



CEITEC

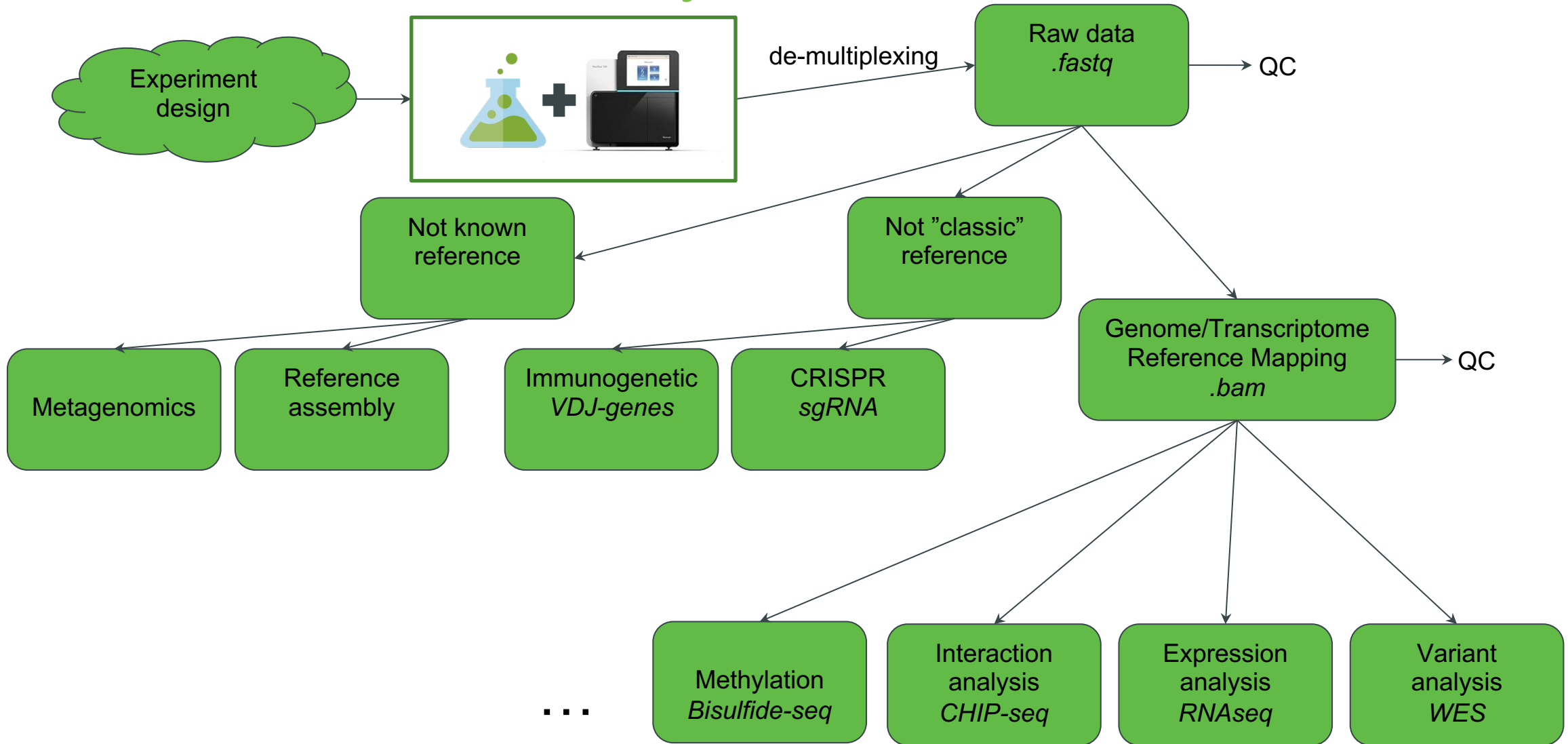
Central European Institute of Technology  
BRNO | CZECH REPUBLIC



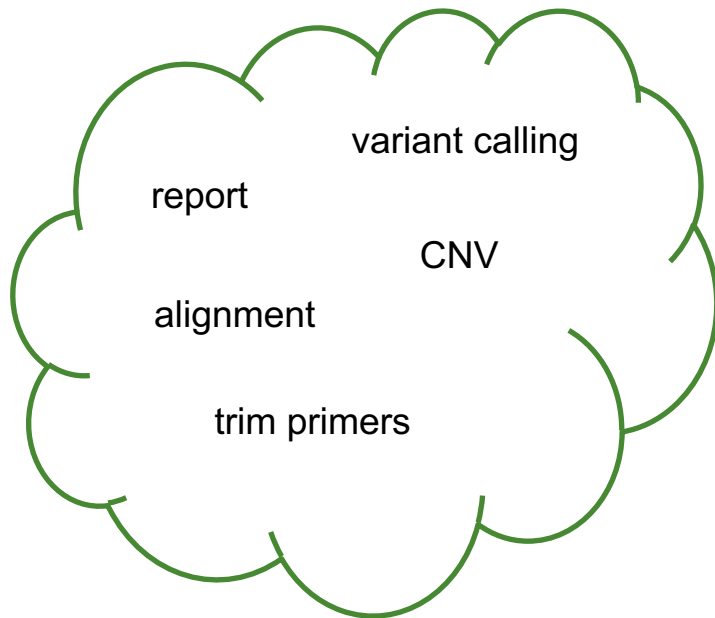
# Bioinformatics workflow management tools

Vojta Bystry  
[vojtech.bystry@ceitec.muni.cz](mailto:vojtech.bystry@ceitec.muni.cz)

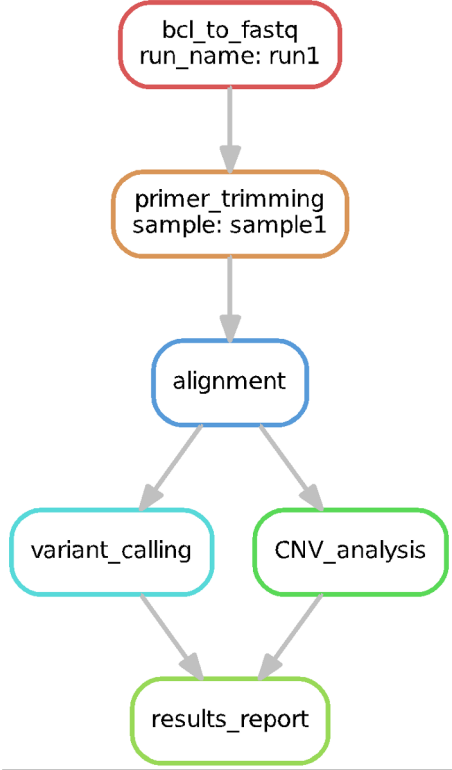
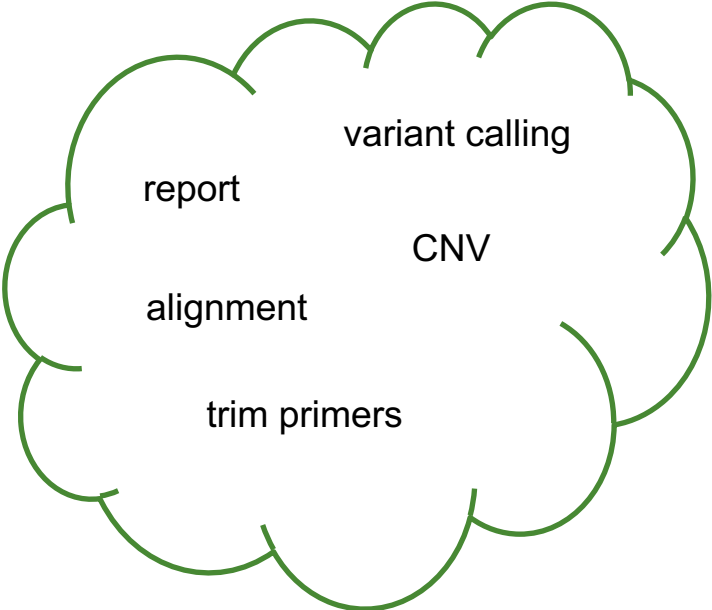
# NGS data analysis



# Bioinformatics workflow (pipeline)

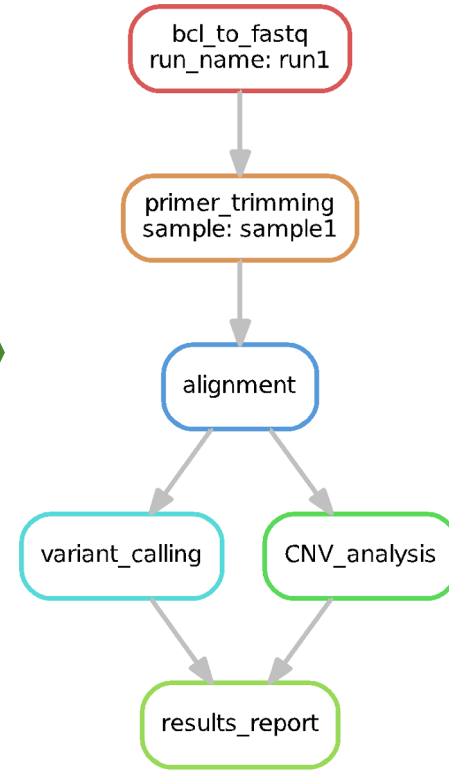
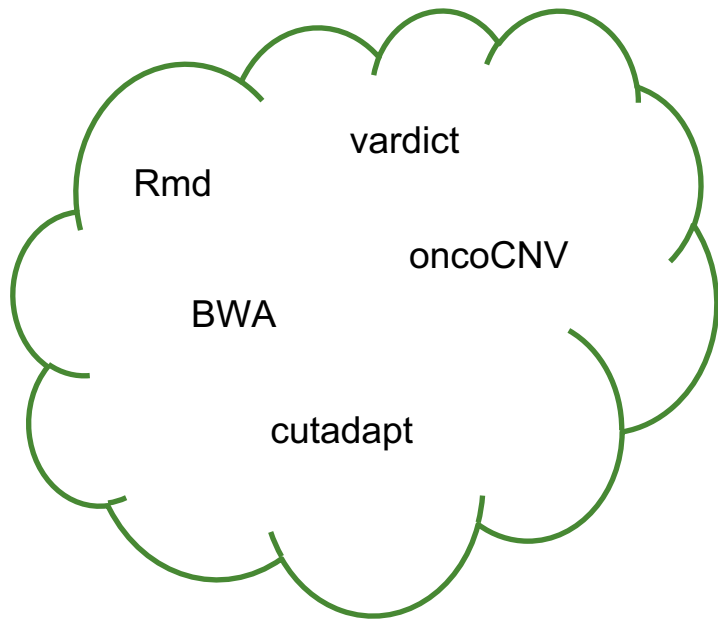


