DMP du projet "IFB_Training_plant"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "INRAE - Trame générique projet v1" fourni par INRAE - Institut national de recherche pour l'agriculture l'alimentation et l'environnement.

Renseignements sur le plan

Titre du plan	DMP du projet "IFB_Training_plant"
Langue	fra
Date de création	2020-11-12
Date de dernière modification	2021-03-19
Identifiant	

Renseignements sur le projet

Titre du projet	IFB_Training_plant
Résumé	PGD exemple dédié à la formation IFB online 2021

Produits de recherche :

1. RNA Sequences (Jeu de données)
2. SNP and indel analysis (Jeu de données)
3. IFB Training PGD (Jeu de données)
4. Whole Genome Sequencing (Jeu de données)

Contributeurs

Nom	Affiliation	Rôles
Hélène Chiapello		• Coordinateur du projet • Personne contact pour les données (Variant_calling, IFB_training, RNAseq, WGS) • Responsable du plan

DMP du projet "IFB_Training_plant"

Informations sur le plan de gestion

Alain Souchon

SACEM

20/12/2020

V1

Informations sur le projet

Question sans réponse.

Question sans réponse.

BIDON

INRAE, France

CNRS, INSERM

UMS 3601

Question sans réponse.

Présentation générale des données du projet

IFB Training PGD

Question sans réponse.

Whole Genome Sequencing

Mode d'obtention : séquençage Illumina
Origine : vitis vinifera (la vigne)
Type : données de séquences au format fastq

RNA Sequences

Mode d'obtention : séquençage Illumina
Origine : vitis vinifera (la vigne)
Type : données de transcrits au format fastq

SNP and indel analysis

Mode d'obtention : sortie d'un workflow de variant calling
Origine : vitis vinifera (la vigne)
Type : variants au format VCF

Droits de propriété intellectuelle

par défaut INRAE sinon voir convention

Confidentialité

néant

Partage des données à l'issue du projet

IFB Training PGD

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Whole Genome Sequencing

Si financement ANR ou INRAE, par défaut partage obligatoire

Peut être important si un des génomes est issu d'un partenariat avec le privé ou le séquençage est cofinancé par le privé

Analyses bioinformatiques variées : variations structurales, génomique comparée (pan génomique),....

Non pas d'outil spécifique nécessaire en théorie (format texte) mais Oui car inexploitable en pratique sans outil bioinformatique dédié à ce type de données.

Le données seront rendues publiques au moment de la publication via une soumission à une banque de données internationale type ENA (EBI) ou NCBI.

To do : vérifier la licence sur les données quand on soumet à l'ENA (EBI)

A partir du moment où la publication est sortie

Si publication à l'EBI ou au NCBI : accessibilité long terme sans limite de durée à ce jour
Si DOI : minimum 10 ans

A l'EBI et au NCBI : oui via des numéros d'accession
Dans un entrepôt : oui via un DOI

RNA Sequences

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

SNP and indel analysis

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Description et organisation des données

IFB Training PGD

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Whole Genome Sequencing

- Format : fastq, outil : séquenceur Illumina HiSeq 3500
- Format : fasta, outil : assembleur SOAP de novo
- Format : gff3, outil : Eugene

Inclure un schéma du workflow (en fournissant le lien)

Voir par exemple : https://fairsharing.org/FAIRsharing.q9nh66
ENA Checklists spécialisée plantes : https://www.ebi.ac.uk/ena/browser/view/ERC000037

- Chaque entrepôt (ENA, NCBI ,...)proposes ses propres outils proposés pour générer des métadonnées et préparer la soumission mais c'est généralement utilisé à la fin du projet au moment de la soumission
- Actuellement pas d'outil satisfaisant de gestion des métadonnées en génomique
- Suite ISAtools https://isa-tools.org/software-suite/ : trop complexe à utiliser
- Eudat en cours de test à l'IFB

---

cf. Bonnes pratiques de nommage et de gestion des fichiers
- réfléchir en amont du projet au nommage des noms des échantillons et aux règles documentaires pour l'organisation des fichiers de données
- ex: données brutes en lecture seule dans un répertoire dédié partagé, nom de l'échantillon inclu dans le nom de fichier, suppression des fichiers intermédiaires non essentiels et faciles à régénérer, etc.

Exemple
- checksum/md5 pour contrôler l'intégrité des fichiers de données (received=sent)
- ?

RNA Sequences

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

SNP and indel analysis

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Question sans réponse.

Stockage et sécurité des données

Préciser l'infrastructure et la plateforme bioinformatique utilisée (IFB, Southgreen, migale,...)

-> lien avec le workflow bioinfo
1. production par une PF de séquençage
2. analyse sur une PF bioinfo
3. soumission à une banque internationale ou à un entrepôt

Distinguer
- Données brutes (seront soumises à un entrepôt international) : exemple = fichiers fastq
- Données temporaires : données intermédiaires faciles à régénérer qui n'ont pas vocation à être gardées
- Données finales (données issues d'une analyse qui seront préservées dans le temps) : exemple = assemblage, annotation

En général: plusieurs Go par run et plutôt 1 To au minimum de données pour un projet type de génomique

En général les plateformes bioinfo sont hébergées dans un datacenter ou au moins une salle machine

---

En France métropolitaine

---

2 types de protection
- accès physique aux serveurs
- accès informatique aux serveurs

- En général pendant le projet les données sont partagées en interne aux partenaires du projet
- A la fin du projet, ouverture à la communauté via soumission une banque publique

---

En générale sur les plateformes bioinfos
- espace projet dédié pour un groupe (à Southgreen, avec l'outil my à l'IFB,...) sur les serveurs
- contrôle d'accès par les comptes utilisateurs d'accès à l'infra (identique pour tous les membres du projet)

---

2 choses
- Accès aux données en cours d'analyse sur le cluster via des comptés dédiés
- Accès aux données une fois qu'elles sont intégrées dans des ressources dédiées (ex: GenomeHub à Southgreen). Peut permettre une gestion plus fine des accès

---

En bioinformatique il existe des formations au niveau de chaque plateforme pour permettre une utilisation correcte des ressources informatiques et pour promouvoir les bonnes pratiques de la science ouverte et reproductible. Ex: notebooks, GitHub ou gitlab pour l'archivage et le versioning du code et des notebooks

---

Archivage et conservation des données après la fin du projet

- Données à conserver = données brutes + données élaborées répondant à une question scientifique
- Données à détruire = résultats intermédiaires d'analyse

En biologie ce sont jusqu'à maintenant les banques internationales (ENA, ArrayExpress,...) qui assurent l'archivage pérenne des données
Pour les données de phénotypages : quels sont les usages ? Plutôt accès via des bases de données dédiées ? Pérennité des données peut être variable

---

- banques internationales : illimitée
- autres ressources : variable, dépend de la ressource

Alain Souchon (PI)
Laurent Voulzy (pour la bioinfo)
Johnny Halliday (avant sont décès en tant que DU)

Réponse à la fin du projet :-)

- Le stockage à long terme est payant (environ 30 €/an/To) mais dépend du type d'archivage (bande,...)
- Il faut intégrer ce coût aux budgets des projets ou instruire la question de qui paye à une échelle