Plan de gestion de données CO-LOcATION V1

Agnès Bonnet, Laure Gress, Sarah Maman-Haddad, Nathalie Vialaneix

CO-LOCATION

Plan de gestion de données créé à l'aide de DMP OPIDoR

Créateur du PGD : Agnès Bonnet

Affiliation du créateur principal : INRAE - Institut national de recherche pour l'agriculture l'alimentation et l'environnement

Modèle du PGD : INRAE - Trame générique projet

Dernière modification du PGD : 23/04/2021

Financeur : ANR

Numéro de subvention : ANR-20-CE20-0020

Résumé du projet :
Chez le porcelet, la mortalité néonatale est un problème économique, éthique et entre dans les questions liées au bien-être animal. Une part de cette mortalité est due à un défaut de maturité des nouveau-nés à la naissance. Cette maturité est le résultat du développement en fin de gestation et dépend, entre autres, des interactions fœto-maternelles qui régulent la répartition des ressources entre la mère et le fœtus, impactant le développement fœtal et la santé du nouveau-né. Pour aborder cette question, le projet CO-LOcATION propose l'étude des interactions entre le placenta (tissu maternel) et l'endomètre (tissu fœtal) avec une approche intégrative combinant différents niveaux d'information (transcriptome, métabolome, lipidome, phénotypes de maturité). Il utilisera les échantillons provenant du projet ANR PORCINET, qui a généré des fœtus de génotypes purs et des croisés réciproques des races Large white et Meishan caractérisées par une robustesse contrastée à la naissance. Les objectifs sont l'identification de gènes et des mécanismes biologiques participant au dialogue entre placenta et endomètre, la caractérisation de la contribution des génomes maternel et paternel sur le métabolisme placentaire et des génomes fœtaux sur le métabolisme de l'endomètre. L'ensemble des évènements moléculaires identifiés permettront de poursuivre nos travaux sur la mise en place de la maturité en fin de gestation et la survie à la naissance. En particulier, les résultats seront discutés avec des experts et des professionnels de la nutrition de la truie en gestation et de la sélection en génétique porcine.

Chercheur Principal : Agnès Bonnet

Identifiant ORCID : 0000-0001-7702-4025

Contact pour les Données : Agnès Bonnet

Produits de recherche :

1. Echantillons : sélection des fœtus (Collection)
2. RNAseq : séquençage Nouvelle Génération d'ARN (Jeu de données)
3. Métabolome : métabolome RMN H+ (Jeu de données)
4. Lipidomique : lipides neutres GC/FID (Jeu de données)
5. Statistique : analyses statistiques (Workflow)

CO-LOCATION

1- ECHANTILLONS : SELECTION DES FOETUS

A- INFORMATIONS SUR LE PLAN DE GESTION

Date de création du PGD
11/12/2020

Version en cours
V1

Date de la dernière version
23/04/2021.

B- INFORMATIONS SUR LE PROJET

Identifiant de l'appel à projet (call for proposal)
AAPG 2020

Financeur(s) du projet
AAPG JCJC

Nom du programme de recherche
AAPG JCJC CE20 "Biologie des animaux, des organismes photosynthétiques et des microorganismes"

Référence de la convention de financement
ANR-20-CE20-0020

Acronyme du projet
COLOcATION

Nom du projet de recherche
Maturité fœtale à l'interface fœto-maternelle : contribution des génomes fœtal et maternel et perturbations des métabolismes tissulaires

Institution leader du projet, coordinateur bénéficiaire (nom, pays)
Institut National de la Recherche pour l'Agriculture, l'alimentation et l'Environnement, France

Unité de rattachement du responsable du projet
GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France

Dates et durée du projet
01/01/2021 au 31/12/2023 durée 3 ans

C- PRESENTATION GENERALE DES DONNEES DU PROJET

Description de la banque de tissus préexistante:
Nous allons utiliser les échantillons collectés dans le projet PORCINET (ANR-09-GENM-005). Ce projet avait pour objectif d'étudier l'acquisition de la maturité fœtale chez 4 génotypes de porcelets: les purs Large White (LW - lignée sélectionnée de façon intense) qui se caractérisent par une mortalité néonatale importante, les purs Meishan (MS - lignée rustique) qui ont une forte vitalité à la naissance et sont très peu sujets à la mortalité, ainsi que les croisés réciproques LWxMS et MSxLW. L'expérimentation sur des animaux vivants et les procédures mises en œuvre au sein du projet PORCINET ont été accréditées par le comité d'éthique de Midi-Pyrénées (MP/01/01/01/11). Une très importante banque de tissus a été générée et une base de données regroupant des données phénotypiques et physiologiques, des profils métabolomiques (urine, plasma et liquide amniotique; DOI: 10.1038/s41598-020-76709-8) , protéomiques et transcriptomiques sur 6 tissus (DOI: 10.1074/mcp.M116.066357, DOI: 10.1186/s40104-018-0244-2, DOI: 10.1038/s41598-020-76709-8 , DOI: 10.1186/s12864-017-4001-2, DOI: 10.1186/1471-2164-15-797) a été construite. Les échantillons ont été collectés aux deux temps de gestation qui bornent la maturation fœtale (90 et 110 jours de gestation, la durée de gestation étant de 114j) et sont conservés depuis leur obtention en 2010 dans un congélateur à -80°C sous forme de poudre ou de morceaux. La traçabilité et la gestion de ces échantillons sont assurées par le système BARCODE (interface web + base de données localisée à la plateforme GenoToul Bioinfo: E-SITOul barcode).
Nous disposons des 2 tissus adjacents acteurs des interactions mère-fœtus, le placenta (ou chorion, tissu fœtal) et l'endomètre (tissu maternel) en contact direct pour 404 fœtus afin d'étudier le dialogue fœto-maternel en lien avec la maturité fœtale.

