Snakemake projects at the Institut Curie

Journée Snakemake at Institut Pasteur
12/12/2016

Mathieu Valade
Jocelyn Brayet
Bioinformatics platform, Institut Curie, INSERM U900

NGS pipelines currently in production

→Data sets :

- From Institut Curie Research Center and Hospital
- Mainly Illumina Technology

→ Pipelines :

- 7 research pipelines (exome seq, target seq, RNA seq, ...)
- 8 diagnostic pipelines (target seq, RNA seq)

→ Informatics :

- Bash scripts, config files
- Scheduler : Torque

NGS workflow improvement

→ Pipelines evolution :

- Output data quality
- Reproductibility
- Automatic pipelines run

→ Clinical context :

- Increase of NGS data
- Rapid evolution of the high-throughput technologies
- Personalized medicine

Snakemake vs. bash

→ Testing the continuity of the pipeline before launching
- rules
- inputs/outputs
- rule settings

→ Stop process if error and delete incomplete files

→ Error recovery

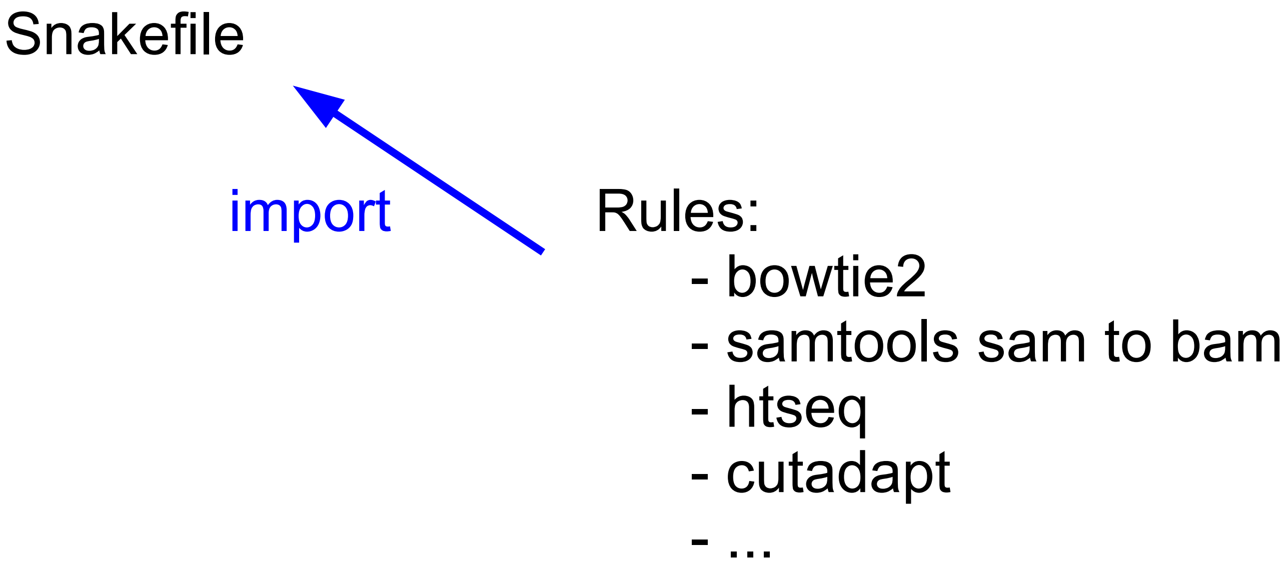
→ Automatic parallelization

→ Cluster (Torque, ...) management by rule

Snakemake pipeline organization

Snakefile

Snakemake pipeline organization



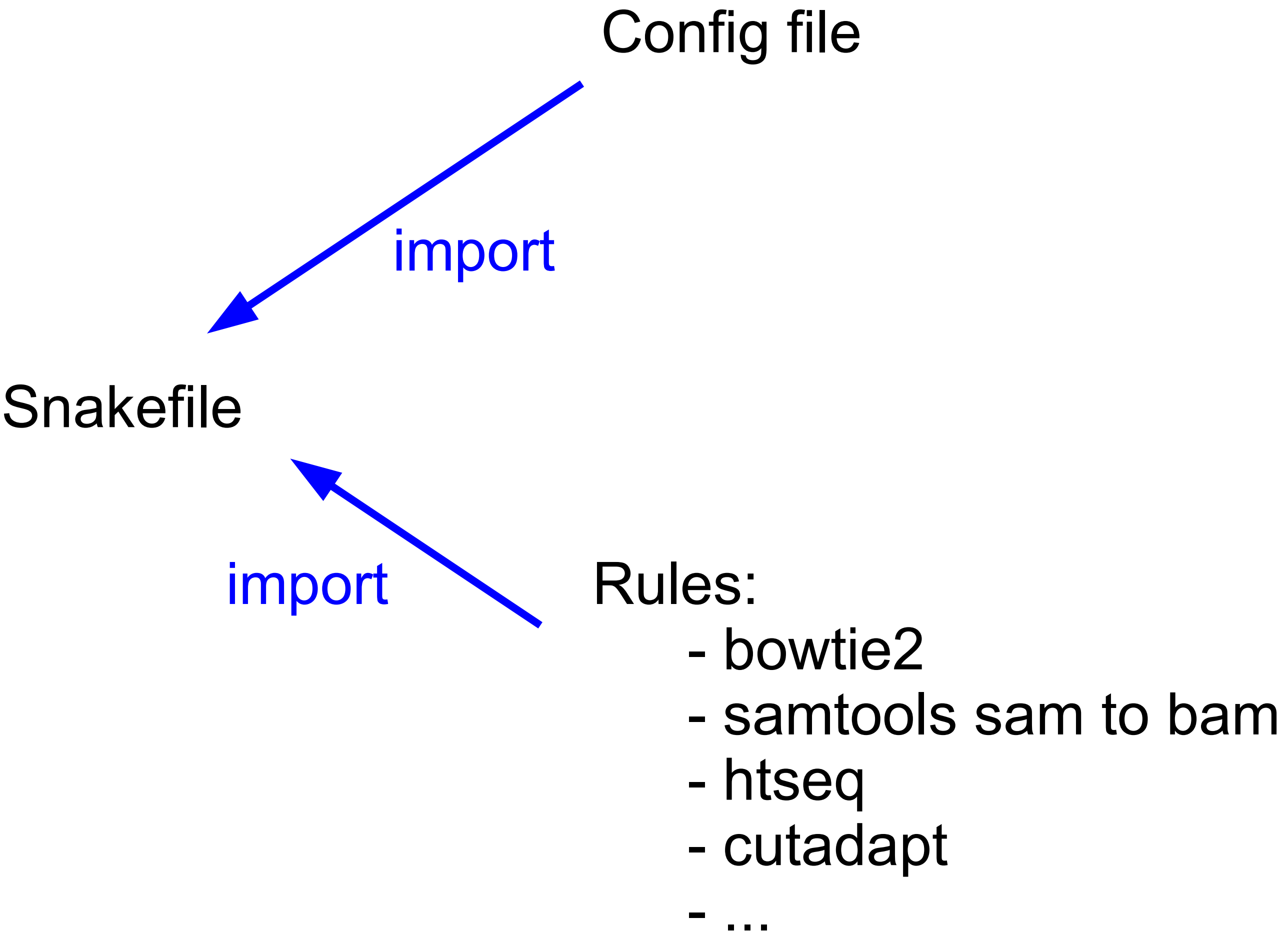
Snakefile

import

Rules:

- bowtie2
- samtools sam to bam
- htseq
- cutadapt
- ...

