DE LA RECHERCHE À L'INDUSTRIE

www.cea.fr

G.R.A.A.L

Global **R**eads **A**n**A**lysis & visua**L**ization

Snakemake day at Pasteur | Hugo PEREIRA

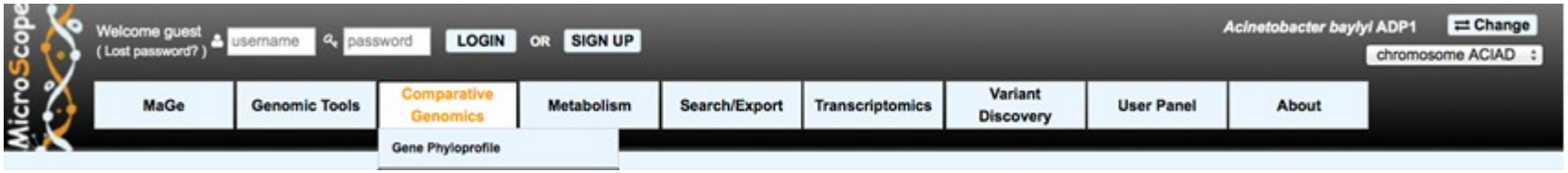
GENOSCOPE | LABGEM

SUMMARY

WHY DID WE CHOOSE SNAKEMAKE?

MICROSCOPE'S SERVICES PRESENTATION

Microbial Genome Annotation & Analysis Plateform



- Plateform for annotation & comparative analysis of bacterial genomes.

- Plateform certified ISO 9001

Genomes :

- > 5000 genomes, ~ 4 genomes / day.

NGS Analysis (polymorphism & RNAseq) :

- 300 fastq files analysed in 2016.
- Increasing each year

WORKFLOW PRESENTATION

Type of Analysis :

- Differential Expression
- Variant Calling

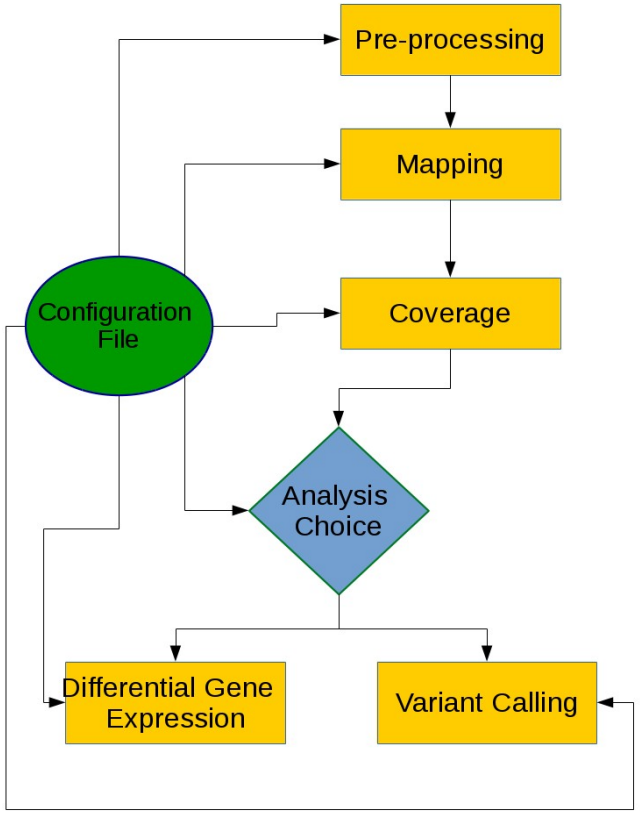
Type of Data :

