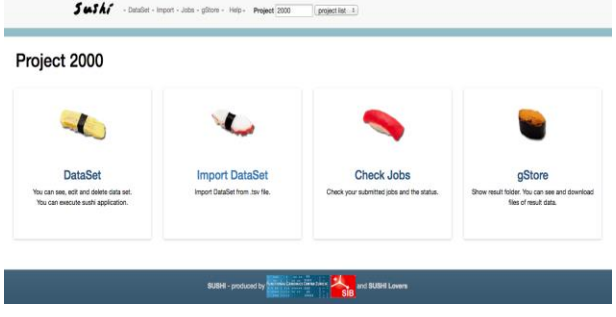
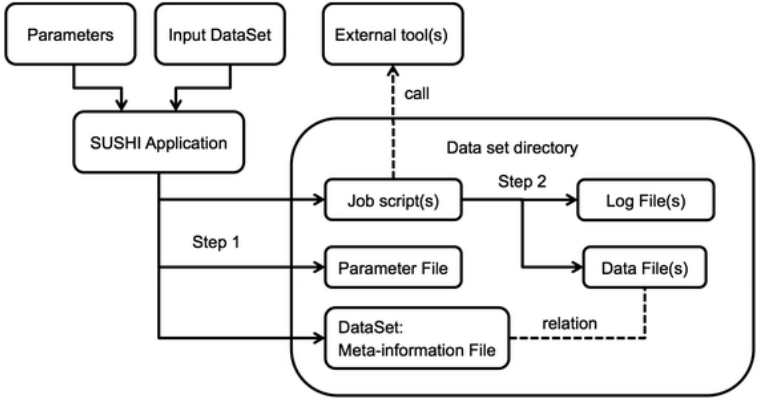
SUSHI can Support yoU by SHell-script Integration

Masaomi Hatakeyama
University of Zurich / Functional Genomics Center Zurich
25 Dec. 2025, 14:00-





University of Zurich[UZH]
ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

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	**Lecutre/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**

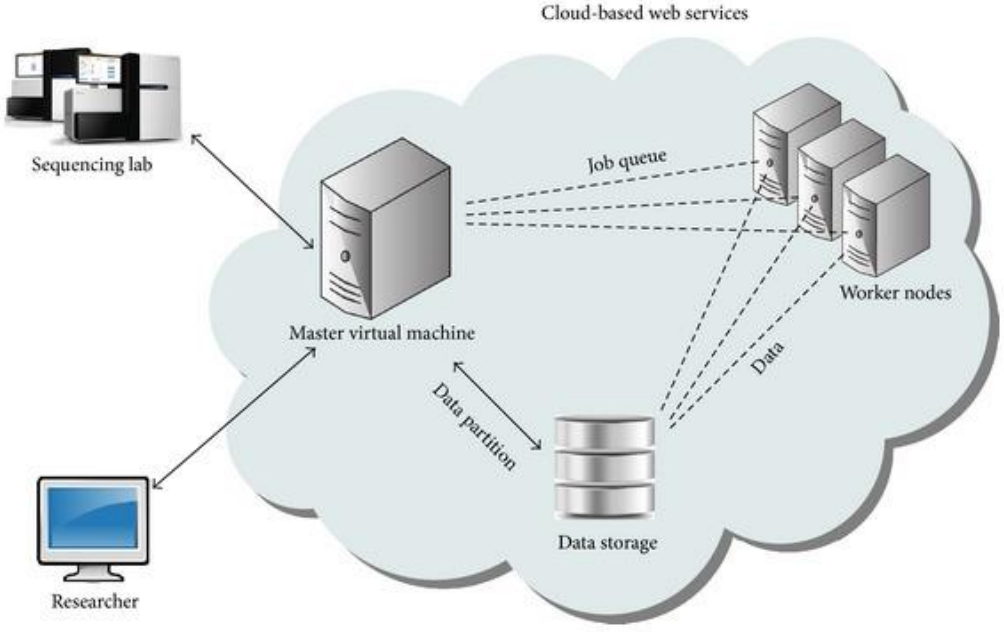
https://github.com/masaomi/bio610_2025_sushi_practice

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

- Submit the job script to cluster

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

- 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
```

User friendly NGS data analysis pipelines

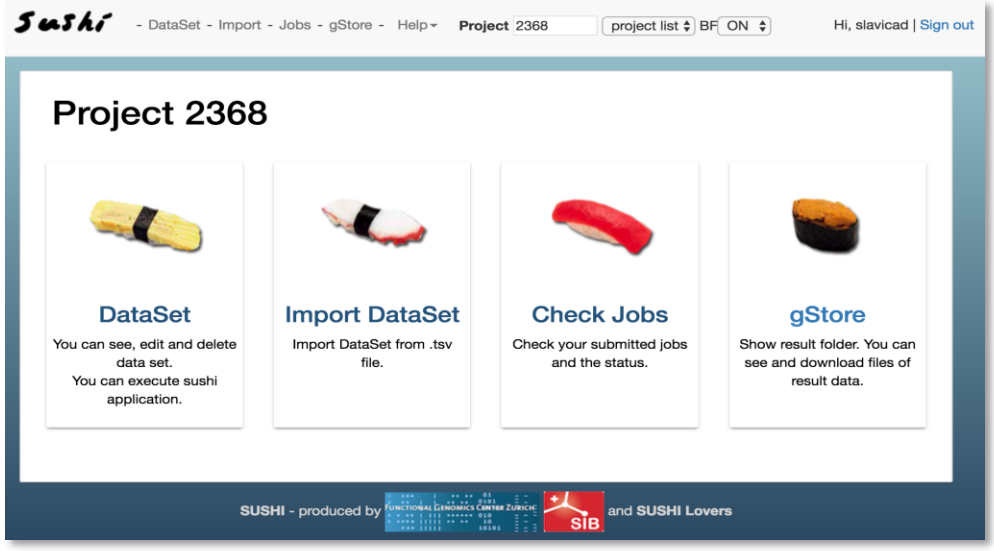
- Graphical User Interface (GUI)
- Commercial packages
  - CLC Genomics Workbench
- Scientific workflow managers/data analysis frameworks
  - **SUSHI**
  - Galaxy
  - Chipster
  - …

- Pros
  - Easy to use
  - Flat learning curve
  - Documented, reproducible workflows
  - Usually implemented following the best-practice within community
- Cons
  - Limited by supported algorithms/modules

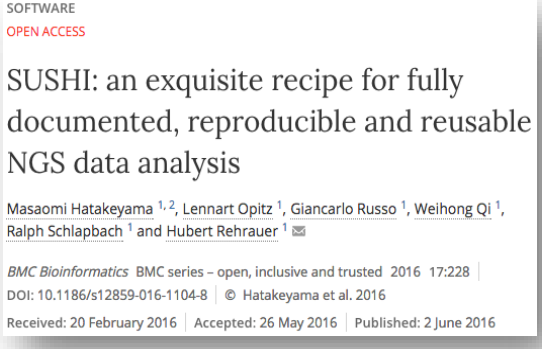
Why SUSHI?