Snakemake pipeline organization



Config file

import

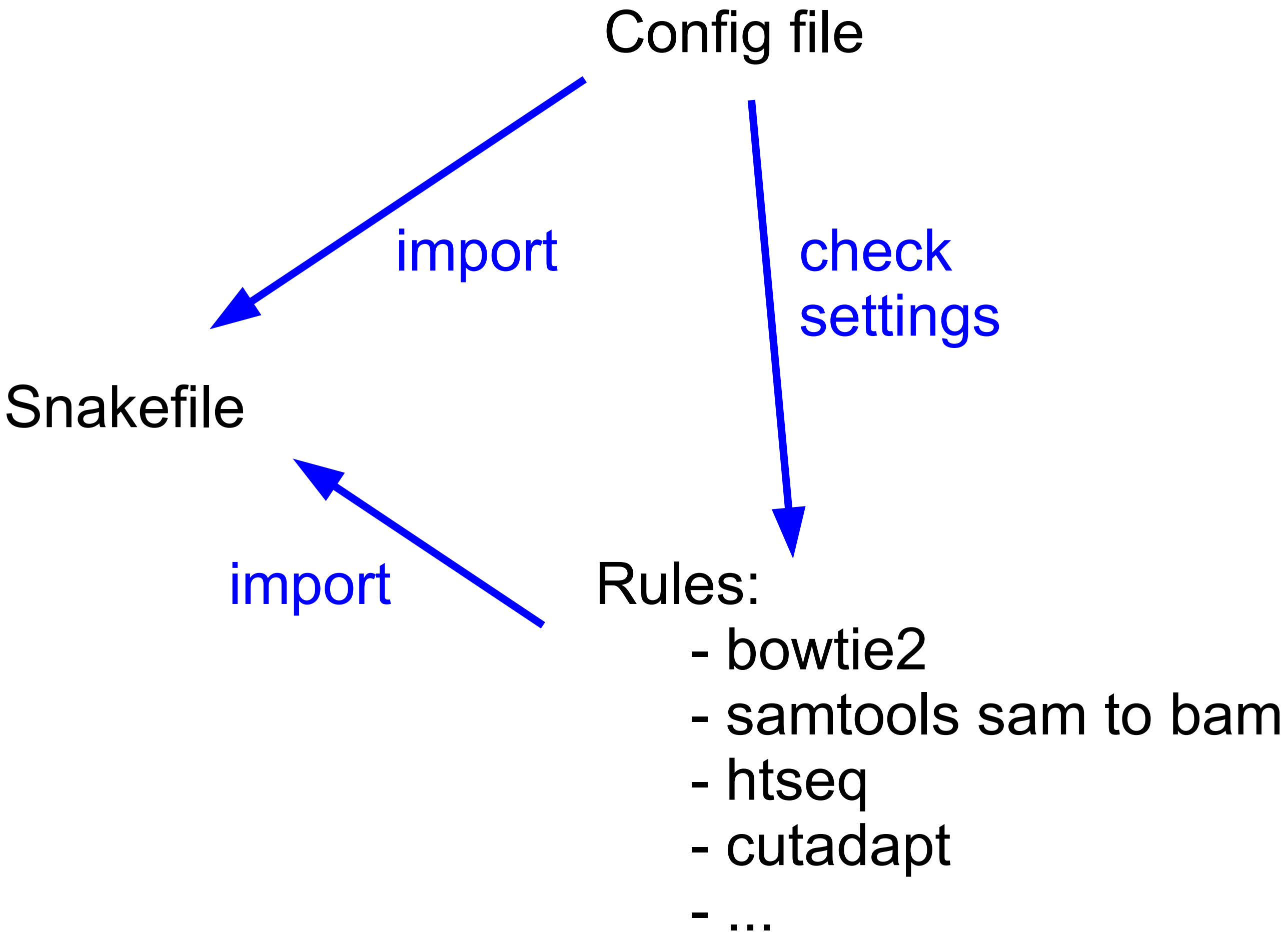
Snakefile

import

Rules:

- bowtie2
- samtools sam to bam
- htseq
- cutadapt
- ...

