Sequana: Motivations and Overview

Thomas Cokelaer and Dimitri Desvillechabrol

Institut Pasteur

March 22d 2016

NGS at Biomics

The Biomics Pole at Pasteur Institute is responsible for Next Generation Sequencing. Many aspects are covered including :

https://research.pasteur.fr/en/team/biomics/

- De novo and targeted sequencing of viruses, prokaryotes and eukaryotes
- Variant (SNP, indel, large rearrangements) detection
- Human and Mouse SNP detection by array
- Transcriptional analysis (RNA-Seq) for both prokaryotes and eukaryotes
- 16S and deep-sequencing metagenomic studies (mouse, human, and other environments)
- Bottom-up whole proteomic analysis and quantification
- Analysis of a wide range of post-translational modifications
- Determination of the dynamics of protein complexes.
- Epigenetics (methylation studies)
- Projects involving two or more techniques (i.e. proteogenomics, single-cell DNA/RNA analysis)

NGS at Biomics

The Biomics Pole at Pasteur Institute is responsible for Next Generation Sequencing. Many aspects are covered including :

https://research.pasteur.fr/en/team/biomics/

- De novo and targeted sequencing of viruses, prokaryotes and eukaryotes
- Variant (SNP, indel, large rearrangements) detection
- Human and Mouse SNP detection by array
- Transcriptional analysis (RNA-Seq) for both prokaryotes and eukaryotes
- 16S and deep-sequencing metagenomic studies (mouse, human, and other environments)
- Bottom-up whole proteomic analysis and quantification
- Analysis of a wide range of post-translational modifications
- Determination of the dynamics of protein complexes.
- Epigenetics (methylation studies)
- Projects involving two or more techniques (i.e. proteogenomics, single-cell DNA/RNA analysis)

We are developing NGS pipelines like many others and have started to gather tools and information in a common repository called **Sequana**.

Needs

What do we have . . . or not ?

- ☑ A bunch of pipelines dedicated to NGS data
- ☑ Expertise
- ☒ Lack of
  - traceability ?
  - reproducibility ?
  - co-development ?
  - common framework ?

What do we need ?

- A framework to combine or re-use existing pipelines
- Fast development (iterative process)
- Continuous Integration and Quality Software (reproducibility, traceability, test, documentation)

Why Sequana ?

Enforce a common framework

- Using Snakemake as a common language to design new pipelines
- Provide reusable snakemake rules and modules

A toolbox in sequana to parse and analyse various data sets

- Include pandas for data mining
- matplotlib for further visualisation

A set of reports to improve

- Software Quality
- Diffusion
- reproducibility

Pipelines included

Snakefile

Snakefile are stored in directories called pipelines and accessible by name in Python

```
>>> from sequana import snakemake
>>> snakemake.rules.keys()
['dag', 'biomics', 'variant']
>>> snakemake.rules['variants']
'/home/cokelaer/Work/github/sequana/pipelines/variants/Snakefile'
```

It is therefore easy to include them in your own Snakefile:

```
import sequana.snakemake as sm
include: sm.rules['dag']
include: sm.rules['variants']
```

Report

We will provide a system of HTML reporting using sequana and JINJA templating

Snakefile

```
rule report:
input:
    dag = "dag.svg"
output: "report/index.html"
run:
    from sequana import report_main
    s = report_main.SequanaReport()
    s.create_report()
    shell("cp Snakefile report/")
    shell("cp dag.svg report/")
```

Toolbox

In addition to pipelines and reports, multi-purpose codes can be included within Sequana. We currently have some tools to handle BAM, FastQ but tend to rely on existing packages such as pysam and pyVCF. Here is a simple function that retrieves the flags of a BAM file into a Pandas DataFrame

Snakefile

```
>>> # BAM is a class that inherits from pysam.Alignment and
>>> # add a couple of functions
>>> from sequana import BAM
>>> b = BAM("filename.bam")
>>> df = b.get_flags_as_df()
>>> df.sum()
1        1526795
2           2703
4        1523785
8        1523785
16          1513
32          1513
64        763395
128       763400
256           13
512            0
1024           0
2048           0
dtype: int64
```

High code quality

Continuous Integration on Travis with currently 50% coverage

README.rst

SEQUANA

pypi package 0.0.3 | build passing | coverage unknown | docs latest

Python version:	Python 2.7, 3.4 and 3.5
Online documentation:	On readthedocs
Issues and bug reports:	On github

Sequana includes a set of pipelines related to NGS (new generation sequencing).

It will provide a set of modular pipelines and reports associated to them.

Open the file report/index.html for example of the current report (v0.0.3)

How to contribute ?

1. transform existing pipelines into Snakefiles and add them to Sequana.
2. Identify parts that can be transformed into modules
3. add tests
4. Benchmarks ?
5. complete documentation
6. Add data processing or visualisation tools
7. Other ideas welcome

Summary

- Sequana is used at biomics to share NGS pipelines (currently variants and fix contamination)
- Ease design of new pipelines.
- Single entry point
- share expertise and existing pipelines
- automatic reports
- reaching a high quality and trustable code

Links

- Join the github : https://github.com/sequana/sequana
- Doc on line on sequana.readthedocs.org