MUNI | RECETOX

Research infrastructure

Workflow & Computatinal Environmnet

E5444 Analysis of sequencing data

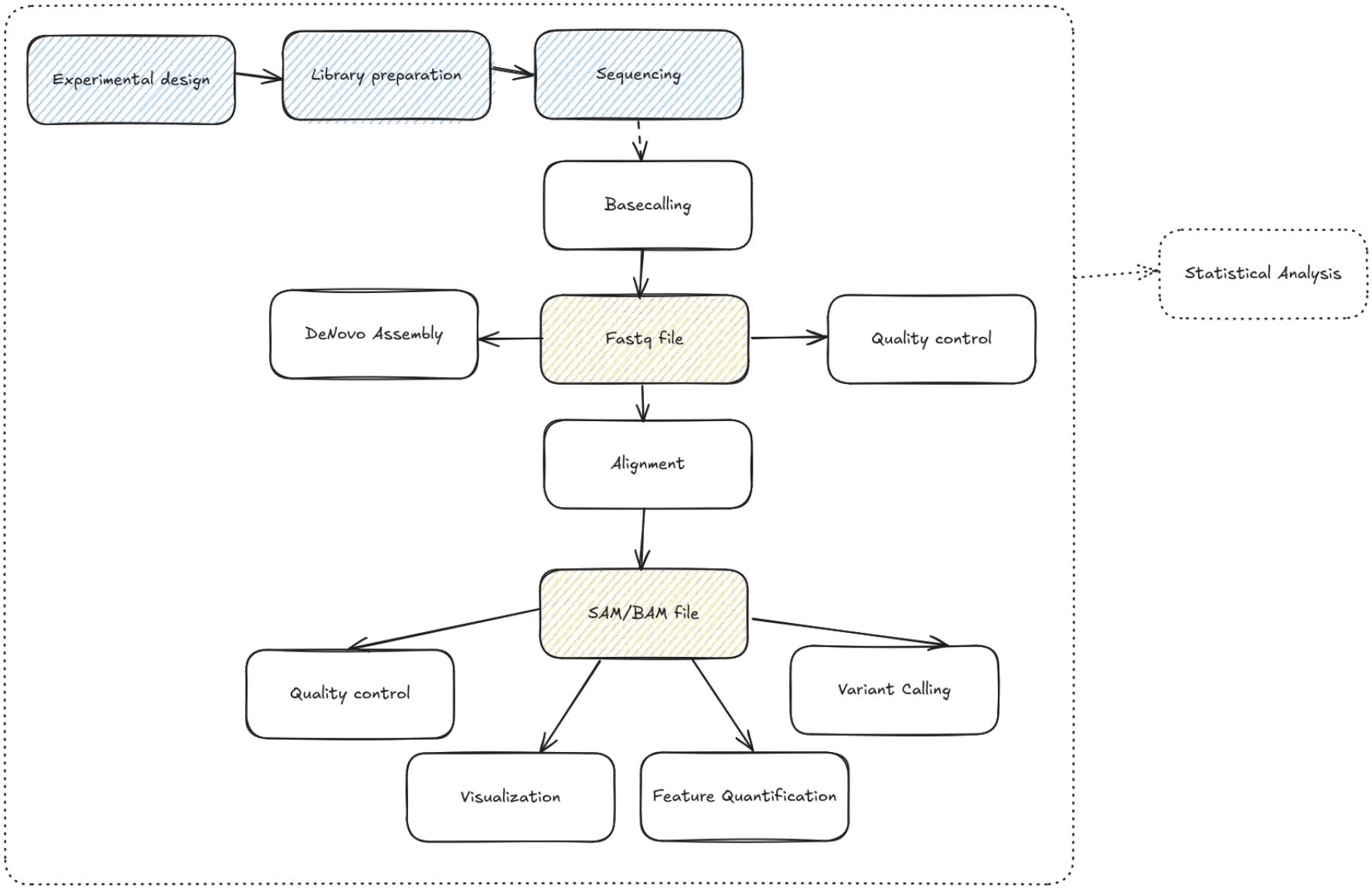
Vojtěch Bartoň
vojtech.barton@recetox.muni.cz

