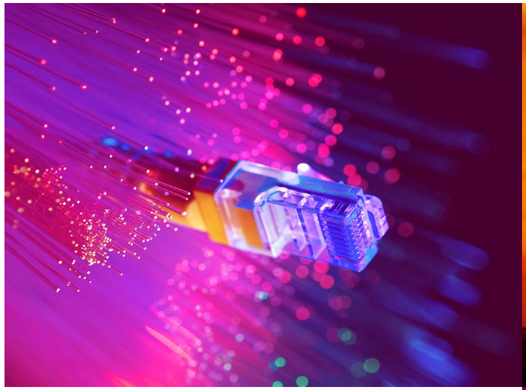


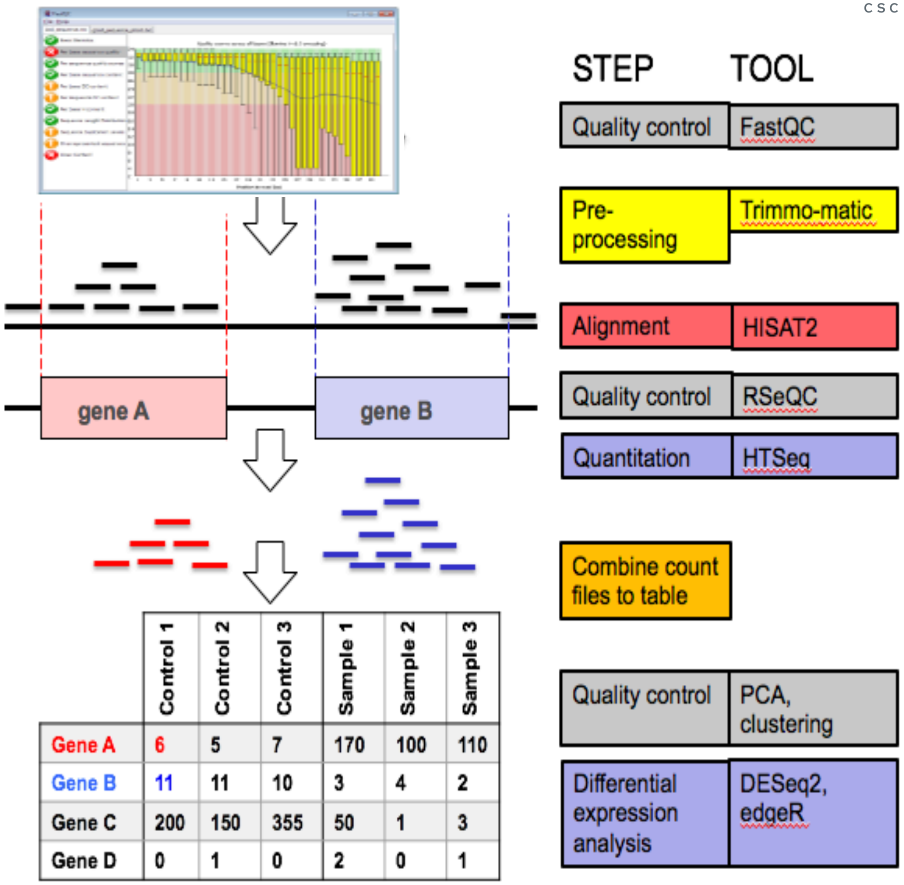
CSC

Introduction to Workflows

CSC – Suomalainen tutkimuksen, koulutuksen, kulttuurin ja julkishallinnon ICT-osaamiskeskus

What is a workflow ?

- A workflow is a collection of several analysis steps
- Steps are linked by input/output files
- One often needs to run the same workflow on several samples



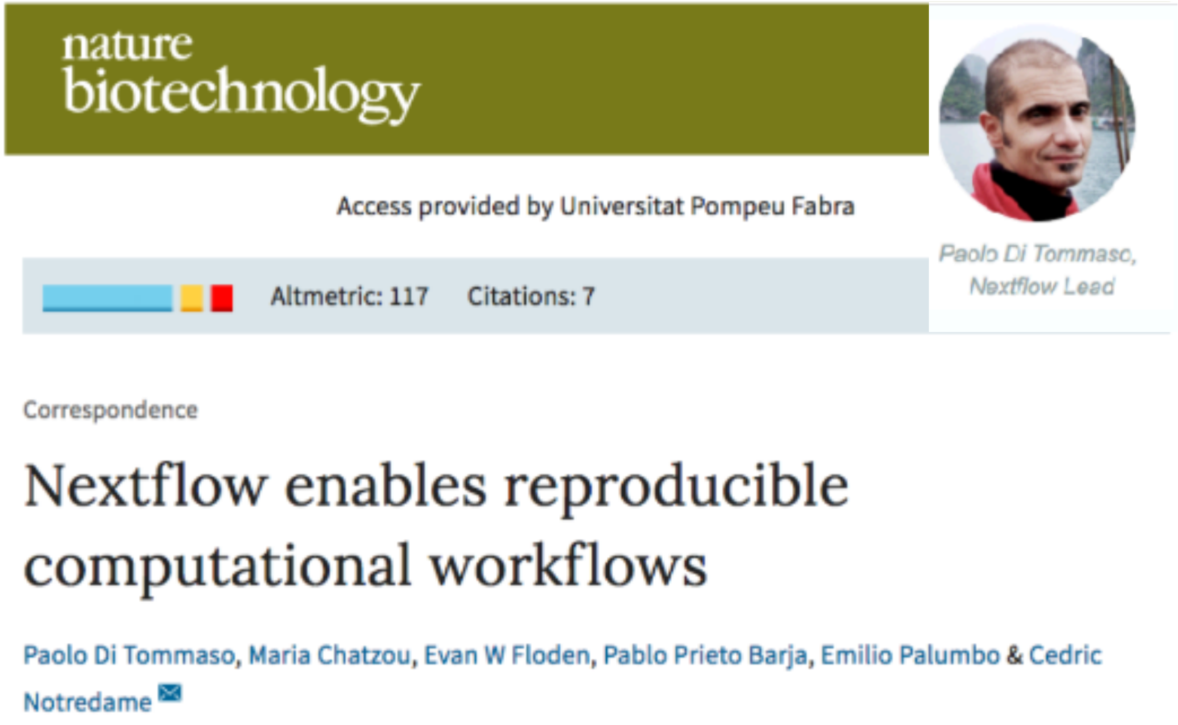
STEP	TOOL
Quality control	FastQC
Pre-processing	Trimmo-matic
Alignment	HISAT2
Quality control	RSeQC
Quantitation	HTSeq
Combine count files to table	
Quality control	PCA, clustering
Differential expression analysis	DESeq2, edgeR

