INRA-PACA

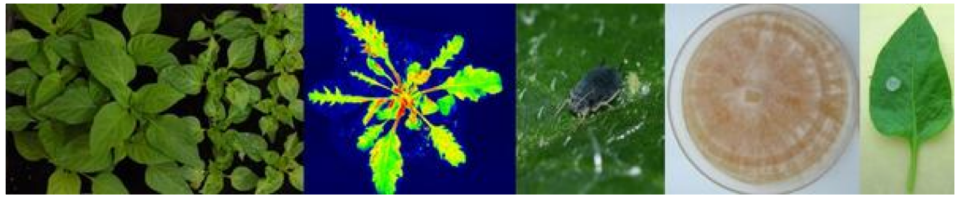
Formation et support

Bioinformatique et reproductibilité au sein de l'unité Génétique et Amélioration des Fruits et Légumes (GAFL)

Jacques Lagnel
Jérémy Verrier

Journée Calcul IRSTEA/INRA *14 juin 2019*

GAFL
INRA AVIGNON



2 équipes de recherches

DADI: Diversité, Adaptation, Déterminants et Intégration

ReDD: Résistance aux pathogènes et aux ravageurs, Diversité et Durabilité

CRB => 10,000 accessions
Prunus
Piments
Tomates
Melons
Aubergines
Arabidopsis
Puceron
Oomycètes

- Génotypages: GBS, capture
- Re-séquençage (WGS Illumina)
- Séquençage de novo (long reads, 10X genomics, optical mapping,…)
- Transcriptomics: RNAseq, IsoSeq
- Analyses de QTL, GWAS
- Phénotypages

Besoin de « reproductibilité »
et application du principe FAIR au GAFL

Bioinformatique/bio-analyse

Projet développé au GAFL pour remplir 4 critères :

- Reproductibilité des résultats
- Un « formalisme » adapté est la clef pour assurer la reproductibilité des expériences
- Portabilité entre environnements Linux et plateformes de calculs (clusters)
- Facilité d'utilisation, bonnes pratiques et accompagnement

Problématiques rencontrées

- Pas d'analyse de "routine"
- Impossibilité de reproduire les résultats de l'analyse computationnelle
- Portabilité
- Ressources de calcul limité au GAFL (assemblages de novo)
- Impossibilité d'installer des outils, OS non compatible
- Dépendances complexes ou plus disponible
- Mise à jour de l'outil rendant inutilisable les codes
- Changement des arguments des fonctions utilisées (R, python,...)
- Version des packages

Environnement virtuel

Nombreux logiciels d'analyses

- Impossibilité d'installer des outils, OS non compatible
- Dépendances complexes ou plus disponible
- Mise à jour de l'outil rendant inutilisable les codes
- Changement des arguments des fonctions utilisées (R, python,...)
- Version des packages

Plus de 40 containers créés au GAFL
une application (bwa) à un regroupement d'applications (>10) /container

