SCIENCE MEETS LIFE

Nextflow at the VSC

VIB-KU LEUVEN CENTER FOR BRAIN & DISEASE RESEARCH

KU LEUVEN

25/11/2024

Kris Davie

Workflow Managers

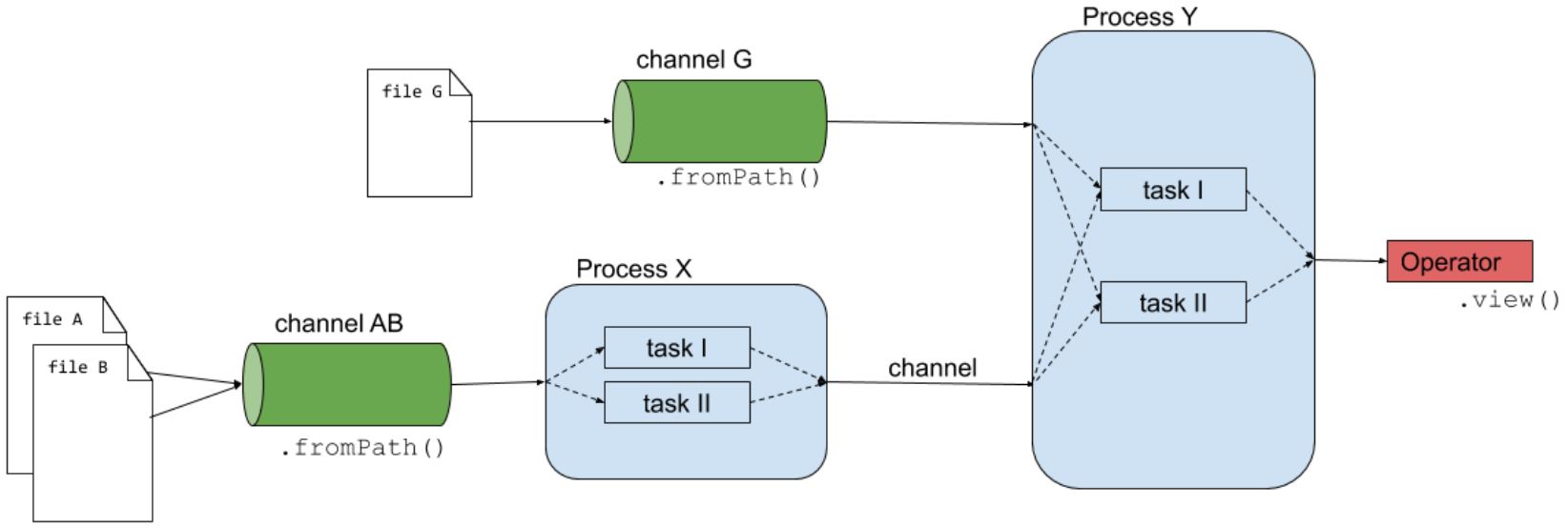
- Nextflow is a bioinformatics workflow manager that enables the development of portable and reproducible workflows.

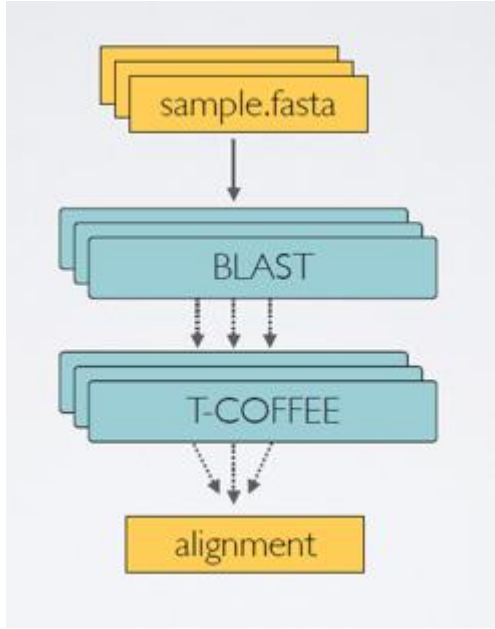
- Nextflow is a reactive workflow framework and a programming Domain Specific Language that eases the writing of data-intensive computational pipelines