# Bioinformatics workflow (pipeline)



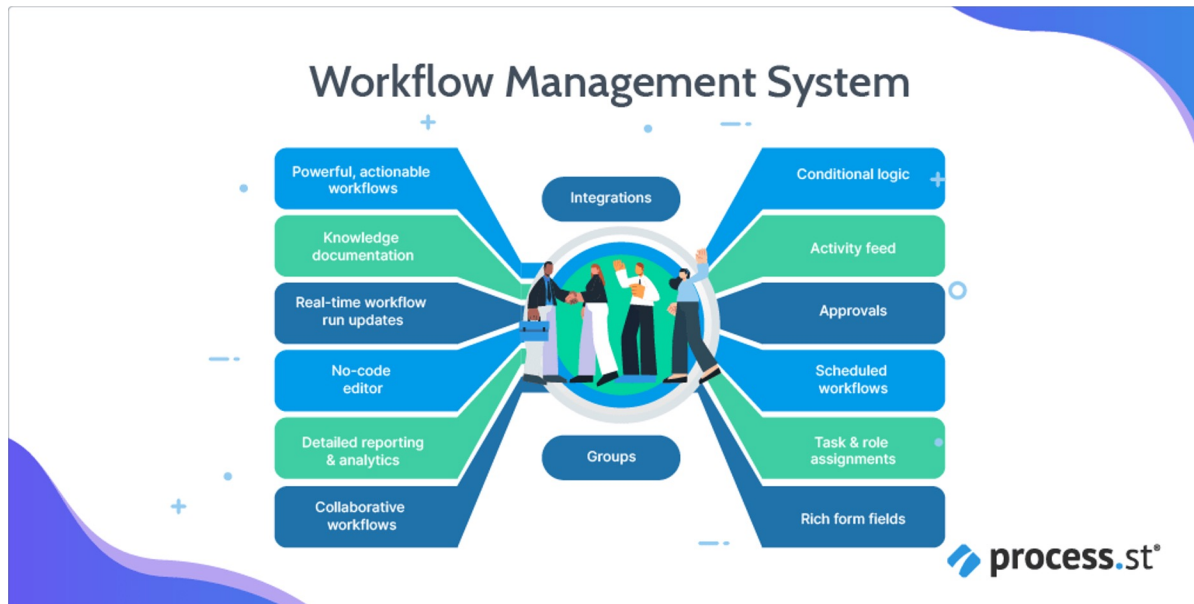


# Bioinformatics workflow (pipeline)

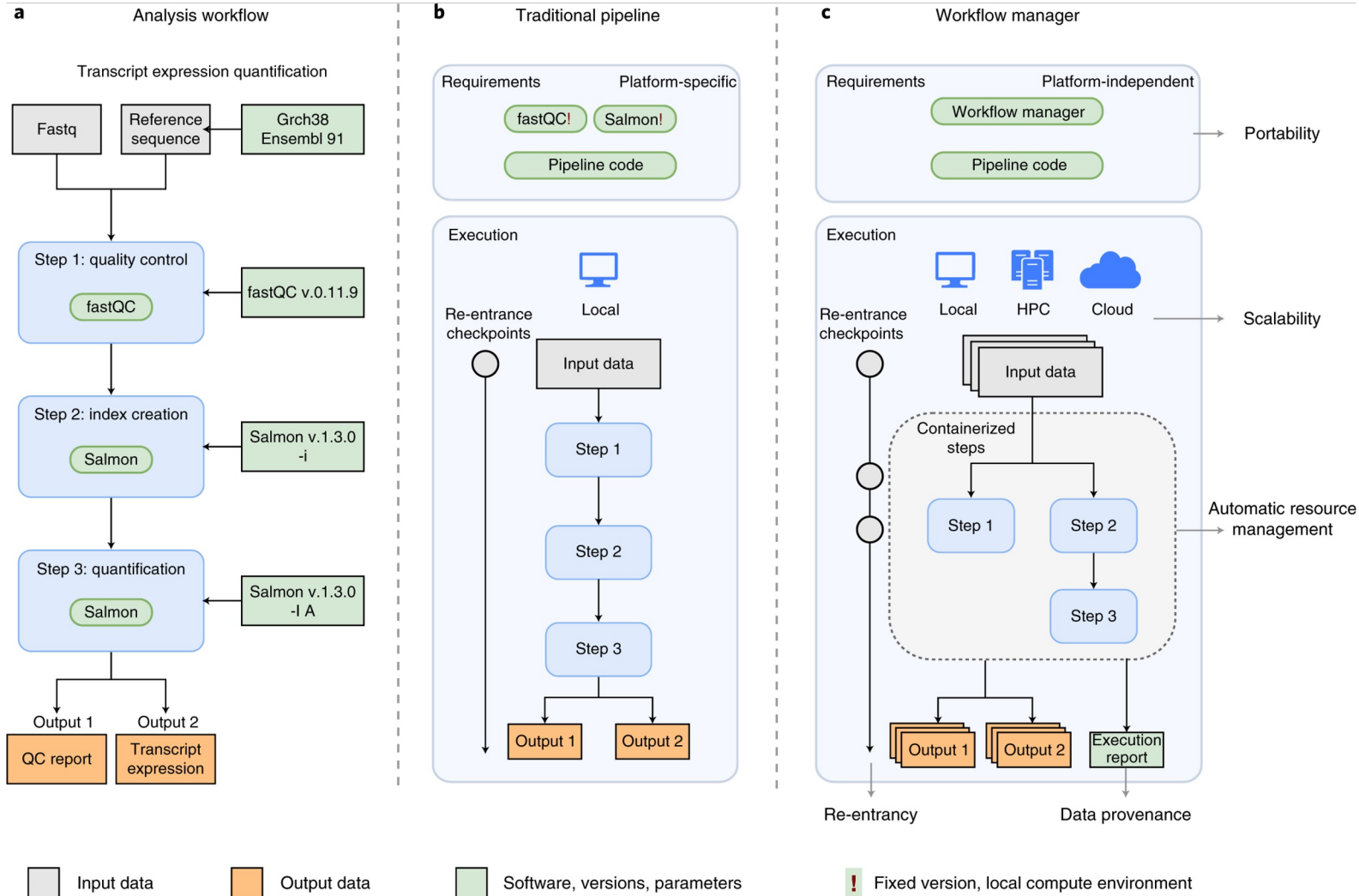


# Workflow management

- Workflow management is **the discipline of creating, documenting, monitoring and improving upon the series of steps, or workflow, that is required to complete a specific task.**



# Bioinformatic workflow management



# Bioinformatic workflow management

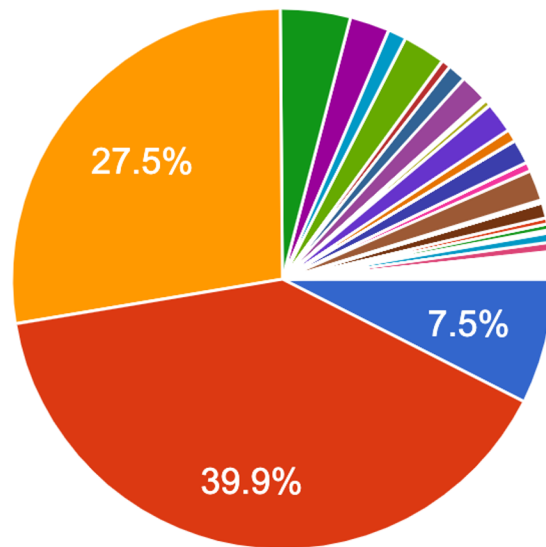
- Reusability and Reproducibility
- Parallelization and Scale
- Error solving / debugging

# Bioinformatic workflow managers

Which Bioinformatics Workflow Manager / Tool / Platform / Standard do you use or prefer?

[bit.ly/biowl](http://bit.ly/biowl)

549 responses



- None
- SnakeMake
- NextFlow
- Common Workflow Language
- Seven Bridges Genomics
- Luigi/Scilugli
- Niassa / Apache Oozie
- GalaxyProject

▲ 1/6 ▼

# Common Workflow Language (CWL)



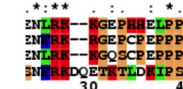
- Pushed by EU projects
- Not big grassroots community
- Scripts in .yaml format

```
hello_world.cwl

cwlVersion: v1.2

# What type of CWL process we have in this document.
class: CommandLineTool
# This CommandLineTool executes the linux "echo" command-line tool.
baseCommand: echo

# The inputs for this process.
inputs:
  message:
    type: string
    # A default value that can be overridden, e.g. --message "Hola mundo"
    default: "Hello World"
    # Bind this message value as an argument to "echo".
    inputBinding:
      position: 1
outputs: []
```





# Galaxy project



- Workflow manager with GUI
- Biologists can do their own analysis ???
- It can work - EMBL

A screenshot of the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. On the left, a 'Tools' sidebar lists various categories like 'Get Data', 'Text Manipulation', and 'Statistics'. The main content area is divided into two panels: 'Edit Attributes' and 'Change data type'. The 'Edit Attributes' panel shows fields for 'Name' (containing 'Join two Queries on data 3 and data 1'), 'Info', 'Database/Build', and 'Number of comment lines'. The 'Change data type' panel shows a 'New Type' dropdown set to 'tabular'. On the right, a 'History' sidebar shows a list of workflow steps, such as '14: Draw phylogeny on data 12', '13: Draw phylogeny on data 11', and '12: Find lowest diagnostic rank on data 10'. Each step includes icons for refresh, delete, and expand.

