene expression analysis and often lack cell deconvolution and variant calling. To address these limitations, we present RnaXtract, a comprehensive and user-friendly pipeline designed to maximize extraction of valuable information from bulk RNA-seq data. RnaXtract automates an entire workflow, encompassing quality control, gene expression quantification, variant calling, and the cell-type deconvolution. Built on the Snakemake framework, RnaXtract ensures robust reproducibility, efficient resource management, and flexibility to adapt to diverse research needs. The pipeline integrates state-of-the-art tools, from quality control to the new updates on variant calling and cell-type deconvolution tools such as EcoTyper and CIBERSORTx, enabling researchers to extract biological insights with precision. By providing an end-to-end solution for bulk RNA-seq, RnaXtract addresses critical gaps in existing workflows, empowering researchers to explore gene expression, genetic variation, and cellular heterogeneity within a single cohesive framework.

Keywords Bulk RNA-seq, Transcriptomics, Snakemake pipeline, Gene expression, Variant calling, Cell deconvolution, Computational biology, RNA sequencing analysis

Despite the emergence of new technologies, such as single-cell and spatial transcriptomics, bulk RNA sequencing (RNA-seq) remains the most widely employed due to its cost-effectiveness in providing a comprehensive tissue overview[1]. Consequently, short-read sequencing technologies, which typically produce tens of millions of reads per sample, are the predominant methods for gene expression profiling[2]. The widespread adoption of bulk RNA sequencing has led to the development of standardized workflows for data analysis. These workflows typically begin with quality control, assessing the quality and accuracy of bases in each read, followed by a trimming step to remove low-quality reads[3,4]. The subsequent step involves mapping the reads to a reference genome or transcriptome, which is used for gene expression quantification by counting the reads mapped to each gene[5,6]. This alignment also facilitates variant calling, allowing for the identification of individual-specific mutations[7,8]. The extensive use and standardization of these methods have resulted in numerous analytical pipelines[9].

The advent of single-cell sequencing has provided new insights into cellular composition and characterization, opening the opportunity for analyzing tissue composition and its differentiation under various conditions[10]. Single-cell transcriptomics is usable to deconvolute bulk RNA-seq to estimate cellular composition of the tissue and their specific states[11,12]. However, these methods are relatively new, and not included in existing bulk RNA-

[1]Doctoral Program in Molecular Medicine, Faculté de Médecine, Université Laval, Québec, Québec G1V 0A6, Canada. [2]Endocrinology and Nephrology Axis, Centre de recherche du CHU de Québec—Université Laval, Québec, Québec G1V 4G2, Canada. [3]Computational Biology Laboratory, Centre de recherche du CHU de Québec—Université Laval, Québec, Québec G1V 4G2, Canada. [4]Department of Molecular Biology, Medical Biochemistry and Pathology, Faculté de Médecine, Université Laval, Québec, Québec G1V 0A6, Canada. [5]Oncology Axis, Centre de recherche du CHU de Québec—Université Laval, Québec, Québec G1J 1Z4, Canada. [6]Centre de Recherche sur le Cancer de l'Université Laval, Québec, Québec G1R 3S3, Canada. [7]Centre de Recherche en Données Massives de l'Université Laval, Québec, Québec G1V 0A6, Canada. [8]Department of Molecular Medicine, Faculté de Médecine, Université Laval, Québec, Québec G1V 0A6, Canada. ✉email: steve.bilodeau@crchudequebec.ulaval.ca; arnaud.droit@crchudequebec.ulaval.ca

seq analysis pipelines. Additionally, many of these pipelines are optimized for gene expression analysis and do not align with current gold-standard methods for variant calling (Table 1).

With the quantity of data generated from bulk RNA sequencing, machine learning (ML) methods are used for biomarker discovery, enabling the identification of molecular signatures that correlate with specific conditions or therapeutic responses[13,14]. Consequently, its use is becoming prevalent in biological analyses to identify significant genes, pathways and disease mechanisms[14,15]. However, many current pipelines are not optimized for ML integration. This often necessitates data transformation by the user, which can be burdensome and may deter biologists from using these types of tools.

In this study, we present RnaXtract, a complete workflow designed to maximize extraction of the information contained in RNA-Seq data. RnaXtract not only generates the conventional normalized gene expression matrix but also performs state-of-the-art variant calling analysis and cellular deconvolution using CIBERSORTx and its extension, EcoTyper[11,16]. This cellular composition analysis provides an additional layer of feature extraction beyond genomics and gene expression.

Methods

Workflow implementation

The complete pipeline was developed using Snakemake, a Python-based workflow definition language, due to its efficiency, scalability, and ease of use[17]. Snakemake facilitates seamless workflow management and reproducibility in bioinformatics analyses. To ensure long-term reproducibility, all tools and environments were containerized using Singularity images[18]. A script was created to simplify the generation of a config.yml file to manage user interactions. Running RnaXtract requires four dependencies: R (v3.6.0)[19], Snakemake[17], Python3[20], and a container runtime engine: Singularity. The list of Docker containers utilized by RnaXtract is provided in Supplementary Table.S1.

