



# Introduction to Nextflow

Albert Palleja – Jakob Jespersen  
Group meeting 2024-01-25

# Outline

- Behind nextflow
- Why do we need a Community effort?
- Nextflow - Basic concepts
- Nextflow's core features
- A Nextflow script
- Installing Nextflow and running the first scripts
- Nextflow and Azure batch
- nf-metagenomics pipeline
- Resources



# Behind Nextflow



## Open Science

Make science more open!



## Open source

What this software is doing?  
(parameters, versions, functions reused/inherited)



## Open Data

Which data are we handling?  
How was measure, which technology, units, etc...  
FAIR principles



## Open community

A community that their members interact to each other and collaborate  
nf-core → pipelines  
Slack chat community



## Genomics workflows

First pipelines meant for that, large files per samples with many rows and columns, different formats (binary, tab separated, etc..) parallel processes, different software, languages (python, R, matlab, ...), takes decisions for the next processes

# Why do we need a Community effort?

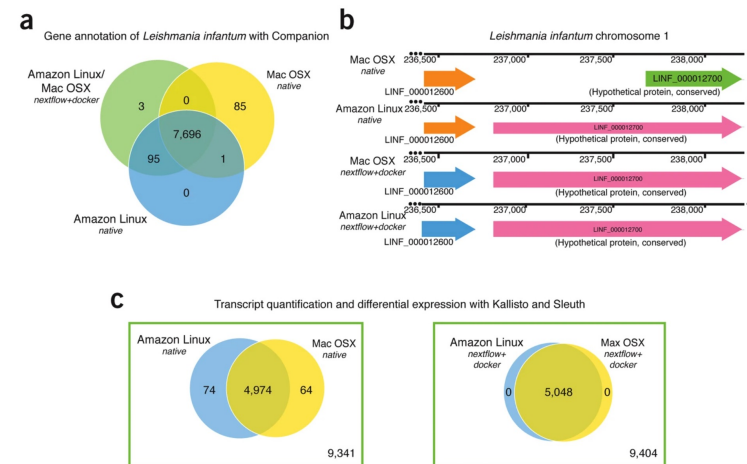
Reproducibility and even repeatability is challenging even if you are very careful keeping the same, params, versions, etc...!!!

## Nextflow enables reproducible computational workflows

[Paolo Di Tommaso](#), [Maria Chatzou](#), [Evan W Floden](#), [Pablo Prieto Barja](#), [Emilio Palumbo](#) & [Cedric Notredame](#) 

[Nature Biotechnology](#) **35**, 316–319 (2017) | [Cite this article](#)





Platform	Mac OSX	Amazon Linux	Debian Linux	Mac OSX	Amazon Linux
Execution	Native	Native	Native	NF+Docker	NF+Docker
number of chromosomes	36	36	36	36	36
overall length (bp)	32,032,223	32,032,223	32,032,223	32,032,223	32,032,223
number of genes	7,771	7,781	7,783	7,783	7,783
gene density	236.32	236.64	236.64	236.64	236.64
number of coding genes	7570	7,580	7,580	7,580	7,580
average coding length (bp)	1,762	1,764	1,764	1,764	1,764
number of genes with multiple CDS	111	113	113	113	113
number of genes with known function	4,142	4,147	4,147	4,147	4,147
number of t-RNAs	88	88	90	90	90



- Different results with different environments
- different gene annotations working in different environments. This does not happen working in containers