# Nextflow

- Great deployability
- Great existing workflow repository



nextflow

```
1  #!/usr/bin/env nextflow
2
3
4  params.in = "$baseDir/data/sample.fa"
5
6  /*
7   * Split a fasta file into multiple files
8   */
9  process splitSequences {
10
11     input:
12     path 'input.fa'
13
14     output:
15     path 'seq_*'
16
17     """
18     awk '/^>/{f="seq_"++d} {print > f}' < input.fa
19     """
20 }
21
22 /*
23  * Reverse the sequences
24  */
25 process reverse {
26
27     input:
28     path x
29
30     output:
31     stdout
32
33     """
34     cat $x | rev
35     """
36 }
```

# Snakemake



**BIOINFORMATICS APPLICATION NOTE**

Vol. 28 no. 19 2012, pages 2520–2522  
doi:10.1093/bioinformatics/bts480

*Genome analysis*

Advance Access publication August 20, 2012

## **Snakemake—a scalable bioinformatics workflow engine**

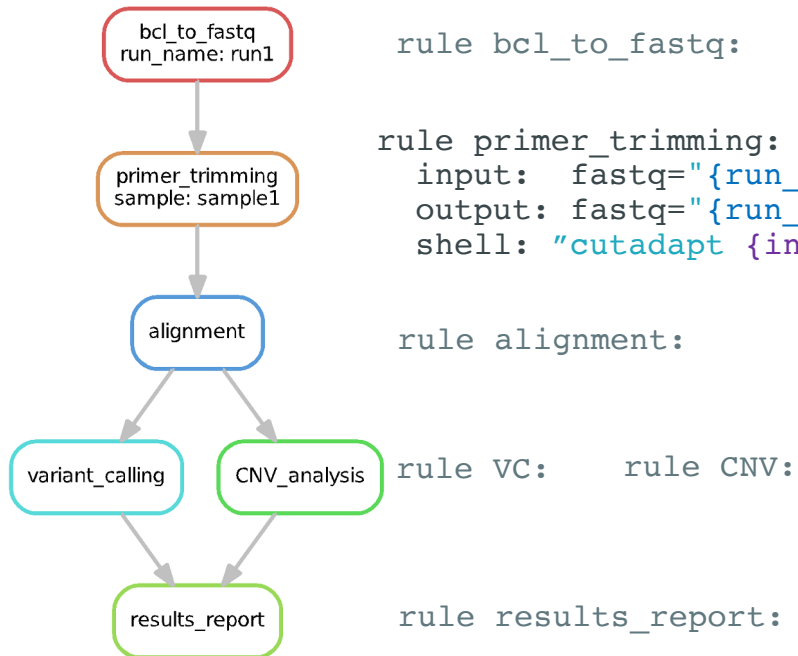
Johannes Köster<sup>1,2,\*</sup> and Sven Rahmann<sup>1</sup>

<sup>1</sup>Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen and <sup>2</sup>Paediatric Oncology, University Childrens Hospital, 45147 Essen, Germany

Associate Editor: Alfonso Valencia

- make + python = Snakemake

# Snakemake



```
rule bcl_to_fastq:
```

```
rule primer_trimming:
```

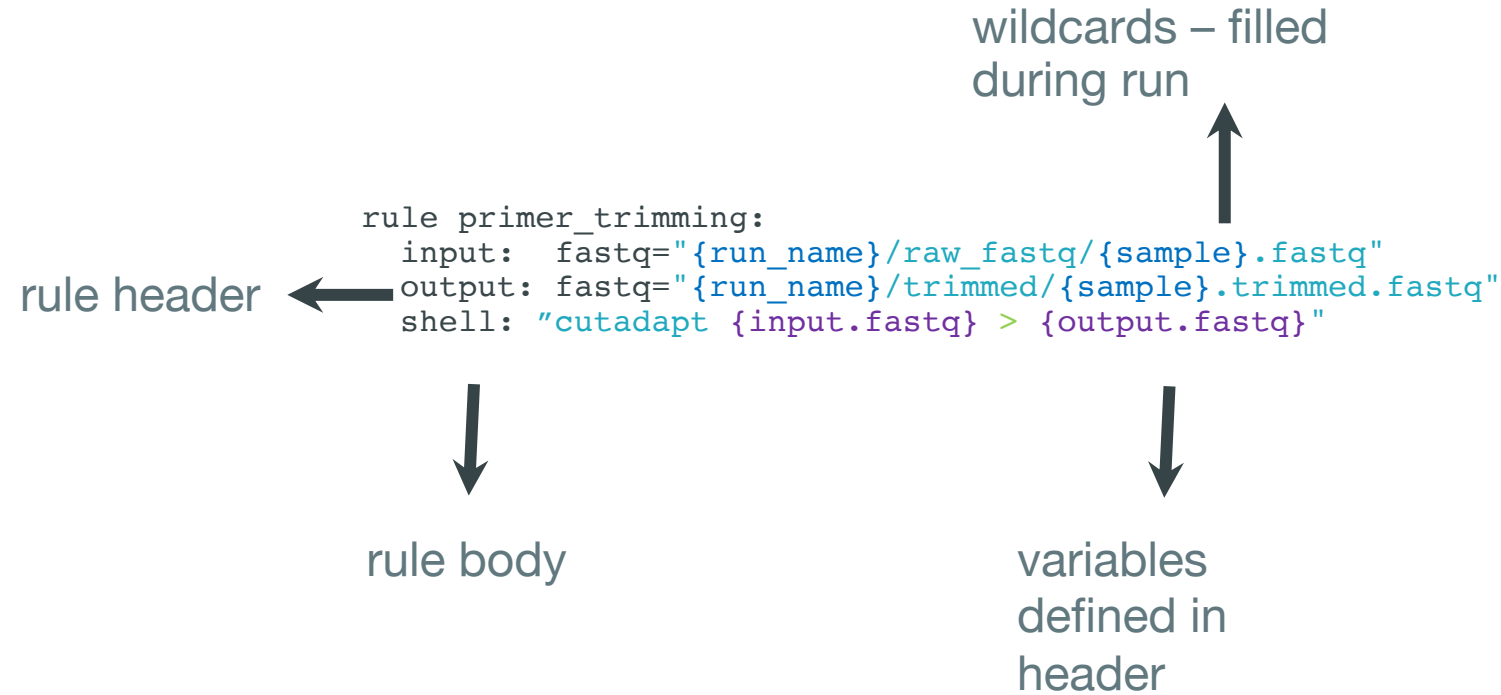
```
input: fastq="{run_name}/raw_fastq/{sample}.fastq"  
output: fastq="{run_name}/trimmed/{sample}.trimmed.fastq"  
shell: "cutadapt {input.fastq} > {output.fastq}"
```

```
rule alignment:
```

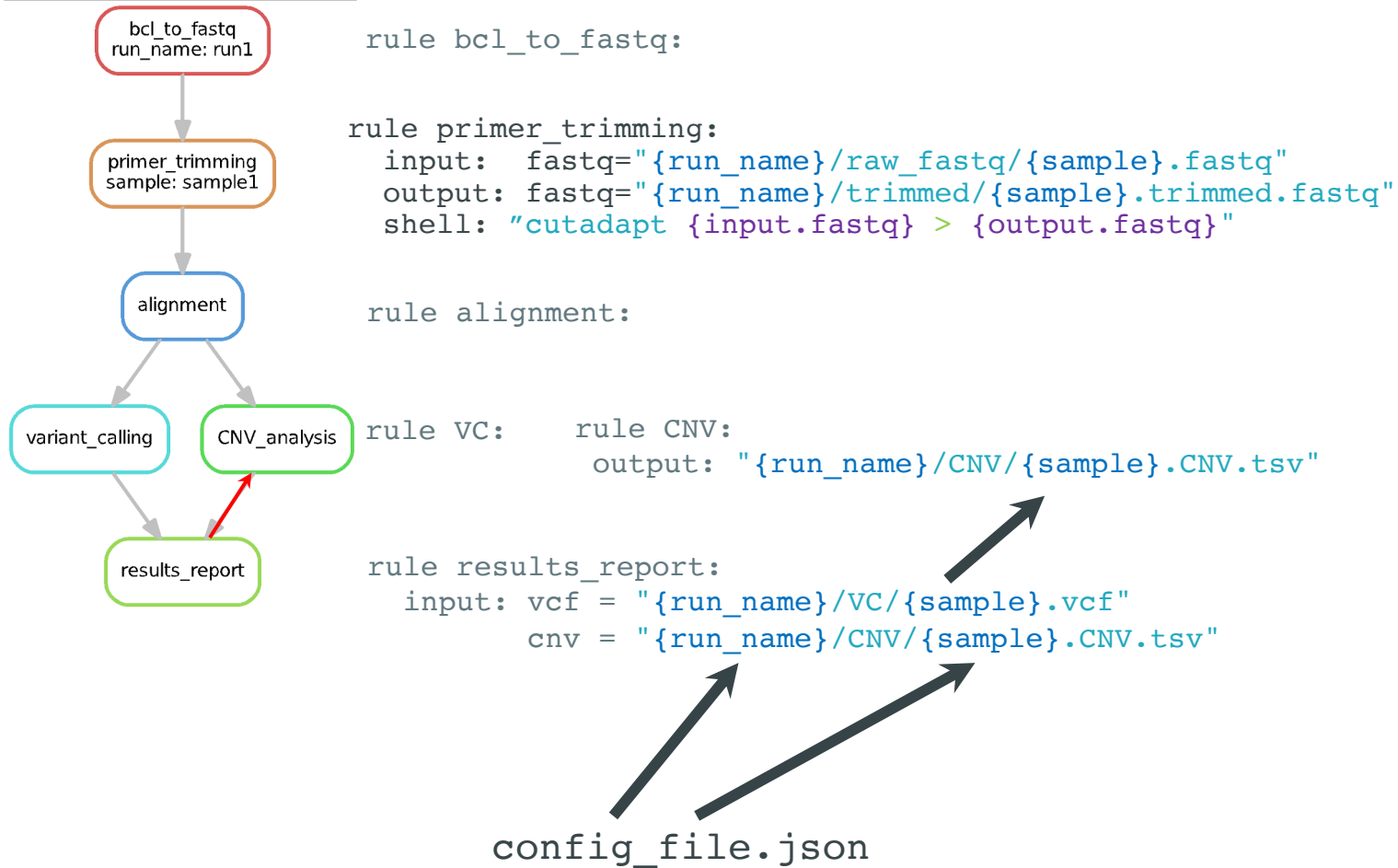
```
rule VC:    rule CNV:
```

```
rule results_report:
```

# Snakemake



# Snakemake





# Snakemake



- ▶ Simple shell script

```
shell: "mv -R {input} {output}"
```

- ▶ Combine languages

```
run:  
if {params.cluster} is TRUE:  
    R("cutree(hclust({input}),h = 7)")  
else:  
    shell("mv -R {input} {output}")
```

python  
R  
shell

Three arrows originate from the code: one from the R function call points to the label "python", one from the shell function call points to the label "R", and one from the shell function call points to the label "shell".