D- DROITS DE PROPRIETE INTELLECTUELLE

Qui détiendra les droits sur les données et les autres informations créées lors du projet ?
Le détenteur de la propriété intellectuelle des données est INRAE dont relève le porteur de projet.

Du matériel protégé par des droits spécifiques sera t-il utilisé au cours du projet ? Dans ce cas, qui s'occupe des formalités à accomplir, obtient les autorisations d'utilisation et de diffusion éventuelle …
Non

E- CONFIDENTIALITE

Identification des jeux de données confidentielles
Les jeux de données bruts et traités qui seront acquis ne sont pas confidentiels. Ils resteront privés au sein du consortium jusqu'à publication des résultats dans le cadre du projet ou de nouvelles méthodes statistiques pour le projet, dans des revues à comité de lecture. La décision de publication des données sera prise par le consortium.

Quelles sont les mesures prises et les normes auxquelles il est nécessaire de se conformer pour garantir cette confidentialité ?
Question sans réponse.

Le cas échéant, comment la confidentialité de données fournies par des personnes sera garantie lorsque les données seront partagées ou rendues disponibles pour une analyse de second niveau ?
Question sans réponse.

F- PARTAGE DES DONNEES A L'ISSUE DU PROJET

Y a t'il une obligation de partage (ou à l'inverse une interdiction ou une restriction)
L'ANR, en tant que financeur, demande de privilégier un accès libre aux données de la recherche.

Quelles données seront partagées à l'issue du projet ? Si toutes les données ne sont pas disponibles de la même façon, ou en même temps, le préciser
Toutes les données brutes et traitées seront partagées après publication dans des revues à comité de lecture (~2024).
Une période d'exclusivité d'accès aux données réservée au consortium est nécessaire pour la valorisation scientifique des données.

Données brutes disponibles :	format brut	Format fichiers traités
sélection des fœtus	csv	-
RNAseq	fq	bam; gtf; csv
métabolome	fid / ser	csv
lipidome	cmbx /csv	csv
statistiques	fichiers traités des autres produits de recherche	jeux de données normalisés format csv, listes de gènes différentiels format csv Réseaux de gènes R markdown et notebook associés, scripts R

Quelles sont les réutilisations potentielles de ces données ?
Les données pourront être réutilisées pour des études ciblées ou de l'intégration de données hétérogènes.

La lecture des données nécessite-t-elle le recours à un logiciel ou un outil spécifique ? Si oui, lequel et comment y accéder ?
Le tableau suivant décrit le type de fichier brut pour chaque produit de recherche ainsi que les formats des données.

selection des foetus	csv	format tableur
RNAseq	fastq	format texte
Metabolome	fid / ser	format propriétaire Bruker : logiciel Topspin gratuit
lipidome	csv	format tableur
statistique	R markdown, csv	format texte et tableur

Comment les données seront-elles partagées ?
Les données brutes seront déposées dans des entrepôts dédiés dès la fin de l'acquisition et seront rendues accessibles après une période d'exclusivité d'accès réservée au consortium permettant la publication des résultats et la valorisation. Les jeux de données traités, les tables de comptage (RNAseq), de quantification (métabolome, lipidome) seront déposées dans l'entrepôt Data INRAE et décrites dans des Data papers associés aux publications.

produits de recherche	types de données	entrepôt
sélection des fœtus	csv	Biosamples (EBI)
RNAseq	fq.gz	ArrayExpress (EBI)
metabolome	fid /ser	MetaboLight (EBI)
lipidome	csv	MetaboLight (EBI)
statistique	R markdown, csv	Forge MIA(GitLab), Dataverse INRAE

Avec qui ? sous quelle licence ?
Les jeux de données brutes seront diffusés en libre accès sous licence CC BY SA 4.0 (creative commons) après publications.
Les jeux de données finaux (comptages, quantifications) seront diffusés en libre accès sous licence ouverte etalab-2.0 après publications.

A partir de quand ?
Les données brutes et finales seront disponibles après publications des résultats sur les analyses prévues dans le projet (~2024).

Pendant combien de temps ?
A ce jour, la durée de stockage des données dans les entrepôts cités est illimitée.

