DolphinNext: A distributed data processing platform for high throughput genomics

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

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

*corresponding: alper.kucukural@umassmed.edu or manuel.garber@umassmed.edu

ABSTRACT

The emergence of high throughput technologies, that produce vast amounts of genomic data, such as next-generation sequencing (NGS) are transforming biological research. The dramatic increase in the volume of data makes 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 a highly-parallel, flexible, portable, and reproducible data processing frameworks. Flexibility ensures that pipelines can support a variety of applications without requiring one-off modifications. Portability ensures that users can leverage computationally available resources and work within economic constraints. Reproducibility warrants credibility to the results, and is particularly challenging in the face of the sheer volume of data and the complexity of processing pipelines that vary widely between users.

Several platforms currently exist for the design and execution of complex pipelines (e.g. Galaxy, GenePattern, GeneProf). Unfortunately, these platforms lack the necessary combination of parallelism, portability, flexibility and/or reproducibility that are required by the current research environment. To address these shortcomings, we have created DolphinNext, an easy-to-use platform for creating and deploying complex workflows for high throughput data processing. The guiding principle of DolphinNext is to facilitate the building and deployment of complex pipelines using a modular approach implemented in a graphical interface. DolphinNext provides: 1. A drag and drop user interface that abstracts pipelines and allows users to create pipelines without familiarity with underlying programming languages. 2. A user interface to monitor pipeline execution that allows the re-initiation of pipelines at intermediate steps 3. Reproducible pipelines with version tracking and stand-alone versions that can be run independently. 4. Seamless portability to distributed computational environments such as high performance clusters or cloud computing environments.

KEYWORDS

Pipeline, Workflow, Genome analysis, big data processing, Sequencing

INTRODUCTION

Analysis of high-throughput data is now widely regarded as the major bottleneck in modern biology. In response, resource allocation has dramatically skewed towards computational power, with significant impacts on budgetary decisions (1). 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 most simple 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 (2) and Affymetrix (3), that provide 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 (4), GenomicScape (5), and other platforms (2, 6–13). 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.

The paradigm used by previous platforms required a central server to which data could be uploaded. This implementation hampers the ability of these frameworks to address three fundamental needs of current and future data driven biological research. First different organizations have implemented specific data privacy and protection policies regarding the location, storage and accessibility to patient and other sensitive data. Uploading sensitive data to an external server is discouraged in these environments. Second, there is a need to improve on the reproducibility of analyses by the research community at large (14). 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.

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 can be run as a centralized server, a stand-alone application on a personal computer, or as a graphical interface to create workflows and then to monitor their execution. DolphinNext can be installed and run within the organization IT premises and integrated with its data security needs.

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. To easily enable third parties to reproduce analyses, complete pipelines along with the computing environment used must be available and need to be easily shared (15–18). DolphinNext allows users to package pipelines into portable containers (19, 20) 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.

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 on 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 vast sequencing data such as ENCODE (21), GTex (22), 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.

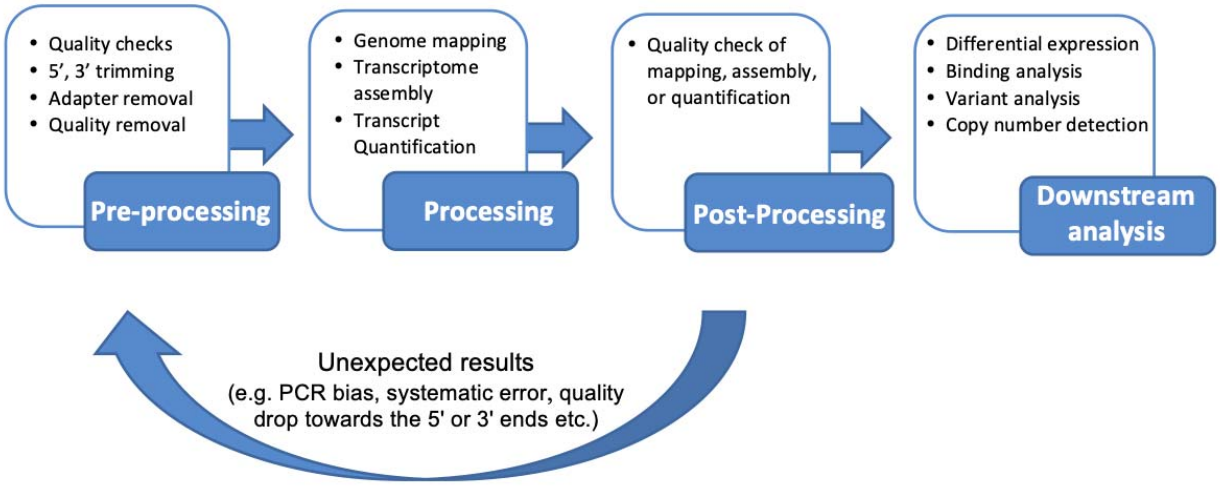

Figure 1. General diagram of a typical data processing and analysis cycle

Third, DolphinNext implements pipelines that incorporate best practices derived from the previous decade of genomics research. It is now clear that analysis of high-throughput sequencing data consists of two main tasks: data processing and data analysis (**Figure 1**). 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) (23). 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. Data processing also greatly benefits from parallelization, and for large projects it is generally carried out on a high performance computing clusters or cloud computing environments. 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 (24) and Wiggle Track Format (WIG, or BigWIG) for the UCSC genome browser (25, 26). 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.