SUSHI can **S**upport yo**U** by **SH**ell-script **I**ntegration

- *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 Opitz [1], Giancarlo Russo [1], Weihong Qi [1], Ralph Schlapbach [1] and Hubert Rehrauer [1]

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

DataSets - delete selected

Search:

- 2 HiSeq_short_dataset_20200609
  - 1 Kallisto_55548
    - 0 CountQC_55550
  - 0 Fastqc_55548

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

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	Treate		20130830.A-hal_w...
Ahal_L48hZ3	Treated		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
ATAC	AtacENC
Assemble	CanuApp SpadesApp TrinityApp
Count	KallistoApp RSEMApp
Map	BWAApp BismarkApp Bowtie2App Bowtie2TranscriptomeApp BowtieApp STARApp
Metagenomics	DADA2Step1SampleApp KrakenApp MegahitApp MetaspadesApp MothurApp VirDetectApp

2. SUSHI App

Question?

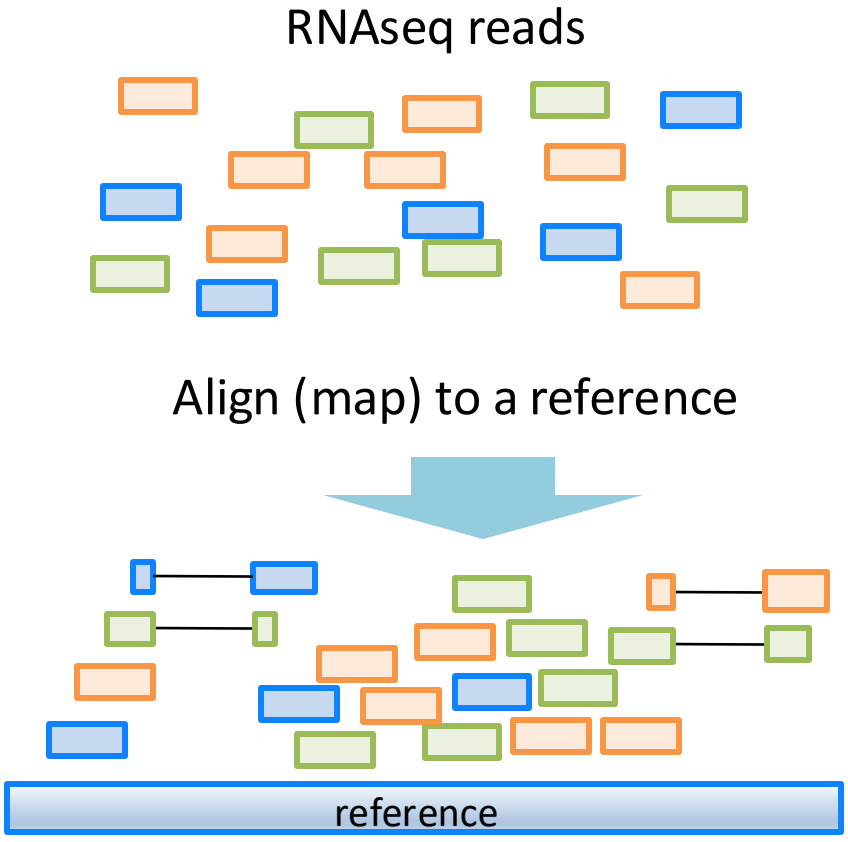
Demonstration
RNAseq analysis

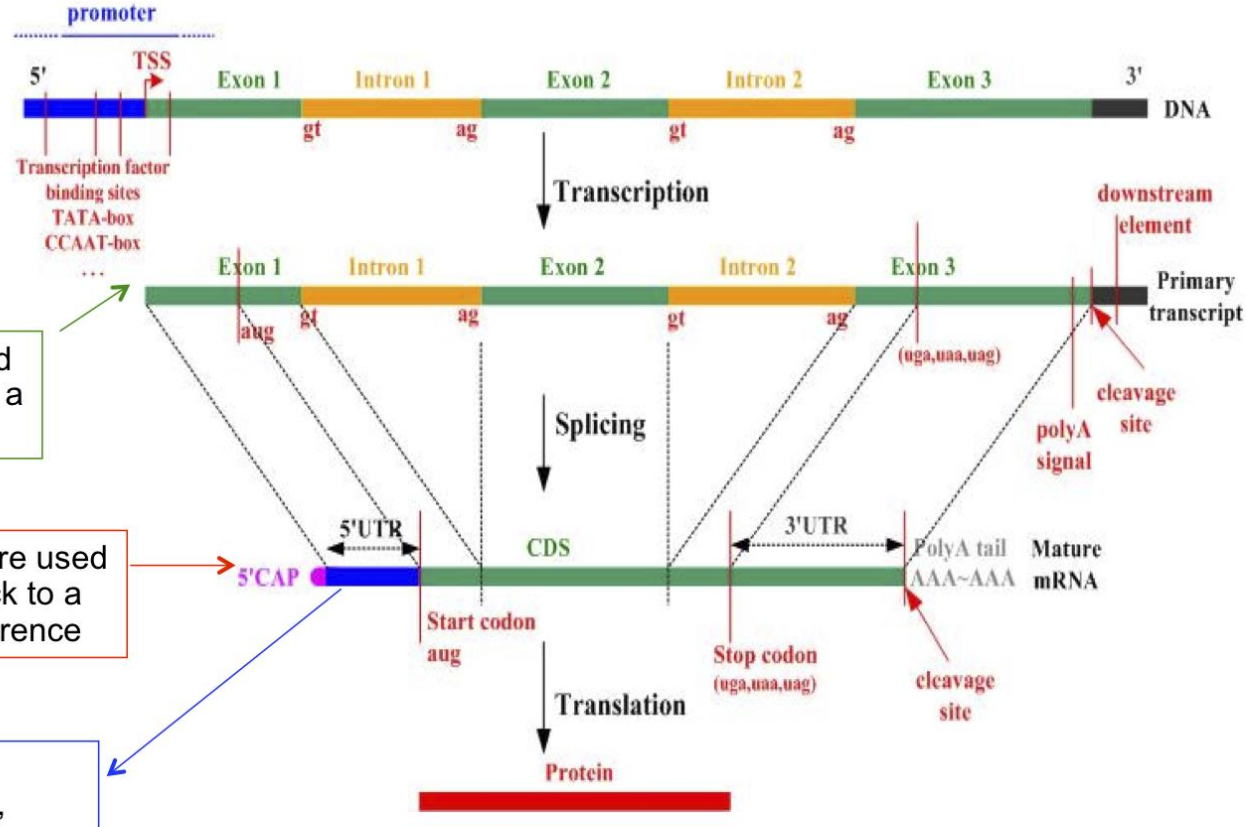
RNAseq analysis, transcriptomics, gene expression analysis

RNAseq analysis core steps

1. Align (map) a lot of short reads (transcripts) onto a referece sequence (genome, transcriptome)

2. Count abundance of each gene (transcript)
   =Estimate expression level



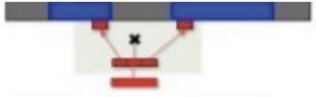