Les données seront-elles identifiées par un identifiant pérenne (DOI ou autre) ?
Elles seront identifiées par un numéro d'accession pérenne lié à l'entrepôt.

produits de recherche	entrepôt	code/lien d'accession
échantillons biologiques	BioSamples (EBI)	numéro d'accession de BioSamples
RNAseq	ArrayExpress (EBI)	E-XXXX-n pour les expériences ENA accession pour les séquences
métabolome	MetaboLight (EBI)	MTBLS-n
lipidome	MetaboLight (EBI)	MTBLS-n
statistique	ForgeMIA(GitLab), Dataverse INRAE	Forge MIA : pas d'identifiant pérenne. Data INRAe : DOI

Quel est l'organisme qui se chargera de la demande d'identifiant dans le cas de projets multi-partenaires ?
Question sans réponse.

G- DESCRIPTION ET ORGANISATION DES DONNEES

Quels méthodes et outils sont utilisés pour acquérir et traiter les données ? Précisez les différents formats dans lesquels les données seront disponibles aux différentes phases de la recherche

Sélection des fœtus du projet:
Les fœtus ont été sélectionnés selon deux modes pour répondre aux objectifs du projet:
i. contribution des génomes : Pour cela il est nécessaire d'avoir un maximum de fœtus par portée et des portées avec une bonne répartition des génotypes.
ii. l'étude de la maturité: la sélection des classes de maturité (matures, peu mature ou moyen) s'est effectuée par une analyse en composante principale sous environnement R en utilisant les coordonnées des axes (format csv) (PCA; packages factominer et factoextra implémentés dans Bioconductor).
Les scripts de l'analyse statistique ont été sauvegardés en R Markdown HTML et sont stockés ainsi que les fichiers finaux (Fichiers plats ;csv) dans le SharePoint "COLOcATION" dédié au projet.

Documentation associée aux données

Procédure de sélection sous format PowerPoint.
Scripts de l'analyse statistique sous format R Markdown.
Fichiers finaux (Fichiers plats .csv)

Quels types de métadonnées seront produites pour accompagner les données ? Quels sont les standards et les vocabulaires ou taxonomies qui seront utilisés pour décrire les données ?
Métadonnées qui décrivent l'origine des échantillons en format csv .
Utilisation des standards FAANG (EBI; BioSamples)

Animal	identification GEMA (Foetus+numéro)
Mère	Numero de l'animal
Père	Numero de l'animal
Sexe	M/F
dg	temps de gestation (90/110)
TG	type génétique (L(Large White pur),MS (Meishan pur), LM (de père LW et de mère MS), ML(de père MS et de mère LW))
Maturité	stade de maturité foetale (M- (Maturité faible), N (maturité normale), M+ (maturité élevée))
TGmere	type génétique de la mère (MS,LW)
TGpere	ype génétique du père (MS,LW)

Comment les métadonnées seront elles produites ?
Le fichier de métadonnées n'est pas produit par un outil mais utilise un standard FAANG.

Comment les fichiers de données sont-ils gérés et organisés au cours du projet : contrôle des versions, conventions de nommage des fichiers, organisation des fichiers…
Le contrôle des versions s'effectue par addition de la date de création et de modification.

Quelle est la procédure de contrôle qualité des données ? joindre éventuellement le plan d'assurance qualité
La sélection des fœtus a été validée par une analyse statistique multivariée (ACP).

H- STOCKAGE ET SECURITE DES DONNEES

Stockage : Quels seront les supports utilisés pour les données au cours du projet ?

- Partage interne au projet :
  - Espace projet du cluster GenoToul Bioinfo
  - Lien vers le wiki de la Forge DGA
- Partage externe :
  - Les échantillons seront déposés dans l'entrepôt BioSamples EBI (et reversés sur FAANG) (responsable du dépôt : Agnès Bonnet) avec un embargo jusqu'à publication d'au moins un article scientifique sur ces données. Les identifiants nationaux des parents seront anonymisés.
- Analyses
  - Espace projet du cluster GenoToul Bioinfo
  - Disque réseau P

Stockage : Quels seront les types de flux empruntés par les données au cours du projet ?

Les données transiteront par flux https.
Voir schéma fonctionnel produit de recherche 2.

Stockage : Quelle est la volumétrie prévisionnelle ?
~1Mo

Stockage : Où sont hébergées physiquement les données, sur quel type d'hébergement ?
Les données sont hébergées sur l'arche de données INRAE Toulouse- Occitanie.

Stockage : Où sont localisées géographiquement les données ?
Auzeville, centre INRAE Toulouse-Occitanie

Sécurité : L'entité hébergeant physiquement les données a-t-elle une politique de sécurité pour son système d'information ?
La charte de la plateforme BioInfo Genotoul: http://bioinfo.genotoul.fr/index.php/news/new-genotoul-bioinfo-charter/?highlight=charte définit les conditions d'accès au serveur et les règles d'utilisation et est subordonnée à la charte informatique de l'INRAe. Elle applique la politique Sécurité des Systèmes d'Information de l'état (PSSIE).

