Institut Pasteur

Sequana: a set of flexible genomic pipelines for processing and reporting NGS analysis

Dimitri Desvillechabrol

Institut Pasteur

Dec 12th 2016, Journée Snakemake

Motivation

NGS at Biomics (Sean Kennedy)

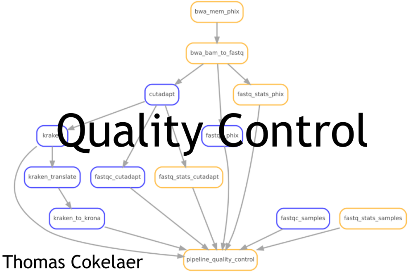
Development driven by the Biomics Pole at Pasteur Institute, which involves many aspects of NGS including :

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 (CHIP-Seq, methylation studies)
- Projects involving two or more techniques (i.e. proteogenomics, single-cell DNA/RNA analysis)

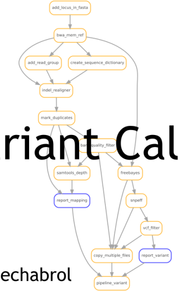
Pipelines

Pipelines available in Sequana



Quality Control

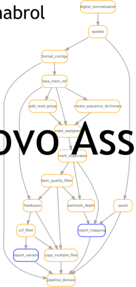
Thomas Cokelaer



Variant Calling

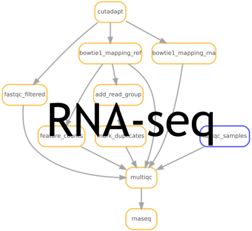
Dimitri Desvillechabrol

Dimitri Desvillechabrol



De Novo Assembly

Rachel Legendre



RNA-seq

Rules specificity: input and output are variables

- Rules are generic and easily reusable

mark_duplicates.rules

```
rule mark_duplicates:
    input:
        __mark_duplicates__input
    output:
        bam = __mark_duplicates__output,
        metrics = __mark_duplicates__metrics
    log:
        out = __mark_duplicates__log_std,
        err = __mark_duplicates__log_err
```

Dynamics rules

- Each rule must be unique in a pipeline

Some pipelines must use multiple times one rules like fastqc in quality control pipeline

- These rules are templatized to become dynamic:

quality_control.rules

```
exec(open(sequana.modules["fastqc"], "r").read())
        ...
include: fastqc_dynamic("samples", manager)
        ...
include: fastqc_dynamic("phix", manager)
        ...
include: fastqc_dynamic(adapter_removal, manager)
```

Usage

Using command line

- One command line to initiate the pipeline

Shell

```
sequana --pipeline variant_calling \
        --input-directory path/to/sample/ \
        --reference sequence.fasta \
        --output-directory analysis/
```

- The sequana executable creates a directory with the project name
- The directory contains all the necessary files (config, snakefile)

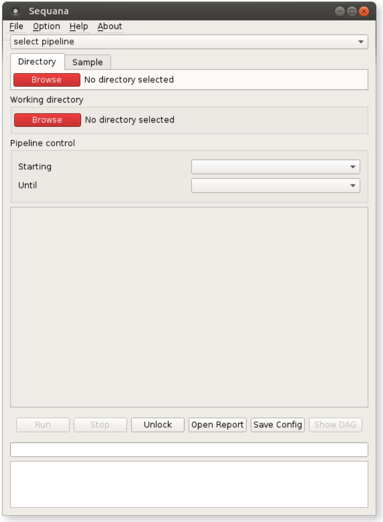
Continuous Integration

Versioning, Test and Documentation

- Sequana is available on GitHub (github.com/sequana/sequana)
- Continuous Integration on Travis with 60 tests with 60% coverage
- Documentation available on sequana.readthedocs.org .
  - Uses Sphinx (RST syntax) to document the source code and provides user guide.
  - Updated automatically at each commits