To accomplish its goal of providing a flexible, reproducible and scalable platform for sequencing data or any type of computationally intensive data processing, DolphinNext is built on top of the Nextflow platform (27). As such, it readily integrate with high performance and cloud computing environments. More importantly, 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.

In summary, 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
- Portability across computational environments such as workstations, computing clusters, or cloud-based servers
- Built-in pipeline revision control
- Full access to application run logs
- Parallel execution of non-dependent processes
- Integrated data analysis and reporting interface with R markdown support

Finally, DolphinNext can be easily packaged in portable containers that include the exact programs and versions of the pipelines, making it easy to distribute pipelines (including relevant datasets) to ensure easy replication of computational results.

RESULTS

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 a 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 (27) is a language to create scalable and reproducible scientific workflows. In creating DolphinNext, we aimed 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 simplify 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 with all users in the system. Users can repurpose existing processes belonging to any pipeline, 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.

DolphinNext pipelines are versioned, allowing users to improve pipelines or processes, while allowing the previous, stable versions to still be available. All the pipelines and processes in DolphinNext carry unique revision IDs (UIDs) that enable users to run an older version of a pipeline or process should a specific reason to reproduce an old result emerge.

To facilitate reproducibility of data processing and the execution of pipelines in any computing environment, DolphinNext supports Singularity and Docker container technologies (19, 20). 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. Because all other dependencies are included in the image, DolphinNext pipelines can be executed in remote machines or clusters without the need to 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 re-running of a pipeline, DolphinNext provides the ability to re-execute or resume a pipeline from any of its steps. Thus, pipelines can be resumed from a failure point after correcting the parameters that resulted in the failure. Similarly, previously executed pipelines can 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 admin. 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 (28) are defined as a default option, the user can change it to Ensembl annotations (29) 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 were, a new, dedicated computer cluster can easily be set up within DolphinNext with just a few clicks. 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 (30–32).

General Implementation and Structure

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

1. Profile Module:

Users may have access to a wide range of different computing environments: A workstation, Cloud Computing, or a high performance computing clusters where jobs are submitted through a job scheduler such as IBM's LSF, SLURM or Apache Ignite. DolphinNext relies on NextFlow (27) to encapsulate computing environment settings and allows administrators to rely on a single configuration file that enables users to run 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.

2. 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 become complex, pipeline maintenance can be a daunting task.