Sécurité - Confidentialité : les données feront-elles l'objet d'échange ou de partage avec de tiers acteurs ?
Les données concernant l'origine des échantillons seront disponibles pour l'ensemble des personnes impliquées dans les analyses dans l'espace projet du cluster GenoToul Bioinfo et, par un lien, dans l'espace projet sur le wiki de la Forge DGA (https://forge-dga.jouy.inra.fr/).

Sécurité - Confidentialité : comment sont déterminés les droits d'accès aux données pendant les recherches ?
Les accès au sharepoint collaboratif et au wiki de la Forge DGA ainsi que la Forge MIA seront restreints aux personnels INRAE par le login et le mot de passe institutionnel et sur demande d'autorisation d'accès au responsable du projet. Les accès à l'espace projet du cluster Genotoul Bioinfo est restreinte aux membres du projet via leur login.

Sécurité - Confidentialité : De quelle manière l'ensemble des chercheurs partenaires du projet auront-ils accès aux données pendant la recherche ?
Les accès aux données et aux différents espaces de travail et de stockage se feront par un login et mot de passe après autorisation du responsable du projet.
Espace sharepoint COLOCATION
Espace projet sur le cluster de la plateforme GenoToul Bioinfo
Wiki sur la Forge DGA
Git sur la Forge MIA

Sécurité - Intégrité - Tracabilité : Quelles sont les mesures de protection mises en œuvre pour suivre la production et l'analyse des données au cours du projet ?
Cahiers de laboratoire, R markdown.
La traçabilité et la gestion de ces échantillons sont assurées par le système BARCODE (interface web + base de données localisé à la plateforme GenoToul Bioinfo : E-SIToul barcode). Un fichier readme par dossier de l'espace projet sur le serveur GenoToul permettra de tracer les versions.

I- ARCHIVAGE ET CONSERVATION DES DONNEES APRES LA FIN DU PROJET

Quelles sont les données à conserver sur le moyen ou le long terme et quelles sont les données à détruire ?
Les données conservées sur le long terme seront les métadonnées décrivant les échantillons et les protocoles, telles que déposées dans l'entrepôt BioSamples (EBI).

Sur quelle plateforme d'archivage pérenne seront archivées les données à conserver sur le long terme ? Sinon, quelles procédures seront mises en place pour la conservation à long terme ?
Plateforme d'archivage de l'EBI (BioSample)s + référencement sur le portail données de FAANG.

Quelle est la durée de conservation des données ?
Les données seront conservées de manière pérennes aussi longtemps que les entrepôts de données de l'EBI existeront.

Qui sera responsable de la conservation à long terme ? nommer un contact individuel.
Le coordinateur du projet est responsable de la gestion des données durant le projet et de leur archivage à la fin de celui-ci. Le directeur de l'unité ayant coordonné le projet peut être responsable sur le long terme.

Le responsable de l'archivage à long terme est Laurence Liaubet (INRAE GenPhySE) ou, en son absence, le directeur/trice de l'Unité GenPhySE.

Quel sera le volume de ces données ?
~1Mo

Quelles garanties de financements couvriront les coûts associés à la conservation à long terme ?
Le stockage dans les entrepôts de données de l'EBI est gratuit.

2- RNASEQ : SEQUENÇAGE NOUVELLE GENERATION D'ARN

A- INFORMATIONS SUR LE PLAN DE GESTION

Voir produit de recherche 1

B- INFORMATIONS SUR LE PROJET

Voir produit de recherche 1.

C- PRESENTATION GENERALE DES DONNEES DU PROJET

Brève présentation des données générées, collectées ou réutilisées :
Les ARN seront extraits avec le kit Macherey-Nagel nucleospin RNA (ref 740955.50) à partir de 80 mg de poudre par échantillon. Les 448 échantillons d'ARN messagers seront ensuite séquencés en paired end (longueur attendue des lectures 2x150pb) sur le novaseq 6000 (flow cell S4) par la plateforme GeT-PlaGe (Plateforme ISO9001 et NFX50-900, relabellisée IBiSA).
Le séquençage sera effectué à raison de 50 échantillons par ligne afin d'atteindre une profondeur de 40 millions de lectures minimum par échantillon, ce qui représente 0.8 TO d'espace par ligne et 9 lignes de séquençage pour la totalité du projet.

D- DROITS DE PROPRIETE INTELLECTUELLE

Voir produit de recherche 1.

E- CONFIDENTIALITE

Voir produit de recherche 1.

F- PARTAGE DES DONNEES A L'ISSUE DU PROJET

Voir produit de recherche 1.

