*Published online 25 July 2022*

*NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 3* **1**
*https://doi.org/10.1093/nargab/lqac053*

# iCOMIC: a graphical interface-driven bioinformatics pipeline for analyzing cancer omics data

**Anjana Anilkumar Sithara[1,2,3], Devi Priyanka Maripuri[1,2,3], Keerthika Moorthy[1,2,3], Sai Sruthi Amirtha Ganesh[1,2,3], Philge Philip[2,3], Shayantan Banerjee[1,2,3], Malvika Sudhakar[1,2,3] and Karthik Raman[1,2,3,*]**

[1]Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology (IIT) Madras, Chennai 600036, India, [2]Centre for Integrative Biology and Systems mEdicine, IIT Madras, India and [3]Robert Bosch Centre for Data Science and Artificial Intelligence (RBCDSAI), IIT Madras, India

Received December 17, 2021; Revised June 17, 2022; Editorial Decision June 27, 2022; Accepted July 04, 2022

## ABSTRACT

**Despite the tremendous increase in omics data generated by modern sequencing technologies, their analysis can be tricky and often requires substantial expertise in bioinformatics. To address this concern, we have developed a user-friendly pipeline to analyze (cancer) genomic data that takes in raw sequencing data (FASTQ format) as input and outputs insightful statistics. Our iCOMIC toolkit pipeline featuring many independent workflows is embedded in the popular Snakemake workflow management system. It can analyze whole-genome and transcriptome data and is characterized by a user-friendly GUI that offers several advantages, including minimal execution steps and eliminating the need for complex command-line arguments. Notably, we have integrated algorithms developed in-house to predict pathogenicity among cancer-causing mutations and differentiate between tumor suppressor genes and oncogenes from somatic mutation data. We benchmarked our tool against Genome In A Bottle benchmark dataset (NA12878) and got the highest F1 score of 0.971 and 0.988 for indels and SNPs, respectively, using the BWA MEM—GATK HC DNA-Seq pipeline. Similarly, we achieved a correlation coefficient of $r = 0.85$ using the HISAT2-StringTie-ballgown and STAR-StringTie-ballgown RNA-Seq pipelines on the human monocyte dataset (SRP082682). Overall, our tool enables easy analyses of omics datasets, significantly ameliorating complex data analysis pipelines.**

## INTRODUCTION

Over the past couple of decades, genomic research has developed tremendously due to the rise of Next-Generation Sequencing (NGS) technologies. These rapid advances have had a considerable impact in the realm of sequence-based analysis: NGS has enabled researchers to discover novel DNA and RNA variants (1), along with differentially expressed genes (2). Whole Genome/Exome Sequencing (DNA-Seq) pipelines identify nucleotide variants, while RNA Sequencing (RNA-Seq) enables quantification of gene expression. Whole Genome Sequencing further serves as a powerful tool in analyzing mutations in the context of cancer (3,4) and is also a bedrock for personalized medicine (5). RNA-Seq further refines the essential interpretation of various biological phenomena.

Various bioinformatics tools have been developed to analyze the large amounts of data generated by NGS technologies. Data analysis poses a major hindrance to biologists, exacerbating the need for an automated pipeline. Even though new tools for genomic data analysis are being developed from time to time, a comprehensive toolkit does not exist. Extensive comparative studies that deal with different combinations of tools have been conducted (6–8). Although software suites consisting of a combination of few tools exist (9–14), a user-friendly toolkit to aid non-programmers, incorporating a surplus number of bioinformatics tools, is missing (15). Moreover, there is a dearth of open-source bioinformatics pipelines that enables comprehensive analysis of large (cancer) genomic datasets (16).

To make life easier for clinical researchers and biologists, we developed iCOMIC (**i**ntegrating the **CO**ntext of **M**utations **I**n **C**ancer), an open-source, standalone tool characterized by a Python-based Graphical User Interface and automated Bioinformatics pipelines for DNA-Seq and RNA-Seq data analysis. It serves as a point-and-click application facilitating genomic data analysis accessible to users

*To whom correspondence should be addressed. Tel: +91 44 2257 4139; Email: kraman@iitm.ac.in
Present address: Devi Priyanka Maripuri, Faculty of Medicine & Dentistry, University of Alberta, Edmonton T6G 2R3, Canada.

© The Author(s) 2022. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

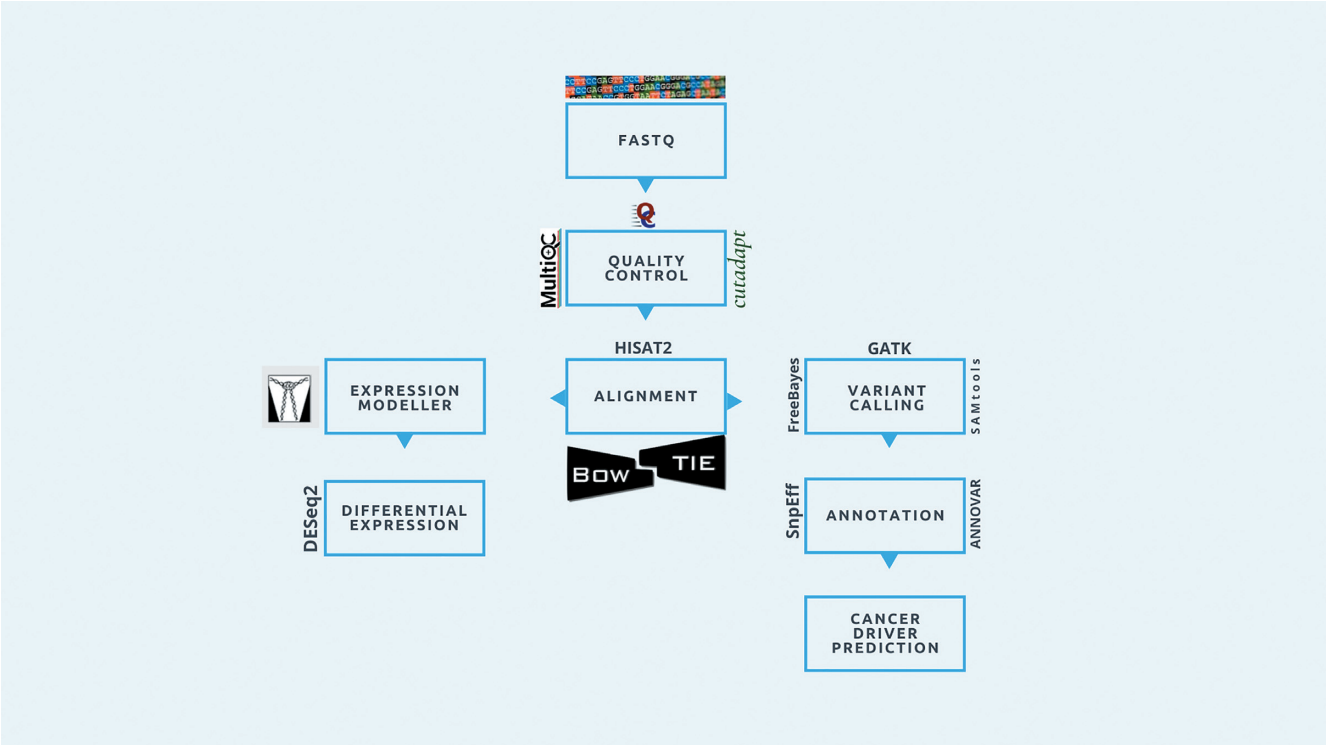


**Figure 1.** Schema for iCOMIC pipeline. Multiple workflows are embedded in iCOMIC providing users with the complete freedom to choose from the integrated tools. Both DNA-Seq and RNA-Seq pipelines take in raw FASTQ files as input. Quality control and alignment are common steps in both pipelines. FastQC and Cutadapt are the Quality control tools used and MultiQC is used to generate a consolidated report on Quality statistics. Analysis of RNA-Seq data includes mapping of sequencing reads to a reference genome using Aligner, Quantification of expression levels using Expression modeller and Differential expression analysis. On the other hand, steps in DNA-Seq analysis include Alignment followed by identifying the variants and annotating them. Tools incorporated in iCOMIC are listed in Table 1.