RECETOX, Masaryk University

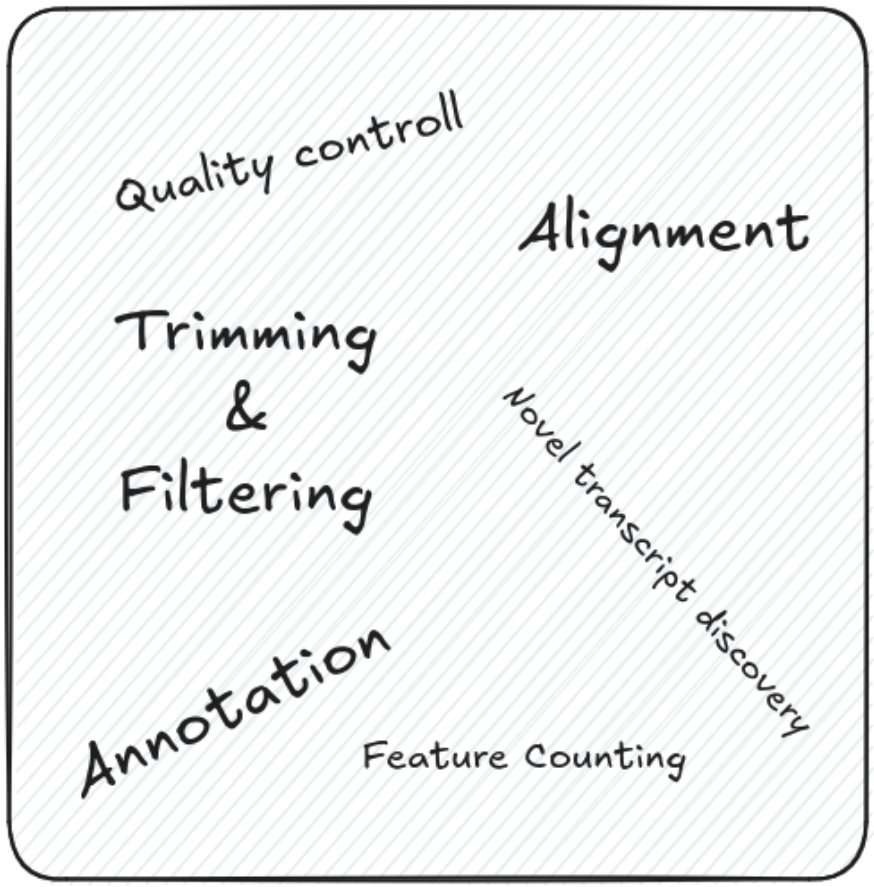
October 2, 2024

Table of Contents

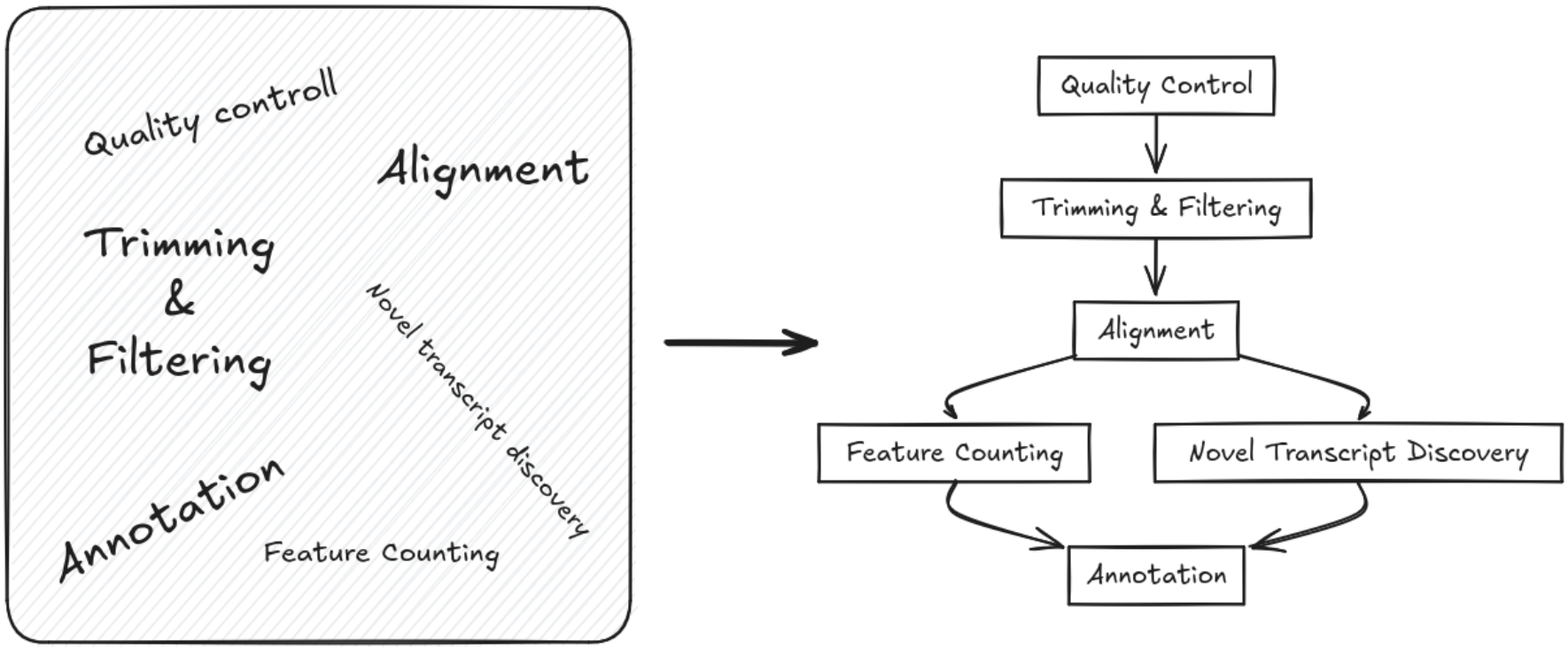
Bioinformatics workflow



Bioinformatics workflow



Bioinformatics workflow



Bioinformatics Morkflow Managers

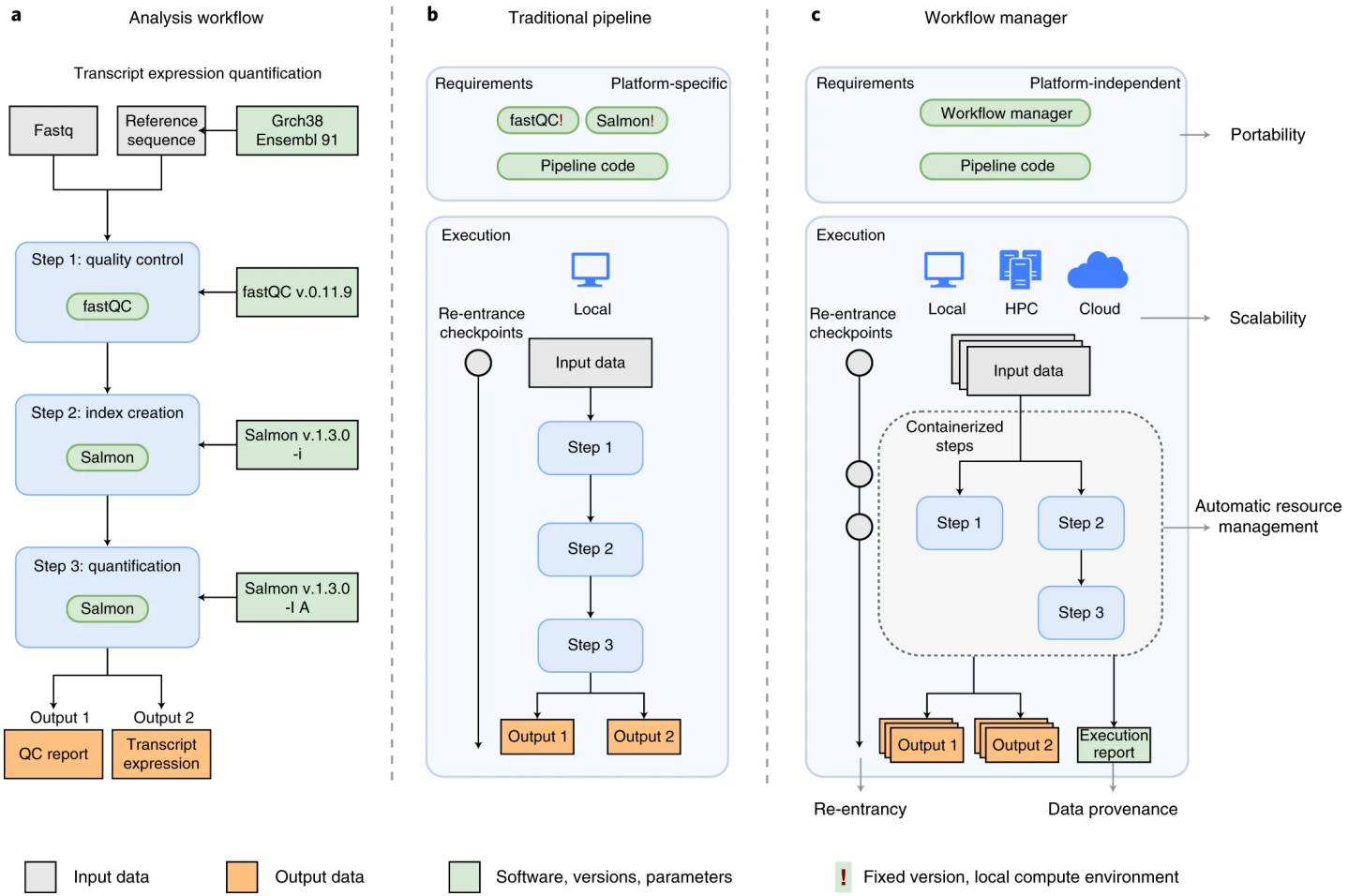


Figure: Wratten, L., Wilm, A. & Göke, J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods 18, 1161–1168 (2021). https://doi.org/10.1038/s41592-021-01254-9

Bioinformatics Workflow Managers

- Modularity
- Scalability
- Reusability & Reproducibility
- Stability
- Loging, debuging

Bioinformatics Workflow Managers

- CWL
- SnakeMake
- Nextflow
- Galaxy project

CWL (Common Worflow Language)

- General worflow definition language
- format: YAML

```
#!/usr/bin/env cwl-runner

cwlVersion: v1.0
class: CommandLineTool
baseCommand: [bwa, mem]
requirements:
  DockerRequirement:
    dockerPull: biocontainers/bwa:v0.7.17_cv1
inputs:
  reference_fasta:
    type: File
    inputBinding:
      position: 1
  input_reads:
    type: File
    inputBinding:
      position: 2
outputs:
  output_bam:
    type: File
    outputBinding:
      glob: "*.bam"
  secondaryFiles:
    - samtools_sort:
        run: samtools
        in:
          input_bam: $(inputs.output_bam)
        out: [sorted_bam]
```

