UMass Chan Medical School
eScholarship@UMassChan

Open Access Publications by UMMS Authors

2020-04-19

# DolphinNext: a distributed data processing platform for high throughput genomics

Onur Yukselen
*University of Massachusetts Medical School*

*Et al.*

Let us know how access to this document benefits you.
Follow this and additional works at: https://escholarship.umassmed.edu/oapubs

 Part of the Bioinformatics Commons, Computational Biology Commons, Computer Sciences Commons, Genomics Commons, Integrative Biology Commons, and the Systems Biology Commons

## Repository Citation

Yukselen O, Turkyilmaz O, Ozturk AR, Garber M, Kucukural A. (2020). DolphinNext: a distributed data processing platform for high throughput genomics. Open Access Publications by UMMS Authors. https://doi.org/10.1186/s12864-020-6714-x. Retrieved from https://escholarship.umassmed.edu/oapubs/4205

Creative Commons License

This work is licensed under a Creative Commons Attribution 4.0 License.
This material is brought to you by eScholarship@UMassChan. It has been accepted for inclusion in Open Access Publications by UMMS Authors by an authorized administrator of eScholarship@UMassChan. For more information, please contact Lisa.Palmer@umassmed.edu.

Yukselen et al. BMC Genomics (2020) 21:310
https://doi.org/10.1186/s12864-020-6714-x

BMC Genomics

SOFTWARE Open Access

# DolphinNext: a distributed data processing platform for high throughput genomics

Onur Yukselen[1], Osman Turkyilmaz[2], Ahmet Rasit Ozturk[2], Manuel Garber[1,3,4*] and Alper Kucukural[1,3,4*]

## Abstract

**Background:** The emergence of high throughput technologies that produce vast amounts of genomic data, such as next-generation sequencing (NGS) is transforming biological research. The dramatic increase in the volume of data, the variety and continuous change of data processing tools, algorithms and databases make analysis the main bottleneck for scientific discovery. The processing of high throughput datasets typically involves many different computational programs, each of which performs a specific step in a pipeline. Given the wide range of applications and organizational infrastructures, there is a great need for highly parallel, flexible, portable, and reproducible data processing frameworks.
Several platforms currently exist for the design and execution of complex pipelines. Unfortunately, current platforms lack the necessary combination of parallelism, portability, flexibility and/or reproducibility that are required by the current research environment. To address these shortcomings, workflow frameworks that provide a platform to develop and share portable pipelines have recently arisen. We complement these new platforms by providing a graphical user interface to create, maintain, and execute complex pipelines. Such a platform will simplify robust and reproducible workflow creation for non-technical users as well as provide a robust platform to maintain pipelines for large organizations.

**Results:** To simplify development, maintenance, and execution of complex pipelines we created DolphinNext. DolphinNext facilitates building and deployment of complex pipelines using a modular approach implemented in a graphical interface that relies on the powerful Nextflow workflow framework by providing 1. A drag and drop user interface that visualizes pipelines and allows users to create pipelines without familiarity in underlying programming languages. 2. Modules to execute and monitor pipelines in distributed computing environments such as high-performance clusters and/or cloud 3. Reproducible pipelines with version tracking and stand-alone versions that can be run independently. 4. Modular process design with process revisioning support to increase reusability and pipeline development efficiency. 5. Pipeline sharing with GitHub and automated testing 6. Extensive reports with R-markdown and shiny support for interactive data visualization and analysis.

**Conclusion:** DolphinNext is a flexible, intuitive, web-based data processing and analysis platform that enables creating, deploying, sharing, and executing complex Nextflow pipelines with extensive revisioning and interactive reporting to enhance reproducible results.

**Keywords:** Pipeline, Workflow, Genome analysis, Big data processing, Sequencing

* Correspondence: manuel.garber@umassmed.edu; alper.kucukural@umassmed.edu
[1]Bioinformatics Core, University of Massachusetts Medical School, Worcester, MA 01605, USA
Full list of author information is available at the end of the article

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Analysis of high-throughput data is now widely regarded as the major bottleneck in modern biology [1]. In response, resource allocation has dramatically skewed towards computational power, with significant impacts on budgetary decisions [2]. One of the complexities of high-throughput sequencing data analysis is that a large number of different steps are often implemented with a heterogeneous set of programs with vastly different user interfaces. As a result, even the simplest sequencing analysis requires the integration of different programs and familiarity with scripting languages. Programming was identified early on as a critical impediment to genomics workflows. Indeed, microarray analysis became widely accessible only with the availability of several public and commercial platforms, such as GenePattern [3] and Affymetrix [4], that provided a user interface to simplify the application of a diverse set of methods to process and analyze raw microarray data.

A similar approach to sequencing analysis was later implemented by Galaxy [5], GenomicScape [6], Terra (https://terra.bio) and other platforms [3, 7–14]. Each of these platforms has a similar paradigm: Users upload data to a central server and apply a diverse, heterogeneous set of programs through a standardized user interface. As with microarray data, these platforms allow users without any programming experience to perform sophisticated analyses on sequencing data obtained from different protocols such as RNA Sequencing (RNA-Seq) and Chromatin Immunoprecipitation followed by Sequencing (ChIP-Seq) and carry out sophisticated analysis. Users are able to align sequencing reads to the genome, assess differential expression, and perform gene ontology analysis through a unified point and click user interface.

While current platforms are a powerful way to integrate existing programs into pipelines that carry end-to-end data processing, they are limited in their flexibility. Installing new programs is usually only done by administrators or advanced users. This limits the ability of less skilled users to test new programs or simply add additional steps into existing pipelines. This development flexibility is becoming ever more necessary as genome-wide assays are becoming more prevalent and data analysis pipelines becoming increasingly creative [15].

Similarly, computing environments have also grown increasingly complex. Institutions rely on a diverse set of computing options ranging from large servers, higher performance computing clusters, to cloud computing. Data processing platforms need to be easily portable to be used in different environments that best suit the computational needs and budgetary constraints of the project. Further, with the increased complexity of analyses, it is important to ensure reproducible analyses by making analysis pipelines easily portable and less dependent on the computing environment where they were developed [16]. Lastly, it is necessary to have a flexible and scalable pipeline platform that can be used both by individuals with smaller sample sizes as well as by medium and large laboratories that need to analyze hundreds of samples a month, or centralized informatics cores that analyze data produced by multiple laboratories.

Nextflow is a recently developed workflow engine built to address many of these needs [17]. The Nextflow engine can be configured to use a variety of executors (e.g. SGE, SLURM, LSF, Ignite) in a variety of computing environments. A pipeline that leverages the specific multi-core architecture of a server can be written on a workstation and easily re-used on a high-performance cluster environment (e.g. Amazon and Google cloud) whenever the need for higher parallelization arises. Further, Nextflow allows in-line process definition that simplifies the incorporation of small processes that implement new functionality. Not surprisingly, Nextflow has quickly gained popularity, as reflected by several efforts to provide curated and revisioned Nextflow-based pipelines such as nf-core [18], Pipeliner [19] and CHIPER [20], which are available from a public repository.

In spite of its simplicity, Nextflow can get unwieldy when pipelines become complex, and maintaining them becomes taxing. Here we present DolphinNext, a user-friendly, highly scalable, and portable platform that is specifically designed to address current challenges in large data processing and analysis. DolphinNext builds on Nextflow as shown in Fig. 1. To simplify pipeline design and maintenance, DolphinNext provides a graphical user interface to create pipelines. The graphical design of workflows is critical when dealing with large and complex workflows. Both advanced Nextflow users as well as users with no prior experience benefit from the ability to visualize dependencies, branch points, and parallel processing opportunities. DolphinNext goes beyond providing a Nextflow graphical design environment and addresses many of the needs of high-throughput data processing: First, DolphinNext helps with reproducibility enabling the easy distribution and running of pipelines. In fact, reproducible data analysis requires making both the code and the parameters used in the analysis accessible to researchers [21–24]. DolphinNext allows users to package pipelines into portable containers [25, 26] that can be run as stand-alone applications because they include the exact versions of all software dependencies that were tested and used. The automatic inclusion of all software dependencies vastly simplifies the effort needed to share, run and reproduce the exact results obtained by a pipeline.