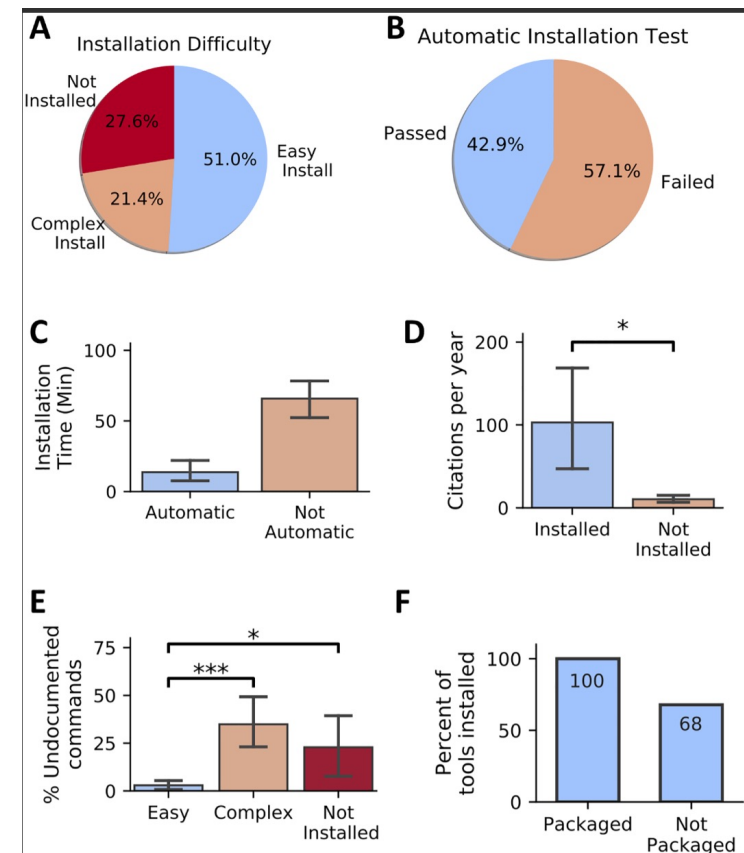
# Why do we need a Community effort?

## Challenges and recommendations to improve the installability and archival stability of omics computational tools

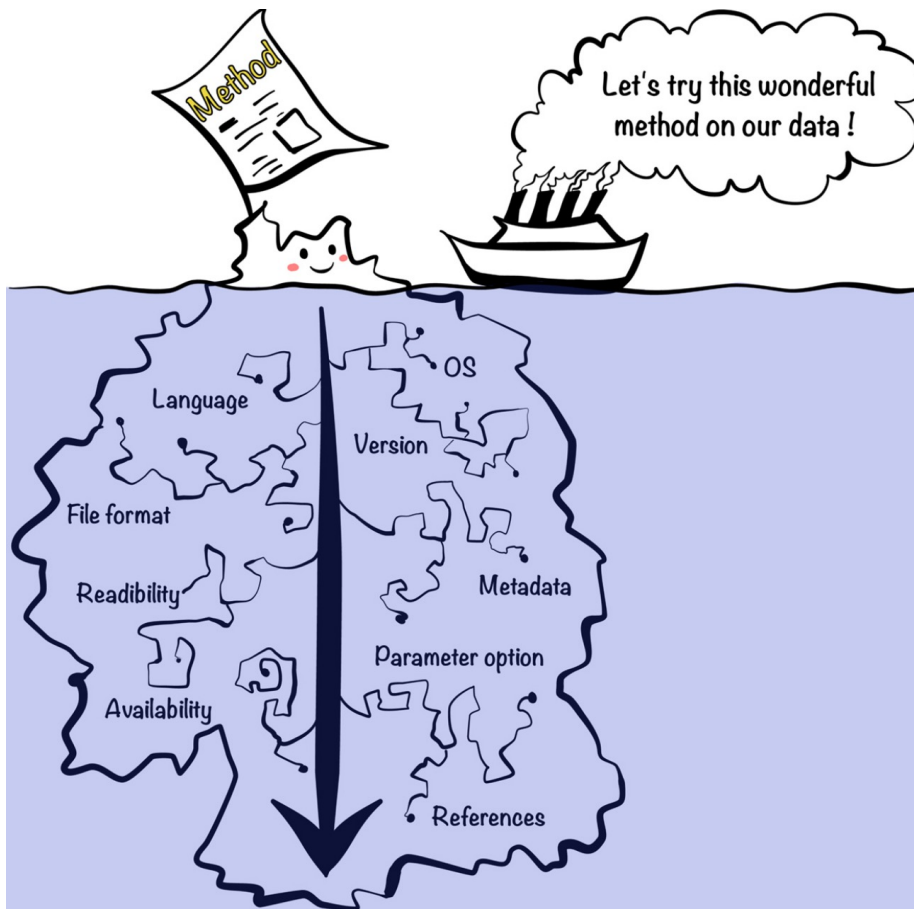
Serghei Mangul  , Thiago Mosqueiro , Richard J. Abdill, Dat Duong, Keith Mitchell, Varuni Sarwal, Brian Hill, Jaqueline Brito, Russell Jared Littman, Benjamin Statz , Angela Ka-Mei Lam, Gargi Dayama, Laura Grieneisen, [ ... ], Ran Blekhman [ view all ]

Version 2  Published: June 20, 2019 • <https://doi.org/10.1371/journal.pbio.3000333>

- 28% of all omics software resources are currently not accessible through URLs published
- Among the tools found, 49% were difficult to install or could not be installed at all!