	Control 1	Control 2	Control 3	Sample 1	Sample 2	Sample 3
Gene A	6	5	7	170	100	110
Gene B	11	11	10	3	4	2
Gene C	200	150	355	50	1	3
Gene D	0	1	0	2	0	1

RNAseq pipeline for differential gene regulation

Popular Choices for Bioinformatics Workflows

- Workflows
  - Snakemake
  - Cromwell
  - Nextflow
  - Galaxy



nature biotechnology

Access provided by Universitat Pompeu Fabra

Altmetric: 117 Citations: 7

Paolo Di Tommaso, Nextflow Lead

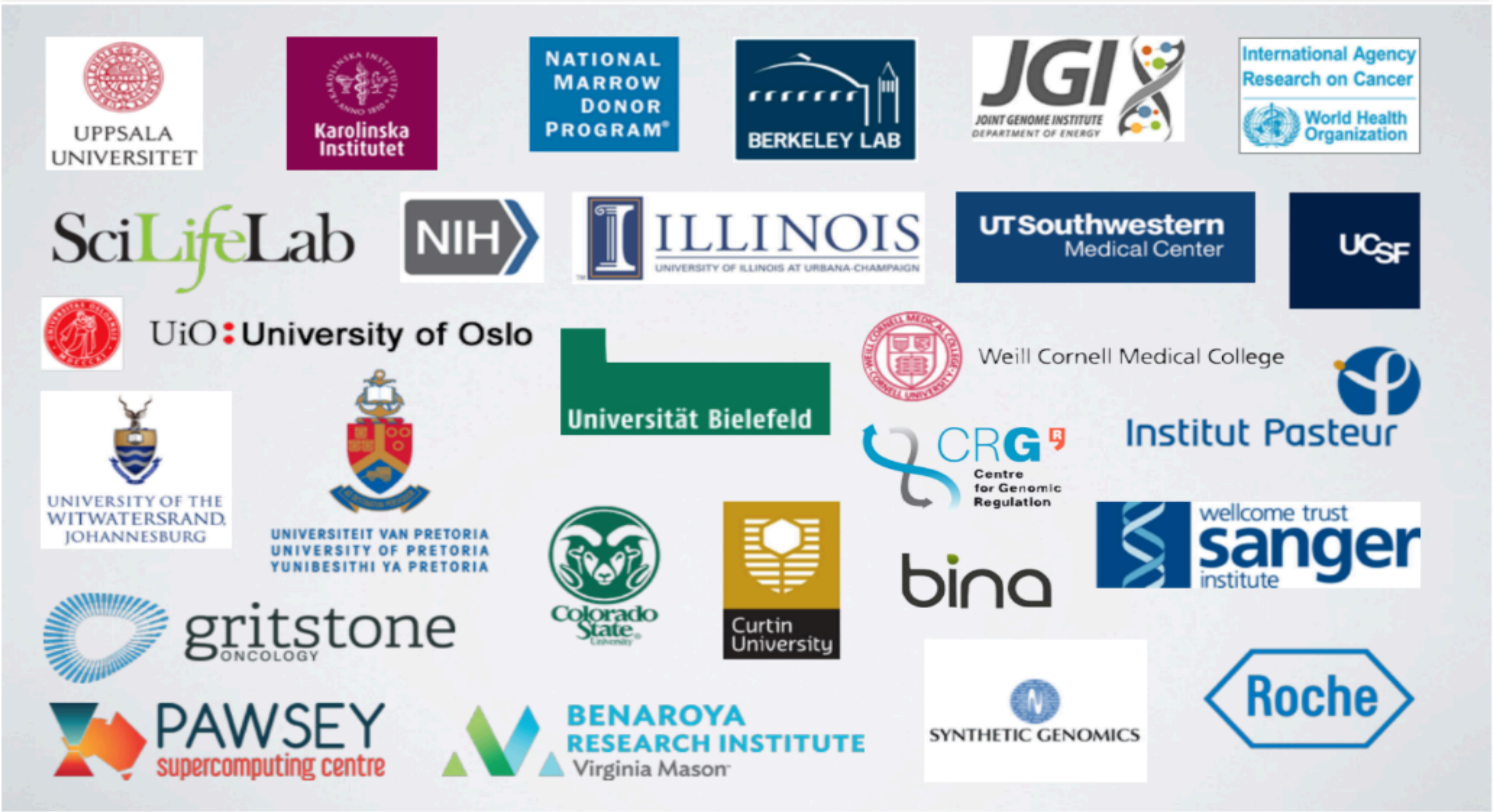
Correspondence

Nextflow enables reproducible computational workflows

Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo & Cedric Notredame

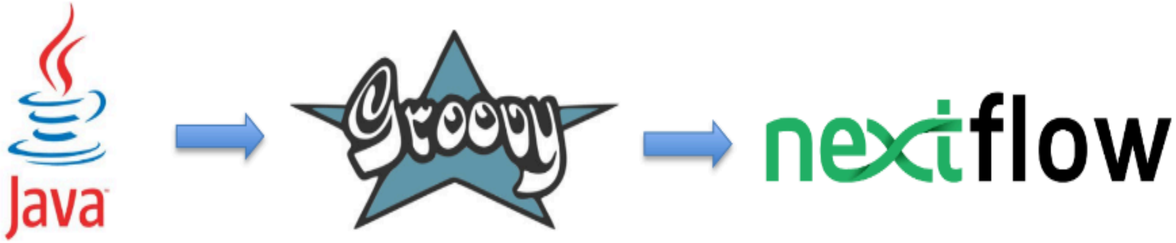
nextflow

Nextflow is Getting Popular


UPPSALA UNIVERSITET
Karolinska Institutet
NATIONAL MARROW DONOR PROGRAM®
BERKELEY LAB
JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY
International Agency Research on Cancer
World Health Organization
SciLifeLab
NIH
ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
UT Southwestern Medical Center