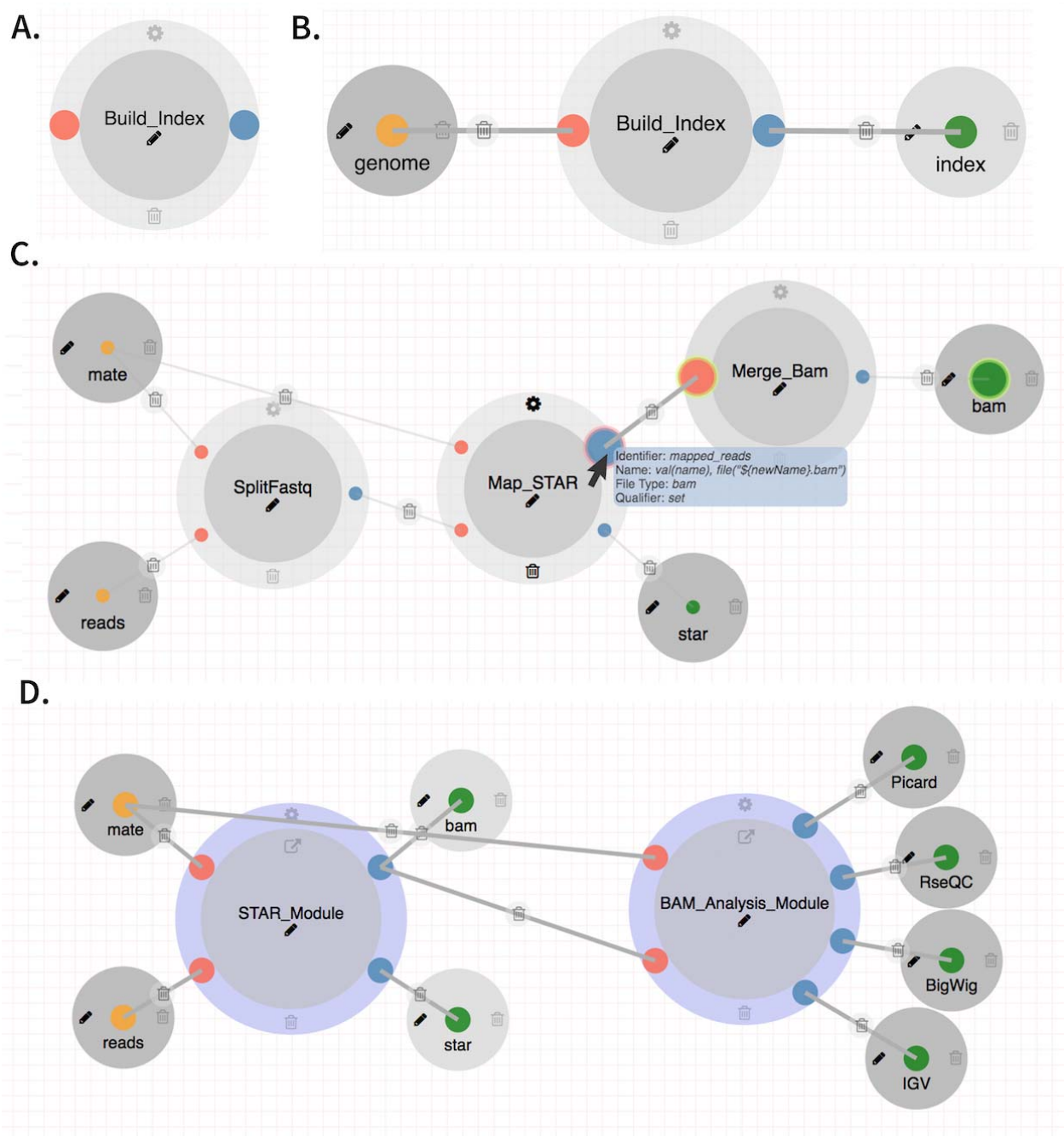

Figure 2. A. A process for building index files **B.** Input and output parameters attached to a process **C.** The STAR alignment module connected through input/output with matching parameter types. **D.** The RNA-Seq pipeline can be designed using two nested pipelines: the STAR pipeline and the BAM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

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 **(Figure 2. A)** and pipelines and use drag and drop functionality to create new workflows by connecting processes through their input and output parameters **(Figure 2. B)**. Two processes can only be connected when the output data type of one is compatible with the input data type of the second **(Figure 2. C)**. Upon connecting two compatible processes DolphinNext creates all necessary execution dependencies.

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:

2. Pipeline Builder:

Processes are the core units in a pipeline they carry out self-contained operations that carry out 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 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 necessary information to run. The 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. For example, if a pipeline designer needs to create a dropdown selection method for the genome reference input parameter she would specify this parameter by adding the directives below:

```
params.genome_build = "" //* @dropdown @options:"human_hg19, mouse_mm10, custom"
```

Similarly, to specify whether the transcript quantification program, RSEM (33) should be run in a pipeline, the pipeline designer can capture this information by specifying a run_RSEM parameter with a default option of running and two possible options ‘yes’ and ‘no’:

```
params.run_RSEM = "yes" //* @dropdown @options:"yes","no”
```

In both cases, these directives ensure that the process form includes the options and defaults intended by the designer, irrespective of the pipeline where the process is included.

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. This allows process builders to include directives and obtain default parameters for general usage, such as RSEM parameters to control the number of threads or to determine whether alignment files will be generated upon completion of the run. In this way, this pipeline can be run with the default parameters. For example, the default parameters for an RSEM process can be set to run with four processors and no alignment files will be generated to reduce the space usage by specifying:

```
//*autofill
RSEM_parameters = "-p 4" //* @input @description:"RSEM params" @tooltip:"RSEM v1.2.28"
no_bam_output = "true" //* @dropdown @options:"true","false" @description:"If true is
selected, RSEM will not output any BAM file (default=false)"
//*
```

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. In this way, we keep pipelines safe from modifications yet allowing for improvements to be available in new revisions.

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 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.

Pipelines can also be downloaded in NextFlow format for documentation, distribution and execution outside DolphinNext. To allow users and administrators to make pipelines available across installations, DolphinNext supports pipeline import and export. Further, all pipelines are given a unique identifier (uid). This mechanism guarantees that the same pipeline is imported only once and provides a robust way to compare pipeline versions across DolphinNext installations.

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 **(Figure 2. D)**. It relies on RseQC (34) and Picard (35) 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 maintenance of data processing pipelines.

