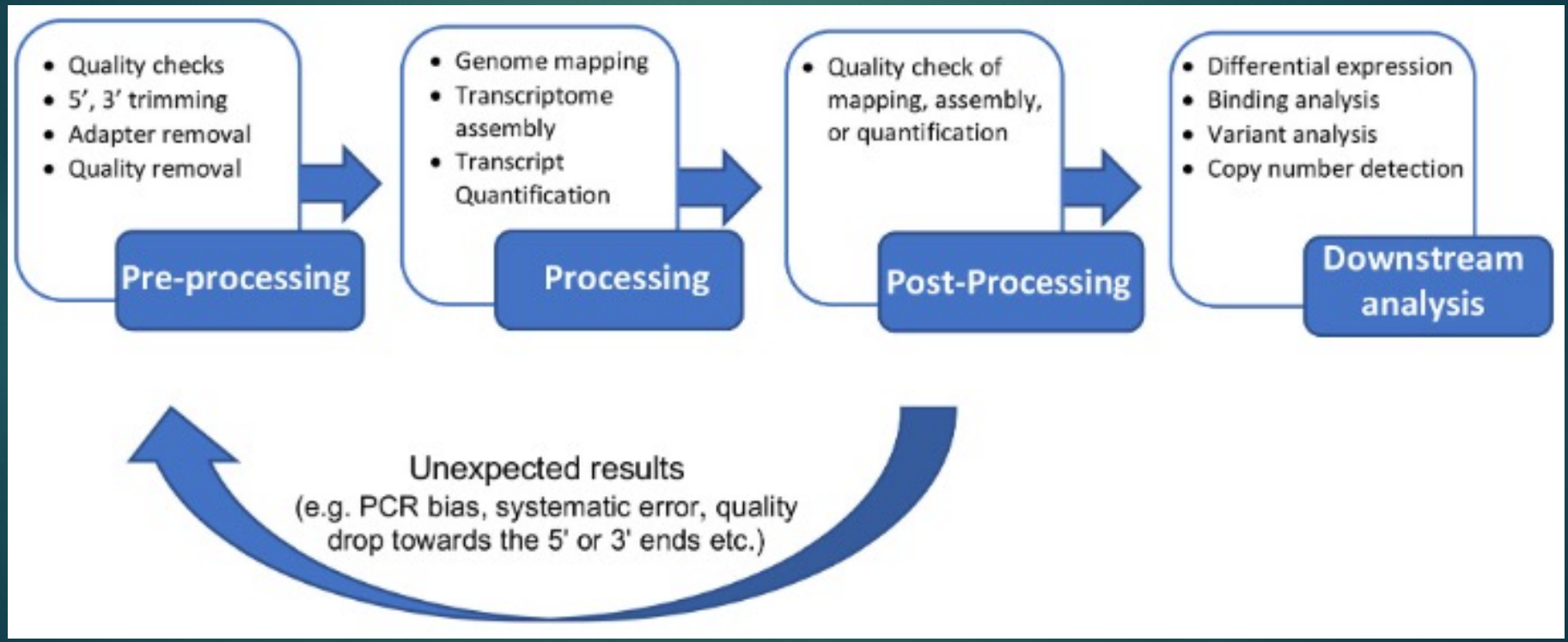
# DolphinNext: A graphical user interface for reproducible pipelines

Onur Yukselen, Artur Manukyan,
Alper Kucukural
May 6, 2021

nextflow

University of Massachusetts Medical School

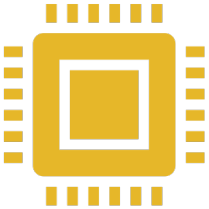
# Sequence analysis: mostly reusable steps



# Large projects require

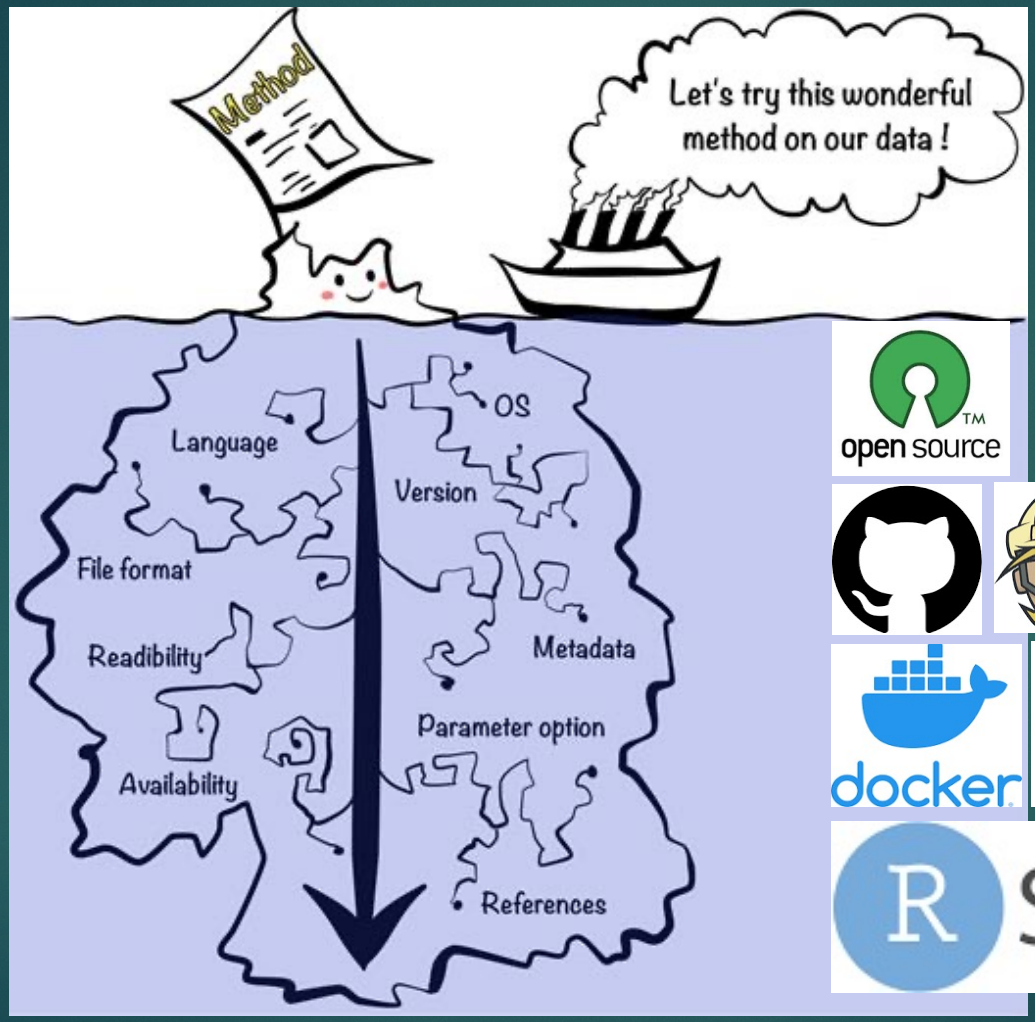
## Capture sample metadata

•Identify and correct batches
•Troubleshoot
•Drive processing



## Pipelines should be

- Highly parallel
- Self monitoring → restart on hardware failures
  - Easily restartable from mid points
- Built from reusable components
- Versioned
- Handle large number of samples