RnaXtract is structured into four modules: preprocessing with quality control, gene expression analysis, cell deconvolution and variant calling with an optional recalibration step for variant calling (Fig. 1).

Quality control and alignment

Initially, raw sequences are inputted into the workflow, and fastp (v0.20.0)[4] is employed to perform quality trimming and filtering of reads, utilizing default parameters for window size, mean quality, and length. Subsequently, FastQC (v0.11.5) generates a quality report for each pre-processed read file, and MultiQC (v0.11.5) compiles a comprehensive report across all samples. Following quality control, alignment is conducted using STAR (v2.7.2b)[6] against a reference genome (or transcriptome), which is essential for variant calling. For the gene expression quantification Kallisto (v0.48.0)[21] was chosen as it is the quickest tool for this step while having equivalent quality to its concurrents[22].

Variant calling and transformation

Genomic variants can be inferred from RNA data through variant calling. Due to the number of intermediate steps carried out during this process, it was encapsulated in an independent sub-workflow run at sample level. The input data consists of the alignments generated with STAR and the output is a list of variants in the variant call format (VCF) and a table format file. The whole process is performed using the Genome Analysis ToolKit (v4.1.3.0)[7,21], and it was designed following the GATK best practices for the variant calling from RNA-Seq data which is the gold standard in the field. The variant calling is done on all the samples together instead of them separately using the "HaplotypeCaller" and "GenotypeGVCFs" functions following GATK and genotyping recommendations[23,24]. This reduces the direct dependency of the depth and enhances the quality of the mutation identification.

Variants are separated between single nucleotide polymorphism (SNP) and insertion/deletion (INDEL) to be filtered using a hard filter. The filters differ between the two types of variants and were chosen following GATK guidelines. The filtering differences exist because SNPs and INDELs have distinct characteristics, error profiles, and detection challenges. Filtering criteria were tailored to maximize sensitivity and specificity for each type of variant while minimizing false positives from sequencing or alignment artifacts.

A primary challenge in machine learning applications is overfitting, where models capture dataset-specific features rather than generalizable biological signals. To reduce the risk of overfitting when using RnaXtract for

Tool	Gene expression	Variant calling (joint genotyping)	Cell deconvolution
MIGNON	Yes	Yes (no)	No
Latch	Yes	No	No
RNAflow	Yes	No	No
SPEAQeasy	Yes	Yes (no)	No
sequana_rnaseq	Yes	No	No
GenPipes RNA	Yes	No	No
RnaXtract	Yes	Yes (yes)	Yes

Table 1. Comparative analysis of RnaXtract and other RNA-Seq processing pipelines. Table placing RnaXtract in the various RNA-seq analysis tools in performing three key tasks: gene expression, variant calling, and cell deconvolution.

