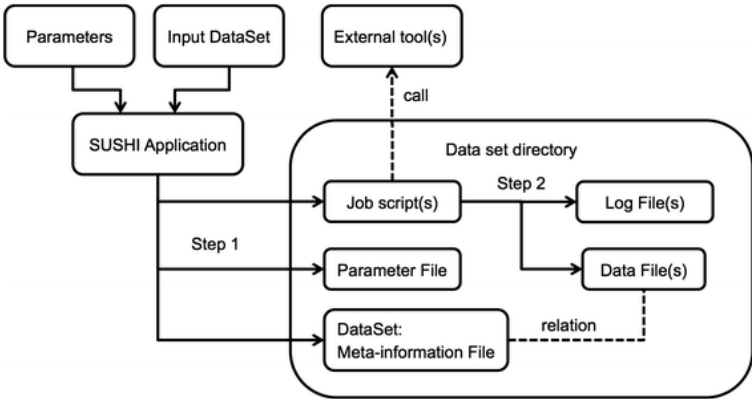


SUSHI can Support yoU by SHell-script Integration

Masaomi Hatakeyama

University of Zurich / Functional Genomics Center Zurich

25 Dec. 2025, 14:00-



Scope and aims

1. Understand the functions of **SUSHI** system
2. Understand typical workflows of NGS data analyses
3. Perform SUSHI applications on RNAseq

Note

NOT look deeply into

- NGS technologies
- Background theories

<https://fgcz.ch/education.html>

- Genomics course (Bio680)
- Transcriptomics courses (Bio675)
- Integrative-Omics course (Bio393)



Outline

Afternoon	
14.00-14.15	Guidance, Introduction to SUSHI
14.15-14.45	Lecture/Demonstration, RNAseq analysis
14.45-15.00	break
15.00-16.00	Hands-on practice, RNAseq, Variant analysis, Free Q&A, break
16.00-16.30	Reproducible research by SUSHI, Wrap-up



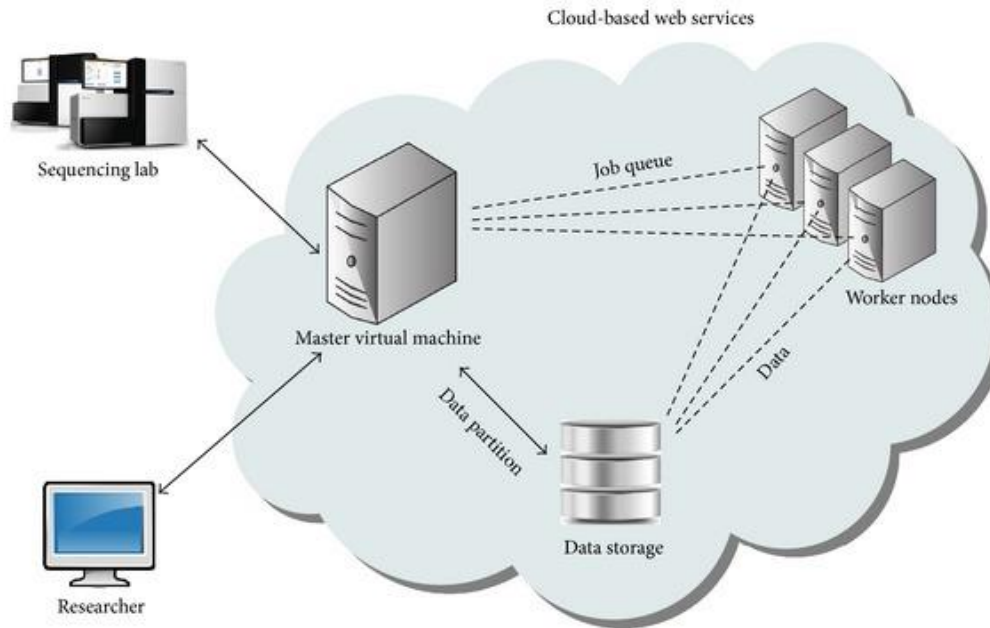
https://github.com/masaomi/bio610_2025_sushi_practice

Major challenges

in working with NGS data

Major challenges in working with NGS data

1. *Cluster/cloud computing* is often needed
2. *Reproducibility* – standardized and documented workflows



Analyzing NGS data with flexibility and reproducibility

- Command Line Interface (CLI)
 - Unix
 - Cluster computing
 - Scripting (R, python, ruby)
- Pros
 - Flexible, reproducible when workflows are scripted down
- Cons
 - Steep learning curves

- Example job script: bwa.sh

```
#!/bin/bash

# Create a working directory
# Make an index (files) for BWA
cd /scratch
mkdir -p test
cd test
cp LOCATION_OF_REF/reference.fasta .
bwa index reference.fasta

# Align reads using 8 cores
cp LOCATION_OF_DATA/data.fasta.gz .
bwa -t 8 mem reference.fasta data.fastq.gz > aln.sam

# Convert SAM to a sorted, indexed BAM file
samtools view -bT reference.fasta -o aln.bam aln.sam
samtools sort aln.bam aln.sorted
samtools index aln.sorted.bam

# Copy result files to storage
cp aln.sorted.bam LOCATION_OF_STORAGE
cp aln.sorted.bam.bai LOCATION_OF_STORAGE
```



- Submit the job script to cluster

```
sbatch -p user -n 8 -mem=100G bwa.sh
```

User friendly NGS data analysis pipelines

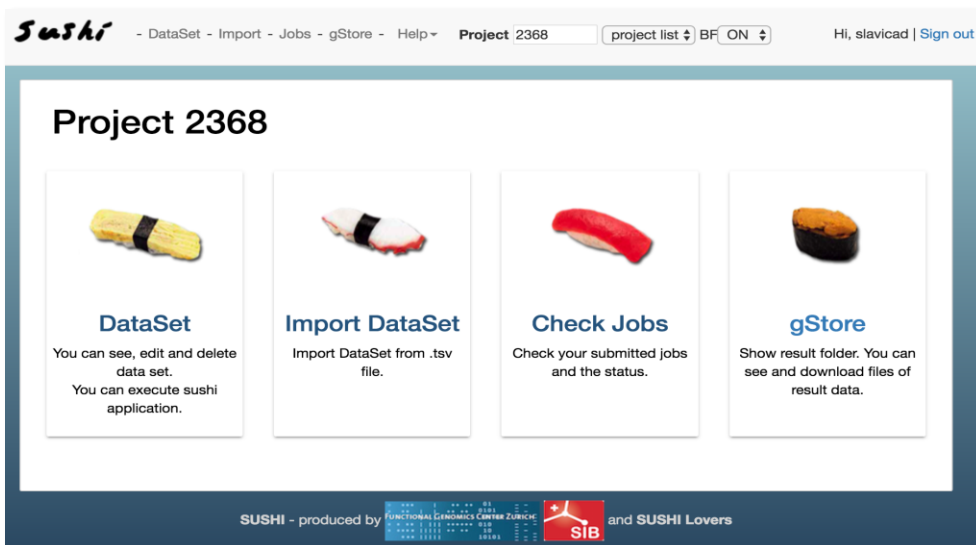
- Graphical User Interface (GUI)
- Commercial packages
 - CLC Genomics Workbench
- Scientific workflow managers/data analysis frameworks
 - **SUSHI**
 - Galaxy
 - Chipster
 - ...
- Pros
 - Easy to use
 - Flat learning curve
 - Documented, reproducible workflows
 - Usually implemented following the best-practice within community
- Cons
 - Limited by supported algorithms/modules



Why SUSHI?

SUSHI can Support you by SHell-script Integration

- SUSHI does make a job script (programming code) for you*



- InHouse Tool
- Easy to use
- Available for all FGCZ-Users
- Perfectly fitting to FGCZ environment

SOFTWARE

OPEN ACCESS

SUSHI: an exquisite recipe for fully documented, reproducible and reusable NGS data analysis

Masaomi Hatakeyama^{1,2}, Lennart Opitz¹, Giancarlo Russo¹, Weihong Qi¹, Ralph Schlapbach¹ and Hubert Rehrauer¹ ✉

BMC Bioinformatics BMC series – open, inclusive and trusted 2016 17:228 |

DOI: 10.1186/s12859-016-1104-8 | © Hatakeyama et al. 2016

Received: 20 February 2016 | Accepted: 26 May 2016 | Published: 2 June 2016

<https://fgcz-sushi.uzh.ch>

<http://fgcz-sushi-demo.uzh.ch>

SUSHI features

1. **Generating a job script for you only by clicking**
 - The generated job script can run independently from SUSHI
 - You can keep the script for *reproducibility*
2. **Result files are full self-contained and documented**
 - Scripts are available as part of the *documentation*
3. **Managing meta-data by meta-process**
 - projects, samples, inputs/outputs, software versions, parameters, etc.
 - SUSHI application generates a job script by integrating the *meta-data*

SUSHI operation

SUSHI important concepts

Selection step

1. **DataSet**: Table, A set of meta-information
2. **SUSHIApp**: Link to a DataSet, corresponding to one job (calculation)



HiSeq_short_dataset_20200609 Content [Hide](#) [- edit table](#) [- data folder](#)

Name	Group [Factor]	Read1 [File]	Read2 [File]
AhaL_L1	Control	20130830.A-hal_w...	20130830.A-hal_w...
AhaL_L3	Control	20130830.A-hal_w...	20130830.A-hal_w...
AhaL_L4	Control	20130830.A-hal_w...	20130830.A-hal_w...
AhaL_L48hZ1	Treated	20130830.A-hal_w...	20130830.A-hal_w...
AhaL_L48hZ3	Treated	20130830.A-hal_w...	20130830.A-hal_w...
AhaL_L48hZ4	Treated	20130830.A-hal_w...	20130830.A-hal_w...

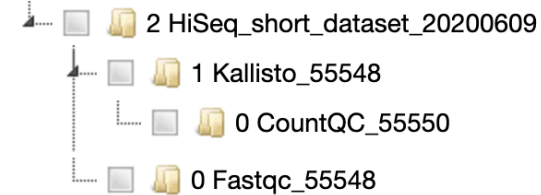
Showing 1 to 6 of 6 entries

Applications - [refresh](#)

Category	Application
ATAC	AtacENCODEApp
Assemble	CanuApp SpadesApp TrinityApp
Count	KallistoApp RSEMApp
Map	BWAApp BismarkApp Bowtie2App Bowtie2TranscriptomeApp BowtieApp STARApp
Metagenomics	DADA2Step1SampleApp KrakenApp MegahitApp MetaspadesApp MothurApp VirDetectApp

DataSets - [delete selected](#)

Search:



Note

- The DataSet tree shows DataSets relationship and analysing steps

SUSHI operations

1. Select a DataSet
2. Select a SUSHI application
3. Set job parameters
4. Submit a job
5. (coffee break)
6. Check job status
7. Check the next DataSet (result)

.... Repeating 1-7



HiSeq_short_dataset_20200609 Content [Hide](#) [- edit table](#) [- data folder](#)

Name	Group [Factor]	Read1 [File]	Read2 [File]
Ahal_L1	Control	20130830.A-hal_w...	20130830.A-hal_w...
Ahal_L3	Contr		20130830.A-hal_w...
Ahal_L4	Contr		20130830.A-hal_w...
Ahal_L48hZ1	Treat		20130830.A-hal_w...
Ahal_L48hZ3	Treated	20130830.A-hal_w...	20130830.A-hal_w...
Ahal_L48hZ4	Treated	20130830.A-hal_w...	20130830.A-hal_w...

Showing 1 to 6 of 6 entries

1. DataSet



Applications -

Category	Applicati	2. SUSHI App —	
ATAC	AtacENC		
Assemble	CanuApp	SpadesApp	TrinityApp
Count	KallistoApp	RSEMApp	
Map	BWAApp	BismarkApp	Bowtie2App
Metagenomics	DADA2Step1SampleApp	KrakenApp	MegahitApp
		MetaspadesApp	MothurApp
			VirDetectApp

Question?