→ **Bonnes pratiques**

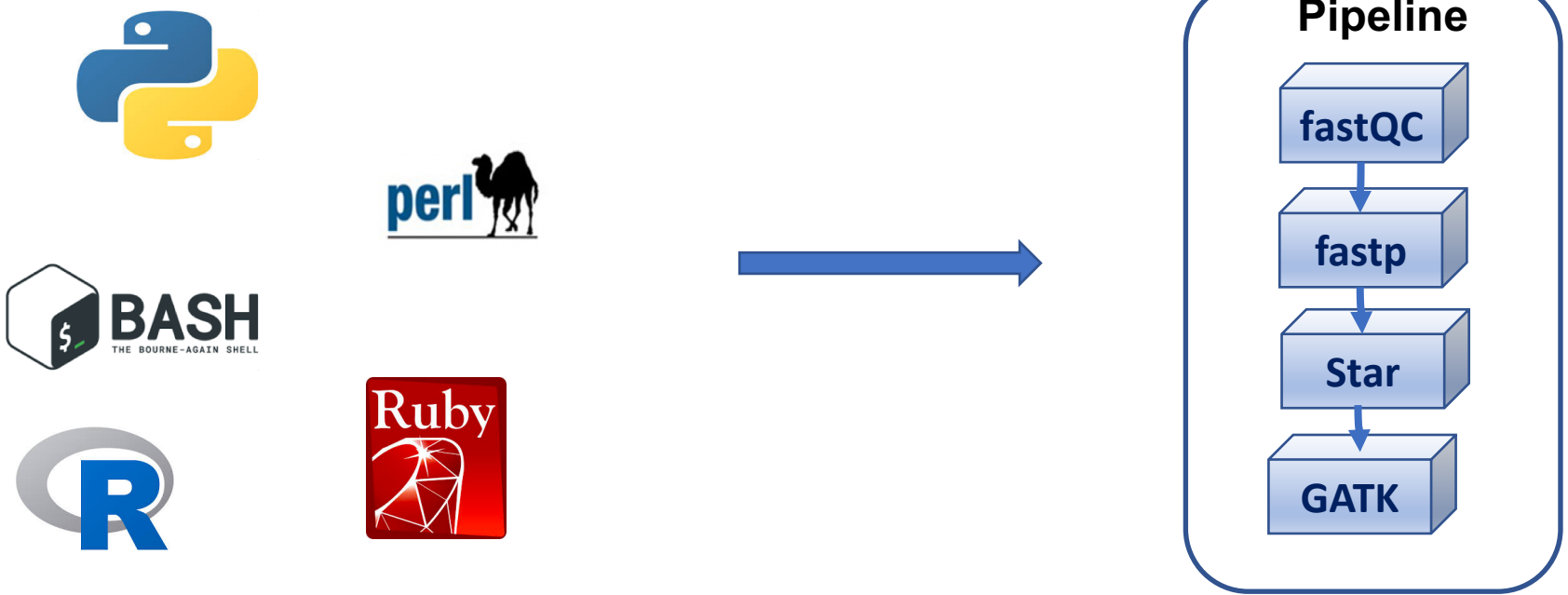
Chaine de traitement des données

- Grande quantité de données
  >100 fichiers avec >10M Seq./>1Go/fichier
- Enchainement de nombreuses étapes d'analyses

Findable
Accessible
Interoperable
Reusable

Pour chaque étape => différents logiciels

Création de scripts d'analyse (Pipeline)



Chaine de traitement des données

- Grande quantité de données
  >100 fichiers, >10M items/>1Go/fichier
- Enchainement de nombreuses étapes d'analyses

Findable
Accessible
Interoperable
Reusable

Pour chaque étape => différents logiciels

Création de scripts d'analyse (Pipeline)

Manque de standardisation
Problèmes de reproductibilité
- **Interopérabilité => Coopération des outils**
- **Réutilisation. => Protocole rejouable simplement**

