Snakemake workflows deployed on virtual environments: a promising way of integrating high-throughput data in RegulonDB

Claire Rioualen[1], Lucie Khamvongsa[1], Alberto Santos-Zavaleta[2], Mishael Sánchez-Pérez[2], Jocelyn Brayet[3], Julio Collado-Vides[2], Jacques van Helden[1]

1. **Technologies avancées pour le génome et la clinique** (TAGC), INSERM U1090 , Aix-Marseille Université, Marseille F-13288, France
2. **Computational Genomics Research program**, Center for Genomics Sciences - UNAM, México
3. **Institut Curie**, PSL Research University, Mines Paris Tech, Inserm, U900, F-75005, Paris, France.

Next-generation sequencing (NGS) has become a mainstream technology in genomics, and it has gotten increasingly cheaper and faster to obtain genomic data. However, the development of reliable tools for the analysis of the huge amount of data generated is still lagging behind[1].

We have developed workflows using the Snakemake environment[2], for its scalability and ease of implementation. Then we proceeded with building a catalogue of tools for ChIP-seq analyses, that were implemented in custom workflows. This work was integrated in different virtual environments, and used to re-analyse published data[3] as well as manually curated data from RegulonDB [4].

Snakemake, a flexible environment for workflow development

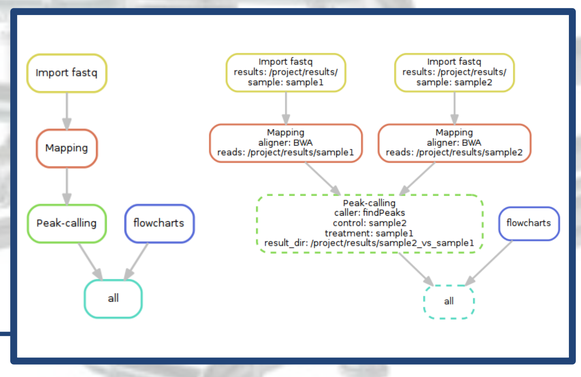
- **python** library for building workflows
- inherits concepts from the **GNU make** sofware:
  - **target** files or operations to be performed
  - **rules** describing how to produce these targets
- operations can be done in **python**, **R**, **shell** languages
- dependencies between rules are defined by their **inputs** and **outputs**
- **wildcards** can be used for automatization

Workflow example

```
## workflow.py
SAMPLES = ["sample1", "sample2"]
CHIP = "sample2"
CONTROL = "sample1"

rule all:
    input:
        expand("{sample}.bam", sample = SAMPLES)
        expand("{chip}_vs_{control}",
                    chip = CHIP, control = CONTROL)
rule mapping:
    input: "{file}.fastq"
    output: "{file}.bam"
    shell: "bowtie {input} > {output}"

rule peak_calling:
    input: chip = "{chip}.bam", control = "{chip}.bam"
    output: "{file}.bam"
    shell: "macs2 {input.chip} {input.control} > {output}"
```



We developed a public library of re-usable rules[5] which can be combined into different workflows: trimming, mapping, peak-calling, FastQC reports, motif search, IGV sessions... Ongoing is the development of rules to handle RNA-seq data and pipelines combining both types of data.

Virtual environments

It can be cumbersome to install all the tools that are required for an NGS analysis, especially when they come in different versions, with a number of dependencies, on different operating systems...

There's a crucial need for tools with a perfect **reproducibility**, and there's a need for the simplification of their use.

In order to facilitate the distribution of our library of rules and workflows, we have decided to rely on **virtualization**. We have created virtual solutions, and we have also written tutorials[4] that can help users setting up their own virtual environment and run small study cases.

Docker image[6] — IFB cloud appliance[7] — VirtualBox appliance[8]

These **virtual environments** are made of 4 principal components:
- Operating system (Ubuntu 14.04)
- Programming tools (snakemake, python, linux packages)
- NGS tools (mapping, peak-calling, file conversion, etc)
- Gene-regulation git repository (rules, workflows, configfiles)

The **git repository** contains a makefile that can install all the necessary programming and

ChIP-seq pipeline: E. coli study case compared with literature and RegulonDB data

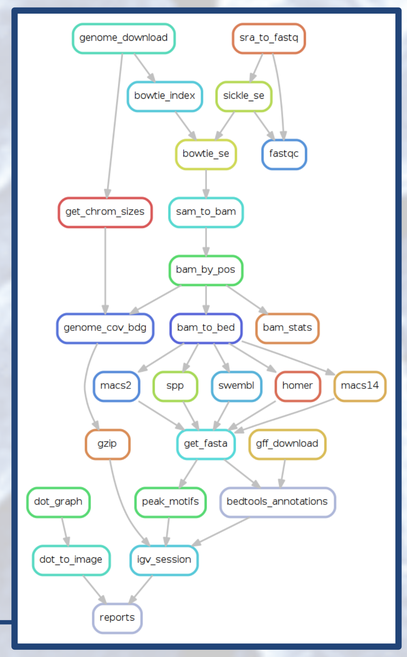
RegulonDB (http://regulondb.ccg.unam.mx/) is the primary database on transcriptional regulation in *Escherichia coli* K-12 containing knowledge manually curated from original scientific publications, experimental evidence, complemented with computational predictions.

