

DolphinNext

A graphical user interface for reproducible pipelines

Onur Yukselen, Alper Kucukural
Bootcamp, May 05, 2021



Content

- ▶ Overview
- ▶ Public Pipelines
- ▶ Running Pipelines
- ▶ Reports

UMASS Web Server

► <https://dolphinnext.umassmed.edu>

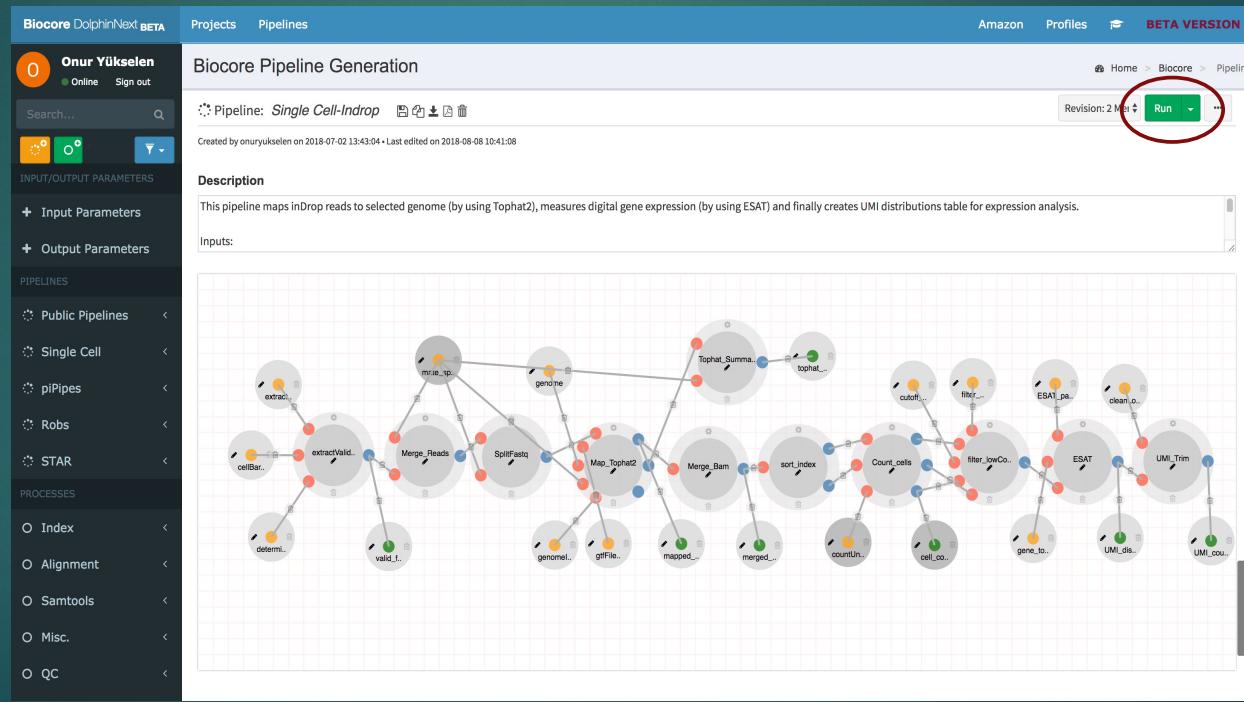
The screenshot shows the 'Public Pipelines' section of the Biocore Generation page. It features six pipeline cards:

- Single Cell-10X Genomics**: Maps 10X Genomics reads to selected genome (using Tophat2), measures digital gene expression (using ESAT) and creates UMI distributions table for expression analysis. Inputs: "Determined_fastq". A 'LEARN MORE' button is highlighted with a red circle.
- Single Cell-Indrop**: Maps inDrop reads to selected genome (using Tophat2), measures digital gene expression (using ESAT) and creates UMI distributions table for expression analysis. Inputs: "Determined_fastq". A 'LEARN MORE' button is visible.
- RSEM Pipeline**: Tophat pipeline includes Quality Control, rRNA filtering, Genome Alignment using Tophat. Steps: 1) For Quality Control, we use FastQC to create qc outputs. 2) Bowtie2 is used to filter out rRNA reads. 3) TopHat is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splicing. A 'LEARN MORE' button is visible.
- HISAT2 pipeline**: Tophat pipeline includes Quality Control, rRNA filtering, Genome Alignment using Tophat. Steps: 1) For Quality Control, we use FastQC to create qc outputs. 2) Bowtie2 is used to filter out rRNA reads. 3) TopHat is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splicing. A 'LEARN MORE' button is visible.
- STAR Pipeline**: Tophat pipeline includes Quality Control, rRNA filtering, Genome Alignment using Tophat. Steps: 1) For Quality Control, we use FastQC to create qc outputs. 2) Bowtie2 is used to filter out rRNA reads. 3) TopHat is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splicing. A 'LEARN MORE' button is visible.
- Tophat2 Pipeline**: Tophat pipeline includes Quality Control, rRNA filtering, Genome Alignment using Tophat. Steps: 1) For Quality Control, we use FastQC to create qc outputs. 2) Bowtie2 is used to filter out rRNA reads. 3) TopHat is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splicing. A 'LEARN MORE' button is visible.

A navigation bar at the bottom shows pages 1, 2, 3, and >.

UMASS Web Server

► <https://dolphinnext.umassmed.edu>



Public Pipelines

RNA-Seq Pipeline:

- Genomic Alignment: HISAT2, STAR or Tophat2
- Estimation of gene/isoform expression levels: RSEM or Feature Counts
- Differential expression analysis: DEseq2/EdgeR/Limma with DEBrowser

ATAC-Seq/ChIP-Seq Pipeline

- Genomic Alignment: Bowtie2
- Peak Calling: MACS2
- Differential analysis: DEseq2/EdgeR/Limma with DEBrowser

Sub modules:

Pre-processing:

- Adapter Removal
- Trimmer
- Quality Filtering
- Sequential Mapping (rRNA, piRNA etc.)

Quality Control :

- FastQC
- RSeQC
- Picard

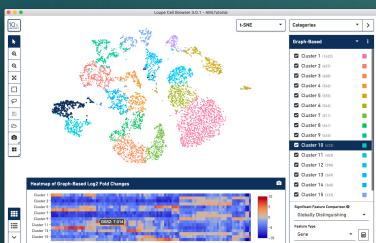
Visualization:

- IGV (.tdf)
- UCSC (bigWig) genome browser

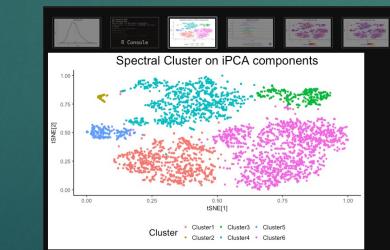
Public Pipelines

Single Cell Pipelines:

- Cell Ranger Pipeline
- Cell Ranger ATAC Pipeline
- 10x Genomics Pipeline
- Indrop Pipeline



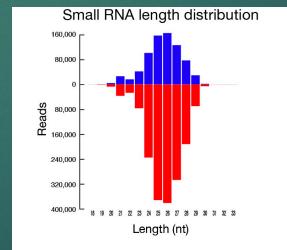
[Loupe Cell Browser](#)



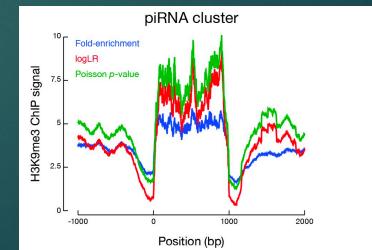
E.g. R packages: Seurat or SignallingSingleCell

piPipes Pipelines:

- smallRNA
- RNA-seq
- Degradome/RACE/CAGE
- ChIP-seq
- Genome-seq



<https://github.com/bowhan/piPipes>



Running Pipelines

The screenshot shows the Biocore platform interface. At the top, there is a navigation bar with tabs for "Pipelines", "Projects", and "Run Status". On the right side of the navigation bar, it says "ALPHA VERSION". Below the navigation bar, the main content area is titled "Biocore" and "Public Pipelines". There are six pipeline cards displayed:

- Single Cell-10X Genomics**: This pipeline maps 10X Genomics reads to selected genome (by using Tophat2, Hisat2 or STAR), measures digital gene expression (by using ESAT) and finally creates UMI distributions table for expression analysis. Steps: 1) UmiTools (<https://github.com/CGATOxford/UMI-tools>) is used to extract val
- Single Cell-Indrop**: This pipeline maps inDrop reads to selected genome (by using Tophat2), measures digital gene expression (by using ESAT) and finally creates UMI distributions table for expression analysis. Inputs: *Determined_fastq: output of Bcl2fastq software which has a structure as shown below. In order t
- RNA-seq Pipeline**: RNA-seq pipeline includes Quality Control, rRNA filtering, Genome Alignment using HISAT2, STAR and Tophat2, and estimating gene and isoform expression levels by RSEM. Steps: 1) For Quality Control, we use FastQC to create qc outputs. 2) Bowtie2 is used to filtered out rRNA reads. 3) RSEM is a pr
- RSEM Pipeline**: RSEM pipeline includes Quality Control, rRNA filtering, Genome Alignment using Bowtie, and estimating gene and isoform expression levels by RSEM. Steps: 1) For Quality Control, we use FastQC to create qc outputs. 2) Bowtie2 is used to filtered out rRNA reads. 3) RSEM is a program that aligns RN
- STAR Pipeline**: STAR pipeline includes Quality Control, rRNA filtering, Genome Alignment using STAR. Steps: 1) For Quality Control, we use FastQC to create qc outputs. 2) Bowtie2 is used to filter ed out rRNA reads. 3) STAR is a program that aligns RNA-Se q reads to a genome.
- HISAT2 pipeline**: HISAT2 pipeline includes Quality Control, rRNA filtering, Genome Alignment using HISAT2. Steps: 1) For Quality Control, we use FastQC to create qc outputs. 2) Bowtie2 is used to filtered out rRNA reads. 3) HISAT2 is a program that aligns R NA-Seq reads to a genome in order to identify exon-exon sp

At the bottom of the page, there is a navigation bar with page numbers 1, 2, 3, and a "»" symbol. On the right side, there is a vertical "Feedback" button.

Adding NCBI/GEO Files

The screenshot shows the Biocore DolphinNext software interface. The top navigation bar includes 'Biocore DolphinNext', 'Pipelines', 'Projects', 'Run Status', and user information ('Onur Yükselen', 'Online', 'Sign out'). The top right corner displays 'VERSION 1.1.7'. The main content area is titled 'Run Settings' under 'Run Description'.

Run Description: Enter run description here..

Work Directory (Full path): /home/oy28w/nextflowruns/test

Run Environment: Cluster(lsf) (Remote machine: oy28w@ghpcc06.umassrc.org)

Use Docker Image Use Singularity Image

Image Path: /project/umw_biocore/singularity/UMMS-Biocore-rna-seq-1.0.img

RunOptions: -bind /share -bind /nl --bind /project

Inputs:

Given Name	Select Items
~ User Inputs ~	
reads	Enter File
mate	pair
run_FeatureCounts_after_Tophat2	no
run_FeatureCounts_after_STAR	no
run_FeatureCounts_after_Hisat2	no
run_FeatureCounts_after_PSEM	

A vertical 'Feedback' button is located on the right side of the interface.

Adding Paired Files

Enter run description here..

Work Directory (Full path)
/home/oy28w/test

Run Environment ⓘ
Cluster(lsf) (Remote machine: oy28w@ghpcc06.umassrc.org)

Use Docker Image Use Singularity Image

Image Path ⓘ /project/umw_biocore/singularity/UMMS-Biocore-singularitysc-master-latest.sim

RunOptions ⓘ -bind /project --bind /share --bind /nl

Inputs

Given Name	Select Items
~ User Inputs ~	
reads	<input type="button" value="Enter File"/>
mate	<input type="button" value="Choose Value"/>
genome_build	<input type="button" value="Choose Value"/>
run_Tophat	no <input type="button" value=""/>
run_RSEM	yes <input type="button" value=""/> <input type="button" value=""/> <input type="button" value=""/>
run_HISAT2	no <input type="button" value=""/>
run_STAR	yes <input type="button" value=""/> <input type="button" value=""/>
run_IGV_TDF_Conversion	no <input type="button" value=""/>
run_RSeQC	no <input type="button" value=""/>