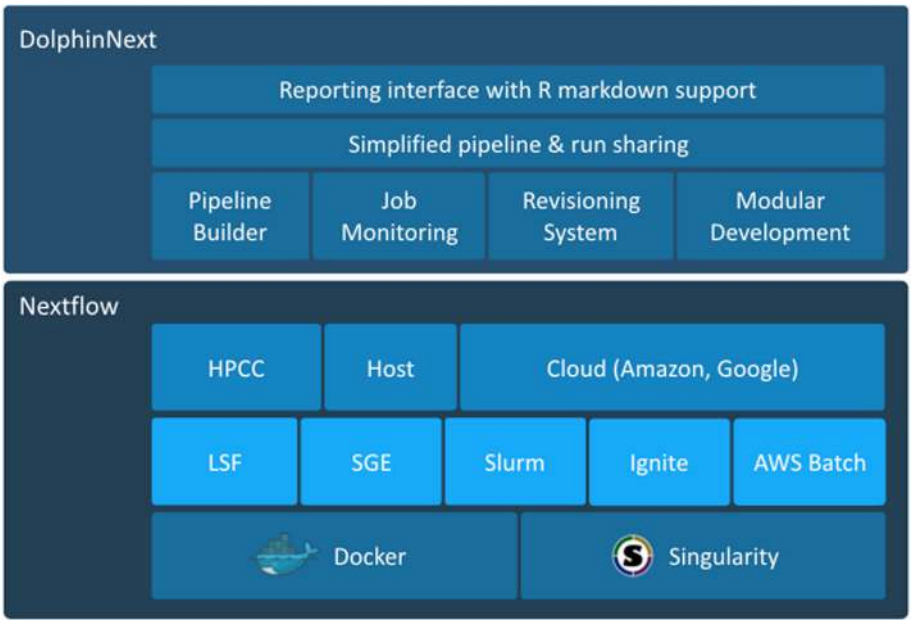


**Fig. 1** DolphinNext builds on Nextflow and simplifies creating complex workflows

Second, DolphinNext goes beyond existing data processing frameworks: Rather than requiring data to be uploaded to an external server for processing, DolphinNext is easily run across multiple architectures, either locally or in the cloud. As such it is designed to process data where the data resides rather than requiring users to upload data into the application. Further, DolphinNext is designed to work on large datasets, without needing customization. It can thus support the needs of large sequencing centers and projects that generate a vast amount of sequencing data such as ENCODE [27], GTex [28], and TCGA (The Cancer Genome Atlas) Research Network (https://www.cancer.gov/tcga) that have had the need to develop custom applications to support their needs. DolphinNext can also readily support smaller laboratories that generate large sequencing datasets.

Third, as with Nextflow, DolphinNext is implemented as a generic workflow design and execution environment. However, in this report, we showcase its power by implementing sequencing analysis pipelines that incorporate best practices derived from our experience in genomics research. This focus is driven by our current use of DolphinNext, but its architecture is designed to support any workflow that can be supported by Nextflow.

In conclusion, DolphinNext provides an intuitive interface for weaving together processes each of which have dependent inputs and outputs into complex workflows. DolphinNext also allows users to easily reuse existing components or even full workflows as components in new workflows; in this way, it enhances portability and helps to create more reproducible and easily customizable workflows. Users can monitor job status and, upon identifying errors, correct parameters or data files and restart pipelines at the point of failure. These features save time and decrease costs, especially when processing large data sets that require the use of cloud-based services.

The key features of DolphinNext include:

- Simple pipeline design interface
- Powerful job monitoring interface
- User-specific queueing by job submissions tied to user accounts
- Easy re-execution of pipelines for new sets of samples by copying previous runs
- Simplified sharing of pipelines using the GitHub repository hosting system (github.com)
- Portability across computational environments such as workstations, computing clusters, or cloud-based servers
- Built-in pipeline and process revision control
- Full access to application run logs
- Parallel execution of non-dependent processes
- Integrated data analysis and reporting interface with R markdown support
- Launching cloud clusters on Amazon (AWS) and Google (GCP) with backup options to S3 and google buckets

## Implementation

The DolphinNext workflow system, with its intuitive web interface, was designed for a wide variety of users, from bench biologists to expert bioinformaticians. DolphinNext is meant to aid in the analysis and management of large datasets on High Performance Computing (HPC) environments (e.g LSF, SGE, Slurm Apache Ignite), cloud services, or personal workstations.

DolphinNext is implemented with PHP, MySQL and Javascript technologies. At its core, it provides a drag-and-drop user interface for creating and modifying Nextflow pipelines. Nextflow [17] is a language to create scalable and reproducible scientific workflows. In creating DolphinNext, we aim to simplify Nextflow pipeline building by shifting the focus from software engineering to bioinformatics processes using a graphical interface that requires no programming experience. DolphinNext supports a wide variety of scripting languages (e.g. Bash, Perl, Python, Groovy) to create processes. Processes can be used in multiple pipelines, which increases the reusability of the process and simplifies code sharing. To that end, DolphinNext supports user and group level permissions so that processes can be shared among a small set of users or all users in the system. Users can repurpose existing processes used in any other pipelines, which eliminates the need to create the same process multiple times. These design features allow users to focus on only their unique needs rather than be concerned with implementation details.

To facilitate the reproducibility of data processing and the execution of pipelines in any computing environment, DolphinNext leverages Nextflow's support for Singularity and Docker container technologies [25, 26]. This allows the execution of a pipeline created by DolphinNext to require only Nextflow and a container software (Singularity or Docker) to be installed in the host machine. Containerization simplifies complex library, software and module installation, packaging, distribution and execution of the pipelines by including all dependencies. When distributed with a container, DolphinNext pipelines can be readily executed in remote machines or clusters without the need to manually install third-party software programs. Alternatively, DolphinNext pipelines can be exported as Nextflow code and distributed in publications. Exported pipelines can be executed from the command line upon ensuring that all dependencies are available in the executing host. Moreover, multiple executors, clusters, or remote machines can easily be defined in DolphinNext in order to perform computations in any available Linux-based cluster or workstation.

User errors can cause premature failure of pipelines, while also consuming large amounts of resources. Additionally, users may want to explore the impact of different parameters on the resulting data. To facilitate rerunning of a pipeline, DolphinNext builds on Nextflow's ability to record a pipeline execution state, enabling the ability to re-execute or resume a pipeline from any of its steps, even after correcting parameters or correcting a process. Pipelines can also be used as templates to process new datasets by modifying only the dataset-specific parameters.

In general, pipelines often require many different parameters, including the parameters for each individual program in the pipeline, system parameters (e.g. paths, commands), memory requirements, and the number of processors to run each step. To reduce the tedious set-up of complex pipelines, DolphinNext makes use of extensive pre-filling options to provide sensible defaults. For example, physical paths of genomes, their index files, or any third-party software programs can be defined for each environment by the administrator. When a pipeline uses these paths, the form loads pre-filled with these variables, making it unnecessary to fill them manually. The users still can change selected parameters as needed, but the pre-filling of default parameters speeds up the initialization of a new pipeline. For example, in an RNA-Seq pipeline, if RefSeq annotations [29] are defined as a default option, the user can change it to Ensembl annotations [30] both of which may be located at predefined locations. Alternatively, the user may specify a custom annotation by supplying a path to the desired annotation file.

Finally, when local computing resources are not sufficient, DolphinNext can also be integrated into cloud-based environments. DolphinNext readily integrates with Amazon AWS and Google GCP where, a new, dedicated computer cluster can easily be set up within DolphinNext with Nextflow's Amazon and Google cloud support. On AWS, necessary input files can be entered from a shared file storage EFS, EBS, or s3, and output files can also be written on s3 or other mounted drives [31–33]. On GCP, the input files can be selected from a Google bucket and the output files are exported to another Google bucket.

## General implementation and structure

DolphinNext has four modules: The **profile module is** specifically designed to support a multi-user environment and allows an administrator to define the specifics of their institutional computing environment. A **pipeline builder** is to create reusable processes and pipelines. A **pipeline executor** is created to run pipelines, and lastly the reports section is to monitor the results.

### Profile module

Users may have access to a wide range of different computing environments: A workstation, Cloud Computing, or a high-performance computing cluster where jobs are

submitted through a job scheduler such as IBM's LSF, SLURM or Apache Ignite. DolphinNext relies on Nextflow [17] to encapsulate computing environment settings and allows administrators to rely on a single configuration file that enables users to run the pipelines on diverse environments with minimal impact on user experience. Further, cloud computing and higher performance computing systems keep track of individual user usage to allocate resources and determine job scheduling priorities. DolphinNext supports individual user profiles and can transparently handle user authentication. As a result, DolphinNext can rely on the underlying computing environment to enforce proper resource allocation and fair sharing policies. By encapsulating the underlying computing platform and user authentication, administrators can provide access to different computing architectures, and users with limited computing knowledge can transparently access a vast range of different computing environments through a single interface.

### Pipeline builder

While Nextflow provides a powerful platform to build pipelines, it requires advanced programming skills to define pipelines as it requires users to use a programming language to specify processes, dependencies, and the execution order. Even for advanced users, when pipelines are becoming complex, pipeline maintenance can be a daunting task.