Spliced aligners are used to mapped them back to a GENOME reference

Unspliced aligners are used to mapped them back to a TRANSCRIPTS reference

In RNA-seq these are sequenced, i.e., contiguous sequences of exons

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

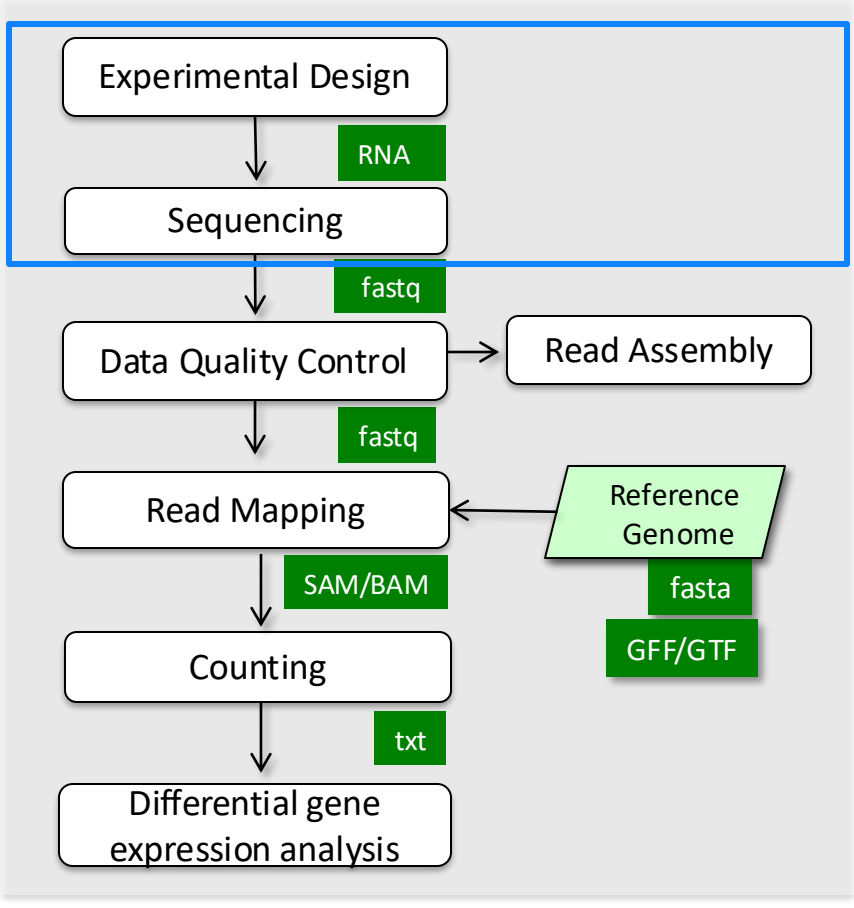
- Un-spliced mappers (DNA-Seq)
  - Align continuous reads (not containing gaps as a result of splicing)
  - When encountering an intron, the aligner **stops and trims the rest of the read**
  - BWA, **BOWTIE2**, BLAST
  - Splice junctions are impossible to detect

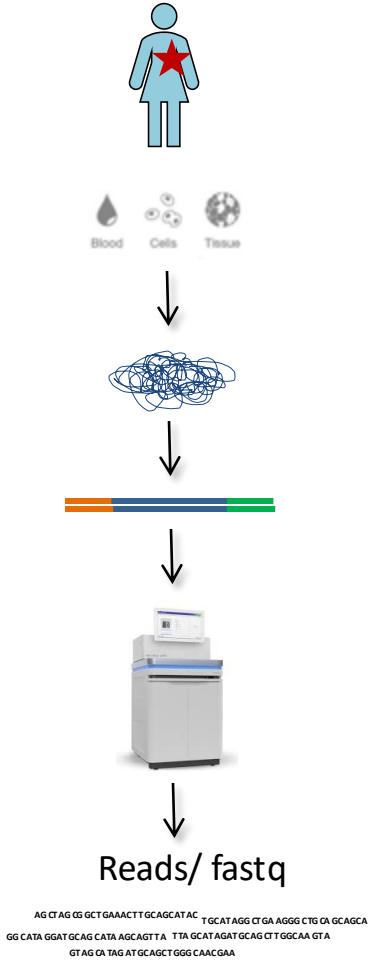
Sequence / Alignment (SAM) files

- **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

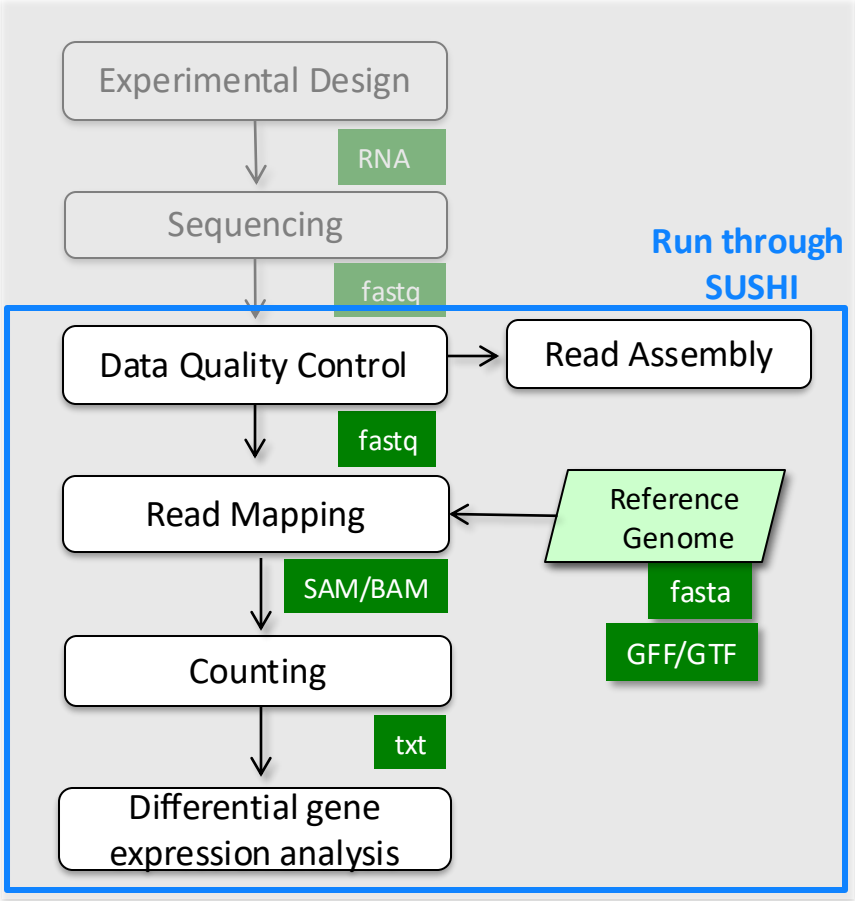
- **BAM (Binary Alignment/Map)**
  - Binary equivalent of SAM
  - **Compressed data** plus index (bai)
  - Developed for fast processing/indexing
  - RNA-Seq: >2GB
  - Exome: 2-10GB, Whole Genome: 100-300GB

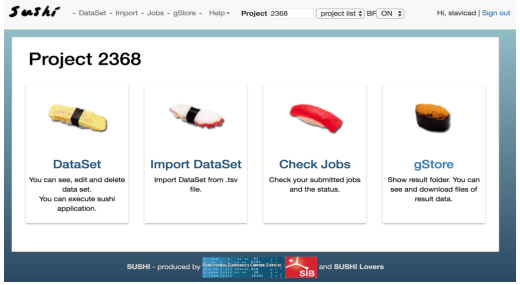
Workflows (A typical RNAseq workflow)





Workflows (A typical RNAseq workflow)

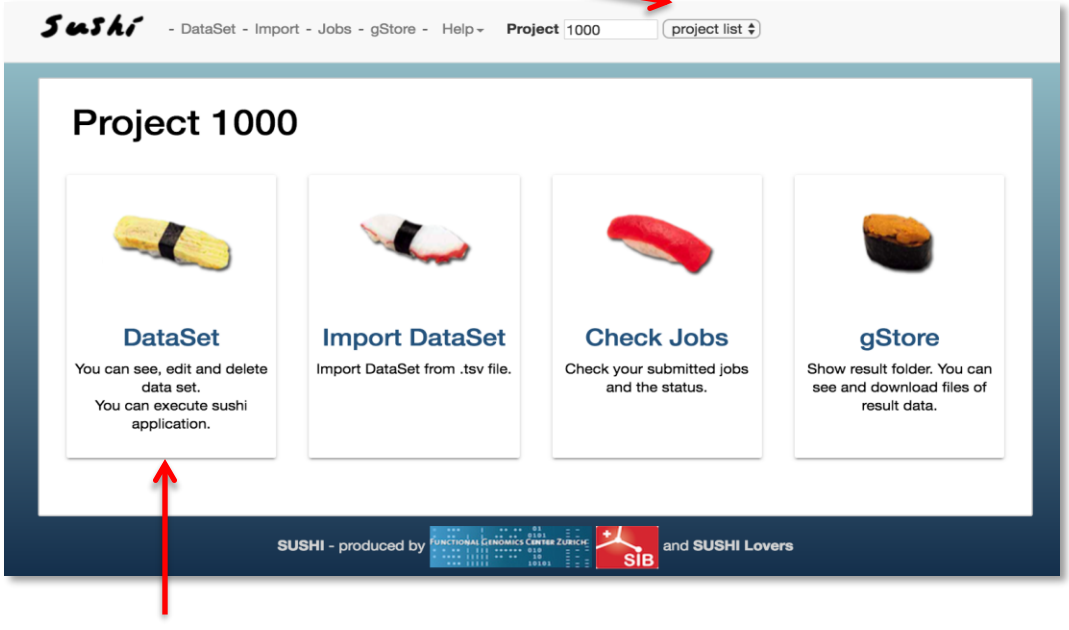




https://fgcz-sushi.uzh.ch/

SUSHI (course) server

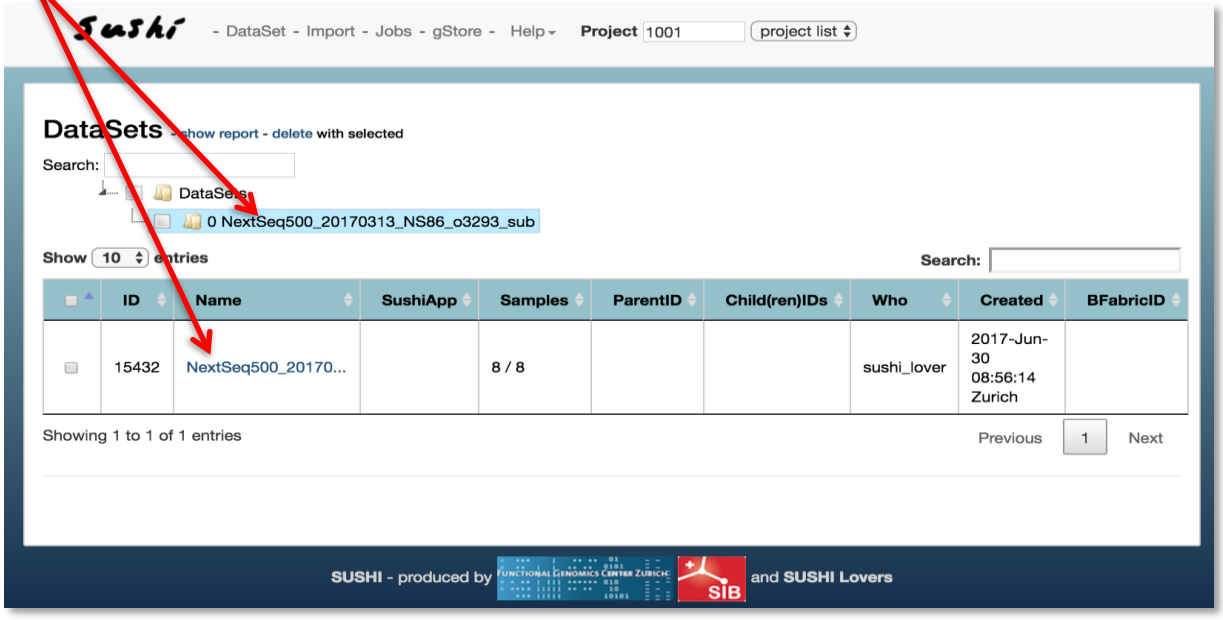
1. SUSHI course Server: http://fgcz-course1.bfabric.org
2. Select your project