**Table 1.** List of tools incorporated in iCOMIC along with their corresponding functions

| Function | DNA-Seq tools | RNA-Seq tools |
|---|---|---|
| Quality control | FastQC, MultiQC, Cutadapt | FastQC, MultiQC, Cutadapt |
| Alignment | GEM-Mapper v3, BWA-MEM, Bowtie2, | STAR, HISAT2 |
| Variant calling | GATK HC, samtools mpileup, FreeBayes, GATK Mutect2 | - |
| Annotation | Annovar, SnpEff | - |
| Quantification of expression levels | - | StringTie, HTSeq |
| Differential expression | - | DESeq2, ballgown |

with minimal programming expertise. iCOMIC provides a versatile, fully automated pipeline for the analysis of genomic data, with a user-defined combination of tools and a set of easily tunable parameters. In addition, iCOMIC grants users the privilege to customize pipelines in less than five simple steps, integrating a wide range of bioinformatics tools.

## MATERIALS AND METHODS

The first step towards developing iCOMIC was to compile a set of tools following best practices for DNA-Seq and RNA-Seq analysis. We identified the most widely used tools for alignment of reads, variant calling, variant annotation, expression modeling and differential expression analysis (Figure 1, Table 1) (17). Two of the most cited and widely used tools, FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Cutadapt (18), were chosen for quality control.

Aligner tools incorporated in iCOMIC for DNA-Seq data include BWA-MEM (19), GEM-Mapper v3 (20) and Bowtie 2 (21). Apart from the three major steps involved in analyzing whole-genome sequence data, several preprocessing steps are carried out intermediately following the GATK framework. This included sorting and indexing alignment files generated by the aligner, marking duplicates using Picard markduplicates (https://github.com/broadinstitute/picard), and base quality score recalibration followed by filtration of the variants called. Variant callers include GATK Mutect2 (22), samtools mpileup (23), FreeBayes (24) and GATK Haplotype-Caller (25). Mutect2 identifies variants in normal-tumor pairs while the rest of the tools perform variant calling in a given sample analogous to the reference genome. SnpEff (26) and ANNOVAR

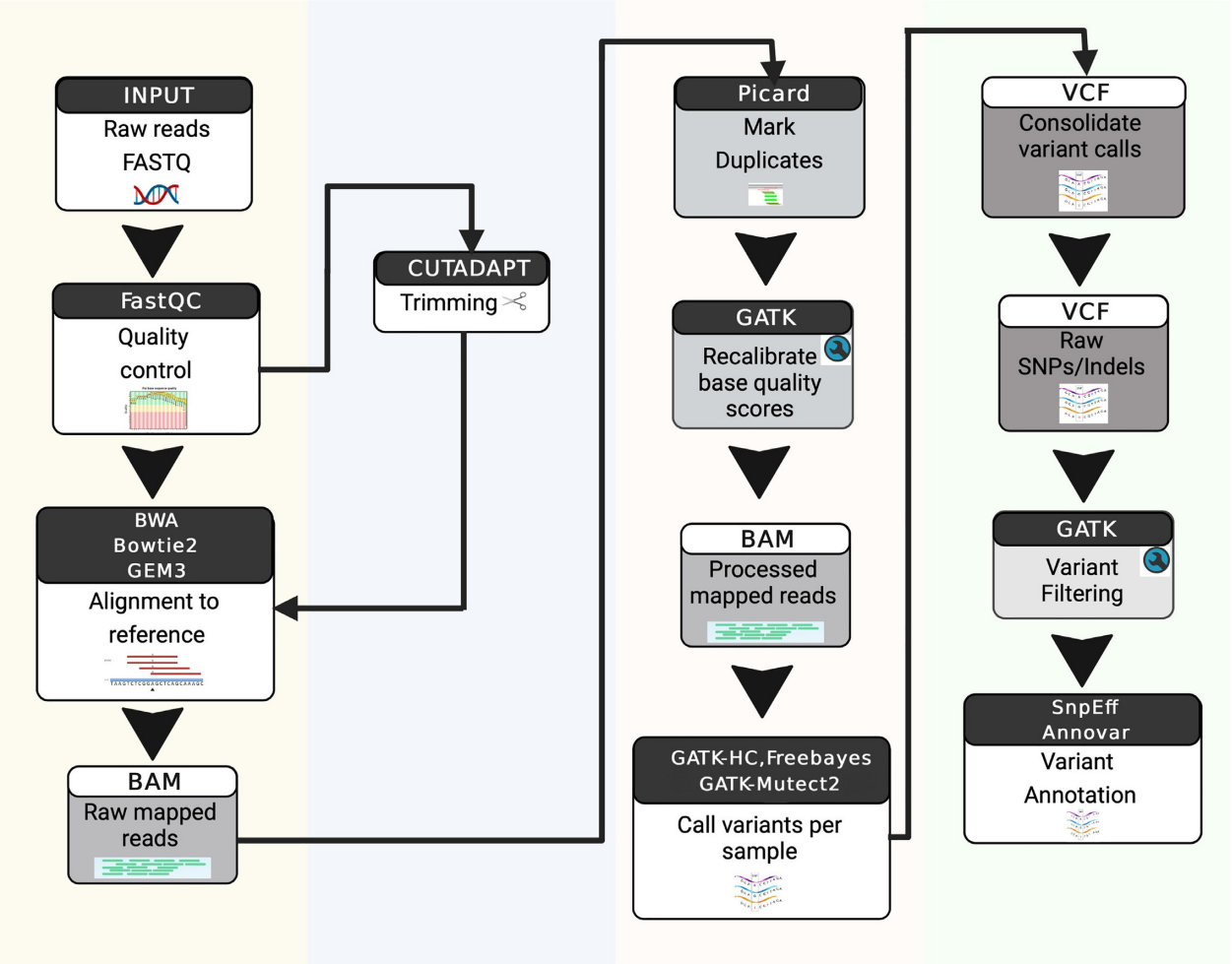


**Figure 2.** Schematic diagram of DNA-Seq pipeline. The input, followed by the application of various quality control techniques, alignment to the reference genome, variant calling, filtering and annotation are indicated in this figure.