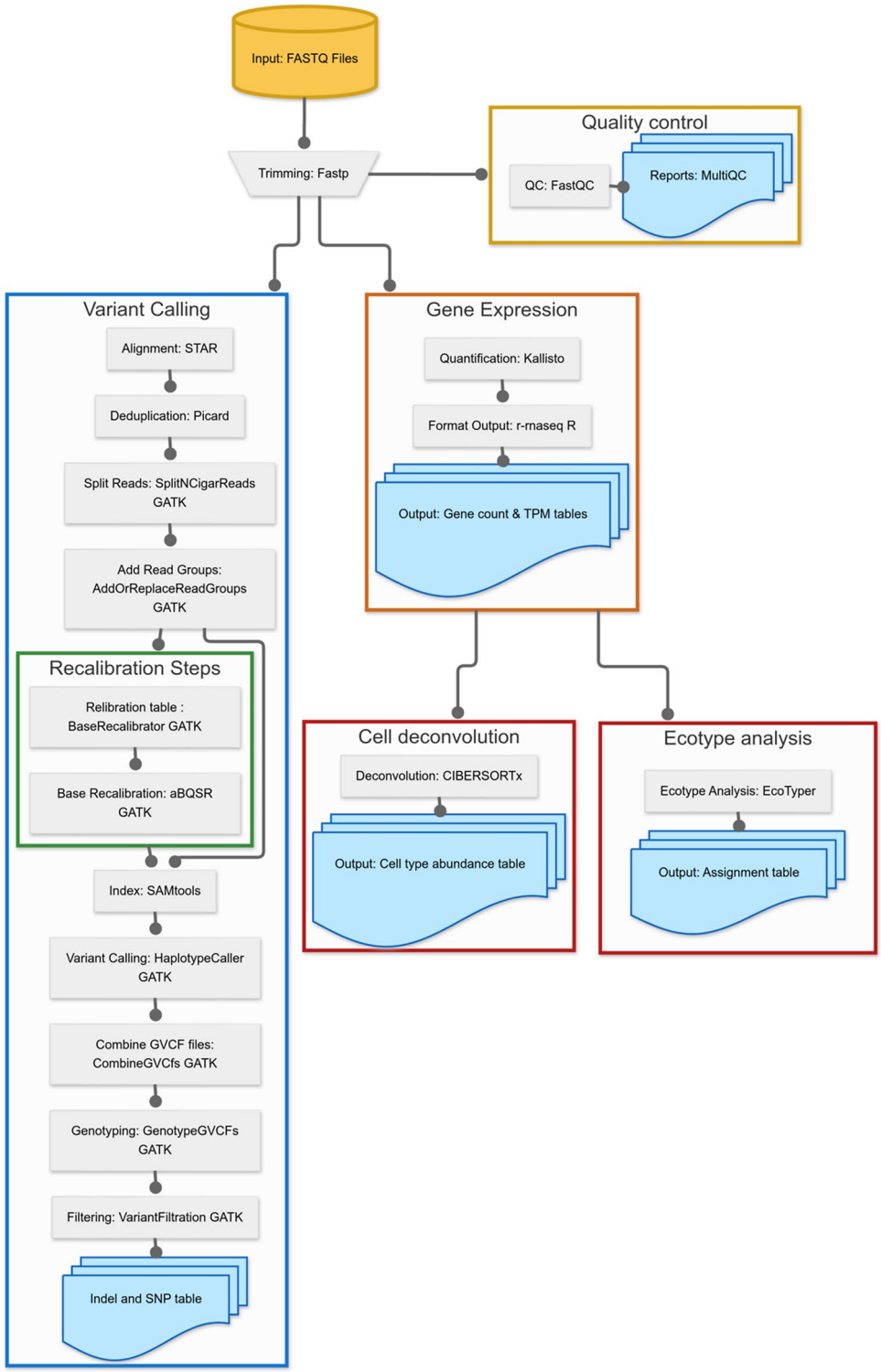

Fig. 1. RnaXtract workflow. Overview of the RNA-seq analysis pipeline, including quality control, pre-processing, variant calling, gene expression quantification, and cell deconvolution. The workflow begins with quality control using FastQC and MultiQC, generating a quality report. Pre-processing includes trimming and preparing FASTQ files. Variant calling involves alignment (STAR), deduplication (Picard), read splitting, and read group addition (GATK), followed by base quality score recalibration (BQSR) and variant identification using HaplotypeCaller and GenotypeGVCFs. Gene expression analysis is performed using Kallisto for quantification and subsequent formatting with the r-rnaseq package, generating gene count and TPM tables. Cell deconvolution is conducted with EcoTyper for ecotype analysis and CIBERSORTx for cell type abundance estimation. Final outputs include tables summarizing gene expression, variant calls, and inferred cell type compositions.

downstream machine learning approaches, we implemented a user-defined script to filter variants based on their frequency within the dataset. The output VCFs are then transformed into a table format. This table has as lines the sample identification and as columns the variant which name is composed of: chromosome:position_reference>mutation. The value of the column is represented by the two numbers associated with mutation present at this position separated by an underscore with 0 representing the absence of mutation. Mutations with absence of information for at least one sample at the position of the mutation were removed. Also, a possibility to exclude variants based on a threshold of presence in the cohort was added.

Normalization

The use of Transcripts per Million (TPM) was chosen as a normalization method for gene expression data. TPM accounts for both sequencing depth and gene length, ensuring comparability across samples by normalizing gene expression values relative to the total transcript count. This normalization method is essential for building machine learning models, as it enhances the model's ability to generalize and reduces its susceptibility to batch effects. In addition, the cell deconvolution step in the pipeline required TPM, as it is one of the two input formats supported by EcoTyper and CIBERSORTx.

Cell deconvolution

The cell deconvolution is achieved through the utilization of EcoTyper, a tool specifically designed to decode cellular heterogeneity within bulk RNA-seq data. EcoTyper is a pre-trained deep learning tool capable of estimating the cell composition and states which compose the ecotype of the bulk sample. Moreover, in the case that the user possesses their own single-cell RNA-seq datasets as reference, another step using CIBERSORTx was added. CIBERSORTx enables the creation of custom signature matrices from single-cell data, thereby improving the accuracy of cell type quantification and functional analysis. CIBERSORTx has been shown to be one of the best tools for deconvolution[25] and its extension offers a pre-trained model for cancer research which enhances its granularity by going into cell ecotypes. The combined use of EcoTyper and CIBERSORTx provides insight on the cellular landscapes in bulk RNA-seq datasets.

Organisms

RnaXtract is mostly oriented for the analysis of human transcriptomic data. However, it can be used for other species that possess both a reference genome and transcriptome, with only minor adjustments, primarily the selection of appropriate reference annotations and, when relevant, species-specific single-cell reference datasets. For cell deconvolution, CIBERSORTx has demonstrated utility across multiple organisms[26–28]. There are certain limitations in the use of EcoTyper, as its models were trained on human single-cell data. However, an option was included for users to apply EcoTyper if they supply a conversion table mapping their organism's cell types to corresponding human counterparts.