G- DESCRIPTION ET ORGANISATION DES DONNEES

Quels méthodes et outils sont utilisés pour acquérir et traiter les données ? Précisez les différents formats dans lesquels les données seront disponibles aux différentes phases de la recherche
Les données brutes sont obtenues sur le novaseq6000 flowcell S4: format de sortie bcl puis fastq.
Les données issues du séquençage seront traitées en ligne de commande selon le pipeline Nextflow RNAseq suivant : https://github.com/nf-core/rnaseq (doi: 10.5281/zenodo.1400710). Le pipeline sera validé en terme d'outils et de paramètres, sur un nombre limité d'échantillons, l'ensemble des échantillons pourront être traités ensuite.
Principales étapes du pipeline bio-informatique:

Nettoyage des adaptateurs: outil Trim Galore: format fastq

Alignement sur génome de référence (version la plus récente) : outil STAR: formats bam/sam et quantification de l'expression des gènes : outil RSEM: formats gff/gtf

Reconstruction des transcrits connus, nouveaux et potentiel codant: outil StringTie (formats bam/sam ou gff/gtf) puis quantification des transcripts avec RSEM.

Prédiction des ARN long non codants : outil FEELnc puis quantification des transcripts non codants avec RSEM.

Les fichiers archivés seront :

- Les fichiers de séquences fastq sous forme compressées comme données brutes de départ.
- Les fichiers bam issus de l'alignement.
- Les fichiers gtf issus de StringTie.
- Les matrices de comptages (tables csv)

Documentation associée aux données
Origine des échantillons
Contrôle qualité d'extraction des ARN totaux: format csv
Protocole d'extraction des ARN: fomat texte
Traçabilité des échantillons format csv
Contrôle qualité des ARNs au fragment analyser: format texte et csv
Protocole des librairies
Contrôle qualité du séquençage : format texte
Traçabilité des traitements bioinformatiques , des versions et des options mise à jour sur la forge DGA (https://forge-dga.jouy.inra.fr/) pour s'assurer d'un travail collaboratif (https://forge-dga.jouy.inra.fr/projects/colocation/wiki).

Quels types de métadonnées seront produites pour accompagner les données ? Quels sont les standards et les vocabulaires ou taxonomies qui seront utilisés pour décrire les données ?
1- Métadonnées associées à l'origine de l'échantillon: format csv (fichier samples FAANG data base (EBI))
2- Métadonnées associées à l'extraction des ARN totaux: format csv
3- Métadonnées associées à la génération des données brutes : format csv ((selon le guide de soumission FAANG data base pour soumission dans EBI (https://data.faang.org/home)
4- Métadonnées de contrôle qualité des séquences accompagneront les données brutes (Fastqc report et contaminants), format texte.
3- Métadonnées de traitements bio-informatiques issus des fichiers de logs des outils bio informatiques.
Utilisation des standards FAANG

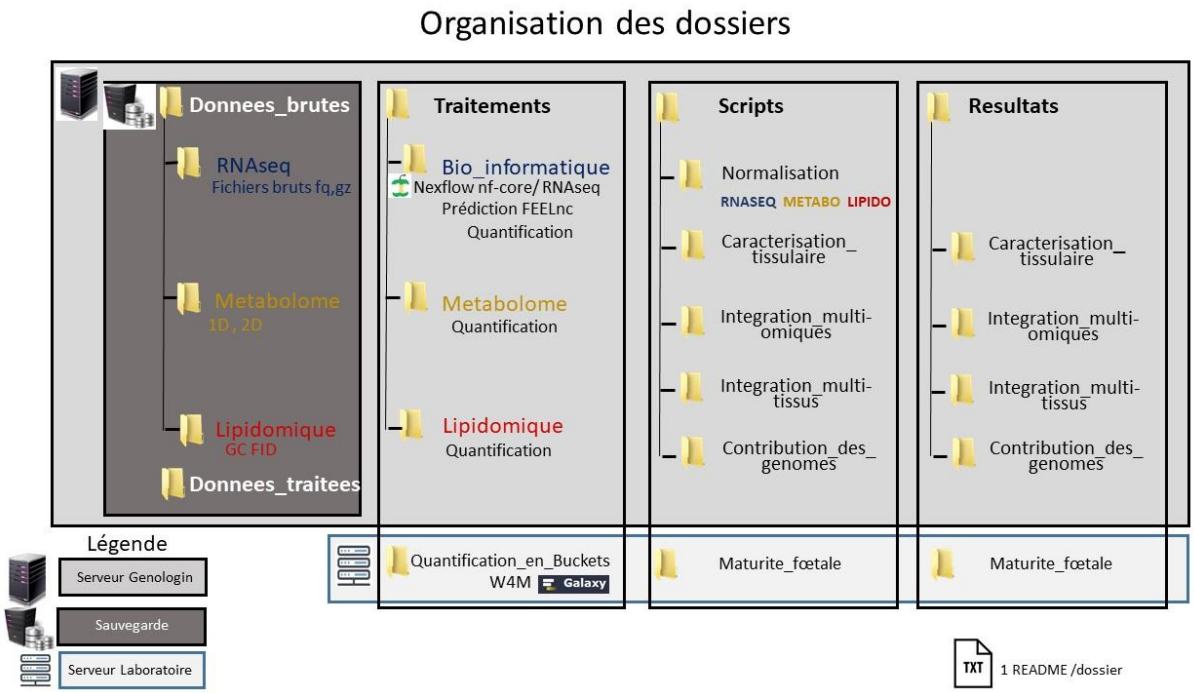
Comment les métadonnées seront elles produites ?
Les métadonnées associées aux données brutes seront produites en utilisant les standards FAANG (EBI).
Les métadonnées bio-informatiques seront fournies par les outils utilisés.