UCSF
UiO : University of Oslo
Universität Bielefeld
Weill Cornell Medical College
Institut Pasteur
UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
CRG Centre for Genomic Regulation
Colorado State University
Curtin University
bina
wellcome trust sanger institute
gritstone ONCOLOGY
PAWSEY supercomputing centre
BENAROYA RESEARCH INSTITUTE Virginia Mason
SYNTHETIC GENOMICS
Roche

What is Nextflow?

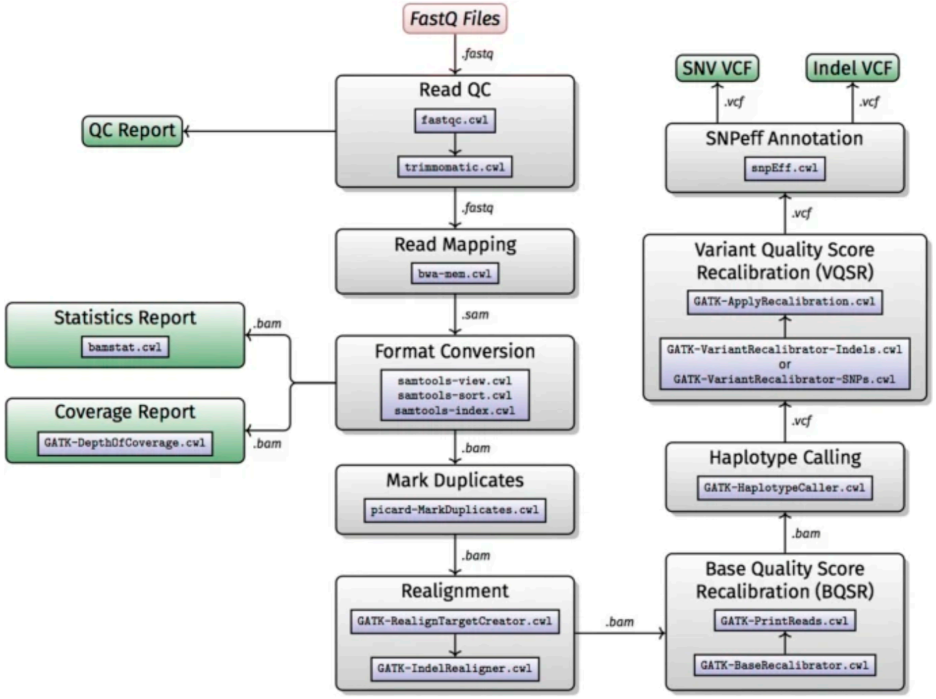
- A tool for managing scientific workflows, written in groovy, a language for java program



- A dataflow programming model
  - Communication by dataflow vairables (channels)
  - Processes (softwares/scripts) receiving (inputs) and emitting (outputs) through channels

Why using Nextflow ?

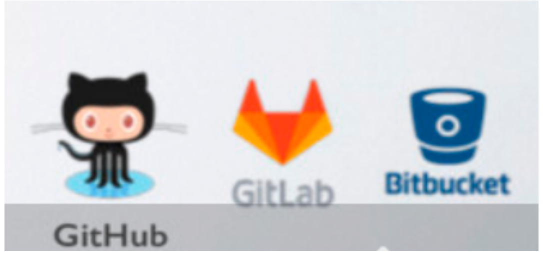
- Workflow Management
- Reproducibility
- Portability
- Scalability
- Parallelisation
- Easy resumption
- Fast prototyping



WGS/WES example from
Baichoo et.al., BMC Bioinformatics,19,457 (2018)

Why using Nextflow ?