Snakemake pipeline organization



Config file

import

check settings

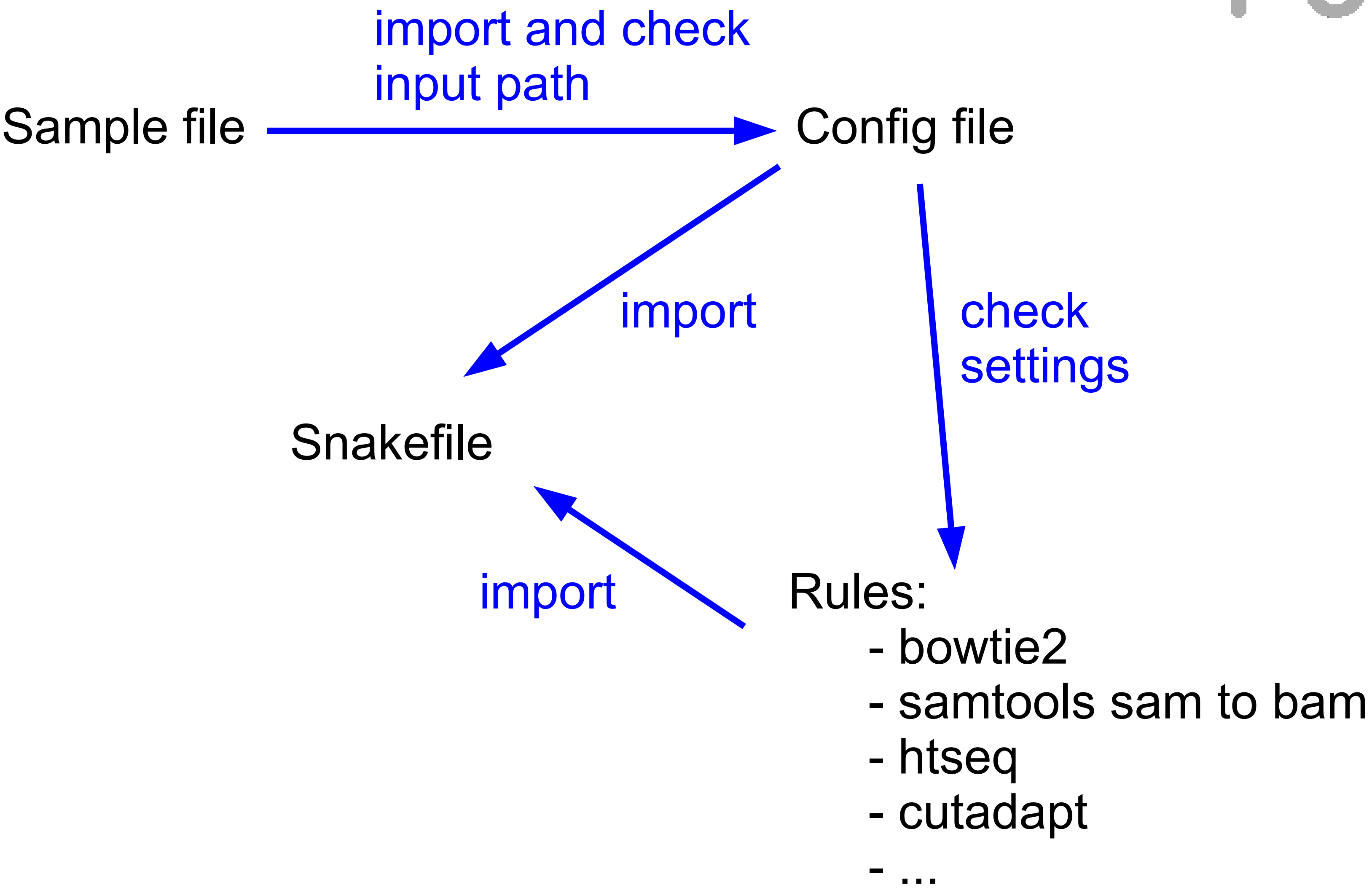
Snakefile

import

Rules:

- bowtie2
- samtools sam to bam
- htseq
- cutadapt
- ...