(27) were the tools selected for variant annotation. The tool MultiQC (28) was used to aggregate the results obtained from the numerous tools in the workflow. In the case of whole genome/exome data, the MultiQC report was generated based on the results from tools such as FastQC and SnpEff. Two cancer-related data analysis tools, NBDriver (29), which uses a machine learning approach to identify the context of mutations, and cTaG (30), a tool to predict whether a given gene is a tumor suppressor (TSG) or an oncogene (OG), were also incorporated in the final pipeline (Figure 2).

For RNA-seq data, HISAT2 (31) and STAR (32) were selected to perform alignment following quality control. Expression modellers such as StringTie (33) and HTSeq (34) were included to count the number of reads aligned to the genome. For differential expression analysis, ballgown (35) and DESeq2 (36) were incorporated. The MultiQC report for RNA-Seq data includes results from tools such as FastQC, Cutadapt and STAR. Normalization is inherently carried out by the tool used for differential expression (Figure 3).

iCOMIC is embedded in the popular workflow management system, Snakemake (37). The analysis workflow has 'rules' as the building blocks, which describe the connection between the input and output (38) (Figure 4). It enables easy connectivity between the different tools/software within the workflow. The 'rules' specify input files, output files, log files and wrapper/shell commands. The Snakemake wrapper scripts together with the conda environment manage the automated installation of software and their dependencies. Tools without a wrapper script are configured separately, and shell commands are used for its execution. Parallelization of workflows is managed by Snakemake. iCOMIC users have the privilege to choose the number of cores they need to run iCOMIC with, thereby allowing multiple jobs to be executed together (this option is given in the input tab). The number of cores provided by the user is passed as an input to each tool. The list of dependencies and the versions can be exported from the conda environment file (https://github.com/RamanLab/iCOMIC/blob/main/icomic_env.yml).

The details of individual tools incorporated in iCOMIC are available here (https://icomic-doc.readthedocs.io/en/latest/walkthrough.html).

According to the user's selection, appropriate rules are combined in a 'Snakefile' to generate the target output. The

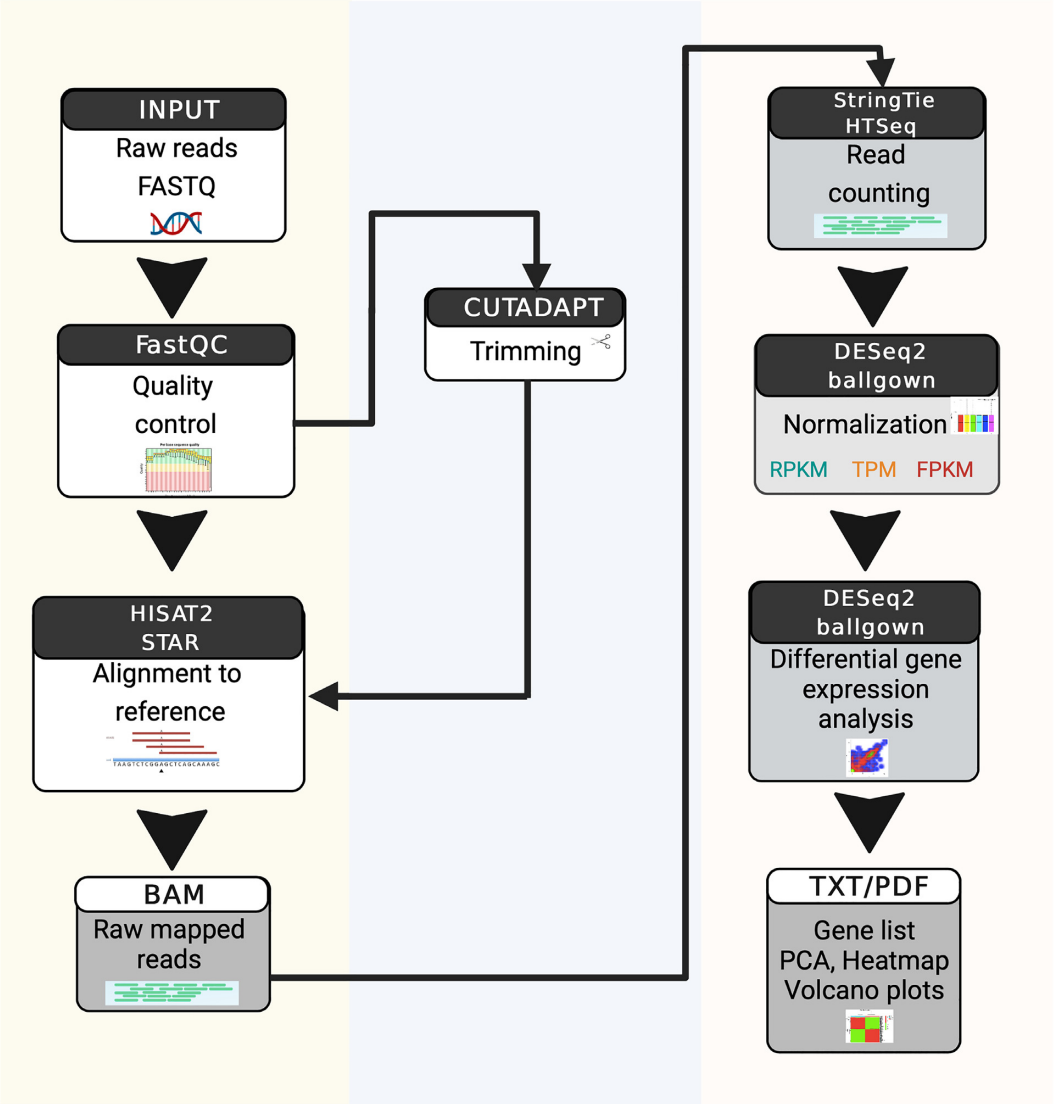


**Figure 3.** Schematic diagram of RNA-Seq pipeline. The input, followed by the application of various quality control techniques, alignment to the reference genome, counting the mapped reads, normalization, and differential expression analysis, ultimately generating the TXT/PDF output is detailed in this figure.

input file paths and parameters set by the user are automatically fed into a configuration file, referred to as the 'config file.' Multiple config files and Snakefiles are auto-generated for the quality check of the input data, generating genome index, and executing the main workflow.

We implemented the iCOMIC pipeline in the form of a Graphical User Interface (GUI). The GUI has been developed using PyQt5, a Python binding of the cross-platform GUI toolkit Qt. The GUI retrieves the user input files and the parameters and communicates with the Snakemake rules to set up the analysis. A Python wrapper binds together the PyQt5 GUI and the Snakemake workflow of iCOMIC. After completing the initial requirements and creating the conda environment, the iCOMIC GUI can be accessed using a single command, 'icomic'. Additionally, running iCOMIC in Linux platform is supporte by docker.