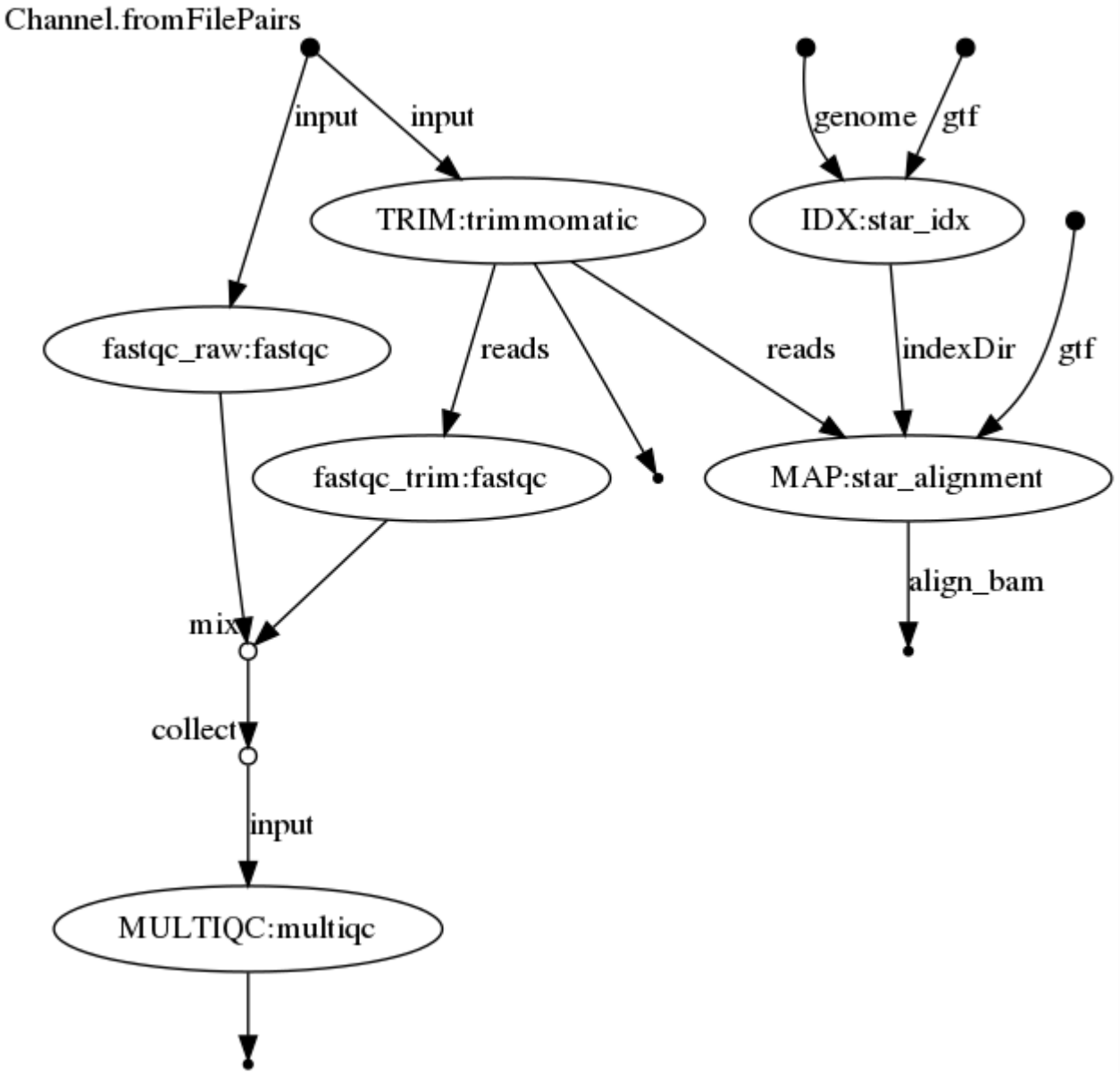
Nextflow Pros and Cons

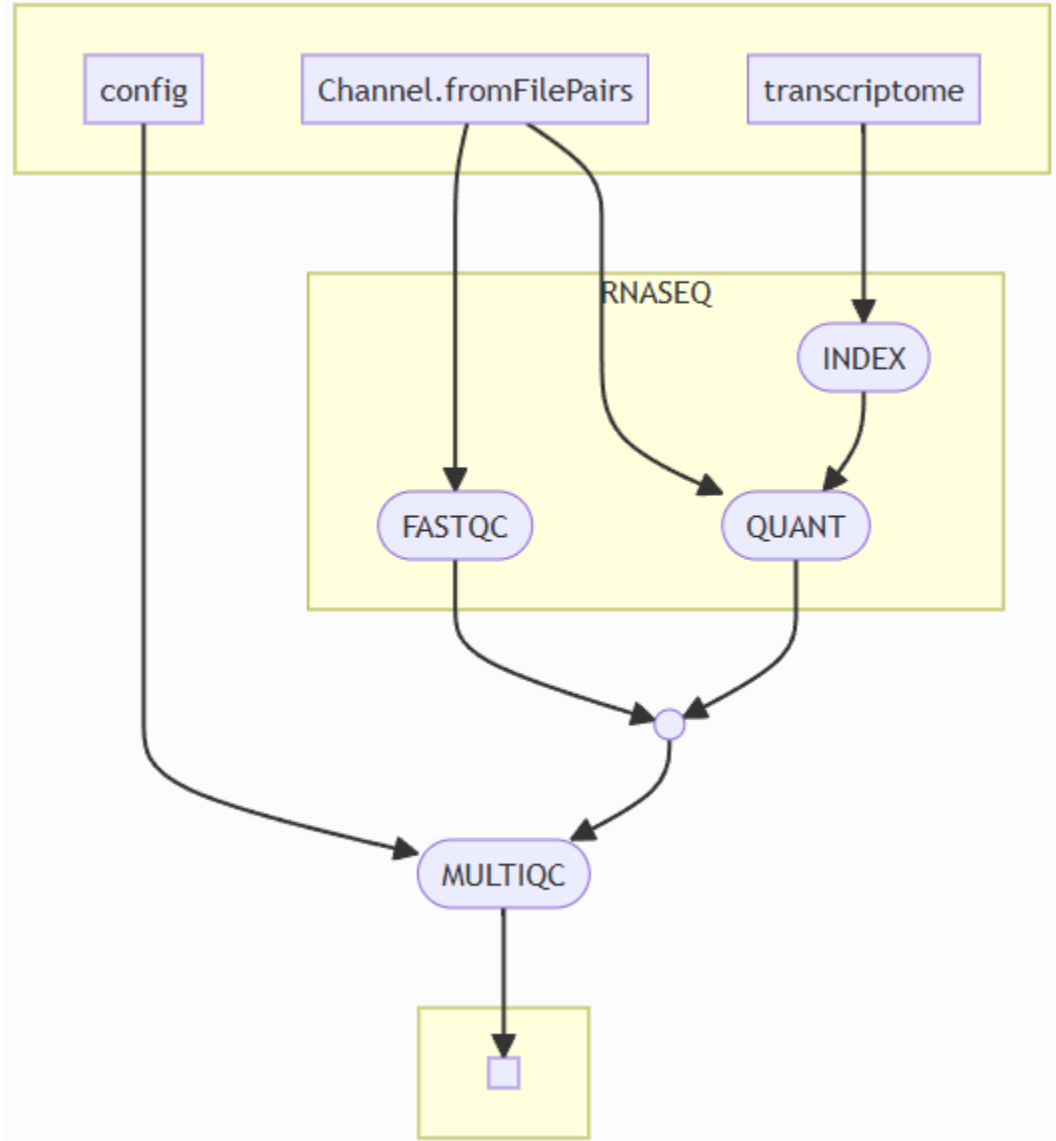
- Pros:
  - Parallelization
  - Scalability
  - Portability
  - Reproducible
  - Continuous checkpoints
  - Modularity
  - Community
- Cons:
  - Groovy -> Another, non-bioinformatics language
  - Complexity
  - Debugging

Nextflow Concepts









VIB-KU LEUVEN CENTER FOR BRAIN & DISEASE RESEARCH

KU LEUVEN

SCIENCE MEETS LIFE

Example Script

shebang

Channel definitions

Process definition
Input channel definitions

Output channel definitions

Process script

Workflow definition
Process call

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

// Creating channels
numbers_ch = Channel.of(1,2,3)
strings_ch = Channel.of('a','b')