open source

docker

nextflow

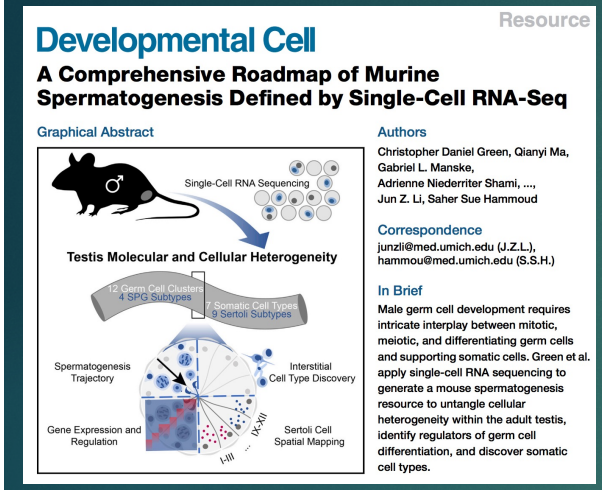
R Studio

jupyter

Kim *et. al, Jul 2018, Giga Science*

# Replicating a Method

## Dependency issues



Resource

**Developmental Cell**

**A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell RNA-Seq**

Graphical Abstract




Authors

Christopher Daniel Green, Qianyi Ma, Gabriel L. Manske, Adrienne Niederriter Shami, ..., Jun Z. Li, Saher Sue Hammoud

Correspondence

junzli@med.umich.edu (J.Z.L.), hammou@med.umich.edu (S.S.H.)



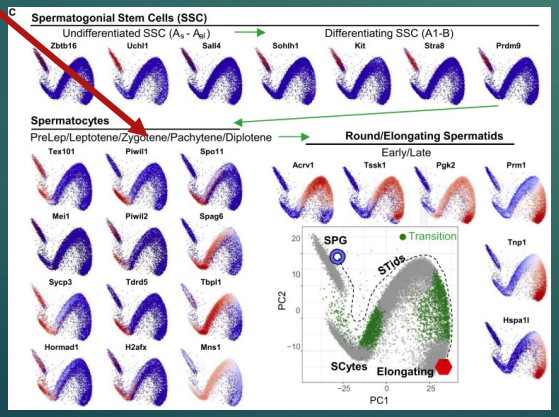
In Brief

Male germ cell development requires intricate interplay between mitotic, meiotic, and differentiating germ cells and supporting somatic cells. Green et al. apply single-cell RNA sequencing to generate a mouse spermatogenesis resource to untangle cellular heterogeneity within the adult testis, identify regulators of germ cell differentiation, and discover somatic cell types.


```
checking for sys/types.h... yes
checking for sys/stat.h... yes
checking for stdlib.h... yes
checking for string.h... yes
checking for memory.h... yes
checking for strings.h... yes
checking for inttypes.h... yes
checking for stdint.h... yes
checking for unistd.h... yes
checking for h5cc... no
checking for h5pcc... no
checking for HDF5 libraries... no
configure: WARNING:
Unable to locate HDF5 compilation helper scripts 'h5cc' or 'h5pcc'.
Please specify --with-hdf5=<LOCATION> as the full path to h5cc or h5pcc.
HDF5 support is being disabled (equivalent to --with-hdf5=no).

configure: error: hdf5 does not seem to be installed on your platform.
  Please install the hdf5 library.
  The required HDF5 library files can be installed as follows:
      - Debian-based (e.g. Debian >= 8.0, Ubuntu >= 15.04): 'sudo apt-get install libhdf5
-dev'
      - Old Debian-based (e.g Debian < 8.0, Ubuntu < 15.04): Install from source (see I
NSTALL)
      - OS X using Homebrew: 'brew install homebrew/science/hdf5'
      - RPM-based (e.g Fedora): 'sudo yum install hdf5-devel'
ERROR: configuration failed for package 'hdf5r'
* removing '/home/dell/R/x86_64-pc-linux-gnu-library/3.5/hdf5r'
Warning in install.packages :
  installation of package 'hdf5r' had non-zero exit status
ERROR: dependency 'hdf5r' is not available for package 'Seurat'
* removing '/home/dell/R/x86_64-pc-linux-gnu-library/3.5/Seurat'
Warning in install.packages :
  installation of package 'Seurat' had non-zero exit status

The downloaded source packages are in
        '/tmp/RtmpryfH6z/downloaded_packages'
```

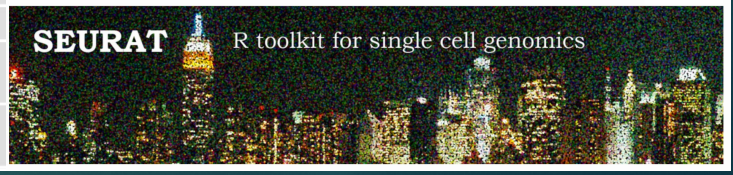
## Hardcoded Scripts

```
### R script for pseudotemporal ordering of germ cells using Monocle in Nov 2017 by Qianyi
### Related to Figure S2C right panel: germ cells (N=20,646) with >1k detected genes

### Read Data
home="/scratch/junzli_flux/qzm/Dropseq_analysis/"
file="figJul2017_10CellTypes_MouseAdultST24mergedclusters/monocle_"
load(file =paste0(home,"data_DGE/MouseAdultST24genesUMI20cell15.Robj"))
dge20=dge    # Gene Filter 1 (N=24482): genes >20totalUMIsOverAllCells & >15Cells +31 genes
table(dge@data.info[,52:53])
# Cluster31SeriationOLO_GeneFilter1 is the reordered cluster ID 1-31
# Cluster 8-31 are germ cells
```





| Software and Algorithms |
|---|
| Drop-seq_tools (v1.12) |
| Picard Tools (v2.6.0) |
| Samtools (v1.2) |
| STAR (v2.5.2b) |
| R (v3.3.3) |
| Seurat (v1.4.0.3) |
| Seriation (v1.2-2) |
| Monocle 2 |





**R Markdown Code for Reproducing Clustering Analysis**

As an accompaniment to this paper, we provide an R markdown file that describes step-by-step procedures, including the loading and preprocessing of the Drop-seq digital expression matrix, PCA, Louvain-Jaccard clustering, data visualization, ordering by seriation, and differential expression tests. The R commands are provided at https://github.com/qianqianshao/Drop-seq_ST.