## RESULTS AND DISCUSSION

### Major features of the pipeline

NGS data has become an indispensable tool for biological research, although the data analysis can be daunting for non-bioinformaticians. iCOMIC has been introduced to overcome this concern to a certain extent. It serves as a stand-alone end-to-end analysis toolkit for DNA-Seq and RNA-Seq data. The DNA-Seq component of iCOMIC supports both germline and somatic variant calling. In conventional analysis pipelines like Galaxy, workflows need to be built, whereas iCOMIC has various inbuilt pipelines that automatically transfer output from one tool to the next. iCOMIC provides an interactive and user-friendly GUI, specifically created to accommodate users with minimal programming expertise. On another note, iCOMIC allows expert bioinformaticians to perform analysis incorporat-

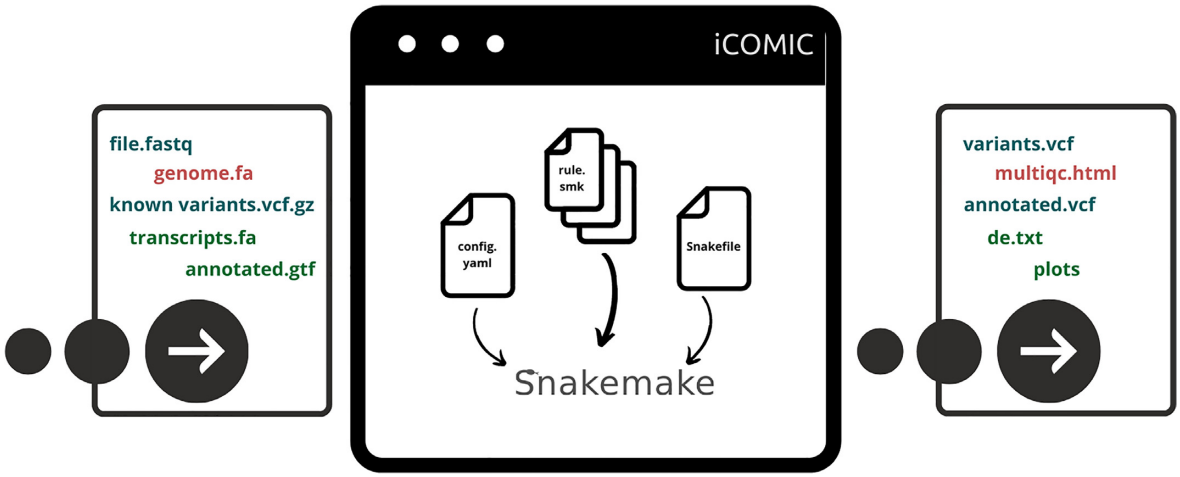


**Figure 4.** Snakemake workflow management system. All the input and output files in blue colour are those corresponding to DNA-Seq analysis and those in green correspond to RNA-seq analysis. The common files for DNA and RNA-Seq analysis are represented in red. 'Rule' files specifying the input, output and the shell/wrapper script form the basic units of Snakemake. Each rule corresponds to individual tools. The additional parameters for the tools are indicated in the 'config' file. According to the choice of tools made by the user, rules are integrated into the Snakefile and the workflow is executed.

ing additional tools and advanced parameters, saving time building the pipeline. The steps to be followed for writing new rules for integrating additional tools are detailed in the documentation at https://icomic-doc.readthedocs.io/en/latest/. The GUI is embedded in a Python wrapper script that connects the Snakemake workflows. Users can select an array of tools from the predesigned combinations best suited for their requirements. The best connectivity between the tools has been taken into account to design these individual workflows. iCOMIC provides the user with the necessary means to replace modules or alter the pipeline. Furthermore, the conda environment ensures easy installation of tools and dependencies. A detailed description of the structure of GUI together with step-by-step screenshots of analyzing an example pipeline is provided in the Supplementary Methods [Sections 1–3].

### Prediction of tumour suppressor genes and oncogenes using the cTaG algorithm

The cTaG (classify TSG and OG) model identifies driver genes by classifying them as tumor suppressor genes (TSGs) and oncogenes (OGs). Given a cohort of samples, the pan-cancer model calculates ratio-metric features from somatic mutation data, capturing mutations' functional impact. Unlike other computational methods that use background mutation rate (BMR) to identify genes with a higher mutation rate as driver genes, cTaG captures the effect of a mutation on the gene's functionality. Methods using BMR are biased towards genes with high mutation rates (39), and we know that while few driver genes have a high mutation rate, most do not. The mutations in TSG and OG differ; we found nonsense mutations more commonly found in TSGs than OGs. The cTaG method contains binary classifiers that classify genes as TSG or OG. The genes are labeled as TSGs or OGs based on consensus across various models.

To build a pan-cancer model, the model was trained on somatic mutation data from COSMIC (v79) (40) from different cancer types. We used ratio-metric and entropy features to classify genes as TSG or OG. The cTaG model uses the random forest classification algorithm to generate the pan-cancer model. A filtered list of genes from the Cancer Gene Census (CGC) (41) is used to label genes as TSG or OG, used to train the pan-cancer model. The pan-cancer model successfully identified tissue-specific driver genes when employed on a cohort from single tissue of origin. The method takes a single maf file with annotated mutations for the cohort of samples as input and generates each gene's ratio-metric and entropy features. Additional arguments such as the percentile and threshold to define highly-mutated samples can be specified. The percentile argument defines the top percentile genes to be considered from predictions made by each model for the final consensus. The default value is 5. Highly mutated samples are skipped from the analysis. By default, samples with >2000 mutations are omitted during analysis. The cTaG model labels these genes as TSG and OG based on the consensus across the random forest models. The model returns the list of all genes and their labels predicted by each model, along with its presence in the top percentile. Our method identifies genes with high as well as low mutation rates. The pan-cancer model also predicts tissue-specific driver genes.

### Prediction of driver and passenger mutations using NBDriver

Differentiating between driver and passenger mutations from sequenced cancer genomes is essential to targeted therapy and precision medicine. Despite the dramatic advances in developing predictive algorithms to differentiate between driver and passenger mutations, very few have concentrated on utilizing the local sequence context as potential features for further analysis. To capture this information, we built a robust machine learning model called NBDriver, which uses raw nucleotide sequences surrounding cancer-causing mutations as features to build machine learning models.