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

1. Select a DataSet

Click a DataSet ether in the tree or in the table



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

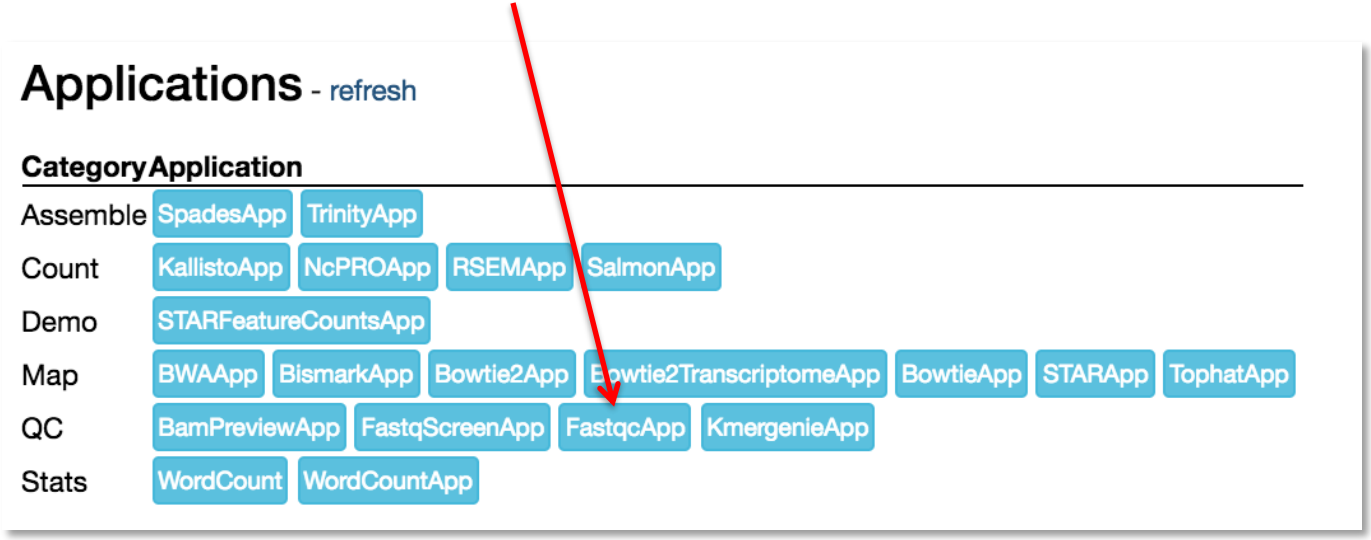
HiSeq_short_dataset_20200609 Content Hide ▲ - edit table - data folder

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

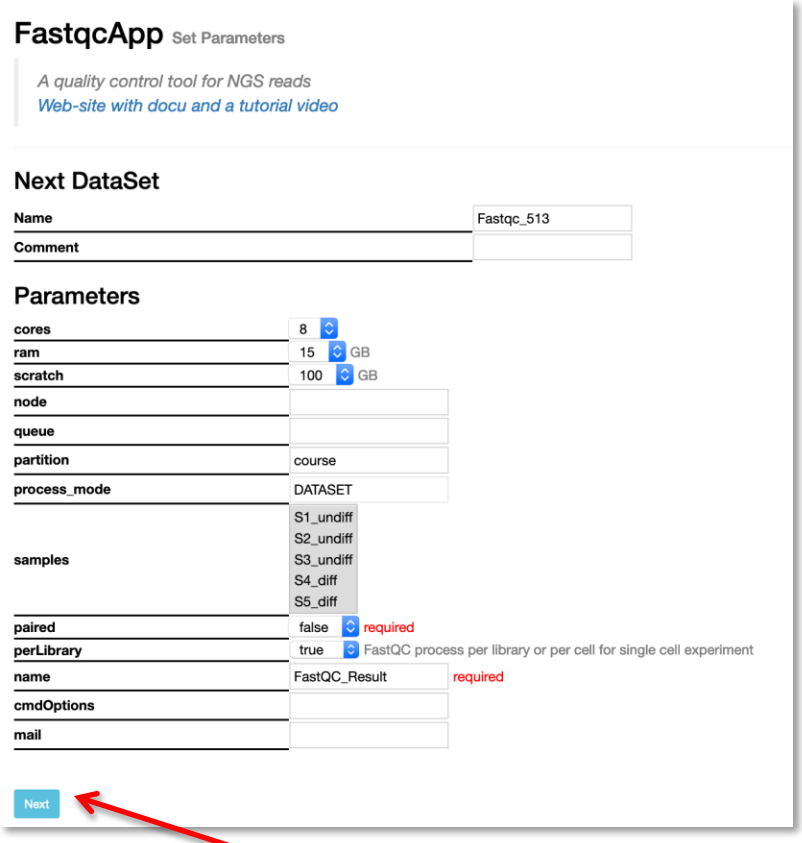
To run FastqcApp, click on "**FastqcApp**" in the Application section



Note

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

3. Set job parameters



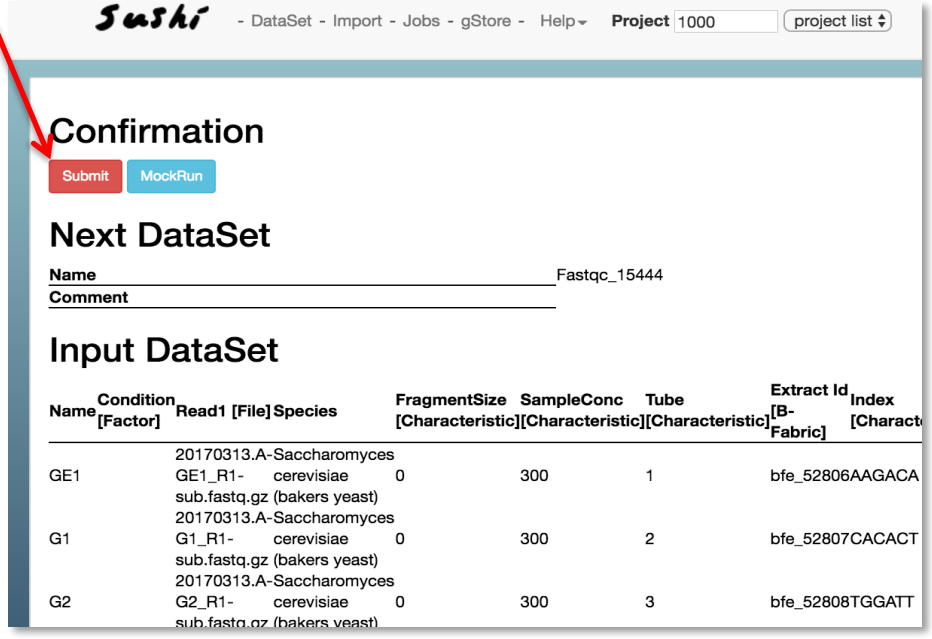
Common parameters

- ***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

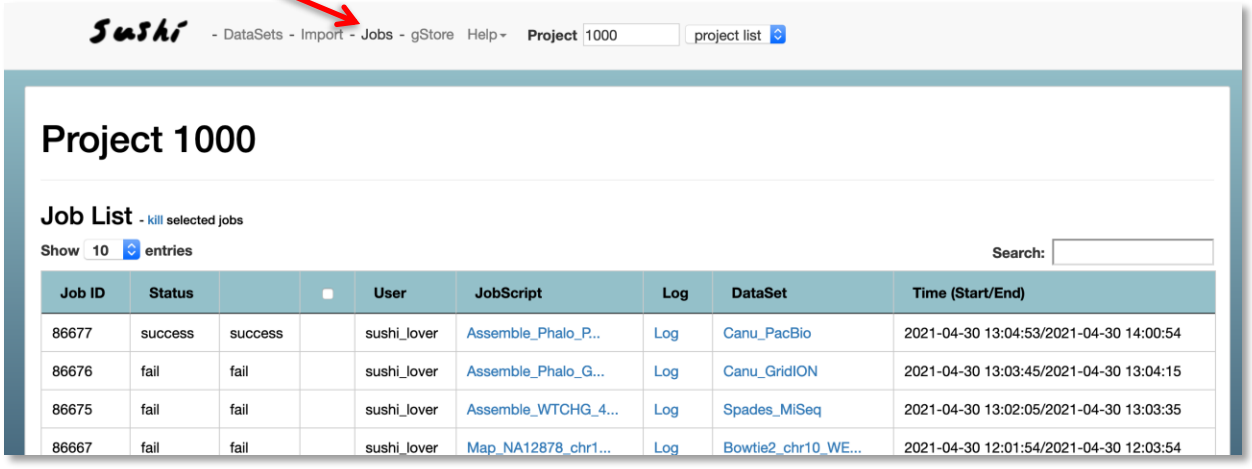


Note

- 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

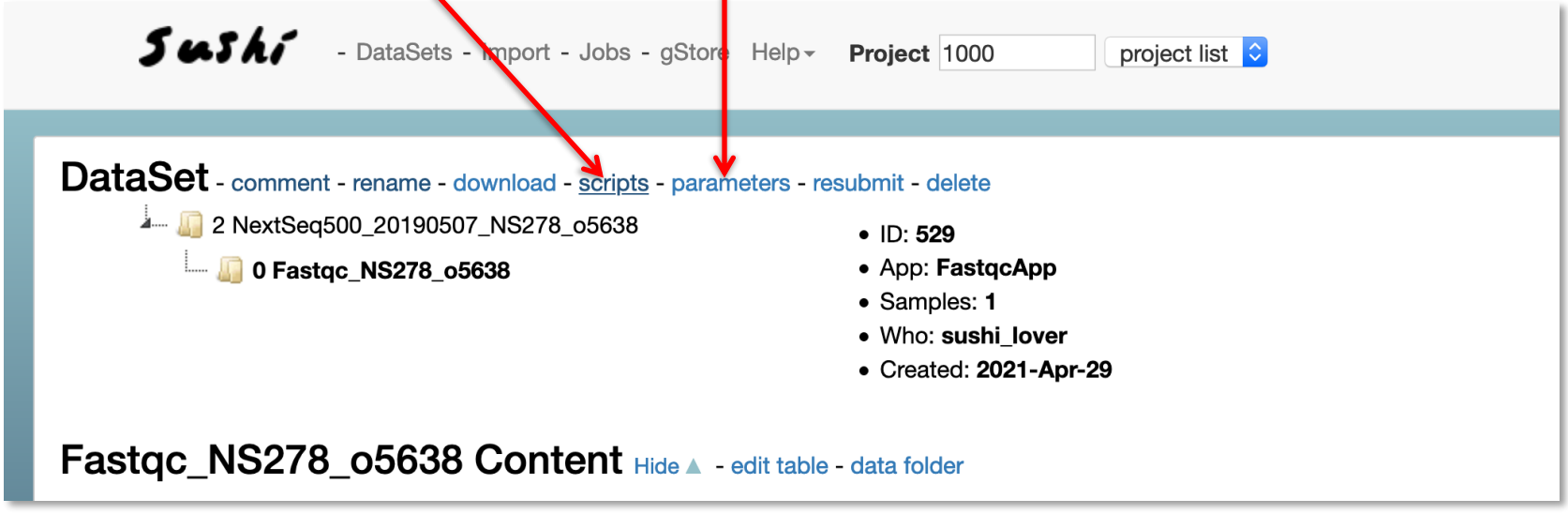


Job status

- *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**"



FastQC result

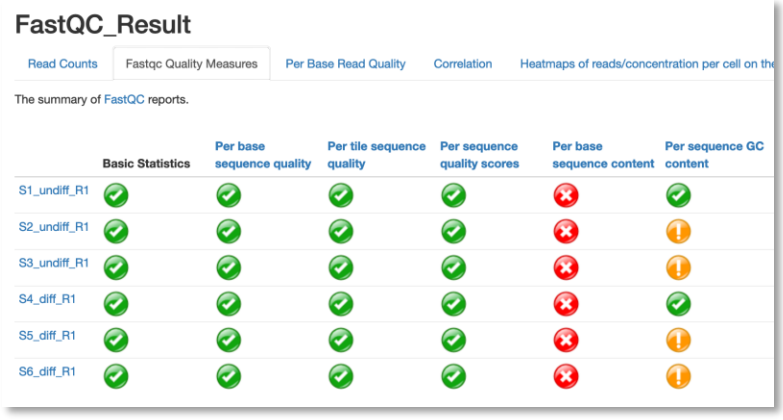
Two types of results: 1. **Normal FastQC result**, 2. **MultiQC result**

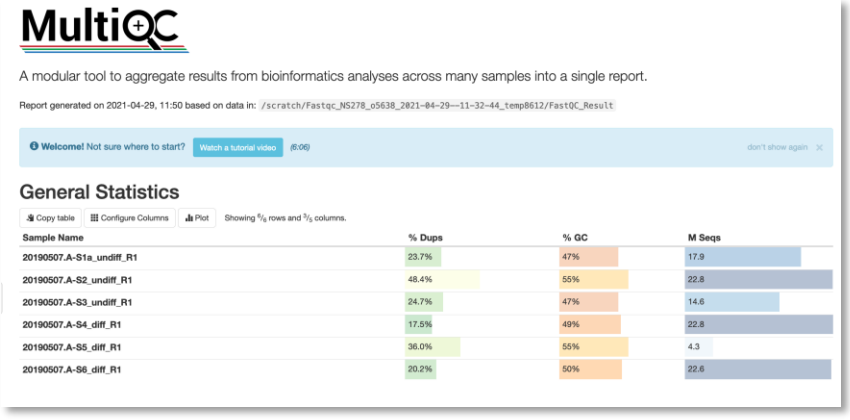
Fastqc_NS278_o5638 Content Hide ▲ - edit table - data folder

Search:

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

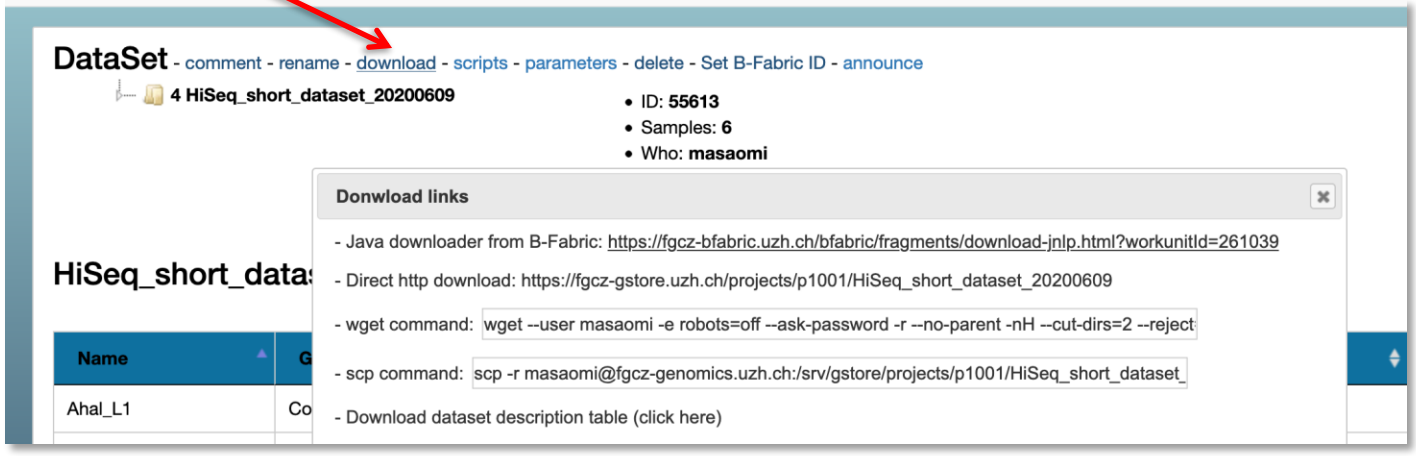
Showing 1 to 1 of 1 entries





Download result