- ▶ Wrap it in separate script

```
script: "my_script.py"
```

- ▶ Separation of logic and functionality

- ▶ Organization

- ▶ Re-usability

# Snakemake - running



- Nice logs and error reporting

## Job counts:

count	jobs
1	all
16	filter_variants
15	igv_picture_print_germline
1	igv_picture_print_somatic
16	merge_variant_callers
32	normalize_variants
16	variant_annotation
97	

```
rule igv_picture_print_germline:
  input: DB_9905/s1/a1/variant_calling/vcf/DB_9905.control.germline.annotated.filtered.vcf,
  e, DB_9905/s1/raw_data/mapped/DB_9905.control.bam
  output: DB_9905/s1/a1/variant_calling/igv/DB_9905.control.germline
  jobid: 13
  reason: Missing output files: DB_9905/s1/a1/variant_calling/igv/DB_9905.control.germline;
  g/vcf/DB_9905.control.germline.annotated.filtered.vcf
  wildcards: donor=DB_9905, sampling=s1, analysis=a1, tag=control
```

```
MissingInputException in line 1787 of /mnt/nfs/shared/999993-Bioda/scripts/vojta/snakemake/primary.smk:
Missing input files for rule ref_info_copy:
/mnt/ssd/ssd_3/references/homsap/TAIR10-31/info.txt
```

```
rule variant_annotation:
  input: DB_9905/s1/a1/variant_calling/vcf/DB_9905.somatic.vcf, /mnt/ssd/ssd_3/references/homsap/GRCh37-p13/seq/GRCh37-p13.fa
  output: DB_9905/s1/a1/variant_calling/vcf/DB_9905.somatic.annotated.vcf
  jobid: 72
  wildcards: donor=DB_9905, sampling=s1, analysis=a1, tag=somatic

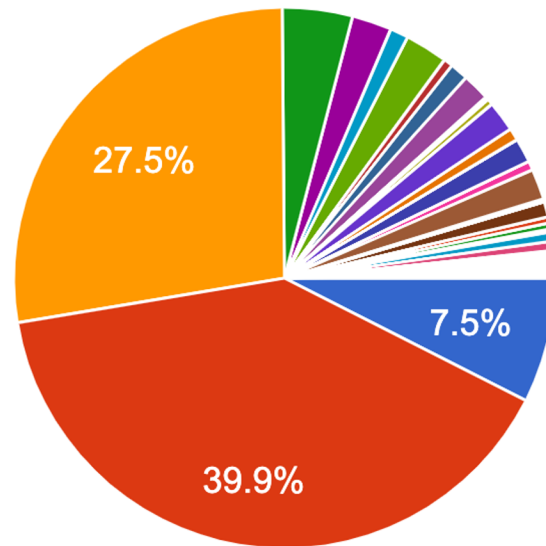
ERROR: Unrecognised consequence term type specified "CombineVariants" - must be one of ensembl, so, ncbi
Error in job variant_annotation while creating output file DB_9905/s1/a1/variant_calling/vcf/DB_9905.somatic.annotated.vcf.
RuleException:
CalledProcessError in line 137 of /mnt/nfs/shared/999993-Bioda/scripts/vojta/snakemake/variant_calling.smk:
Command 'variant_effect_predictor.pl --dir /opt/reference/GRCh37/vep/ -T CombineVariants --show_cache_info --everything --fasta /mnt/ssd/ssd_3/references/homsap/GRCh37-p13/seq/GRCh37-p13.fa --offline --cache_version 83 --assembly GRCh37 --input_file DB_9905/s1/a1/variant_calling/vcf/DB_9905.somatic.vcf /mnt/ssd/ssd_3/references/homsap/GRCh37-p13/seq/GRCh37-p13.fa --output_file DB_9905/s1/a1/variant_calling/vcf/DB_9905.somatic.annotated.vcf --force_overwrite --stats_text ' returned non-zero exit status 2.
File "/mnt/nfs/shared/999993-Bioda/scripts/vojta/snakemake/variant_calling.smk", line 137, in __rule_variant_annotation
File "/opt/install/dir/anaconda/envs/bioda/lib/python3.6/concurrent/futures/thread.py", line 55, in run
Exiting because a job execution failed. Look above for error message
Will exit after finishing currently running jobs.
```

# Bioinformatic workflow managers

Which Bioinformatics Workflow Manager / Tool / Platform / Standard do you use or prefer?

[bit.ly/biowl](http://bit.ly/biowl)

549 responses



- None
- SnakeMake
- NextFlow
- Common Workflow Language
- Seven Bridges Genomics
- Luigi/Scilugli
- Niassa / Apache Oozie
- GalaxyProject

▲ 1/6 ▼

# Reproducibility

ISMB 2016 - 47 open-access publications:

status	count
properly documented, easy to install	4
web service	4
Docker image	1
R packages, not (yet) on CRAN or Bioconductor	3
no software implementation	7
MATLAB code	4
available upon request	2
collection of scripts without proper way to install	12
demo only although README promises a release before ISMB	1
either unclear, no, or erroneous installation instructions	3
missing download URL	1
invalid links	4
build error	1

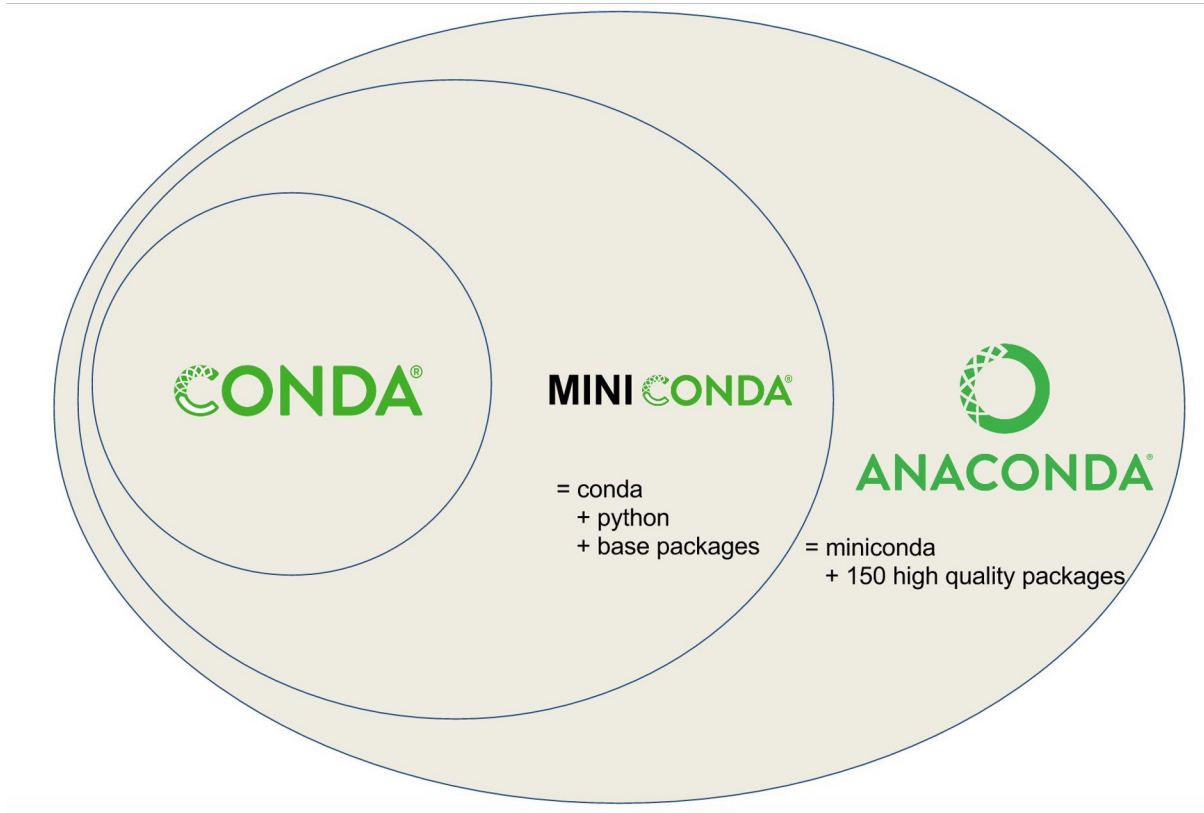
good (26%)

bad (53%)

ugly (21%)