- Workflow Management
- Reproducibility
- Portability
- Scalability
- Parallelisation
- Easy resumption
- Fast prototyping



Version control

Suppports integration with containers and Github

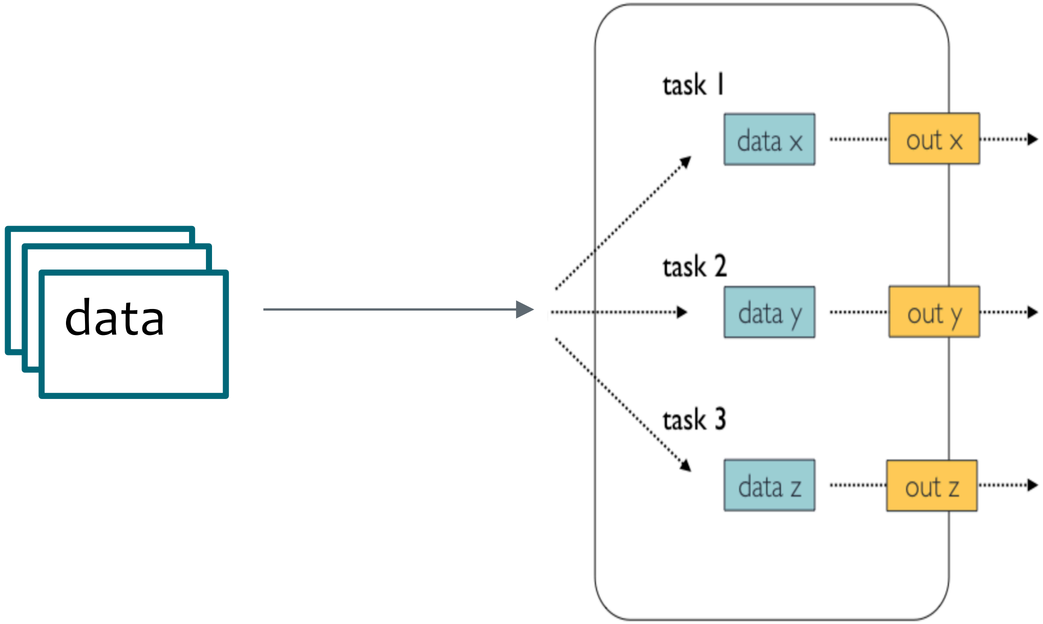
Why using Nextflow ?

- Workflow Management
- Reproducibility
- Portability
- Scalability
- Parallelisation
- Easy resumption
- Fast prototyping

Schedulers
GRID ENGINE
Sun Grid Engine (SGE)
slurm
workload manager
LSF
Platform Load Sharing Facility
PBS Works
Portable Batch System
HTCondor
High Throughput Computing
Cloud platforms
kubernetes
amazon web services

Why using Nextflow ?

- Workflow Management
- Reproducibility
- Portability
- Scalability
- Parallelisation
- Easy resumption
- Fast prototyping



Why using Nextflow ?

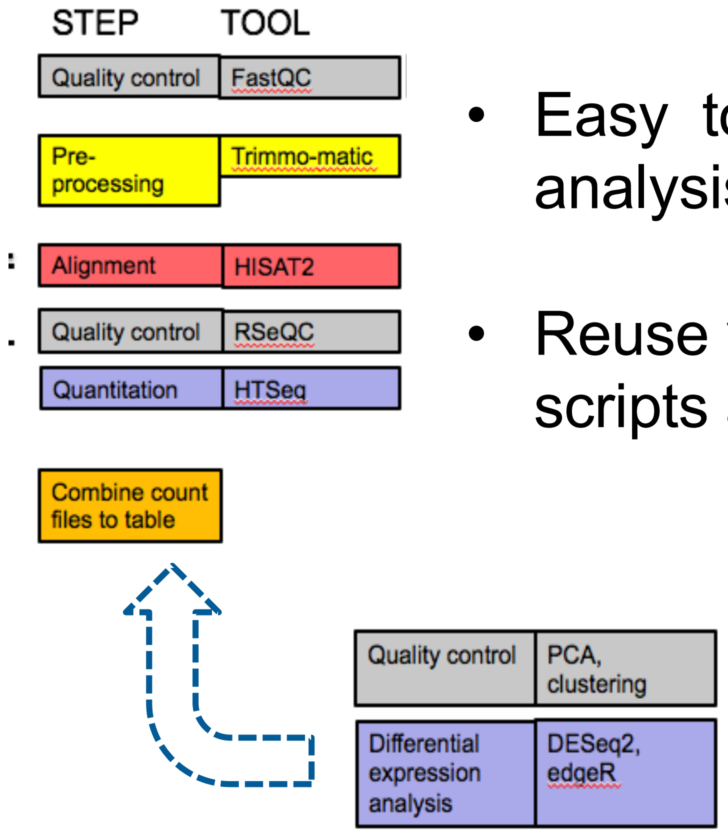
- Workflow Management
- Reproducibility
- Portability
- Scalability
- Parallelisation
- Easy resumption
- Fast prototyping

Call caching

Monitors each chunk/file and process

Why using Nextflow ?