"**Download**" menu shows several options

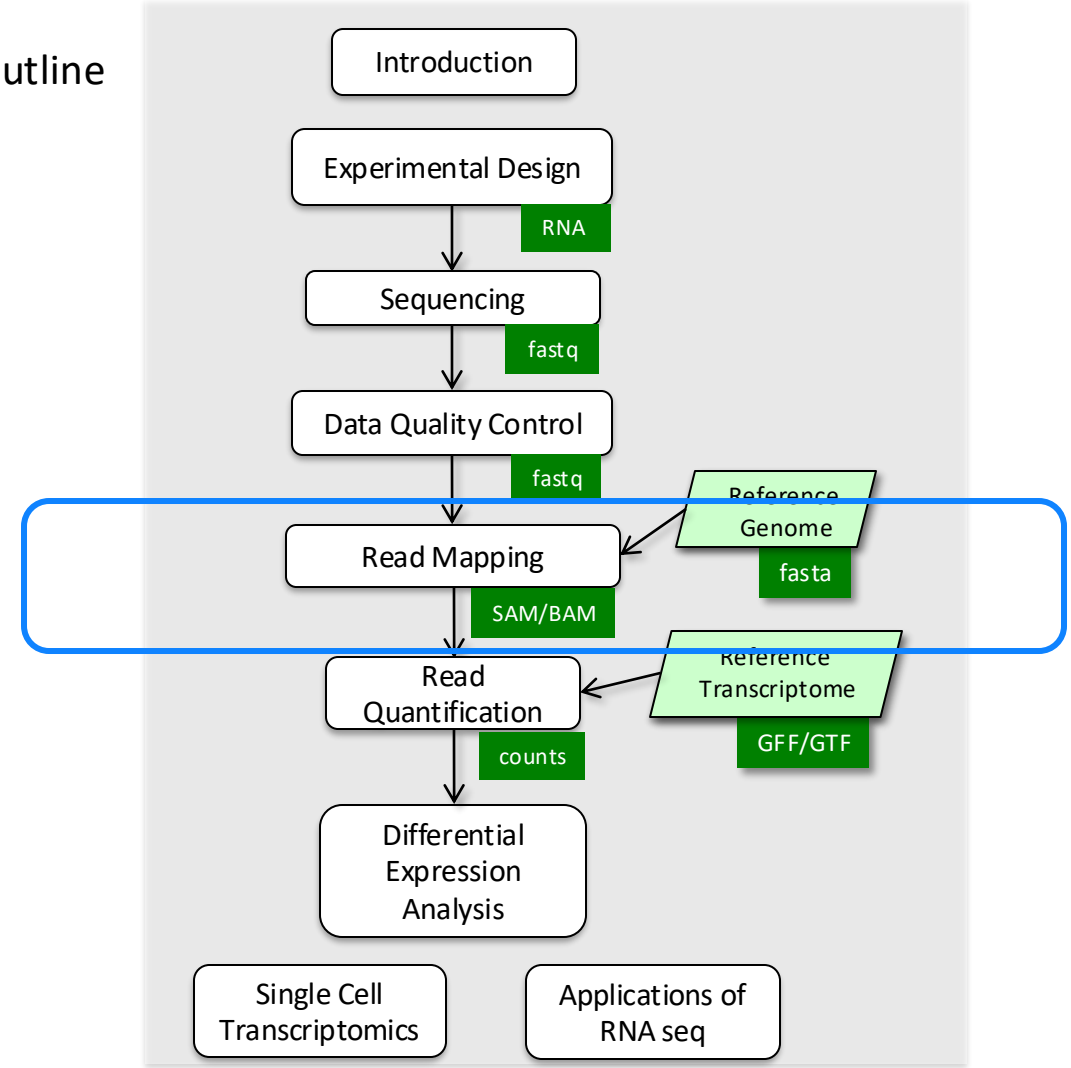


Options

- *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.)

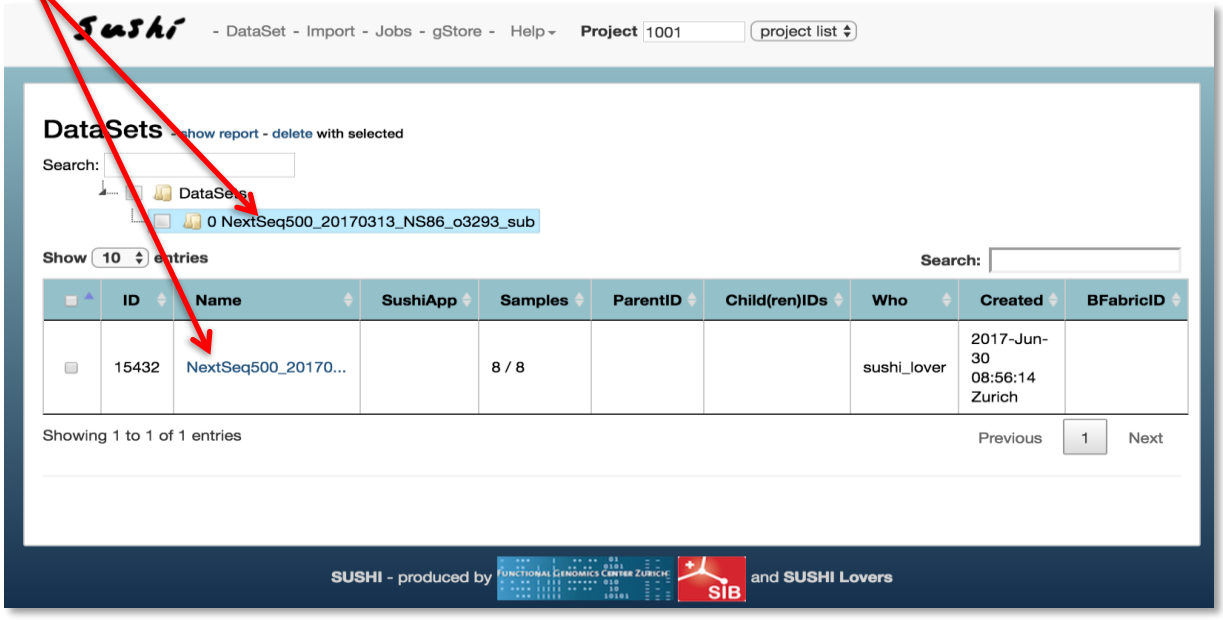
Question?

Outline


Introduction
Experimental Design
RNA
Sequencing
fastq
Data Quality Control
fastq
Reference Genome
fasta
Read Mapping
SAM/BAM
Reference Transcriptome
GFF/GTF
Read Quantification
counts
Differential Expression Analysis
Single Cell Transcriptomics
Applications of RNA seq

1. Select a DataSet

Click a DataSet ether in the tree or in the table

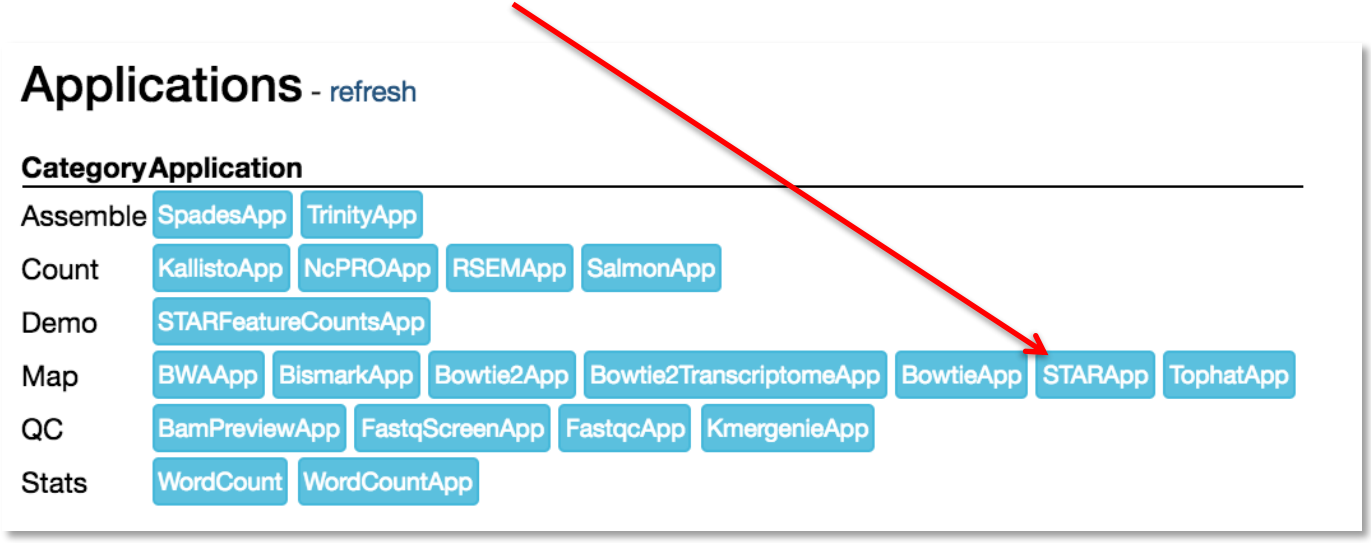


Note

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

2. Select a SUSHI Application

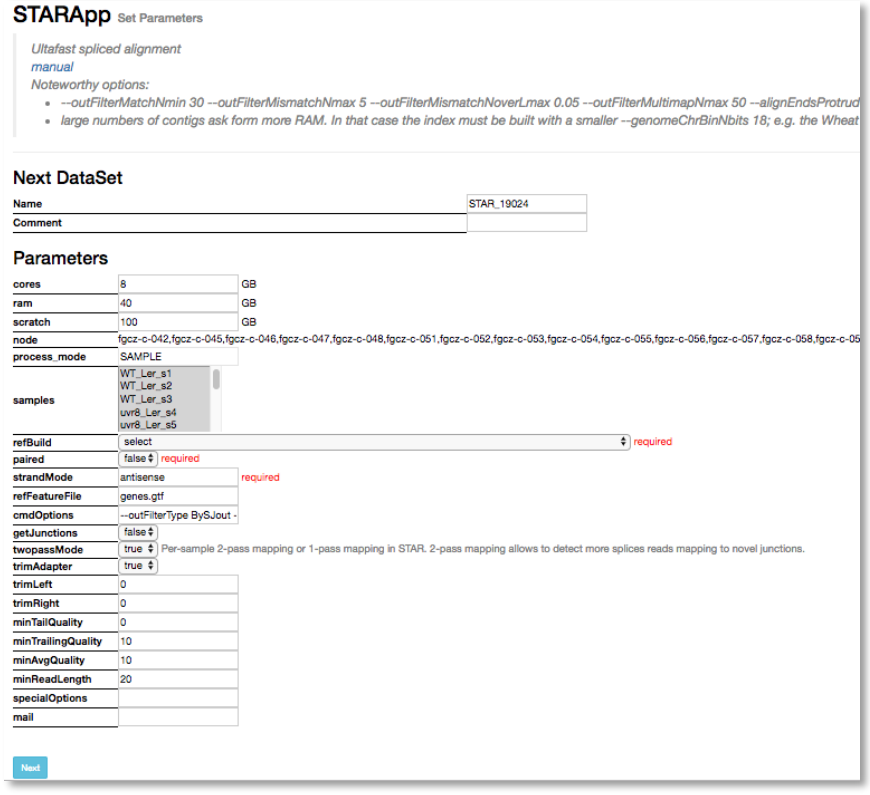
To align the reads, click on "**STARApp**" in the Application section



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 critical parameters

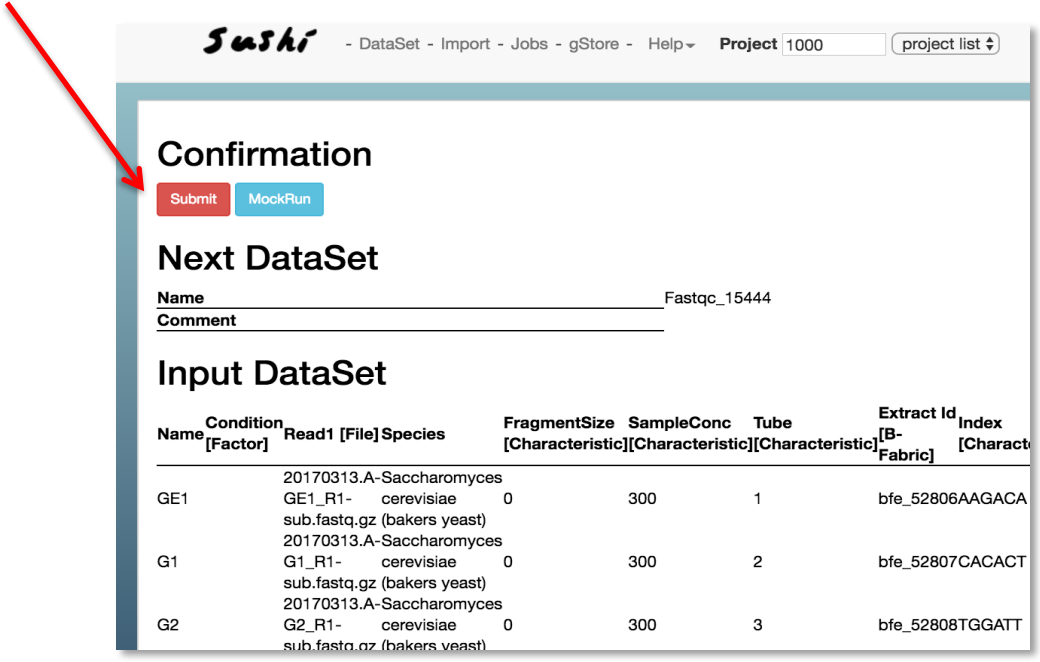
- ***refBuild***: reference genome
- ***Paired***: true(Paired-end)/false (Single-end)
- ***StrandMode***: (only in single-end)
- ***TrimAdapter***: true/false

Note

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

4. Submit a job

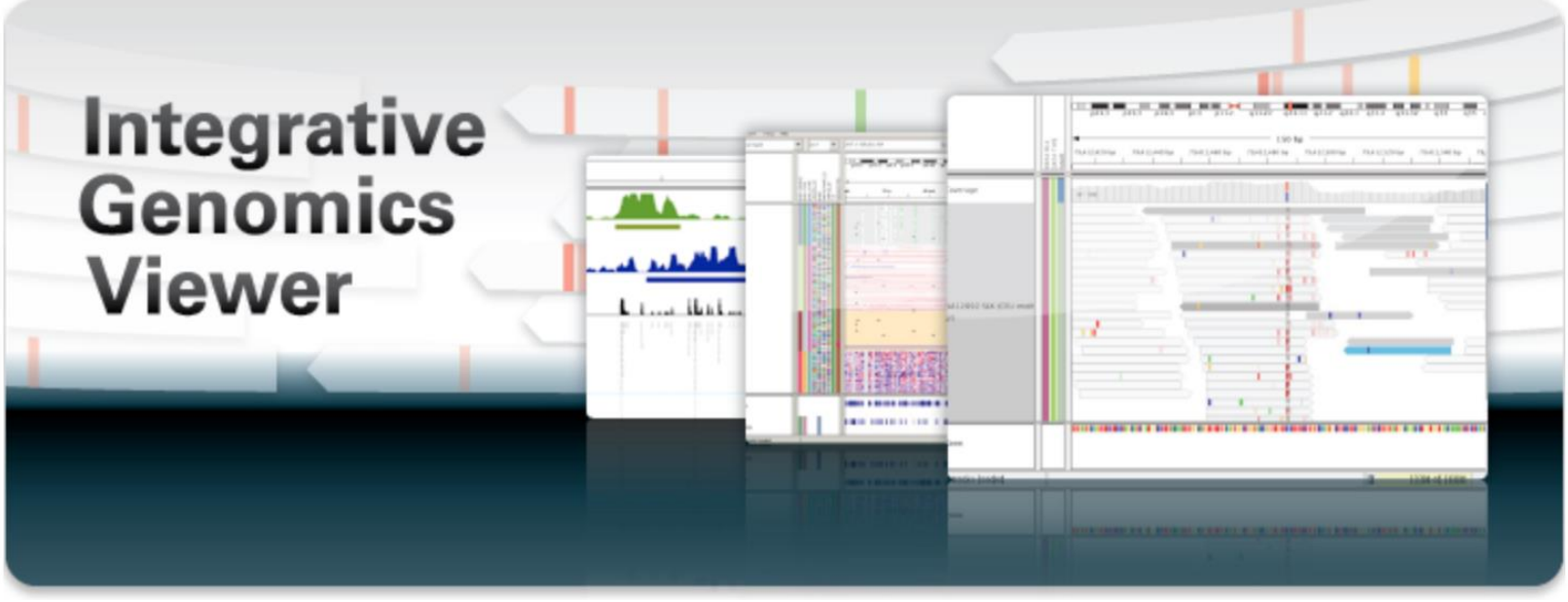
Start the actual job with ‚**Submit**'



Visualization
Mapping result

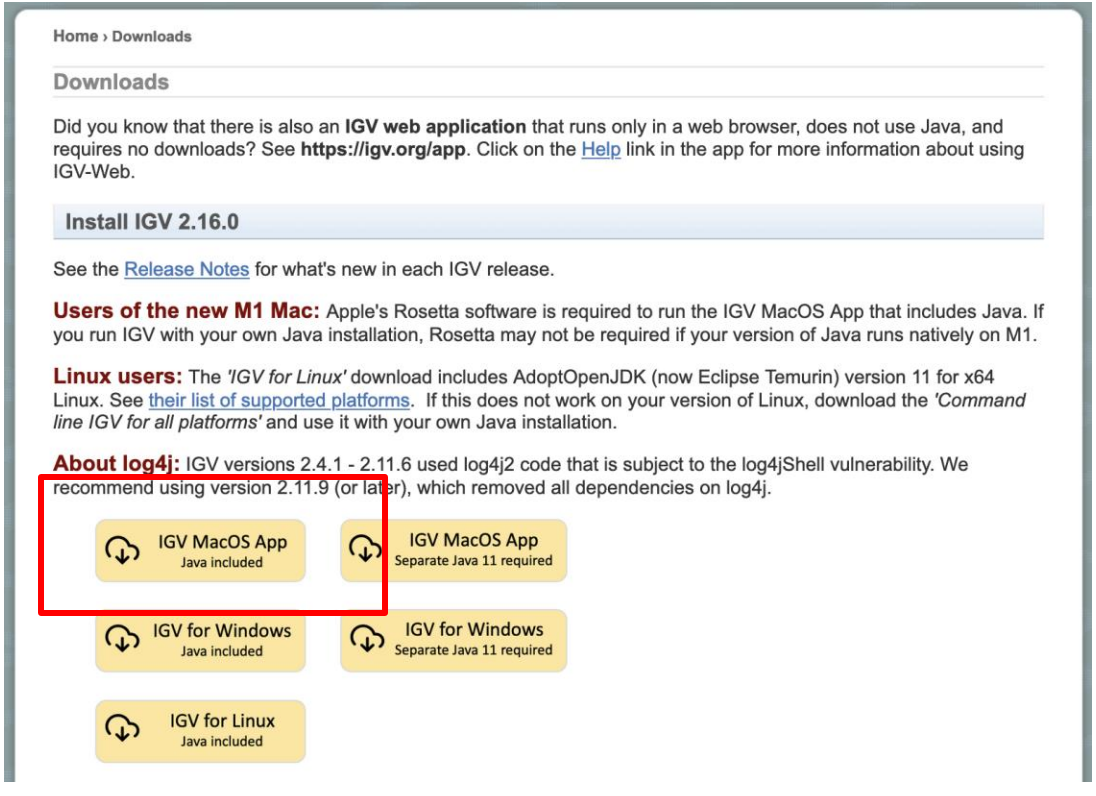