- Single-End
- Paired-End
- Mate Pair

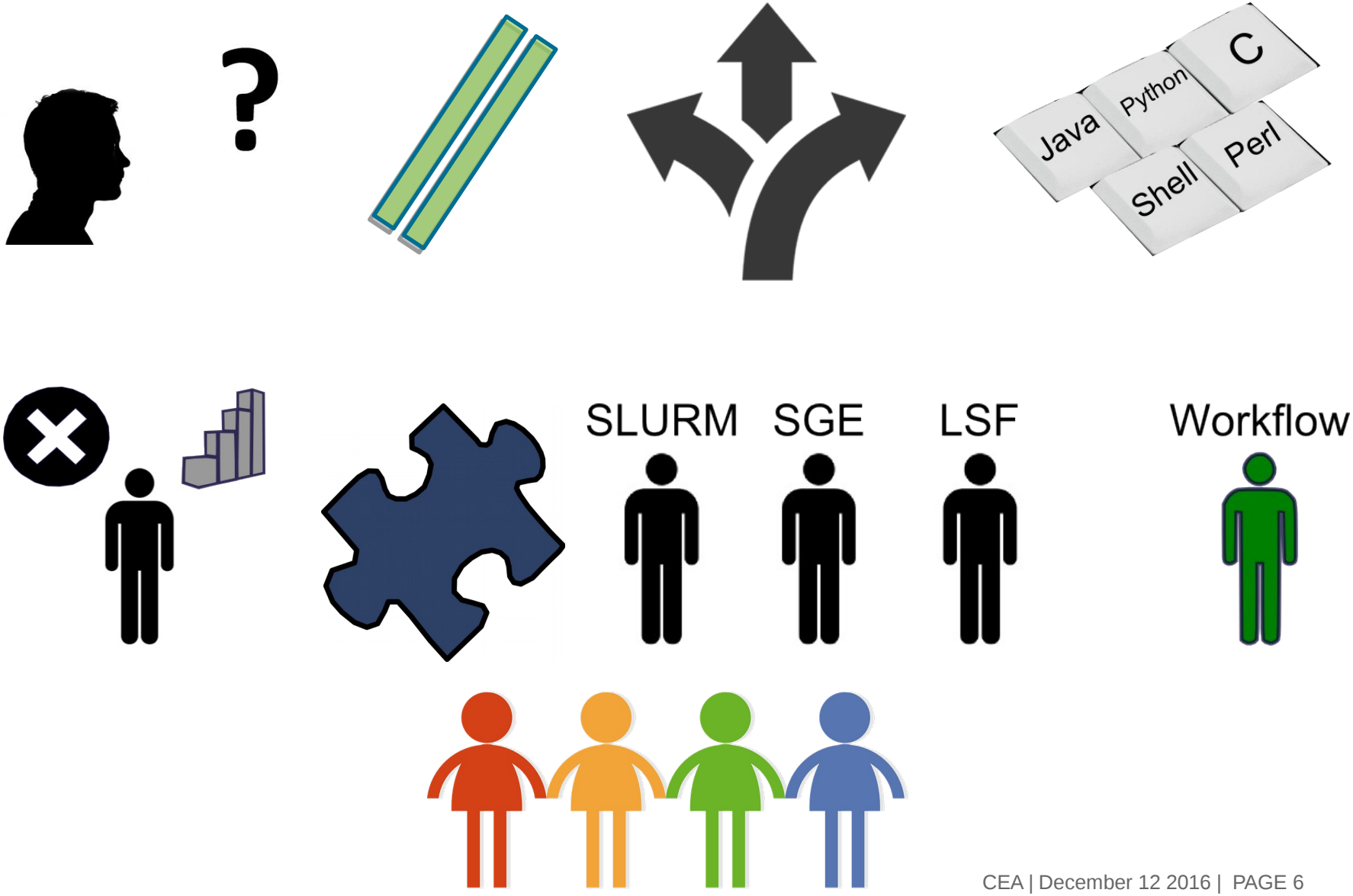
Multiple CheckPoint

Configuration File

All write in bash



WORKFLOW IMPROVEMENT WITH WORKFLOW ENGINE



IMPLEMENTATION OF SNAKEMAKE

COMPLEXITY'S DETERMINATION

Goal? - Type of Tool? - How many Rules?

Low Complexity :

```
Rule all :
      [ test.txt ]

Rule ProduceText :
 Input :
      getready.txt

 Params :
      config[text_into_test]

 Output :
      test.txt

 Shell/Script :
      'echo {params.text_into_test} > test.txt'
```

High Complexity :

```
# Include all rules
Include : AddingWords.rules

Include : RemovingWords.rules

Include : ProduceText.rules

# File to produce
Rule all :
      [ test.txt ]
```

RULES' LIBRARY

Complexity Separation :