Results

To evaluate RnaXtract's ability to extract relevant data, we analyzed human breast cancer samples collected pre-chemotherapy to identify biomarkers predictive of treatment failure using machine learning. We selected 24 tumor samples from the European Nucleotide Archive project PRJNA1004593[29] (Supplementary Table S2). These samples included 12 patients who succeeded and 12 who failed the post-neoadjuvant chemotherapy. A random selection of 2,000 cells from a breast cancer single-cell dataset, available at CELLxGENE[30,31], were used to create the reference for CIBERSORTx. For the variant analysis, a threshold of 0.1 was applied, retaining only those variants present in at least 10% of the cohort.

A gene expression table in TPM format, SNP and INDEL tables, and custom-generated cell composition and ecotype data were generated. These outputs were used to develop individual models with BioDiscML[32], a machine learning tool capable of identifying biomarkers. The most relevant features were combined to create a final model combining accuracy and optimal feature usage.

Multiple models predicting chemotherapy success were derived with varying model performance. A broad range of Matthews Correlation Coefficient (MCC) scores were observed with the lowest for the EcoTyper deconvolution model (MCC = 0.029, 4 features) and the highest for the Gene Expression model (MCC = 0.762, 75 features). Integrating all four data types allowed us to reduce the number of features while building a high quality model. This integrated model achieved an MCC of 0.737—comparable to the Gene Expression model—but significantly lowering the standard deviation (STD) from 0.287 to 0.042 and requiring only 2 features (Fig. 2): a mutation within the 3'UTR of Syntaxin-16 (*STX16*) and the RNA level of ELMO Domain Containing 1 (*ELMOD1*).

STX16 is primarily involved in intracellular transport, particularly within the endosomal and Golgi apparatus pathways. It has been reported to be overexpressed in cervical cancer[33], and data from The Cancer Genome Atlas Breast Invasive Carcinoma suggest that patients with *STX16* amplification exhibit poorer disease-free survival (Supplementary Fig. 1). Mutations affecting the 3'UTR could impact mRNA stability, potentially enhancing cancer cell survival. In our cohort, although no consistent pattern was observed between the number of adenine repeats in the STX16-associated variant and its expression level, patients who experienced treatment failure consistently displayed higher STX16 expression compared to those who responded well (Fig. 3a). This observation aligns with our initial hypothesis that elevated STX16 expression may be linked to poor outcomes.

Similarly, *ELMOD1* is a GTPase-activating protein for the Arf family of small G proteins[34], playing a key role in cellular signaling regulation. Recent studies indicate that *ELMOD1* and *ELMOD2* are upregulated in breast cancer cell lines and are essential for cancer cell migration[35]. Low *ELMOD1* RNA expression may suggest reduced cancer cell migration and an improved likelihood of treatment success as seen in the cohort (Fig. 3b).