**Table 2.** Summary of germline variant benchmarking with NA12878/HG001 dataset

| Workflow | Type | Recall | Precision | *F*1 score |
|---|---|---|---|---|
| BWA MEM-GATK HC-SnpEff | INDEL | 0.967 | 0.976 | 0.971 |
| | SNP | 0.978 | 0.998 | 0.988 |
| BWA MEM-freebayes-SnpEff | INDEL | 0.931 | 0.917 | 0.924 |
| | SNP | 0.979 | 0.997 | 0.988 |
| BWA MEM-GATK HC-Annovar | INDEL | 0.967 | 0.976 | 0.971 |
| | SNP | 0.978 | 0.998 | 0.988 |
| BWA MEM-Bcftools-Annovar | INDEL | 0.741 | 0.838 | 0.789 |
| | SNP | 0.976 | 0.996 | 0.986 |
| Gem3-GATK HC-SnpEff | INDEL | 0.964 | 0.978 | 0.971 |
| | SNP | 0.977 | 0.999 | 0.988 |
| Gem3-Freebayes-SnpEff | INDEL | 0.934 | 0.92 | 0.927 |
| | SNP | 0.978 | 0.998 | 0.988 |
| BWA MEM-GATK HC-Annovar | INDEL | 0.967 | 0.976 | 0.971 |
| | SNP | 0.978 | 0.998 | 0.988 |
| BWA MEM-Freebayes-Annovar | INDEL | 0.931 | 0.917 | 0.924 |
| | SNP | 0.979 | 0.997 | 0.988 |
| Gem3-GATK HC-Annovar | INDEL | 0.964 | 0.978 | 0.971 |
| | SNP | 0.977 | 0.999 | 0.988 |
| Gem3-Freebayes-Annovar | INDEL | 0.934 | 0.92 | 0.927 |
| | SNP | 0.978 | 0.998 | 0.988 |
| BWA MEM-Bcftools-SnpEff | INDEL | 0.741 | 0.838 | 0.789 |
| | SNP | 0.976 | 0.996 | 0.986 |
| Gem3-Bcftools-SnpEff | INDEL | 0.781 | 0.353 | 0.486 |
| | SNP | 0.975 | 0.997 | 0.986 |
| Bowtie2-GATK HC-SnpEff | INDEL | 0.847 | 0.978 | 0.908 |
| | SNP | 0.953 | 0.998 | 0.975 |
| Bowtie2-GATK HC-Annovar | INDEL | 0.847 | 0.978 | 0.908 |
| | SNP | 0.953 | 0.998 | 0.975 |
| Bowtie2-Freebayes-SnpEff | INDEL | 0.717 | 0.909 | 0.802 |
| | SNP | 0.945 | 0.996 | 0.97 |
| Bowtie2-Freebayes-Annovar | INDEL | 0.717 | 0.908 | 0.802 |
| | SNP | 0.945 | 0.996 | 0.97 |
| Bowtie2-Bcftools-SnpEff | INDEL | 0.648 | 0.891 | 0.75 |
| | SNP | 0.944 | 0.985 | 0.964 |
| Bowtie2-Bcftools-Annovar | INDEL | 0.648 | 0.891 | 0.75 |
| | SNP | 0.944 | 0.985 | 0.964 |

Our training data consisted of missense mutations from 58 genes containing experimentally validated functional impacts from several studies. To obtain a numerical representation of the sequence features surrounding the mutations, we used commonly used natural language processing tools such as the TF-IDF Vectorizer, the Count Vectorizer and the One-Hot Encoder. Using kernel density estimation techniques, we showed that the underlying distributions of the neighbourhood sequences surrounding driver and passenger mutations are significantly different from one another. We utilized this information to build robust machine learning models using a repeated cross-validation strategy and report the median values of the performance metrics for each feature-classifier pair. To increase the prediction performance, we integrated sequence features derived from raw nucleotide sequences with other genomic, structural, and evolutionary features, resulting in the development of a pan-cancer mutation effect prediction tool, NBDriver, which was highly efficient in identifying pathogenic variants from five independent validation datasets. An ensemble predictor containing NBDriver, CONDEL (42) and MutationTaster (43) outperformed existing pan-cancer models in prioritizing a literature-curated list of driver and passenger mutations. Considering only the true positive mutation predictions from NBDriver, we identified a list of 138 driver genes with known functional evidence from multiple sources. Overall, our study underpinned the efficacy of utilizing raw nucleotide sequences as features for building robust machine learning models to distinguish between driver and passenger mutations.

## Benchmarking of tools

*Evaluating the performance of germline variant calling pipelines.* Performance validation of the germline variant calling workflows available with iCOMIC was done using Genome In A Bottle (GIAB) benchmark sets, based on Zook *et al.* (44). To this end, we ran 18 different combinations of aligners, germline variant callers, and annotators integrated with iCOMIC on the widely used benchmark dataset NA12878/HG001 to obtain separate vcfs (Table 2). Generated vcfs were then compared with GIAB/NIST HG001 v2.19 truth data, restricting the comparison to the GIAB v2.19 BED file coordinates. We adopted the benchmark framework developed by the Global Alliance for Genomics and Health (GA4GH) benchmarking team (45) for vcf comparison. The method involved the generation of an intermediate vcf with a standardized variant representation using Vcfeval by Real-Time Genomics (RTG) tools (46) followed by quantification of performance metrics using qfy.py, which is a part of hap.py benchmarking toolkit. The

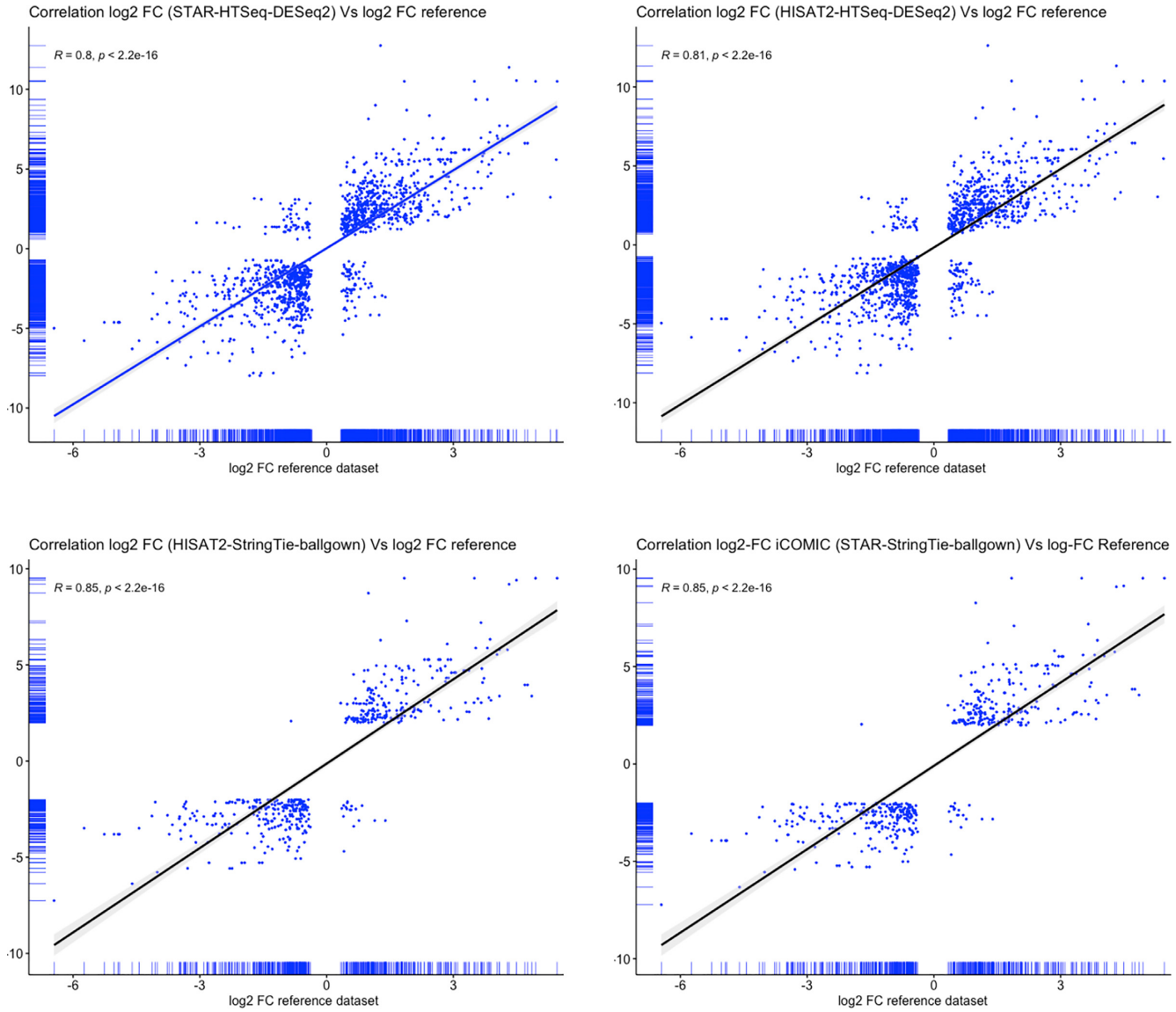


**Figure 5.** Fold change correlation between iCOMIC and reference dataset for the four workflows. The Pearson correlation coefficient was used to calculate fold changes.