This database could be further developed by integrating high-throughput data from experiments such as ChIP-seq or RNA-seq. Thus we decided to confront our pipeline to the data curated in RegulonDB, and with the results published by Myers et al.[3] which combined ChIP-seq and RNA-seq to detect direct target genes of **FNR** in aerobic and **anaerobic conditions**.

We ran the following workflow on samples GSM1010219 (FNR IP ChIP-seq Anaerobic A) and GSM1010224 (Anaerobic input DNA), from GEO subseries GSE41187.

Reads were trimmed using Sickle, and Bowtie was chosen for the alignment after Bowtie2, BWA and subread showed similar mapping rates with twice as much computing time.

Applying all the operations above to samples GSM1010219 and GSM1010224 took about **24mn** to complete on a virtual machine with 32 Go RAM.



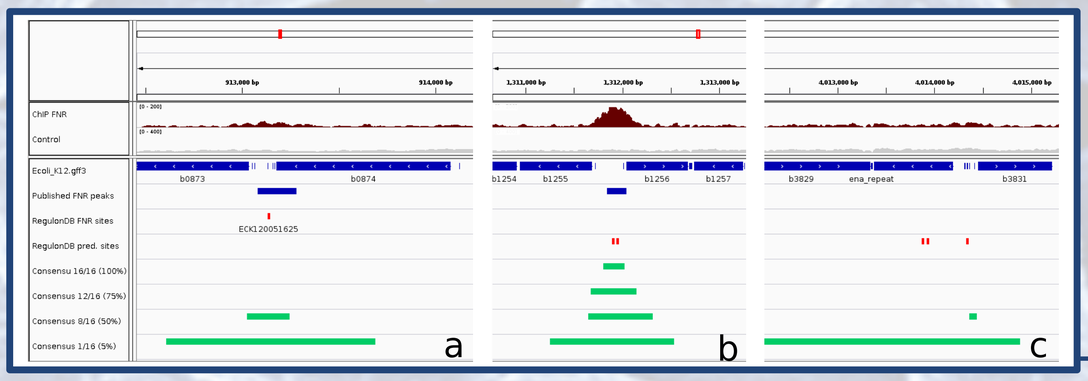
Flowchart generated by snakemake for the ChIP-seq workflow

Thanks to the Snakemake environment, we were able to run 5 different peak-callers, using a variety of parameters.

Myers et al.[3] ran 3 peak-callers: CisGenome, Mosaics and NCIS.

IGV session generated by the workflow

We kept 16 bed files whose peak count was between 100 and 500, and generated "consensus peak sets" using several thresholds. These peak sets were compared with published peaks[3] and with the motifs found in RegulonDB[4].



IGV

- 49 peaks correlated with known sites from RegulonDB, out of 79 sites manually checked (a)
- 29 peaks came as **new evidence** for RegulonDB predicted sites in both published data[3] and our pipeline (b)
- 16 predicted sites were supported by our **consensus peaks only** (c), meaning that many peak-callers couldn't back them up.
- More than 200 peaks were not associated with known or predicted sites in RegulonDB. These peaks will be further investigated by running a motif search algorithm available in our pipeline[9] and combining **RNA-seq** data to check for **expression changes**

Developing Snakemake workflows and virtual environments allows to distribute ready-to-use analysis systems with perfect **reproducibility**. Scientists can thus run analyses rigorously on any computer, at any time.

Our workflow library allows to achieve better **modularity**, **portability** either on personal computers or servers, and **flexibility** thanks to the configuration files. It also enables the **benchmarking** of tools (aligners, peak-callers).

We were able to show that combining several peak-calling approaches can bring a lot of information otherwise undetected.

We are now working on the **integration** of transcriptomic data such as **RNA-seq**. Hopefully this work will contribute to the integration of HT-generated data with literature gathered data, and enrich our overall knowledge of the transcriptional network of *E.coli*.

References

1. Bailey T, et al. (2013) Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. PLoS Comput Biol 9(11): e1003326. doi:10.1371/journal.pcbi.1003326
2. Johannes Köster and Sven Rahmann. Snakemake: a scalable bioinformatics workflow engine. Bioinformatics (2012) 28 (19): 2520-2522. doi: 10.1093/bioinformatics/bts480
3. Myers KS, et al. Genome-scale Analysis of Escherichia coli FNR Reveals Complex Features of Transcription Factor Binding. PLoS Genet (2013) 9(6): e1003565. doi:10.1371/journal.pgen.1003565 - Data: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41187
4. Gama-Castro S et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. J.Nucleic Acids Res. 2016 Jan 4;44(D1):D133-43. doi: 10.1093/nar/gkv1156. Epub 2015 Nov 2.
5. GitHub Gene-regulation: https://github.com/rioualen/gene-regulation
6. Docker image: https://hub.docker.com/r/rioualen/gene-regulation
7. Institut Français de Bioinformatique: http://france-bioinformatique.fr
8. VirtualBox sofware: https://www.virtualbox.org/
9. Medina-Rivera Aet al. (2015) RSAT 2015: Regulatory Sequence Analysis Tools . Nucleic Acids Res. 2015 (Web Server issue) in press.

Acknowledgments: This work is funded by France Génomique. Collaboration with the RegulonDB team was possible thanks to NIH grant GM0110597 & FOINS-CONACYT - Fronteras de la Ciencia 2015 - ID 15