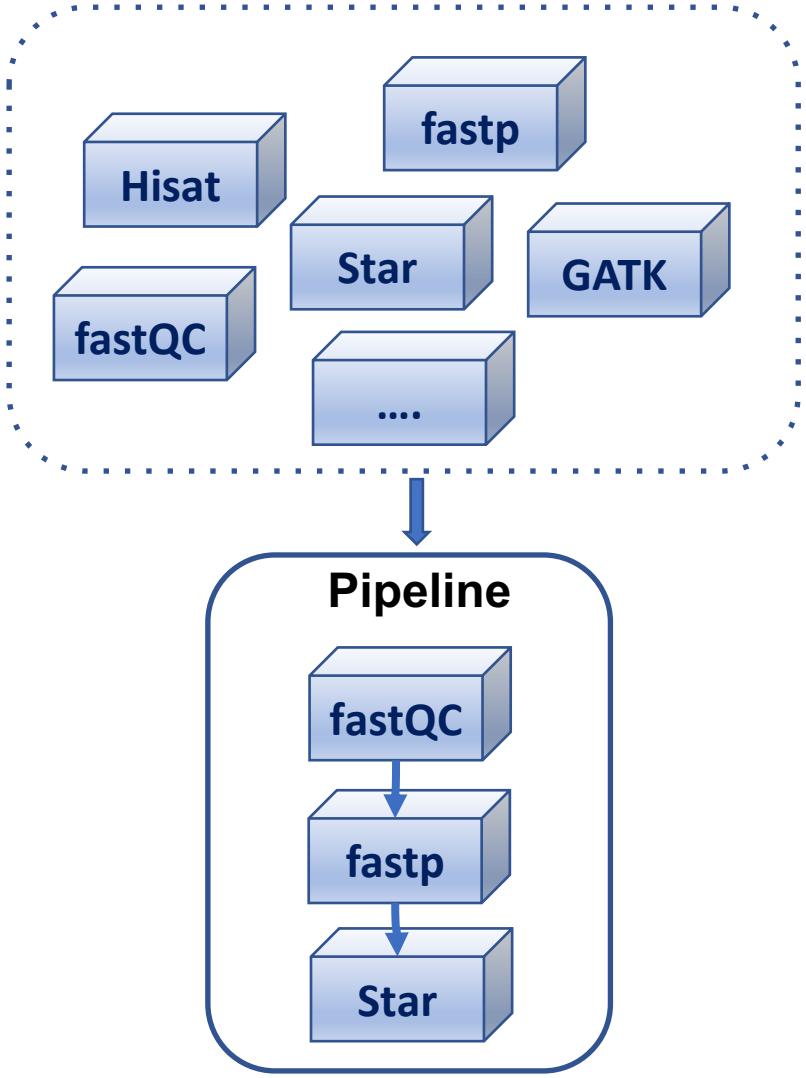
GATK

Création de scripts d'analyse
Utilisation de workflow management system

Snakemake

- Pipeline framework basé sur +
- Facilite la création des pipelines
- Facilite la traçabilité et la reproductibilité des analyses
- Briques indépendantes (bash, python, R,...)
- Utilisation de containers (Singularity)
- Parallélisations
- utilisation de SLURM, SGE,…
- Reprise sur erreur
- Portabilité (fichier de configuration)
- Facile à lancer