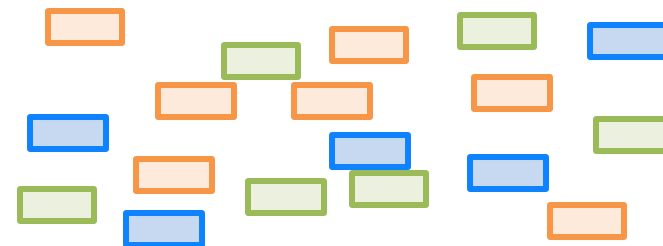
Demonstration RNAseq analysis

RNAseq analysis, transcriptomics, gene expression analysis

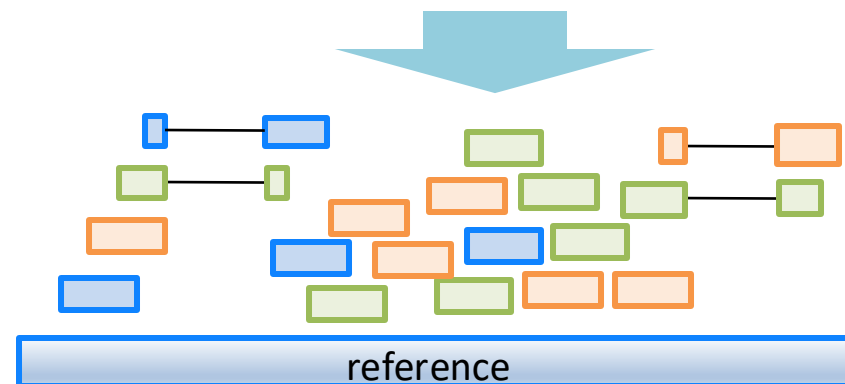
RNAseq analysis core steps

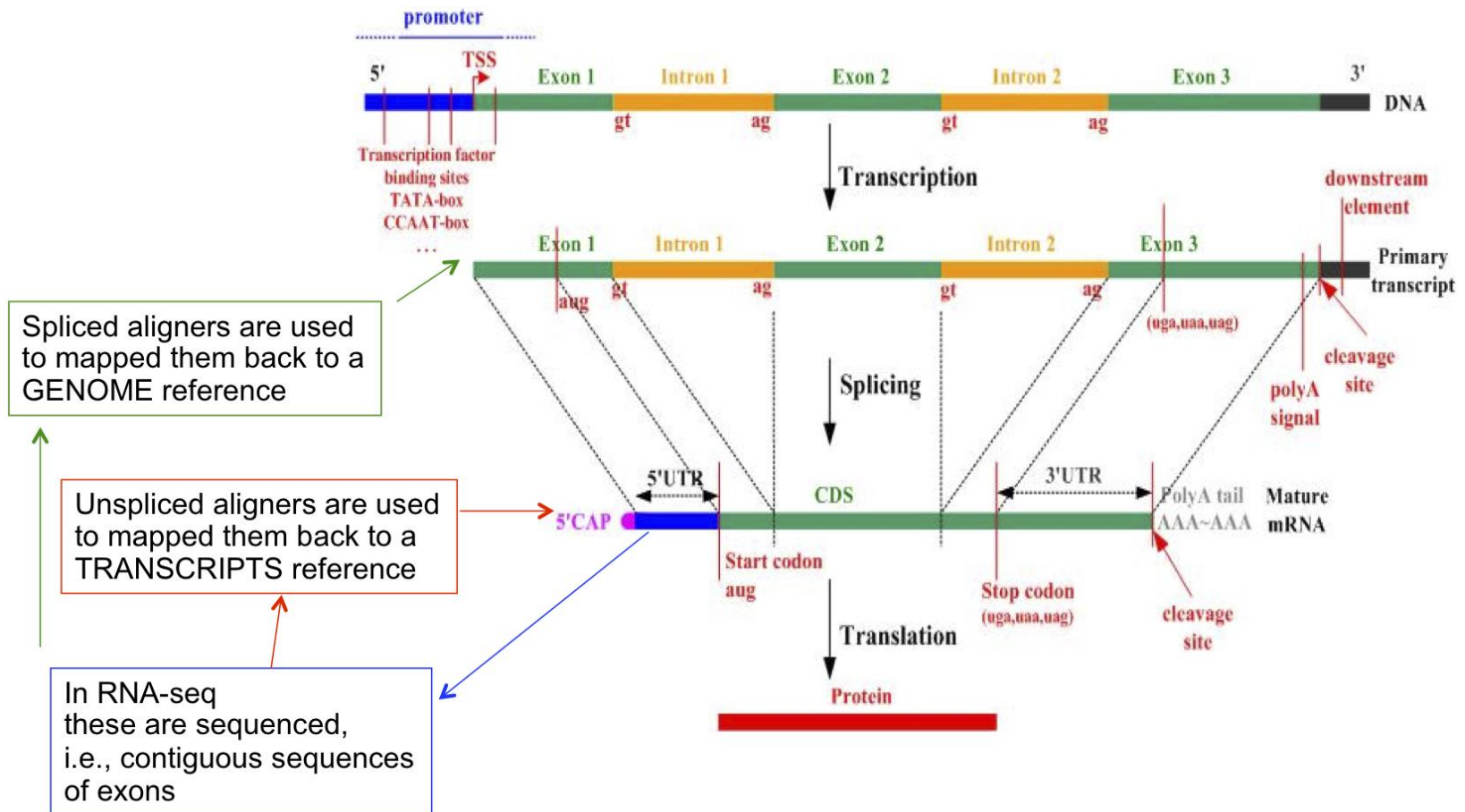
1. Align (map) a lot of short reads (transcripts) onto a reference sequence (genome, transcriptome)
2. Count abundance of each gene (transcript)
=Estimate expression level

RNAseq reads



Align (map) to a reference





Un-spliced vs. spliced mappers

- Spliced mappers (RNA-Seq)

- Aware of the presence of introns
- When encountering an intron, the aligner **does not stop** to trim the rest of the read but continues to find the next exon
- TopHat, **STAR**, , HISAT2, GMAP, BLAT



- Un-spliced mappers (DNA-Seq)

- Align continuous reads (not containing gaps as a result of splicing)
- When encountering an intron, the aligner **stops and trims the rest of the read**
- BWA, **BOWTIE2**, BLAST
- Splice junctions are impossible to detect



Sequence / Alignment (SAM) files

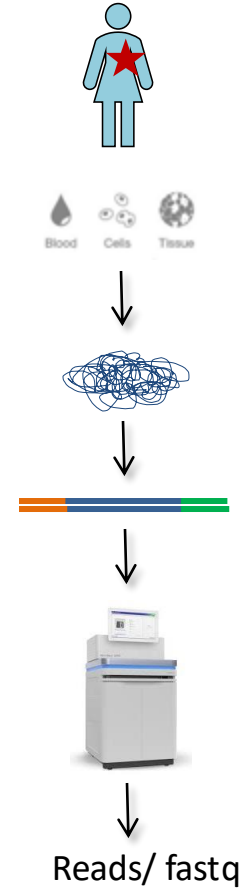
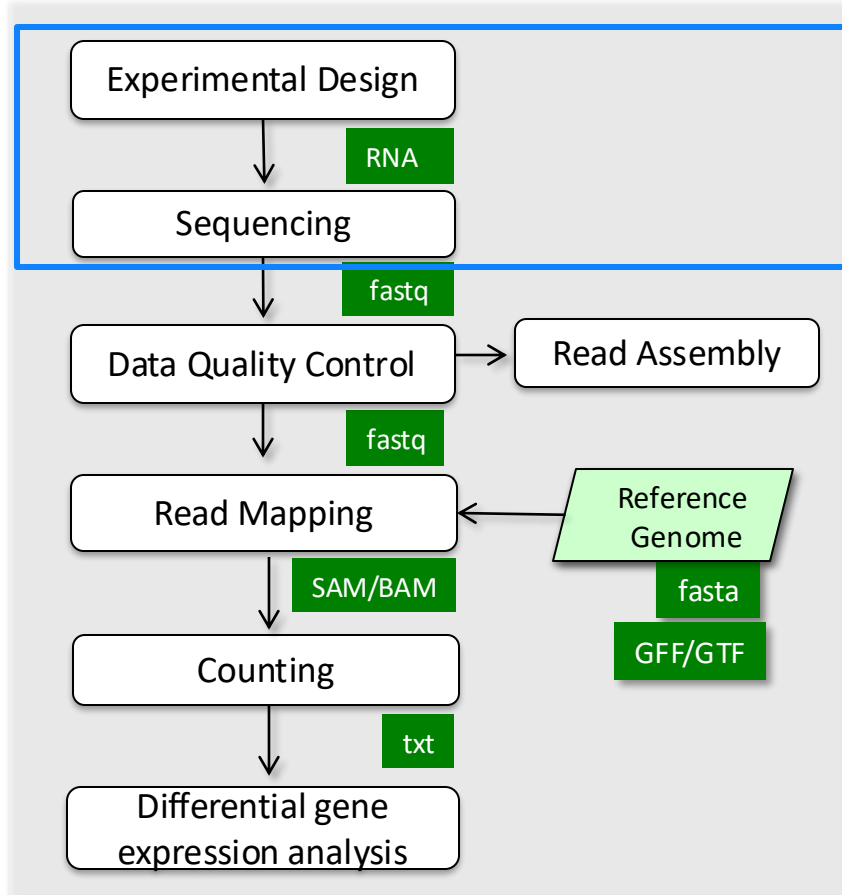
- **SAM (Sequence Alignment/Map)**

- Single unified format for storing read alignments to a reference genome
- Large plain **text file**
- RNA-Seq: >10GB
- Exome: >50GB, Whole Genome: 0.8-1TB

- **BAM (Binary Alignment/Map)**

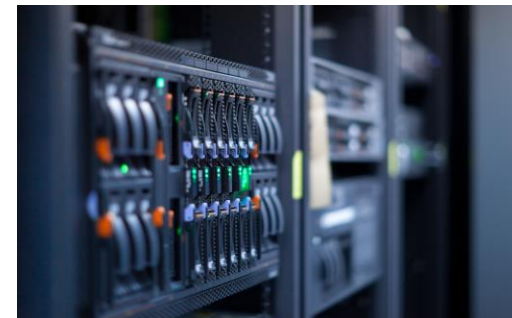
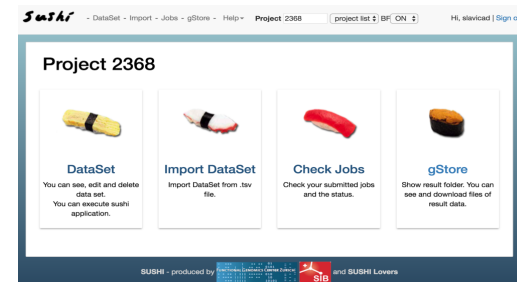
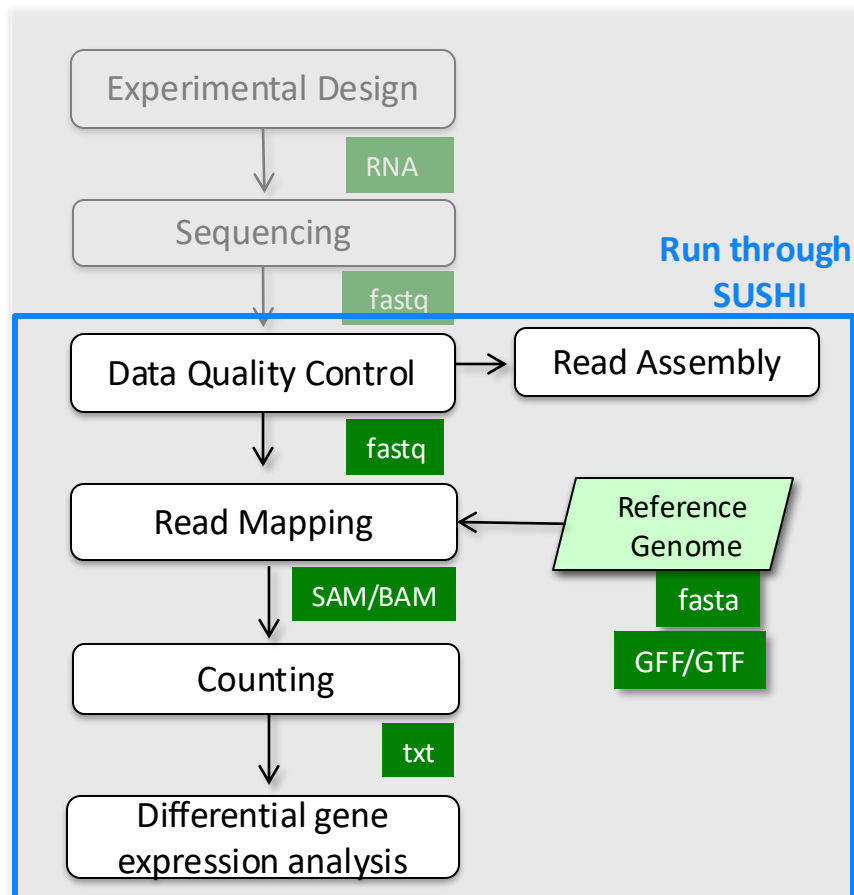
- Binary equivalent of SAM
- **Compressed data** plus index (bai)
- Developed for fast processing/indexing
- RNA-Seq: >2GB
- Exome: 2-10GB, Whole Genome: 100-300GB