GUI

GUI to simplify the usage of snakemake

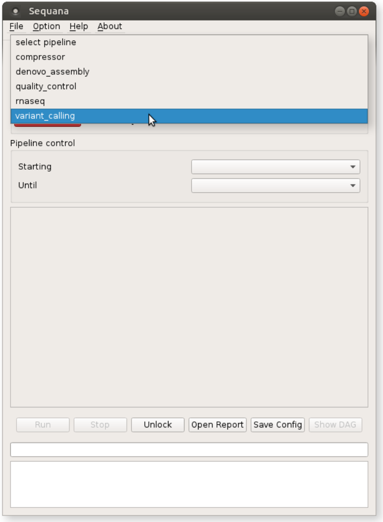


Directory: No directory selected

Working directory: No directory selected

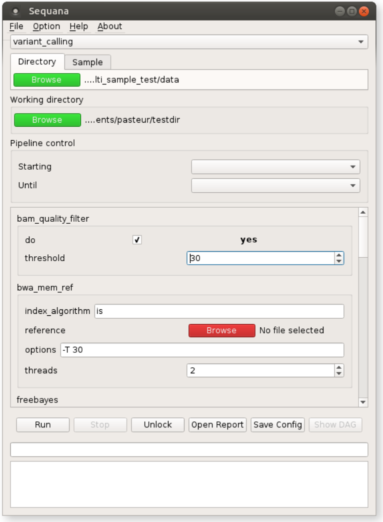
- Interface developed with PyQT5 and python
- Wrap our snakemake pipelines to ease the usage
- Usable on our cluster, which allows X11