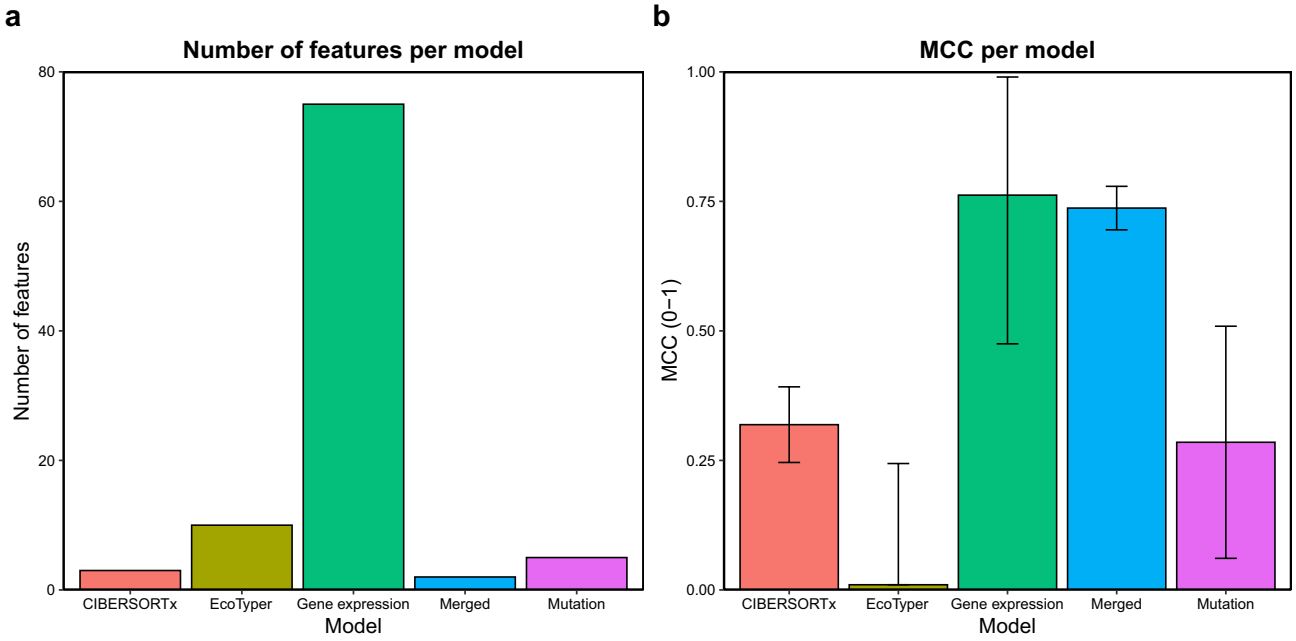

Fig. 2. Quality of machine learning model. (**a**) Comparison of the number of selected features and (**b**) model performance measured by the Matthews Correlation Coefficient (MCC) across different feature sets. The "Merged" model, which integrates all data types, demonstrates a reduced number of features compared to the "Expression" model while maintaining a high MCC. Additionally, the STD (error bars) in MCC is lower in the "Merged" model compared to individual models, indicating increased stability and robustness. This suggests that integrating multiple data sources improves classification performance while reducing feature redundancy and variability.

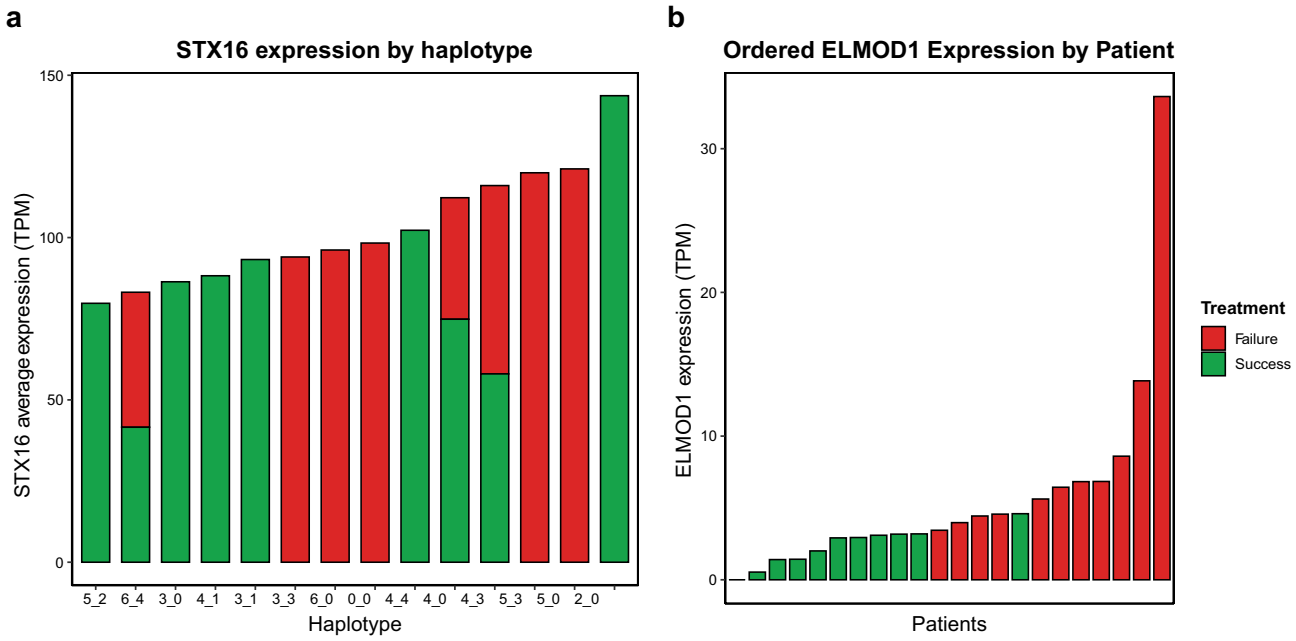

Fig. 3. Interpretation of features from the machine learning model. (**a**) Mean STX16 expression (TPM) grouped by haplotype, with bars colored according to treatment outcome (green: success; red: failure). (**b**) ELMOD1 expression levels (TPM) ordered by individual patients and stratified by treatment outcome. Each bar represents a single patient.