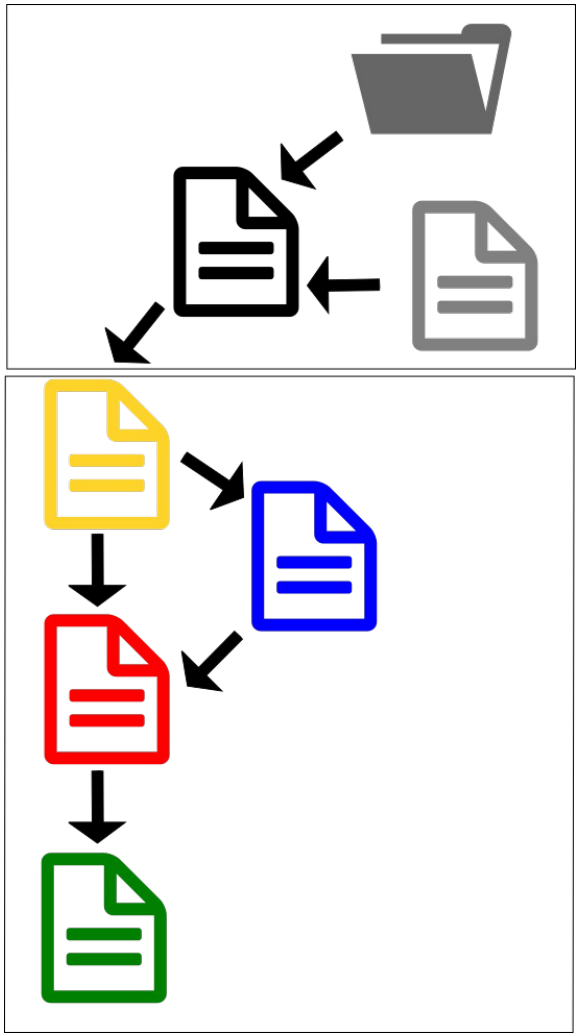
- Easy to understand
- Easy for bug fix

Reusable rules

- Multiple workflow
- Exchange of rules

Need modular rules

- Separate some informations
- Adaptable variables
- Call script bash



CODE CONVENTION'S ESTABLISHMENT

Configuration file - yml

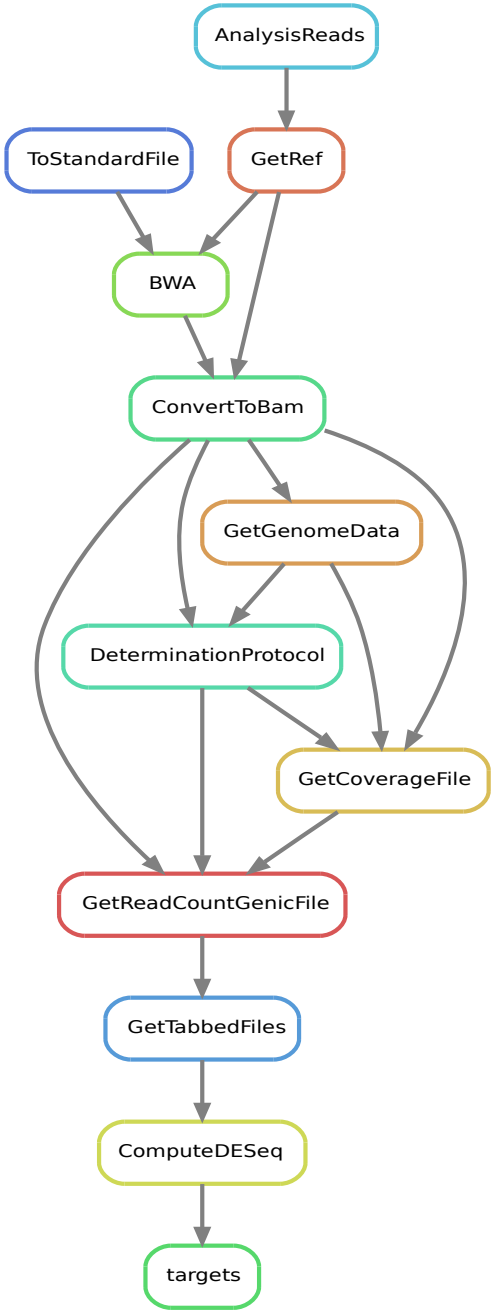
```
# ~~~~~ Names files and path associated corresponding to input
files :
      file1 : path/to/file1.txt
      file2 : path/to/file2.txt
      file3 : path/to/file3.txt

# ~~~~~ Path for rules and scripts
path :
      rules : path/to/rules/

# ~~~~~ Rules Parameters
# ===== Transformations
# ~~~~~ Parameters for TransformText
TransformText :
      repertory : Transformed/
      extension : .txt
      suffix : _trf
```

