CRCM
Centre de Recherche
en Cancérologie de Marseille

Les mardis de la technologie gourmande de DISC

Snakemake et conda, pourquoi, comment ?

Lucie Khamvongsa et Arnaud Guille

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale
INSTITUT PAOLI-CALMETTES
cnrs
Aix*Marseille université

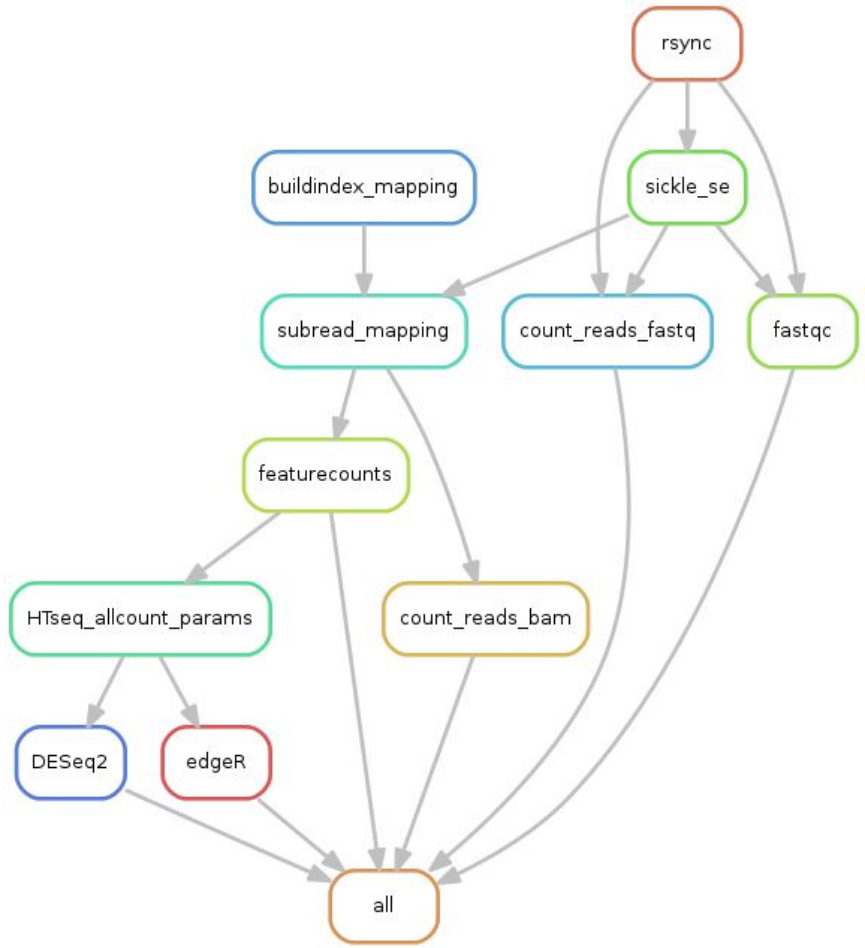
Snakemake qu’est-ce que c’est ?

- **Gestionnaire de pipeline**

Snakemake qu’est-ce que c’est ?

