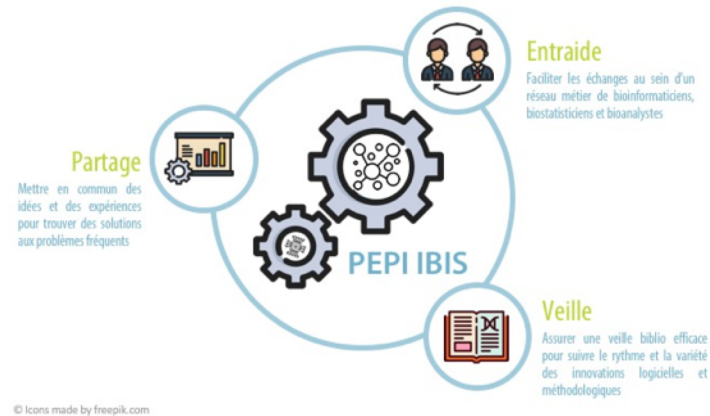


Bioinformatique et reproductibilité au sein de l'unité Génétique et Amélioration des Fruits et Légumes (GAFL): mise en place et retour d'expérience.

Jacques Lagnel



[Journées du PEPI IBIS - 6 & 7 juin 2019](#)

2 équipes de recherches

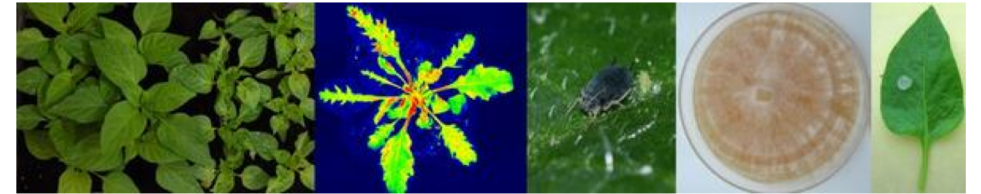
DADI: Diversité, Adaptation, Déterminants et Intégration