DolphinNext facilitates pipeline building, maintenance, and execution by providing a graphical user interface to create and modify Nextflow pipelines. Users choose from a menu of available processes (Fig. 2a) and use drag and drop functionality to create new pipelines by connecting processes through their input and output parameters (Fig. 2b). Two processes can only be connected

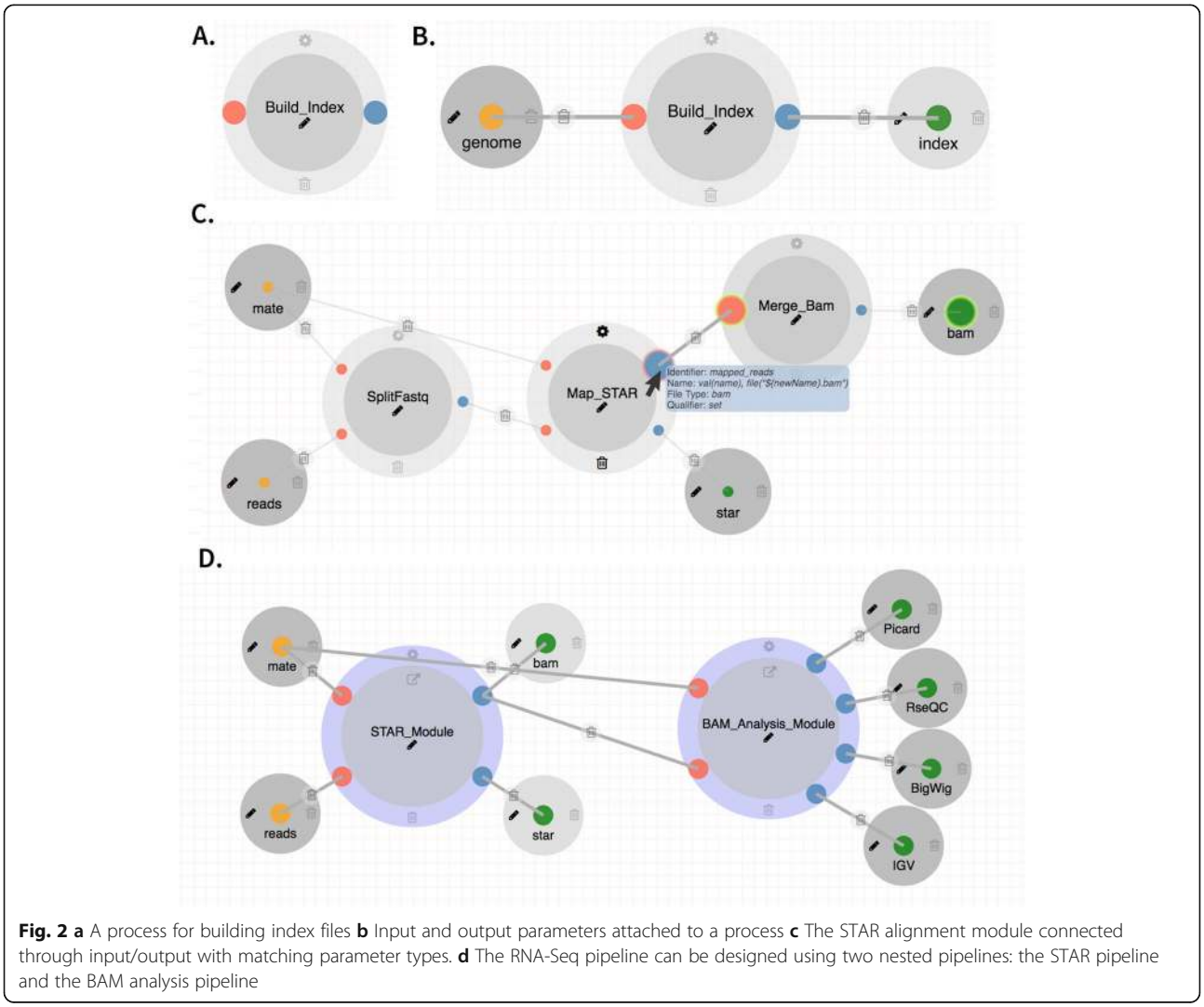


**Fig. 2** **a** A process for building index files **b** Input and output parameters attached to a process **c** The STAR alignment module connected through input/output with matching parameter types. **d** The RNA-Seq pipeline can be designed using two nested pipelines: the STAR pipeline and the BAM analysis pipeline

when the output data type of one is compatible with the input data type of the second (Fig. 2c). Upon connecting two compatible processes DolphinNext creates all necessary execution dependencies. Users can readily create new processes using the process design module (see below). Processes created in the design module are immediately available to the pipeline designer without any installation in DolphinNext.

The UI supports auto-saving to avoid loss of work if users forget to save their work. Once a pipeline is created, users can track revisions, edit, delete and share either as a stand-alone container Nextflow program, or in PDF format for documentation purposes.

The components of the pipeline builder are the process definition module, the pipeline designer user interface, and the revisioning system:

### Process design module

Processes are the core units in a pipeline, they perform self-contained and well-defined operations. DolphinNext users designing a pipeline can define processes using a wide variety of scripting languages (e.g. Shell scripting, Groovy, Perl, Python, Ruby, R). Once a process is defined, it is available to any pipeline designer. A pipeline is built from individual processes by connecting outputs with inputs. Whenever two processes are connected, a dependency is implicitly defined whereby a process that consumes the output of another only runs once this output is generated. Since each process may require specific parameters, DolphinNext provides several features to simplify the maintenance of processes and input forms that allow the user to select parameters to run them.

*Automated input form generation.* Running all processes within a pipeline requires users to specify many different parameters ranging from specifying the input (e.g. input reads in fastq format, path to reference files) to process specific parameters (e.g. alignment maximum mismatches, minimum base quality to keep). To gather this information, users fill out a form or set of forms to provide the pipeline with all the necessary information to run. A large number of parameters makes designing and maintaining the user interfaces that gather this information time consuming and error-prone. DolphinNext includes a meta-language that converts defined input parameters to web controls. These input parameters are declared in the header of a process script with the help of additional directives. *Form autofill support.* The vast majority of users work with default parameters and only need to specify a small fraction of all the parameters used by the pipeline. To simplify pipeline usage, we designed an autofill option to provide sensible process defaults and compute environment information.

Autofill is meant to provide sensible defaults; however, users can override them as needed. The descriptions of parameters and tooltips are also supported in these directives. Figure 3 shows the description of a defined parameter in RSEM settings.

### Revisioning, reusability and permissions system

DolphinNext implements a revisioning system to track all changes in a pipeline or a process. In this way, all versions of a process or pipeline are accessible and can be used to reproduce old results or to evaluate the impact of new changes. In addition, DolphinNext provides safeguards to prevent the loss of previous pipeline versions. If a pipeline is shared (publicly or within a group), it is not possible to make changes on its current revision. Instead, users must create a new version to make changes. Hence, we keep pipelines safe from modifications yet allowing for improvements to be available in new revisions. Unlike nf-core or other Nextflow based pipeline repositories [18–20], DolphinNext keeps track of revisions for each of the processes within a pipeline rather than keeping revisions for each pipeline. In this way, the right combination of process revisions in a pipeline can be used to reproduce previously generated results. DolphinNext uses a local database to assign and store a unique identifier (UID) to every process and pipeline created and every revision made. A central server may be configured to assign UIDs across different DolphinNext installations so that pipelines can be identified from the UID, regardless of where they were created. Pipeline designers and users can select any version of a pipeline for execution or editing. In addition to database support, DolphinNext integrates with a GitHub repository so that pipelines can be more broadly shared. DolphinNext can seamlessly push pipelines to a specified repository or branch. In addition to storing the pipeline code, DolphinNext updates its own pipeline or revision database record with the GitHub commit id to keep the revisions that have been synced with a GitHub repository. To support tests and continuous integration of pipelines, we have integrated Travis-ci (travis-ci.org), the standard for automated testing. Pipeline designers can define the Travis-ci test description document within the DolphinNext pipeline builder. When a pipeline is updated and pushed to GitHub, it automatically triggers the Travis-ci tests. To enable Travis-ci automation, pipeline designers specify a container [25, 26] within the pipeline builder.