Pipeline: *RSEM Pipeline*

Revision: 1 wh

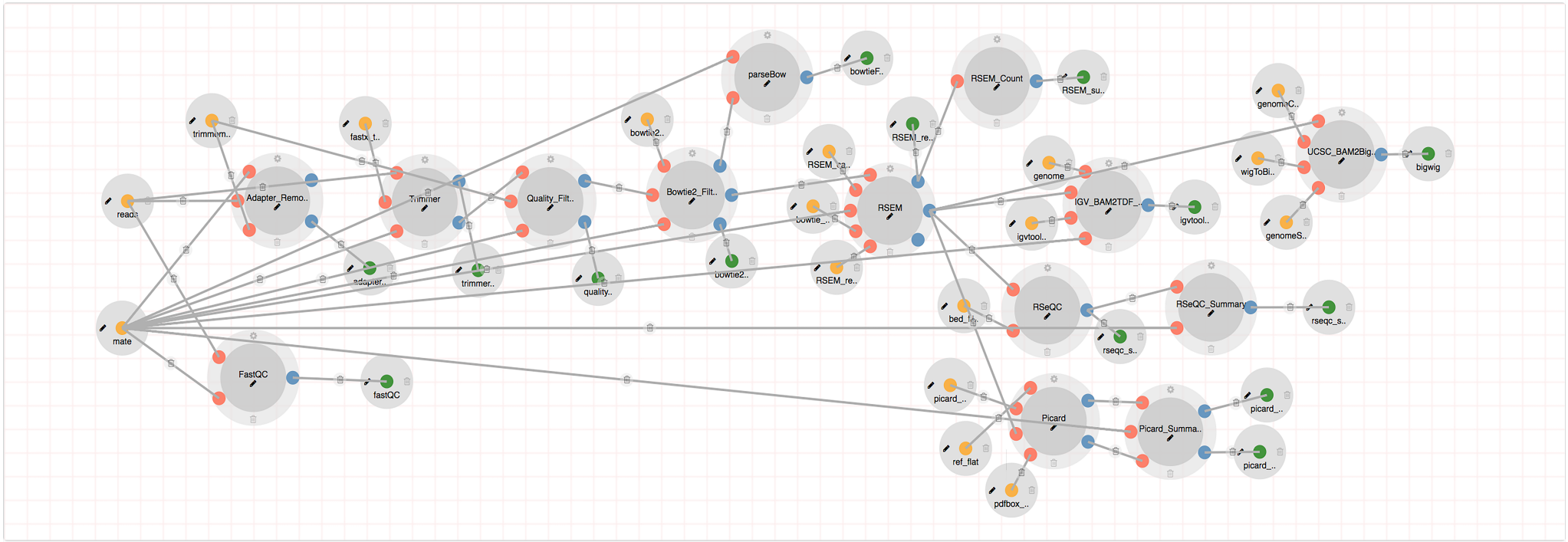
Run

Created by onuryukselen on 2018-09-10 17:37:33 • Last edited on 2018-10-11 15:24:00

## Description

RSEM pipeline includes Quality Control, rRNA filtering, Genome Alignment using Bowtie, and estimating gene and isoform expression levels by RSEM.

Steps:



# Best practices for reproducible data processing

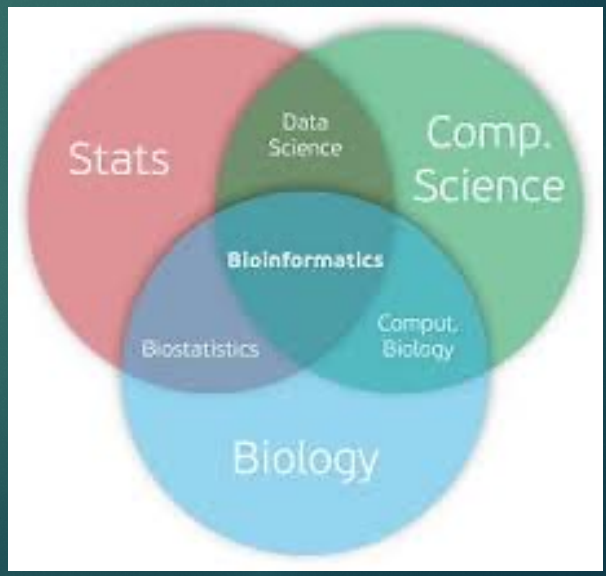
- For every result, keep
  how it was produced
- Avoid manual data
- Archive the exact ve
  programs used
- Version control all cu
- Provide (public) acc
  runs, and results in FA


https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx

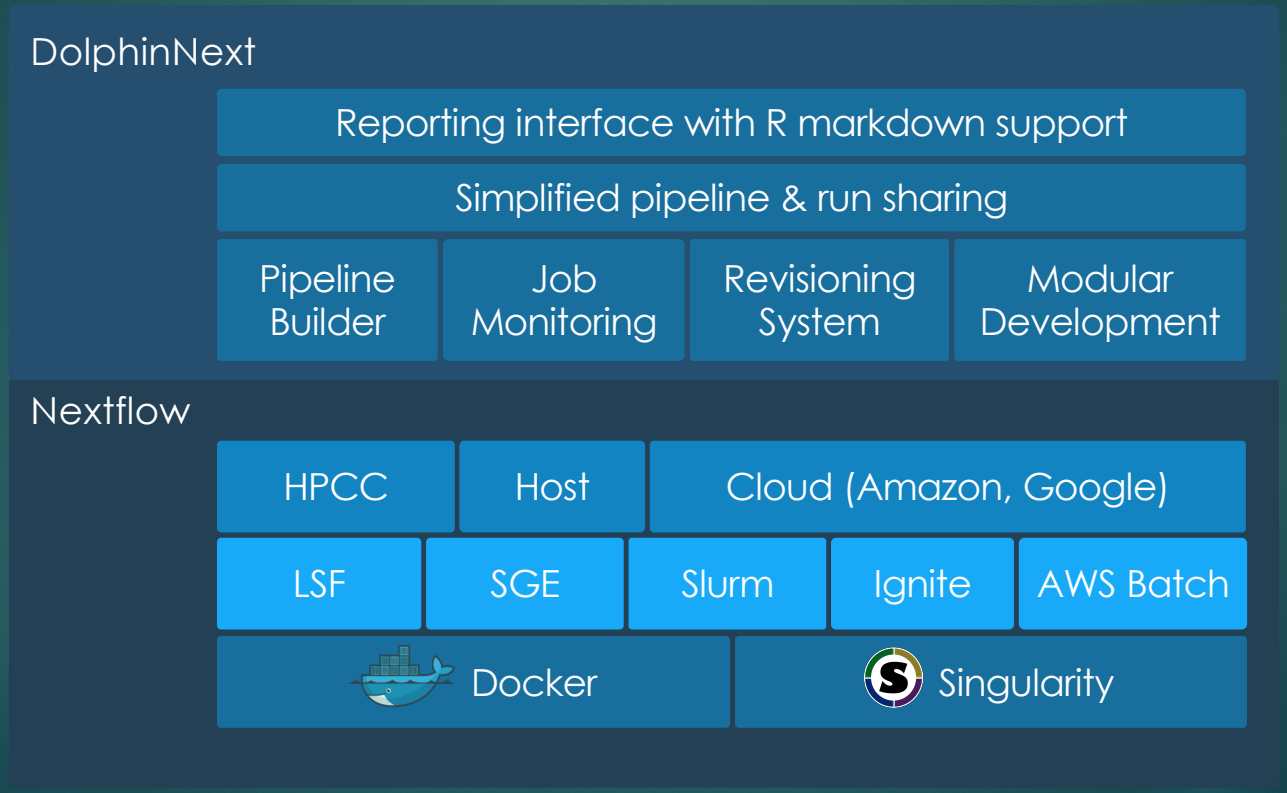
Kulkarni *et. al*, Oct 2019, BMC Bioinformatics

- DolphinNext is an easy-to-use web platform for creating, deploying, and executing complex pipelines for high throughput data processing and analysis.