3. 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 permission, 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 (**Figure 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:

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.

Accessibility to all run logs. To monitor run specific details and troubleshoot failed runs, DolphinNext provides access to all application logs. Further, DolphinNext 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.

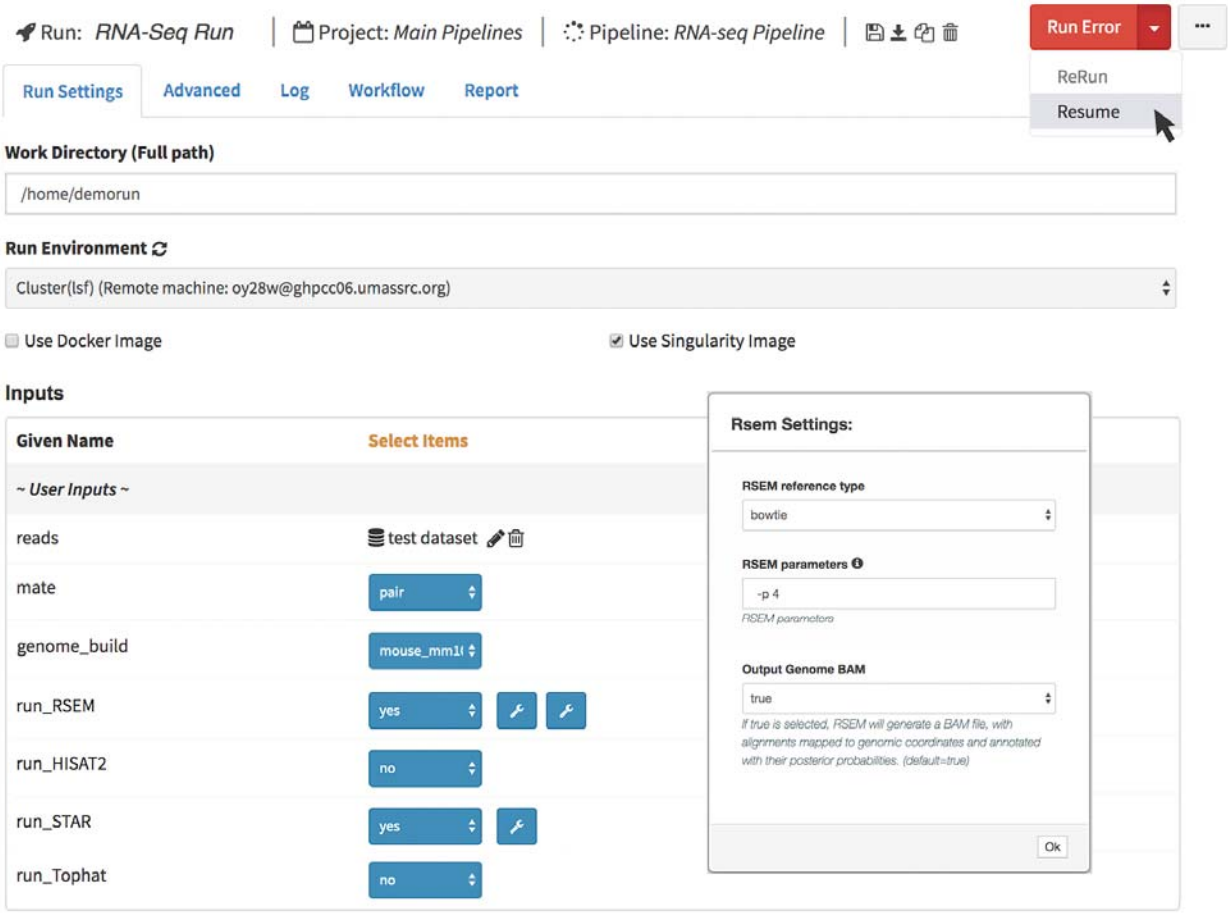

Figure 3. Resuming RNA-Seq pipeline after changing RSEM parameters

4. 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 (36). 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, these data can also be seamlessly loaded and analyzed using an embedded DEBrowser, an interactive visualization tool for count data (23). PDF and HTML files can be visualized within DolphinNext or downloaded for offline analysis (**Figure 4**).