SnakeMake

- General worflow manager
- Pythonic
- Large community

```
rule all:
    input:
        "results/sorted.bam"

rule bwa_mem:
    input:
        reference="data/reference.fasta",
        reads="data/reads.fastq"
    output:
        "results/aligned.sam"
    shell:
        "bwa mem {input.reference} {input.reads} > {output}"

rule samtools_sort:
    input:
        "results/aligned.sam"
    output:
        "results/sorted.bam"
    shell:
        "samtools sort {input} -o {output}"

rule index_bam:
    input:
        "results/sorted.bam"
    output:
        "results/sorted.bam.bai"
    shell:
        "samtools index {input}"
```

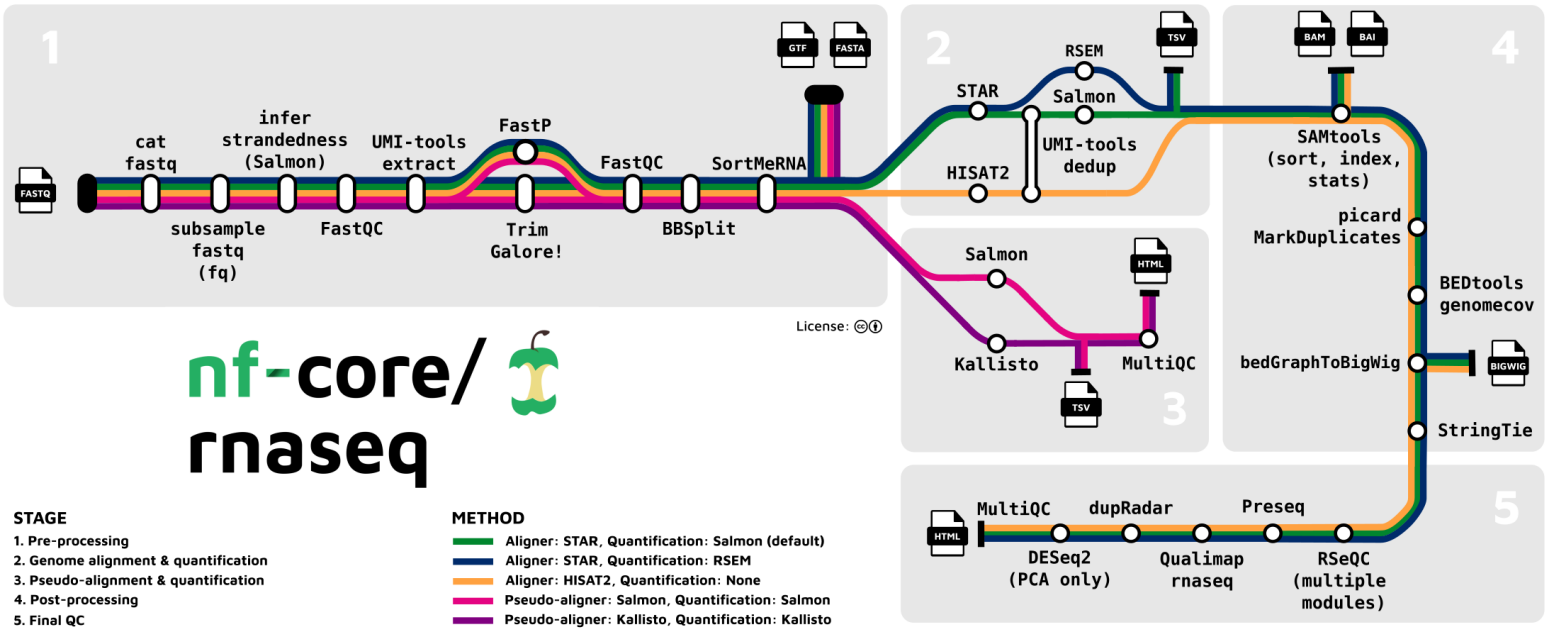
NextFlow

- Bioinformatics manager
- Language: Groovy
- Large community
- Designed for bioinformatics

```
#!/usr/bin/env nextflow

params.reads = "data/reads.fastq"
params.reference = "data/reference.fasta"

process BwaAlign {
    input:
    path reference
    path reads

    output:
    path "aligned.sam"

    """
    bwa mem ${reference} ${reads} > aligned.sam
    """
}

process SamtoolsSort {
    input:
    path "aligned.sam"

    output:
    path "sorted.bam"

    """
    samtools sort aligned.sam -o sorted.bam
    """
}

workflow {
    BwaAlign(params.reference, params.reads)
    SamtoolsSort()
}
```