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



CRB => 10,000 accessions

Prunus

Piments

Tomates

Melons

Aubergines

Arabidopsis

Puceron

Oomycètes



**DADI: Diversité, Adaptation, Déterminants et
Intégration**

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

- **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



- ✓ **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

Findable 
Accessible 
Interoperable 
Reusable 



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

Plans de gestion de données



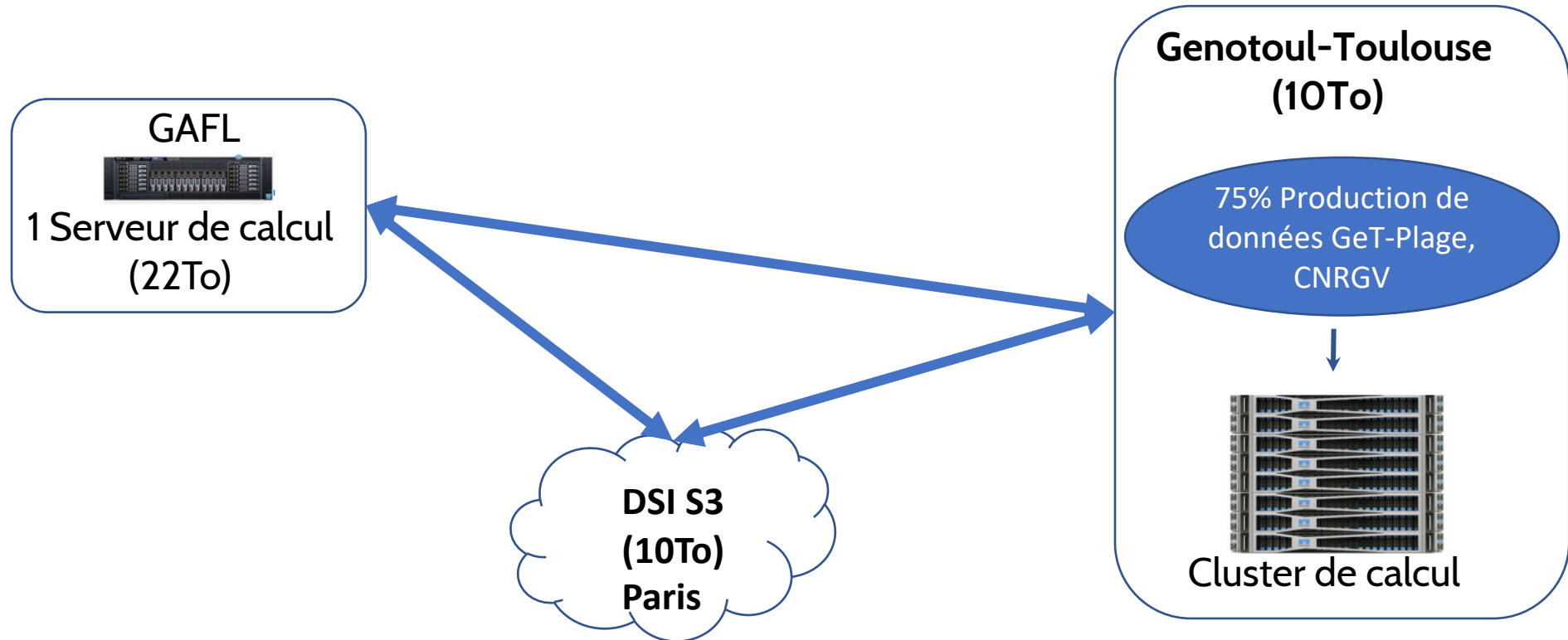
Génotypages et phénotypages



Gestion des données brutes du GAFL et leur réutilisation

Accessibilité, sécurisation des données et proximité données/calcul

6To données brutes NGS Tri-réplication



Bioinformatique/bio-analyse



Projet développé au GAFL pour remplir 3 critères :

- Reproductibilité des résultats
- 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

Bioinformatique/bio-analyse



gestionnaire de ressources



Environnement virtuel



Création de scripts d'analyse



Versionning



rapport d'analyse





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

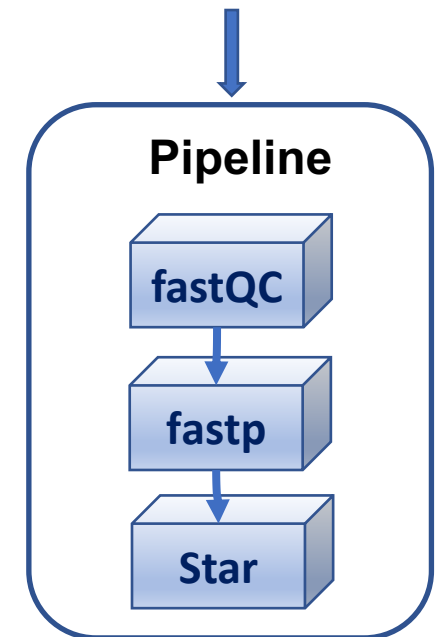
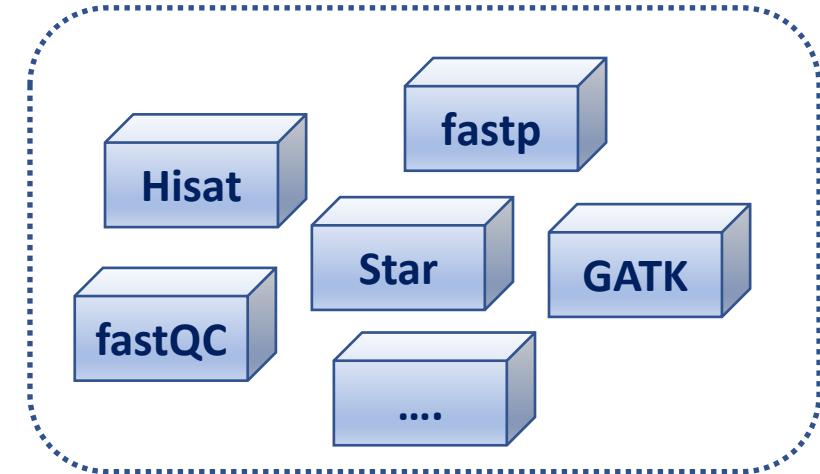
Bonnes pratiques