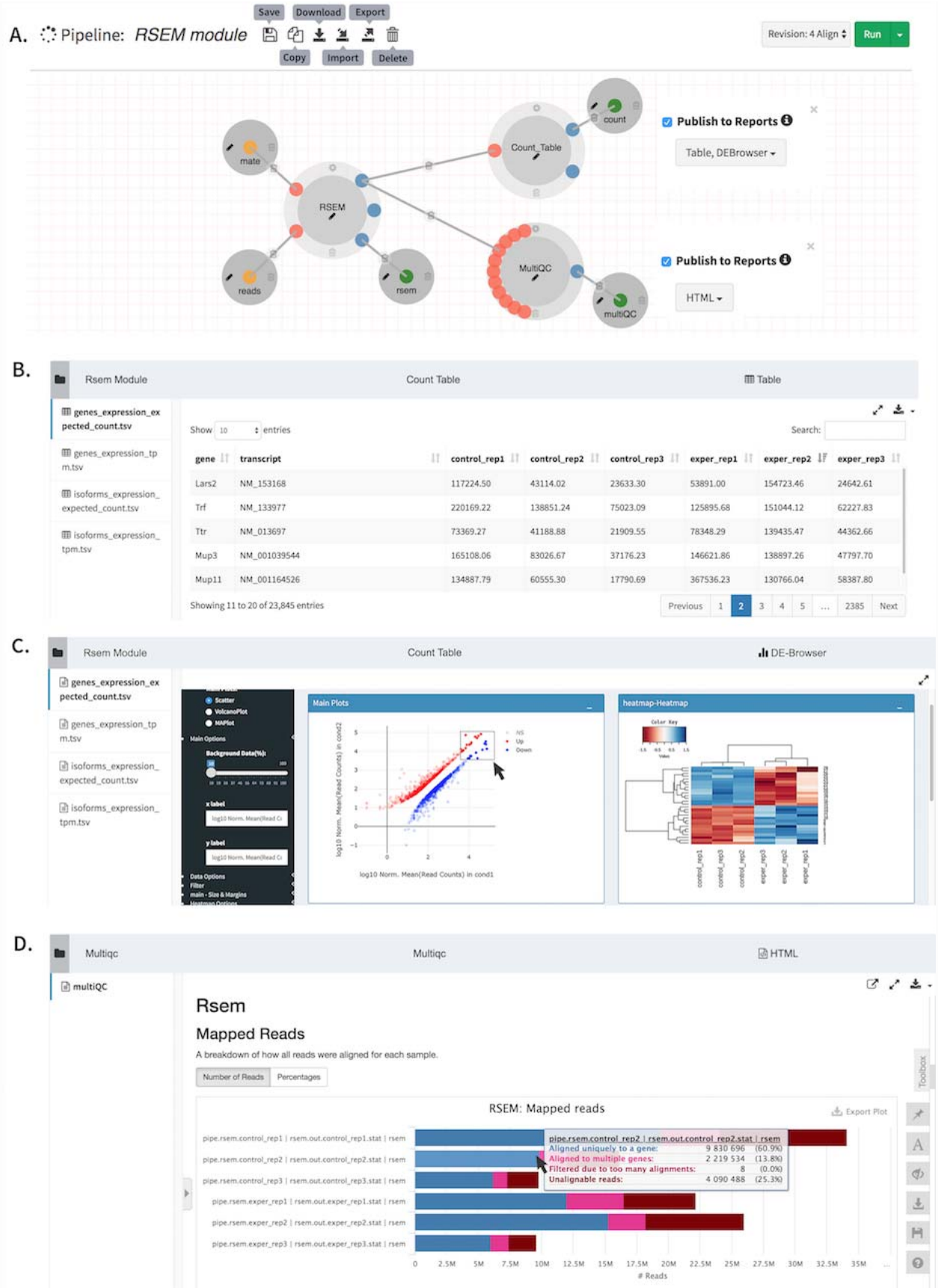
A.
Pipeline: RSEM module
Save
Download
Export
Copy
Import
Delete
Revision: 4 Align
Run
mate
reads
RSEM
rsem
Count_Table
count
MultiQC
multiQC
Publish to Reports
Table, DEBrowser
Publish to Reports
HTML
B.
Rsem Module
Count Table
Table
genes_expression_expected_count.tsv
genes_expression_tpm.tsv
isoforms_expression_expected_count.tsv
isoforms_expression_tpm.tsv
Show 10 entries
Search:
gene | transcript | control_rep1 | control_rep2 | control_rep3 | exper_rep1 | exper_rep2 | exper_rep3
Lars2 | NM_153168 | 117224.50 | 43114.02 | 23633.30 | 53891.00 | 154723.46 | 24642.61
Trf | NM_133977 | 220169.22 | 138851.24 | 75023.09 | 125895.68 | 151044.12 | 62227.83
Ttr | NM_013697 | 73369.27 | 41188.88 | 21909.55 | 78348.29 | 139435.47 | 44362.66
Mup3 | NM_001039544 | 165108.06 | 83026.67 | 37176.23 | 146621.86 | 138897.26 | 47797.70
Mup11 | NM_001164526 | 134887.79 | 60555.30 | 17790.69 | 367536.23 | 130766.04 | 58387.80
Showing 11 to 20 of 23,845 entries
Previous 1 2 3 4 5 ... 2385 Next
C.
Rsem Module
Count Table
DE-Browser
genes_expression_expected_count.tsv
genes_expression_tpm.tsv
isoforms_expression_expected_count.tsv
isoforms_expression_tpm.tsv
Main Plots
heatmap-Heatmap
log10 Norm. Mean(Read Counts) in cond1
log10 Norm. Mean(Read Counts) in cond2
D.
Multiqc
Multiqc
HTML
multiQC
Rsem
Mapped Reads
A breakdown of how all reads were aligned for each sample.
Number of Reads
Percentages
RSEM: Mapped reads
Export Plot
pipe.rsem.control_rep1 | rsem.out.control_rep1.stat | rsem
pipe.rsem.control_rep2 | rsem.out.control_rep2.stat | rsem
pipe.rsem.control_rep3 | rsem.out.control_rep3.stat | rsem
pipe.rsem.exper_rep1 | rsem.out.exper_rep1.stat | rsem
pipe.rsem.exper_rep2 | rsem.out.exper_rep2.stat | rsem
pipe.rsem.exper_rep3 | rsem.out.exper_rep3.stat | rsem
Aligned uniquely to a gene: 9 830 696 (60.9%)
Aligned to multiple genes: 2 219 534 (13.8%)
Filtered due to too many alignments: 8 (0.0%)
Unalignable reads: 4 090 488 (25.3%)
Reads
Toolbox