*User permissions and pipeline reusability.* To increase reusability, DolphinNext supports pipeline sharing. DolphinNext relies on a permissions system similar to that used by the UNIX operating systems. There are three levels of permissions: user, group and world. By default, a new pipeline is owned and only visible to the user who created it. The user can change this default by creating a group of users and designating pipelines as visible to

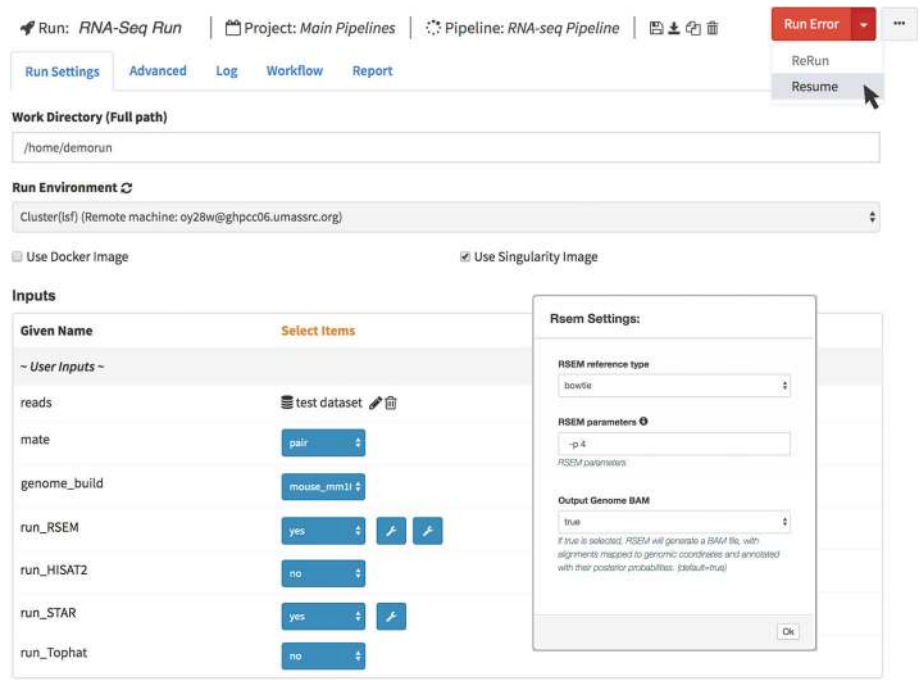


**Fig. 3** Resuming RNA-Seq pipeline after changing RSEM parameters

users within that group. Alternatively, the user can make a pipeline available to all users. DolphinNext further supports a refereed workflow by which pipelines can only be made public after authorization by an administrator, this is useful for organizations that desire to maintain strict control of broadly available pipelines.

Although integration with GitHub makes sharing and executing possible, pipelines can also be downloaded in Nextflow format for documentation, distribution and execution outside of DolphinNext. To allow users and administrators to make pipelines available across installations, DolphinNext supports pipeline import and export.

### Nested pipelines

Many pipelines share not just processes, but subcomponents involving several processes. For instance, BAM quality assurance is common to most sequence processing pipelines (Fig. 2d). It relies on RSeQC [34] and Picard (http://broadinstitute.github.io/picard) to create read quality control reports. To minimize redundancy, these modules can be encapsulated as pipelines and re-used as if they were processes. The pipeline designer module supports drag and drop of whole pipelines and in a similar way as it supports individual processes. Multiple pipelines such as RNA-Seq, ATAC-Seq, and ChIP-Seq can, therefore, have the same read quality assurance logic (Figure S4). Reusing complex logic by encapsulating it in a pipeline greatly simplifies and standardizes the maintenance of data processing pipelines.

## Pipeline executor

One of the most frustrating events in data processing is an unexpected pipeline failure. Failures can be the result of an error in the parameters supplied to the pipeline (e.g. an output directory with no write permissions, or incompatible alignment parameters) or because of computer system malfunctions. Restarting a process from the beginning when an error occurred at the very end of the pipeline can result in days of lost computing time.

DolphinNext provides a user interface to monitor pipeline execution in real-time. If an error occurs the pipeline is stopped; the user, however, can restart the pipeline from the place where it stopped after changing the parameters that caused the error (Fig. 3). Users can also assign pipeline runs to projects so that all pipelines associated with a project can be monitored together.

In addition to providing default values for options that are pipeline specific, administrators can provide default values for options common to all pipelines, such as resource allocation (e.g., memory, CPU, and time), and access level of pipeline results.

Specific features of pipeline running are:

1. ***Run status page*:** DolphinNext provides a "Run Status" page for monitoring the status of all running jobs that belong to the user or the groups to which

the user belongs. This searchable table includes the name of the run, working directory, description, and status. Users can also access the run page of a specific run to make a change, terminate, resume, copy, delete, or re-execute.

2. ***Accessibility to all run logs*:** To monitor run specific details and troubleshoot failed runs, DolphinNext provides access to all application logs. Further, it also gives access to Nextflow's execution reports, which include the execution time, resource utilization (e.g. CPU or memory for each process), process duration time (both clock and CPU time) and memory utilization. In this way, users can optimize their computational resources by reducing unnecessary overhead based on the data available from past executions.

## Reports

Most processes in a pipeline produce interim data that is used by downstream processes. For example, in an RNA-Seq analysis pipeline, subtracting ribosomal reads prior to genomic alignment reduces processing time. This is done by aligning directly to the ribosomal RNA genes, and keeping reads with no matches for alignment to the genome or transcriptome [35]. In addition to reducing compute time, the fraction of reads that align to ribosomal genes is an important metric to assess the technical efficiency of library preparation, specifically to measure how well ribosomal depletion worked. DolphinNext allows users to access and visualize interim pipeline results, such as intermediate alignments like alignment to the ribosomal RNAs, that are not part of the final result. Process designers can define any number of secondary outputs, which the DolphinNext report module can process and present in a user-friendly format. While there is an integrated table viewer for tabular data, this data can also be seamlessly loaded and analyzed using an embedded DEBrowser, an interactive visualization tool for count data [36]. PDF and HTML files can be visualized within DolphinNext or downloaded for offline analysis (Fig. 4).

For more flexible reports, DolphinNext supports an R-markdown defined output. Process designers can include a custom R markdown script specifically designed to handle and visualize the process output. Pipeline designers can make this report available to users for interactive analysis of the process results (Fig. 5).

## Results

Sequence analysis requires an iterative approach that consists of two main tasks: data processing and data analysis [38] (Fig. 6). Data processing involves taking raw sequencing data to produce a count matrix in which features (genes, peaks or any genomic annotation) are scored across samples. Data analysis uses this count matrix to perform comparisons across conditions (differential analysis), unbiased analysis (clustering, principal component analysis), or supervised analysis (data modeling, classifier building) [36]. An informative count matrix requires systematic data processing steps that are consistent across samples and even projects. As opposed to data analysis, which involves more ad-hoc, and data-driven decisions, data processing can be standardized into a workflow or pipeline that, once established, needs very few changes. In general, data processing requires great computing power and storage needs relative to data analysis.

A typical data processing pipeline can be broken down into three steps: pre-processing, processing, and post-processing. Pre-processing: includes eliminating low-quality reads or low-quality bases through filtering, read trimming and adapter removal. These steps are critical to improve alignment and assembly operations, which make up the central processing step. In general, processing of sequencing reads involves alignment to a genome or reference annotations (e.g. a transcriptome) or an assembly process. Post-processing involves evaluating the quality of mapping or assembly before creating a count table, quality checks of alignment and/or assembly steps, and outputs the quantification of genomic features as consolidated count tables. To enable post-processing and data quality assessment, DolphinNext automatically creates genome browser loadable files such as Tile Data Files (TDF) for the Integrative Genome Viewer [39] and Wiggle Track Format (WIG, or BigWIG) for the UCSC genome browser [40, 41]. In addition, DolphinNext produces alignment reports that summarize read coverage of genomic annotations (e.g. exonic, intergenic, UTR, intronic) in order to help assess whether the library captured expected genomic regions.

Here, we highlight the flexibility of DolphinNext's pipeline builder by providing a detailed description of end-to-end RNA-Seq and ChIP-Seq/ATAC-Seq processing pipelines. Each of these pipelines provides unique features not found in other publicly available workflows such as: 1. Extensive support for pre-processing (read trimming by quality or adapter sequence, iterative removal of specific RNA or DNA species, such as rRNA and repetitive sequences), 2. Support for different aligners and quantification methods in the processing steps, and 3. An extensive reports and quality control checks in the post-processing steps.