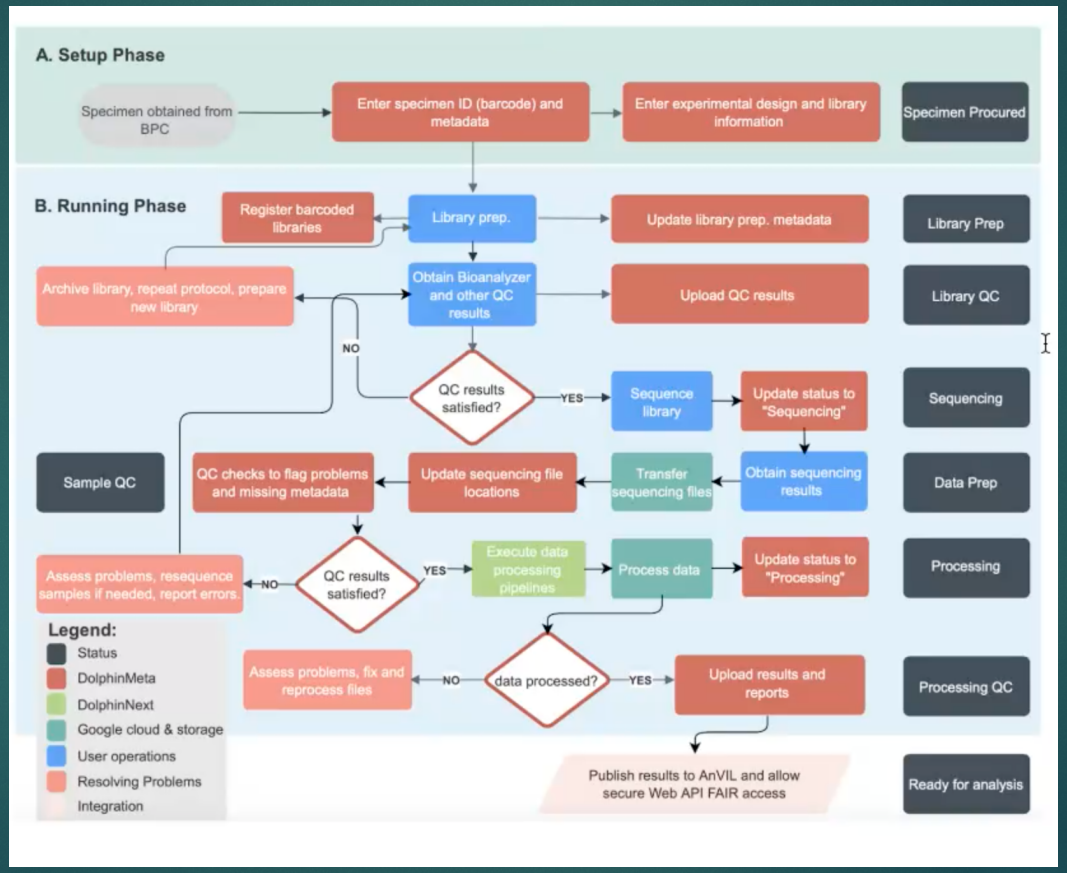
For all skill levels and areas to perform bioinformatics analysis



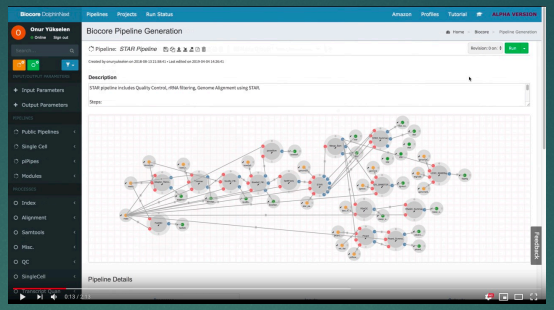
# Structure


DolphinNext
Reporting interface with R markdown support
Simplified pipeline & run sharing
Pipeline Builder
Job Monitoring
Revisioning System
Modular Development
Nextflow
HPCC
Host
Cloud (Amazon, Google)
LSF
SGE
Slurm
Ignite
AWS Batch
Docker
Singularity

# DolphinLIMS

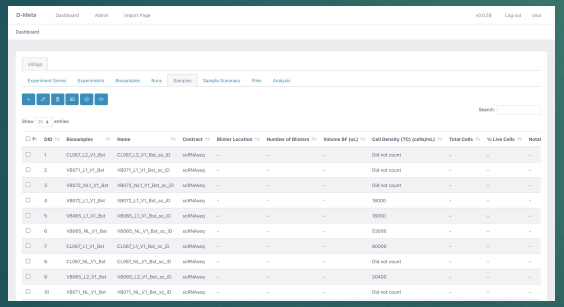



# D-meta and D-portal

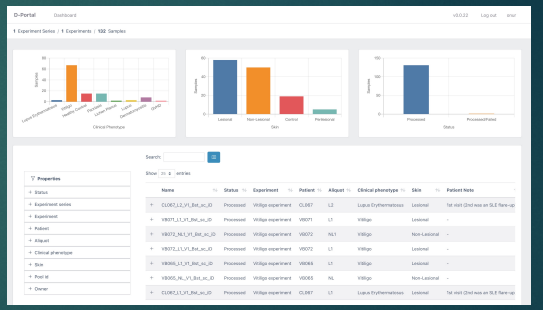



# Dnext

Auth Server




# Dmeta



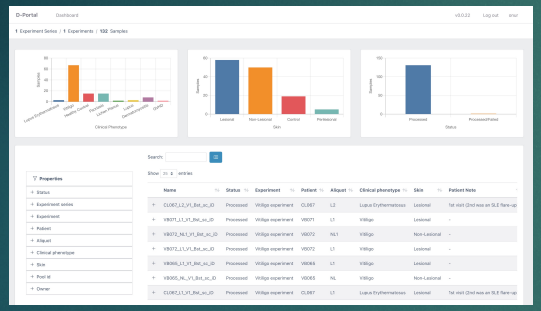

# Dportal

# Use Case

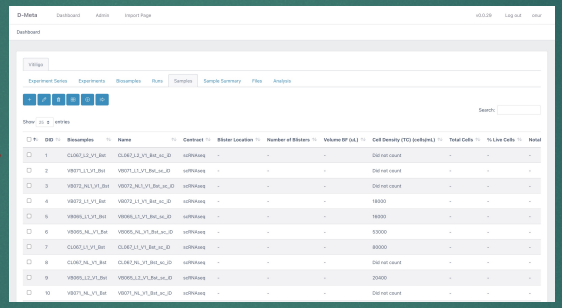
| total reads | valid reads | %valid reads | unique alignment | alignment Rate(%) | UMI | # Cell | UMI / Cell |
|---|---|---|---|---|---|---|---|
| 57963319 | 53803911 | 92.824 | 23527725 | 43.72865199 | 91590 | 9159 | 10 |
| 19753628 | 18251929 | 92.398 | 7505518 | 41.12177951 | 32803 | 612 | 53.6 |
| 98329966 | 90939492 | 92.484 | 37666497 | 41.41929559 | 157886 | 21336 | 7.4 |
| 49282431 | 45654122 | 92.638 | 19789130 | 43.34576843 | 73380 | 6115 | 12 |

Portal update

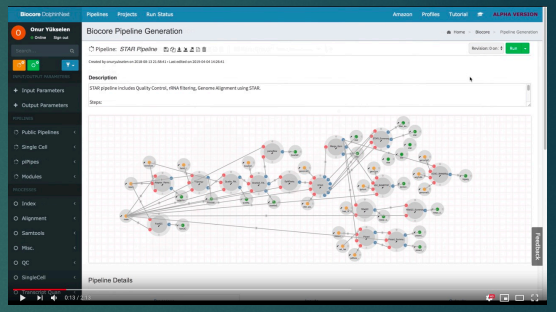
Metadata update



**D**portal



**D**meta-skin



**D**next

- Users enter metadata

| B | E | | G |
|---|---|---|---|
| **Patient** | **Skin** | **S3Bucket** | **FastqName** |
| **CB17** | **Healthy** | **s3://biocorebackup/dolphin_import** | **CB17_H2.fastq.gz** |
| **CB19** | **Healthy** | **s3://biocorebackup/dolphin_import** | **CB19_HA.fastq.gz** |
| **CB19** | **Healthy** | **s3://biocorebackup/dolphin_import** | **CB19_H2.fastq.gz** |
| **CB20** | **Healthy** | **s3://biocorebackup/dolphin_import** | **CB20_H1.fastq.gz** |
| **CB20** | **Healthy** | **s3://biocorebackup/dolphin_import** | **CB20_H2.fastq.gz** |