Figure 4. A. RSEM module which involves Count_Table (where consolidated count table prepared) and multiQC processes. Table and Debrowser report format have been set for the Count_Table process, whearase HTML is selected for multiQC process. **B.** Count table report **C.** MultiQC report, and **D.** DEBrowser module for interactive differential expression analysis.

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 (**Figure 5**).

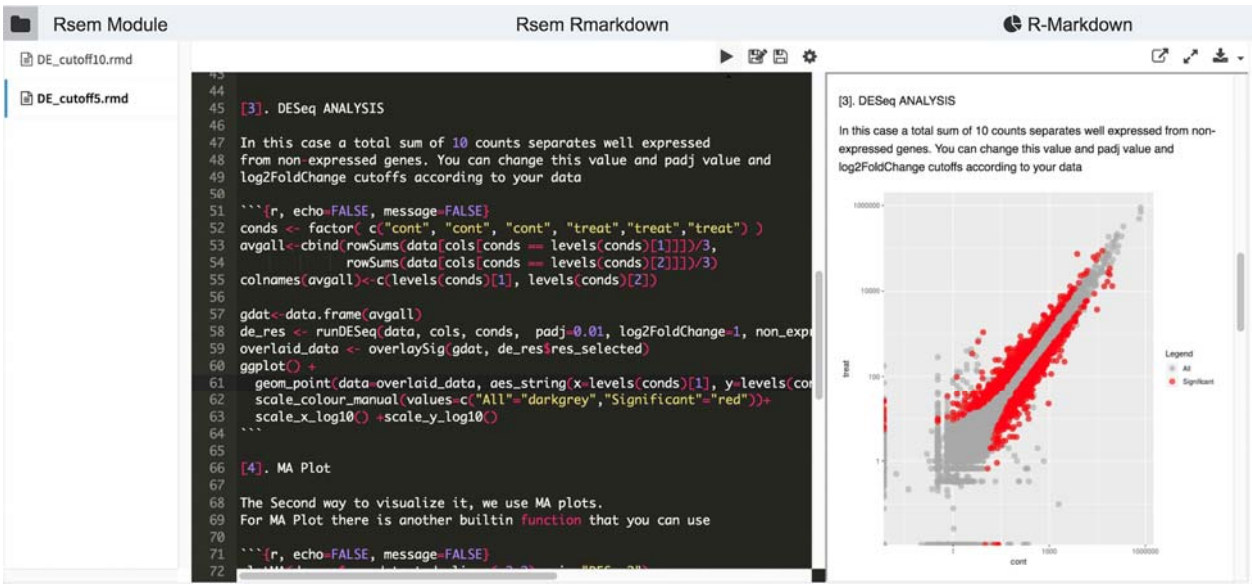

Figure 5. R markdown report for RNA-Seq analysis. Users can adapt this template according to their needs.

PIPELINE EXAMPLES

The pipeline examples here are publicly available and can be downloaded from the examples section in the GitHub repository. (https://github.com/UMMS-Biocore/dolphinnext)

1. RNA-Seq Pipeline (Figure S1):

All supported 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 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 generic quality control, 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 (37), mobile elements). Before any such alignment, reads may be trimmed to a 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 (38) 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

(39), adapter removal using cutadapt (40), 5' and 3' trimming, quality removal using trimmomatic (41), and Bowtie2 (42) 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 (33) which aligns reads to a predefined set of transcripts. The user can use available transcript sets (i.e Ensemble (43), GENCODE (44, 45), RefSeq (46)) 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 (47), Hisat2 (48) and Tophat2 (49). Resulting alignments are then processed to generate genome browser friendly formats: bigwig (for UCSC genome browser (25) or TDF (for the integrative genome viewer (IGV, (24)).