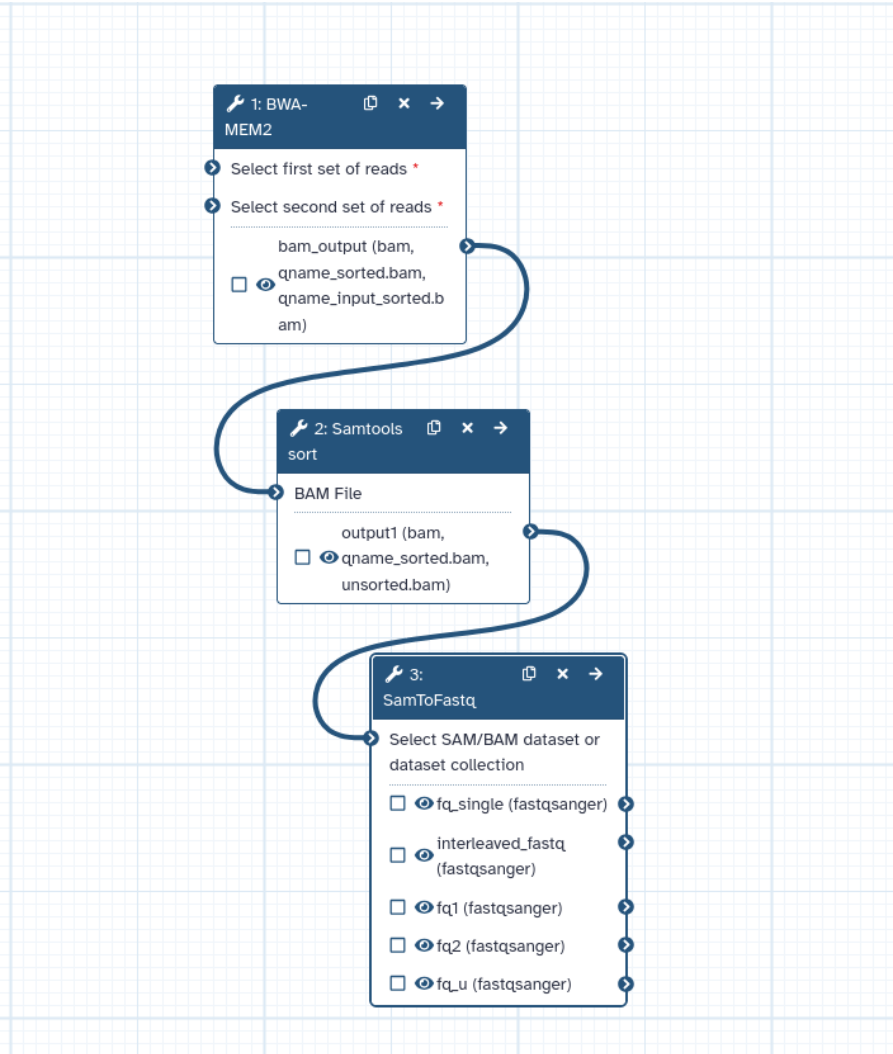
Nextflow - nf-core

- Open community project
- Large repository of pipelines



Galaxy project

- Bioinformatics manager
- Graphical
- Large community
- usegalaxy.eu | usegalaxy.cz