Modulaire →



Besoin de « reproductibilité »
et application du principe FAIR au GAFL

GAFL

1 Serveur de calcul
40 cores, 512Go RAM,22To

Findable
Accessible
Interoperable
Reusable

Technologies mises en place

Versionning (git) → git, GitLab

gestionnaire de ressources → slurm workload manager

Environnement virtuel (container) →  , Singularity

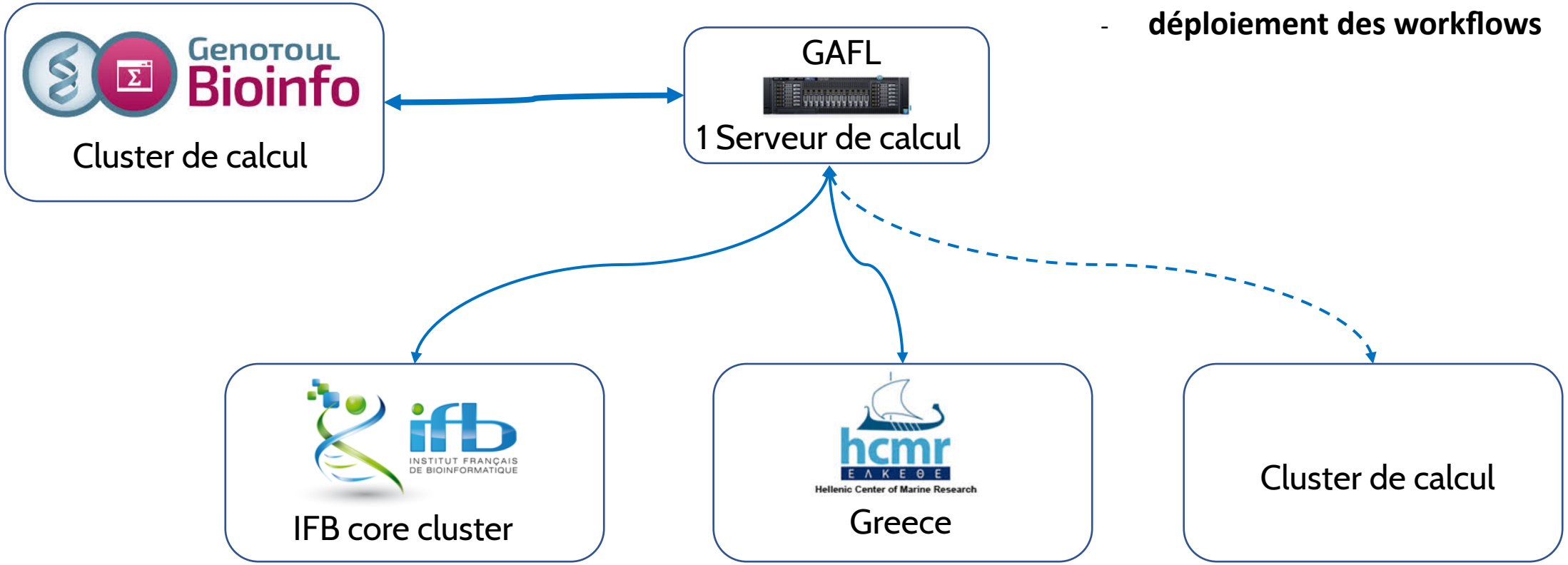
Création de scripts d'analyse: (workflow management system) → Snakemake

Rapport d'analyse (notebook & reporting) → rmarkdown

Déporter les analyses

condition requise: Singularity, Snakemake et SLURM

- **Accessibilité des données**
- **Déploiement des images singularity**
- **déploiement des workflows**



Besoin de « reproductibilité »
et application du principe FAIR au GAFL

Appui/accompagnement

- Accompagnement dans la gestion des données
- Accompagnement et support dans l'analyse de données
- Montée en compétence en bioinfo/bioanalyse des agents de l'unité
- Assurer les bonnes pratiques
- Formation des utilisateurs à des outils de reproductibilité

Animation bioinformatique

- Organisation de workshops
- Création de tutoriels

→

Apprentissage d'un seul workflow management system
=> Snakemake
Utilisation de containers
=> Singularity

⇒ **Apprentissage d'outils interopérables adaptés à la majorité des moyens de calcul**
⇒ **« Réduire » la courbe d'apprentissage et assurer une meilleure reproductibilité**

MERCI DE VOTRE ATTENTION!

Membres de l'équipe support:
BITTON Frédérique,
CHADOEUF Joël,
ELBELT Sonia,
LAGNEL Jacques,
LE-CALONNEC Emmanuel

Backup slides

Besoin de « reproductibilité »
et application du principe FAIR au GAFL

Appui/accompagnement

Appui/accompagnement:

- Proposition de formation
- Elaboration de workflows d'analyses (Snakemake)
- Autonomie en bio-anayse des utilisateurs

Formations locales: Ateliers

- Travail distant (connections, SLURM)
- Snakemake
- Création de containers (pour les dev)
- Utilisation de containers
- Rmarkdown
- analyses thématiques

Formations extérieures:

- Linux
- Bioinformatique
- Biostatistique
- …

Environnement virtuel

Plus de 40 containers créés au GAFL
une application (bwa) à un regroupement d'applications (>20) /container