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 (35) and the RSeQC (34) program.

Finally, the RNA-Seq pipeline returns a quantification matrix that includes the estimated counts and transcripts per million (TPM) 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 software, which allows interactive exploration of the resulting data (23).

2. 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 (42). 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 (50) merge function. The pipeline then allows users to estimate and remove PCR duplicates using the Picard's mark duplicates (35) 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 (36, 51). Once each read has been shortened, peaks are called identically in both the ChIP-Seq and ATAC-Seq using MACS2 (52).

When processing several samples together, the ATAC and ChIP pipelines provide consensus peak calls by merging all peaks individually called in each sample using Bedtools (53). The number of reads in each peak location are then quantified using Bedtools (53) 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 (23) to perform differential analysis or downloaded to perform other analysis.

Table 1. Comparison to related applications.

	DolphinNext	Galaxy	Sequanix	Taverna	Arvados
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	No	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
Container Support					
Docker support	Yes	Yes	Yes	No	Yes
Singularity support	Yes	No	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
Autoscaling	Yes	No	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 module, 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.

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 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 (DEBrower (23) 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

DolphinNext is an open source collaborative initiative available in the GitHub repository (https://github.com/UMMS-Biocore/dolphinnext). A documentaion, quick start guide and video tutorials are availbale in readthedocs. (https://dolphinnext.readthedocs.org). A docker image to run dolphinNext is in the github repository (https://github.com/UMMS-Biocore/dolphinnext-studio). The web server is in https://dolphinnext.umassmed.edu.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGMENTS

We would like to thank Phil Zamore and members of his lab for their advice and suggestions and Athma Pai, Alan Derr, Rachel Murphy, Kyle Gellatly, 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. We would also like to thank Craig C. Mello and his lab to initiate the project.

FUNDING

This work was 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 (M.G). A part of the visualization module was funded by the Howard Hughes Medical Institute (C.C.M.), and the National Institutes of Health (NIH) P01# HD078253 (C.C.M).

Conflict of Interest: none declared.

AUTHORS' CONTRIBUTIONS

OY and AK implemented project. AK and MG supervised the project and wrote the manuscript. OT and AO helped developing pipeline builder visualization module. All authors have read and approved the final manuscript.

REFERNCES

1. Muir,P., Li,S., Lou,S., Wang,D., Spakowicz,D.J., Salichos,L., Zhang,J., Weinstock,G.M., Isaacs,F., Rozowsky,J., *et al.* (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.*, **17**, 53.

2. Reich,M., Liefeld,T., Gould,J., Lerner,J., Tamayo,P. and Mesirov,J.P. (2006) GenePattern 2.0. *Nat. Genet.*, **38**, 500–501.

3. Clevert,D.-A. and Rasche,A. (2009) The Affymetrix GeneChip® Microarray Platform. *Handbook of Research on Systems Biology Applications in Medicine*, 10.4018/978-1-60566-076-9.ch014.

4. Afgan,E., Baker,D., Batut,B., van den Beek,M., Bouvier,D., Cech,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.

5. Kassambara,A., Rème,T., Jourdan,M., Fest,T., Hose,D., Tarte,K. and Klein,B. (2015) 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.*, **11**, e1004077.

6. Halbritter,F., Vaidya,H.J. and Tomlinson,S.R. (2011) GeneProf: analysis of high-throughput sequencing experiments. *Nat. Methods*, **9**, 7–8.

7. Hoon,S., Ratnapu,K.K., Chia,J.-M., Kumarasamy,B., Juguang,X., Clamp,M., Stabenau,A., Potter,S., Clarke,L. and Stupka,E. (2003) Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Res.*, **13**, 1904–1915.

8. Wolstencroft,K., Haines,R., Fellows,D., Williams,A., Withers,D., Owen,S., Soiland-Reyes,S., Dunlop,I., Nenadic,A., Fisher,P., *et al.* (2013) The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res.*, **41**, W557–61.

9. 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. *Bioinformatics*, **34**, 1934–1936.

10. Kluge,M. and Friedel,C.C. (2018) Watchdog - a workflow management system for the distributed analysis of large-scale experimental data. *BMC Bioinformatics*, **19**, 97.