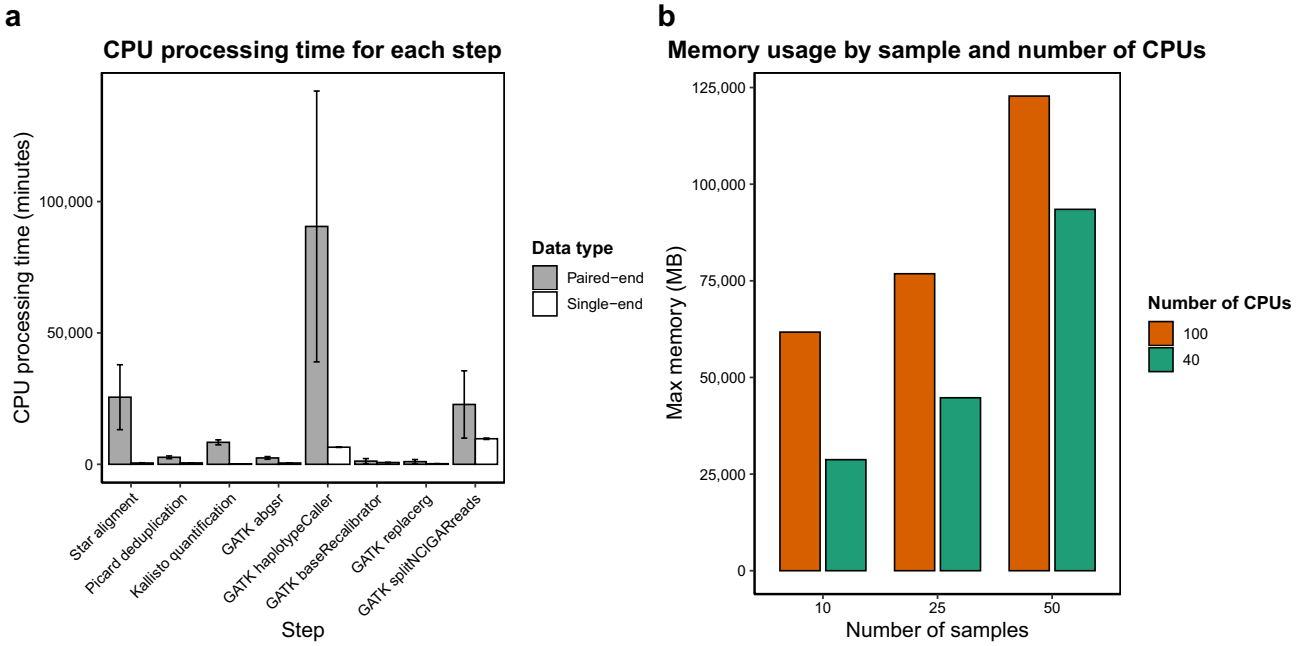

Fig. 4. RnaXtract performance results. (**a**) The barplot illustrates the comparison of data processing steps for paired-end and single-end sequencing data. The y-axis represents the CPU-times in minutes while the x-axis denotes the different steps involved in multi-samples RnaXtract's steps. Two distinct data sets were used: the breast cancer SRA cohort, paired-end (dark grey) and a cohort using single-end sequencing (white). The bars for each step show the average performance with the standard deviation. (**b**) Maximum memory usage (in MB), measured as peak Resident Set Size (RSS), for pipeline executions with 40 and 100 CPU cores across datasets of 10, 24, and 50 samples.

This example highlights RnaXtract's effectiveness in multi-layered biomarker analysis for cancer research, offering a comprehensive and automated tool for RNA-seq data extraction with applications in machine learning. However, it is important to note that the chemotherapy response prediction model presented here is based on a limited sample size (n = 24). Further validation on larger, independent datasets will be necessary to confirm the predictive value. It has to be noted that it's the same patient which has results indicating a failure but had the treatment succeeding, showing the heterogeneity in data and the need to be careful with these models.

Workflow performance evaluation

To assess the performance and resource utilization of the RnaXtract pipeline, we conducted analyses on multiple human datasets with diverse characteristics, encompassing a total of 30 samples derived from two distinct datasets (24 paired-end and 6 single-end)[29,36]. The pipeline was executed on an AMD EPYC 7601 32-Core Processor[37], utilizing the default core configuration specified by the pipeline. Initial evaluations revealed that the variant calling step was the most computationally intensive, requiring an average of 29 h of CPU time (Fig. 4a). Subsequent steps, including dataset integration, genotyping, and variant filtration, each consumed approximately 10 h of CPU time and demonstrated potential for parallelization. In contrast, EcoTyper and CIBERSORTx exhibited significantly lower computational demands, completing within 20 min of CPU time, with EcoTyper further benefiting from multi-threaded execution. The RNA-seq function, which processes Kallisto output into tabular format, was the least resource-intensive, completing in under 2 min.

To characterize memory usage during the most demanding step of the pipeline, we benchmarked RnaXtract on the PRJNA1004593 dataset of increasing the number of samples (10, 25, and 50), using either 40 or 100 CPU cores with smem tool. A peak memory consumption was measured using Resident Set Size (RSS), which reflects the actual physical memory held in RAM during execution (Fig. 4b). Memory usage scaled with the number of samples and was consistently higher with 100 cores, due to greater parallelism and concurrent memory allocation. At the largest scale (50 samples), maximum RSS exceeded 125 GB with 100 cores, compared to under 100 GB with 40 cores. These results show a linear correlation between the number of CPUs and samples for the memory usage. The step which consumed the most memory was HaplotypeCaller.