// Defining the process that is executed
process valuesToFile {
    input:
    val nums
    val strs

    output:
    path 'result.txt'

    """
    echo $nums and $strs > result.txt
    """
}

// Running a workflow with the defined processes
workflow {
    valuesToFile(numbers_ch, strings_ch)
}
```

Processes in other Languages

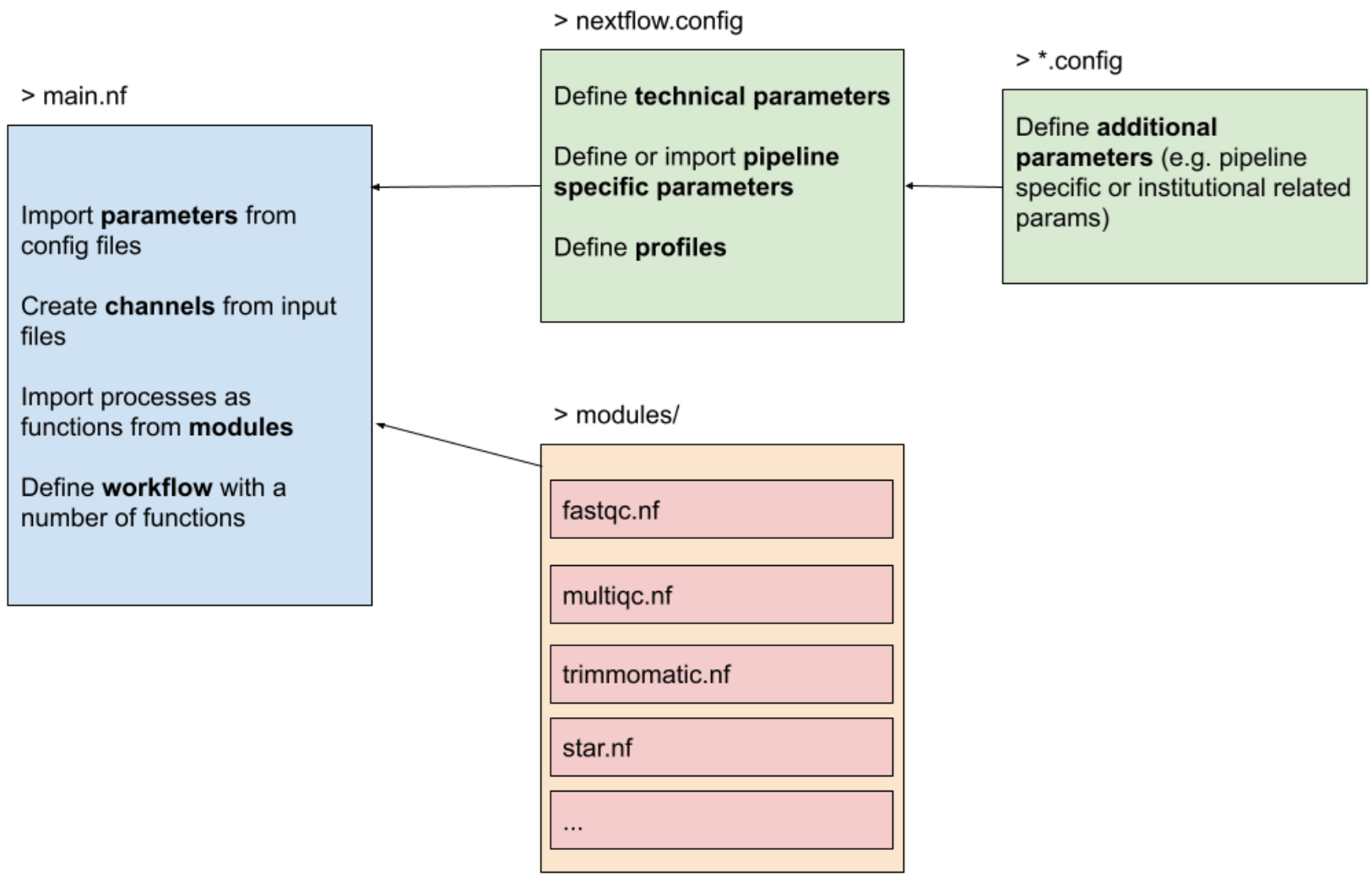
A script, as part of the process, can be written in any language (bash, Python, Perl, Ruby, etc.). This allows to add self-written scripts in the pipeline.

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

process python {

    script:
    """
    #!/usr/bin/env python3

    firstWord = 'hello'
    secondWord = 'folks'
    print(f'{firstWord} {secondWord}')
    """
}
```