- Now it is much better

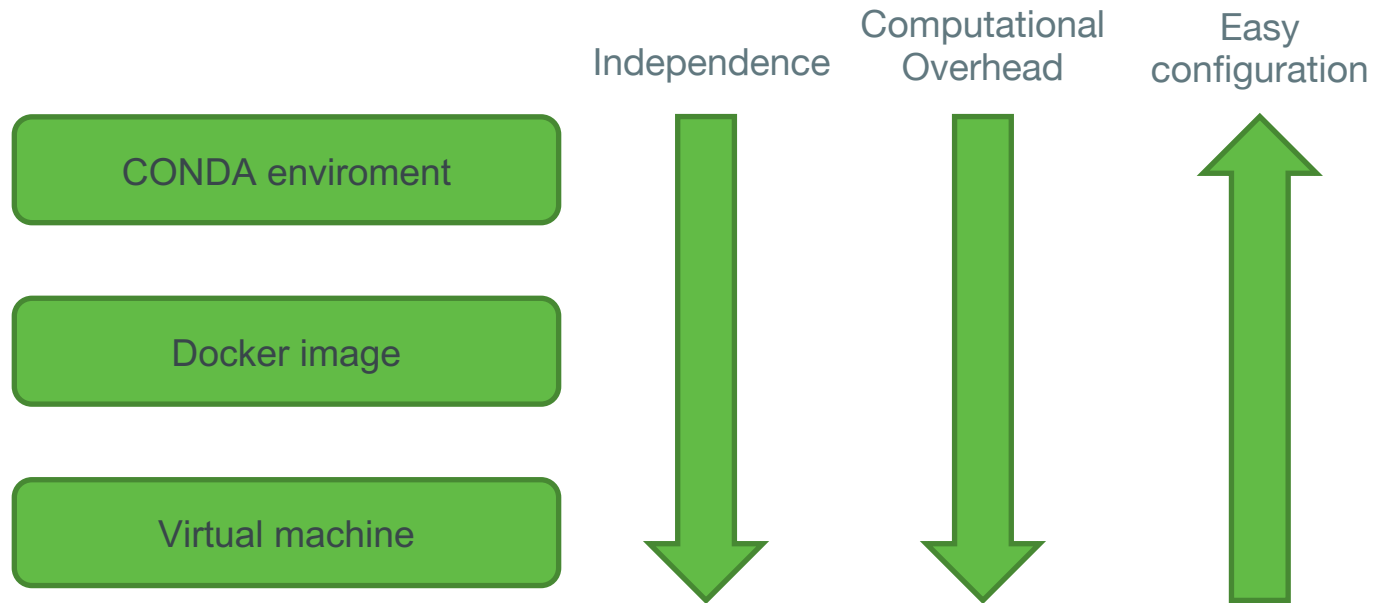
# Conda / Anaconda / Bioconda



## BIOCONDA®

Bioconda is a distribution of bioinformatics software realized as a channel for the versatile Conda package manager.

# Conda





- Easy installation and management
- Installation recipes:

```
conda install vardict
conda update vardict
conda remove vardict
conda env create -f myenv.yaml -n myenv
```

- Isolated environments:

```
channels:
- conda-forge
- defaults
dependencies:
- pandas ==0.20.3
- statsmodels ==0.8.0
- r-dplyr ==0.7.0
- r-base ==3.4.1
```

# Conda

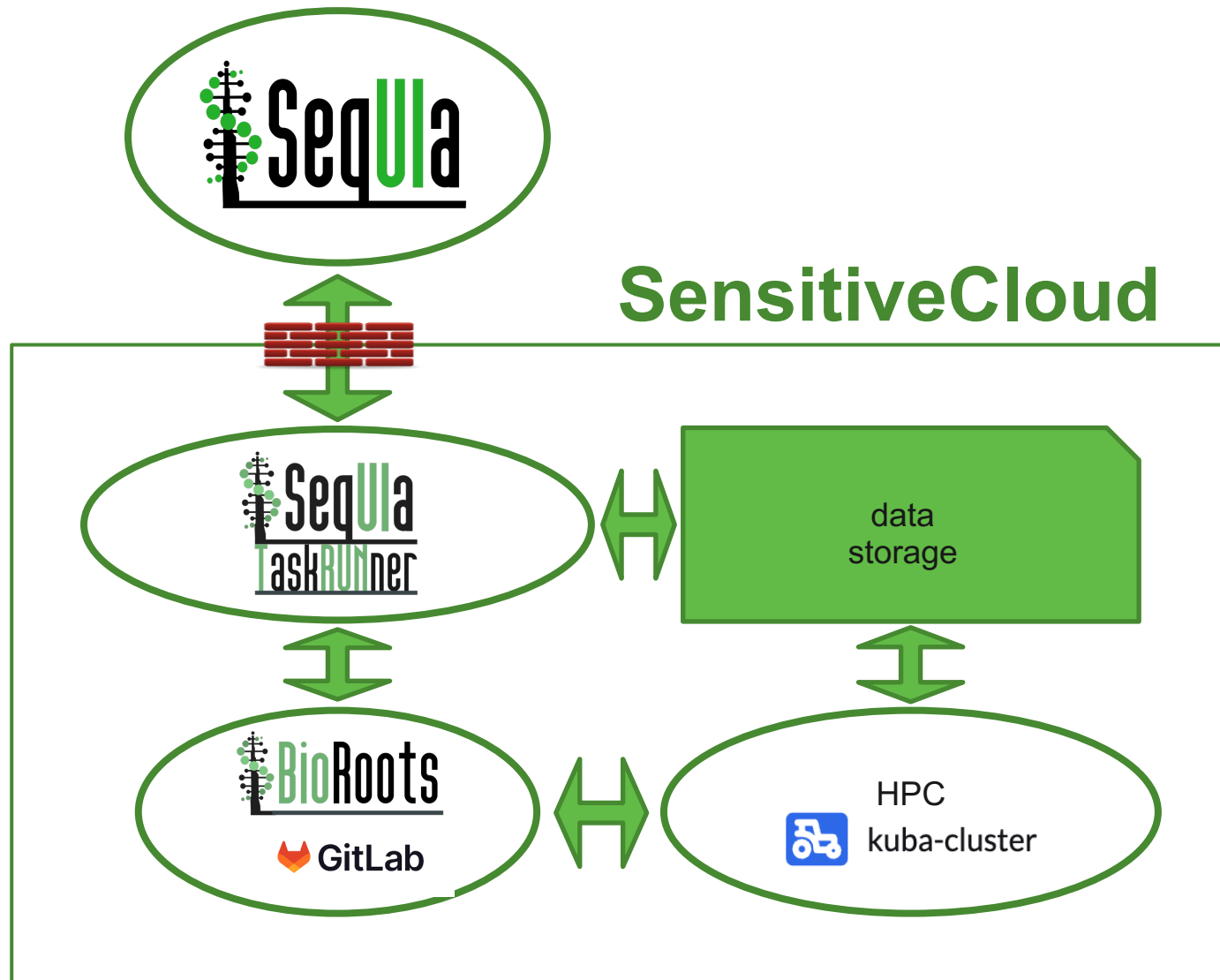


- Cheat sheet
  - [https://docs.conda.io/projects/conda/en/4.6.0/\\_downloads/52a95608c49671267e40c689e0bc00ca/conda-cheatsheet.pdf](https://docs.conda.io/projects/conda/en/4.6.0/_downloads/52a95608c49671267e40c689e0bc00ca/conda-cheatsheet.pdf)
- Google it
  - conda [bioinformatics tool name]

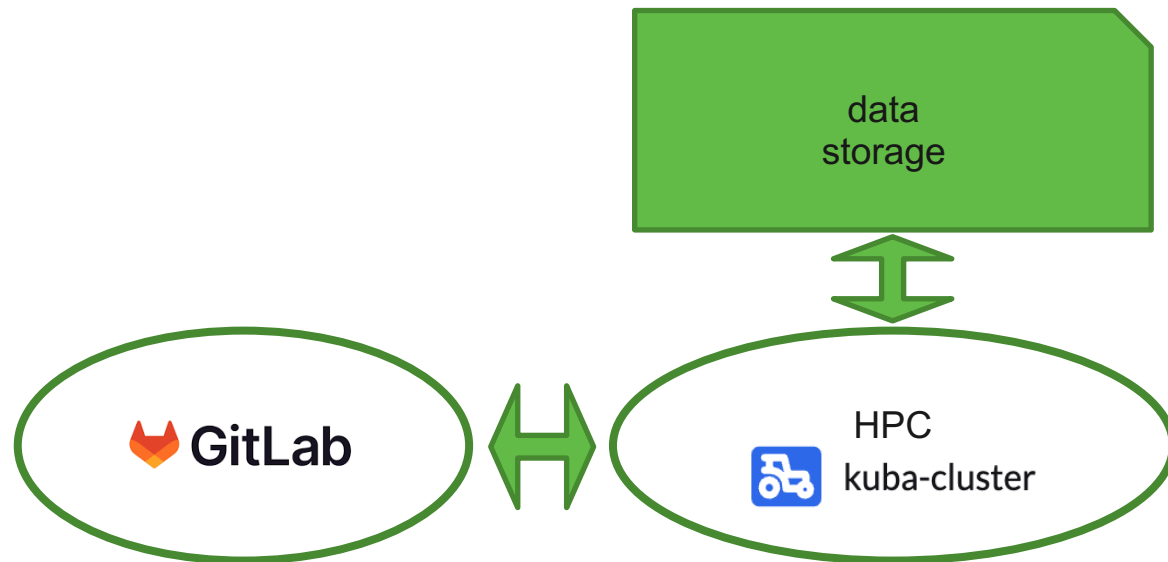
# Computational resources and execution

- Snakemake is quite flexible in cluster execution
  - <https://snakemake.readthedocs.io/en/stable/executing/cloud.html>
  - ! Nothing works as advertise 😊

# Computational resources and execution





# Computational resources and execution










# Computational resources and execution

<https://rancher.cloud.e-infra.cz/>

**B BioRoots**   
Group ID: 11949278  [Leave group](#)

**Subgroups and projects** | Shared projects | Archived projects

- S** small RNA analysis 
- F** Fastq merge 
- C** cleaned\_fastq\_qc 
- A** Alignment ChIP 
- T** Transcriptome assembly 
- S** Snakemake workflow template 
- C** Count\_feature\_rna 