Automatic run submission

# Dmeta

- Ontology based, metadata tracking system

Collecting Sample | Sequencing | Data Processing | Analysis

- FAIR standards:

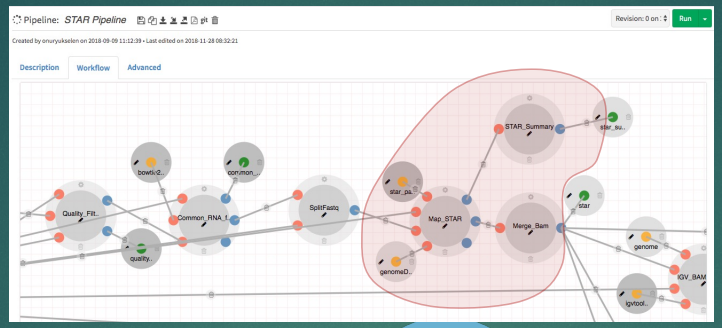
| Findable | Accessible | Interoperable | Reusable |
|---|---|---|---|
| Searchable metadata accurately describes dataset | - Standard APIs<br>- Authentication<br>- Permission control | Integration with Ontology servers | Data and metadata are well described to be reused |

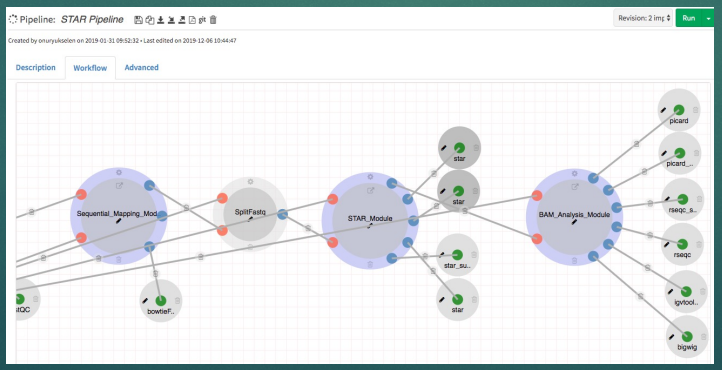
# Dnext

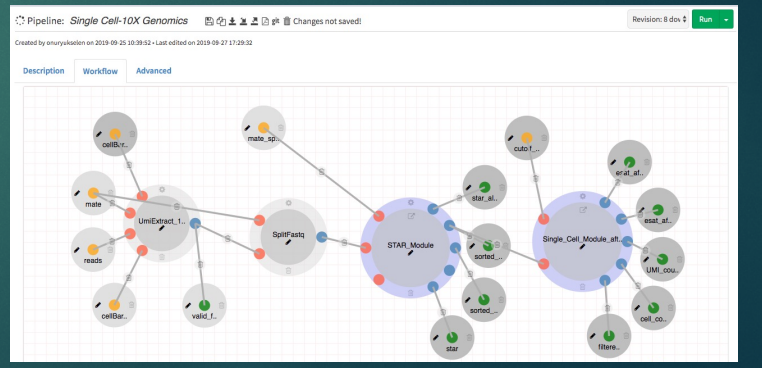
Processes

Modules

Reusability



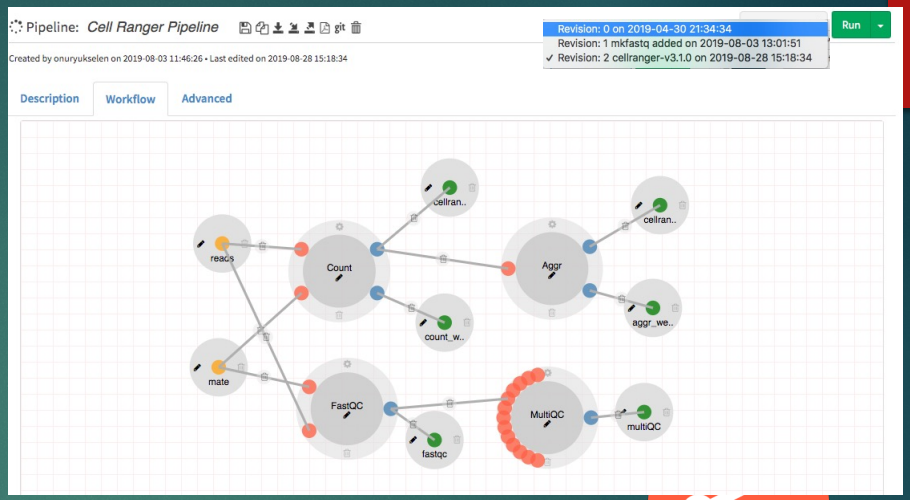



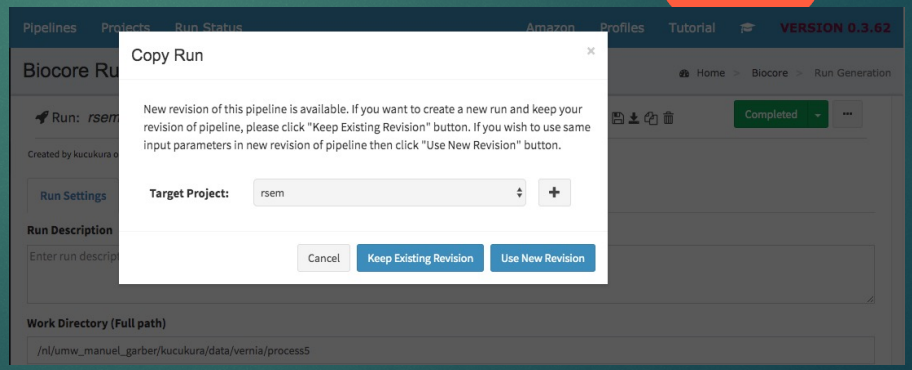



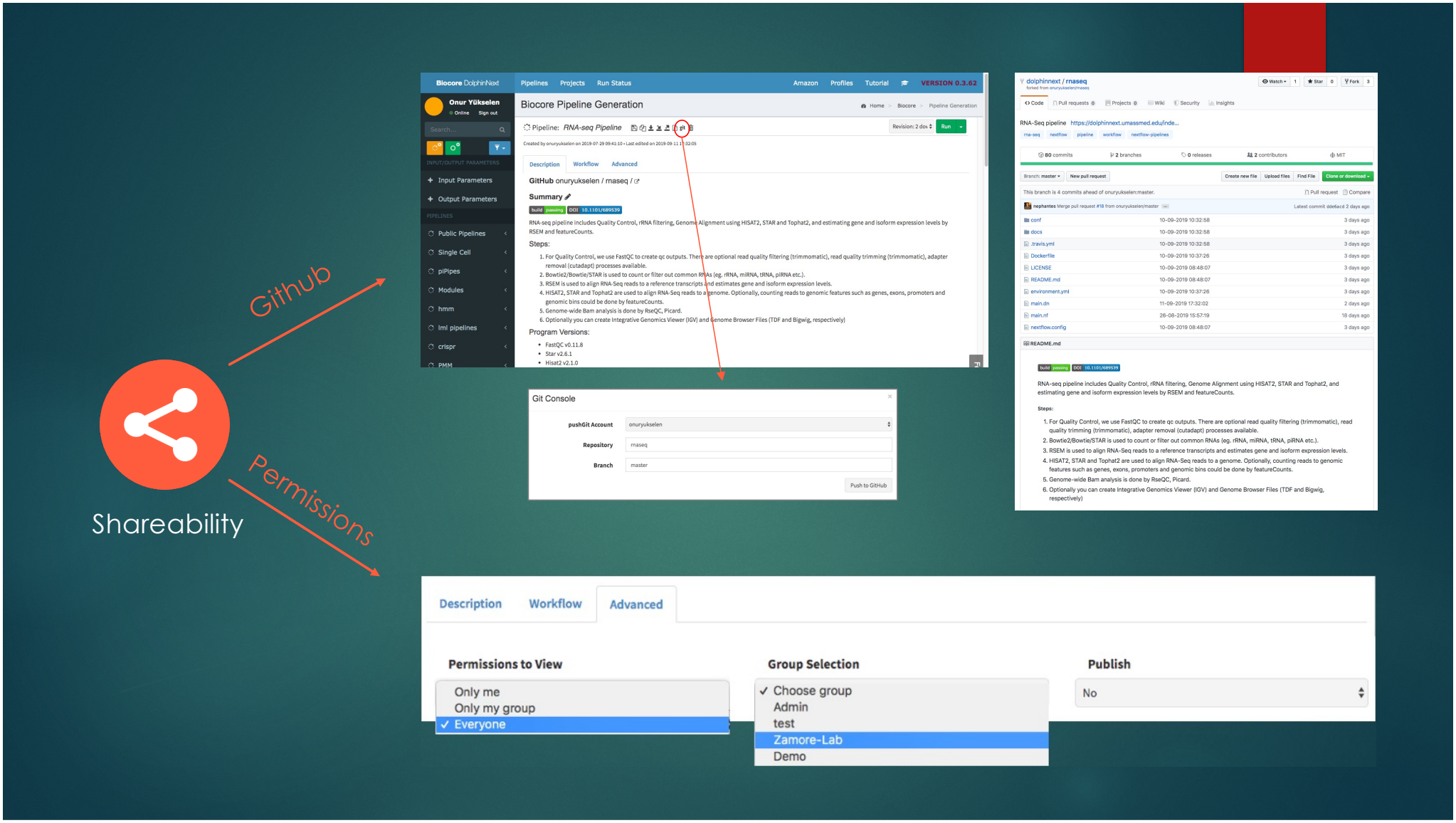
Versioning

Copy Run

Reproducibility

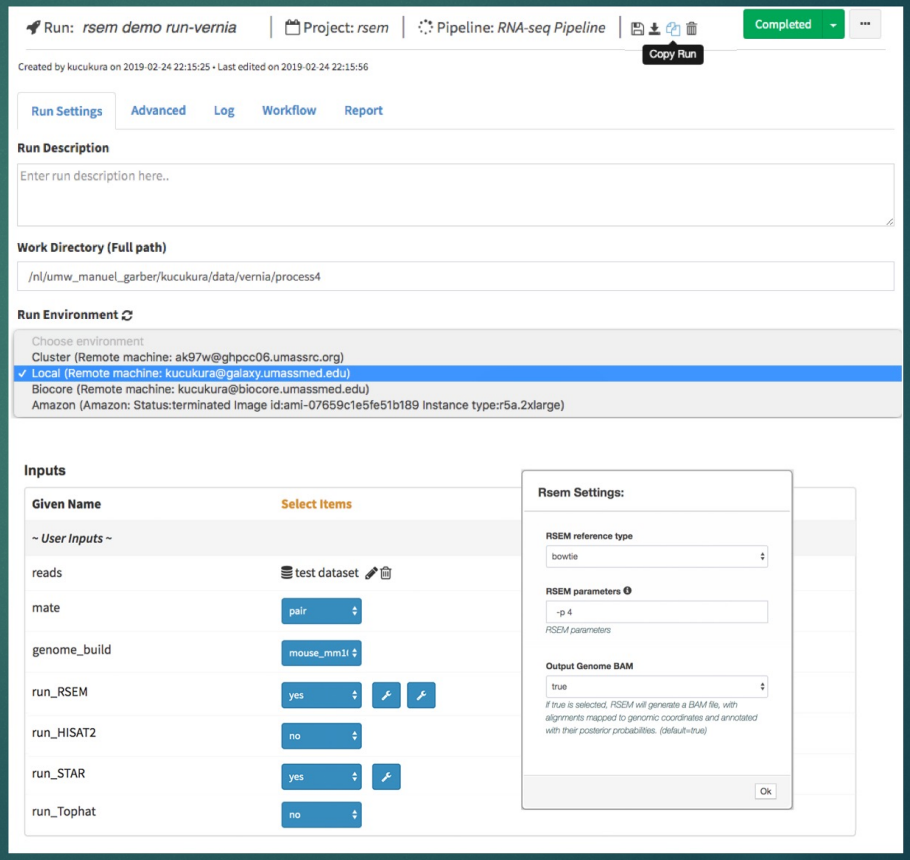



Github
Permissions
Shareability
Biocore DolphinNext
Pipelines
Projects
Run Status
Amazon
Profiles
Tutorial
VERSION 0.3.62
Onur Yükselen
Online
Sign out
Biocore Pipeline Generation
Pipeline: RNA-seq Pipeline
Revision: 2 dos
Run
Description
Workflow
Advanced
Input Parameters
Output Parameters
Public Pipelines
Single Cell
piPipes
Modules
hmm
lml pipelines
crispr
PMM
GitHub onuryukselen / rnaseq /
Summary
RNA-seq pipeline includes Quality Control, rRNA filtering, Genome Alignment using HISAT2, STAR and Tophat2, and estimating gene and isoform expression levels by RSEM and featureCounts.