Bonnes pratiques

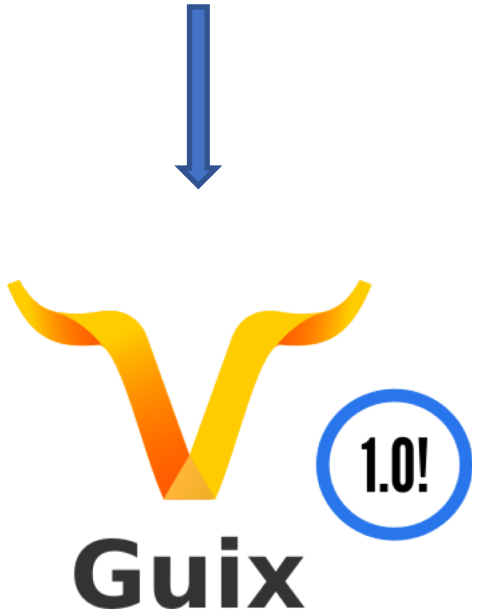
- **Création d'image à partir de recettes**
- **User et abuser des %labels et %help**
- **Une section %test**
- **Convertir la recette Dockerfile en recette Singularity et non l'image docker**
- **Versionner les recettes**
- **Spécifier la version du package bioinfo (source/conda/pip)**
- **Favoriser 1 app/container**

Singularity:
problème de sécurité
setUID

Singularity:
Commercial:
Sylabs


```
BootStrap: docker
From: debian:9.5

%help
Run bwa v0.7.17
Usage: bwa -h

%setup
mkdir ${SINGULARITY_ROOTFS}/data

%labels
jacques lagnel <jacques.lagnel@inra.fr>
Date modification: 06/06/2019
Version v1.0
Software bwa v0.7.17

%environment
    export LC_ALL=en_US.utf8
    export PATH=/usr/local/bin:$PATH

%post
apt-get -y update
apt -y install bwa

%runscript
bwa "$@"
```

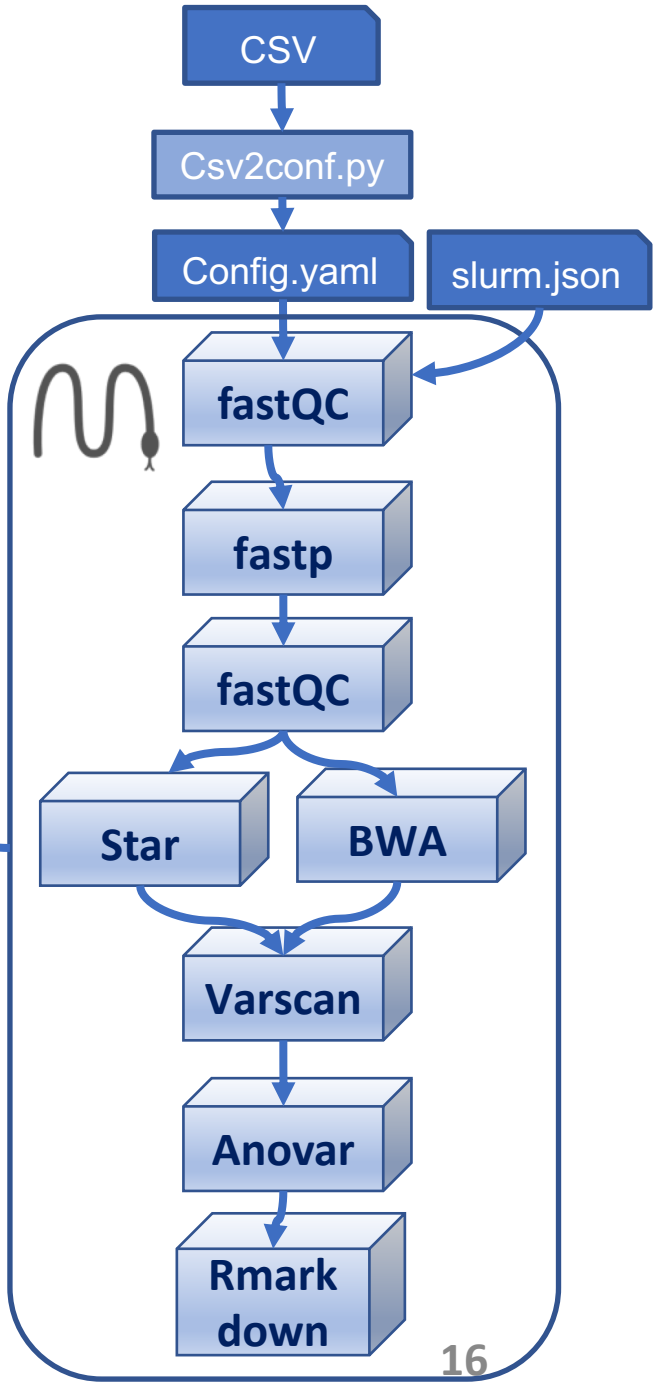
Singularity:
problème de reproductibilité