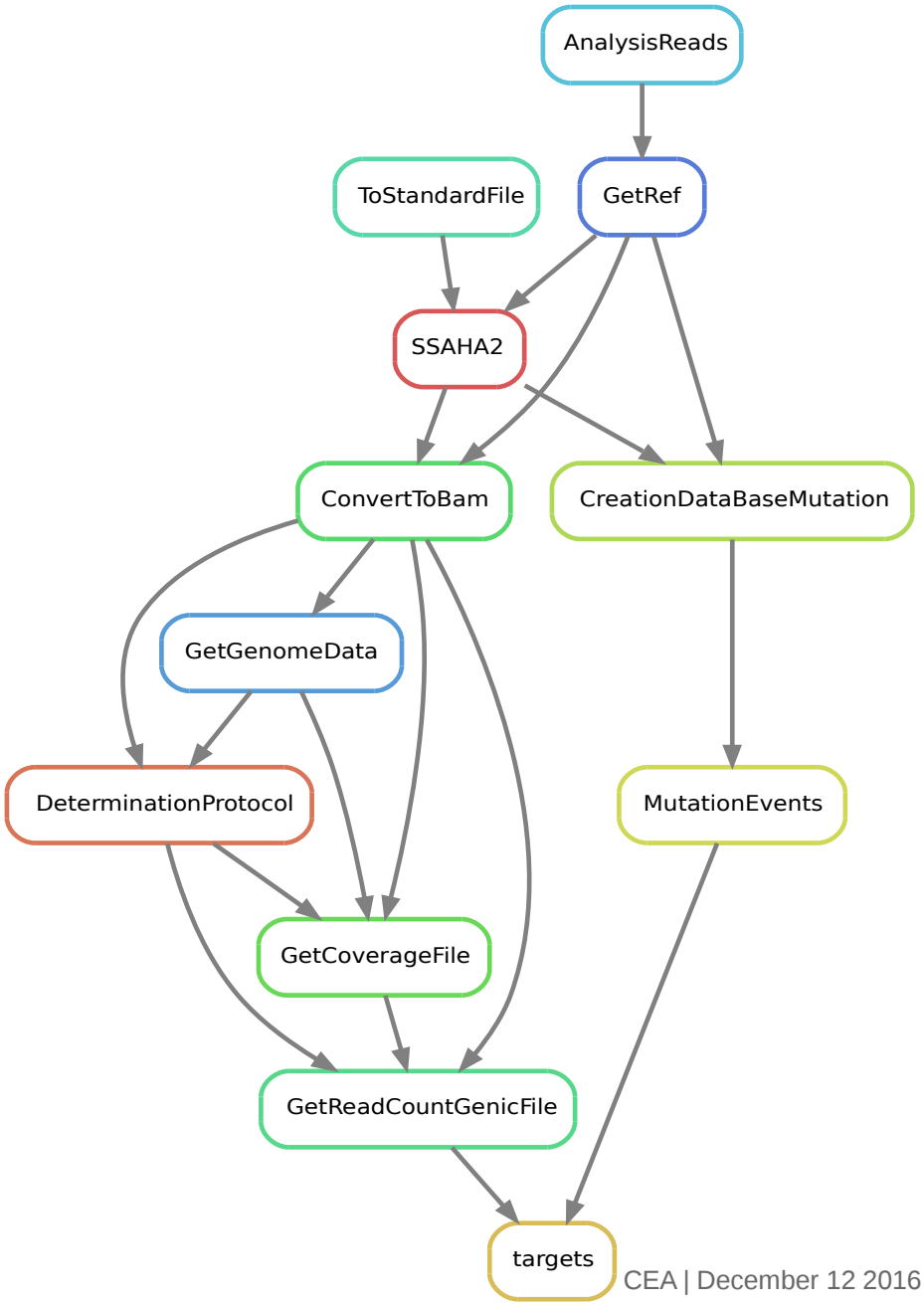
Snakefile

```
'''
    General description & License
'''
#####    Imports
'''
 Import descriptions
'''

import some_library

####    Rules input & Output
'''
    Input & Output Description
'''

_input_transformtext = lamba : config['files'][wildcard.input]

_output_transformtext = config['AddingWords']['repertory'] + '{input}' + \
    config['TransformText']['suffix'] + config['AddingWords']['extension']
```

```
#####    Include All Rules
'''
  Include descriptions
'''
path_to_rules = config['path']['rules']

include: path_to_rules + 'TransformText.rules'

#####    Finals Targets
'''
  Describe and Define all final targets
'''


final = expand('{repertory}{filename}{suffix}{extension}',
               repertory = config['TransformText']['repertory'],
               filename = config['files'].keys(),
               suffix = config['TransformText']['suffix']
               extension = config['TransformText']['extension'])

#####    Rule Targets
rule targets:
  input: [final]
```

FINALS WORKFLOWS


AnalysisReads
ToStandardFile
GetRef
BWA
ConvertToBam
GetGenomeData
DeterminationProtocol
GetCoverageFile
GetReadCountGenicFile
GetTabbedFiles
ComputeDESeq
targets


AnalysisReads
ToStandardFile
GetRef
SSAHA2
ConvertToBam
CreationDataBaseMutation
GetGenomeData
DeterminationProtocol
MutationEvents
GetCoverageFile
GetReadCountGenicFile
targets

Commissariat à l'énergie atomique et aux énergies alternatives
Centre de Saclay | 91191 Gif-sur-Yvette Cedex

Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019

CEA
Genoscope
LABGeM