Feedback

Logs - Timeline

The screenshot shows the Biocore Run Generation interface. At the top, there's a navigation bar with tabs for Pipelines, Projects, and Run Status, and links for Amazon, Profiles, and a red ALPHA VERSION badge. On the left, a sidebar titled 'PROJECTS' lists various projects: Single Cell, Build Index, ltest, piPipes, Dylan_10X, Main Pipelines, kraken, modules, and test. The 'test' project is currently selected. The main content area is titled 'Biocore Run Generation' and shows details for a run named 'test run'. It includes fields for 'Run Description' (with placeholder 'Enter run description here...'), 'Work Directory (Full path)' (set to '/home/oy28w/test'), 'Run Environment' (set to 'Cluster(lsf) (Remote machine: oy28w@ghpcc06.umassrc.org)'), and options for 'Use Docker Image' (unchecked) and 'Use Singularity Image' (checked). The 'Image Path' is set to '/project/umw_biocore/singularity/UMMS-Biocore-singularitysc-master-latest.simg'. Under the 'RunOptions' section, the value is '--bind /project --bind /share --bind /nl'. At the bottom, there's a 'Feedback' button and a table for 'Inputs' with rows for 'reads' (selected item: 'test collection'), 'mate' (status: 'pair'), and 'genome_build'.

Biocore DolphinNext

Pipelines Projects Run Status

Amazon Profiles ALPHA VERSION

Onur Yükselen Online Sign out

Search...

PROJECTS

- Single Cell
- Build Index
- ltest
- piPipes
- Dylan_10X
- Main Pipelines
- kraken
- modules
- test

Biocore Run Generation

Run: test run | Project: test | Pipeline: RNA-seq Pipeline |

Created by onuryukseLEN on 2019-04-01 00:22:37 • Last edited on 2019-04-01 00:23:42

Ready to Run ...

Run Settings Advanced Workflow

Run Description

Enter run description here...

Work Directory (Full path)

/home/oy28w/test

Run Environment

Cluster(lsf) (Remote machine: oy28w@ghpcc06.umassrc.org)

Use Docker Image Use Singularity Image

Image Path /project/umw_biocore/singularity/UMMS-Biocore-singularitysc-master-latest.simg

RunOptions --bind /project --bind /share --bind /nl

Inputs

Given Name	Select Items
~ User Inputs ~	
reads	test collection
mate	pair
genome_build	

Feedback



Run Reports

Reports – Run Notes

The screenshot shows the Biocore DolphinNext software interface. The top navigation bar includes 'Pipelines', 'Projects', 'Run Status', and a user profile for 'Onur Yükselen'. The version 'VERSION 1.1.7' is also visible. The main content area is titled 'Biocore Run Generation' and displays a run named 'rna-seq test v3'. The pipeline used is 'Main Pipelines' (Pipeline: RNA-seq Pipeline (Rev 2)). The status is 'Completed'. A search bar and a 'Report' tab are present. The left sidebar lists various projects: Single Cell, Build Index, ltest, piPipes, Dylan_10X, Main Pipelines, kraken, modules, rsem, test, amazon test, and test_20190710_1. The right side shows a table of 'Run Report 18' items:

PROCESS	PUBLISHED DIRECTORY	VIEW FORMAT
Run Notes	Description	Markdown
Multiqc	Multiqc	HTML
Overall Summary	Summary	Table
Rsem Module	Rsem Summary	Table
Rsem Module	Rsem Summary	DE-Browser
Rsem Module	Rsem Rmarkdown	R-Markdown

A 'Feedback' button is located on the right edge of the report area. A URL at the bottom left is: <https://dolphinnext.umassmed.edu/index.php?np=3&id=1546#reportTab>.

Reports - MultiQC

Biocore DolphinNext

Pipelines Projects Run Status

Amazon Profiles ALPHA VERSION

Onur Yükselen Online Sign out

Search...