11. DNAnexus, Saphetor Partner on Genomic Analysis Solution (2017) *Clinical OMICs*, **4**, 31–31.

12. Lau,J.W., Lehnert,E., Sethi,A., Malhotra,R., Kaushik,G., Onder,Z., Groves-Kirkby,N., Mihajlovic,A., DiGiovanna,J., Srdic,M., *et al.* (2017) The Cancer Genomics Cloud: Collaborative, Reproducible, and

Democratized-A New Paradigm in Large-Scale Computational Research. *Cancer Res.*, **77**, e3–e6.

13. Illumina to Integrate Watson for Genomics into BaseSpace Sequence Hub (2017) *Clinical OMICs*, **4**, 32–32.

14. Baker,M. (2016) 1,500 scientists lift the lid on reproducibility. *Nature*, **533**, 452–454.

15. Stodden,V., Leisch,F. and Peng,R.D. (2014) Implementing Reproducible Research CRC Press.

16. Peng,R.D. (2011) Reproducible research in computational science. *Science*, **334**, 1226–1227.

17. Baichoo,S., Souilmi,Y., Panji,S., Botha,G., Meintjes,A., Hazelhurst,S., Bendou,H., Beste,E. de, Mpangase,P.T., Souiai,O., *et al.* (2018) Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics. *BMC Bioinformatics*, **19**, 457.

18. Kulkarni,N., Alessandrì,L., Panero,R., Arigoni,M., Olivero,M., Ferrero,G., Cordero,F., Beccuti,M. and Calogero,R.A. (2018) Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics*, **19**, 349.

19. Kurtzer,G.M., Sochat,V. and Bauer,M.W. (2017) Singularity: Scientific containers for mobility of compute. *PLoS One*, **12**, e0177459.

20. Hale,J.S., Li,L., Richardson,C.N. and Wells,G.N. (2017) Containers for Portable, Productive, and Performant Scientific Computing. *Computing in Science Engineering*, **19**, 40–50.

21. de Souza,N. (2012) The ENCODE project. *Nat. Methods*, **9**, 1046.

22. Carithers,L.J., Ardlie,K., Barcus,M., Branton,P.A., Britton,A., Buia,S.A., Compton,C.C., DeLuca,D.S., Peter-Demchok,J., Gelfand,E.T., *et al.* (2015) A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreserv. Biobank.*, **13**, 311–319.

23. Kucukural,A., Yukselen,O., Ozata,D.M., Moore,M.J. and Garber,M. (2019) DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics*, **20**, 6.

24. Robinson,J.T., Thorvaldsdóttir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24.

25. Karolchik,D., Barber,G.P., Casper,J., Clawson,H., Cline,M.S., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M., *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, **42**, D764–70.

26. Kent,W.J. (2002) The Human Genome Browser at UCSC. *Genome Research*, **12**, 996–1006.

27. Di Tommaso,P., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.

28. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D., *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–45.

29. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G., *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.

30. Gulabani,S. (2017) Hands-on Elastic Compute Cloud. *Practical Amazon EC2, SQS, Kinesis, and S3*, 10.1007/978-1-4842-2841-8_2.

31. Documentation Team (2018) Amazon Elastic Compute Cloud User Guide for Linux Instances

Samurai Media Limited.

32. Emeras,J., Varrette,S., Plugaru,V. and Bouvry,P. (2016) Amazon Elastic Compute Cloud (EC2) vs. in-House HPC Platform: a Cost Analysis. *IEEE Transactions on Cloud Computing*, 10.1109/tcc.2016.2628371.

33. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

34. Wang,L., Wang,S. and Li,W. (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, **28**, 2184–2185.

35. Broad Institute Picard Tools. *Broad Institute, GitHub repository*.

36. Donnard,E., Vangala,P., Afik,S., McCauley,S., Nowosielska,A., Kucukural,A., Tabak,B., Zhu,X., Diehl,W., McDonel,P., *et al.* (2018) Comparative Analysis of Immune Cells Reveals a Conserved Regulatory Lexicon. *Cell Syst*, 10.1016/j.cels.2018.01.002.

37. Munro,S.A., Lund,S.P., Pine,P.S., Binder,H., Clevert,D.-A., Conesa,A., Dopazo,J., Fasold,M., Hochreiter,S., Hong,H., *et al.* (2014) Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nat. Commun.*, **5**, 5125.

38. Bioinformatics,B. (2011) FastQC: a quality control tool for high throughput sequence data. *Cambridge, UK: Babraham Institute*.

39. Andrews,S. and Others (2010) FastQC: a quality control tool for high throughput sequence data.

40. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.

41. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

42. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

43. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S., *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–9.

44. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.-K., Chrast,J., Lagarde,J., Gilbert,J.G.R., Storey,R., Swarbreck,D., *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7 Suppl 1**, S4.1–9.

45. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S., *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.

46. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M., *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–63.

47. 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. *Bioinformatics*, **29**, 15–21.

48. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

49. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

50. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

51. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.

52. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

53. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.