Workflows (A typical RNAseq workflow)



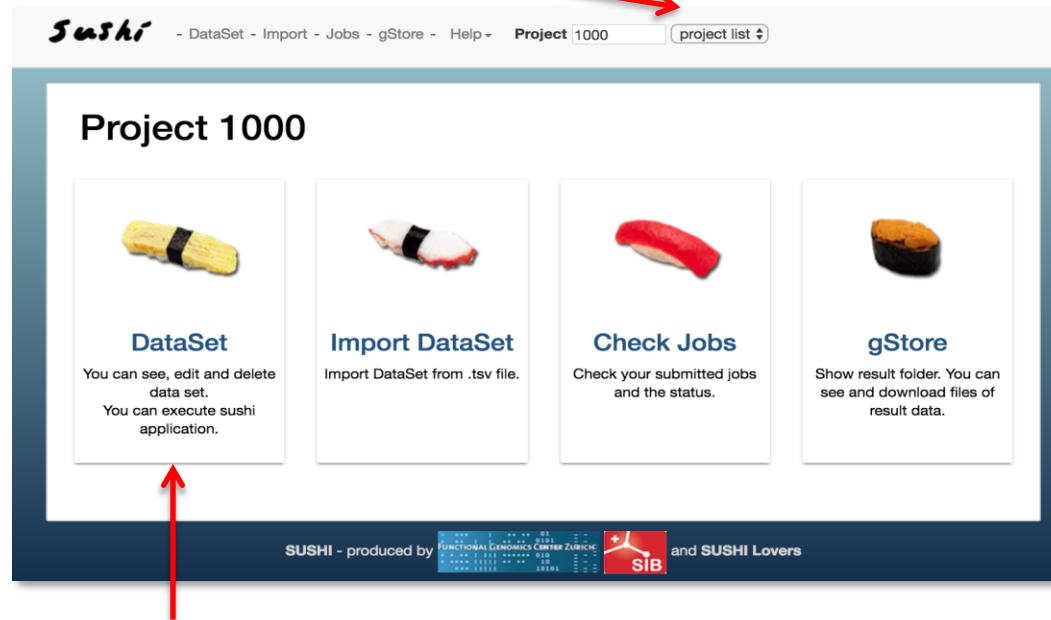
```
AG CT AG GG GCT GAACTT GCAGCATAC T GCAT AGG CT GA AGGG CTG CA GCAGCA
GG CATA GGAT GCAG CATA AGCAGTT A TTA GCAT AGAT GCAG CT TGGCAA GTA
GT AG CA TAG AT GCAGCT GGG CAACGAA
```

Workflows (A typical RNAseq workflow)



SUSHI (course) server

1. SUSHI course Server: <http://fgcz-course1.bfabric.org>
2. Select your project



3. Click on „**DataSet**“ to get a listing of existing DataSets in your Project.

1. Select a DataSet

Click a DataSet either in the tree or in the table



SUSHI - DataSet - Import - Jobs - gStore - Help - Project 1001 project list

DataSets show report - delete with selected

Search:

DataSet

0 NextSeq500_20170313_NS86_o3293_sub

Show 10 entries Search:

ID	Name	SushiApp	Samples	ParentID	Child(ren)IDs	Who	Created	BFabricID
15432	NextSeq500_20170...		8 / 8			sushi_lover	2017-Jun-30 08:56:14 Zurich	

Showing 1 to 1 of 1 entries Previous 1 Next

SUSHI - produced by  and SUSHI Lovers 

Note

- DataSet instances have a tree structure
 - Parental node = input DataSet, Child node = output DataSet
 - Connected by SUSHI Application

Check samples

Just look at the samples in the table

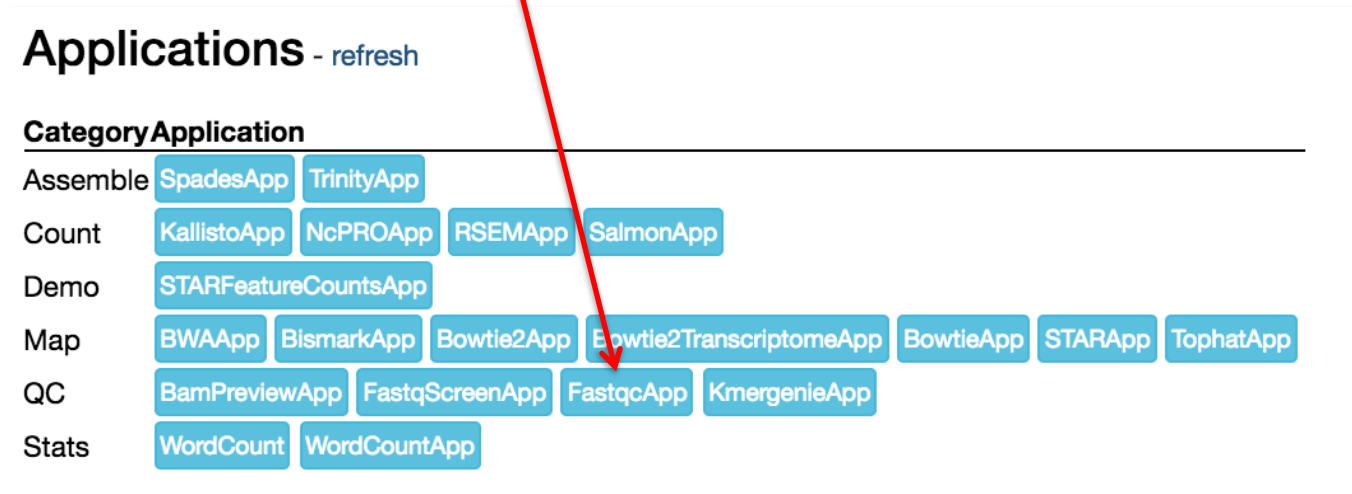
HiSeq_short_dataset_20200609 Content [Hide ▲](#) - [edit table](#) - [data folder](#)

Name ▲	Group [Factor] ▼	Read1 [File] ▼	Read2 [File] ▼
Ahal_L1	Control	20130830.A-hal_w...	20130830.A-hal_w...
Ahal_L3	Control	20130830.A-hal_w...	20130830.A-hal_w...
Ahal_L4	Control	20130830.A-hal_w...	20130830.A-hal_w...
Ahal_L48hZ1	Treated	20130830.A-hal_w...	20130830.A-hal_w...
Ahal_L48hZ3	Treated	20130830.A-hal_w...	20130830.A-hal_w...
Ahal_L48hZ4	Treated	20130830.A-hal_w...	20130830.A-hal_w...

Showing 1 to 6 of 6 entries

2. Select a SUSHI Application

To run FastqcApp, click on “**FastqcApp**” in the Application section



Applications - refresh

Category	Application
Assemble	SpadesApp TrinityApp
Count	KallistoApp NcPROApp RSEMAApp SalmonApp
Demo	STARFeatureCountsApp
Map	BWAApp BismarkApp Bowtie2App Bowtie2TranscriptomeApp BowtieApp STARApp TophatApp
QC	BamPreviewApp FastqScreenApp FastqcApp KmergenieApp
Stats	WordCount WordCountApp

Note

- Possible applications are automatically selected by DataSet type (Meta-data)

3. Set job parameters

FastqcApp Set Parameters

A quality control tool for NGS reads
[Web-site with docu and a tutorial video](#)

Next DataSet

Name

Comment

Parameters

cores	<input type="text" value="8"/>
ram	<input type="text" value="15"/> GB
scratch	<input type="text" value="100"/> GB
node	<input type="text"/>
queue	<input type="text"/>
partition	<input type="text" value="course"/>
process_mode	<input type="text" value="DATASET"/>
samples	<div><div>S1_undiff</div><div>S2_undiff</div><div>S3_undiff</div><div>S4_diff</div><div>S5_diff</div></div>
paired	<input type="text" value="false"/> required
perLibrary	<input checked="" type="checkbox"/> FastQC process per library or per cell for single cell experiment
name	<input type="text" value="FastQC_Result"/> required
cmdOptions	<input type="text"/>
mail	<input type="text"/>

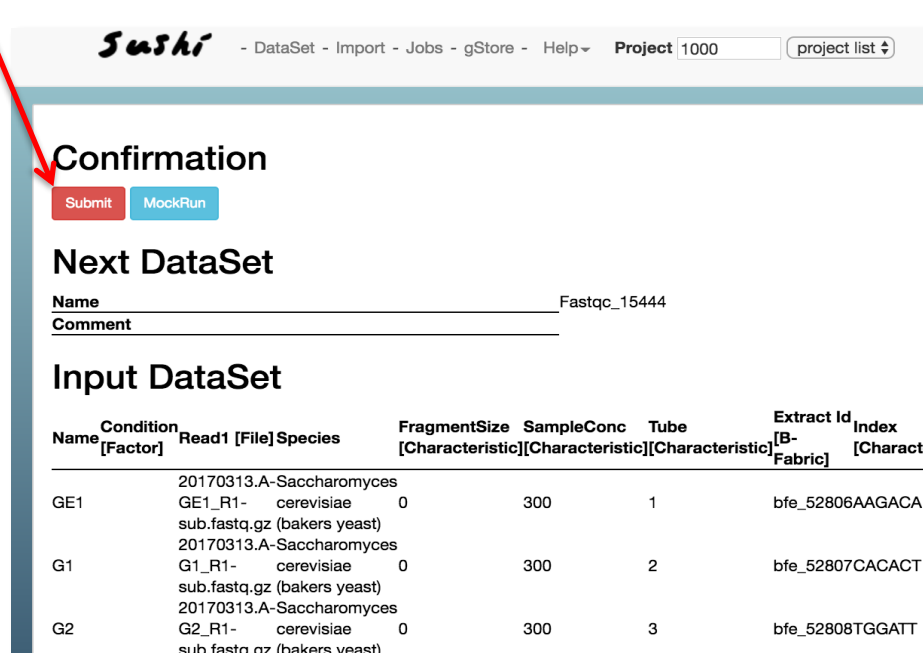
Common parameters

- **Name:** job name = folder base name
- **Cores:** how many CPU cores reserved
- **RAM:** how much RAM reserved
- **Scratch:** how much disk space reserved
- **Samples:** which samples are used

If all parameters are fine, continue with ,Next'

4. Submit a job

Click "Submit" and submit a job



SUSHI - DataSet - Import - Jobs - gStore - Help - Project: 1000 [project list]

Confirmation

Submit **MockRun**

Next DataSet

Name: Fastqc_15444
Comment:

Input DataSet

Name	Condition [Factor]	Read1	File	Species	FragmentSize [Characteristic]	SampleConc [Characteristic]	Tube [Characteristic]	Extract	Id [B-Fabric]	Index [Character]
GE1		20170313.A-Saccharomyces GE1_R1-	cerevisiae	0	300	1	bfe_52806AAGACA			
G1		20170313.A-Saccharomyces G1_R1-	cerevisiae	0	300	2	bfe_52807CACACT			
G2		20170313.A-Saccharomyces G2_R1-	cerevisiae	0	300	3	bfe_52808TGGATT			

Note

- SUSHI produces a job script (meta-process)
 - The shell script takes care of all detail calculation (base-process)

5. Check job status

Click “**Jobs**” and you can check the submitted job status

Sushi - DataSets - Import - **Jobs** - gStore - Help - Project 1000 project list

Project 1000

Job List - kill selected jobs

Show 10 entries Search:

Job ID	Status		User	JobScript	Log	DataSet	Time (Start/End)
86677	success	success	sushi_lover	Assemble_Phalo_P...	Log	Canu_PacBio	2021-04-30 13:04:53/2021-04-30 14:00:54
86676	fail	fail	sushi_lover	Assemble_Phalo_G...	Log	Canu_GridION	2021-04-30 13:03:45/2021-04-30 13:04:15
86675	fail	fail	sushi_lover	Assemble_WTCHG_4...	Log	Spades_MiSeq	2021-04-30 13:02:05/2021-04-30 13:03:35
86667	fail	fail	sushi_lover	Map_NA12878_chr1...	Log	Bowtie2_chr10_WE...	2021-04-30 12:01:54/2021-04-30 12:03:54

Job status

- *Pending*: Under job script preparation or waiting in a queue
- *Running*: the job is running, and please be patient
- *Success*: job succeeds, and you can see the result DataSet
- *Fail*: the job failed, and please look at the job log

Job script and parameters

You can check job “**scripts**” and “**parameters**”

Sushi - DataSets - Import - Jobs - gStore - Help - Project 1000 project list

DataSet - comment - rename - download - scripts - parameters - resubmit - delete

2 NextSeq500_20190507_NS278_o5638

0 Fastqc_NS278_o5638

- ID: 529
- App: FastqcApp
- Samples: 1
- Who: sushi_lover
- Created: 2021-Apr-29

Fastqc_NS278_o5638 Content Hide ▲ - edit table - data folder

FastQC result

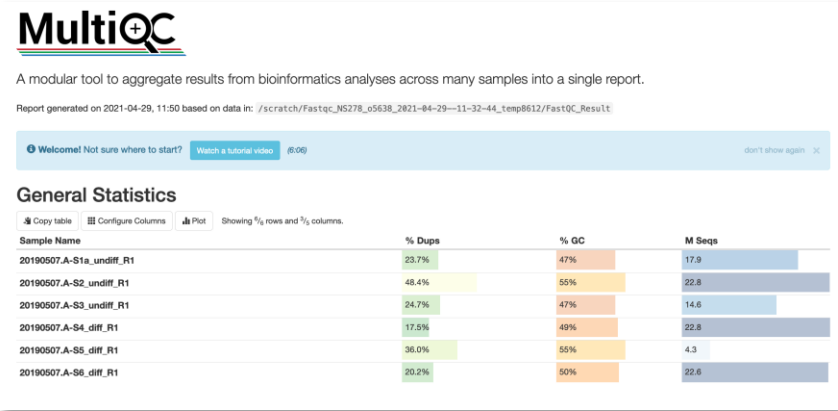
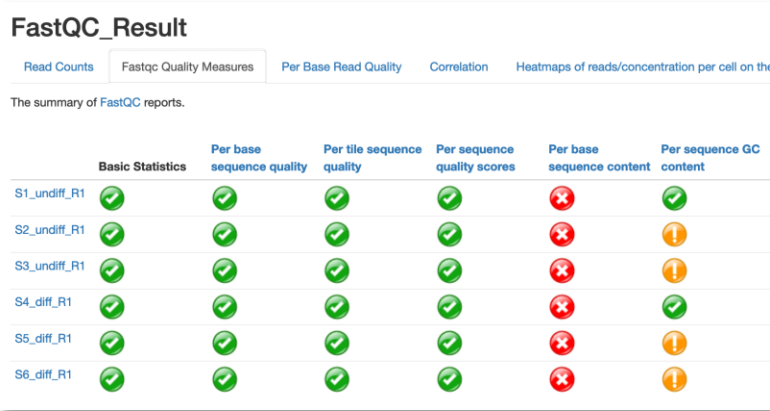
Two types of results: 1. Normal FastQC result, 2. MultiQC result

Fastqc_NS278_o5638 Content [Hide](#) [- edit table](#) [- data folder](#)

Search:

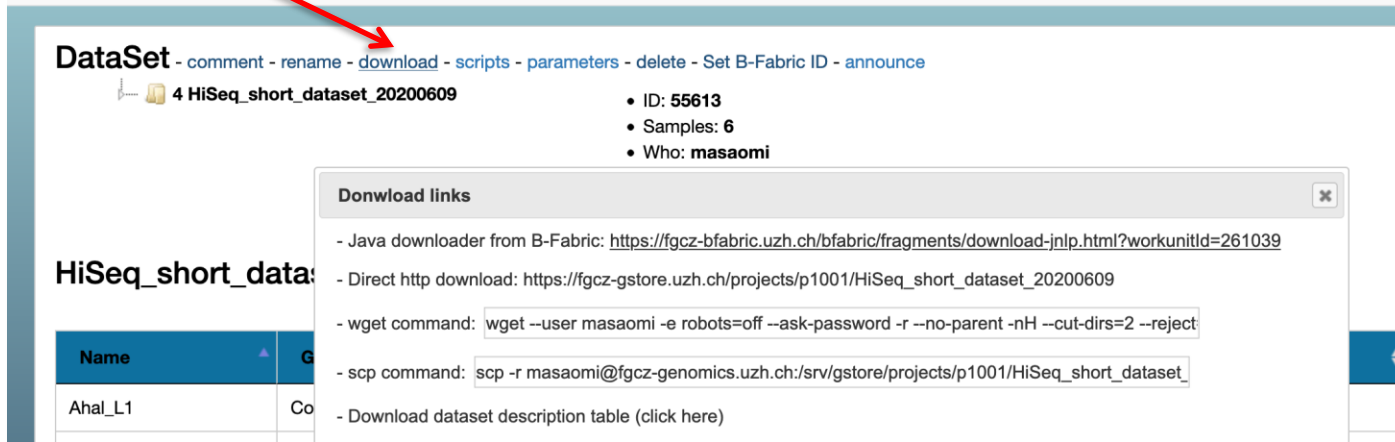
Name	Report [File]	Html [Link]	MultiQC [Link]
FastQC_Result	FastQC_Result	00index.html	multiqc_report.h...

Showing 1 to 1 of 1 entries



Download result

“Download” menu shows several options



DataSet - comment - rename - **download** - scripts - parameters - delete - Set B-Fabric ID - announce

4 HiSeq_short_dataset_20200609

- ID: 55613
- Samples: 6
- Who: masaomi

HiSeq_short_data

Name	Group
Ahal_L1	Co

Download links

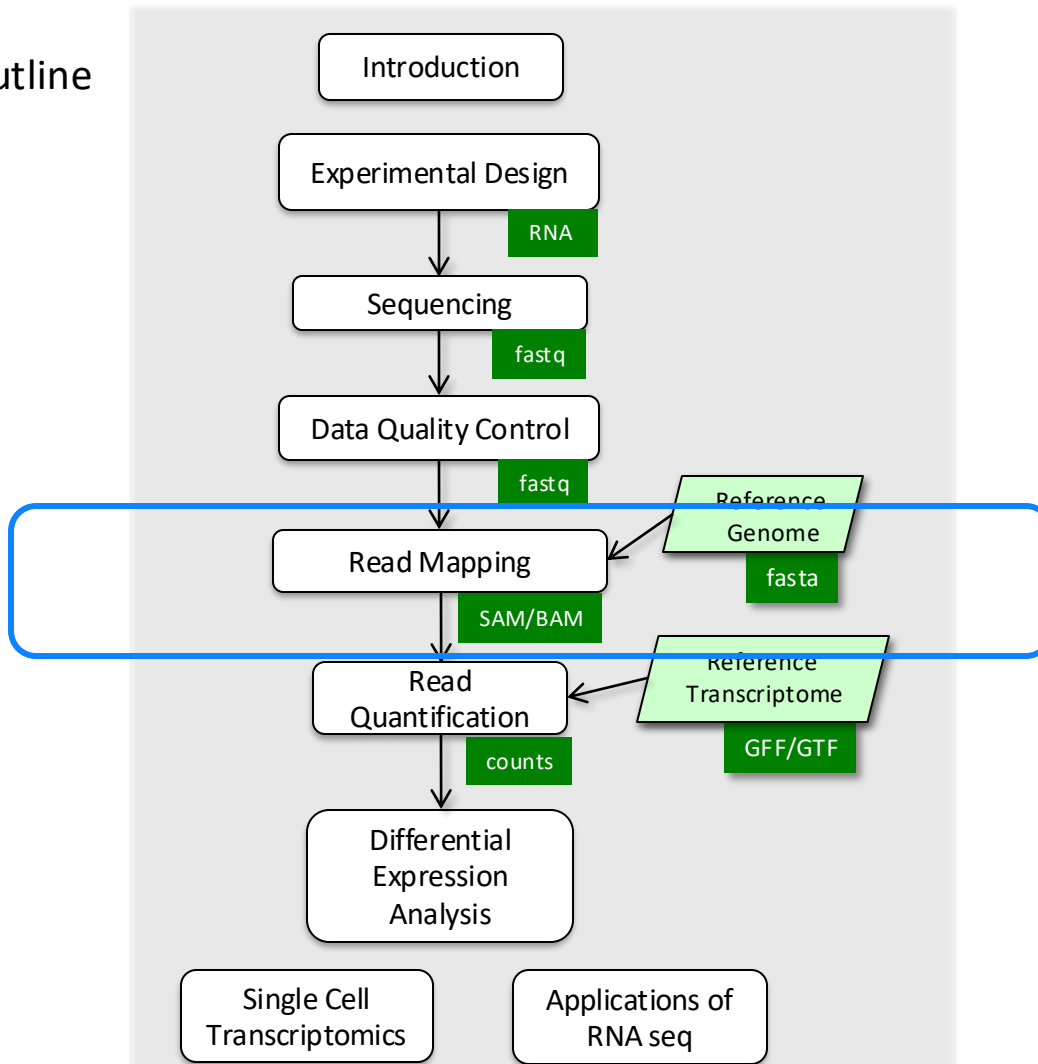
- Java downloader from B-Fabric: <https://fgcz-bfabric.uzh.ch/bfabric/fragments/download-jnlp.html?workunitId=261039>
- Direct http download: https://fgcz-gstore.uzh.ch/projects/p1001/HiSeq_short_dataset_20200609
- wget command: `wget --user masaomi -e robots=off --ask-password -r --no-parent -nH --cut-dirs=2 --reject`
- scp command: `scp -r masaomi@fgcz-genomics.uzh.ch:/srv/gstore/projects/p1001/HiSeq_short_dataset_`
- Download dataset description table (click here)

Options

- *Java downloader*: via BFabric
- *Direct http*: via File server
- *wget command*: in terminal via File server
- *scp command*: in terminal via File server
- *Download dataset*: DataSet sample table (.tsv) (**NOT actual data file(s)**, only DataSet table info.)

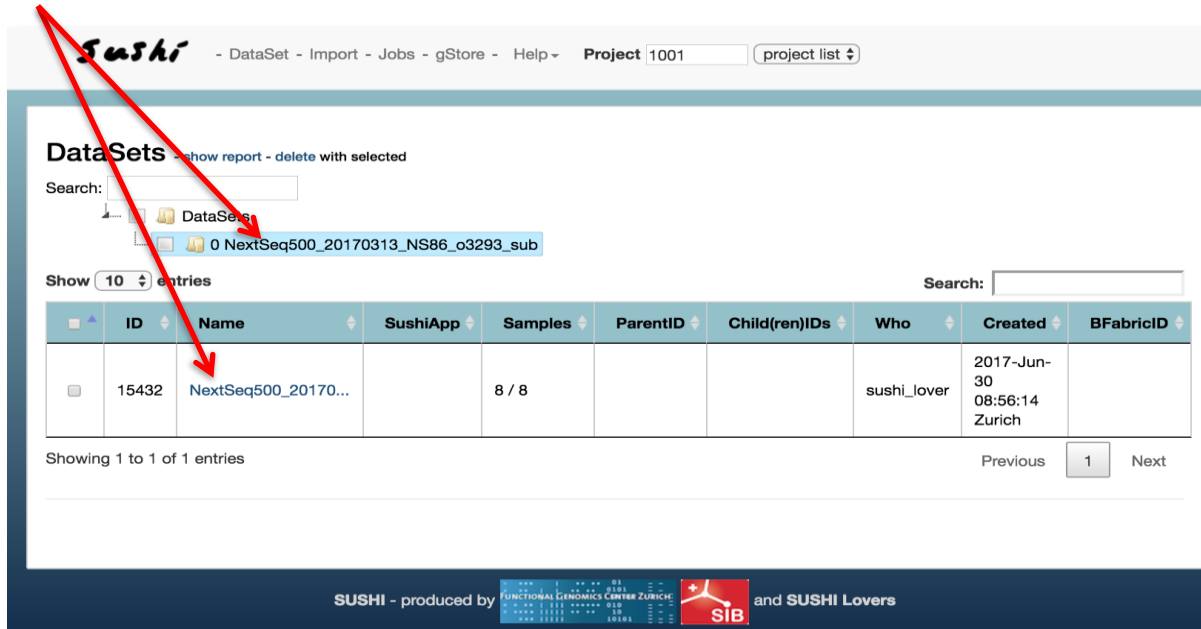
Question?