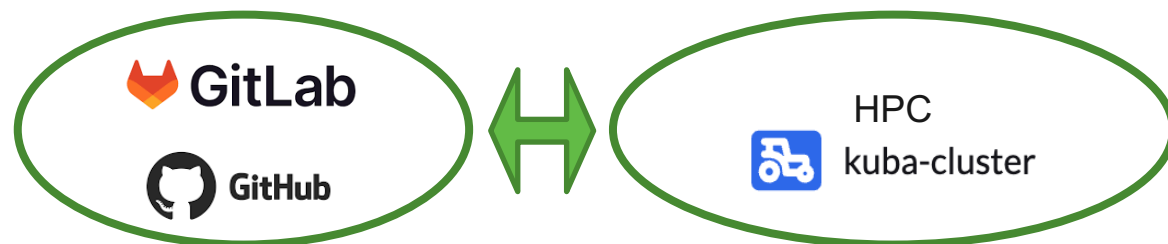
**kuba-cluster**

- Starred
- Cluster
- Workloads
  - CronJobs 1
  - DaemonSets 0
  - Deployments 1
  - Jobs 6**
  - StatefulSets 1
  - Pods 7
- Apps
- Service Discovery
- Storage
- Policy
- Monitoring
- Logging
- More Resources

**Jobs** ☆



[Download YAML](#) [Delete](#)

State	Name	Namespace
Active	12054--raw-fastq-qc--navrk67--231002-1696232891	sequia-ns
Active	taskrunner-cron-update-basic-resource-cache-28227506	sequia-ns
Failed	taskrunner-cron-update-basic-resource-cache-28269248	sequia-ns
Job Failed. failed: 7/1		
Active	taskrunner-cron-update-basic-resource-cache-28270546	sequia-ns
Active	taskrunner-cron-update-basic-resource-cache-28270548	sequia-ns
Active	taskrunner-cron-update-basic-resource-cache-28270550	sequia-ns

















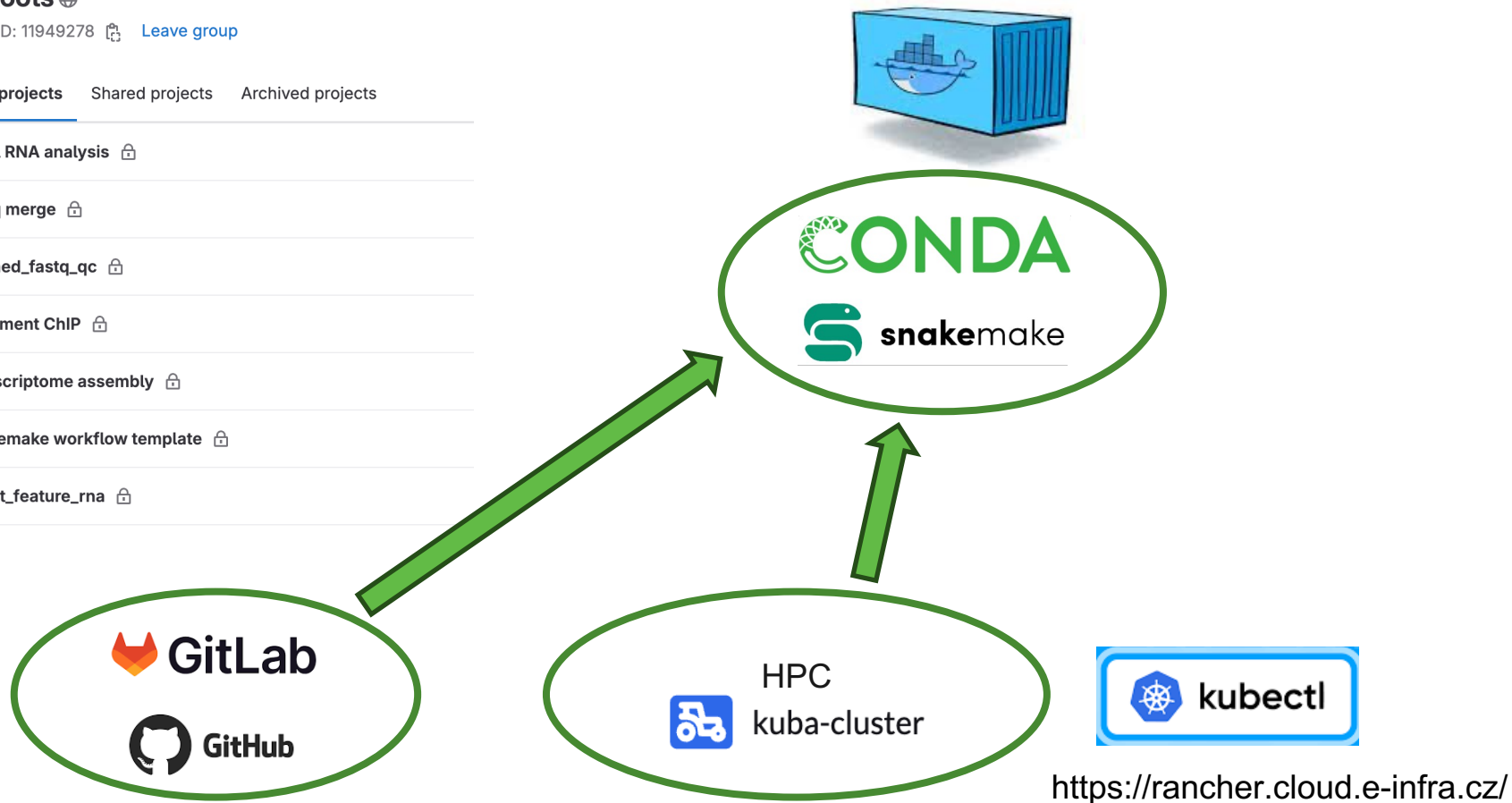


# Computational resources and execution

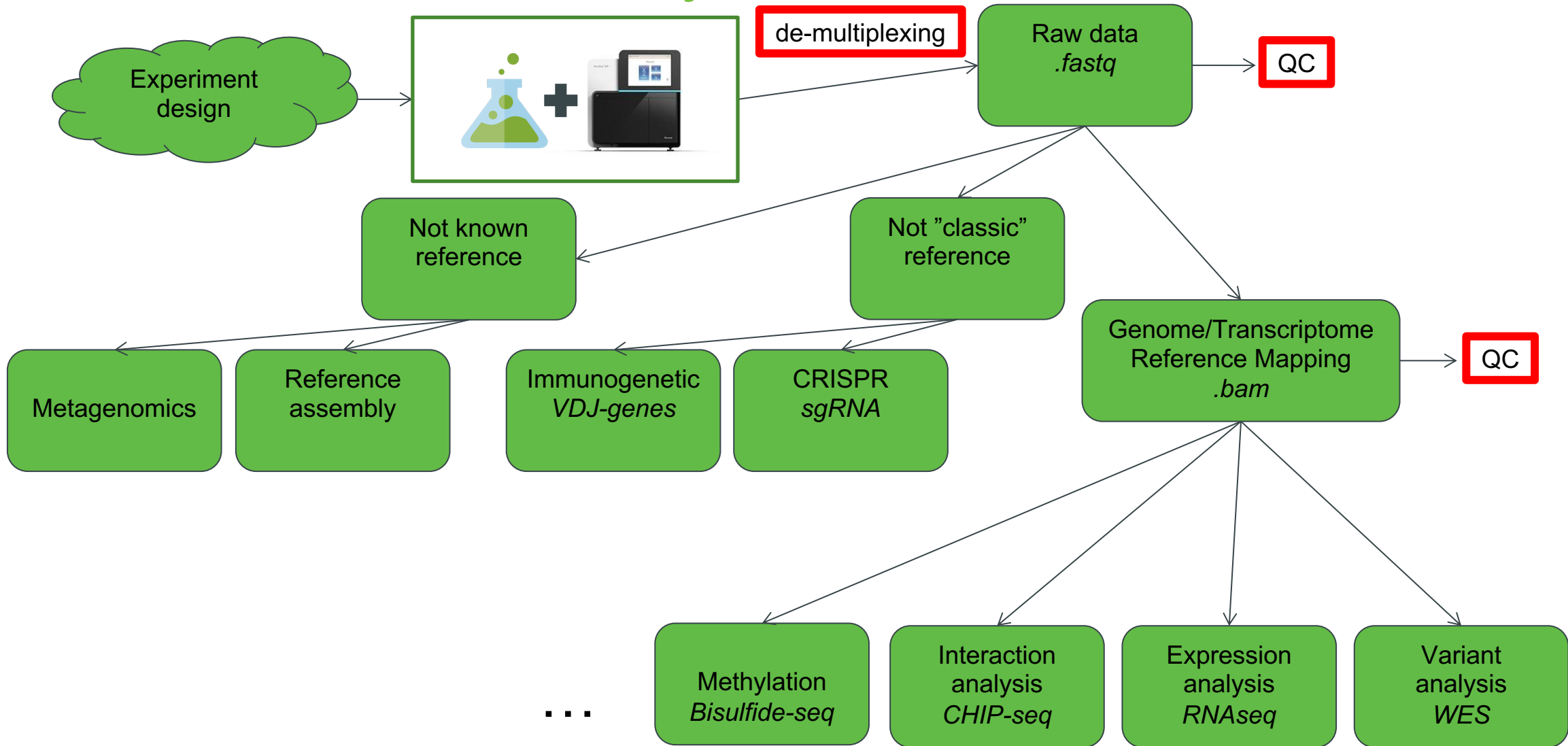
**B BioRoots**    
 Group ID: 11949278  [Leave group](#)