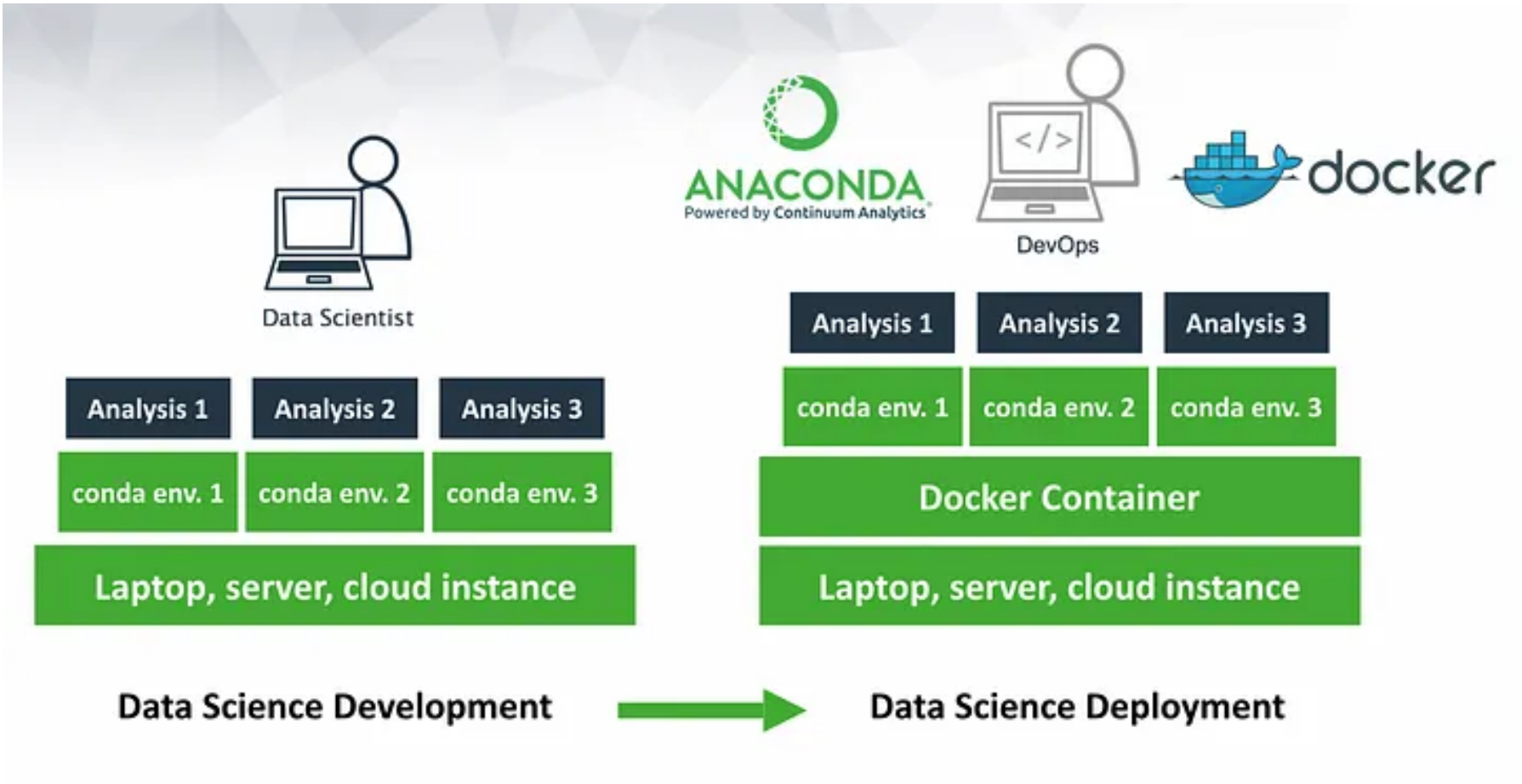


Workflow Environments

- **Conda**:
  - Lightweight package/environment management.
  - Ensures reproducibility by creating isolated environments with specific dependencies.
- **Docker**:
  - Containerization for full software environments.
  - Ensures portability and reproducibility across different systems.
- **Virtual Machines (VMs)**:
  - Full operating system virtualization.
  - Ideal for running workflows with complex or legacy dependencies on isolated OS environments.
- **Grid Computing**:
  - Distributes tasks across a network of computers.
  - Facilitates high-throughput and large-scale computations.