PROJECTS

- Single Cell
- Build Index
- Istest
- piPipes
- Dylan_10X
- Main Pipelines
- kraken
- modules
- test

Biocore Run Generation

Run: test run | Project: test | Pipeline: RNA-seq Pipeline | Completed

Created by onuryukseLEN on 2019-04-01 00:22:37 • Last edited on 2019-04-01 01:23:54

Run Report 4:

Run Report 4 c

PROCESS	PUBLISHED DIRECTORY	VIEW FORMAT
Multiqc	Multiqc	HTML
Alignment Summary	Summary	Table
Sequential Mapping Module	Sequential Mapping Summary	Table
Rsem Module	Rsem Summary	Table
Rsem Module	Rsem Summary	DE-Browser
Rsem Module	Rsem Rmarkdown	R-Markdown
Sequential Mapping Module	Sequential Mapping Counts	Table

Feedback

<https://dolphinnext.umassmed.edu/index.php?np=3&id=978#reportTab>

Reports – Alignment Summary

The screenshot shows the Biocore Run Generation interface. On the left, there is a sidebar with a user profile icon and the name "Onur Yükselen" (Online). Below it is a search bar and a "PROJECTS" section listing various pipeline configurations. The main content area is titled "Biocore Run Generation" and shows details for a run named "test run". The run was created by "onuryuksele" on 2019-04-01 00:22:37 and last edited on 2019-04-02 15:48:07. The status is "Completed". The interface includes tabs for "Run Settings", "Advanced", "Log", "Workflow", and "Report". The "Report" tab is selected, displaying a table of generated reports:

PROCESS	PUBLISHED DIRECTORY	VIEW FORMAT
Multiqc	Multiqc	HTML
Alignment Summary	Summary	Table
Sequential Mapping Module	Sequential Mapping Summary	Table
Rsem Module	Rsem Summary	Table
Rsem Module	Rsem Summary	DE-Browser
Rsem Module	Rsem Rmarkdown	R-Markdown
Sequential Mapping Module	Sequential Mapping Counts	Table

A "Run Report 4 c" button is visible on the right side of the report table. A "Feedback" button is located at the bottom right of the main content area.

Reports – Count Tables

Biocore Run Generation

Run: test run | Project: test | Pipeline: RNA-seq Pipeline | [Run Report](#) [Log](#) [Workflow](#) [Report](#) [Completed](#) ...

Created by onuryukselen on 2019-04-01 00:22:37 • Last edited on 2019-04-01 01:23:54

Run Report 4:

Run Report 4 c

PROCESS	PUBLISHED DIRECTORY	VIEW FORMAT
Multiqc	Multiqc	HTML
Alignment Summary	Summary	Table
Sequential Mapping Module	Sequential Mapping Summary	Table
Rsem Module	Rsem Summary	Table
Rsem Module	Rsem Summary	DE-Browser
Rsem Module	Rsem Rmarkdown	R-Markdown
Sequential Mapping Module	Sequential Mapping Counts	Table

Feedback

The screenshot shows the Biocore Run Generation interface. At the top, there's a navigation bar with a user icon, 'Online' status, and 'Sign out' option. Below it is a search bar. The main area has tabs for 'Run Settings', 'Advanced', 'Log', 'Workflow', and 'Report'. The 'Report' tab is selected. A message indicates the run is 'Completed'. On the left, a sidebar lists 'PROJECTS' with items like 'Single Cell', 'Build Index', 'ltest', 'piPipes', 'Dylan_10X', 'Main Pipelines', 'kraken', 'modules', and 'test'. The main content area displays a table of generated reports. The table has columns for 'PROCESS', 'PUBLISHED DIRECTORY', and 'VIEW FORMAT'. The 'VIEW FORMAT' column includes icons for HTML, Table, DE-Browser, and R-Markdown. The table rows correspond to the items listed in the sidebar.

Reports – DEBrowser

The screenshot shows the DEBrowser interface for a completed RNA-seq Pipeline. The top navigation bar includes 'Run: test run', 'Project: test', 'Pipeline: RNA-seq Pipeline', and status indicators for 'Completed' and '...'.

The left sidebar lists 'PROJECTS' with items like Single Cell, Build Index, ltest, piPipes, Dylan_10X, Main Pipelines, kraken, modules, and test. The 'Report' tab is selected in the top menu.