Comment les fichiers de données sont-ils gérés et organisés au cours du projet : contrôle des versions, conventions de nommage des fichiers, organisation des fichiers…
Les fichiers bruts seront nommés selon le code-barres spécifique à l'animal, à la nature de la donnée (transcriptome, métabolome, lipidome) et au type de tissu (endomètre/placenta).

Le pipeline Nextflow permet de gérer la traçabilité des outils et les options utilisées lors du traitement bio-informatique.
Les différents fichiers de traitement seront nommés suivant la convention "code barre échantillon"_""numéro de version" et classé dans des dossiers correspondants au niveau dans la chaine de traitement: A : nettoyage et contrôle qualité, B: alignement STAR ;C : reconstruction des transcrits ; D: comptages.

- numéro de version: V0,1,2,3,... pour les modifications importantes et V1.1... pour les modifications minimes.
La version finale sera notée _F.

Les fichiers bruts de séquençage seront stockés dans la partition save de l'espace projet ouvert sur l'arche des données Toulouse-Occitanie. Les traitements seront effectués dans la partition work de l'espace projet
Les fichiers bruts seront accessibles à partir d'un lien symbolique.
Le suivi et la traçabilité des traitements bio-informatiques se feront via forge DGA (https://forge-dga.jouy.inra.fr/).
Démarche qualité interne,
Organisation des dossiers: voir graphique
Chaque dossier contiendra un readme.txt décrivant les fichiers.



Quelle est la procédure de contrôle qualité des données ? joindre éventuellement le plan d'assurance qualité

Le contrôle qualité des données s'effectue à plusieurs étapes :

- Analyse de la qualité des ARN totaux (nanodrop et Fragment analyser)
- Analyse de la qualité des séquences avec l'outil FASTQC Report.
- Analyse des contaminants (E. coli, phi, yeast)
- Analyse des logs des outils bio-informatiques à l'aide de MultiQC.
- Statistiques de base sur les alignements (flagstats et idxstats)
- Graphes de profondeurs, couvertures et modèles de gènes

la plateforme de séquençage est certifiée ISO9001 et NFX50-900 et relabellisée IBiSA en 2008.

H- STOCKAGE ET SECURITE DES DONNEES

Stockage : Quels seront les supports utilisés pour les données au cours du projet ?

Les données brutes seront sauvegardées dans NG6 GENOTOUL pendant 3 mois puis dans l'espace de sauvegarde du projet de la plateforme Genotoul Bioinfo, labellisée IBISA en lecture seulement. Les données de comptages issues du traitement bio-informatique seront aussi positionnées dans l'espace collaboratif d'archivage du serveur.

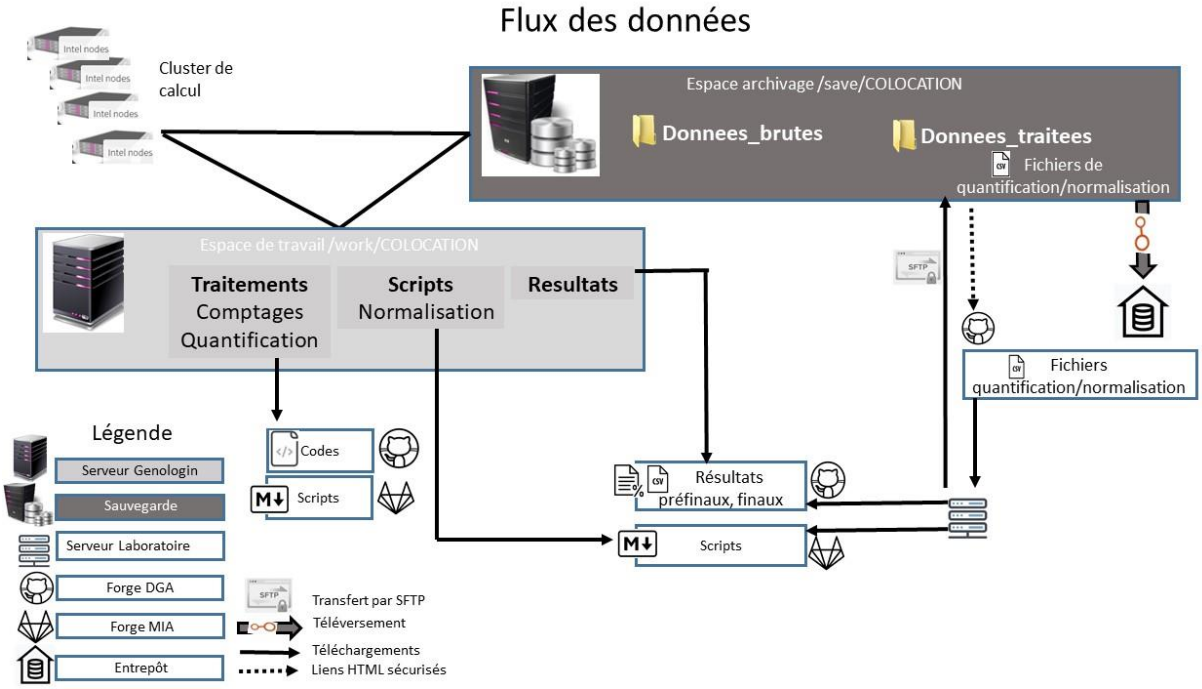
Les données brutes seront déposées dans l'entrepôt ENA (EBI) et référencés dans FAANG à la fin de l'acquisition (responsables du dépôt : Nathalie Vialaneix, Agnès Bonnet) avec un embargo jusqu'à publication d'au moins un article scientifique sur ces données.