summary of accuracy measures is highlighted in Table 2. The analysis was performed, specifying ten threads for running each tool for all the workflow combinations. iCOMIC provided the highest *F*1 score of 0.971 and 0.988 for indels and SNPs, respectively, for the BWA MEM - GATK HC pipeline among all the combinations (Table 2). Time consumed for analyzing the GIAB benchmark set for BWA MEM—GATK HC—SnpEff pipeline is provided in Section 4a of the Supplementary Methods.

*Evaluating the performance of RNA-Seq pipelines.* The validation of the performance of RNA-Seq workflows available with iCOMIC was done using the Human monocyte RNA-Seq dataset from NCBI-SRA (SRP082682). We ran four different combinations of aligners, expression modellers and differential expression tools integrated with iCOMIC on the RNA-Seq benchmark dataset to obtain differentially expressed genes between classical and non-classical monocytes. Comparison of differentially expressed genes between the genes identified and the non-classical and classical monocytes from the reference microarray dataset (GSE34515) obtained from the NCBI GEO database was performed. The fold change correlation (47) was calculated between reference microarray and RNA-Seq data. Based on the fold change correlation computed for the four different pipelines, it is evident that HISAT2-StringTie-ballgown and STAR-StringTie-ballgown pipelines performed the best with a correlation coefficient of $r = 0.85$, higher than the values obtained for HISAT2-HTSeq-DESeq2 and STAR-HTSeq-DESeq2 pipelines (Figure 5). Time consumed for analyzing the benchmark dataset using STAR-HTSeq-DESeq2 is available in Section 4b of the Supplementary Methods.

### Feature comparison between iCOMIC and other bioinformatics pipelines

There exist many pipelines that integrate different bioinformatics tools for genomic data analysis. A comprehensive feature comparison with previously developed workflows (snakePipes, Sequanix, Omics Pipe, Galaxy (48), GenPipes (49), CANEapp, ARMOR (50), Galaxy, VIPER (51), systemPipeR (52), CLC Genomics Workbench and nf-core (53)) was performed to highlight the significance of iCOMIC (Table 4). The accessibility of the pipeline through GUI, availability to the public, cloud support, the ability of automated execution of an entire pipeline, user freedom to choose tools of interest, and programming skills required for the analysis include some of the features used for comparison. Only those tools that share common features with iCOMIC were considered for comparison. In addition, a comparison with the most popular bioinformatics tools like Galaxy and CLC genomics workbench was performed. Galaxy, one of the customary scientific platforms for bioinformatics data analysis, does not afford in-built pipelines, whereas CLC genomics workbench is not open source. The comparative analysis highlights the open-source GUI with end-to-end automated analysis integrating a plethora of tools as the strongest aspects of iCOMIC.

*Performance comparisons between iCOMIC and GALAXY.* Galaxy is a popular publicly available web-based interface consisting of many tool combinations and workflows to perform various studies. It offers ease of access to complex computational analyses to users with minimal programming expertise and the functionality to examine large datasets in a multi-step process. Considering the similarities in features between iCOMIC and Galaxy, pipeline validation was conducted using the same methods to establish a comparison scale for performance.

Pipelines are fully automated in iCOMIC and the most predominantly used workflows are pre-built. While workflows need to be specified in Galaxy in order to automate, iCOMIC comes with a set of preinstalled pipelines for both DNA and RNA-Seq analysis. In iCOMIC, the user has the option to specify a large number of files in the form of a table, while no such provision is available on Galaxy. Although extensive resources and tutorials are provided for the use of Galaxy, we believe it has a steeper learning curve than iCOMIC, which is mostly a point-and-click type application. Certain tools are unavailable on the Galaxy server and need to be installed from the tool shed, while iCOMIC installations are readily done via the conda environment, with minimal user input/interference.

While analysing multiple samples on a workflow in Galaxy, it is required to rename the files prior to passing it as input to the next tool in the pipeline which can be tedious when there is a large number of samples. On the other hand with iCOMIC, the entire process is automated. Undoubtedly, Galaxy has its advantages, and is a very popular pipeline for genomic data analysis. Yet, iCOMIC excels in its simplicity, and we believe it will be more inviting for biologists and clinical researchers to quickly analyse their data.

***Comparison in terms of germline variant calling.*** DNA-Seq analysis was performed using BWA-MEM, Freebayes, and SnpEff for each step in the analysis workflow using GIAB dataset. Performance validation for the same was done using a process similar to that of iCOMIC. The variant files obtained as output from this pipeline were compared with GIAB/NIST HG001 v2.19 truth data restricting the comparison to the GIAB v2.19 BED file coordinates. Vcfeval method was used for vcf comparison, and the quantification of performance metrics was computed using the hap.py algorithm. Comparison of the performance metrics such as *F*1 and precision scores between iCOMIC and Galaxy indicates that the values are pretty similar. Considering the SNP *F*1 score, iCOMIC has 0.988 and Galaxy has 0.980, proving that the performance of iCOMIC is on par with that of Galaxy (Table 3).

**Table 3.** Summary of germline variant benchmarking with NA12878/HG001 dataset using Galaxy

| Workflow | Type | Recall | Precision | *F*1 score |
|---|---|---|---|---|
| BWA MEM-freebayes-SnpEff | INDEL | 0.887 | 0.948 | 0.917 |
| | SNP | 0.976 | 0.984 | 0.980 |

***Comparison in terms of differential expression analysis.*** RNA-Seq analysis was performed using the tools STAR, HTSeq, and DESeq2 for each step in the analysis workflow using the Human monocyte RNA-Seq dataset from NCBI-SRA (SRP082682). The fold change correlation was calculated between the reference microarray dataset (GSE34515) obtained from the NCBI GEO database and RNA-Seq data for iCOMIC and Galaxy. The values were found to be 0.8 and 0.66, respectively. Comparison of the fold change correlation between iCOMIC and Galaxy indicates that the value of iCOMIC is higher than that of GALAXY in the particular pipeline (Figure 6).