Outline



1. Select a DataSet

Click a DataSet either in the tree or in the table



The screenshot shows the SUSHI web interface. At the top, there is a navigation bar with links: DataSet, Import, Jobs, gStore, Help, and a Project dropdown set to 1001. Below this, the 'DataSets' section is active, showing a search bar and a tree view on the left. The tree view shows a folder 'DataSets' containing a file '0 NextSeq500_20170313_NS86_o3293_sub'. Below the tree, there is a table with columns: ID, Name, SushiApp, Samples, ParentID, Child(ren)IDs, Who, Created, and BFabricID. The table contains one entry with ID 15432 and Name 'NextSeq500_20170...'. A red arrow points from the tree view to this entry. At the bottom, there is a footer with logos for SUSHI, Functional Genomics Center Zurich, SIB, and SUSHI Lovers.

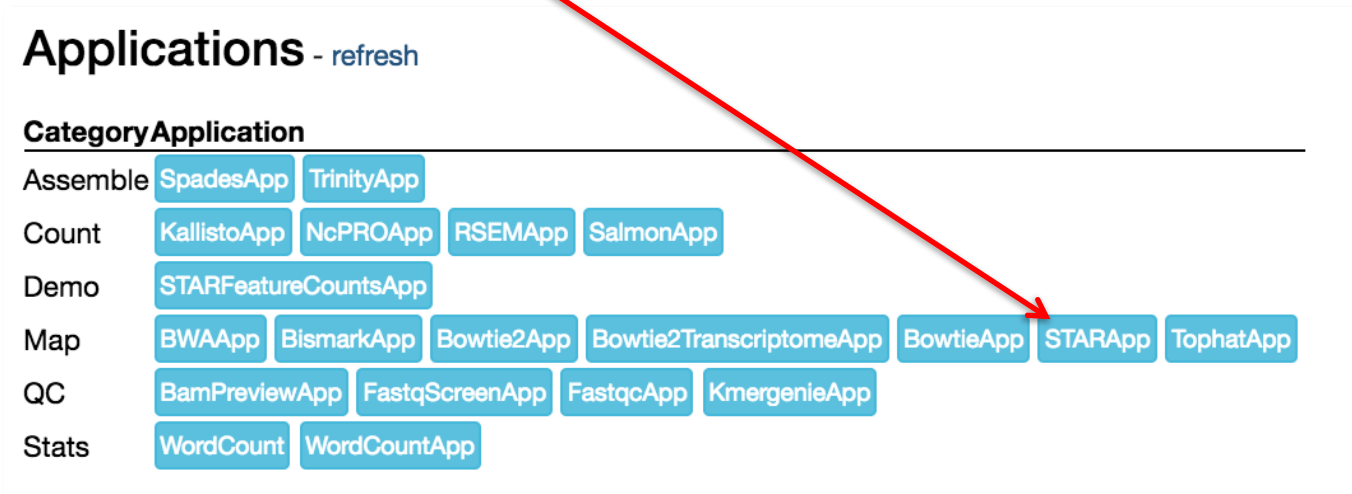
ID	Name	SushiApp	Samples	ParentID	Child(ren)IDs	Who	Created	BFabricID
15432	NextSeq500_20170...		8 / 8			sushi_lover	2017-Jun-30 08:56:14 Zurich	

Note

- DataSet instances have a tree structure
 - Parental node = input DataSet, Child node = output DataSet
 - Connected by SUSHI Application

2. Select a SUSHI Application

To align the reads, click on “**STARApp**” in the Application section



Applications - refresh

Category	Application
Assemble	SpadesApp TrinityApp
Count	KallistoApp NcPROApp RSEMAApp SalmonApp
Demo	STARFeatureCountsApp
Map	BWAApp BismarkApp Bowtie2App Bowtie2TranscriptomeApp BowtieApp STARApp TophatApp
QC	BamPreviewApp FastqScreenApp FastqcApp KmergenieApp
Stats	WordCount WordCountApp

Note

- Possible applications are automatically selected by DataSet type (Meta-data)
- *Kallisto* could be the second option for gene expression analysis

3. Set job parameters

STARApp

Set Parameters

Ultrafast spliced alignment

manual

Noteworthy options:

- outFilterMatchNmin 30 --outFilterMismatchNmax 5 --outFilterMismatchNoverLmax 0.05 --outFilterMultimapNmax 50 --alignEndsProtrud
- large numbers of contigs ask form more RAM. In that case the index must be built with a smaller --genomeChrBinNbits 18; e.g. the Wheat

Next DataSet

Name

STAR_19024

Comment

Parameters

cores

8

GB

ram

40

GB

scratch

100

GB

node

fgcz-c-042,fgcz-c-045,fgcz-c-046,fgcz-c-047,fgcz-c-048,fgcz-c-051,fgcz-c-052,fgcz-c-053,fgcz-c-054,fgcz-c-055,fgcz-c-056,fgcz-c-057,fgcz-c-058,fgcz-c-05

process_mode

SAMPLE

samples

WT_Ler_s1

WT_Ler_s2

WT_Ler_s3

uvr8_Ler_s4

uvr8_Ler_s5

refBuild

select

required

paired

false

required

strandMode

antisense

required

refFeatureFile

genes.gtf

cmdOptions

--outFilterType BySJout -

getJunctions

false

twopassMode

true

Per-sample 2-pass mapping or 1-pass mapping in STAR. 2-pass mapping allows to detect more splices reads mapping to novel junctions.

trimAdapter

true

trimLeft

0

trimRight

0

minTailQuality

0

minTrailingQuality

10

minAvgQuality

10

minReadLength

20

specialOptions

mail

Next

STARApp critical parameters

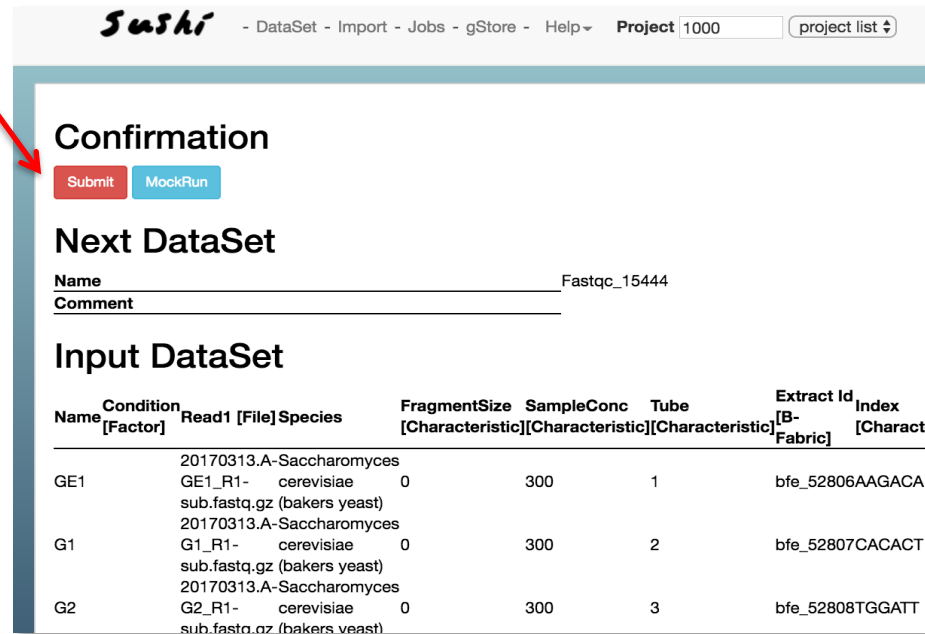
- **refBuild**: reference genome
- **Paired**: true(Paired-end)/false (Single-end)
- **StrandMode**: (only in single-end)
- **TrimAdapter**: true/false

Note

- **TrimAdapter** option is **true** (by default)

4. Submit a job

Start the actual job with **,Submit'**



Sushi - DataSet - Import - Jobs - gStore - Help - Project: 1000 [project list]

Confirmation

Next DataSet

Name: Fastqc_15444
Comment:

Input DataSet

Name	Condition [Factor]	Read1 [File]	Species	FragmentSize [Characteristic]	SampleConc [Characteristic]	Tube [Characteristic]	Extract Id [B-Fabric]	Index [Character]
GE1		20170313.A-Saccharomyces GE1_R1- sub.fastq.gz (bakers yeast)	cerevisiae	0	300	1	bfe_52806AAGACA	
G1		20170313.A-Saccharomyces G1_R1- sub.fastq.gz (bakers yeast)	cerevisiae	0	300	2	bfe_52807CACACT	
G2		20170313.A-Saccharomyces G2_R1- sub.fastq.gz (bakers yeast)	cerevisiae	0	300	3	bfe_52808TGGATT	

Visualization Mapping result

Mapping visualization

- IGV (Integrative Genomics Viewer)

<https://software.broadinstitute.org/software/igv/>



Download

- <https://software.broadinstitute.org/software/igv/download>

Home › Downloads

Downloads

Did you know that there is also an **IGV web application** that runs only in a web browser, does not use Java, and requires no downloads? See <https://igv.org/app>. Click on the [Help](#) link in the app for more information about using IGV-Web.


Install IGV 2.16.0


See the [Release Notes](#) for what's new in each IGV release.


Users of the new M1 Mac: Apple's Rosetta software is required to run the IGV MacOS App that includes Java. If you run IGV with your own Java installation, Rosetta may not be required if your version of Java runs natively on M1.


Linux users: The 'IGV for Linux' download includes AdoptOpenJDK (now Eclipse Temurin) version 11 for x64 Linux. See [their list of supported platforms](#). If this does not work on your version of Linux, download the 'Command line IGV for all platforms' and use it with your own Java installation.


About log4j: IGV versions 2.4.1 - 2.11.6 used log4j2 code that is subject to the log4jShell vulnerability. We recommend using version 2.11.9 (or later), which removed all dependencies on log4j.