- **Création d'image à partir de recettes**
- **Utiliser 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**

Snakemake

- Langage basé sur Python
- 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

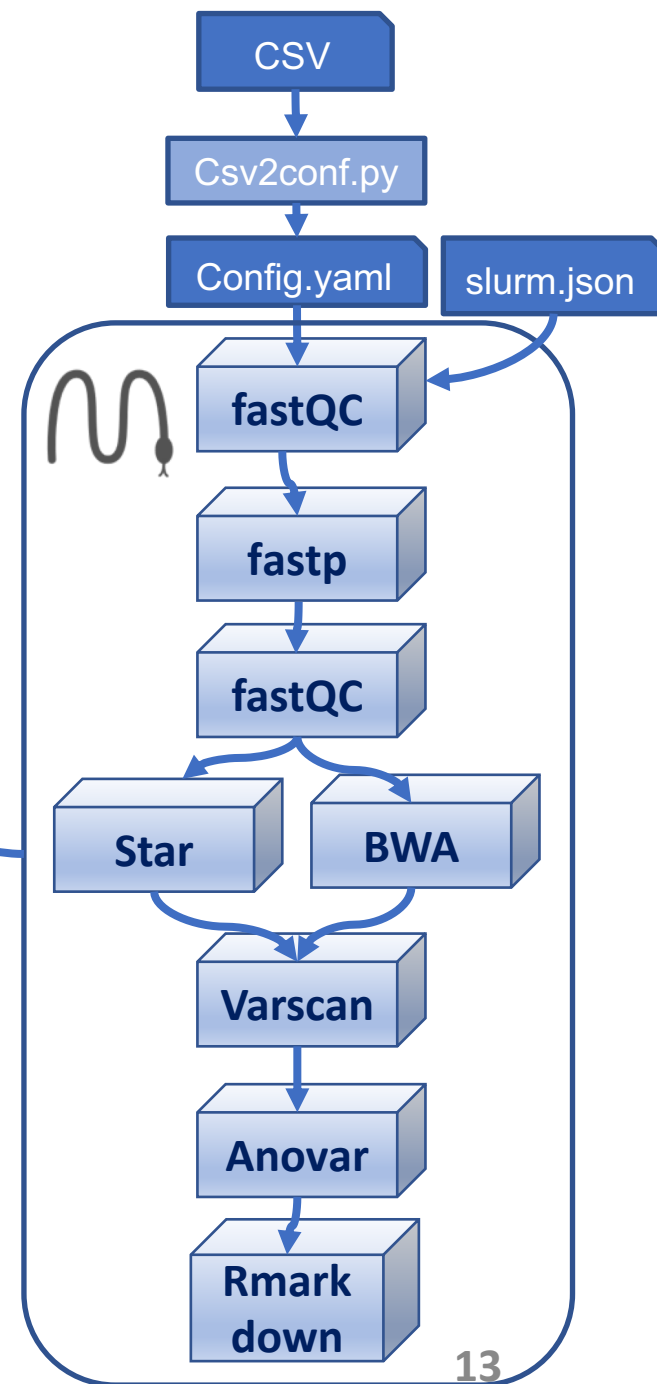
Modulaire →



AllMine, un pipeline flexible d'*Allele mining*

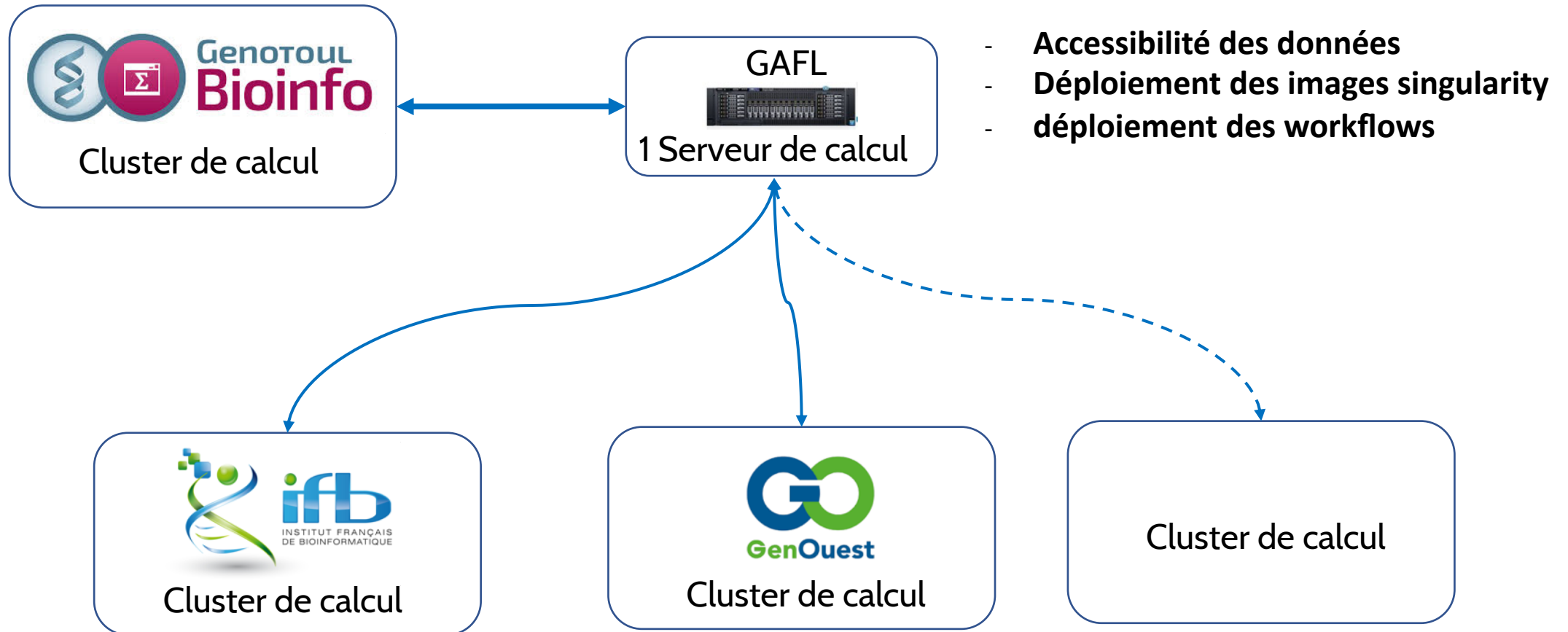
Thomas Bersez M2

- 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 **génomme de référence**.
- Génère un **fichier type tableur** listant les SNPs trouvés et leurs caractéristiques + un rapport d'analyse en **Rmarkdown**.



Déporter les analyses

condition requise: Singularity, Snakemake et SLURM



- **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**
- **Assuré les bonnes pratiques**



Animation bioinformatique

- **Organisation de workshops**
- **Création de tutoriels**



MERCI DE VOTRE ATTENTION!

Membres de l'équipe support:

BITTON Frédérique,

CHADOEUF Joël,

ELBELT Sonia,

LAGNEL Jacques,

LE-CALONNEC Emmanuel



Backup slides



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

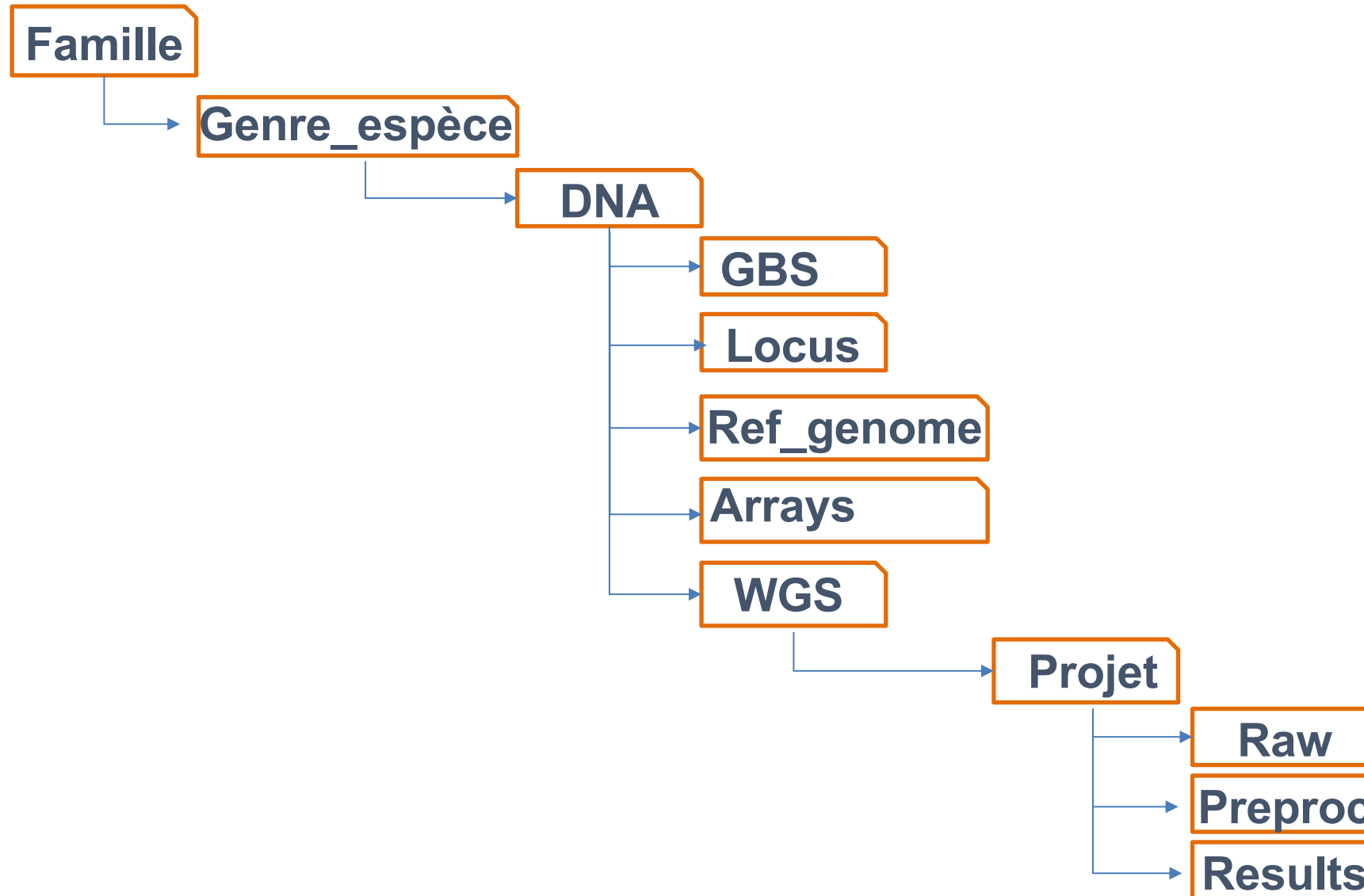
%post
apt-get -y update
apt -y install bwa

%runscript
bwa "$@"
```

Singularity:
problème de reproductibilité



Gestion des données brutes du GAFL et leur réutilisation.



Findable 
Accessible 
Interoperable 
Reusable 



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 dépôt 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, Minlon)
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

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)

DADI

Gestion génotypages
et phénotypages (BF)

I2B

CHADOEUF Joël (CJ), LAGNEL Jacques (LJ),
LE-CALONNEC Emmanuel (LCE)

Correspondants d'équipes:

DADI: BITTON Frédérique (BF)

ReDD: ELBELT Sonia (ES)

ReDD

Gestion données NGS et
développement d'une
base « résistance » (ES)

CRB

- Accompagnement SI de gestion
des ressources CRB (JL, ELC)
- Système embarqué terrain (LCE,LJ)