Mapping visualization

- IGV (Integrative Genomics Viewer)
  https://software.broadinstitute.org/software/igv/



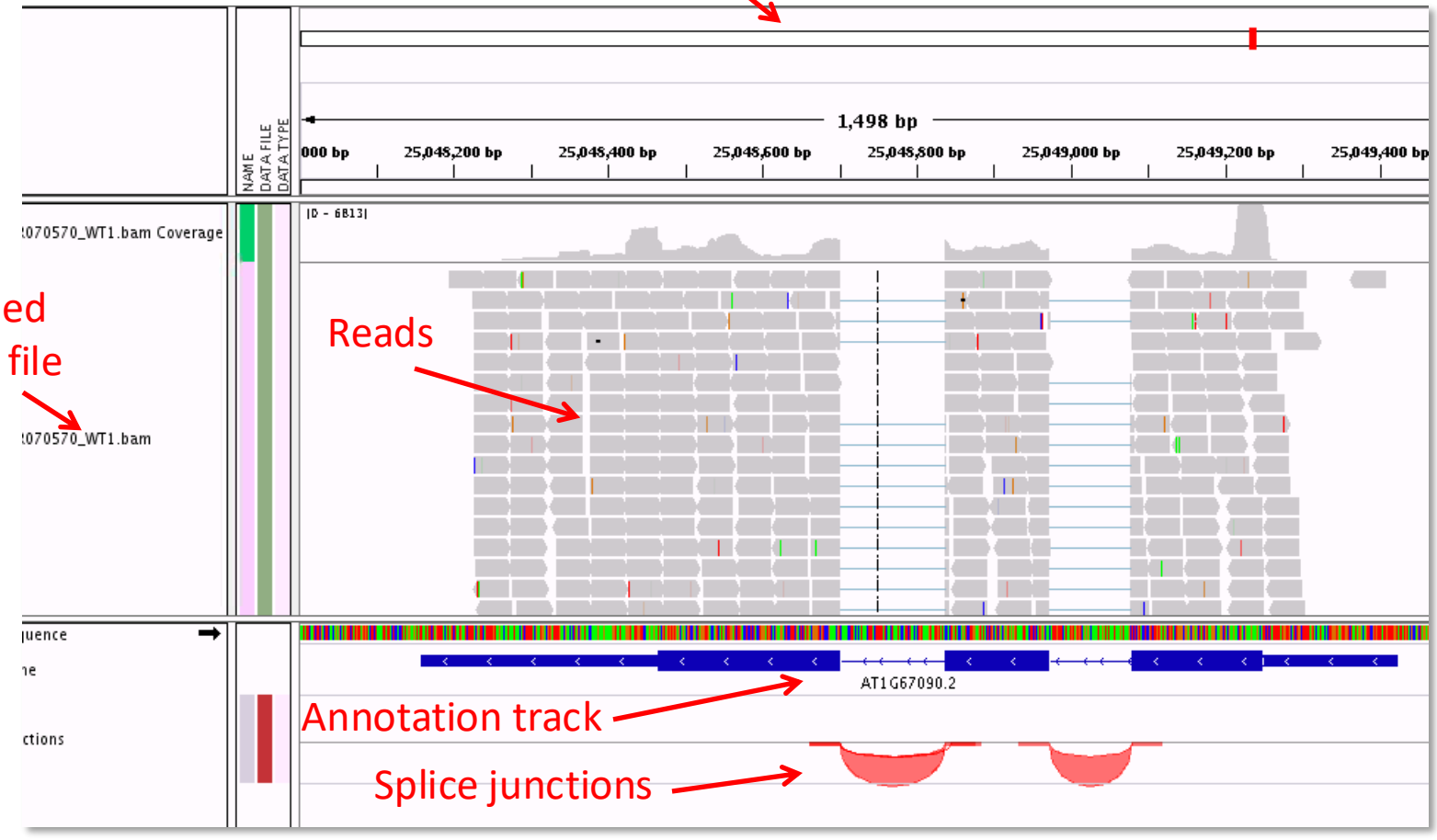
Download

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

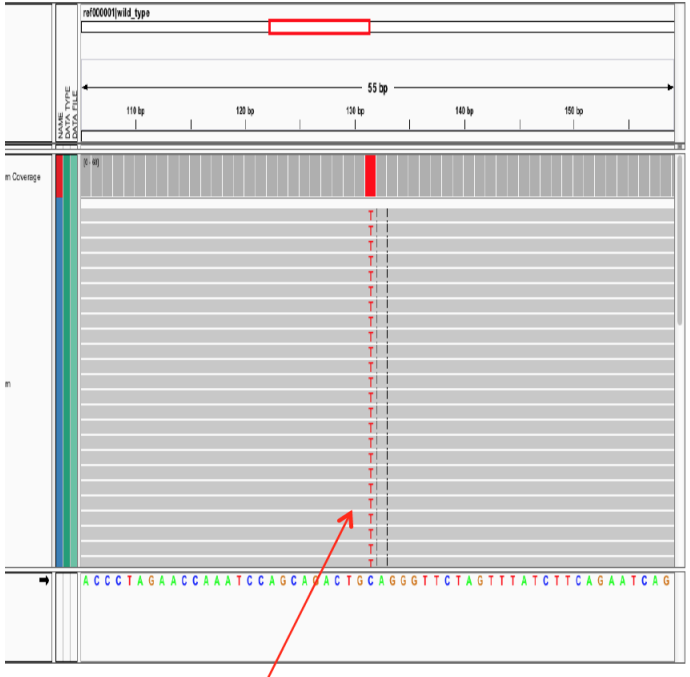


Short demo

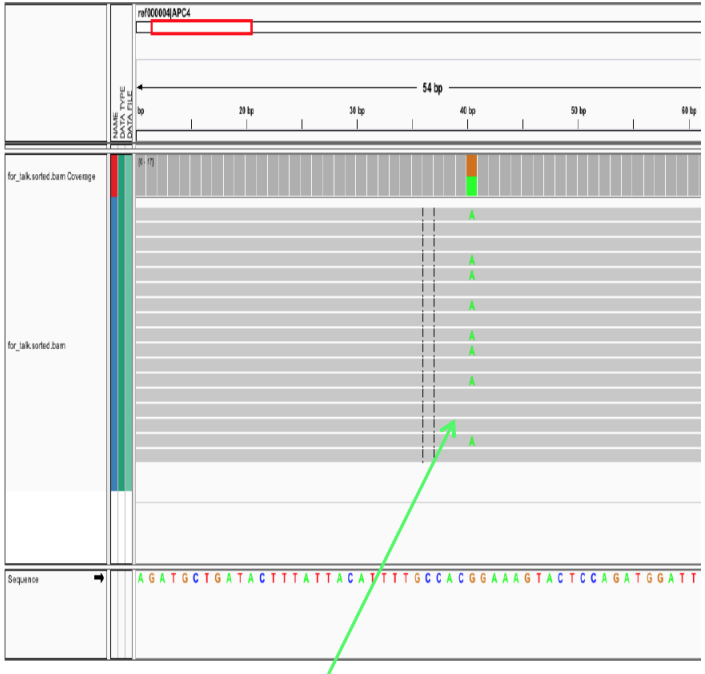
Mapped reads in IGV



SNPs

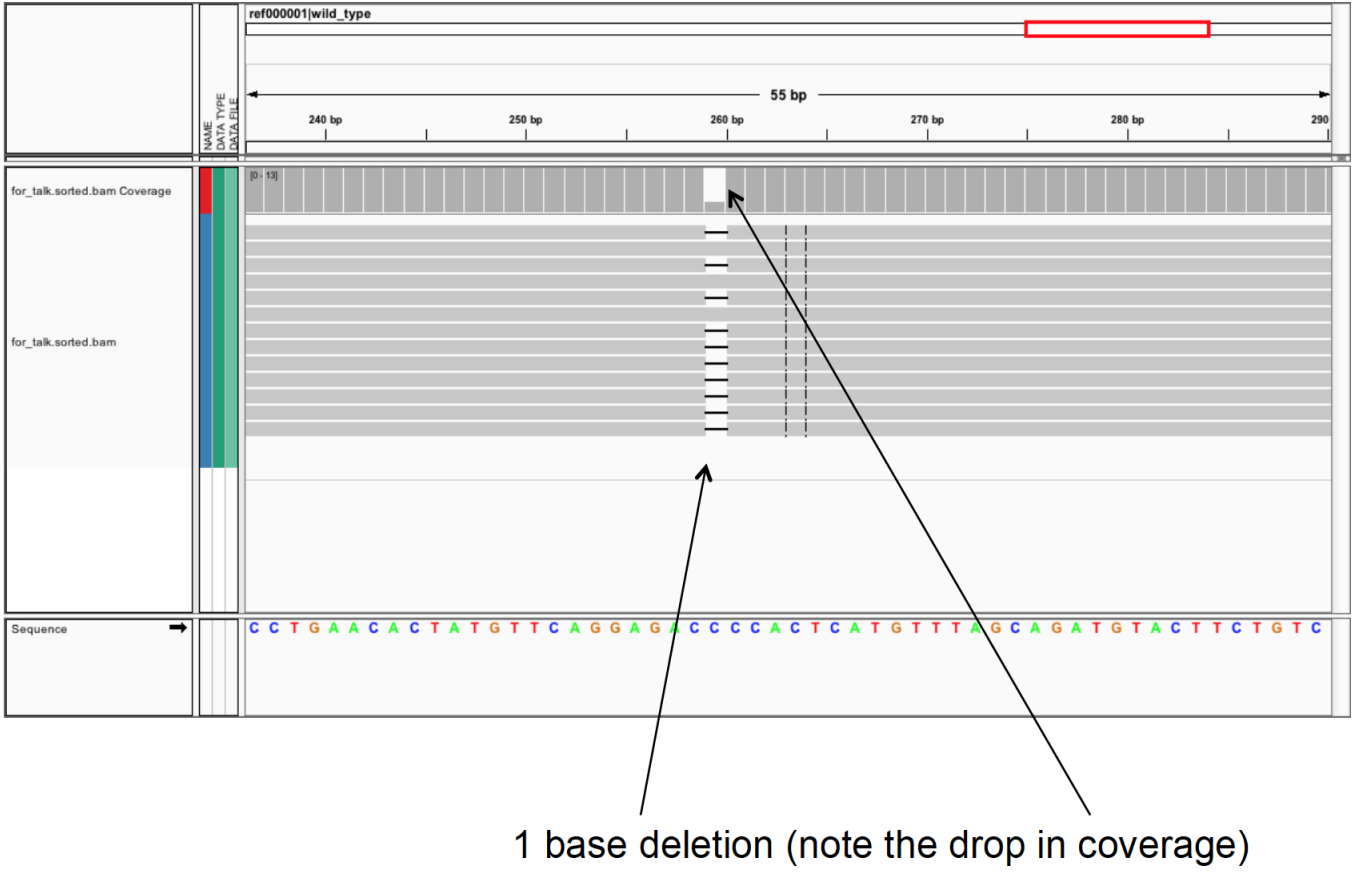


C > T homozygous substitution

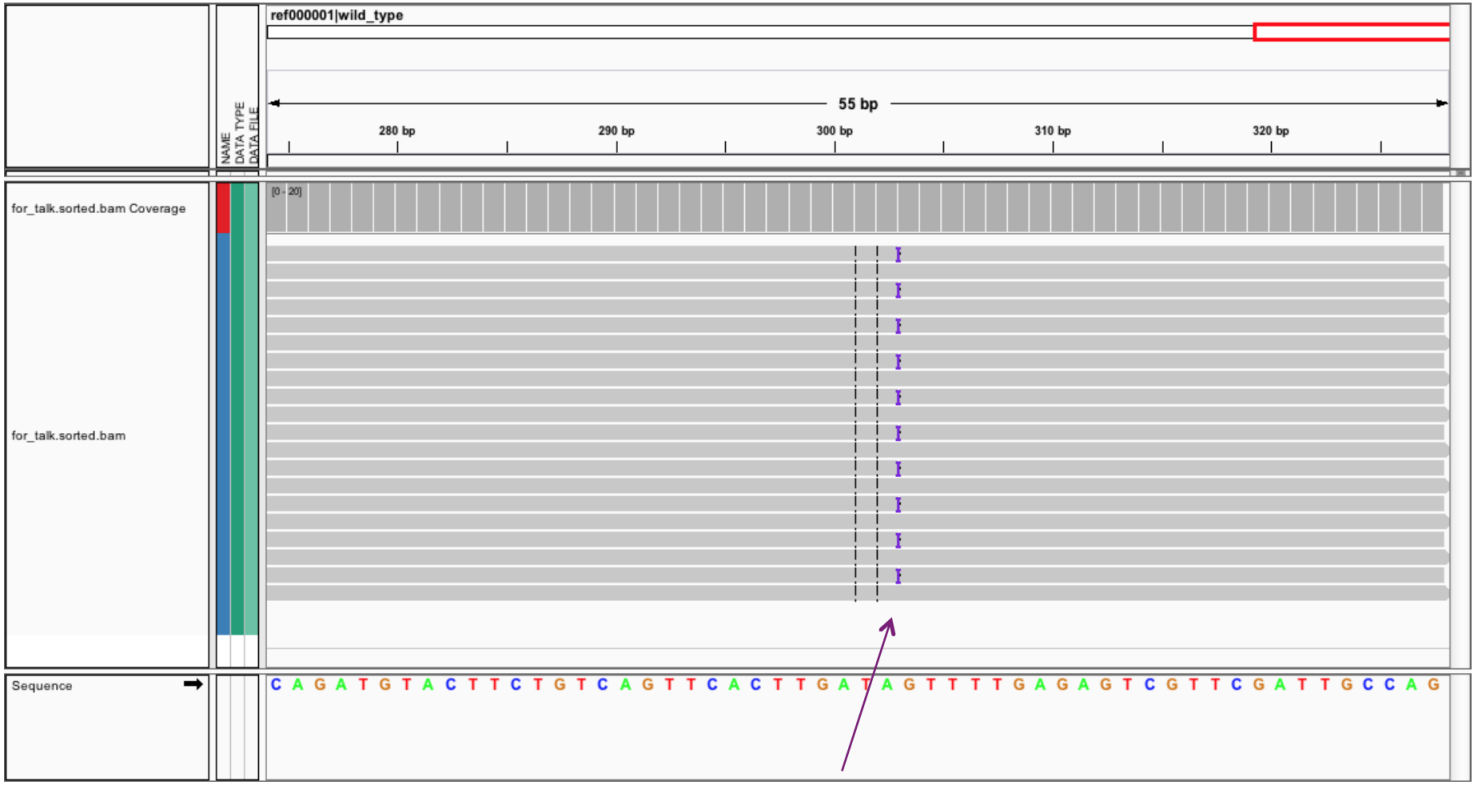


G > A heterozygous substitution

Single base deletions



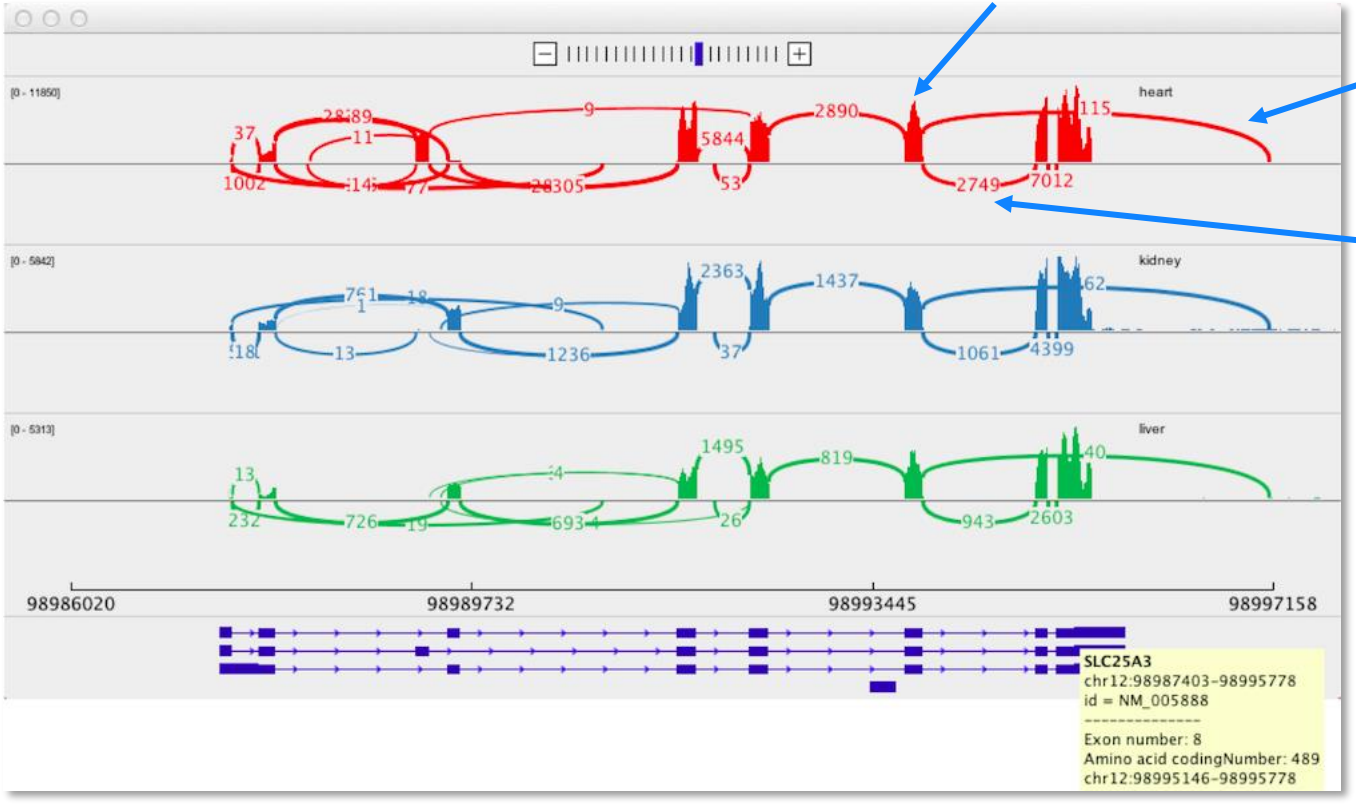
Single base insertions



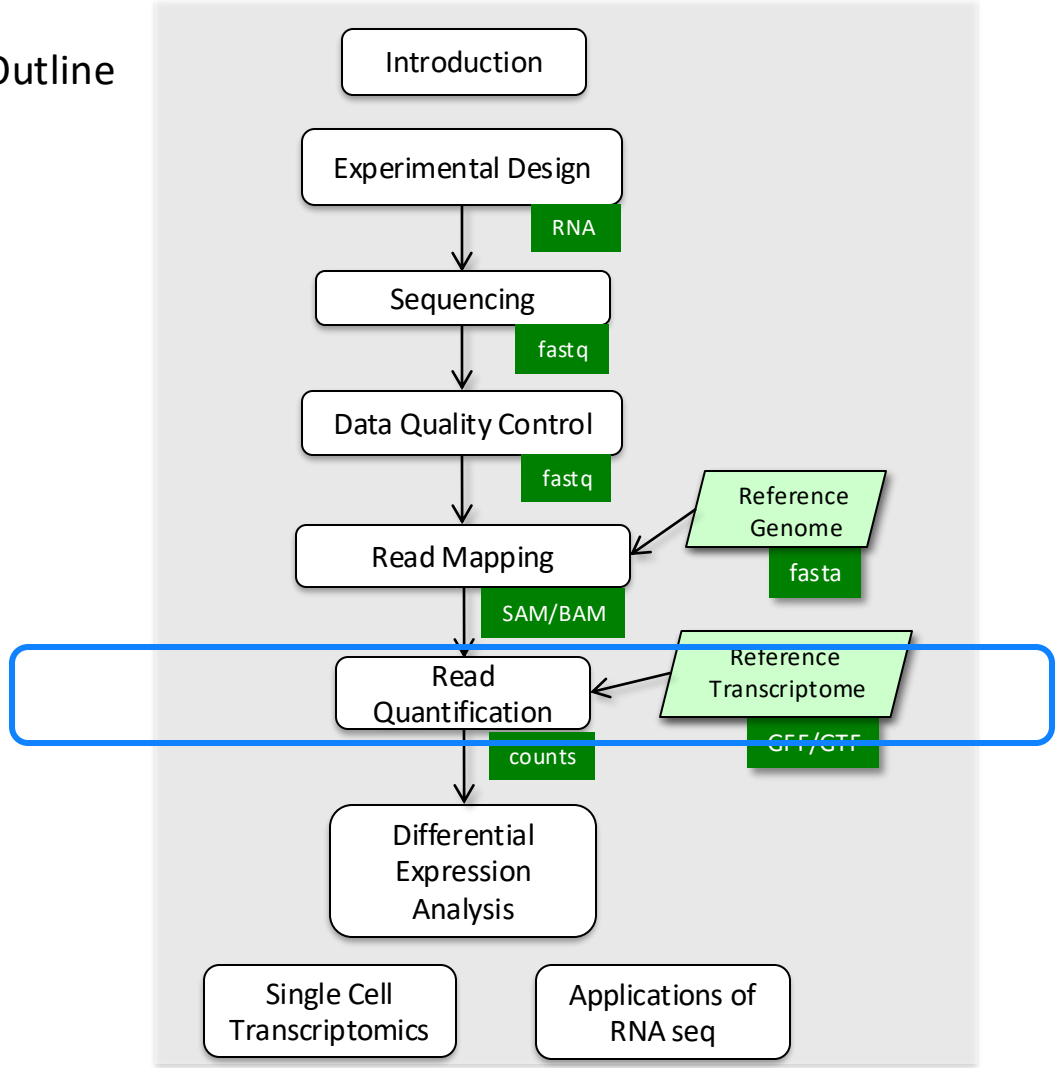
1 base insertion

Sashimi plots

- Quantitatively visualize splice junctions

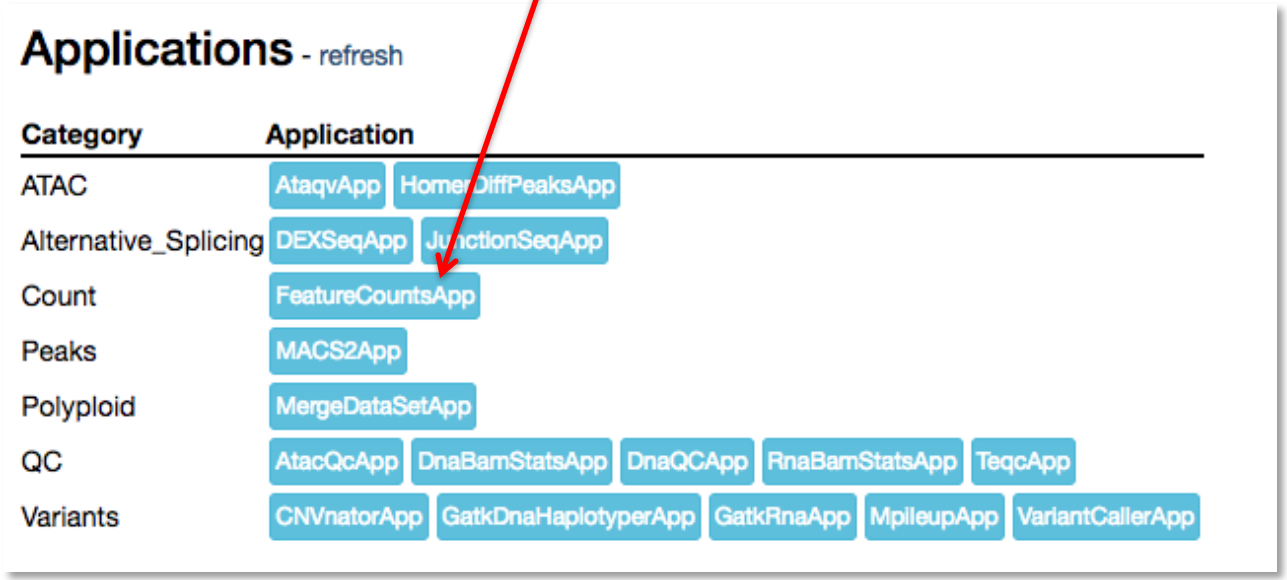


Outline

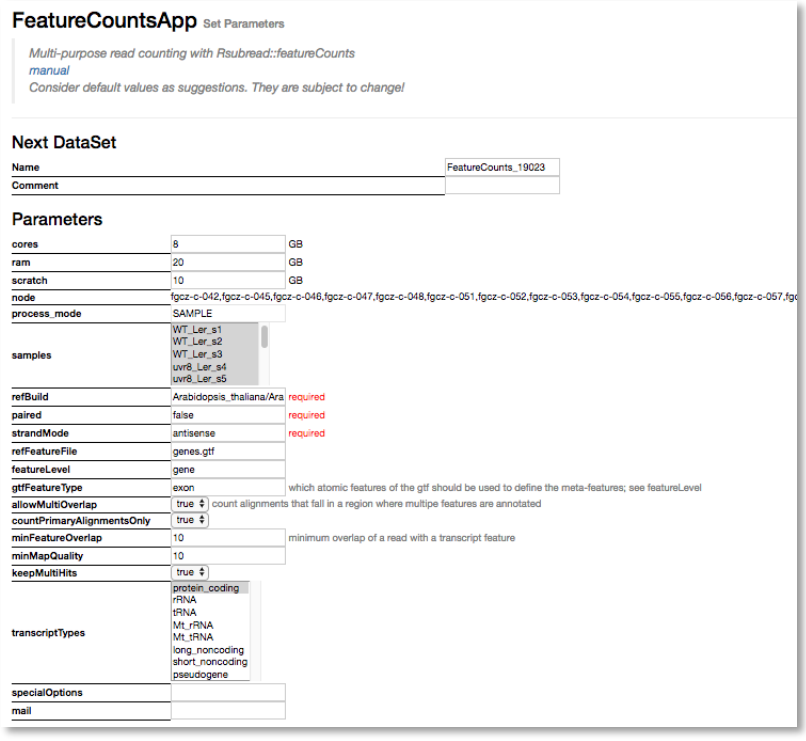

Introduction
Experimental Design
RNA
Sequencing
fastq
Data Quality Control
fastq
Reference Genome
fasta
Read Mapping
SAM/BAM
Reference Transcriptome
Read Quantification
counts
Differential Expression Analysis
Single Cell Transcriptomics
Applications of RNA seq

Read quantification

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



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-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,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 multipe 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		