### RNA-Seq pipeline (Figure S1)

All sequence processing pipelines take one or several fastq input files. This pipeline, like all other high-throughput sequencing processing pipelines (see below for other examples), first performs data preparation, a

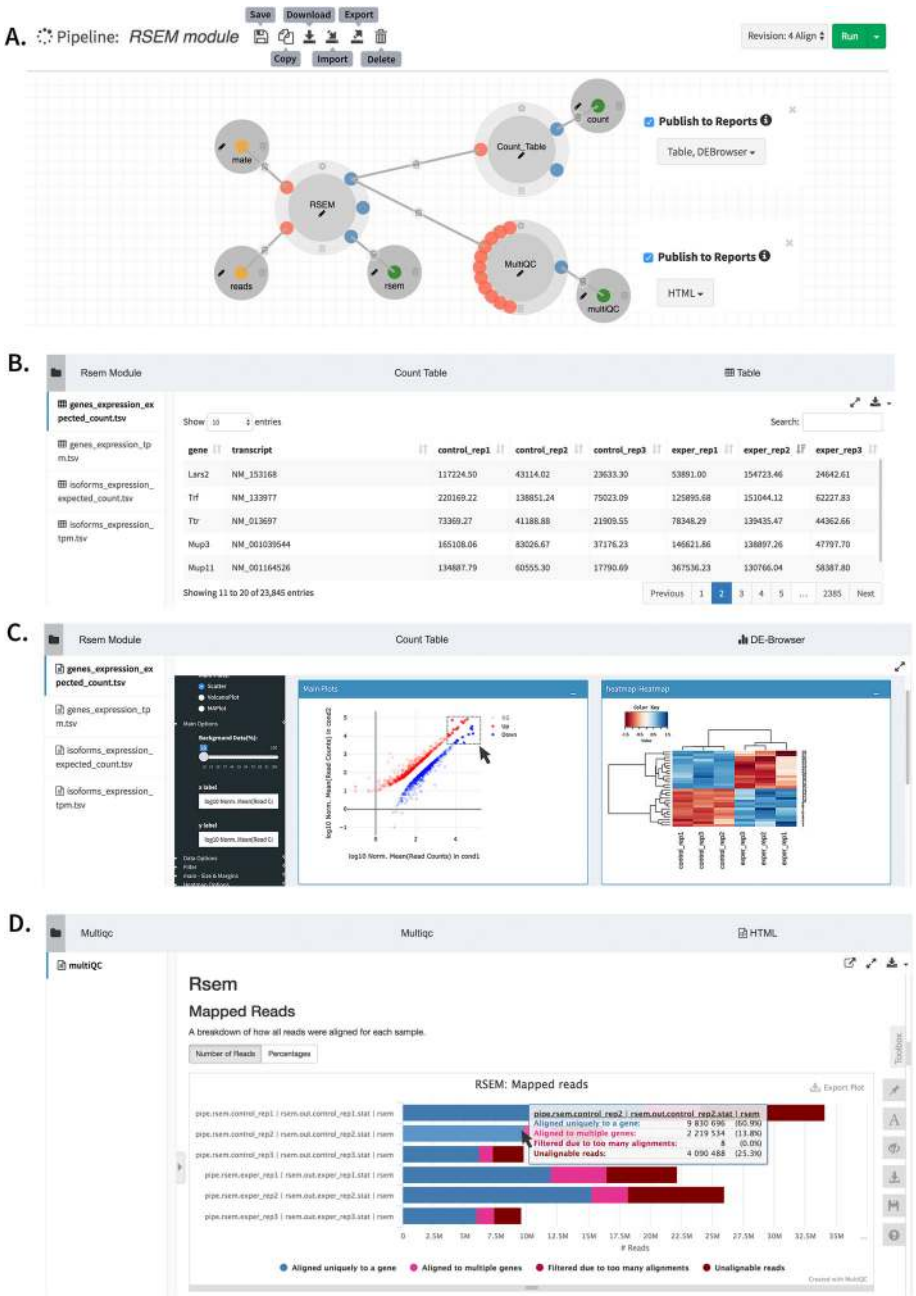


**Fig. 4** **a** RSEM module which involves Count_Table to summarize sample counts into a consolidated count table. This process reports the results with a table or upload the count table to embedded DEBrowser [36], **b** Count table report **c.** MultiQC [37] report to summarize numerous bioinformatics tool results, and **d** Embedded DEBrowser [36] module for interactive differential expression analysis

step that consists of basic read quality control and filtering actions (Figure S2): read quality reports, read quality filtering, read quality trimming, adapter removal. After these quality control steps, the RNA-Seq pipeline offers the user the option to align, filter out, and/or estimate the abundance of both standard and predefined sets of genomic loci (e.g. rRNAs, miRNAs, tRNAs, piRNAs, snoRNAs, ERCC [42], mobile elements). Before any such alignment, reads may be trimmed to the desired length for optimal alignment, especially if quality issues in the 3′ or 5′ ends of the reads are suspected (e.g. to align miRNA or tRNA sequences). Once the data preparation phase of the pipeline is complete, the pipeline produces quality reports including FastQC reports [43] and information about the fraction of reads aligned to each of the genomic loci selected by the user. In its current implementation, data preparation relies on FastQC [44], adapter removal using cutadapt [45], 5′ and 3′

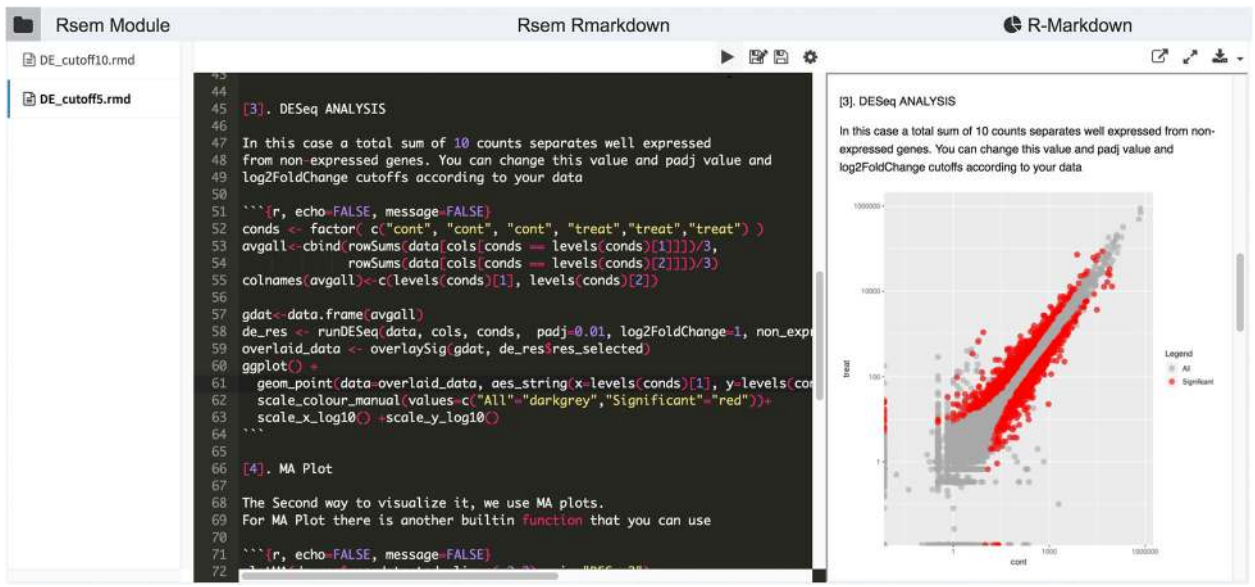


**Fig. 5** R markdown report for RNA-Seq analysis. Users can adapt this template according to their needs

trimming, quality removal using trimmomatic [46], and Bowtie2 [47] for alignments against selected regions or transcripts (Figure S3).

Only reads that pass all filters in the data preparation stage are kept for later steps. To estimate expression levels, the RNA-Seq pipeline uses RSEM [48] which aligns reads to a predefined set of transcripts. The user can use available transcript sets (i.e Ensemble [49], GENCODE [50, 51], RefSeq [52]) or upload their own. The RNA-Seq pipeline also allows the user to align reads against the genome using splicing aware alignment algorithms and generate a genome browser viewable file to visually assess genomic contamination and library quality. To do so, the user may choose between any or all of the most commonly used aligners: STAR [53], Hisat2 [54] and Tophat2 [55]. Resulting alignments are then processed to generate genome browser-friendly formats: bigwig (for UCSC genome browser [40] or TDF (for the integrative genome viewer (IGV, [39]).

If the user opted to perform genomic alignments, the pipeline reports overall quality metrics such as coverage and the number of mapped reads to different genomic and transcriptomic regions (Figure S4). These reports rely on Picard's CollectRNASeqMetrics program (http://broadinstitute.github.io/picard) and the RSeQC [34] program.