- Workflow Management
- Reproducibility
- Portability
- Scalability
- Parallelisation
- Easy resumption
- Fast prototyping



STEP	TOOL
Quality control	FastQC
Pre-processing	Trimmo-matic
Alignment	HISAT2
Quality control	RSeQC
Quantitation	HTSeq
Combine count files to table	
Quality control	PCA, clustering
Differential expression analysis	DESeq2, edgeR

RNAseq pipeline for differential gene regulation

- Easy to add new analysis step
- Reuse your existing scripts and tools

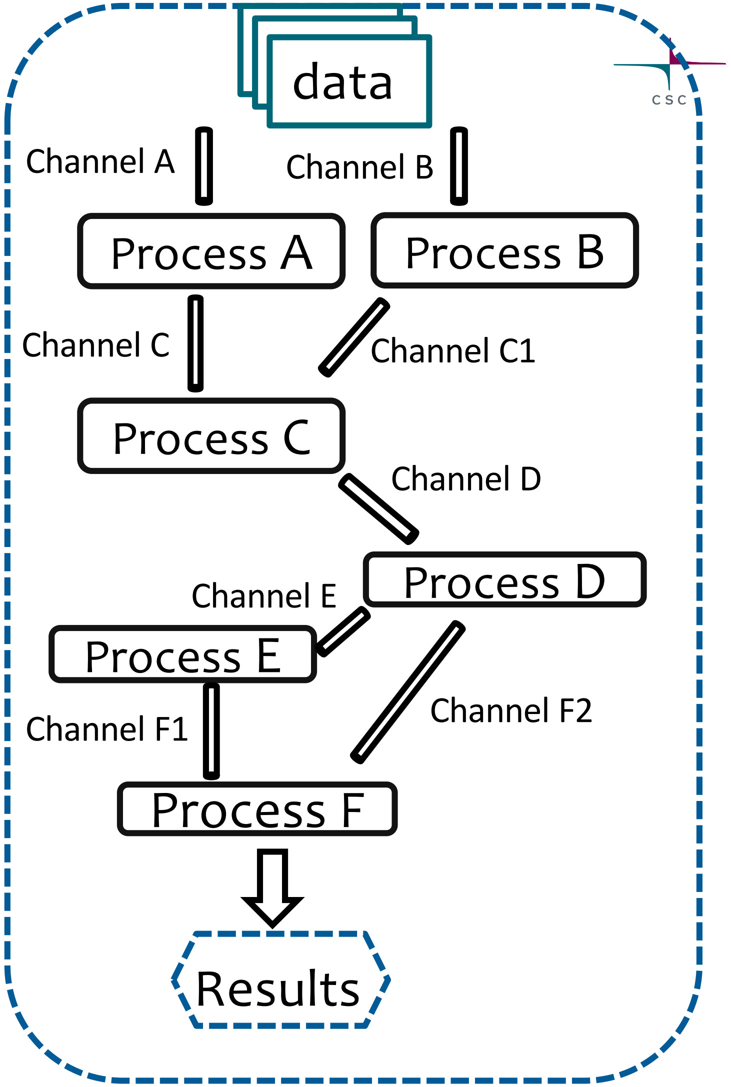
Getting Started with Nextflow

- Required:
  - Posix file system (Linux/OS …)
  - Java 8
- Software installation:
  - curl get.nextflow.io | bash
  - mv nextflow ~/bin
- Software stack you want:
  - Scripts available on PATH or under bin directory
  - Docker engine
  - Singularity
  - Conda

NextFlow: Building Blocks

- **channel**: information flows from one process to another via *Channels* as defined in the input and output sections of each process

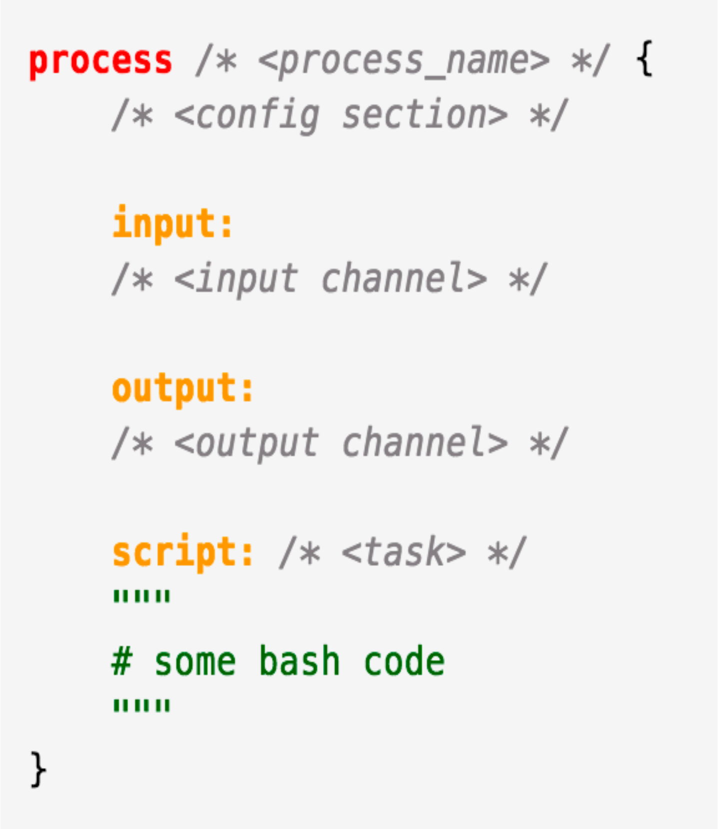
- **process**: one (independent) step in the pipeline block. This is where the execution of code happens