The main content area displays 'Run Report 4:' with a 'Run Report 4 c' button. Below this is a table showing the report structure:

PROCESS	PUBLISHED DIRECTORY	VIEW FORMAT
Multiqc	Multiqc	HTML
Alignment Summary	Summary	Table
Sequential Mapping Module	Sequential Mapping Summary	Table
Rsem Module	Rsem Summary	Table
Rsem Module	Rsem Summary	DE-Browser
Rsem Module	Rsem Rmarkdown	R-Markdown
Sequential Mapping Module	Sequential Mapping Counts	Table

At the bottom, there's a file viewer for 'genes_expressi...tsv' and buttons for 'Show All' and 'Feedback'.

Reports – RMarkdown

The screenshot shows the Biocore Run Generation interface. At the top, there is a navigation bar with a user profile (Alper Kucukural, Online), sign-out link, search bar, and a red sidebar.

The main area displays a "Run Report 4" for a completed run. The report is titled "Biocore Run Generation". It includes tabs for Run Settings, Advanced, Log, Workflow, and Report (which is selected). Below the tabs, it shows the creation details: "Created by nephantes on 2019-02-24 21:09:26 • Last edited on 2019-02-26 10:18:39".

The report content is organized into three columns: PROCESS, PUBLISHED DIRECTORY, and VIEW FORMAT. The rows are:

PROCESS	PUBLISHED DIRECTORY	VIEW FORMAT
Alignment Summary	Summary	Table
Sequential Mapping Module	Sequential Mapping Summary	Table
Rsem Module	Rsem Summary	Table
Rsem Module	Rsem Summary	DE-Browser
Rsem Module	Rsem Rmarkdown	R-Markdown

A "Feedback" button is located on the right side of the report area.

DolphinNext Paper and Repository

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

SOFTWARE **Open Access**

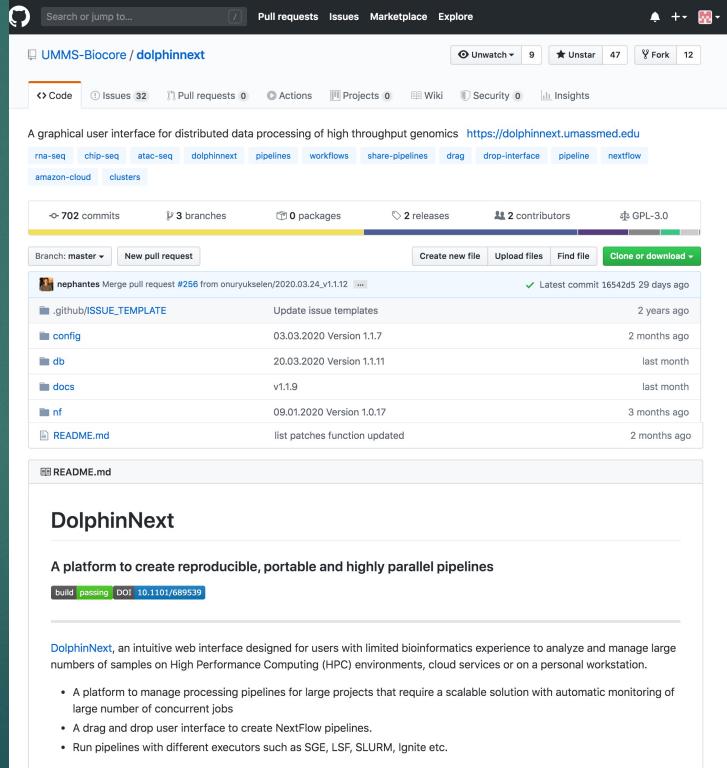
DolphinNext: a distributed data processing platform for high throughput genomics

Onur Yukseken¹, Osman Turkyilmaz², Ahmet Rasit Ozturk², 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 of the pipeline. The main aim of this study was to develop a distributed data processing platform.

▶ github.com/UMMS-Biocore/dolphinnext



The screenshot shows the GitHub repository page for 'UMMS-Biocore / dolphinnext'. The repository has 9 stars, 47 forks, and 12 issues. It contains 702 commits, 3 branches, 0 packages, 2 releases, and 2 contributors. The code is licensed under GPL-3.0. The README.md file describes DolphinNext as a platform to create reproducible, portable and highly parallel pipelines. It includes a build status badge (passing), a DOI link (10.1101/689539), and a list of features: managing processing pipelines for large projects, creating NextFlow pipelines via drag-and-drop, and running pipelines with different executors like SGE, LSF, SLURM, Ignite etc.