Functional overview of existing bulk RNA-seq workflows

In order to evaluate the place of RnaXtract within the field of bulk RNA-seq data processing, we compiled a list of pipelines published after 2020 (Table 1). Six pipelines were selected as they were specialized for bulk RNA-seq and were taking as input raw FASTQ files: MIGNON[9], Latch[38], RNAflow[39], SPEAQeasy[40], sequana_rnaseq[41], and GenPipes RNA[42]. A key observation is that most of these pipelines primarily focus on differential gene expression analysis, with only two (MIGNON and SPEAQeasy) supporting variant calling. This highlights that RnaXtract will operate in a distinct niche compared to existing tools. The primary objective of RnaXtract

is to maximize the extraction of raw information from bulk RNA sequencing, offering outputs such as gene expression, variant calling, and novel cell deconvolution in a format that is easily accessible for machine learning applications. Notably, the other tools performing variant calling present variants in VCF format, which is less suited for downstream machine learning applications.

Interest of cell deconvolution

Cell deconvolution has proven its utility in machine learning frameworks by enhancing predictive accuracy and uncovering biologically meaningful features from bulk RNA-seq data. The integration of immune cell fractions, particularly tumor-associated neutrophils, significantly improved chemotherapy response prediction in gastric cancer, with deconvolution-derived features emerging among the most important variables in the model[43]. Similarly, large-scale deconvolution of AML cohorts using single-cell references has reconstructed differentiation trajectories and predicted drug resistance with high fidelity [44]. Together, these studies demonstrate how cell deconvolution transforms bulk transcriptomic data into a powerful resource bridging tissue-level insights with single-cell resolution.

Conclusions

RnaXtract provides an automated RNA-seq data analysis workflow, transforming raw sequencing reads into structured, machine learning-compatible datasets. It offers a robust framework for extracting genomic, transcriptomic, and cellular composition data efficiently. Leveraging state-of-the-art methods for read processing, RnaXtract integrates the CIBERSORTx/EcoTyper deep learning model to deliver detailed cellular and ecotype profiling of samples. This makes it highly applicable for advancing personalized medicine, particularly in predicting cancer trajectories, due to its machine learning-compatible output format. Furthermore, RnaXtract can be deployed seamlessly across diverse computational environments, ensuring optimized resource utilization. Its modular architecture facilitates straightforward upgrades and maintenance, enhancing adaptability for evolving research needs.

Data availability

The RnaXtract pipeline is publicly available at https://gitlab-pub.compbio.ulaval.ca/sophiane/rnaxtract under the GNU-GPL 3 licence and the documentation is accessible within the repository. The samples used to test RnaXtract accession number of this manuscript are available in the supplementary data file.

Received: 16 May 2025; Accepted: 20 August 2025
Published online: 24 August 2025

References

1. Ximin Li, C.-Y. W. From bulk, single-cell to spatial RNA sequencing. *Int. J. Oral Sci*. **13**, (2021).
2. Mangul, S. et al. ROP: Dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues. *Genome Biol.* **19**, 36 (2018).
3. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
4. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* **2**, e107 (2023).
5. Shibaguchi, H. & Yasutaka, Y. HiSAT: A novel method for the rapid diagnosis of allergy. *Bio. Protoc.* **12**, e4309 (2022).
6. Dobin, A. et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
7. McKenna, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
8. Kim, S. et al. Strelka2: Fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
9. Garrido-Rodriguez, M. et al. A versatile workflow to integrate RNA-seq genomic and transcriptomic data into mechanistic models of signaling pathways. *PLoS Comput. Biol.* **17**, e1008748 (2021).
10. Zhi-Xiong, C. Single-Cell RNA sequencing in ovarian cancer: Current progress and future prospects. *Prog. Biophys. Mol. Biol.* https://doi.org/10.1016/j.pbiomolbio.2025.01.002 (2025).
11. Newman, A. M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
12. Fan, J. *et al.* MuSiC2: Cell-type deconvolution for multi-condition bulk RNA-seq data. *Brief. Bioinform.* **23**, (2022).
13. Zeng, T., Huang, T. & Lu, C. *Machine Learning Advanced Dynamic Omics Data Analysis for Precision Medicine*. (Frontiers Media SA, 2020).
14. El Naqa, I., Li, R. & Murphy, M. J. *Machine Learning in Radiation Oncology: Theory and Applications*. (Springer, 2015).
15. Saghaleyni, R., Sheikh Muhammad, A., Bangalore, P., Nielsen, J. & Robinson, J. L. Machine learning-based investigation of the cancer protein secretory pathway. *PLoS Comput. Biol.* **17**, e1008898 (2021).
16. Luca, B. A. et al. Atlas of clinically distinct cell states and ecosystems across human solid tumors. *Cell* **184**, 5482-5496.e28 (2021).
17. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res* **10**, 33 (2021).
18. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLoS ONE* **12**, e0177459 (2017).
19. The R Project for Statistical Computing. https://www.R-project.org/.
20. Van Rossum, G. & Drake, F. L., Jr. *The Python Language Reference Manual*. (Network Theory, 2011).
21. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
22. Everaert, C. et al. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Sci. Rep.* **7**, 1559 (2017).
23. Heldenbrand, J. R. et al. Recommendations for performance optimizations when using GATK3.8 and GATK4. *BMC Bioinform.* **20**, 557 (2019).
24. Brouard, J.-S., Schenkel, F., Marete, A. & Bissonnette, N. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. *J. Anim. Sci. Biotechnol.* **10**, 44 (2019).
25. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 1–14 (2020).