Steps:
1. For Quality Control, we use FastQC to create qc outputs. There are optional read quality filtering (trimmomatic), read quality trimming (trimmomatic), adapter removal (cutadapt) processes available.
2. Bowtie2/Bowtie/STAR is used to count or filter out common RNAs (eg. rRNA, miRNA, tRNA, piRNA etc.).
3. RSEM is used to align RNA-Seq reads to a reference transcripts and estimates gene and isoform expression levels.
4. HISAT2, STAR and Tophat2 are used to align RNA-Seq reads to a genome. Optionally, counting reads to genomic features such as genes, exons, promoters and genomic bins could be done by featureCounts.
5. Genome-wide Bam analysis is done by RseQC, Picard.
6. Optionally you can create Integrative Genomics Viewer (IGV) and Genome Browser Files (TDF and Bigwig, respectively)
Program Versions:
FastQC v0.11.8
Star v2.6.1
Hisat2 v2.1.0
Git Console
pushGit Account
onuryukselen
Repository
rnaseq
Branch
master
Push to GitHub
dolphinnext / rnaseq
RNA-Seq pipeline https://dolphinnext.umassmed.edu/inde...
80 commits
2 branches
0 releases
2 contributors
MIT
README.md
Description
Workflow
Advanced
Permissions to View
Only me
Only my group
Everyone
Group Selection
Choose group
Admin
test
Zamore-Lab
Demo
Publish
No