## CONCLUSION

Here, we present iCOMIC to analyze genomic data quickly. Analyzing the large amount of data generated by Next Generation Sequencing techniques can be tricky for a non-programmer as it requires computational skills. iCOMIC serves as a robust platform for users with minimal programming skills. iCOMIC enables the user to choose from several pre-configured workflows for analyzing Whole Genome/Exome Sequencing and RNA-Seq data. Notably, iCOMIC also integrates novel algorithms developed in-house to predict cancer driver and passenger mutations, as well as tumor suppressor genes and oncogenes. The iCOMIC toolkit can be downloaded as a package and run on Linux, Windows, or Mac operating systems. iCOMIC enables easy analysis through an interactive GUI and hassle-free installation of software. The features listed above make iCOMIC an easily accessible, open-source pipeline for large genomic datasets. DNA-Seq benchmarking study performed in iCOMIC with GIAB dataset resulted in an *F*1 score of 0.971 and 0.988 for indels and SNPs, respectively. Similarly, for RNA Seq, comparison with a microarray dataset produced a fold change correlation coefficient of 0.85.

**Table 4.** Comparison of iCOMIC with existing tools for genomic data analysis. 'Yes' specifies the presence of the feature, 'No' indicates the absence of a feature, 'Partial' indicates the presence of some aspects of the particular feature, and 'Not Specified' indicates that the information is not available. The features which are compared included: 1) The accessibility of the tool through Graphical User Interface, 2) commercial availability of the tool, 3) ability of the tool to run on the cloud, 4) automated analyses of the pipeline, 5) ability of the user to customize their own pipeline, 6) programming skills required for performing the analysis, 7) availability of DNA-Seq analysis pipeline, 8) availability of RNA-Seq analysis pipeline, 9) compatibility of the tool for cancer data. Furthermore, we have also listed out the platforms supported by the tools, programming language used to build the tool and workflow language used to write pipelines

| Features/tools | iCOMIC | snakePipes | Sequanix | Omics Pipe | GenPipes | CANEapp | Armor | Galaxy | VIPER | systemPipeR | CLC genomics workbench | nf-core |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GUI | Yes | No | Yes | No | No | Yes | No | Yes | No | No | Yes | No |
| Open source | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | No | Yes |
| Cloud support | No | Yes | No | Yes | Yes | Yes | No | Yes | No | Yes | No | Yes |
| Automated analysis | Yes | Yes | Yes | Yes | No | Yes | Partial | No | Not specified | Yes | Yes | Yes |
| Custom pipeline | Yes | Yes | Yes | Yes | Yes | Yes | Partial | Yes | No | Yes | No | Yes |
| Programming Skills not necessary | Yes | Partial | Yes | Partial | Partial | Yes | Partial | Yes | Partial | Partial | Yes | Partial |
| DNA-Seq analysis | Yes | Partial | Yes | Yes | Yes | No | No | Yes | No | No | Yes | Yes |
| RNA-Seq analysis | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Compatible for cancer data | Yes | Not specified | No | Yes | No | No | No | Yes | Yes | No | Yes | Yes |
| Platform supported | Linux, macOS (v10.15.5 and above), Windows OS* | Not specified | Linux | Not specified | Unix | Windows and mac(GUI), Linux(server side pipeline) | Linux, iOS | Linux, iOS | Unix, iOS | Not specified | Windows, iOS, Linux | Linux, iOS |
| Programming language | Python | Python | Python | Python | Python | Python, Java | R | Web based | Multiple | R | Java | Python |
| Workflow language | Snakemake | Snakemake | Snakemake | Ruffus | Not specified | Not specified | Snakemake | - | Snakemake | SYSargs | Not specified | Nextflow |

*The steps required to successfully install iCOMIC on Windows are discussed in the documentation (Section 2.4).




**Figure 6.** Fold change correlation between Galaxy and reference dataset for STAR-HTSeq-DESeq2 workflow**.** The Pearson correlation coefficient was used to calculate fold changes.

## DATA AVAILABILITY

iCOMIC source code is available here https://github.com/RamanLab/iCOMIC. iCOMIC user manual can be accessed using the link, https://icomic-doc.readthedocs.io/en/latest/user-guide.html.

## SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

## ACKNOWLEDGEMENTS

The authors acknowledge Mr. Likith Reddy for preliminary benchmarking studies and other members of the lab for useful discussions and Mr. Vimaladhasan Senthamizhan for his help with creating the dockerized version of the GUI.

## FUNDING

Department of Biotechnology, Government of India (DBT) [BT/PR16710/BID/7/680/2016], IIT Madras, Centre for Integrative Biology and Systems mEdicine (IBSE); Robert Bosch Center for Data Science and Artificial Intelligence (RBCDSAI).

*Conflict of interest statement.* None declared.