AllMine, un pipeline flexible d'*Allele mining*

Thomas Bersez M2

- **Recherche d'allèles de résistance de familles de gènes dans des accessions. Diversité génétique** chez les différentes variétés de plantes de culture : Source de **traits agronomiques d'intérêt.**
- Grands volumes de données de séquençage disponibles publiquement.
- Besoin d'un **outil informatique** de recherche de **variants** potentiellement associés à des résistances pour **validation en wetlab**.
- AllMine, recherche des **SNPs non synonymes** à partir de **données NGS DNA/RNAseq paired/single end** dans des **gènes d'intérêt** par comparaison avec un **génome de référence.**
- Génère un **fichier type tableur** listant les SNPs trouvés et leurs caractéristiques + un rapport d'analyse en **Rmarkdown**.



2 équipes de recherches

DADI: Diversité, Adaptation, Déterminants et Intégration

Objectifs:

- D'étudier la diversité des ressources génétiques au niveau du génome entier
- D'identifier de nouveaux gènes, QTL et géniteurs d'intérêt
- De comprendre la fonction des gènes dans des conditions de stress (biotiques et abiotiques)
- De développer la sélection multi-caractère pour des systèmes de production à faible niveau d'intrants.

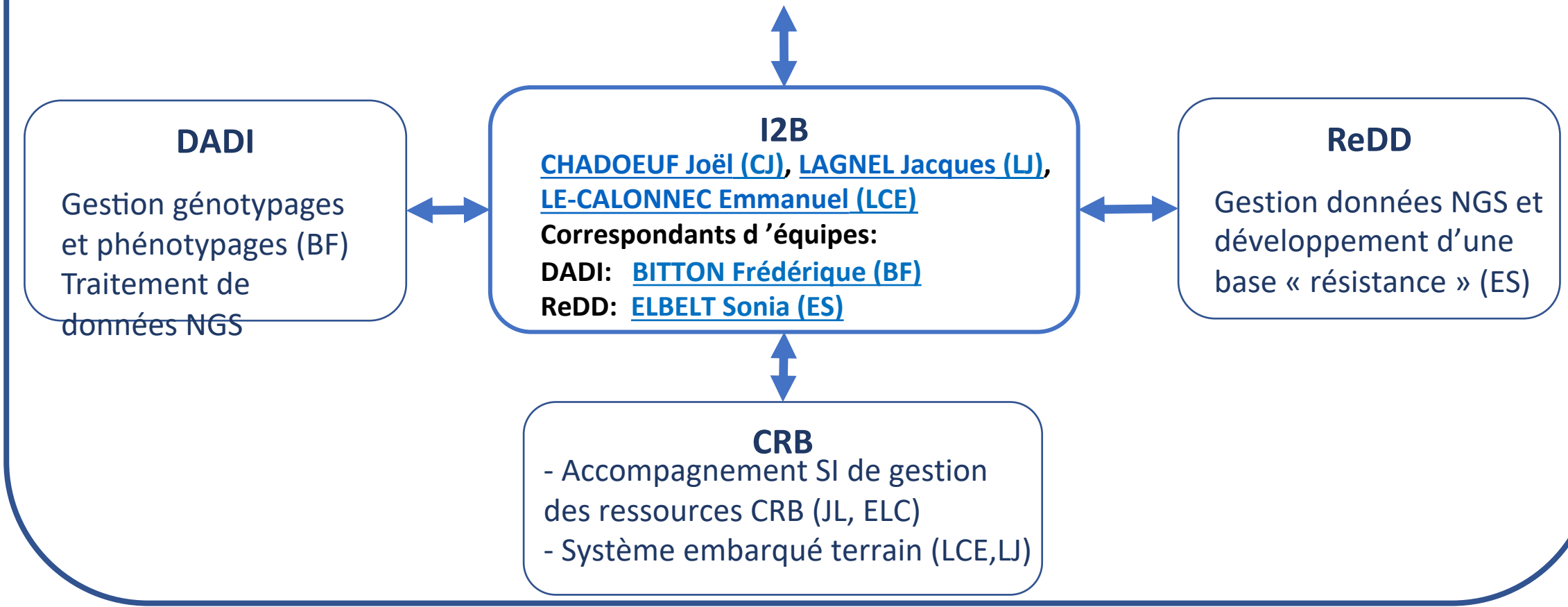
ReDD: Résistance aux pathogènes et aux ravageurs, Diversité et Durabilité

