A novel tool for high scalable computational pipelines

Paolo Di Tommaso[1], Maria Chatzou[1,2], Pablo Prieto Baraja[1,2], Cedric Notredame[1]

[1] Comparative Bioinformatics, Bioinformatics and Genomics Program,
Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain
[2] Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain

What Nextflow is

Nextflow is a data-driven framework for computational pipelines that:

1) Simplifies writing parallel and scalable pipelines in a portable manner.
2) Allows you to use your favourite programming language and tools, exploiting your skills.

Why Nextflow

Parallelisation

Pipeline parallelisation is managed implicitly by Nextflow without adding unnecessary complexity. You can use it to parallelise your existing scripts.

Fault tolerance

The continue checkpoint mechanism automatically tracks all results. You can resume the pipeline execution at any time no matter what the reason was for it stopping.

Polyglot

It's easy to use whether you are a Python geek or a PERL hacker. Nextflow is not meant to replace your favourite tools but to integrate smoothly with them.

Scalability

Develop on your laptop, run in the grid or scale-out in the cloud. The support for Docker container technology allows you to write truly reproducible pipelines.

How it works

A Nextflow pipeline is made up by putting together several processes. Each process can be written in any scripting language that can be executed by the Linux platform (BASH, Perl, Ruby, Python, etc). Parallelisation is automatically managed by the framework and it is implicitly defined by the processes input and output declarations.

A key component of Nextflow is the *dataflow* programming model. Dataflow is a declarative processing model for parallel task executions in which concurrency and synchronisation is managed automatically and tasks communicate by using **asynchronous** queues called *channels*. This is a powerful primitive for distributed computing because it improves latency tolerance and third-party independence.

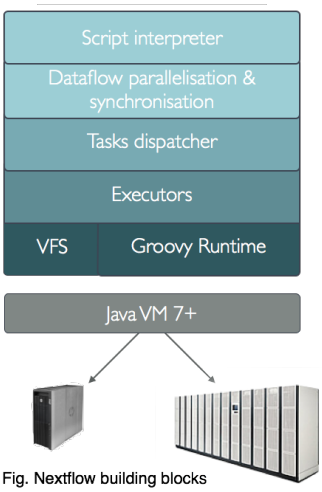
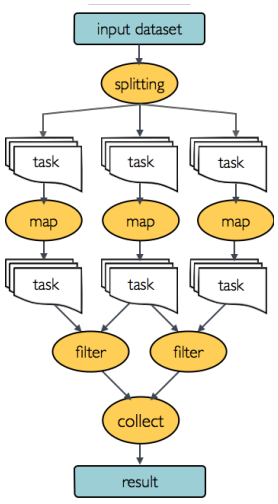
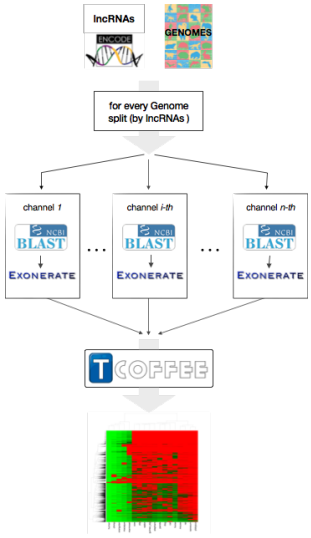


Fig. Nextflow building blocks



Moreover Nextflow provides an abstraction over the underlying executing platform. The resulting pipeline can run on a single workstation, on different grid infrastructures (SGE, LSF, SLURM) and in a cloud environment (DNAnexus).

It includes a rich set of functions for recurrent operations on common bioinformatics data formats (FASTA, FASTQ, etc) such as split, count, filter, combine, etc.



The Nextflow programming model greatly simplifies writing large scalable pipelines that use a scatter-process-gather parallelisation strategy that is quite common in bioinformatics applications.

We used this approach in PIPER, a pipeline for the detection and mapping of long non-coding RNAs. We managed to speed-up the overall pipeline execution by 6 times and to reduce the application code base in a significative manner.