 IGV MacOS App
Java included

 IGV MacOS App
Separate Java 11 required

 IGV for Windows
Java included

 IGV for Windows
Separate Java 11 required

 IGV for Linux
Java included

Mapped reads in IGV

Reference track

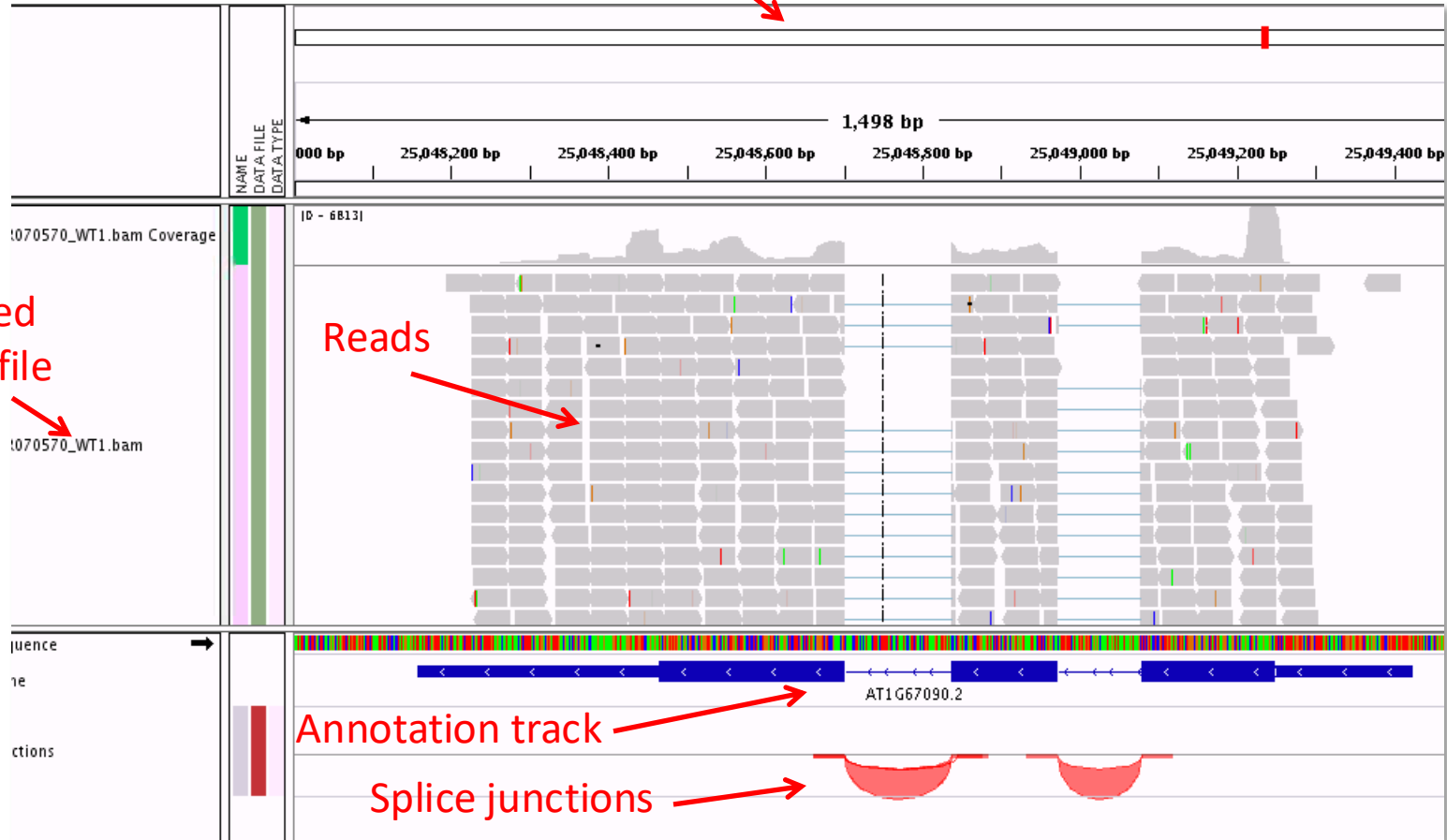
Loaded
BAM file

070570_WT1.bam

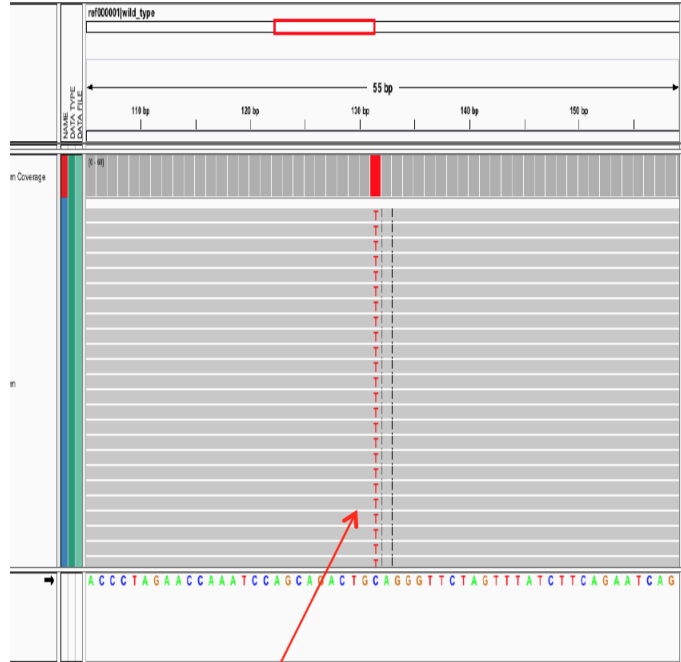
Reads

Annotation track

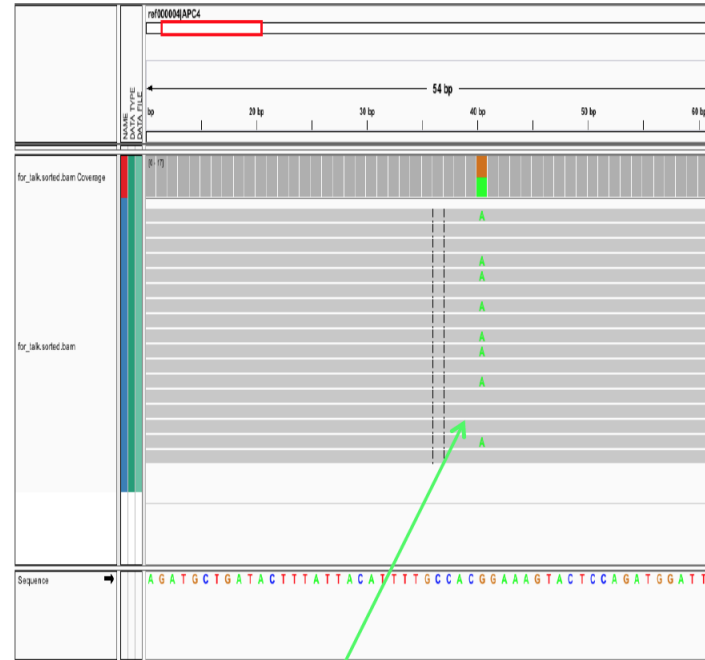
Splice junctions



SNPs

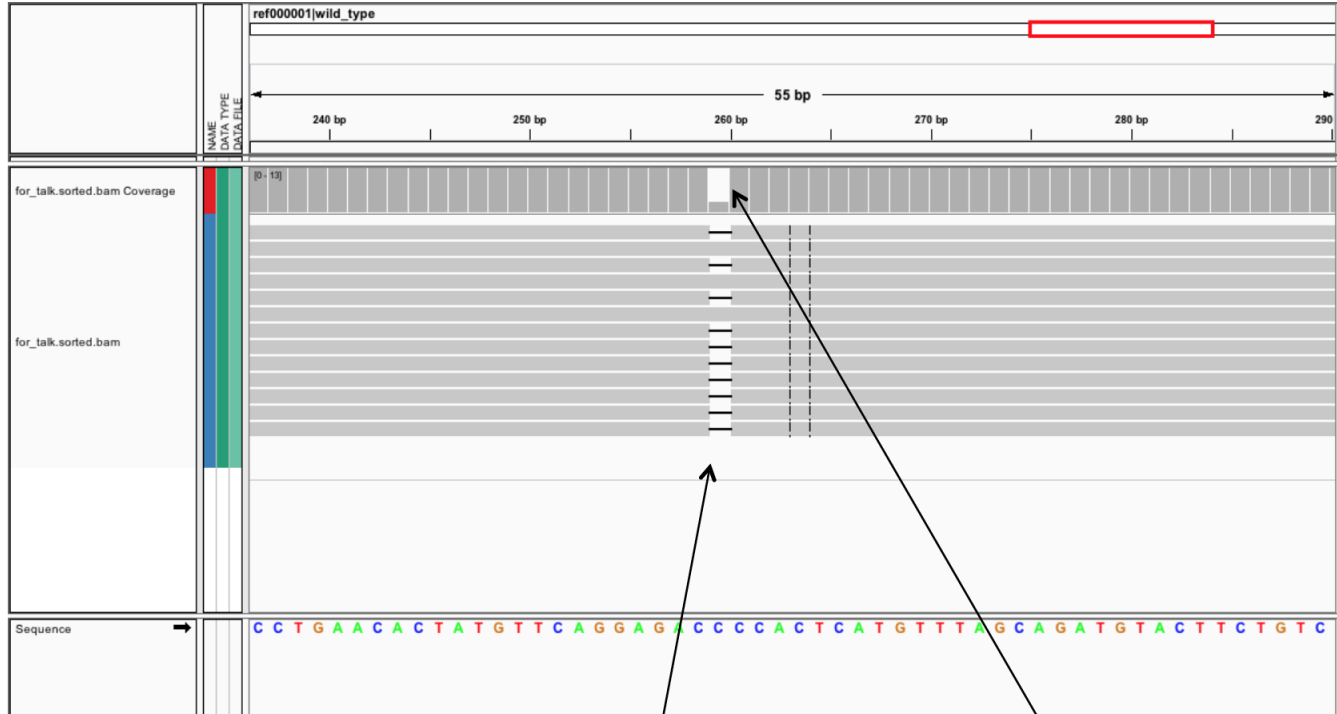


C > T homozygous substitution



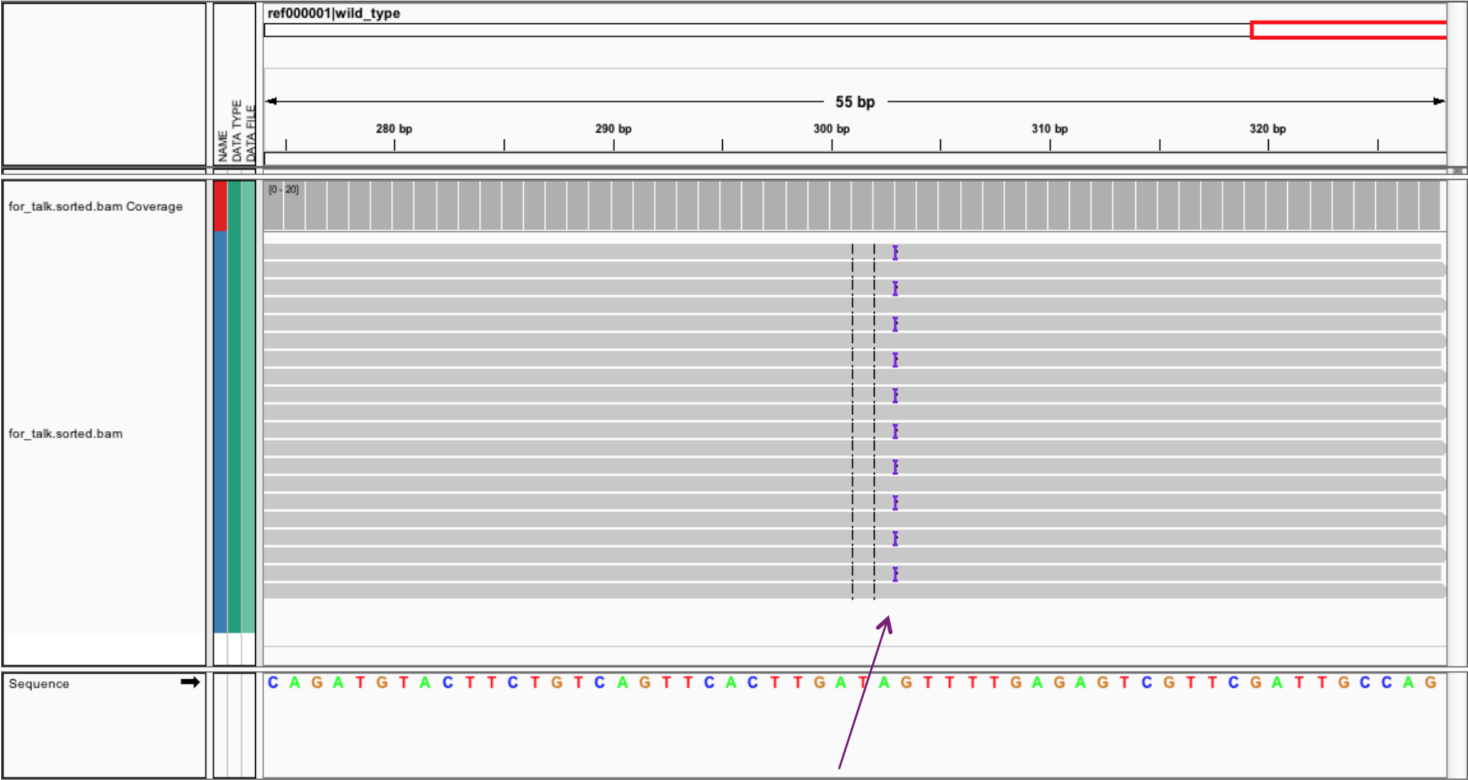
G > A heterozygous substitution

Single base deletions



1 base deletion (note the drop in coverage)