GUI to simplify the usage of snakemake



1. Choose a pipeline

GUI to simplify the usage of snakemake



variant_calling

....lti_sample_test/data

Working directory

....ents/pasteur/testdir

Pipeline control

Starting

Until

bam_quality_filter

do ☑ yes

threshold 30

bwa_mem_ref

index_algorithm is

reference No file selected

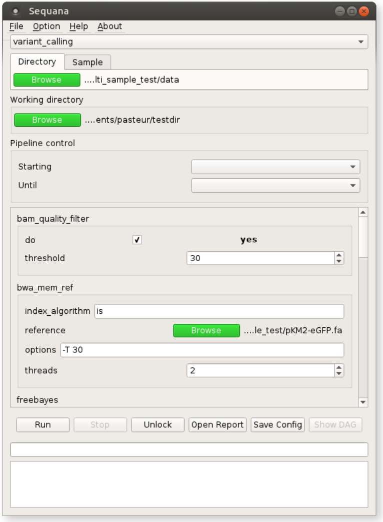
options -T 30

threads 2

freebayes

1. Choose a pipeline
2. Set input and output

GUI to simplify the usage of snakemake



variant_calling

Directory | Sample

....lti_sample_test/data

Working directory

....ents/pasteur/testdir

Pipeline control

Starting

Until

bam_quality_filter

do ☑ yes

threshold 30

bwa_mem_ref

index_algorithm is

reference ....le_test/pKM2-eGFP.fa

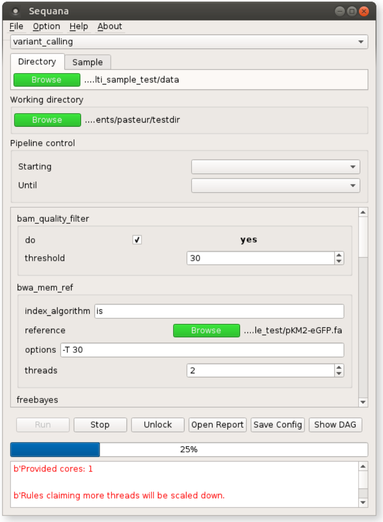
options -T 30

threads 2

freebayes

1. Choose a pipeline
2. Set input and output
3. Fill the config formular

GUI to simplify the usage of snakemake



variant_calling

....lti_sample_test/data

Working directory

....ents/pasteur/testdir

Pipeline control

Starting

Until

bam_quality_filter

do ☑ yes

threshold 30

bwa_mem_ref

index_algorithm is

reference ....le_test/pKM2-eGFP.fa

options -T 30

threads 2

freebayes

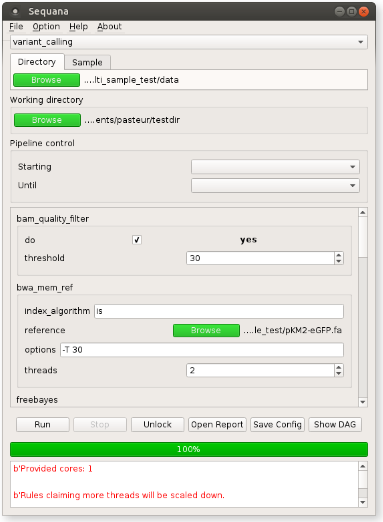
25%

```
b'Provided cores: 1

b'Rules claiming more threads will be scaled down.
```

1. Choose a pipeline
2. Set input and output
3. Fill the config formular
4. Run the pipeline

GUI to simplify the usage of snakemake



variant_calling

Browse ....lti_sample_test/data

Working directory: Browse ....ents/pasteur/testdir

Starting

Until

bam_quality_filter

do ☑ yes

threshold 30

bwa_mem_ref

index_algorithm is

reference Browse ....le_test/pKM2-eGFP.fa

options -T 30

threads 2

freebayes

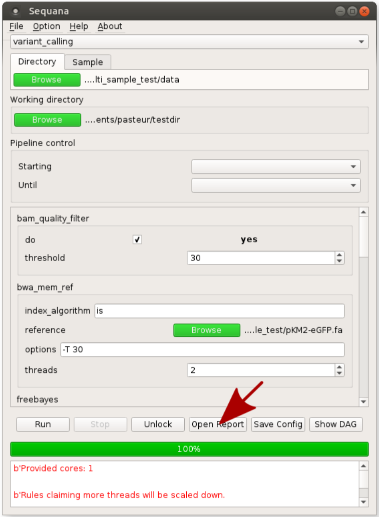
100%

```
b'Provided cores: 1

b'Rules claiming more threads will be scaled down.
```

1. Choose a pipeline
2. Set input and output
3. Fill the config formular
4. Run the pipeline
5. Finished !

GUI to simplify the usage of snakemake



variant_calling

....lti_sample_test/data

Working directory

....ents/pasteur/testdir

Pipeline control

Starting

Until

bam_quality_filter

do ☑ yes

threshold 30

bwa_mem_ref

index_algorithm is

reference ....le_test/pKM2-eGFP.fa

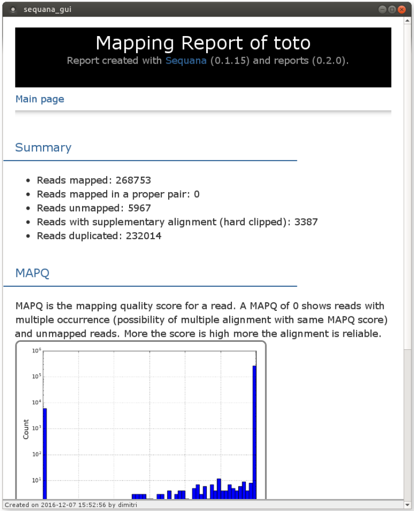
options -T 30

threads 2

freebayes

```
b'Provided cores: 1

b'Rules claiming more threads will be scaled down.
```



Mapping Report of toto

Report created with Sequana (0.1.15) and reports (0.2.0).