Snakemake pipeline organization



Sample file

import and check
input path

Config file

import

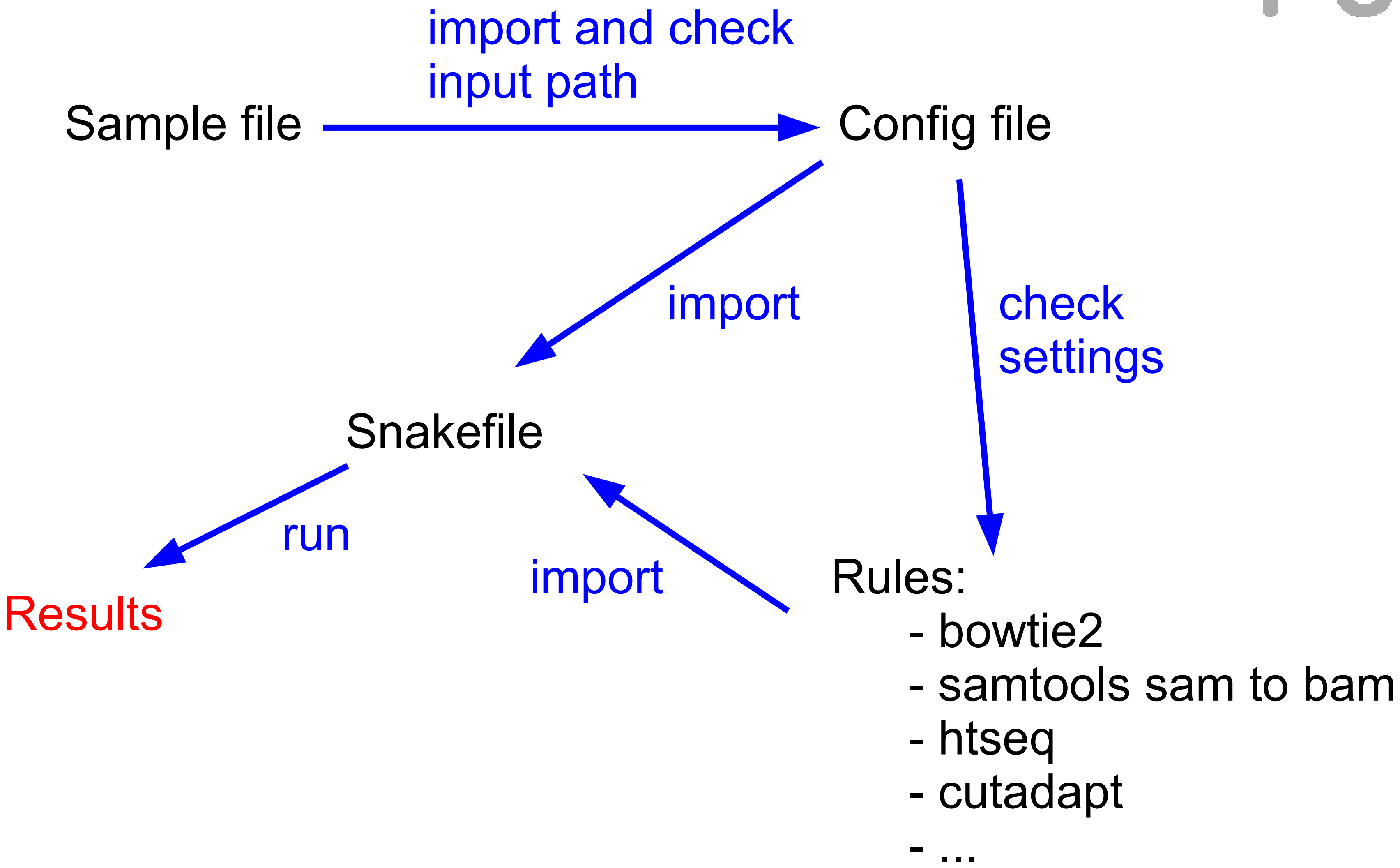
check
settings

Snakefile

import

Rules:
- bowtie2
- samtools sam to bam
- htseq
- cutadapt
- ...

Snakemake pipeline organization



Sample file

import and check input path

Config file

import

check settings

Snakefile

run

Results

import

Rules:

- bowtie2
- samtools sam to bam
- htseq
- cutadapt
- ...