Objectifs:

- De caractériser les bases génétiques et fonctionnelles des résistances
- De mettre en évidence la diversité des facteurs de résistances génétiques chez les plantes
- De mettre à disposition de nouveaux allèles et de nouvelles combinaisons génétiques en vue d'élargir les spectres et niveaux de résistance, ainsi que leur durabilité
- De prédire la capacité des pathogènes et des ravageurs à s'adapter aux plantes en étudiant leurs modes d'interaction

GAFL

- **Accompagnement PGD et données FAIR.**
- **Accompagnement bioinformatique analyse de données**
- **Bonnes pratiques: reproductibilité, traçabilité. Utilisation de workflow manager (Snakemake).**
- **Portabilité des outils et pipelines (Containers)**



Besoin de « reproductibilité » et Application du principe FAIR au GAFL

Gestion des données

Bioinformatique/Bioanalyse

Besoin de « reproductibilité » et Application du principe FAIR au GAFL

✓ Gestion des données

- Comment assurer la traçabilité des données ?
- Comment accéder aux données?
- Comment stocker, protéger et exploiter ces données ?
- Proximité données<->calcul

Gestion des données brutes du GAFL et leur réutilisation.

NGS données brutes

- **Regrouper et Structurer** les données NGS
  Arborescence commune:
  Taxonomie, molécule, technologie,..
- **Métadonnées**
  utilisation d'un vocabulaire contrôlé
- **Accessibilité** et **sécurisation** des données

Gestion des données brutes du GAFL et leur réutilisation.

GAFL INRA AVIGNON

Findable
Accessible
Interoperable
Reusable

NGS données brutes →

- **Regrouper et Structurer** les données NGS
  Arborescence commune:
  Taxonomie, molécule, technologie,..
- **Métadonnées**
  utilisation d'un vocabulaire contrôlé
- **Accessibilité** et **sécurisation** des données