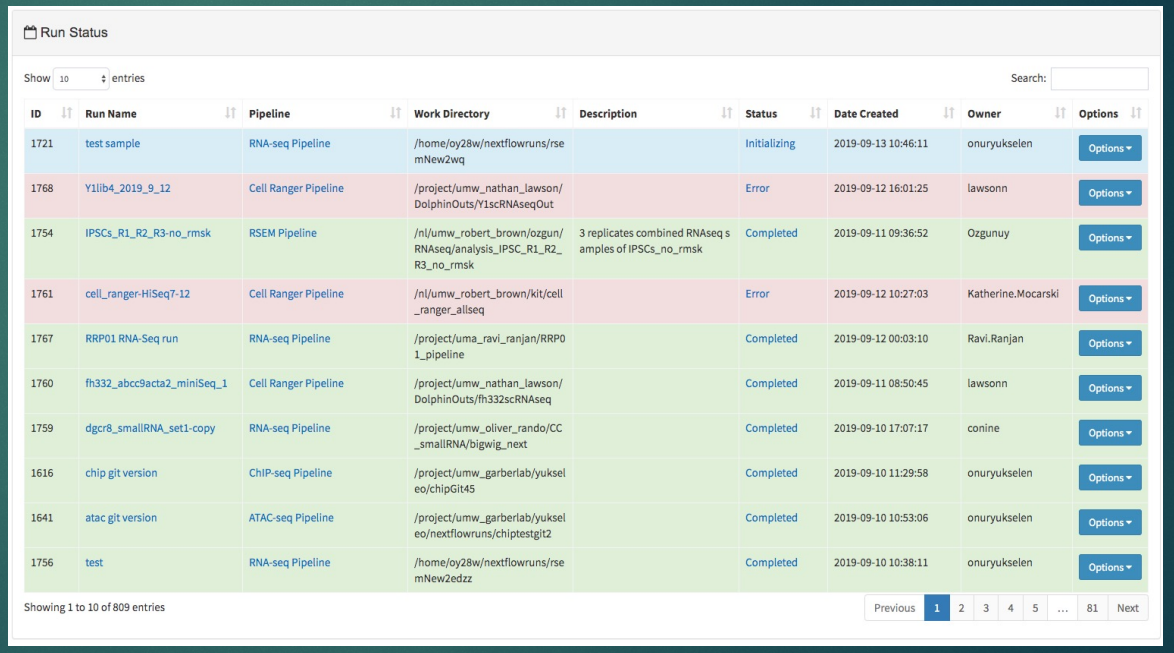
# Increase the team efficiency with

- Easy execution
- Easy monitoring
- Easy reporting



# Increase the team efficiency with

- Easy execution
- Easy monitoring
- Easy reporting



Run Status

Show 10 entries Search:

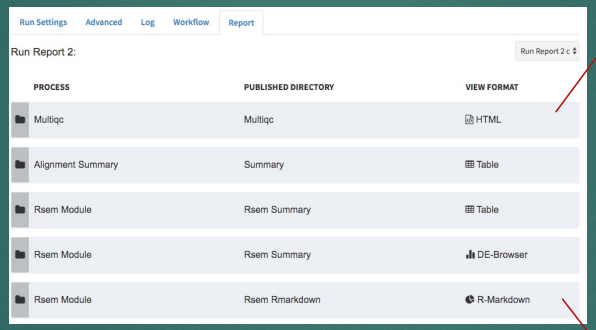
| ID | Run Name | Pipeline | Work Directory | Description | Status | Date Created | Owner | Options |
|---|---|---|---|---|---|---|---|---|
| 1721 | test sample | RNA-seq Pipeline | /home/oy28w/nextflowruns/rseqmNew2wq | | Initializing | 2019-09-13 10:46:11 | onuryukselen | Options |
| 1768 | Y1lib4_2019_9_12 | Cell Ranger Pipeline | /project/umw_nathan_lawson/DolphinOuts/Y1scRNAseqOut | | Error | 2019-09-12 16:01:25 | lawsonn | Options |
| 1754 | IPSCs_R1_R2_R3-no_rmsk | RSEM Pipeline | /nl/umw_robert_brown/ozgun/RNAseq/analysis_IPSC_R1_R2_R3_no_rmsk | 3 replicates combined RNAseq samples of IPSCs_no_rmsk | Completed | 2019-09-11 09:36:52 | Ozgunuy | Options |
| 1761 | cell_ranger-HiSeq7-12 | Cell Ranger Pipeline | /nl/umw_robert_brown/kit/cell_ranger_allseq | | Error | 2019-09-12 10:27:03 | Katherine.Mocarski | Options |
| 1767 | RRP01 RNA-Seq run | RNA-seq Pipeline | /project/uma_ravi_ranjan/RRP01_pipeline | | Completed | 2019-09-12 00:03:10 | Ravi.Ranjan | Options |
| 1760 | fh332_abcc9acta2_miniSeq_1 | Cell Ranger Pipeline | /project/umw_nathan_lawson/DolphinOuts/fh332scRNAseq | | Completed | 2019-09-11 08:50:45 | lawsonn | Options |
| 1759 | dgcr8_smallRNA_set1-copy | RNA-seq Pipeline | /project/umw_oliver_rando/CC_smallRNA/bigwig_next | | Completed | 2019-09-10 17:07:17 | conine | Options |
| 1616 | chip git version | ChIP-seq Pipeline | /project/umw_garberlab/yukseleo/chipGit45 | | Completed | 2019-09-10 11:29:58 | onuryukselen | Options |
| 1641 | atac git version | ATAC-seq Pipeline | /project/umw_garberlab/yukseleo/nextflowruns/chiptestgit2 | | Completed | 2019-09-10 10:53:06 | onuryukselen | Options |
| 1756 | test | RNA-seq Pipeline | /home/oy28w/nextflowruns/rseqmNew2edzz | | Completed | 2019-09-10 10:38:11 | onuryukselen | Options |