Les données de comptage seront déposées dans l'entrepôt Dataverse INRAE en fin de traitement avec un embargo jusqu'à la fin de la valorisation des données prévue dans le projet (responsables du dépôt : Nathalie Vialaneix, Agnès Bonnet).

Stockage : Quels seront les types de flux empruntés par les données au cours du projet ?

Un espace collaboratif COLOCATION sera ouvert sur le cluster GenoToul Bioinfo pour la durée du projet.

Stockage temporaire des données brutes FASTQ dans NG6 puis dans l'espace collaboratif du projet COLOCATION.
En fin d'acquisition des jeux de données, les données seront soumises et sauvegardés dans les entrepôts correspondants de l'EBI et référencées sur le portail données de FAANG.
Il n'y a peu de flux de données (Voir graphique) : les données brutes sont accessibles seulement en lecture via des liens symboliques sur l'espace collaboratif du projet de la plateforme GenoToul Bioinfo. Les traitements bio-informatiques sont effectués sur l'espace de calcul en utilisant les nœuds de calculs avec une interface infiniband. Il y a deux générations de nœuds différentes. Le descriptif des nœuds et du work sont disponibles sur le site BioInfo Genotoul.
A la fin des traitements bio-informatiques et statistiques, les résultats des analyses sont mis à disposition des partenaires dans la forge DGA.
Les jeux de données traités seront déposés de manière pérenne sur un entrepôt de l'EBI.



Stockage : Quelle est la volumétrie prévisionnelle ?
Gestion des ressources informatiques nécessaires prévisionnelles:

type de ressource	stockage	Calcul
année 1	27 To	36 To
année 2	27 To	18 To
année 3	14 To	9 To
année 4	14 To	

Stockage : Où sont hébergées physiquement les données, sur quel type d'hébergement ?
Les données sont hébergées sur l'arche de données INRAE Toulouse- Occitanie.

Stockage : Où sont localisées géographiquement les données ?
Auzeville, centre INRAE Toulouse-Occitanie

Sécurité : L'entité hébergeant physiquement les données a-t-elle une politique de sécurité pour son système d'information ?
La charte de la plateforme BioInfo Genotoul: http://bioinfo.genotoul.fr/index.php/news/new-genotoul-bioinfo-charter/?highlight=charte définit les conditions d'accès au serveur et les règles d'utilisation et est subordonnée à la charte informatique de l'INRAe. Elle applique la politique Sécurité des Systèmes d'Information de l'état (PSSIE).

Sécurité - Confidentialité : les données feront-elles l'objet d'échange ou de partage avec de tiers acteurs ?
Les données brutes ne font pas l'objet d'échanges ou de partage avec des tiers acteurs pendant la durée du projet.
Les données brutes et les matrices de comptages seront disponibles pour l'ensemble des personnes impliquées dans les analyses dans l'espace projet du serveur de GenoToul et via un lien sur le wiki de la Forge DGA (https://forge-dga.jouy.inra.fr/).

Sécurité - Confidentialité : comment sont déterminés les droits d'accès aux données pendant les recherches ?
Les données de séquençage acquises par la plateforme Get-PlaGe seront déposées dans l'espace NG6 de l'équipe Genorobust pour une durée de 3 mois. Les droits d'accès de cet espace sont réservés aux membres de l'équipe. Les données brutes seront également transférées dans l'espace de sauvegarde du projet sur le cluster GenoToul Bioinfo. Les droits d'accès à cet espace de sauvegarde et de travail dédiés au projet sur le cluster GenoToul Bioinfo seront donnés par la plateforme GenoToul aux personnes assurant les traitements bio informatiques et statistiques. L'accès se fera via le login et un mot de passe utilisateur des participants sur le cluster.
Les accès au sharepoint collaboratif et au wiki de la Forge DGA seront restreint aux personnels INRAE par le login et le mot de passe institutionnels et sur demande d'autorisation d'accès au responsable du projet.