# Why do we need a Community effort?



## Experimenting with reproducibility: a case study of robustness in bioinformatics

Yang-Min Kim ✉, Jean-Baptiste Poline, Guillaume Dumas

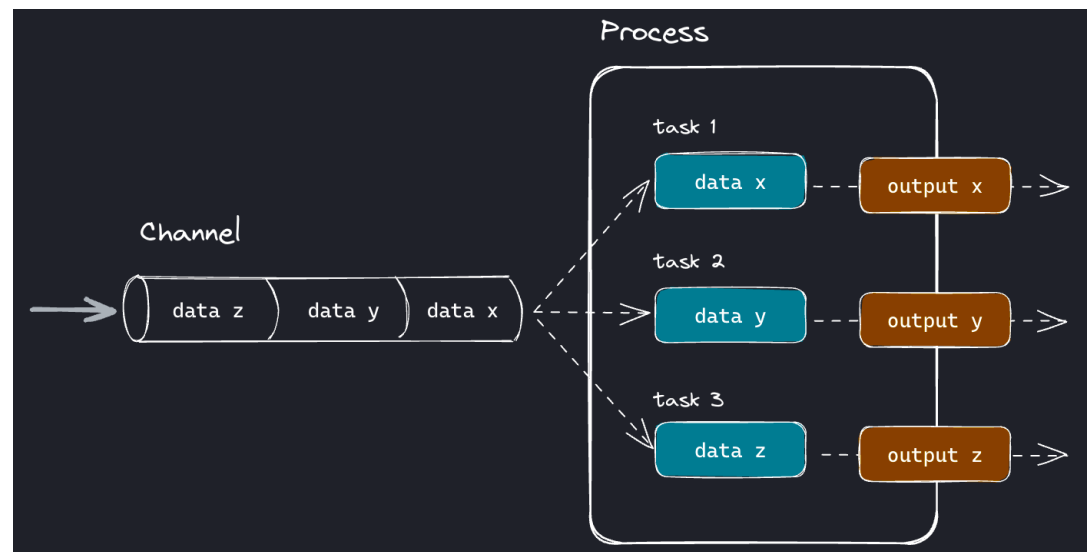
*GigaScience*, Volume 7, Issue 7, July 2018, giy077,  
<https://doi.org/10.1093/gigascience/giy077>

- "First we tried to rerun the analysis with the code and the data provided by the authors. Second we reimplemented the whole method in a python package.."

If you could use a tool to download, organize and orchestrate all these software / params / versions, etc... and achieve repeatable and reproducible results wouldn't you use it?