**Subgroups and projects**   Shared projects   Archived projects

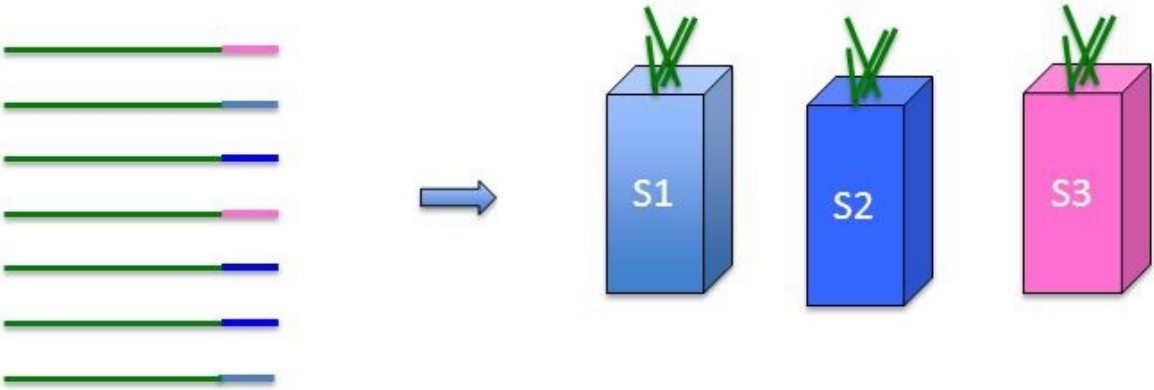
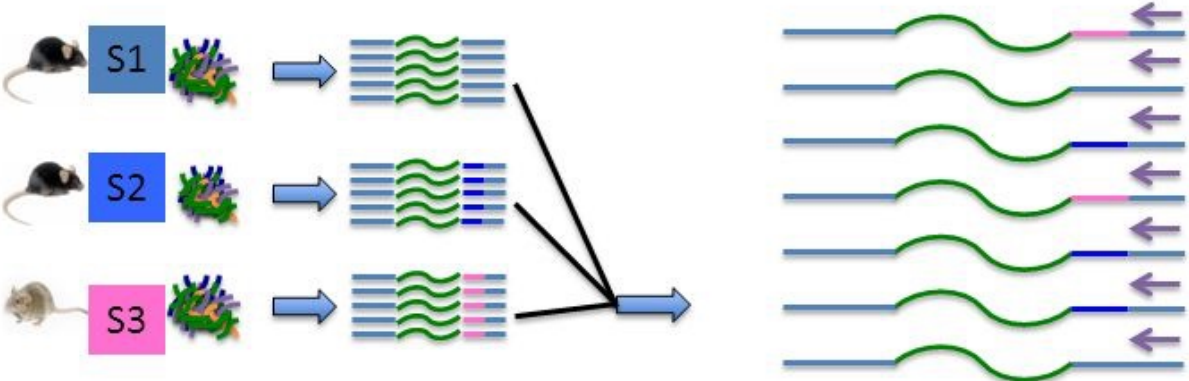
-  **S** small RNA analysis 
-  **F** Fastq merge 
-  **C** cleaned\_fastq\_qc 
-  **A** Alignment ChIP 
-  **T** Transcriptome assembly 
-  **S** Snakemake workflow template 
-  **C** Count\_feature\_rna 



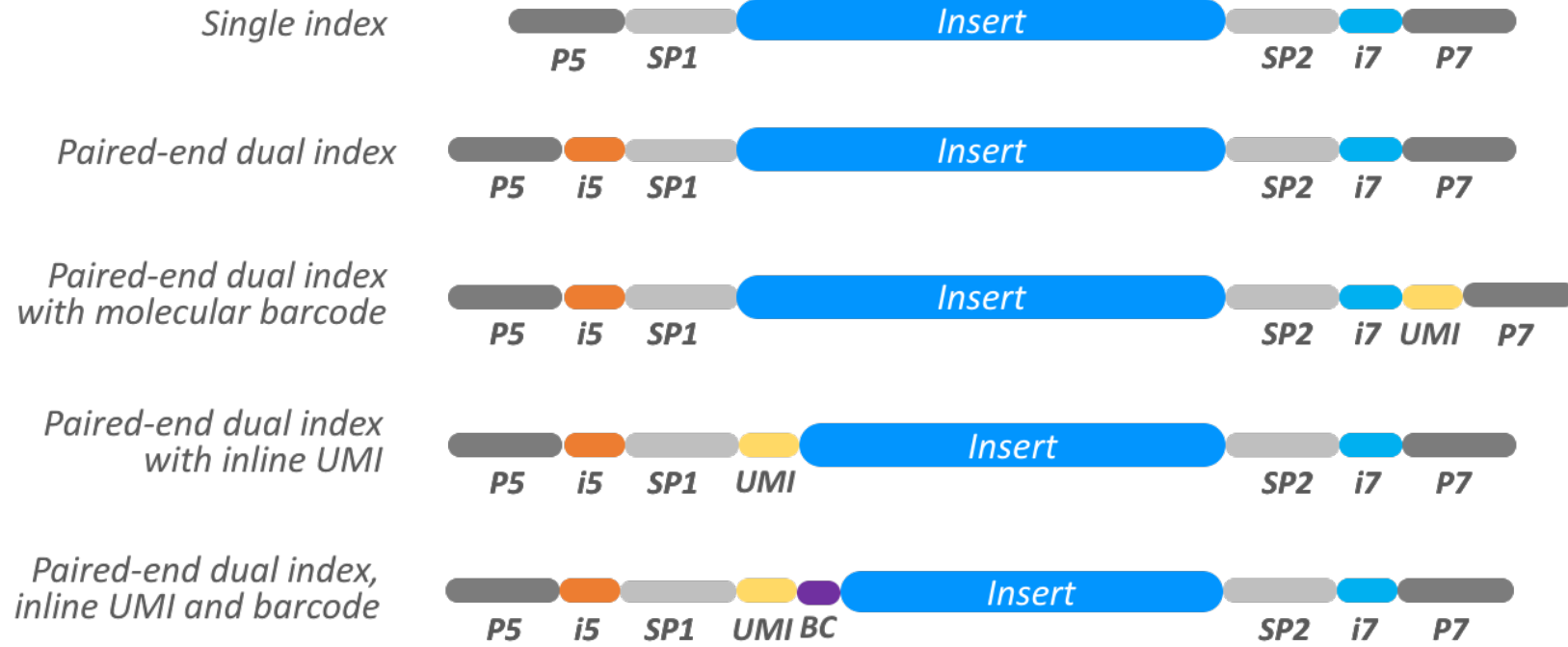
# NGS data analysis



# De-multiplexing

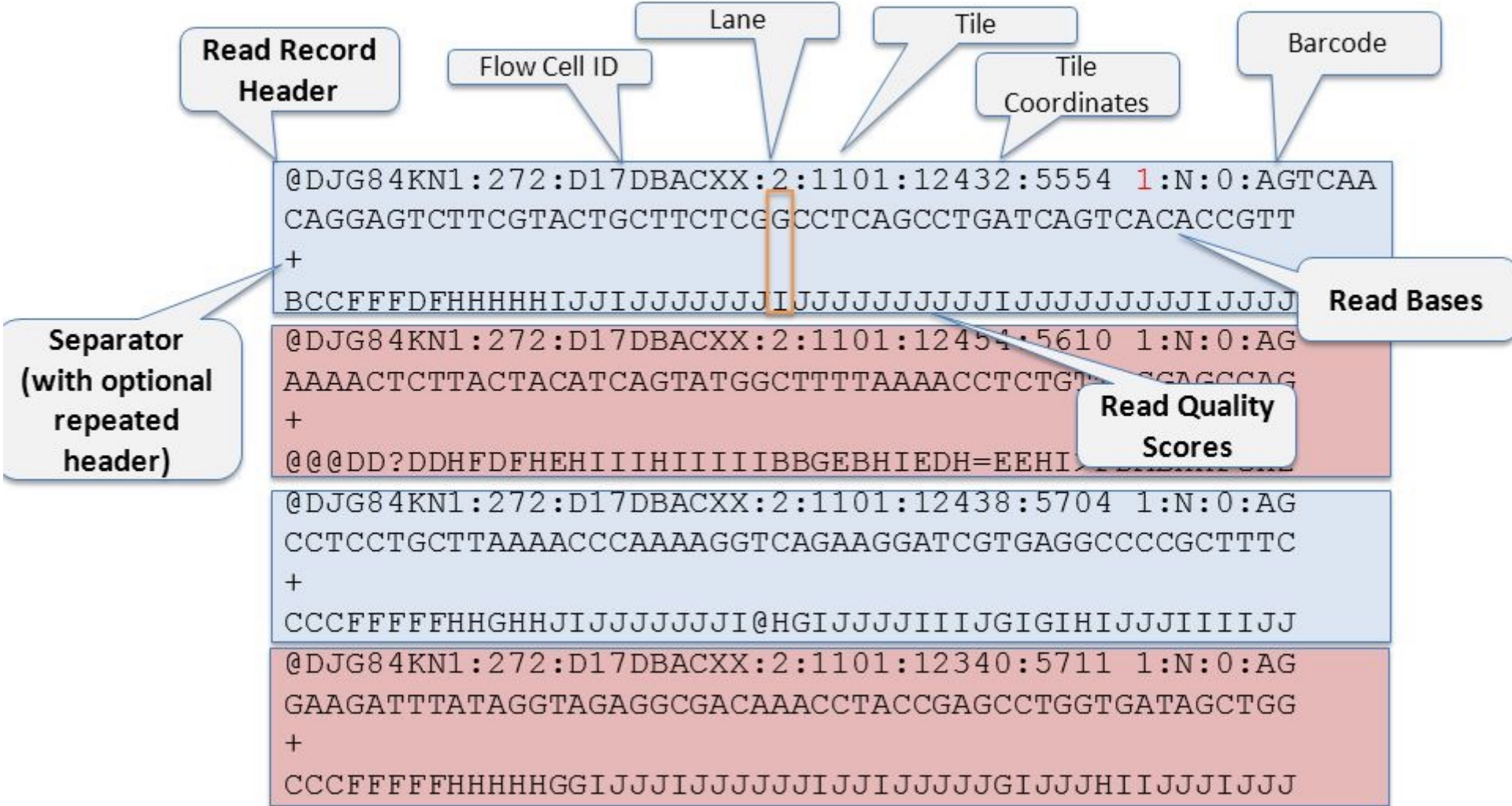


# De-multiplexing



- P5/P7:** Flow cell binding sequences (platform-specific)
- SP1/2:** Sequencing primer binding sites (common for all libraries)
- i5/i7:** Sample Indexes (specific to a particular library)
- UMI:** Unique molecular index (barcode tag for individual molecules)
- BC:** User-defined barcode (unique per sample, single cell, etc.)
- Insert:** Target DNA or cDNA fragment (library-specific)