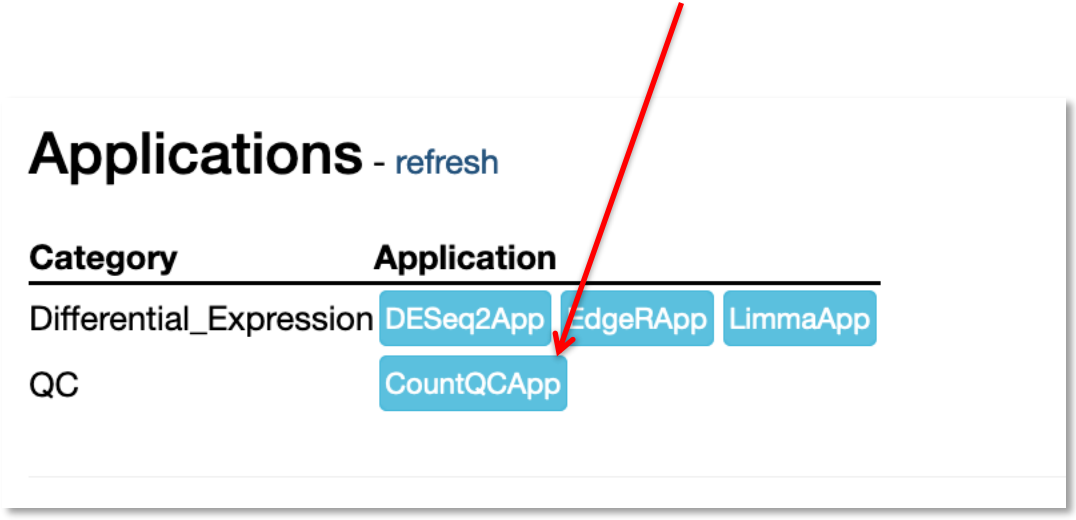
FeatureCountsApp critical parameters

- ***allowMultiOverlap***: true/false
- ***countPrimaryAlignmentsOnly***: true/false

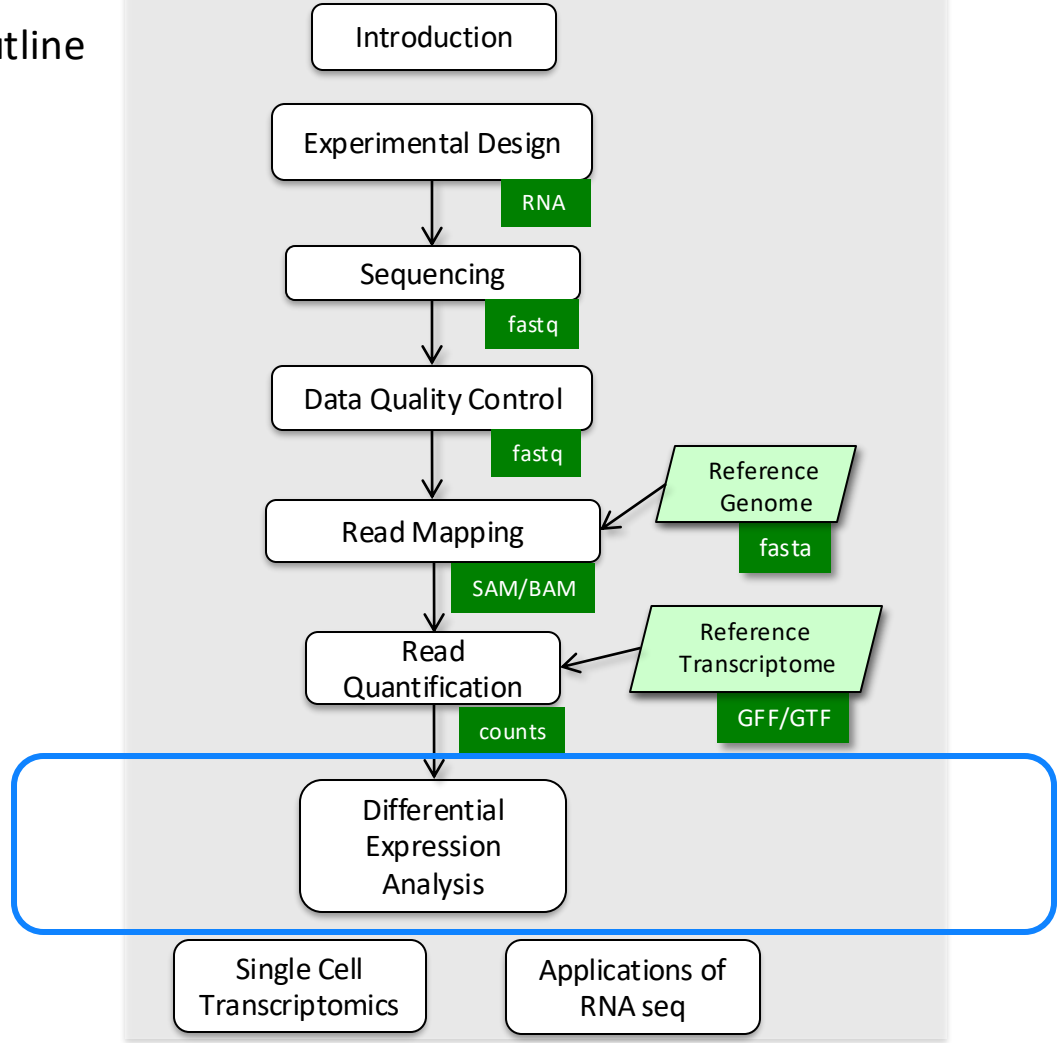
If all parameters are fine, continue with ‚**Next**'

CountQC

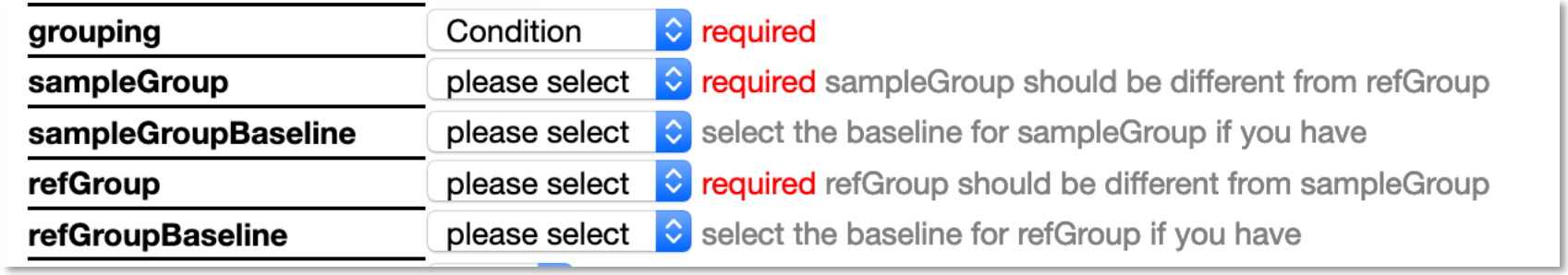
After counting the reads, click on "**CountQCApp**" in the Application section



Outline


Introduction
Experimental Design
RNA
Sequencing
fastq
Data Quality Control
fastq
Reference Genome
fasta
Read Mapping
SAM/BAM
Reference Transcriptome
GFF/GTF
Read Quantification
counts
Differential Expression Analysis
Single Cell Transcriptomics
Applications of RNA seq

Differential expression analysis (EdgeR)



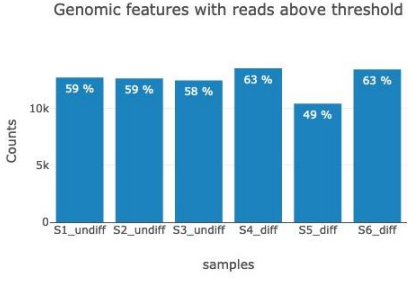
edgeRApp critical parameters

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

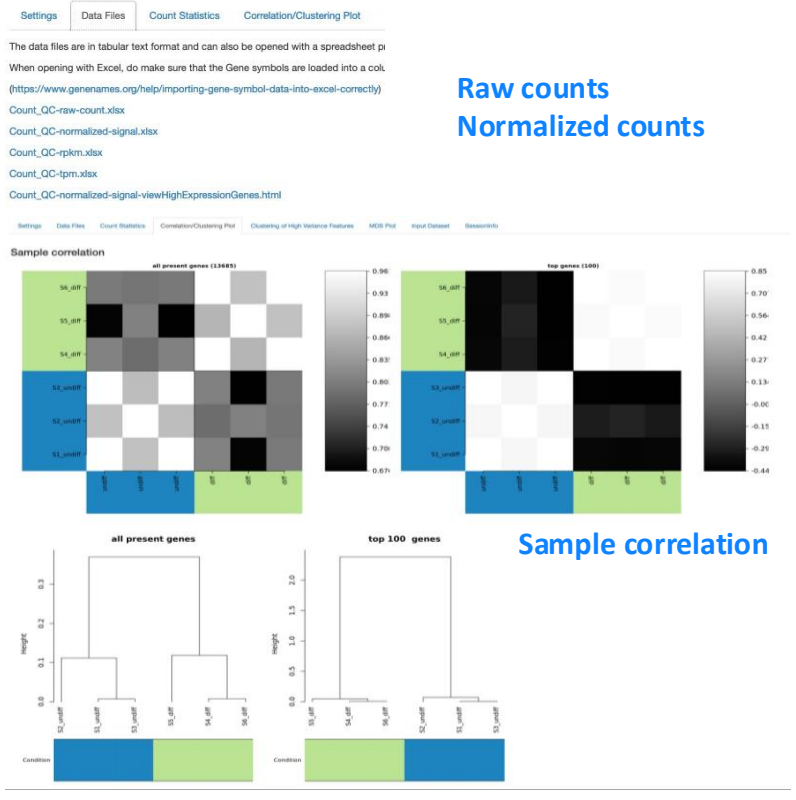
Short demo

CountQC (static report)

Count distribution



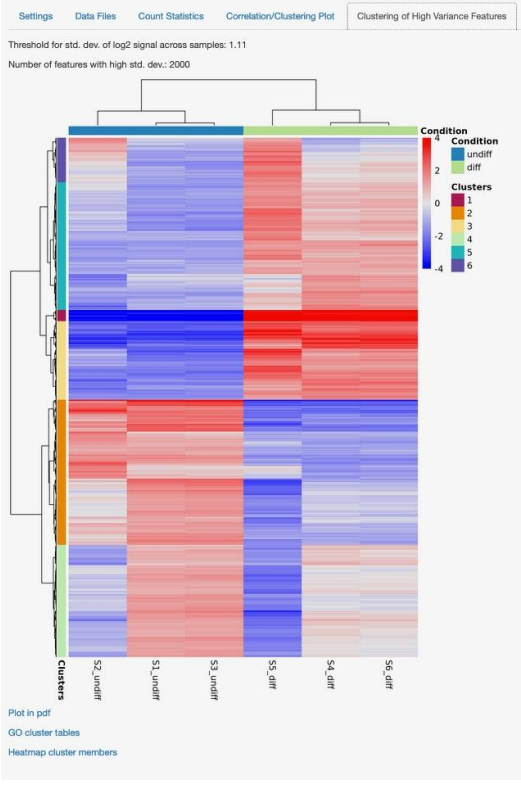
Feature distribution



Raw counts
Normalized counts

Sample correlation

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 significants by p-value and fold-change

	#significants	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