Summary

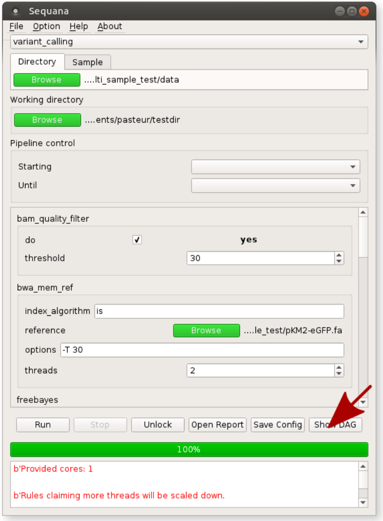
- Reads mapped: 268753
- Reads mapped in a proper pair: 0
- Reads unmapped: 5967
- Reads with supplementary alignment (hard clipped): 3387
- Reads duplicated: 232014

MAPQ

MAPQ is the mapping quality score for a read. A MAPQ of 0 shows reads with multiple occurrence (possibility of multiple alignment with same MAPQ score) and unmapped reads. More the score is high more the alignment is reliable.

Created on 2016-12-07 15:52:56 by dimitri

Ease the pipeline manipulation and viewing



variant_calling

Directory: ....lti_sample_test/data

Working directory: ....ents/pasteur/testdir

Pipeline control

Starting

Until

bam_quality_filter

do ☑ yes

threshold 30

bwa_mem_ref

index_algorithm is

reference ....le_test/pKM2-eGFP.fa

options -T 30

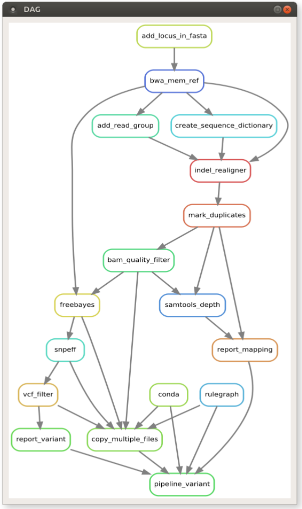
threads 2

freebayes

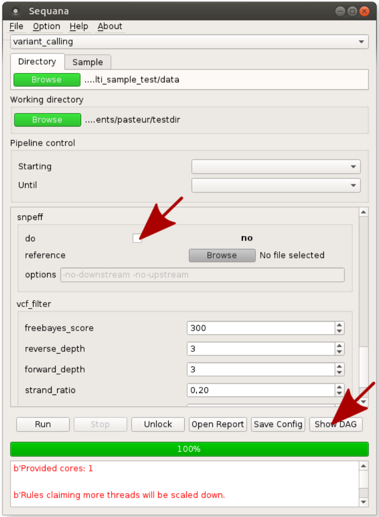
100%

```
b'Provided cores: 1

b'Rules claiming more threads will be scaled down.
```



Ease the pipeline manipulation and viewing



Sequana

variant_calling

Directory: ....lti_sample_test/data

Working directory: ....ents/pasteur/testdir

Pipeline control

Starting

Until

snpeff

do □ no

reference: No file selected

options: -no-downstream -no-upstream

vcf_filter

freebayes_score: 300

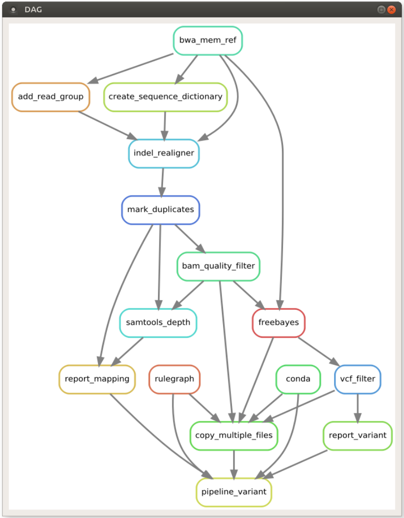
reverse_depth: 3

forward_depth: 3

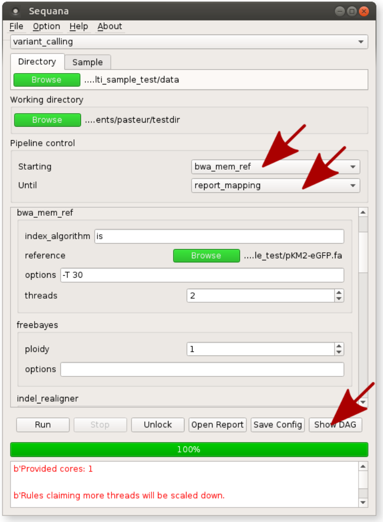
strand_ratio: 0,20

```
b'Provided cores: 1

b'Rules claiming more threads will be scaled down.
```