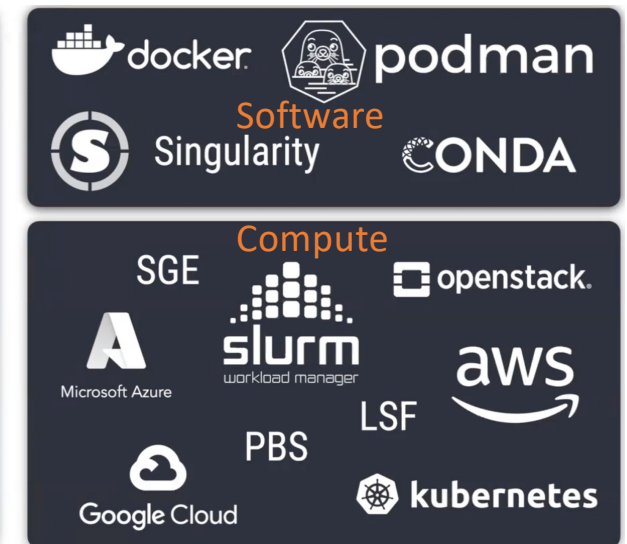
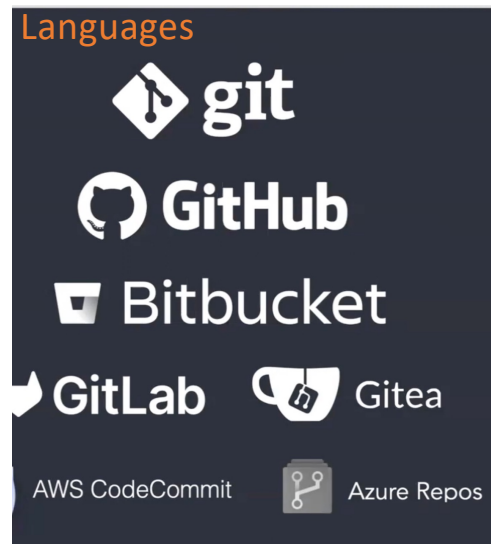
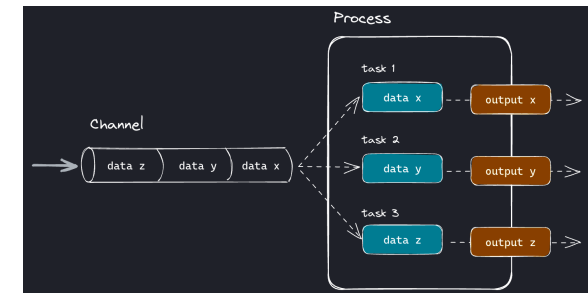
# Nextflow – Basic concepts

- Nextflow is a **software** - Workflow orchestrator engine
  - Nextflow is a Domain-specific **language** (DSL) built on top of Groovy
  - It allows writing data-intensive computational workflows
- 
- Nextflow **workflow** is a set of processes communicated by channels
  - Each **process** is a special function that can be written in any language that can be executed by the Linux platform (Bash, R, Python, etc...)
  - Processes are executed independently and are isolated from each other, and they can communicate by **channels**
  - Any process can define one or more channels as an input and output. The interaction between these processes, and ultimately the workflow execution flow itself, is implicitly defined by these input and output declarations.



# Nextflow – Basic concepts


- Implicit **parallelism** (tasks in a process run in parallel)
- **Reentrancy** (resume partial runs, do not need to rerun the entire pipeline, using -resume in the cmd line)
- **Reusability** (use different modules, subworkflows, written and containerized by the Nextflow community)
- Nextflow works with all main languages and version control providers
- Regarding software environments where to execute the code it works with most of the container solutions
- Same thing for computing environments







# Nextflow's core features

- 
- **Reproducible** between runs: all packages downloaded, organized in containers, control over computing environment,...
  - **Portable** between systems: write it in your laptop and can run everywhere
  - **Scalable** everywhere with implicit parallelism, can be run for 10 or thousands of samples
  - **Integration** of existing tools, systems, and industry standards