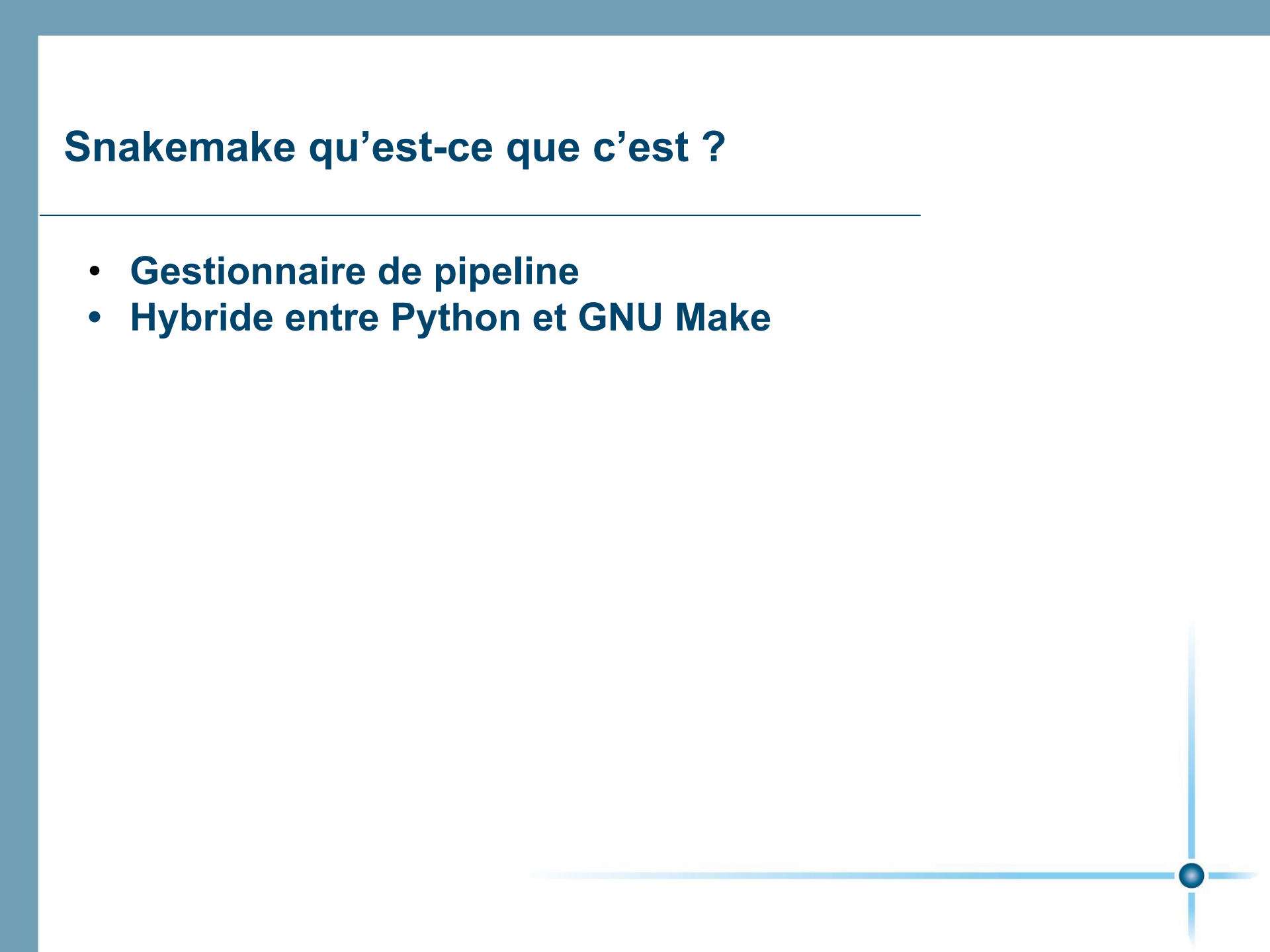
- **Gestionnaire de pipeline**

En informatique il désigne un groupe de logiciels exécutés en série de telle façon que la sortie d'un logiciel sert d'entrée pour le suivant.



Snakemake qu’est-ce que c’est ?

- Gestionnaire de pipeline
- Hybride entre Python et GNU Make



Snakemake qu'est-ce que c'est ?

- **Hybride entre Python et GNU Make**

Héritage GNU Make:
Gestion des opérations pour obtenir un type de fichier (défini par son extension)
Gestion des dépendances entre tâches

Héritage Python:
Syntaxe

```
rule all:
    input:
        "plots/dataset1.pdf",
        "plots/dataset2.pdf"


rule plot:
    input:
        "raw/{dataset}.csv"
    output:
        "plots/{dataset}.pdf"
    shell:
        "somecommand {input} {output}"
```

Avantages additionnels:
Gestion automatique de la création de dossiers
Suppression automatique des fichiers temporaires
Possibilité d'insérer du code en R et Python dans les règles
Personnalisation par un fichier de configuration (YAML)

Snakemake qu'est-ce que c'est ?

- Gestionnaire de pipeline
- Hybride entre Python et Make
- Analyses reproductibles et évolutives

Snakemake qu'est-ce que c'est ?