> nextflow.config
> main.nf
> *.config
Define **technical parameters**
Define or import **pipeline specific parameters**
Define **profiles**
Define **additional parameters** (e.g. pipeline specific or institutional related params)
Import **parameters** from config files
Create **channels** from input files
Import processes as functions from **modules**
Define **workflow** with a number of functions
> modules/
fastqc.nf
multiqc.nf
trimmomatic.nf
star.nf
...

Interacting with Nextflow

```
nextflow [options] COMMAND [arg...]
```

module load Nextflow/23.04.2

e.g.

```
nextflow run hello.nf
```

```
nextflow -C file.config run hello/world
```

```
nextflow -C file.config run hello/world --name Kris
```

Specifying Software

Support for:

- Conda
- Containers
  - Podman (docker)
  - Apptainer (singularity)
- Modules

```
process runWithCondaMakeEnv {
  conda 'bwa=0.7.15'
    or
  conda '/path/to/env.yaml'
```

```
process runWithCondaExistingEnv {
  conda '/path/to/env/dir'
```

```
process runWithDockerOrApptainer {
  container 'quay.io/biocontainers/star:2.7.11a'
```

```
process runWithApptainerImg {
  container '/path/to/images/star_2.7.11a.sif'
```

```
process runWithModules {
  module STAR/2.7.11a:SAMtools/1.13'
```

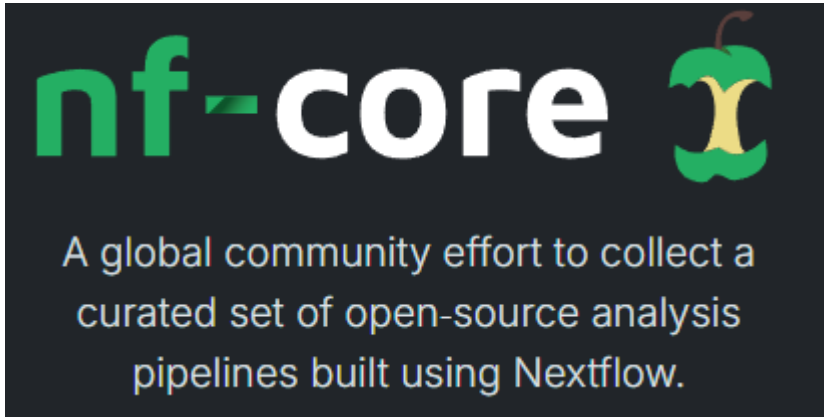
Controlling computational resources

Processes can be labelled to control resources that they request.

```
process runWithHighGPU {
  label ‘high’
```

```
process runWithHighCPU {
  label ‘gpu’
```

```
process {
    withLabel: 'low' {
        memory = ‘10G'
        cpus = ‘8'
        time = '6h'
    }
    withLabel: 'high' {
        memory = ‘180G'
        cpus=‘64'
    }
    withLabel: ‘gpu' {
        clusterOptions = ‘--gpus-per-node=1’
        queue = ‘gpu'
    }
}
```



118 pipelines available
Huge community
VSC institutional config

rnaseq ✓ ☆ 921

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

rna · rna-seq

3.17.0 released 29 days ago

sarek ✓ ☆ 410

Analysis pipeline to detect germline or somatic variants (pre-processing, variant calling and annotation) from WGS / targeted sequencing

annotation · cancer · gatk4 · genomics · germline · pre-processing · somatic · target-panels · variant-calling · whole-exome-sequencing · whole-genome-sequencing

3.4.4 released 3 months ago

atacseq ✓ ☆ 188

ATAC-seq peak-calling and QC analysis pipeline

atac-seq · chromatin-accessibiity

2.1.2 released over 1 year ago

nanoseq ✓ ☆ 180

Nanopore demultiplexing, QC and alignment pipeline

alignment · demultiplexing · nanopore · qc

3.1.0 released over 1 year ago

mag ✓ ☆ 217

Assembly and binning of metagenomes

annotation · assembly · binning · long-read-sequencing · metagenomes · metagenomics · nanopore · nanopore-sequencing