Ease the pipeline manipulation and viewing



variant_calling

Directory: ....lti_sample_test/data

Working directory: ....ents/pasteur/testdir

Pipeline control

Starting: bwa_mem_ref

Until: report_mapping

bwa_mem_ref

index_algorithm: is

reference: ....le_test/pKM2-eGFP.fa

options: -T 30

threads: 2

freebayes

ploidy: 1

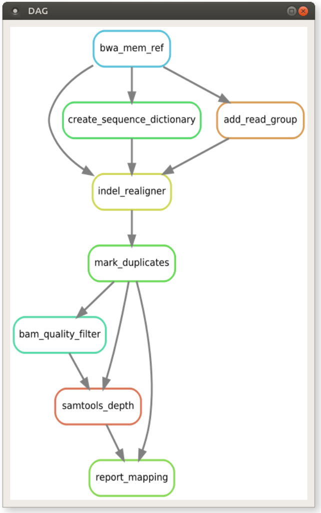
options:

indel_realigner

100%

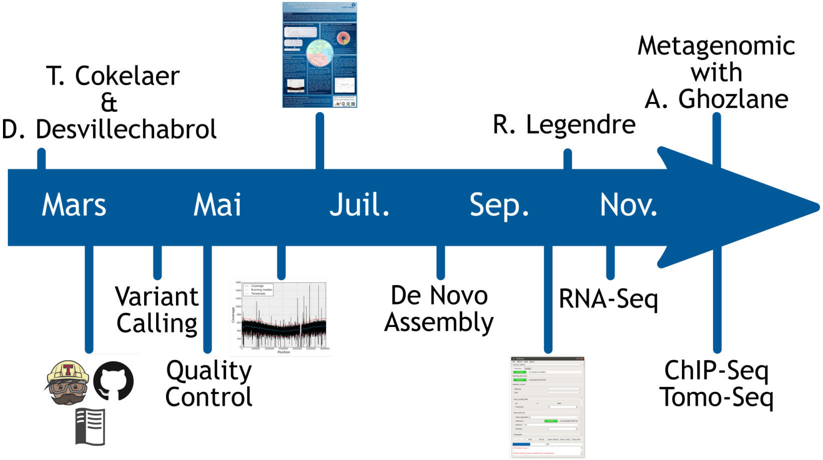
```
b'Provided cores: 1

b'Rules claiming more threads will be scaled down.
```



Summary and Future Directions

Sequana history



[1] Detection and characterization of low and high genome coverage regions using an efficient running median and a double threshold approach. Dimitri Desvillechabrol, Christiane Bouchier, Sean Kennedy, Thomas Cokelaer bioRxiv 092478; doi: http://dx.doi.org/10.1101/092478

Acknowledgement

- Developers:
  - Thomas Cokelaer
  - Rachel Legendre
- Beta-tester:
  - Christiane Bouchier
- Fruitful discussions:
  - Claudia Chica
  - Varun Khanna
  - Pierre Lechat
  - Frédéric Lemoine
  - Hervé Ménager
  - Bioinformatics and Biostatistics HUB
- Biomics team:
  - Jean-Yves Copee
  - Amine Ghozlane
  - Sean Kennedy
  - Béatrice Regnault
  - Hugo Varet
- Organizations:

Citech

C3BI

FRANCE GÉNOMIQUE