- **Analyses reproductibles et évolutives**

```
rule all:
    input:
        "plots/dataset1.pdf",
        "plots/dataset2.pdf"


rule plot:
    input:
        "raw/{dataset}.csv"
    output:
        "plots/{dataset}.pdf"
    shell:
        "somecommand {input} {output}"
```

config.yaml:

```
genome:
  organism: dm6
  fasta_file: genome.fa
  gtf_file: genes.gtf

quality_control: FastQC
trimming: sickle
mapping: subread-align
read_counts: featureCounts
diffexpr: DESeq2
```

Snakemake qu'est-ce que c'est ?

- **Analyses reproductibles et évolutives**

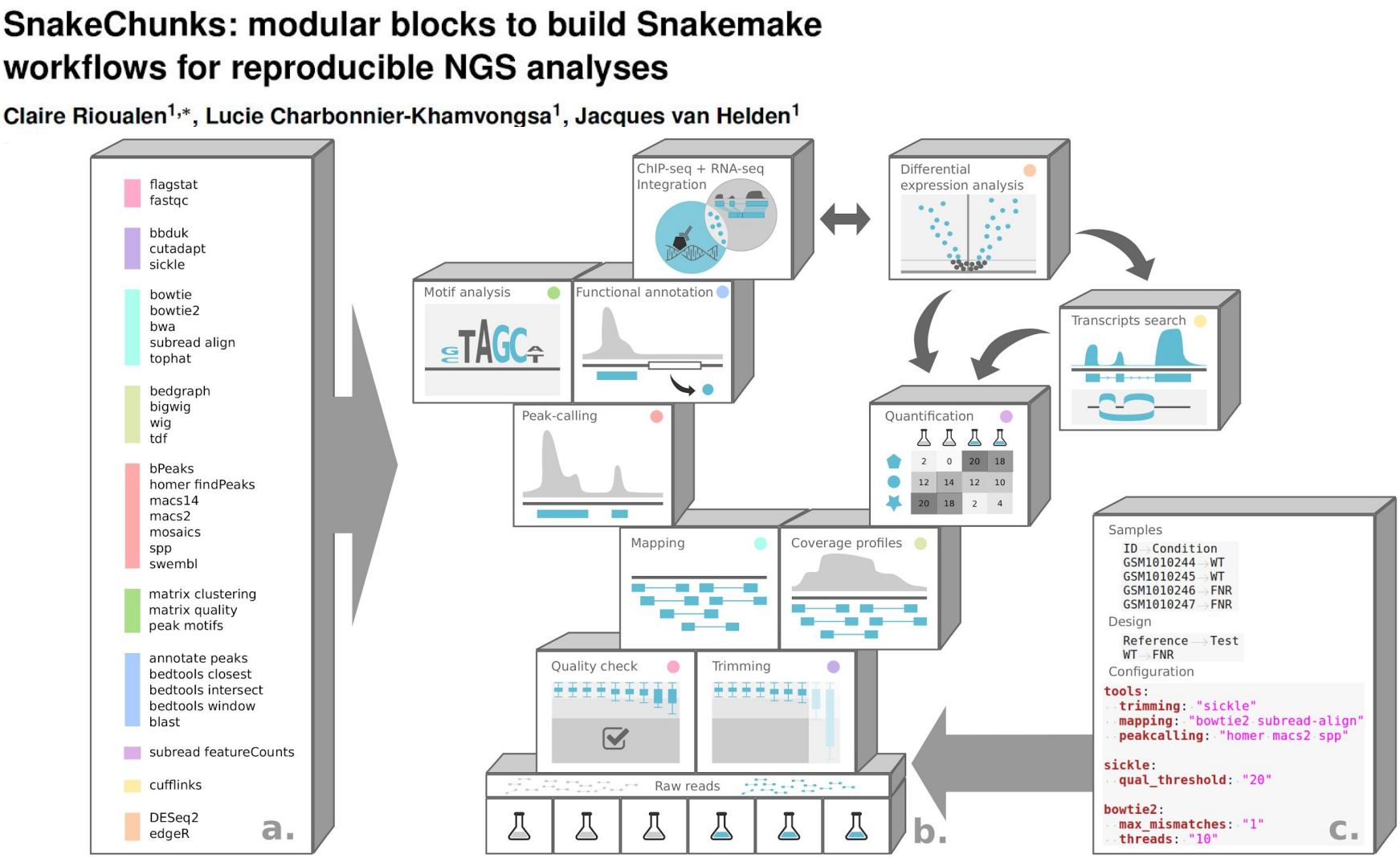
```
rule all:
    input:
        "plots/dataset1.pdf",
        "plots/dataset2.pdf"


rule plot:
    input:
        "raw/{dataset}.csv"
    output:
        "plots/{dataset}.pdf"
    shell:
        "somecommand {input} {output}"
```

```
include:"plot.rules"


rule all:
    input:
        "plots/dataset1.pdf",
        "plots/dataset2.pdf"
```