# Primary data – fastq file



**NOTE:** for paired-end runs, there is a second file with one-to-one corresponding headers and reads.



# Fastq – quality control

- How can we summarize this?
- What QC can be done?

```
@M04743:199:000000000-CGG4F:1:1101:16145:1655 1:N:0:233
GGTGCCAGCCGCCGCGTAATACGAAGGTGGCAAGCGTTGTTCCGATTACTGGGCGTACAGGGAGCGTAGGCGGTTGGGTAAGCCCTCCGTGAAATCTCCGGG
+
ABCCFFFCADBGGGGGGGGHGHGGFHHGHGGGAFFHGGGGHHHHHHHGGGGHGGGGGGGGHGGEGGGGGHHHHHHHGGHGGHHHHHHHGGG
|M04743:199:000000000-CGG4F:1:1101:18938:1729 1:N:0:233
GGTGCCAGCCGCCGCGTAATACGTAGGTCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTCGCCAGGCTGTTTTGTAAGTCAGATGTGAAATCCCCGAG
+
BBBBBFFB4CCGGGGGGGCFHHHHHGGHGGGGGGGGHGGEGFHHHHHHHGGGGHFGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGGG
|M04743:199:000000000-CGG4F:1:1101:14968:1984 1:N:0:233
GGTGCCAGCCGCCGCGTACTACGTAGGTCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCCGCTCGTTGTGAAAACCCGGGG
+
BBBBBFFB4CCGGGGGGGCFHHHHHGGHGGGGGGGAFGHGG?EFHFEHHHHHGGGGFHFHFGHGGHGG3EEEGGGHHEGGGGGGDHEHGHGGGGGGG
F9FFFFFFFFFFFFBFBFBFB; -@DFB-BBBFFFEFF/EBBEFFF/BADFFDFFF.;
|M04743:199:000000000-CGG4F:1:1101:14830:1795 1:N:0:233
GGTGCCAGCCGCCGCGTAATACGTAGGTCGAGCGTTGTCCGATTATTGGGTTTAAAGGTCGCTAGGCGGTTCTTTAAGTCAGTGGTAAATACAGCCG
+
ABBABFBFB?AAEE?EGEFCGGHHFFHGEHFFHHHGGGCFHHGEEGGDFGDHGGGGFDDGGHGGFEGFGGDFGGGGHHFFBGFH34FGBFFHGHGHGFFC
9BD?99-9/9@-BD.;ADFFBF//BBF:FFFFFFED?DFDFF?A.
|M04743:199:000000000-CGG4F:1:1101:14968:1984 1:N:0:233
AGTGCCAGCCGCCGCGTAATACGTAGGTCGAGCGTTGTCCGATTATTGGGTTTAAAGGTCGCTAGGCGGTTCTTTAAGTCAGTGGTAAATACAGCCG
+
BBBBBFFBABBGGGGGGGGHGHGHHGHGGGCFHHGGEGHHHHHGGGGHHHHHGGGGGGGGGGGGHHHHHHHHHHHHHGFHHHHHHHHG
FCHHHGGHHHHHHHHHHHHHHHHHHHHHHHHHGFHHHGGEGFHHGHGGGFEGG9FGGAEGGGGAFDGEFFGGFFBFEFFFFFFFFFFFFFFFFF>DFDBFFI
|M04743:199:000000000-CGG4F:1:1101:12706:2099 1:N:0:233
TGTGCCAGCCGCCGCGTAATACGGAGGAGCTAGCGTTGTTCCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGTTTTTTAAGTCAGAGGTGAAAGCCCGGG
+
BCCCCFFCCCGGGGGGGGGHHEGGGDFGGHHGGGGHGGGGFHHGHHHHHGGGGHHHGGGGGGGGHGHGGGGGGGACGHHHHHHGHGHGHHHHGGGG
BFFFFFFFF9FFFFFFFFFFFFFFFF/
|M04743:199:000000000-CGG4F:1:1101:13747:2260 1:N:0:233
CGTGCCAGCCGCCGCGTAATACGAAGGGGCTAGCGTTGTTCCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCCGCTCGTTGTGAAAACCCGGGG
+
CCCCCFFCABCGGGGGGGGGHGFCEGGGGHHGGGEFHHGGGFHHFHHHHGGGGHH@GHHHGGHGGHGGGGGGF/>CFCGGGGHHHHHFGGGGGG
A@FFFFFFFFFFFFBF9C;=CF.@;CDFFFFFFBDFFFFFFF?BEFFFFFFFFFFFFFFFF?
|M04743:199:000000000-CGG4F:1:1101:20151:2263 1:N:0:233
TGTGCCAGCCGCCGCGTAATACGTAGGTCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTCGCCAGGCTGTTTTGTAAGTCAGATGTGAAATCCCCGAG
+
BBBBBFFBAAADGGGGGGGGGGHHHHHGGHGGGGGGGGGGHGGDFHHHHHHHGGGGGGHGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGFG
|M04743:199:000000000-CGG4F:1:1101:17232:2363 1:N:0:233
GGTGCCAGCCGCCGCGTAATACGGAGGGGCTAGCGTTGTTCCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGATCGGAAAGTCAGAGGTGAAATCCAGGG
+
BBBBBFFB4CCGGGGGGGCFHHHHHGGHGGGGGGGGGGHGGDFHHHHHHHGGGGGGHGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGG
```



# FastQC Report

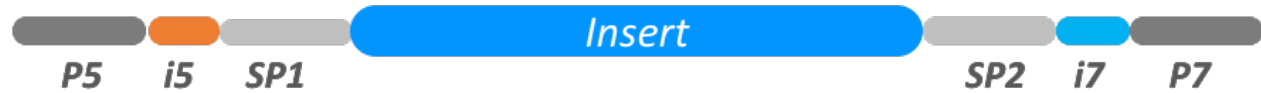
## Summary

Return to [start page](#)

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ! [Per base sequence content](#)
- ! [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Adapter Content](#)

## ✓ Basic Statistics

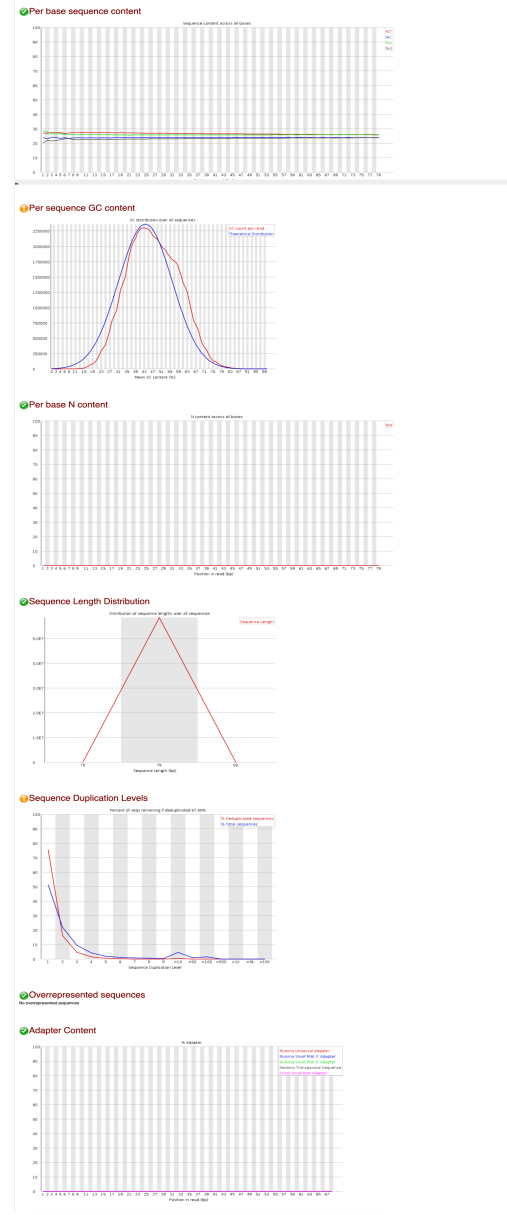
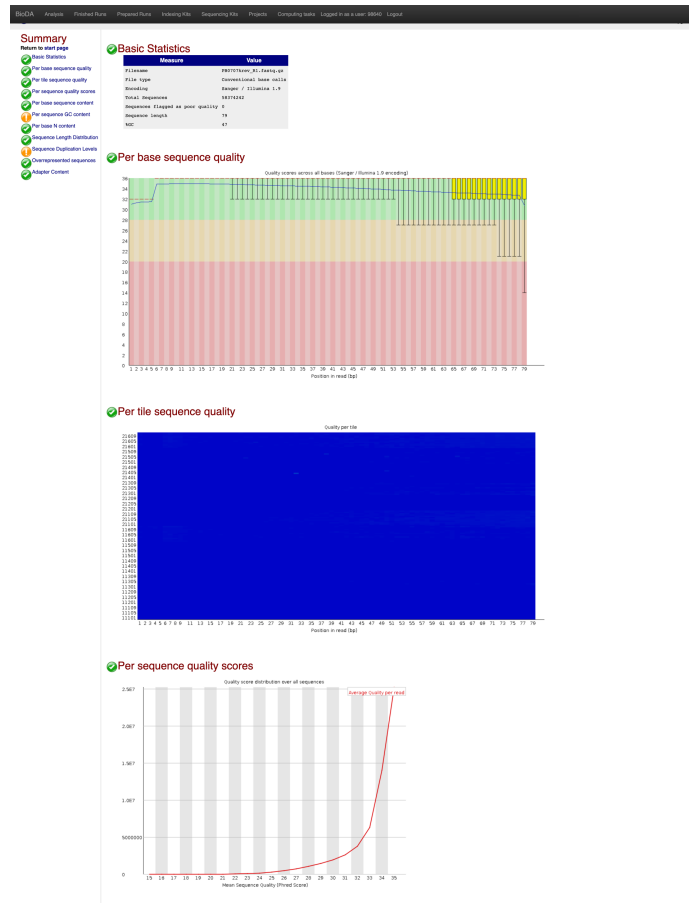
Measure	Value
Filename	MU_a_ytHl_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	252819865
Sequences flagged as poor quality	0
Sequence length	161
%GC	40





# Fastq – quality control

- Fastqc - tool





CEITEC



@CEITEC\_Brno

Thank you for your attention!