## REFERENCES

1. Qin,D. (2019) Next-generation sequencing and its clinical application. *Cancer Biol. Med.*, **16**, 4–10.

2. Kukurba,K.R. and Montgomery,S.B. (2015) RNA sequencing and analysis. *Cold Spring Harb. Protoc.*, **2015**, 951–969.
3. Alioto,T.S., Buchhalter,I., Derdak,S., Hutter,B., Eldridge,M.D., Hovig,E., Heisler,L.E., Beck,T.A., Simpson,J.T., Tonon,L. *et al.* (2015) A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.*, **6**, 10001.
4. Nakagawa,H. and Fujita,M. (2018) Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Sci.*, **109**, 513–522.
5. Nocq,J., Celton,M., Gendron,P., Lemieux,S. and Wilhelm,B.T. (2013) Harnessing virtual machines to simplify next-generation DNA sequencing analysis. *Bioinforma. Oxf. Engl.*, **29**, 2075–2083.
6. Williams,C.R., Baccarella,A., Parrish,J.Z. and Kim,C.C. (2017) Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*, **18**, 38.
7. Fisch,K.M., Meißner,T., Gioia,L., Ducom,J.-C., Carland,T.M., Loguercio,S. and Su,A.I. (2015) Omics pipe: a community-based framework for reproducible multi-omics data analysis. *Bioinforma. Oxf. Engl.*, **31**, 1724–1728.
8. Bhardwaj,V., Heyne,S., Sikora,K., Rabbani,L., Rauer,M., Kilpert,F., Richter,A.S., Ryan,D.P. and Manke,T. (2019) snakePipes: facilitating flexible, scalable and integrative epigenomic analysis. *Bioinformatics*, **35**, 4757–4759.
9. Asmann,Y.W., Middha,S., Hossain,A., Baheti,S., Li,Y., Chai,H.-S., Sun,Z., Duffy,P.H., Hadad,A.A., Nair,A. *et al.* (2012) TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data. *Bioinforma. Oxf. Engl.*, **28**, 277–278.
10. Fischer,M., Snajder,R., Pabinger,S., Dander,A., Schossig,A., Zschocke,J., Trajanoski,Z. and Stocker,G. (2012) SIMPLEX: cloud-enabled pipeline for the comprehensive analysis of exome sequencing data. *PLoS ONE*, **7**, e41948.
11. Germain,P.-L., Vitriolo,A., Adamo,A., Laise,P., Das,V. and Testa,G. (2016) RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Res.*, **44**, 5054–5067.
12. Lam,H.Y.K., Pan,C., Clark,M.J., Lacroute,P., Chen,R., Haraksingh,R., O'Huallachain,M., Gerstein,M.B., Kidd,J.M., Bustamante,C.D. *et al.* (2012) Detecting and annotating genetic variations using the hugeseq pipeline. *Nat. Biotechnol.*, **30**, 226–229.
13. Joo,T., Choi,J.-H., Lee,J.-H., Park,S.E., Jeon,Y., Jung,S.H. and Woo,H.G. (2019) SEQprocess: a modularized and customizable pipeline framework for NGS processing in r package. *BMC Bioinformatics*, **20**, 90.
14. Singer,J., Ruscheweyh,H.-J., Hofmann,A.L., Thurnherr,T., Singer,F., Toussaint,N.C., Ng,C.K.Y., Piscuoglio,S., Beisel,C., Christofori,G. *et al.* (2018) NGS-pipe: a flexible, easily extendable and highly configurable framework for NGS analysis. *Bioinforma. Oxf. Engl.*, **34**, 107–108.
15. Velmeshev,D., Lally,P., Magistri,M. and Faghihi,M.A. (2016) CANEapp: a user-friendly application for automated next generation transcriptomic data analysis. *BMC Genomics*, **17**, 49.
16. Liu,C.-H. and Di,Y.P. (2020) Analysis of RNA sequencing data using CLC genomics workbench. *Methods Mol. Biol. Clifton NJ*, **2102**, 61–113.
17. Hwang,S., Kim,E., Lee,I. and Marcotte,E.M. (2015) Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.*, **5**, 17875.
18. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.*, **17**, 10–12.
19. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinforma. Oxf. Engl.*, **25**, 1754–1760.
20. Marco-Sola,S., Sammeth,M., Guigó,R. and Ribeca,P. (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods*, **9**, 1185–1188.
21. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nat. Methods*, **9**, 357–359.
22. Benjamin,D., Sato,T., Cibulskis,K., Getz,G., Stewart,C. and Lichtenstein,L. (2019) Calling somatic SNVs and indels with mutect2. bioRxiv doi: https://doi.org/10.1101/861054, 02 December 2019, preprint: not peer reviewed.
23. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and and 1000 genome project data processing subgroupand 1000 genome project data processing subgroup (2009) the sequence alignment/map format and SAMtools. *Bioinforma. Oxf. Engl.*, **25**, 2078–2079.
24. Poplin,R., Ruano-Rubio,V., Depristo,M., Fennell,T., Carneiro,M., Auwera,G., Kling,D., Gauthier,L., Levy-Moonshine,A., Roazen,D. *et al.* (2017) Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv doi: https://doi.org/10.1101/201178, 24 July 2018, preprint: not peer reviewed.
25. Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. arXiv doi: https://arxiv.org/abs/1207.3907, 20 Jul 2012, preprint: not peer reviewed.
26. Cingolani,P., Platts,A., Wang,L.L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff. *Fly (Austin)*, **6**, 80–92.
27. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
28. Ewels,P., Magnusson,M., Lundin,S. and Käller,M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.
29. Banerjee,S., Raman,K. and Ravindran,B. (2021) Sequence neighborhoods enable reliable prediction of pathogenic mutations in cancer genomes. *Cancers*, **13**, 2366.
30. Sudhakar,M., Rengaswamy,R. and Raman,K. (2022) Novel ratio-metric features enable the identification of new driver genes across cancer types. *Sci. Rep.*, **12**, 5.
31. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
32. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.*, **29**, 15–21.
33. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.-C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
34. Anders,S., Pyl,P.T. and Huber,W. (2015) HTSeq–a python framework to work with high-throughput sequencing data. *Bioinforma. Oxf. Engl.*, **31**, 166–169.
35. Frazee,A.C., Pertea,G., Jaffe,A.E., Langmead,B., Salzberg,S.L. and Leek,J.T. (2015) Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat. Biotechnol.*, **33**, 243–246.
36. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
37. Köster,J. and Rahmann,S. (2012) Snakemake–a scalable bioinformatics workflow engine. *Bioinforma. Oxf. Engl.*, **28**, 2520–2522.
38. Desvillechabrol,D., Legendre,R., Rioualen,C., Bouchier,C., van Helden,J., Kennedy,S. and Cokelaer,T. (2018) Sequanix: a dynamic graphical interface for snakemake workflows. *Bioinforma. Oxf. Engl.*, **34**, 1934–1936.
39. Lawrence,M.S., Stojanov,P., Polak,P., Kryukov,G.V., Cibulskis,K., Sivachenko,A., Carter,S.L., Stewart,C., Mermel,C.H., Roberts,S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.
40. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
41. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
42. Gonzalez-Perez,A., Deu-Pons,J. and Lopez-Bigas,N. (2012) Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med.*, **4**, 89.
43. Schwarz,J.M., Rödelsperger,C., Schuelke,M. and Seelow,D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.
44. Zook,J.M., McDaniel,J., Olson,N.D., Wagner,J., Parikh,H., Heaton,H., Irvine,S.A., Trigg,L., Truty,R., McLean,C.Y. *et al.* (2019)

An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.*, **37**, 561–566.
45. Krusche,P., Trigg,L., Boutros,P.C., Mason,C.E., De La Vega,F.M., Moore,B.L., Gonzalez-Porta,M., Eberle,M.A., Tezak,Z., Lababidi,S. *et al.* (2019) Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.*, **37**, 555–560.
46. Trigg,L., Cleary,J., Braithwaite,R., Gaastra,K., Hilbush,B., Inglis,S., Irvine,S., Jackson,A., Littin,R., Mehul,R. *et al.* (2015) Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. bioRxiv doi: https://doi.org/10.1101/023754, 02 August 2015, preprint: not peer reviewed.
47. Everaert,C., Luypaert,M., Maag,J.L.V., Cheng,Q.X., Dinger,M.E., Hellemans,J. and Mestdagh,P. (2017) Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Sci. Rep.*, **7**, 1559.
48. Afgan,E., Baker,D., Batut,B., van den Beek,M., Bouvier,D., Čech,M., Chilton,J., Clements,D., Coraor,N., Grüning,B.A. *et al.* (2018) The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.
49. Bourgey,M., Dali,R., Eveleigh,R., Chen,K.C., Letourneau,L., Fillon,J., Michaud,M., Caron,M., Sandoval,J., Lefebvre,F. *et al.* (2019) GenPipes: an open-source framework for distributed and scalable genomic analyses. *GigaScience*, **8**, giz037.
50. Orjuela,S., Huang,R., Hembach,K.M., Robinson,M.D. and Soneson,C. (2019) ARMOR: an automated reproducible MOdular workflow for preprocessing and differential analysis of RNA-seq data. *G3: Genes Genomes Genetics*, **9**, 2089–2096.
51. Cornwell,M., Vangala,M., Taing,L., Herbert,Z., Köster,J., Li,B., Sun,H., Li,T., Zhang,J., Qiu,X. *et al.* (2018) VIPER: visualization pipeline for RNA-seq, a snakemake workflow for efficient and complete RNA-seq analysis. *BMC Bioinformatics*, **19**, 135.
52. Backman,T.W.H. and Girke,T. (2016) systemPipeR: NGS workflow and report generation environment. *BMC Bioinformatics*, **17**, 388.
53. Ewels,P.A., Peltzer,A., Fillinger,S., Patel,H., Alneberg,J., Wilm,A., Garcia,M.U., Di Tommaso,P. and Nahnsen,S. (2020) The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.*, **38**, 276–278.