Snakemake qu'est-ce que c'est ?



Snakemake qu'est-ce que c'est ?

- Gestionnaire de pipeline
- Hybride entre Python et Make
- Analyses reproductibles et évolutives
- Adapté aux environnements serveur, cluster, grid et cloud

Conda : Pourquoi ?

- La difficulté à installer un logiciel :
  - De quel version de python ai-je besoin, 2.7, 3.0 ?
  - Quels sont les packages à installer ?
  - Quels sont les versions des packages à installer ?
  - Comment installer le package sans être root et sans cesse embêter son admin.

- La gestion des différentes versions
  - Faire co-exister python2, python3, R2.7, R3.0 etc...
  - Faire co-exister différentes librairies C, R, Java etc..

Conda : C’est quoi ?

- **Un gestionnaire de paquet et d’environnement pour vos languages préférés (Python, R, Ruby, Lua, Scala, Java, JavaScript, C/ C++, FORTRAN)**
- **Compatible Linux, Mac, Windows**
- **Open source**

https://conda.io/docs/#

Conda : Gestionnaire de paquets (1)

- **Par défaut la recherche de logiciel ou paquet se fait sur le dépôt officiel avec la commande**

conda search

- **Possibilité d'ajouter d'autres dépots comme conda-forge ou bioconda**

```
conda config --add channels conda-forge
conda config --add channels bioconda
```

https://bioconda.github.io/

Conda : Gestionnaire de paquets (2)

- **Installation d'un paquet sans création d'environnement**

```
conda install bwa
```

- **Installation d'un paquet avec création d'un environnement**

```
conda create -n aligners bwa bowtie hisat star
```

Conda : Gestionnaire d'environnements

- Un environnement peut être vu comme un espace de travail isolé du reste qui contient ses propres logiciels, paquets, librairies et chemins.
- Permet de séparer proprement différents projets
- Évite les problèmes de dépendances et conflits entre différentes versions.

Conda : Gestionnaire d'environnements

- **Création d'un environnement et installation des paquets dedans**

```
conda create --name snowflakes biopython
```

- **Activation de l'environnement**

```
source activate snowflakes
```

- **Dé-activation de l'environnement**

```
source deactivate
```

- **Suppression de l'environnement**

```
conda env remove -n snowflakes
```

- **Lister les environnements**

```
conda env list
```

Conda : Gestionnaire d'environnements à partir d'un fichier yaml

```
name: somatic_sv

channels:
   - conda-forge
   - bioconda
dependencies:
   - delly
   - lumpy-sv
   - manta
   - gridss
   - pyvcf
```

- **Création de l'environnement**

```
conda env create -f mon_fichier.yml
```

Snakemake et conda :

- **Un gestionnaire de pipeline + un gestionnaire de paquet et d'environnement = un bioinformaticien heureux**

```
rule NAME:
    input:
        "table.txt"
    output:
        "plots/myplot.pdf"
    conda:
        "envs/ggplot.yaml"
    script:
        "scripts/plot-stuff.R"
```

Snakemake et conda :

- **Un gestionnaire de pipeline + un gestionnaire de paquets et d'environnements = un bio-info heureux**

```
rule NAME:
    input:
        "table.txt"
    output:
        "plots/myplot.pdf"
    conda:
        "envs/ggplot.yaml"
    script:
        "scripts/plot-stuff.R"
```

```
snakemake --snakefile my_snakefile --use-conda
```

Merci !!!

Des questions ???