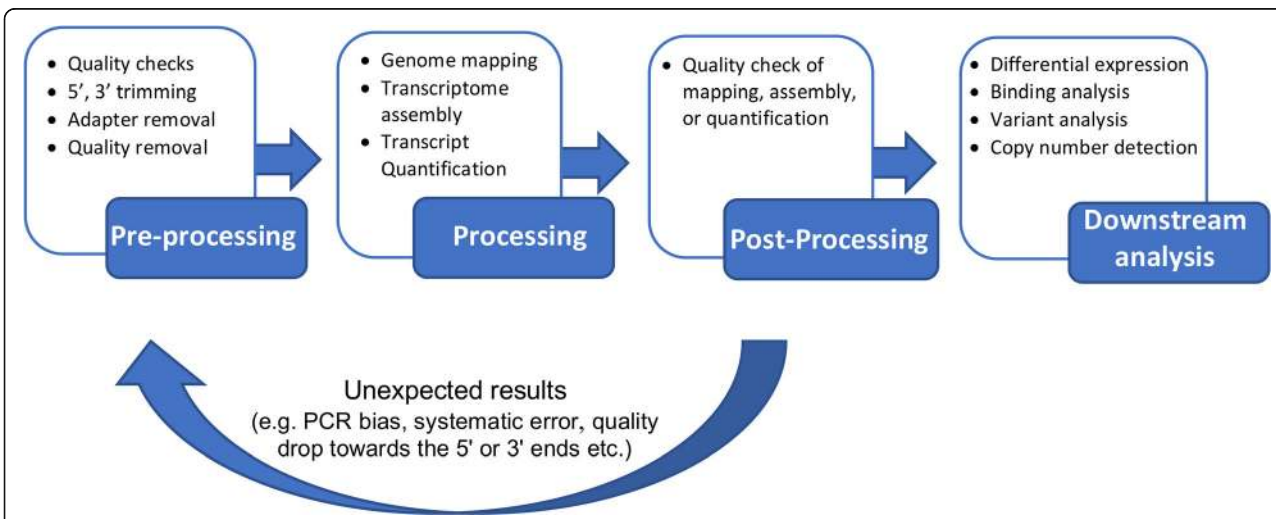


**Fig. 6** DolphinNext allows for implementation of a complete sequencing analysis cycle

Finally, the RNA-Seq pipeline returns a quantification matrix that includes the estimated counts and transcripts per million (TPM) based on RSEM [48] or by simply counting reads using featureCounts [56] for each gene and/or for each annotated isoform. These matrices are used as inputs for differential gene expression analysis and can be uploaded directly to an embedded instance of our DEBrowser [36] software, which allows interactive exploration of the resulting data.

#### ATAC-Seq and ChIP-Seq pipeline (Figure S4-S5)

DolphinNext offers pipelines to process libraries for the analysis of ChIP-Seq and ATAC-Seq data. These two pipelines share most processes and only differ at a few very specific points. They also share all initial data preparation steps with the RNA-Seq pipeline, such that both rely on the very same processes for read filtering, read quality reporting and alignment to desired genomic locations to quantify and filter out the reads mapped to repeat regions or any other loci of interest. Filtering out reads mapping to a predefined set of loci can dramatically speed up the genome-wide alignment that follows.

After data processing, reads are aligned to the genome using a short read aligner such as Bowtie2 [47]. For large data files, such as those typically obtained from ATAC-Seq, alignments can be sped up by splitting the files into smaller chunks that are aligned in parallel. The choice to split and the chunk sizes should be determined by the user based on the specific computing environment. By default, the pipeline creates smaller files that each have 5 million reads. After alignments of each read chunk, the results are automatically merged with the samtools [57] merge function. The pipeline then allows users to estimate and remove PCR duplicates using the Picard's mark duplicates (http://broadinstitute.github.io/picard) function.

For ATAC-Seq, the pipeline calls accessible chromatin regions by estimating the Tn5 transposase cut sites through a read extension process that has been shown to more accurately reflect the exact position that was accessible to transposase [35, 58]. Once each read has been shortened, peaks are called identically in both the ChIP-Seq and ATAC-Seq using MACS 2 [59].

When processing several samples together, the ATAC-Seq and ChIP-Seq pipelines provide consensus peak calls by merging all peaks individually called in each sample using Bedtools [60]. The number of reads in each peak location are then quantified using Bedtools [60] coverage function. As a result, ATAC-Seq and ChIP-Seq pipelines also generate a matrix that has the count values for each peak region and samples. This matrix can be uploaded directly to the embedded version of DEBrowser [36] to perform differential analysis or downloaded to perform other analysis. Finally, to determine motifs under the peaks in ChIP-Seq pipeline, there is also a motif discovery support using HOMER (http://homer.salk.edu/homer/index.html).

## Example use cases and comparisons

There are several features of DolphinNext that make pipeline creation, maintenance and portability much simpler than in other platforms. Here, we compared DolphinNext with one of the most popular processing platforms, Galaxy, to point out the main differences of DolphinNext. Specifically:

#### Distributed pipeline execution with containerization

Large enterprises or universities support many different execution environments: from High Performance Computing Cluster (HPCC) to large workstations and cloud platforms (AWS, GCP, Azure). Galaxy requires an instance to be deployed and launched for each cloud cluster. To make cloud computing available to users, a Galaxy server needs to be continuously running on the cloud or independent instances would need to be spun for each cloud user. While the first solution can significantly increase the cost of cloud computing, the latter requires an administrator to continuously assist users.

On the other hand, using Nextflow's cloud support and containerization, DolphinNext supports all these environments through configurations stored within DolphinNext. Each environment is configured once and made available to all users or desired user groups. Further, users can add other environments to which they have exclusive access. For example, an investigator with dual affiliation who has access to two HPCCs can have both environments configured within DolphinNext. At runtime, the investigator, or any members in his or her group can decide which system to use depending on the specific needs of their processes or the availability of the systems. Similarly, investigators can decide on whether to use their cost-free but busy local HPCC for routine work without tight deadlines or instead use cloud-based computing which is more expensive but ensures timely processing. In fact, once a Google or Amazon cloud authentication is defined in DolphinNext, users can launch a temporary cluster that can dynamically scale depending on load with a couple of clicks.

#### Agile pipeline design

The evolving nature of research constantly requires new ways in which to process data. As such pipelines are constantly evolving and constant changes require the ability to test different parameters and combinations of existing programs and custom scripts. To adapt a pipeline, new parameters may be required or custom scripts that pre-process data may be included. In Galaxy, the forms to execute a tool are defined in static XML files in

a tool wrapper. Any change in these files requires an update by the tool developer and reinstallation of the tool by the admin and a full system restart. Only a user with administrator rights can install a new tool so that it is available to all users. As a result, users cannot rapidly test pipelines without administrator intervention.

DolphinNext simplifies pipeline/process creation for regular users as well as administrators. For example, a user needs to update an RNA-Seq pipeline with additional programs and parameters to support newly available algorithms to add splicing variant discovery (e.g. Using StringTie [61]). To do this, the user would navigate to the pipeline builder module where all processes in the pipeline can be updated or new ones can be added. In Galaxy, the user depends on the tool developer to improve the wrapper so that new parameters are exposed, or the new program becomes ready for Galaxy use. Instead, in DolphinNext, StringTie can be added directly by the user to a private pipeline version and used right away. The user can define the new input parameters to run StringTie as the form elements in the process design module (see above). The defined input elements (drop box, check box, input box etc.) are then added to the run page automatically even while the pipeline is running. The only thing required is for the software (e.g. StringTie) to be installed on a server or container accessible to DolphinNext. An administrator may be involved later if the user wants to make their custom pipeline available to all DolphinNext users with these changes.

### Pipeline sharing

In Galaxy, the investigator can export the pipeline. However, in order for another investigator to run the pipeline, they would need a Galaxy instance installed and configured with all the pipeline dependencies most likely requiring administrative assistance.

When investigators use DolphinNext to process their data, the platform's pipeline sharing functionality is useful for the reproducibility of the results and valuable for writing the methods section of the publication. Once pipelines are finalized, the user can share a link for the pipeline or utilize GitHub integration. When a link is shared, other investigators can easily execute the pipeline in their instance by just defining their profile. If a pipeline is shared through GitHub, investigators can also directly execute the pipeline from the command line using Nextflow without a DolphinNext instance installed. *Automatic report generator:* Galaxy is designed to chain the executed tools to create pipelines, and this platform keeps all the uploaded files, defined collections, and executed tools in its history. This type of history structure that includes all executed pieces of the pipeline makes it harder to create a built-in report. Even though a history file in Galaxy can be shared with collaborators, it requires extra effort to consolidate the results and convert them to a report.

In DolphinNext, when users need to add a new report or customize an existing report section, they can use the pipeline builder module to simply select an existing output or drag and drop a new output parameter to the builder and attach it to the process (Fig. 2a) and select the report type (Fig. 4a) (e.g. Html, table, pdf, R markdown, etc.). When executing the pipeline, only outputs that are set to 'publish' are added as a section into the report page. In this way, users can organize the run results of a large sequencing project involving hundreds of samples and tools (see RNA-Seq example above) in a single report page and easily compare the results of hundred samples using the consolidated tables and plots. Another advantage of customizing the reports page is that the same pipeline can be used to assess a different set of input parameters (e. g. using a new genome assembly or a new transcript annotation set that has improved 3′ UTRs or additional noncoding RNAs). Rerunning the same pipeline with the new set of parameters and comparing the new and previous results are usually not a trivial task without the automatic report generator designed for the pipeline. However, it is an easy task in DolphinNext, once input parameters are changed and the run resumed, Nextflow only executes the affected processes with this change to produce the new reports. The previous reports can still be accessible with a click on the same page that allows for easy comparison. Lastly, the report section in DolphinNext makes sharing all the published results with collaborators in a single page easier. In addition to that, embedded shiny and R-markdown applications in these reports make effortless interactive data visualization and exploration possible.