NextFlow: Building Blocks

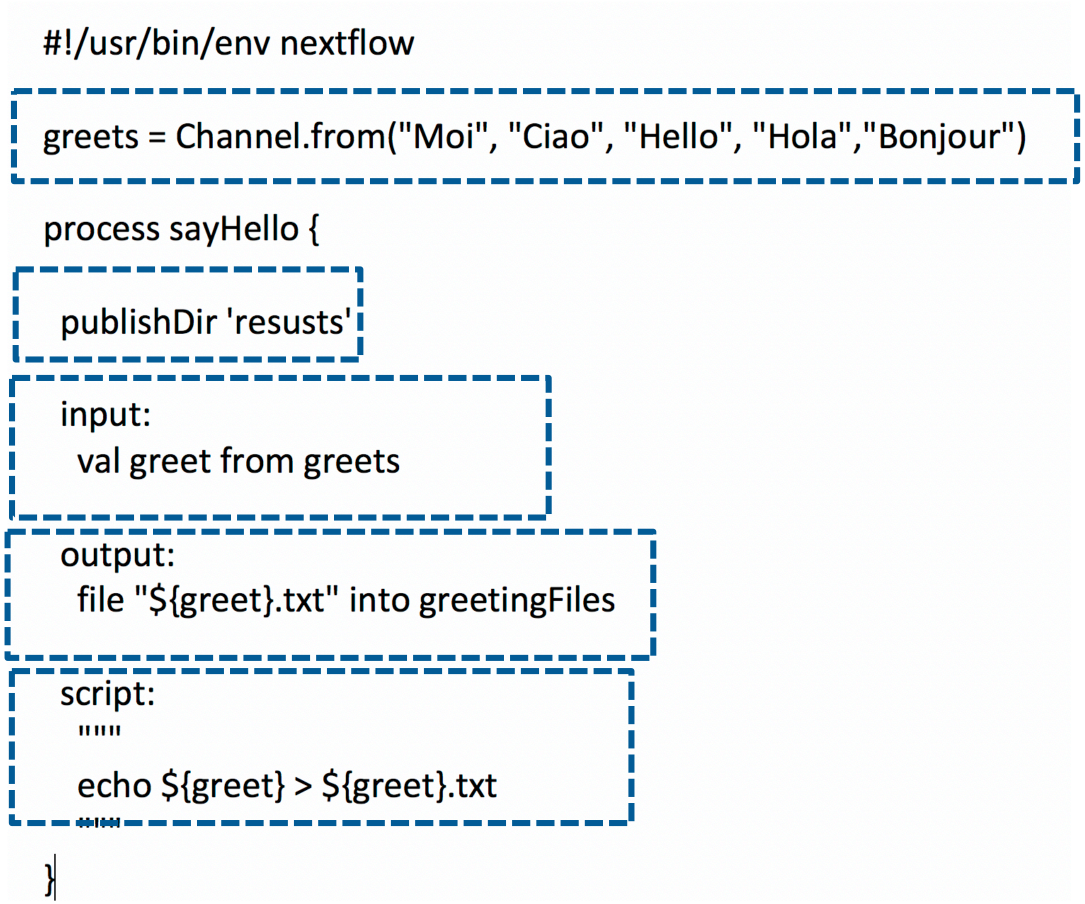
- **channel**: information flows from one process to another via *Channels* as defined in the input and output sections of each process

- **process**: one (independent) step in the pipeline block. This is where the execution of code happens