26. Ren, Y. et al. Deconvolution reveals cell-type-specific transcriptomic changes in the aging mouse brain. *Sci. Rep.* **13**, 16855 (2023).
27. Bouzeraa, L. et al. Decoding epigenetic markers: Implications of traits and genes through DNA methylation in resilience and susceptibility to mastitis in dairy cows. *Epigenetics* **19**, 2391602 (2024).
28. Leitão, A. B. et al. Recognition of nonself is necessary to activate Drosophila's immune response against an insect parasite. *BMC Biol.* **22**, 89 (2024).
29. Derouane, F. et al. Metabolic adaptation towards glycolysis supports resistance to neoadjuvant chemotherapy in early triple negative breast cancers. *Breast Cancer Res.* **26**, 29 (2024).
30. CZI Cell Science Program *et al.* CZ CELLxGENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.* **53**, D886–D900 (2025).
31. Wu, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
32. Leclercq, M. et al. Large-scale automatic feature selection for biomarker discovery in high-dimensional OMICs data. *Front. Genet.* **10**, 452 (2019).
33. Scotto, L. et al. Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: Potential role in progression. *Genes Chromosomes Cancer* **47**, 755–765 (2008).
34. Turn, R. E. *et al.* The ARF GAPs ELMOD1 and ELMOD3 act at the Golgi and cilia to regulate ciliogenesis and ciliary protein traffic. *Mol. Biol. Cell.* **33**, cor1 (2022).
35. Ferreira, A., Castanheira, P., Escrevente, C., Barral, D. C. & Barona, T. Membrane trafficking alterations in breast cancer progression. *Front. Cell Dev. Biol.* **12**, 1350097 (2024).
36. Kosinsky, R. L. et al. USP22-dependent HSP90AB1 expression promotes resistance to HSP90 inhibition in mammary and colorectal cancer. *Cell Death Dis.* **10**, 911 (2019).
37. Website. https://www.techpowerup.com/cpu-specs/epyc-7601.c1920.
38. Le Hannah G. B., H. *et al.* Latch Verified Bulk-RNA Seq toolkit: a cloud-based suite of workflows for bulk RNA-seq quality control, analysis, and functional enrichment. *bioRxiv* (2022). https://doi.org/10.1101/2022.11.10.516016.
39. Lataretu, M. & Hölzer, M. RNAflow: An effective and simple RNA-Seq differential gene expression pipeline using Nextflow. *Genes (Basel).* **11**, (2020).
40. Eagles, N. J. et al. Correction to: SPEAQeasy: A scalable pipeline for expression analysis and quantification for R/bioconductor-powered RNA-seq analyses. *BMC Bioinform.* **22**, 381 (2021).
41. Cokelaer, T., Desvillechabrol, D., Legendre, R. & Cardon, M. Sequana': A set of Snakemake NGS pipelines. *J. Open Source Softw.* **2**, 352 (2017).
42. Bourgey, M. *et al.* GenPipes: An open-source framework for distributed and scalable genomic analyses. *Gigascience*. **8**, (2019).
43. Sasagawa, S. et al. Predicting chemotherapy responsiveness in gastric cancer through machine learning analysis of genome, immune, and neutrophil signatures. *Gastr. Cancer* **28**, 228–244 (2025).
44. Karakaslar, E. O. et al. A transcriptomic based deconvolution framework for assessing differentiation stages and drug responses of AML. *NPJ Precis. Oncol.* **8**, 105 (2024).

Acknowledgements

We thank all lab members for many interesting discussions. This work was supported by the Fonds de recherche du Québec (FRQ) through the research centre grant for the CHU de Québec-Université Laval Research Center (reference: 30641).

Author contributions

SGB software, conceptualization, formal analysis, methodology, writing—original draft. SG software, conceptualization. ML supervision, conceptualization, writing—review & editing CJB software, conceptualization ET conceptualization, methodology SB supervision, fund acquisition, writing—review & editing, supporting conceptualization AD supervision, resources and fund acquisition, writing—review & editing.

Funding

This study was funded by Natural Sciences and Engineering Research Council of Canada (#2019-06490, CG118883) and Canadian Institutes of Health Research (#451568).

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

The datasets used in this study are available under the accession number PRJNA1004593 and PRJEB34009 in the ENA and were published by Derouane et al.[29] (Breast Cancer Res 26, 29) and Kosinsky et al.[36] (Cell Death Dis 10, 911) respectively.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-16875-9.

Correspondence and requests for materials should be addressed to S.B. or A.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.