Plans de gestion de données → DMP OPIDoR

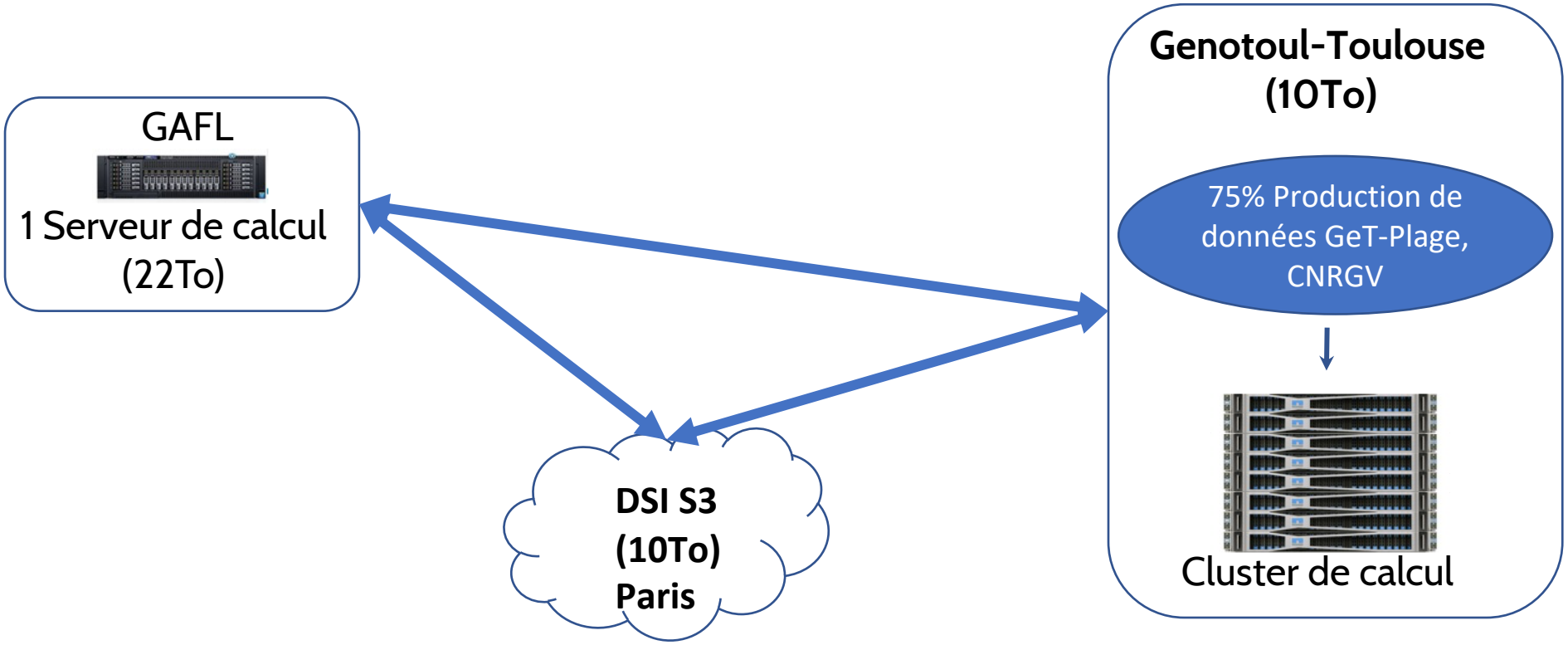
Génotypage et phénotypage → ThaliaDB
A Database for Association Studies

ENA European Nucleotide Archive
NCBI
The Dataverse Project
Datapartage

Gestion des données brutes du GAFL et leur réutilisation

Accessibilité, sécurisation des données et proximité données/calcul

6To données brutes NGS Tri-réplication



Gestion des données brutes du GAFL et leur réutilisation.

- Famille
  - Genre_espèce
    - DNA
      - GBS
      - Locus
      - Ref_genome
      - Arrays
      - WGS
        - Projet
          - Raw
          - Preproc
          - Results

Nécessité d'utiliser des métadonnées

Champ	Vocabulaire contrôlé	Description
Famille	oui	Nom latin de la famille
Espèce	oui	Genre_espèce
Accession	oui	Numéro d'accession (identifiant de l'échantillon)
Projet	oui	Acronyme du projet
Titre/objectif du projet	non	Champ libre avec titre ou objectif du projet
date début projet	oui	MM/AA
date fin projet	oui	MM/AA
responsable GAFL	oui	Permanant pouvant être contacter
Date dépot données	oui	MM/AA
Type de séquence	oui	Génomique, transcriptome, mtDNA,…
Technologie	oui	Technologie RNA-Seq, WGS, GBS, ..
Plateforme de séquençage	oui	Séquenceur utilisé (HiSeq2000, MinIon)
Profondeur	non	En Mreads ou profondeur
longueur read	non	Longueur valide pour des reads courts
Lien sur le serveur	non	Localisation exacte du fichier de séquence sur le serveur
Lien fiche descriptive	non	Optionnel Localisation exacte de la fiche descriptive du projet
Commentaire	non	Commentaire libre