Overall, DolphinNext's main advantage is to simplify creating, deploying, sharing and executing pipelines without administrator rights to democratize the data processing and analysis for all users.

DolphinNext has been deployed in several institutions and companies. A typical application of DolphinNext was showcased in a recent report that identified and characterized artery- and vein-specific endothelial enhancers in Zebrafish [62]. DolphinNext was used to process single cell RNA-Seq and bulk RNA-Seq data in this study. Bioinformatics core scientists from these institutions have particularly benefited from DolphinNext as it allows them to rapidly customize and test pipelines to support ever changing needs. DolphinNext also reduces the burden of supporting concurrent data processing jobs as core scientists can allow users to monitor pipeline execution and re-running of similar pipelines while they focus on designing and testing new solutions.

DolphinNext helps with pipeline development and maintenance by enabling rapid troubleshooting and simpler workflow definitions.

A comparison between widely used platforms is shown in Table 1.

## Current limitations and future plans

DolphinNext currently relies on Nextflow. Although Nextflow is a very successful and widely used pipeline engine, supporting other platforms like SnakeMake [64] and WDL based execution engines [65] would increase the usage of DolphinNext.

Currently, all of our processes and modules are interchangeable which allows reusing existing components. This approach required more structured syntax, which limits the ability to import publicly available Nextflow pipelines such as those available from nf-core [18], Pipeliner [19] or CHIPER [20]. However, in the near future, we plan to support pipeline executions created in other standards.

Managing large datasets that have hundreds of libraries and samples is challenging. Hence a system that stores sample data (metadata) and then relies on metadata to help navigate this complexity and automate data

**Table 1** Comparison of related applications

| | DolphinNext | Galaxy [5] | Sequanix [10] | Taverna [9] | Arvados [63] |
|---|---|---|---|---|---|
| Platform[a] | JS/PHP | Python | Python | Java | Go |
| Workflow management system | Nextflow | Galaxy | Snakemake | Taverna | Arvados |
| Native task support[b] | Yes (any) | No | Yes (bash only) | Yes (bash only) | Yes |
| Common workflow language[c] | No | Yes | No | No | Yes |
| Streaming processing[d] | Yes | No | No | No | Yes |
| Code sharing integration[e] | Yes | No | No | No | Yes (GitHub) |
| Workflow modules[f] | Yes | Yes | Yes | Yes | Yes |
| Workflow versioning[g] | Yes | Yes | No | No | No |
| Automatic error failover[h] | Yes | Yes | No | No | Yes |
| Nested workflows | Yes | Yes | No | Yes | No |
| Used syntax/ semantics | own/own | XML/own | Python/own | own/own | Python/own |
| Web-based | Yes | Yes | No | No | No |
| Web-based process development[i] | Yes | No | No | No | No |
| Distributed pipeline execution[j] | Yes | No | No | No | No |
| Container Support | | | | | |
| Docker support | Yes | Yes | Yes | No | Yes |
| Singularity support | Yes | Yes | Yes | No | No |
| Built-in batch schedulers | | | | | |
| LSF | Yes (Native) | Yes (DRMAA) | Yes (Native) | No | No |
| SGE | Yes (Native) | Yes (DRMAA) | Yes (Native) | Yes (Native) | No |
| SLURM | Yes (Native) | Yes (DRMAA) | Yes (Native) | No | Yes (Native) |
| IGNITE | Yes (Native) | No | No | No | No |
| Built-in cloud | | | | | |
| AWS (Amazon Web Services) | Yes | Yes | No | Yes | Yes |
| GCP (Google Cloud Platform) | Yes | Yes (Partial)[k] | No | No | Yes |
| Autoscaling | Yes | Yes | No | No | Yes |

[a]The technology and the programming language in which each framework is implemented
[b]The ability of the framework to support the execution of native commands and scripts without re-implementation of the original processes
[c]Support for the CWL specification
[d]Ability to process tasks inputs/outputs as a stream of data
[e]Support for code management and sharing platforms, such as GitHub
[f]Support for modules, sub-workflows or workflow compositions
[g]Ability to track pipeline changes and to execute different versions at any point in time
[h]Support for automatic error handling and resume execution mechanism
[i]Ability to add new processes in an embedded web editor without a wrapper or any installation of the wrapper
[j]Support for executing the same pipeline without any change in multiple computing environments to process the data within a single interface (e.g. hpc clusters, a workstation and cloud)
[k]A Galaxy instance can be launched in Google cloud but for one-time use. When it is shut down, they are permanently deleted

processing is critical. There is currently growing support for metadata definitions in FAIR standards [66]. Integration of metadata tracking system into DolphinNext in FAIR standards will increase data portability and enable a simpler interface for users to integrate and compare across diverse datasets.

Lastly, another technical limitation of DolphinNext is related to containerization for advanced users. Even DolphinNext has a GitHub integration that users can push their pipelines to GitHub and test them using Travis-ci (travis-ci.org) for any change in the pipeline with defined Docker containers. Docker Hub is a repository specialized in hosting ready to execute images. Integration with Docker Hub will allow users to deposit, build and test Docker containers directly from Docker Hub. Currently, in order to enable automated builds, users need to use Docker Hub website to create a new repository in Docker Hub from their GitHub repository of a pipeline. With Docker Hub integration, we will reduce the complexity of pipeline containerization one more level in the future.

## Conclusion

DolphinNext is a powerful workflow design, execution, and monitoring platform. It builds on top of recent technological advances such as software containerization and the Nextflow engine to address current data processing needs in high-throughput biological data analysis. Its ability to run anywhere, leverage the computing infrastructure of the institution, and provide an intuitive user interface makes it suitable for both small, large, and complex projects.

Reproducing third party analyses that involve many different programs, each with custom parameters, is a tremendous challenge. We have put special emphasis on the reproducibility of pipelines. To enable this, DolphinNext investigators can distribute their processing pipelines as containers that can run as stand-alone applications including the proper versions of all software dependencies. Further, DolphinNext supports pipeline versioning where users can create and tag a pipeline version with a unique identifier (UID) that includes the general pipeline "graphical representation", the executable Nextflow script, as well as the exact software versions and parameters used when run on a specific dataset.

Users with different computational skills can use DolphinNext to increase productivity. Casual users can rely on previously defined and tested pipelines to process their datasets. Investigators can easily distribute their processing pipeline along with the data for others to reproduce their analyses. Investigators with access to parallel computing systems but without a strong computational background can use DolphinNext to optimally utilize their computing environment. For example, DolphinNext pipelines can automatically split large sequence data files into smaller "chunks" that can be aligned in parallel across hundreds of cluster nodes if such infrastructure is available. Finally, bioinformaticians can easily create new processes and integrate them into custom robust and reproducible pipelines.

DolphinNext offers a highly modular platform. Though this manuscript only focuses on the power of DolphinNext for data processing, we have also integrated two different downstream analysis tools (DEBrowser [36] and Rmarkdown) that take the count matrices output from the data processing steps directly into analysis and data exploration steps. Furthermore, DolphinNext's modular architecture allows for easy integration of any custom data analysis and visualization tool. As such DolphinNext is meant to provide a basic platform to standardize processing and analysis across institutions.

## Availability and requirements

**Project name**: DolphinNext.

**Project home page**: https://github.com/UMMS-Biocore/dolphinnext

**Documentation**: https://dolphinnext.readthedocs.org

**Web server**: https://dolphinnext.umassmed.edu

**Example pipelines:** https://github.com/dolphinnext

**Operation system**s: Platform independent.

**Other requirements:** Nextflow and Java 8 or higher, docker or singularity for containerization.

**Programming language**: PHP, Javascript.

**License**: GPL-v3.

**Restrictions to use by non-academics**: None.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s12864-020-6714-x.

**Additional file 1: Figure S1.** RNA-Seq Pipeline. **Figure S2.** Adapter Removal, Trimmer and Quality Filtering Module. **Figure S3.** Sequential Mapping Module. **Figure S4.** BAM Analysis Module. **Figure S5.** ATAC-Seq pipeline. **Figure S6.** ChIP-Seq pipeline.