Orchestration levels

Orchestration levels



Metacentrum

High-Performance Computing for Research

- **What is MetaCentrum?**
  - National grid infrastructure operated by CESNET.
  - Provides computational resources for research and education in the Czech Republic.
- **Key Features**:
  - **High-Performance Computing (HPC)**: Access to powerful computing clusters.
  - **Cloud Services**: Virtual machines, storage, and customized environments for specific workflows.
  - **Grid Computing**: Distributed computing resources across multiple sites for large-scale projects.

Metacentrum

- **Storage capacity**:
  - Providing storage capacity for data.
  - Several type of storages.
- **Supported Domains**:
  - Bioinformatics, Physics, Chemistry, Climate Modeling, Machine Learning, and more.
- **User Access**:
  - Free for academic institutions in the Czech Republic.
  - Web-based interface, SSH access, and job scheduling via PBS.

On Demand

- Web-based platform for accessing and running applications on MetaCentrum resources without needing command-line expertise.
- Pre-configured Applications: Access to a wide range of scientific applications (e.g., RStudio, Jupyter, MATLAB).
- ondemand.metacentrum.cz

Command-line access

- Front-end
- Storages
- PBS Scheduler
- Batch job | Interactive job

Basic Linux Commands

Overview of Linux Commands

- Command-line interface (CLI) used for interacting with the operating system.
- Efficient for managing files, directories, and processes.
- Essential for bioinformatics workflows and high-performance computing.

File and Directory Navigation

Common Commands:

- `pwd` - Print current working directory.
- `ls` - List files and directories.
- `cd` - Change directory.
- `mkdir` - Create a new directory.
- `rmdir` - Remove an empty directory.

File Manipulation

Common Commands:

- `cp` - Copy files and directories.
- `mv` - Move or rename files and directories.
- `rm` - Remove files or directories.
- `touch` - Create an empty file or update file timestamps.
- `cat` - Concatenate and display file content.

Working with Compressed Files

- `gzip`, `gunzip` – Compress or decompress files (common with FASTQ and VCF formats).
- `tar` – Archive and extract multiple files (tar -xvf, tar -czvf).
- `zcat`, `zgrep` – View or search within compressed files without uncompressing them.

File Permissions

Common Commands:

- `chmod` - Change file permissions (read, write, execute).
- `chown` - Change file owner or group.
- `ls -l` - List files with detailed permissions.
- `umask` - Set default file permissions.

Searching and Finding Files

Common Commands:

- `find` - Search for files in a directory hierarchy.
- `grep` - Search for text patterns within files.
- `locate` - Quickly find file locations using a database.
- `which` - Show the location of an executable command.
- `man` - Display manual pages for command help.

Networking Commands

Common Commands:

- `ping` - Check connectivity to a host.
- `ifconfig` - Display or configure network interfaces.
- `ssh` - Secure shell access to a remote machine.
- `scp` - Securely copy files between hosts.
- `wget` - Download files from the web.

Software Management

- `conda` – Manage bioinformatics software environments.
- `module ava` – Availibility of software modules on HPC systems like MetaCentrum.
- `module load` – Load software modules on HPC systems like MetaCentrum.
- `apt-get`, `yum` – Install software on Linux systems (depends on the package manager).

Job Scheduling (PBS)

- `qsub` - Submit jobs on PBS-based HPC systems.
- `qstat` - Monitor jobs on PBS-based HPC systems.
- `qdel` - Manage jobs on PBS-based HPC systems.

Hands-on

- docs.metacentrum.cz/computing/concepts/
- Try some commands in terminal
- Submit interactive job
- Submit batch job
- Explore outputs

MASARYK
UNIVERSITY