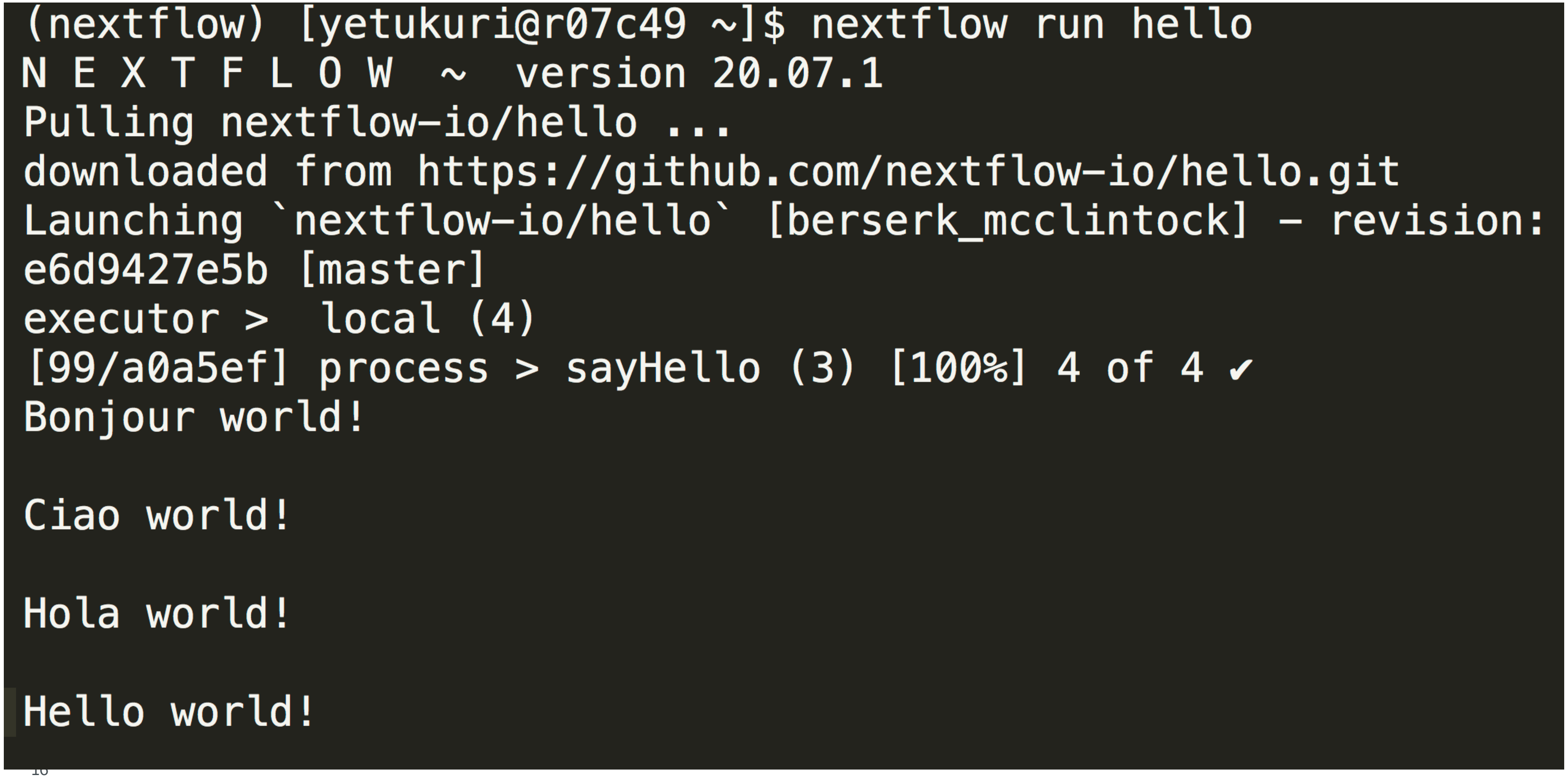
```
process /* <process_name> */ {
    /* <config section> */

    input:
    /* <input channel> */

    output:
    /* <output channel> */

    script: /* <task> */
    """
    # some bash code
    """
}
```

Nextflow : Hello World Example



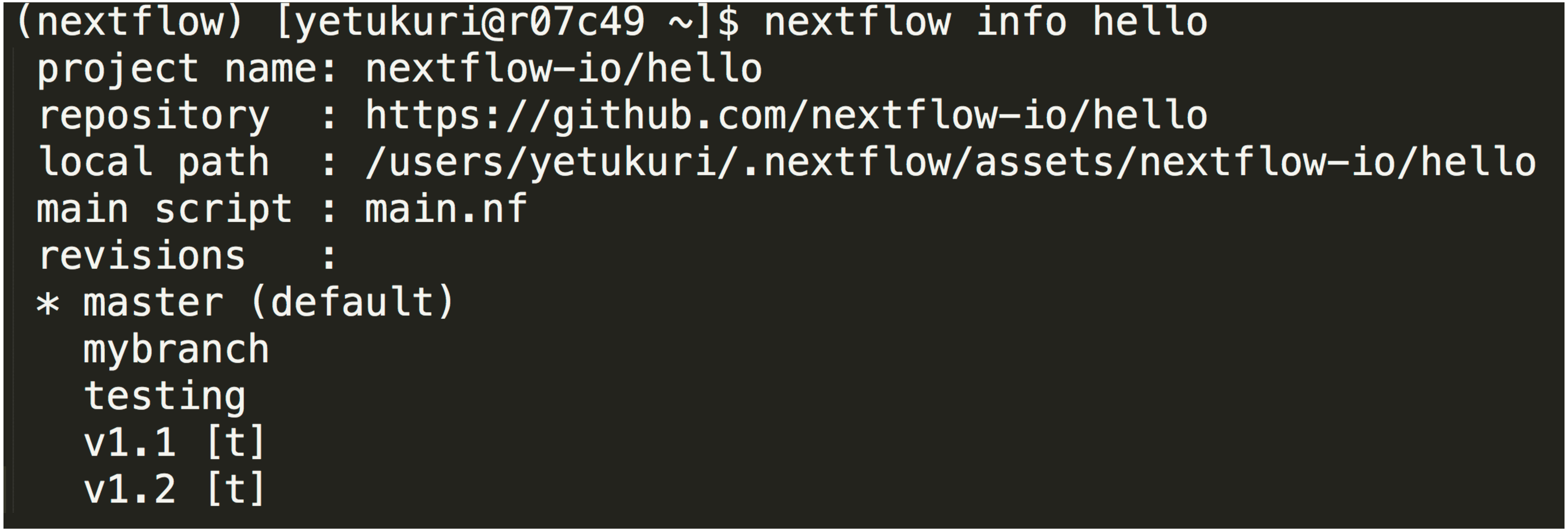
```
#!/usr/bin/env nextflow

greets = Channel.from("Moi", "Ciao", "Hello", "Hola","Bonjour")

process sayHello {

  publishDir 'resusts'

  input:
    val greet from greets

  output:
    file "${greet}.txt" into greetingFiles

  script:
    """
    echo ${greet} > ${greet}.txt
    """
}
```