### Abbreviations

NGS: Next generation sequencing; RNA-Seq: RNA sequencing; ATAC-Seq: Assay for transposase-accessible chromatin using sequencing; ChIP-Seq: Chromatin immunoprecipitation sequencing; IGV: Integrative genome viewer; TCGA: The cancer genome atlas; TDF: Tiled data file; TPM: Transcript per million; UID: Unique identifier; UCSC: University of California, Santa Cruz; DNA: Deoxyribonucleic acid; RNA: Ribonucleic acids; rRNAs: Ribosomal RNAs; miRNAs: Micro RNAs; tRNAs: Transfer RNAs; piRNAs: Piwi-interacting RNAs; snoRNAs: Small nucleolar RNAs; ERCC: External RNA controls consortium; Tn5: Transposase 5; QC: Quality control; PDF: Portable Document Format; HTML: Hypertext Markup Language; CPU: Central processing unit; BAM: Binary alignment map; WIG: Wiggle Track Format; BED: Browser Extensible Data; AWS: Amazon web services; EFS: Elastic file system; EBS: Elastic block store; S3: Simple cloud storage service; GCP: Google cloud platform; SGE: Sun grid engine; SLURM: Simple linux utility for resource management; LSF: Load sharing facility

**Acknowledgements**
We would like to thank Phil Zamore and members of his lab for their advice and suggestions and Elissa Donnard, Kyle Gellatly, Alan Derr, Athma Pai, Rachel Murphy, and Laney Zuerlein for constructive suggestions, to all members of the Garber Lab for testing and improving pipelines and for their comments on the manuscript.

**Rights and permissions**
Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

**Authors' contributions**
OY and AK implemented the project. AK and MG supervised the project and wrote the manuscript. ARO and OT helped in the development of the pipeline builder visualization module. All authors have read and approved the final manuscript.

**Funding**
MG, AK and OY were supported by funds from the National Human Genome Research Institute NHGRI grant #U01 HG007910–01 and the National Center for Advancing Translational Sciences grant #UL1 TR001453–01, ARO and OT were supported by funds from the Howard Hughes Medical Institute and the National Institutes of Health (NIH) P01# HD078253. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

**Availability of data and materials**
Supplementary Material.docx: Supplementary figures.

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare that they have no competing interests.

**Author details**
[1]Bioinformatics Core, University of Massachusetts Medical School, Worcester, MA 01605, USA. [2]RNA Therapeutics Institute, University of Massachusetts Medical School, Worcester, MA 01605, USA. [3]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA. [4]Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA.

Received: 22 October 2019 Accepted: 1 April 2020
Published online: 19 April 2020

**References**

1. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics. 2015;8:33.
2. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. 2016;17:53.
3. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet. 2006;38:500–1.
4. Clevert D-A, Rasche A. The Affymetrix GeneChip® Microarray Platform. In: Handbook of Research on Systems Biology Applications in Medicine; 2009. p. 251–61. https://doi.org/10.4018/978-1-60566-076-9.ch014.
5. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46:W537–44.
6. Kassambara A, Rème T, Jourdan M, Fest T, Hose D, Tarte K, et al. GenomicScape: an easy-to-use web tool for gene expression data analysis. Application to investigate the molecular events in the differentiation of B cells into plasma cells. PLoS Comput Biol. 2015;11:e1004077.
7. Halbritter F, Vaidya HJ, Tomlinson SR. GeneProf: analysis of high-throughput sequencing experiments. Nat Methods. 2011;9:7–8.
8. Hoon S, Ratnapu KK, Chia J-M, Kumarasamy B, Juguang X, Clamp M, et al. Biopipe: a flexible framework for protocol-based bioinformatics analysis. Genome Res. 2003;13:1904–15.
9. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Res. 2013; 41(Web Server issue):W557–61.
10. Desvillechabrol D, Legendre R, Rioualen C, Bouchier C, van Helden J, Kennedy S, et al. Sequanix: a dynamic graphical interface for Snakemake workflows. Bioinformatics. 2018;34:1934–6.
11. Kluge M, Friedel CC. Watchdog - a workflow management system for the distributed analysis of large-scale experimental data. BMC Bioinformatics. 2018;19:97.
12. DNAnexus, Saphetor Partner on Genomic Analysis Solution. Clin OMICs. 2017;4:31. https://doi.org/10.1089/clinomi.04.05.23.
13. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, et al. The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research. Cancer Res. 2017;77:e3–6.
14. Illumina to Integrate Watson for Genomics into BaseSpace Sequence Hub. Clin OMICs. 2017;4:32. https://doi.org/10.1089/clinomi.04.01.28.
15. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017;550:345–53.
16. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533:452–4..
17. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.
18. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38:276–8.
19. Federico A, Karagiannis T, Karri K, Kishore D, Koga Y, Campbell JD, et al. Pipeliner: A Nextflow-Based Framework for the Definition of Sequencing Data Processing Pipelines. Front Genet. 2019;10:614.
20. Guzman C, D'Orso I. CIPHER: a flexible and extensive workflow platform for integrative next-generation sequencing data analysis and genomic regulatory element prediction. BMC Bioinformatics. 2017;18:363.
21. Stodden V, Leisch F, Peng RD. Implementing Reproducible Research. New York: CRC Press; 2014.
22. Peng RD. Reproducible research in computational science. Science. 2011; 334:1226–7.
23. Baichoo S, Souilmi Y, Panji S, Botha G, Meintjes A, Hazelhurst S, et al. Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics. BMC Bioinformatics. 2018;19:457.
24. Kulkarni N, Alessandrì L, Panero R, Arigoni M, Olivero M, Ferrero G, et al. Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. BMC Bioinformatics. 2018;19(Suppl 10): 349.
25. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One. 2017;12:e0177459.
26. Hale JS, Li L, Richardson CN, Wells GN. Containers for Portable, Productive, and Performant Scientific Computing. Computing in Science Engineering. 2017;19:40–50.
27. de Souza N. The ENCODE project. Nature methods. 2012;9:1046.
28. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. Biopreserv Biobank. 2015;13:311–9.
29. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.
30. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018;46:D754–61.
31. Gulabani S. Hands-on Elastic Compute Cloud. In: Practical Amazon EC2, SQS, Kinesis, and S3; 2017. p. 23–88. https://doi.org/10.1007/978-1-4842-2841-8_2.
32. Documentation Team. Amazon Elastic Compute Cloud User Guide for Linux Instances. Samurai Media Limited; 2018.

33. Emeras J, Varrette S, Plugaru V, Bouvry P. Amazon Elastic Compute Cloud (EC2) vs. in-House HPC Platform: a Cost Analysis. IEEE Transactions on Cloud Computing; 2016. p. 1. https://doi.org/10.1109/tcc.2016.2628371.
34. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28:2184–5.
35. Donnard E, Vangala P, Afik S, McCauley S, Nowosielska A, Kucukural A, et al. Comparative Analysis of Immune Cells Reveals a Conserved Regulatory Lexicon. Cell Syst. 2018. https://doi.org/10.1016/j.cels.2018.01.002.
36. Kucukural A, Yukselen O, Ozata DM, Moore MJ, Garber M. DEBrowser: interactive differential expression analysis and visualization tool for count data. BMC Genomics. 2019;20:6.
37. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8.
38. Tripathi R, Sharma P, Chakraborty P, Varadwaj PK. Next-generation sequencing revolution through big data analytics. Frontiers in Life Science. 2016;9:119–49.
39. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24.
40. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014; 42(Database issue):D764–70.
41. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.
42. Munro SA, Lund SP, Pine PS, Binder H, Clevert D-A, Conesa A, et al. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nat Commun. 2014;5:5125.
43. Bioinformatics B. FastQC: a quality control tool for high throughput sequence data. Cambridge: Babraham Institute; 2011.
44. Andrews S, Others. FastQC: a quality control tool for high throughput sequence data. 2010.
45. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–2.
46. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
47. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
48. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
49. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(Database issue):D662–9.
50. Harrow J, Denoeud F, Frankish A, Reymond A, Chen C-K, Chrast J, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006;7(Suppl 1):S4.1–9.
51. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22:1760–74.
52. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014;42(Database issue):D756–63.
53. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
54. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
55. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.
56. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
57. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–9.
58. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10:1213–8.
59. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.
60. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
61. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.
62. Sissaoui S, Yu J, Yan A, Li R, Zhu LJ, Kucukural A, et al. Genomic Characterization of Endothelial Enhancers Reveals a Multifunctional Role for NR2F2 in Regulation of Arteriovenous Gene Expression. Circ Res. 2020. https://doi.org/10.1161/CIRCRESAHA.119.316075.
63. Amstutz P. Portable, Reproducible Analysis with Arvados. F1000Res. 2015;4 https://f1000research.com/assets/download/1110114.
64. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2.
65. Chapman B, Gentry J, Lin M, Magee P, O'Connor B, Prabhakaran A, et al. OpenWDL. 2019.
66. Lamprecht A-L, Garcia L, Kuzak M, Martinez C, Arcila R, Martin Del Pico E, et al. Towards FAIR principles for research software. Data Sci. 2019; vol. Pre-press:1–23.


## Publisher's Note

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


**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

**Learn more** biomedcentral.com/submissions