# A Nextflow script

```

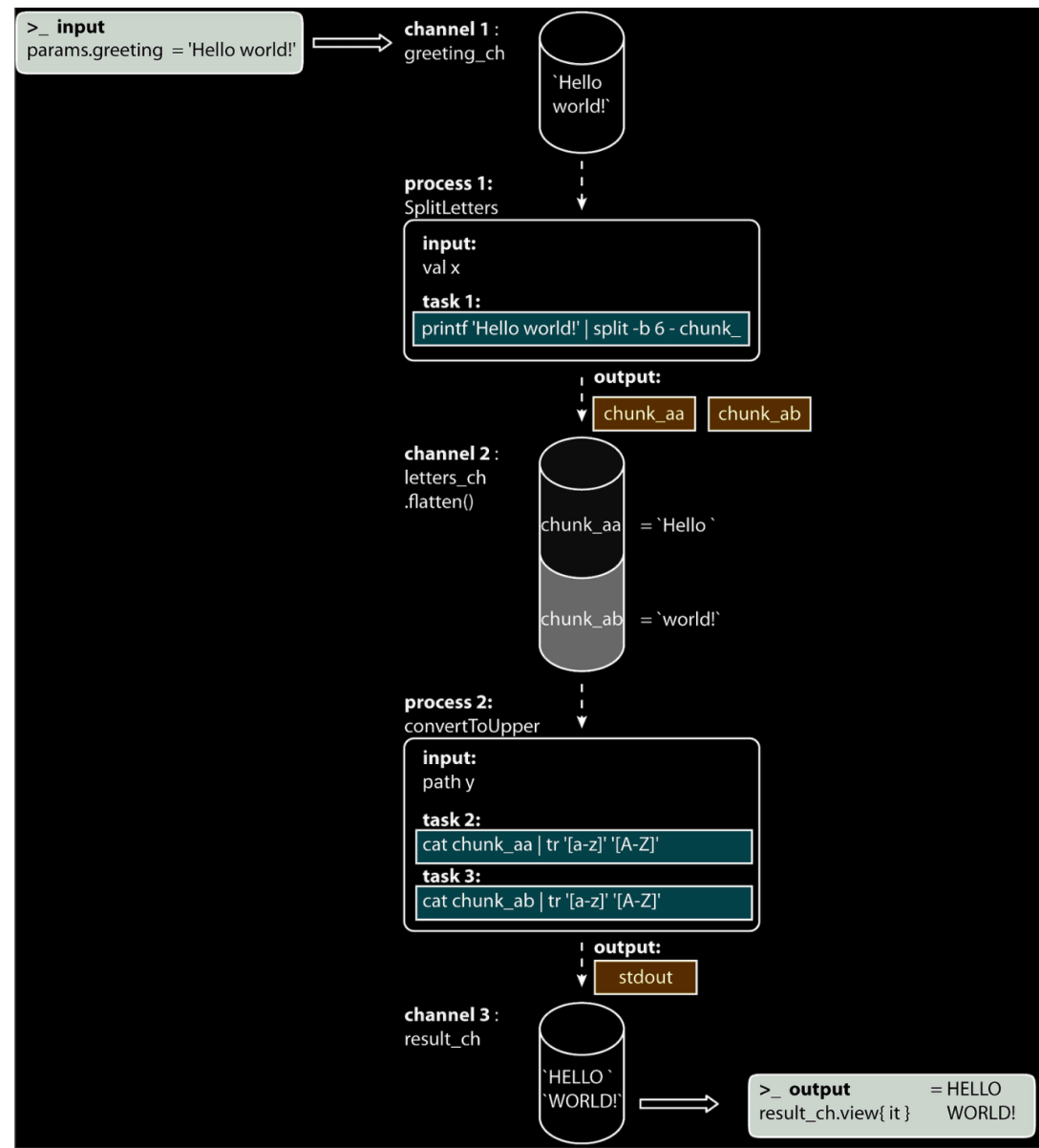
42  /*
43  * Quickly checking raw reads quality
44  */
45  process FASTQC {
46      container "quay.io/biocontainers/fastqc:0.12.1--hdfd78af_0"
47      tag "FASTQC on $sample_id"
48
49      input:
50      tuple val(sample_id), path(reads)
51
52      output:
53      path "fastqc_${sample_id}_logs"
54
55      script:
56      """
57      mkdir fastqc_${sample_id}_logs
58      fastqc -o fastqc_${sample_id}_logs -q ${reads}
59      """
60  }
61
62  workflow {
63      Channel
64      .fromFilePairs(params.reads, checkIfExists: true)
65      .set { read_pairs_ch }
66      fastqc_ch = FASTQC(read_pairs_ch)
67      fastqc_ch.view()
68  }

```

```
nextflow.config
1 | docker.enabled = true
```

```
○ (base) apca@NNFCB-L0989 dsp_nf-metagenomics % nextflow run main.nf -c nextflow.config
```

# Installing Nextflow and running the first scripts



# Resources



## **Nextflow foundational training:**

<https://nextflow.io>

<https://training.nextflow.io> course material



## **How to use it in Azure environment:**

<https://www.nextflow.io/blog/2021/introducing-nextflow-for-azure-batch.html>

<https://segera.io/blog/nextflow-and-azure-batch-part-1-of-2/>

<https://shaunchuah.github.io/posts/setting-up-azure-with-nextflow>