Nextflow : Hello World *run* from github



```
(nextflow) [yetukuri@r07c49 ~]$ nextflow run hello
N E X T F L O W  ~  version 20.07.1
Pulling nextflow-io/hello ...
downloaded from https://github.com/nextflow-io/hello.git
Launching `nextflow-io/hello` [berserk_mcclintock] - revision:
e6d9427e5b [master]
executor >  local (4)
[99/a0a5ef] process > sayHello (3) [100%] 4 of 4 ✔
Bonjour world!

Ciao world!

Hola world!

Hello world!
```

Nextflow : Hello World info



```
(nextflow) [yetukuri@r07c49 ~]$ nextflow info hello
 project name: nextflow-io/hello
 repository   : https://github.com/nextflow-io/hello
 local path   : /users/yetukuri/.nextflow/assets/nextflow-io/hello
 main script : main.nf
 revisions    :
 * master (default)
   mybranch
   testing
   v1.1 [t]
   v1.2 [t]
```

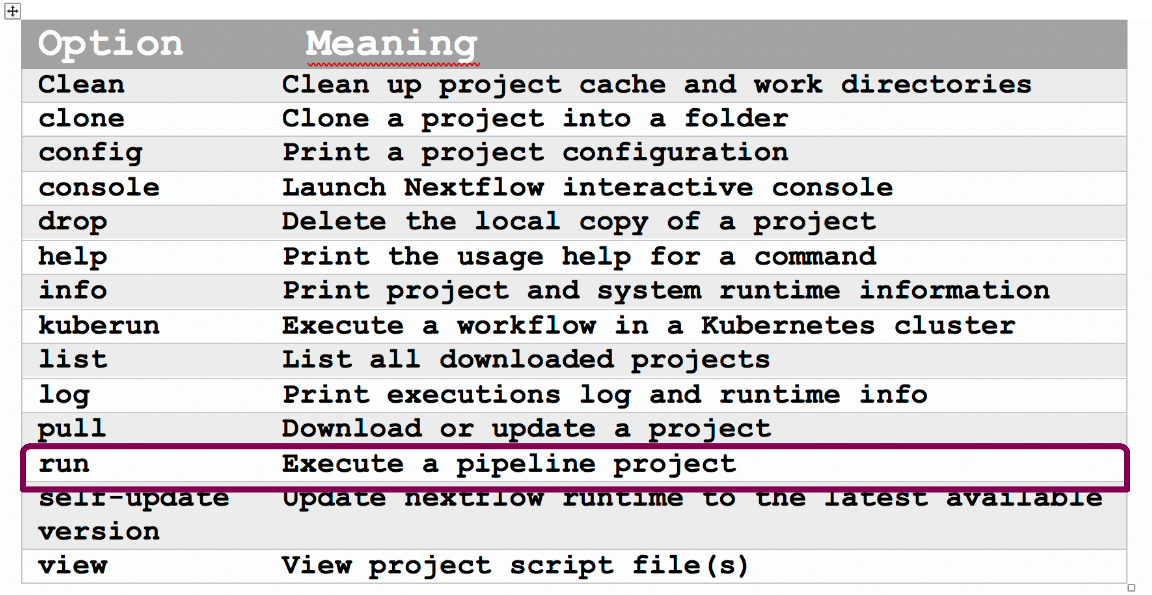
- Think of running above hello world exampl in a reproducible manner

Inspecting Nextflow Results

- Nextflow creates a folder (i.e., inside *work* directory) for each process
- Each folder contains
  - Links to input files
  - Output files
  - Number of hidden files
  - Script used for the process
- You can publish results to a different folder

NextFlow Help in Practice

- Help: nextflow -h
- Nextflow usage: nextflow [options] COMMAND [arg...]



Option	Meaning
Clean	Clean up project cache and work directories
clone	Clone a project into a folder
config	Print a project configuration
console	Launch Nextflow interactive console
drop	Delete the local copy of a project
help	Print the usage help for a command
info	Print project and system runtime information
kuberun	Execute a workflow in a Kubernetes cluster
list	List all downloaded projects
log	Print executions log and runtime info
pull	Download or update a project
run	Execute a pipeline project
self-update version	Update nextflow runtime to the latest available
view	View project script file(s)

Time for practicals !!!

- **Where to run practicals**: Interactive nodes on Puhti
- **Tutorials:** Hello-world and (close to) real-world tutorials
- **Expected outcome from tutorials**:
  - Learn to run a nextflow pipeline interactively (locally)
  - Able to Inspect default output files
  - Move resulting files to a convenient place
- **Set project number appropriately:** (i.e., project_xxxx -> project_ 2002389)