3.2.1 released 23 days ago

scrnaseq ✓ ☆ 214

A single-cell RNAseq pipeline for 10X genomics data

10x-genomics · 10xgenomics · alevin · bustools · cellranger · kallisto · rna-seq · single-cell · star-solo

2.7.1 released 3 months ago

fetchngs ✓ ☆ 151

Pipeline to fetch metadata and raw FastQ files from public databases

ddbj · download · ena · fastq · geo · sra · synapse

1.12.0 released 9 months ago

eager ✓ ☆ 148

A fully reproducible and state-of-the-art ancient DNA analysis pipeline

adna · ancient-dna-analysis · ancientdna · genome · metagenomics · pathogen-genomics · population-genetics

2.5.2 released 5 months ago

chipseq ✓ ☆ 195

ChIP-seq peak-calling, QC and differential analysis pipeline.

chip · chip-seq · chromatin-immunoprecipitation · macs2 · peak-calling

2.1.0 released about 2 months ago

ampliseq ✓ ☆ 188

Amplicon sequencing analysis workflow using DADA2 and QIIME2

16s · 18s · amplicon-sequencing · edna · illumina · iontorrent · its · metabarcoding · metagenomics · microbiome · pacbio · qiime2 · rrna · taxonomic-classification · taxonomic-profiling

2.12.0 released 7 days ago

rnafusion ✓ ☆ 144

RNA-seq analysis pipeline for detection of gene-fusions

fusion · fusion-genes · gene-fusion · rna · rna-seq

3.0.2 released 8 months ago

methylseq ✓ ☆ 140

Methylation (Bisulfite-Sequencing) analysis pipeline using Bismark or bwa-meth + MethylDackel

bisulfite-sequencing · dna-methylation · em-seq · epigenome · epigenomics · methyl-seq · pbat · rrbs

2.7.1 released 26 days ago

VIB-K CENT BRAIN RESEA

MEETS LIFE

Internal Usage

- Aerts Lab
  - SCENIC: https://github.com/aertslab/scenic-nf
  - vib-singlecell-nf (deprecated): https://github.com/vib-singlecell-nf
- Liu Lab:
  - Fly tracking, video processing
  - GPU and CPU node usage
- VIB Nucleomics core:
  - Demultiplexing pipeline from sequencers
  - Auto-process data (watchPath)

Reproducibility

- Versioned Nextflow pipeline
- Containers
- Configuration file

- Together these form fully reproducible and portable pipelines

Nextflow's Working Directory

- All outputs of processes are stored in a working directory
  - Allows resuming an interrupted pipeline
- Wanted outputs can be published elsewhere

- Working directory can become very large, therefore pipelines should be run from $VSC_SCRATCH, or at least put the workdir there

```
-w $VSC_SCRATCH/nextflow_work
```

Apptainer Cache Directory

- Nextflow will auto-pull + build containers for you
- Set the `apptainer.cacheDir` config to a universal location
  - `$VSC_DATA` is a good option
- Prevents having to download containers multiple times

Running Pipelines

- Running locally
  - `process.executor = 'local'`
  - Can also be set using labels
- Using SLURM
  - `process.executor = 'slurm'`
  - `maxForks` option to limit number of jobs
- Start a light job on a compute node
- Trigger the pipeline from that node
- Many other executors available

Reports

```
$ nextflow log
TIMESTAMP           RUN NAME          SESSION ID                            COMMAND
2016-08-01 11:44:51  grave_poincare   18cbe2d3-d1b7-4030-8df4-ae6c42abaa9c  nextflow run hello
2016-08-01 11:44:55  small_goldstine  18cbe2d3-d1b7-4030-8df4-ae6c42abaa9c  nextflow run hello -resume
2016-08-01 11:45:09  goofy_kilby      0a1f1589-bd0e-4cfc-b688-34a03810735e  nextflow run rnatoy -with-docker
```

Nextflow Report Summary Resources Tasks [angry_babbage]

Nextflow workflow report

[angry_babbage]

Workflow execution completed successfully!