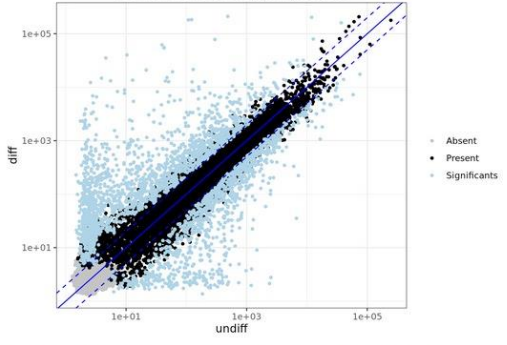
	Number
P-value threshold:	p <= 0.01
Log ratio threshold:	log ratio >= 0.5
Number of significant genes:	2642

p <= 0.01 (reduced FDR)

Subsequent plots highlight significant genes in blue.

Interactive table of significant genes

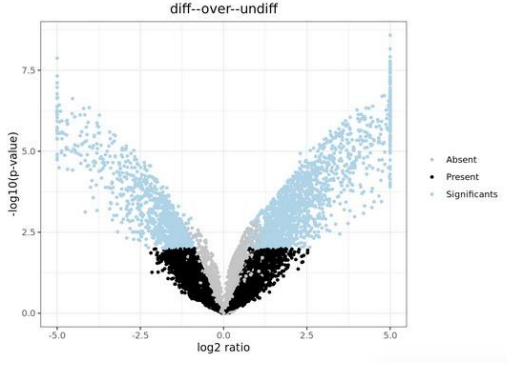
Comparison of average expression



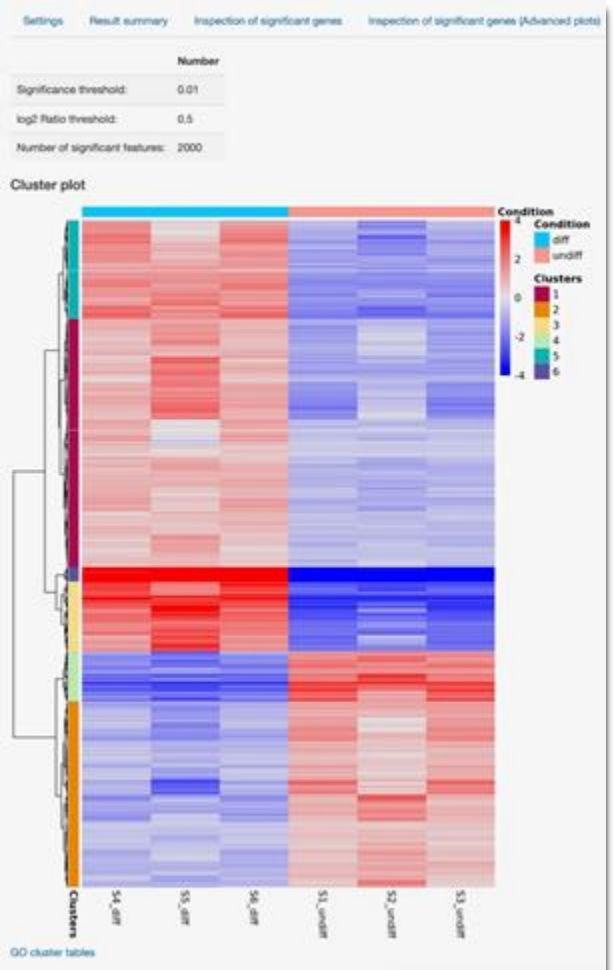
Visualization of significant DE genes

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

diff--over--undiff



EdgeR (static report)



Heat map of DE genes

GO categories of feature clusters

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

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

Settings | Result summary | Inspection of significant genes | Inspection of significant genes (Advanced plots) | Clustering of significant features | MetaCore | Enrichr | Overrepresentation Analysis (ORA) | Gene set enrichment analysis | Technical 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 smalles 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.

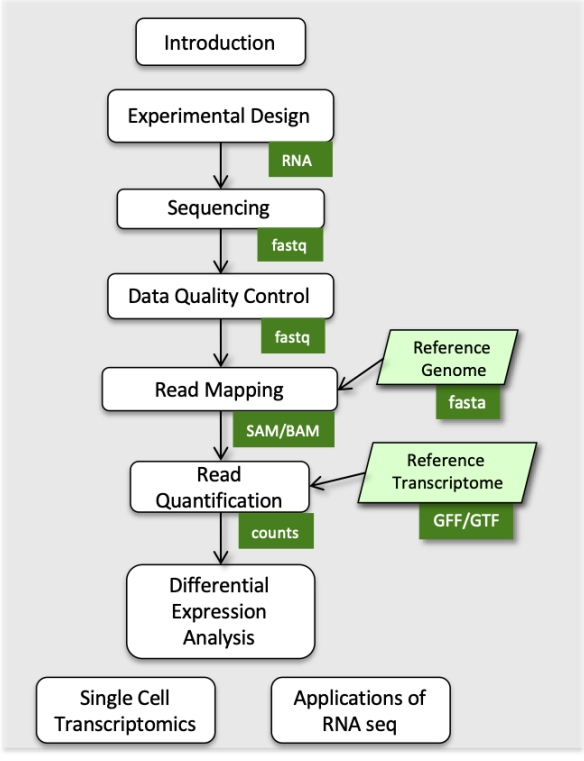
Association test: 1637 up-regulated and 1005 down-regulated genes

	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 detecttion
   - *DESeq2*: the second option



Question?