Single base insertions



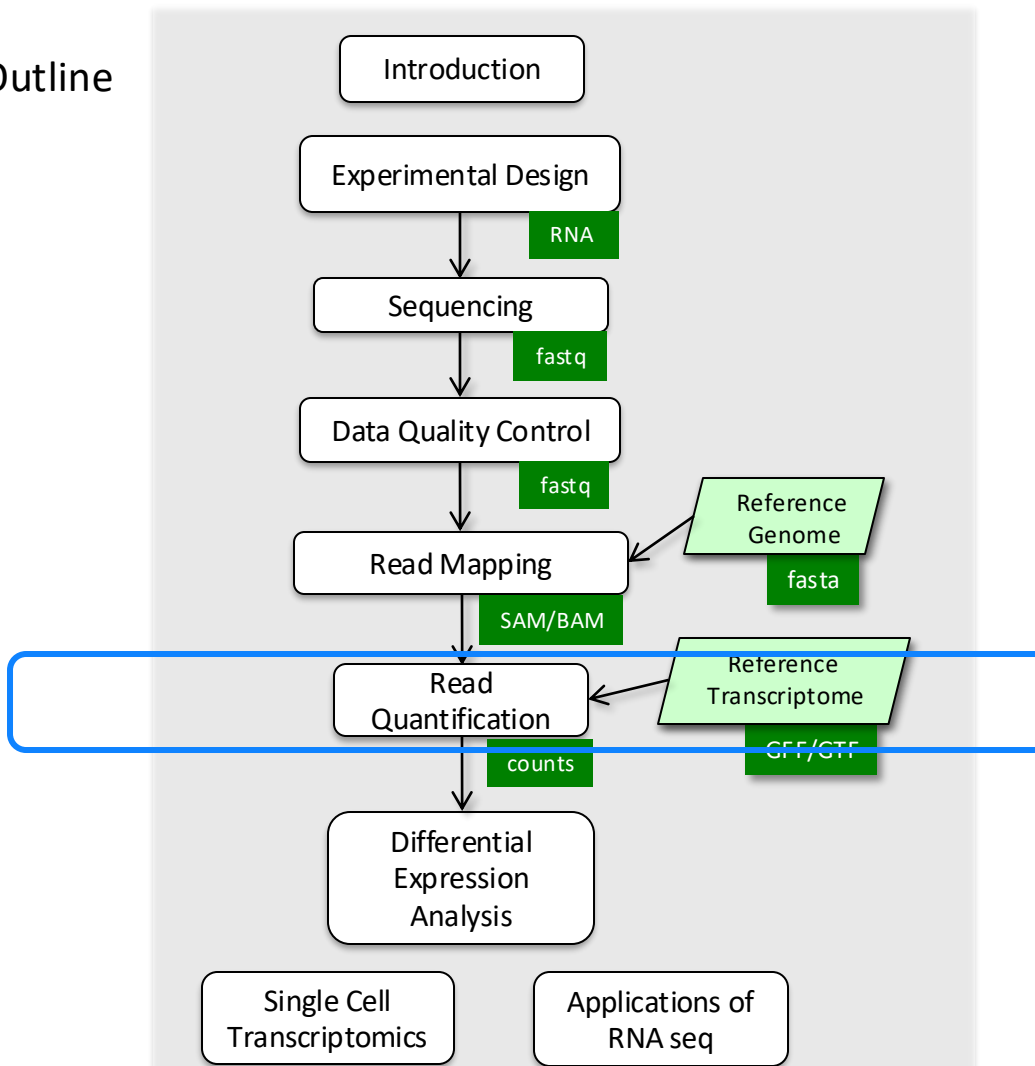
1 base insertion

Sashimi plots

- Quantitatively visualize splice junctions



Outline




Read quantification

To count the mapped reads, click on “**FeatureCountsApp**” in the Application section

Applications - refresh

Category	Application
ATAC	AtaqvApp HomerDiffPeaksApp
Alternative_Splicing	DEXSeqApp JunctionSeqApp
Count	FeatureCountsApp
Peaks	MACS2App
Polyploid	MergeDataSetApp
QC	AtacQcApp DnaBamStatsApp DnaQCApp RnaBamStatsApp TeqcApp
Variants	CNVnatorApp GatkDnaHaplotyperApp GatkRnaApp MpileupApp VariantCallerApp



FeatureCountsApp job parameters

FeatureCountsApp

Set Parameters

Multi-purpose read counting with Rsubread::featureCounts

manual

Consider default values as suggestions. They are subject to change!

Next DataSet

Name

FeatureCounts_19023

Comment

Parameters

cores

8

GB

ram

20

GB

scratch

10

GB

node

fgcz-c-042,fgcz-c-045,fgcz-c-048,fgcz-c-047,fgcz-c-048,fgcz-c-051,fgcz-c-052,fgcz-c-053,fgcz-c-054,fgcz-c-055,fgcz-c-056,fgcz-c-057,fg

process_mode

SAMPLE

samples

WT_Ler_s1

WT_Ler_s2

WT_Ler_s3

uvr8_Ler_s4

uvr8_Ler_s5

refBuild

Arabidopsis_thaliana/Ara

required

paired

false

required

strandMode

antisense

required

refFeatureFile

genes.gtf

featureLevel

gene

gtfFeatureType

exon

which atomic features of the gtf should be used to define the meta-features; see featureLevel

allowMultiOverlap

true

count alignments that fall in a region where multiple features are annotated

countPrimaryAlignmentsOnly

true

minFeatureOverlap

10

minimum overlap of a read with a transcript feature

minMapQuality

10

keepMultiHits

true

transcriptTypes

protein_coding

rRNA

tRNA

Mt_rRNA

Mt_tRNA

long_noncoding

short_noncoding

pseudogene

specialOptions

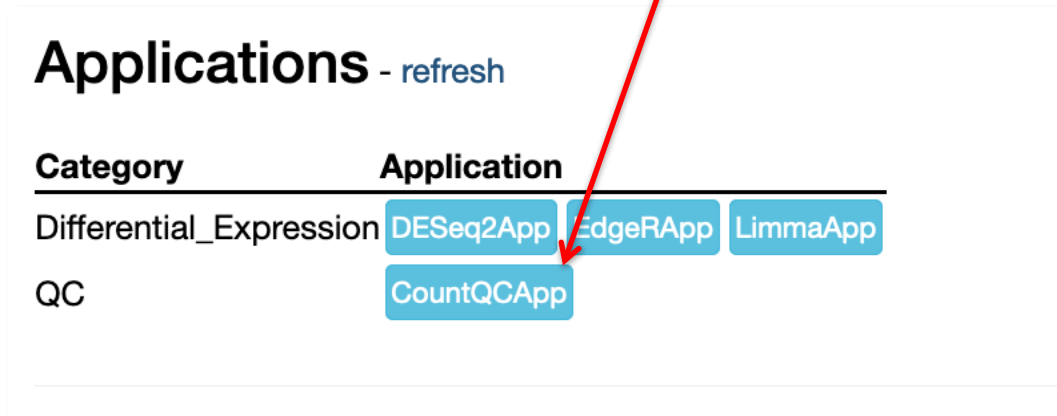
mail

- FeatureCountsApp critical parameters
- allowMultiOverlap***: true/false
 - countPrimaryAlignmentsOnly***: true/false

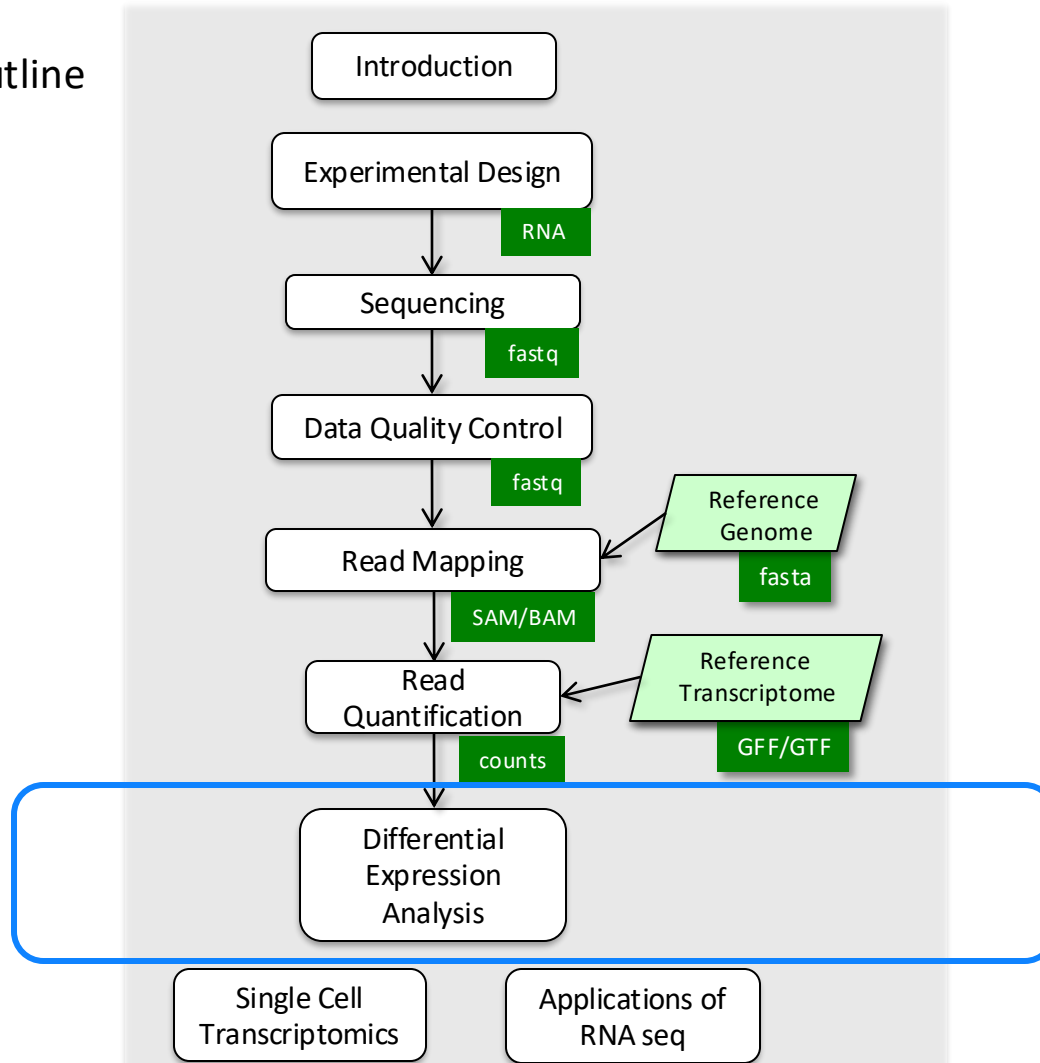
If all parameters are fine, continue with ,Next'

CountQC

After counting the reads, click on “**CountQCAApp**” in the Application section



Outline



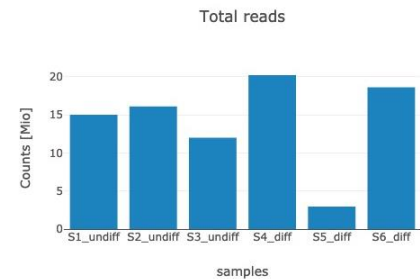
Differential expression analysis (EdgeR)

grouping	Condition	required
sampleGroup	please select	required sampleGroup should be different from refGroup
sampleGroupBaseline	please select	select the baseline for sampleGroup if you have
refGroup	please select	required refGroup should be different from sampleGroup
refGroupBaseline	please select	select the baseline for refGroup if you have

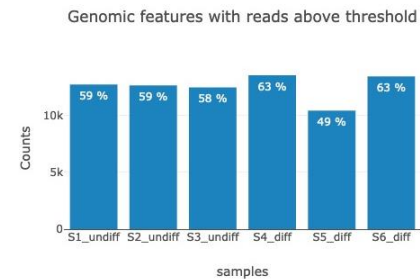
- edgeRApp critical parameters
- Grouping**: corresponding to [Factor] tagged column in DataSet
 - sampleGroup**: comparison group (e.g. treated sample)
 - refGroup**: reference group (e.g. control sample)

CountQC (static report)