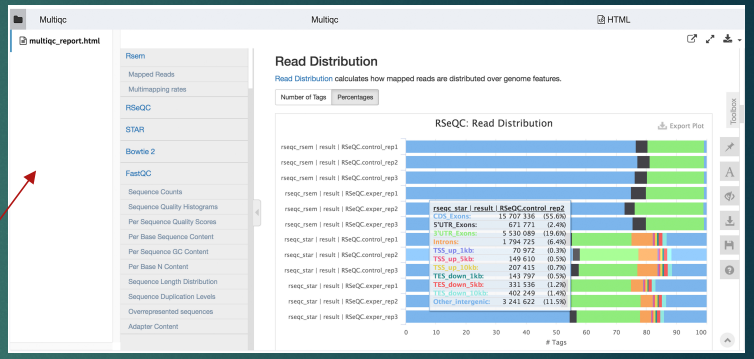
Showing 1 to 10 of 809 entries

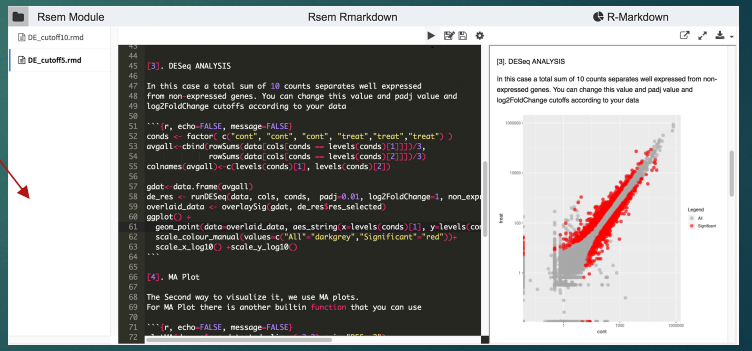
Previous 1 2 3 4 5 … 81 Next

# Increase the team efficiency with

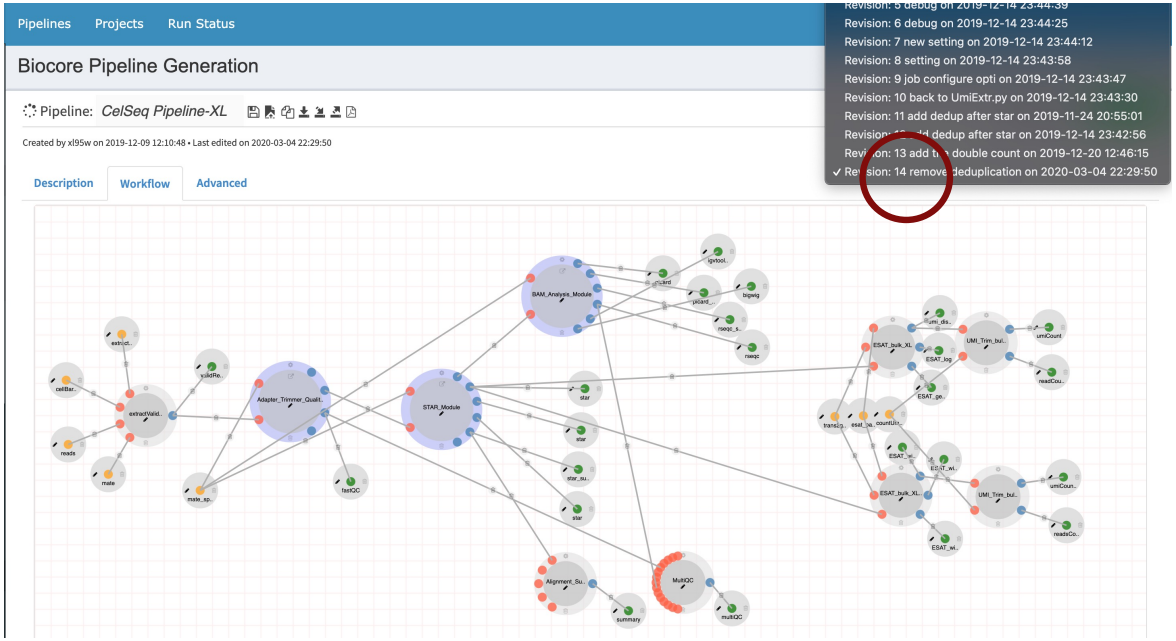
- Easy execution
- Easy monitoring
- Easy reporting



# DolphinNext Use Cases

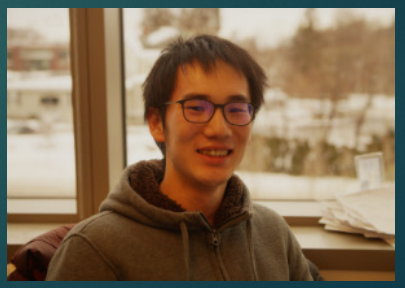



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 4144 | metabolic_lipid_met4_lib1_and_lib6 | CelSeq Pipeline-XL (Rev 14) | /project/umw_marian_walhout/data/xuhangLi/XL_HZ_CelSeq_02082021/dolphinNext/met4_lib1_and_lib6/ | processing BGI-sequenced data of the metabolic RNAi library. Three bio replicates were pooled in the same sequencing li.. | Completed | 2021-02-08 11:48:36 | xl95w | Options |
| 4026 | metabolic_lipid_met3_lib5_and_lib6 | CelSeq Pipeline-XL (Rev 14) | /project/umw_marian_walhout/data/xuhangLi/XL_HZ_CelSeq_12112020/dolphinNext/met3_lib5_and_lib6/ | processing BGI-sequenced data of the metabolic RNAi library. Three bio replicates were pooled in the same sequencing li.. | Completed | 2021-01-07 11:12:22 | xl95w | Options |
| 4025 | metabolic_lipid_met3_lib3_and_lib4 | CelSeq Pipeline-XL (Rev 14) | /project/umw_marian_walhout/data/xuhangLi/XL_HZ_CelSeq_12112020/dolphinNext/met3_lib3_and_lib4/ | processing BGI-sequenced data of the metabolic RNAi library. Three bio replicates were pooled in the same sequencing li.. | Completed | 2021-01-06 21:15:47 | xl95w | Options |

Showing 1 to 10 of 140 entries (filtered from 2,864 total entries)

Previous 1 2 3 4 5 … 14 Next

# Walhout Lab use case



**Xuhang “Hang” Li**

# Reprocessed all data to use updated pipelines

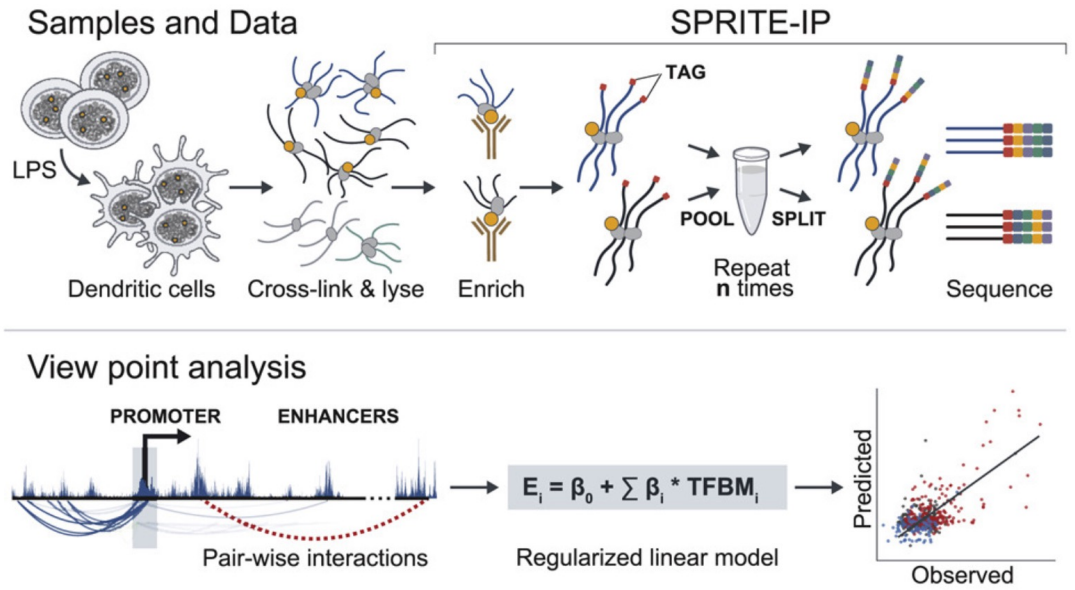
- Updated genome versions
  - mm9 to mm10
  - hg19 to hg38
- 300+ RNA-seq datasets to have same parameters and updated pipelines
- lncRNA quantification with Fitzgerald Lab
  - Easy updating with newer annotations

# Training and Sharing

- Shared the runs and pipelines with
  - Summer interns (undergrads)
  - New grad students
  - Research associates
  - Wet-lab scientists

# Creating new pipelines using pre-built modules

- HiC
- SPRITE



# DolphinNext Demo and Tutorials

https://github.com/dolphinnext/dolphinnext-tutorial

# Thanks to

- **Manuel Garber**
- **Onur Yukselen**
- **Artur Manukyan**
- **Pranitha Vangala**
- Alan Derr
- Elisa Donnard

All Garber Lab Members

- Craig Mello
- Ahmet Ozturk
- Osman Turkyilmaz

- Ruijia Wang
- Xuhang Li

- Nathan Lawson
- Julie Zhu