Run times

Sun Nov 05 11:13:07 CET 2017 - Sun Nov 05 13:50:12 CET 2017 (completed 4 days ago, duration: **2h 37m 5s**)

Nextflow command

```
./nextflow-0.26.0-all run main.nf -profile awsbatch -with-report -with-trace -bg --max_samples 38 -w s3://cbcrg-eu/work
```

Launch directory	/home/pditommaso/projects/rnaseq-encode-nf
Work directory	/cbcrg-eu/work
Project directory	/home/pditommaso/projects/rnaseq-encode-nf
Script path	/home/pditommaso/projects/rnaseq-encode-nf/main.nf
Script name	main.nf
Script hash	17440c7357d1792c8d6be8223aaebe92
Workflow container	nextflow/rnaseq-nf
Workflow profile	awsbatch
Nextflow version	version 0.26.0, build 4710 (03-11-2017 18:14 UTC)
Session ID	54bc9227-daaf-482c-9c62-60ad29af7363

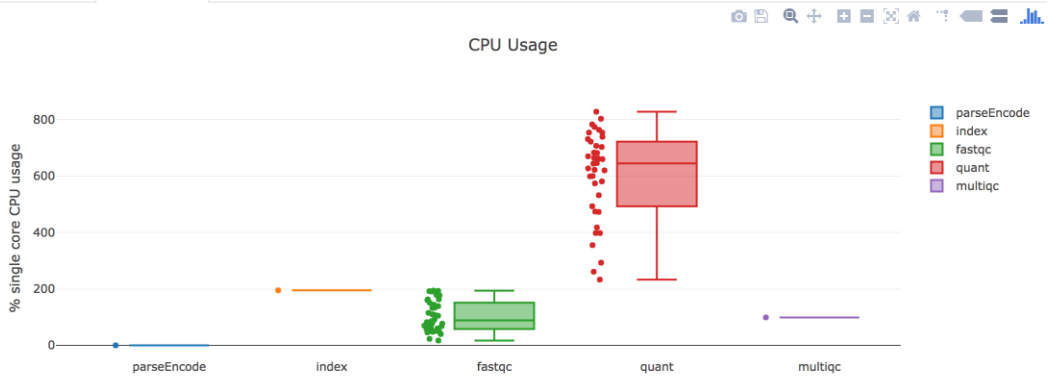
Nextflow Report Summary Resources Tasks [angry_babbage]

Resource Usage

These plots give an overview of the distribution of resource usage for each process.

CPU Usage

% Allocated | Raw Usage



nextflow run <pipeline>
-with-report [file name]

Nextflow Report Summary Resources Tasks [angry_babbage]

Tasks

This table shows information about each task in the workflow. Use the search box on the right to filter rows for specific values. Clicking headers will sort the table by that value and scrolling side to side will reveal more columns.

Show 25 entries Search:

task_id	process	tag	status	hash	allocated cpus	%cpu	allocated memory (bytes)	%mem	vmem	rss
1	index	Homo_sapiens.GRCh38.cdna.all.fa.(	COMPLETED	f4/a72585	2	195.0	8589934592	31.9	5272805376	51318
2	parseEncode	/home/pditommaso/projects/rnasec encode-nf/data/metadata.tsv	COMPLETED	12/bdfd13	1	0.0	-	0.0	17960960	5324
3	fastqc	FASTQC on SRR5210435	COMPLETED	ba/5068a0	2	46.4	6442450944	0.0	4088819712	3685
4	fastqc	FASTQC on SRR3192620	COMPLETED	fa/3e8db3	2	76.7	6442450944	0.0	4089171968	5049
5	fastqc	FASTQC on SRR3192621	FAILED	6b/f753e2	2	-	6442450944	-	-	-
6	fastqc	FASTQC on SRR3192434	COMPLETED	1e/d7f3c2	2	68.8	6442450944	0.0	4088832000	4153
7	fastqc	FASTQC on SRR3192433	COMPLETED	5e/4886ef	2	70.2	6442450944	0.0	4031012864	3843

VIB-KU LEUVEN CENTER FOR BRAIN & DISEASE RESEARCH

KU LEUVEN

SCIENCE MEETS LIFE

Questions?


VIB-KU LEUVEN
CENTER FOR
BRAIN & DISEASE
RESEARCH


KU LEUVEN