Snakemake pipelines development

→ Pre-production :

- RNA seq

Exemple – new RNA pipeline

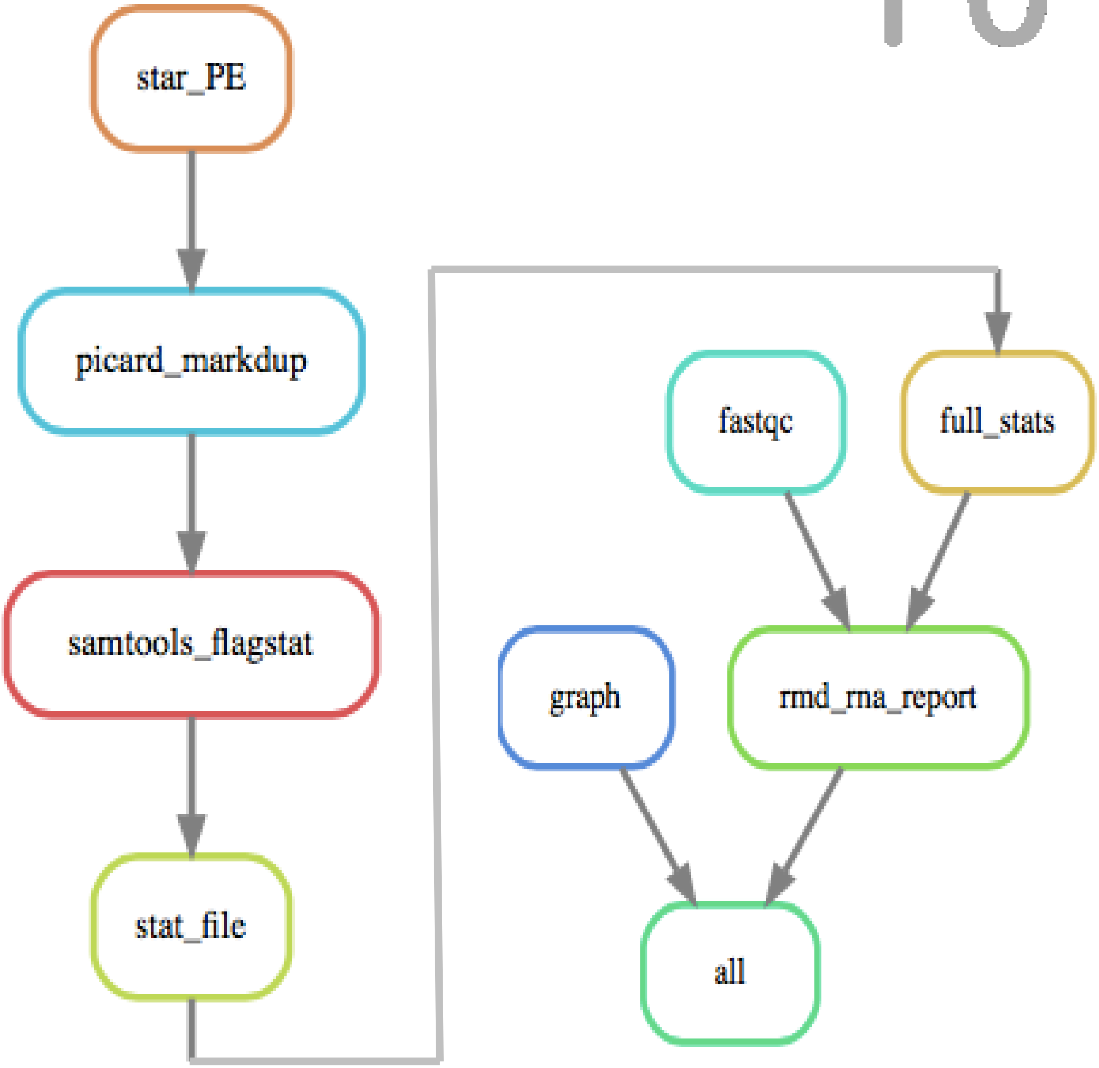
→ 11 rules

→ PE or SE

→ Tophat2 or STAR

→ inputs: fastq.gz / bcl

→ outputs:
-fullStatFile.xls
-report.html
- bam
- ...