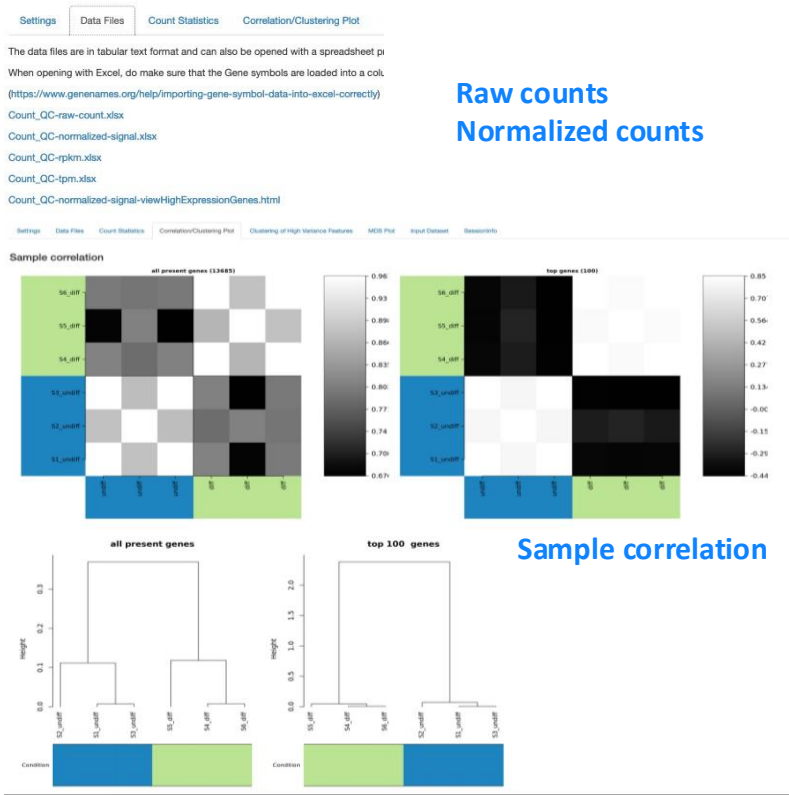
SettingsData FilesCount StatisticsCorrelation/Clustering Plot



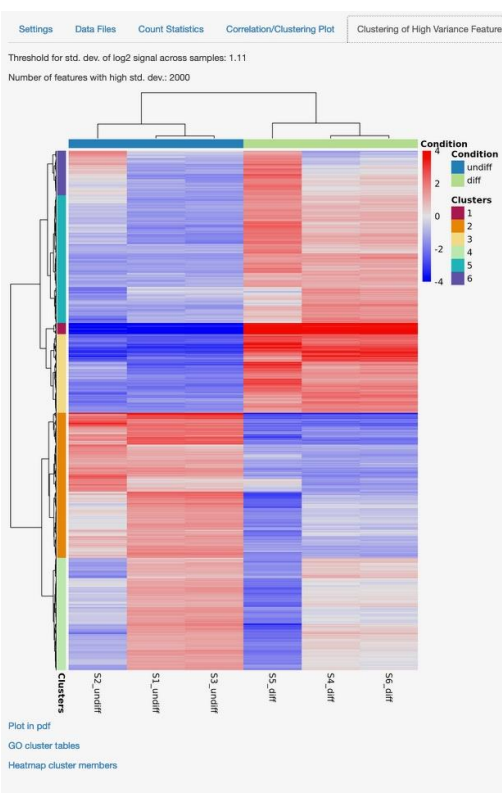
Count distribution



Feature distribution



Sample clustering



Genes with high variance

EdgeR (static report)

TwoGroupsAnalysis_Result

- Settings
- Result summary
- Inspection of significant genes
- Inspection of significant genes (Advanced plots)

	Number
Number of features:	21505
Number of features with counts above threshold:	13685

Number of significant by p-value and fold-change

	#significant	FDR	fc >= 1	fc >= 1.5	fc >= 2	fc >= 3	fc >= 4	fc >= 8	fc >= 10
p < 0.1	5432	0.2519000	5432	5348	3829	1972	1331	596	473
p < 0.05	4295	0.1592000	4295	4295	3554	1962	1330	596	473
p < 0.01	2642	0.0517500	2642	2642	2575	1824	1305	596	473
p < 0.001	1396	0.0097990	1396	1396	1396	1330	1102	583	468
p < 1e-04	706	0.0019380	706	706	706	706	691	508	427
p < 1e-05	344	0.0003963	344	344	344	344	344	332	309

Full result table in xlsx format for opening with a spreadsheet program (e.g. Excel).

- result-diff-over-undiff.xlsx
- Live Report and Visualizations

Differentially expressed gene list

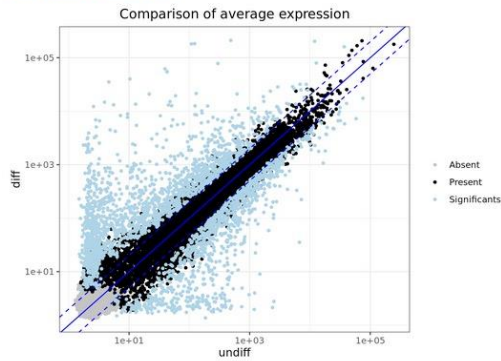
- Settings
- Result summary
- Inspection of significant genes
- Inspection of significant genes (Advanced plots)

Between-group comparison

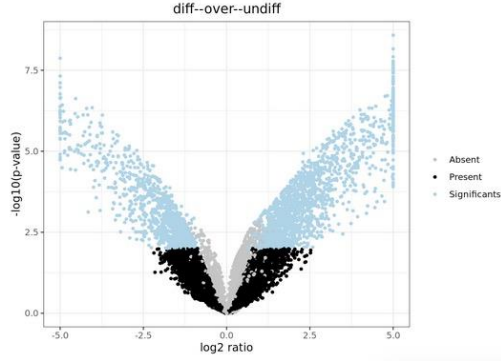
	Number
P-value threshold:	p <= 0.01
Log ratio threshold:	log ratio >= 0.5
Number of significant genes:	2642

p <= 0.01
(reduced FDR)

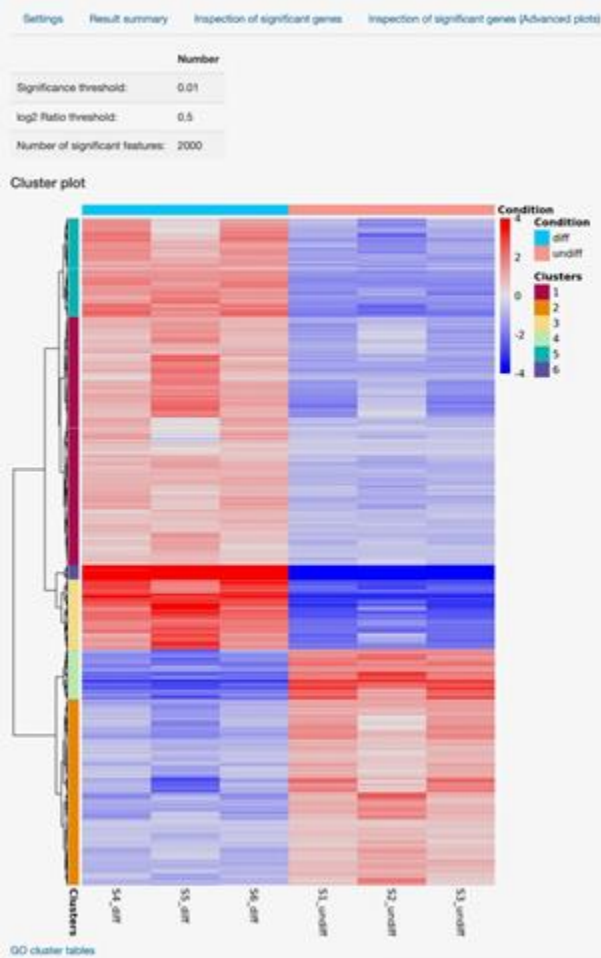
Subsequent plots highlight significant genes in blue.
Interactive table of significant genes



Interactive comparison plot (64-bit Chrome or Safari is recommended for best performance)



EdgeR (static report)



Heat map of DE genes

GO categories of feature clusters

	BP	MF	CC	Enrichment
Cluster 1	Cluster-BP-1.html	Cluster-MF-1.html	Cluster-CC-1.html	Enrichment-1
Cluster 2	Cluster-BP-2.html	Cluster-MF-2.html	Cluster-CC-2.html	Enrichment-2
Cluster 3	Cluster-BP-3.html	Cluster-MF-3.html	Cluster-CC-3.html	Enrichment-3
Cluster 4	Cluster-BP-4.html	Cluster-MF-4.html	Cluster-CC-4.html	Enrichment-4
Cluster 5	Cluster-BP-5.html	Cluster-MF-5.html	Cluster-CC-5.html	Enrichment-5
Cluster 6	Cluster-BP-6.html		Cluster-CC-6.html	Enrichment-6

Note:

Note:

Cluster font color corresponds to the row colors in the heatmap plot.

Rows: features; **Columns:** samples

Cluster colors: GO category table of feature clusters

EdgeR (static report)

Enrichment analysis

- MetaCore:** drug/disease related information (email sequencing@fgcz.ethz.ch for license)
- Enrichr:** multiple annotations from various databases
- ORA (Over Representation Analysis):** GO enrichment for specific set of genes
- GSEA (Gene Set Enrichment Analysis):** GO enrichment score for all genes

Technical bias

SettingsResult summaryInspection of significant genesInspection of significant genes (Advanced plots)Clustering of significant featuresMetaCoreEnrichrOverrepresentation Analysis (ORA)Gene set enrichment analysisTechnical bias

We define 4 gene sets

- high GC: the 5% of the genes with the highest GC content
- low GC: the 5% of the genes with the lowest GC content
- long genes: the 5% of the genes with the biggest length
- short genes: the 5% of the genes with the smallest length

And we test if the up- or down-regulated genes are associated with one of those gene sets. If there is a significant association, some of the significant genes are potentially false positives due to a technical bias.

Tests where the association p-value is below 0.001 are highlighted in red. The column "overlapping/total genes" shows the number of overlapping genes and the total number of genes in that category.

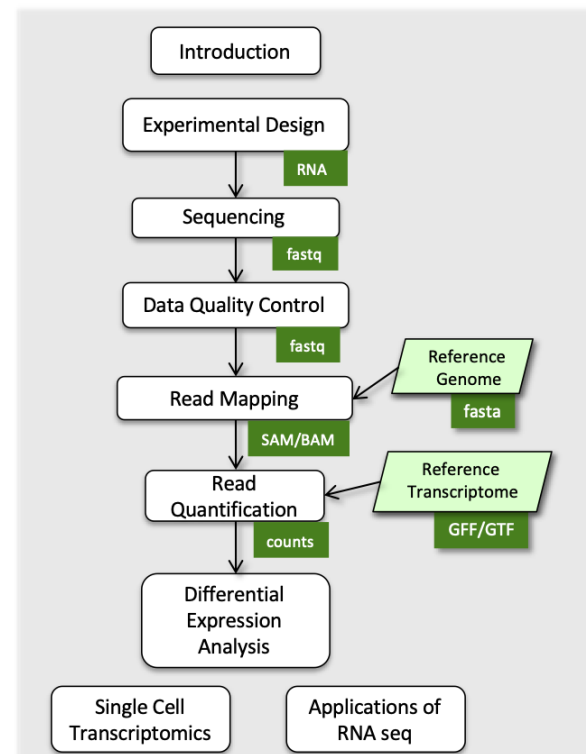
Association test: 1637 up-regulated and 1005 down-regulated genes

	overlapping/total genes	odds ratio	p-value
low GC – Up-regulation	39/685	0.431	1.00e+00
low GC – Down-regulation	55/685	1.107	2.60e-01
high GC – Up-regulation	91/684	1.137	1.47e-01
high GC – Down-regulation	15/684	0.272	1.00e+00
short genes – Up-regulation	78/685	0.943	7.01e-01
short genes – Down-regulation	39/685	0.752	9.66e-01
long genes – Up-regulation	107/683	1.393	1.85e-03
long genes – Down-regulation	75/683	1.601	2.65e-04

Re-run the app with correctBias = true in specialOptions field
Compare the results

Summary: RNAseq (differential gene expression analysis) workflow

1. **FastQC**: Quality control
2. **STAR**: read alignment
 - *Kallisto*: the second option
3. **FeatureCounts**: estimate reads abundance
4. **CountQC**: Quality control, multivariate analysis
5. **edgeR**: differential expressed genes detection
 - *DESeq2*: the second option



Question?