Snakemake pipelines development

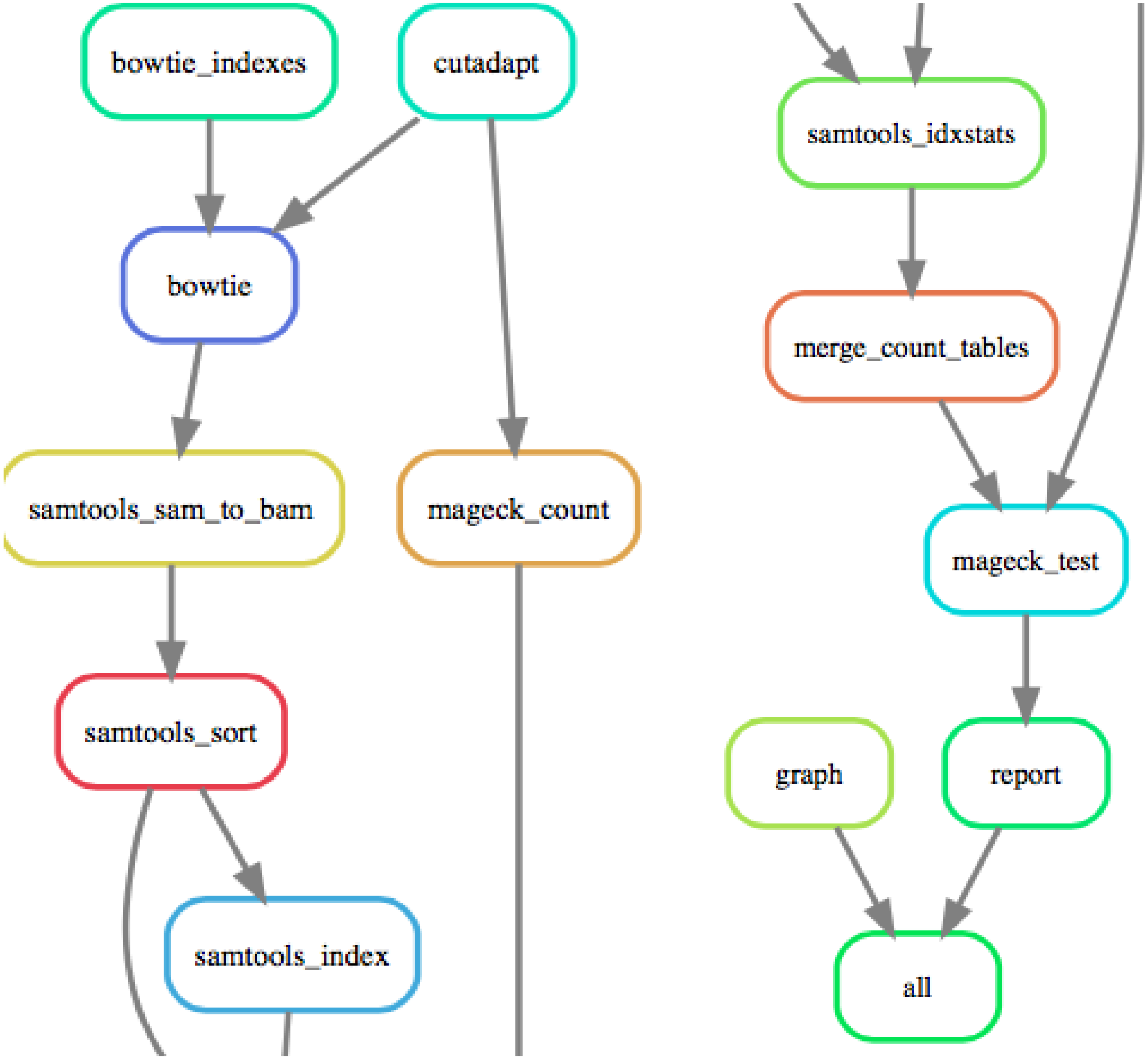
→ Pre-production :

- RNA seq
- GeCKO screen

Exemple – GeCKO screen pipeline

→ 13 rules

→ inputs:
  -fastq
→ outputs:
  -mageck results



Snakemake pipelines development

→ Pre-production :

- RNA seq
- GeCKO screen
- KDI (specific to the Institut Curie)

Snakemake pipelines development

→ Pre-production :

- RNA seq
- GeCKO screen
- KDI (specific to the Institut Curie)

→ Development :

- BRCA SOMATIC (diagnostic)
- TIGER (diagnostic)
- New features for the RNA seq pipeline

Thanks for your attention !