Sécurité - Confidentialité : De quelle manière l'ensemble des chercheurs partenaires du projet auront-ils accès aux données pendant la recherche ?

Méthode d'identification, d'authentification.
Les chercheurs partenaires assurant les analyses auront un accès, protégé par un login/mot de passe, à l'ensemble des données présentes brutes, intermédiaires et finales sur les espaces projet, le sharepoint et la Forge DGA.

Sécurité - Intégrité - Tracabilité : Quelles sont les mesures de protection mises en œuvre pour suivre la production et l'analyse des données au cours du projet ?
Les traçabilités des échantillons et des données générées par les séquenceurs sont assurées par e-SiToul.
La traçabilité des traitements bio informatiques sera effectuée sur la forge DGA (https://forge-dga.jouy.inra.fr/).

Le pipeline Nexflow permet d'assurer la reproductibilité des traitements.

I- ARCHIVAGE ET CONSERVATION DES DONNEES APRES LA FIN DU PROJET

Quelles sont les données à conserver sur le moyen ou le long terme et quelles sont les données à détruire ?

Les données brutes (FASTQ) ainsi que les données de comptages (csv) seront à conserver sur le moyen et long terme pour une ré-exploitation ultérieure. La conservation à long terme sera assurée par dépôt sur un entrepôt de l'EBI.
Les fichiers intermédiaires de traitements bio-informatiques (alignements; reconstruction des transcrits) ne seront conservés que pendant la durée du projet (3 ans).

Sur quelle plateforme d'archivage pérenne seront archivées les données à conserver sur le long terme ? Sinon, quelles procédures seront mises en place pour la conservation à long terme ?
Les données brutes (FASTQ) et les données de comptages seront archivées sur les entrepôts correspondants de l'EBI et référencées sur le portail données de FAANG.

Quelle est la durée de conservation des données ?
Les données seront conservées de manière pérennes aussi longtemps que les entrepôts de données de l'EBI existeront.

Qui sera responsable de la conservation à long terme ? nommer un contact individuel.
Le responsable de l'archivage à long terme est Laurence Liaubet (INRAE GenPhySE).

Quel sera le volume de ces données ?
Le volume des données des fichiers bruts (FASTQ) est de 9 To.

Quelles garanties de financements couvriront les coûts associés à la conservation à long terme ?
Le stockage dans les entrepôts de données de l'EBI est gratuit.

3- METABOLOME : METABOLOME RMN-^{1}H

A- INFORMATIONS SUR LE PLAN DE GESTION

Voir produit de recherche 1.

B- INFORMATIONS SUR LE PROJET

Voir produit de recherche 1.

C- PRESENTATION GENERALE DES DONNEES DU PROJET

Brève présentation des données générées, collectées ou réutilisées :
Chaque échantillon (~100 mg de poudre) sera extrait par une méthode adaptée de Bligh and Dyer (1959) afin de séparer la phase lipidique et aqueuse. Les données spectrales 1D de la fraction aqueuse seront acquises sur le spectromètre Bruker Avance III HD 600 MHz de la plateforme AXIOM (MeTatoul). Des spectres 2D seront acquis pour un pool d'extraits d'endomètre et un pool d'extraits de placenta afin de faciliter l'identification des métabolites.

Les données générées sont des données spectrales. Les spectres bruts seront stockés sur le serveur de l'unité et sur l'espace projet de la plateforme Genotoul Bioinfo puis entreposés dans les archives ouvertes de l'EBI (https://www.ebi.ac.uk/metabolights/) avant publication.

D- DROITS DE PROPRIETE INTELLECTUELLE

Voir produit de recherche 1.

E- CONFIDENTIALITE

Voir produit de recherche 1.

F- PARTAGE DES DONNEES A L'ISSUE DU PROJET

Voir produit de recherche 1.

G- DESCRIPTION ET ORGANISATION DES DONNEES

Quels méthodes et outils sont utilisés pour acquérir et traiter les données ? Précisez les différents formats dans lesquels les données seront disponibles aux différentes phases de la recherche
Les données spectrales 1D de la fraction aqueuse de chaque extraction des échantillons seront acquises sur le spectromètre Bruker Avance III HD 600 MHz équipé d'une cryosonde de la plateforme AXIOM (MetaToul). Les spectres sont acquis et traités par le logiciel TopSpin : format de sortie : fid pour les données brutes et .1rr pour les données traitées
Les spectres seront analysés afin de fournir deux types de quantification:
1- quantification en buckets dans l'instance W4M sous environnement Galaxy et avec le workflow NMR.
2- une quantification relative des métabolites à l'aide du package ASICS implémenté dans BioConductor.
Les fichiers archivés seront les spectres .fid et .1rr sous forme compressés, les fichiers de quantification en buckets et métabolites (format csv) et les métadonnées associées (format csv).

Documentation associée aux données
Origine des échantillons
Procédure d'extraction et cahier de laboratoire
Une procédure de traitement des spectres par W4M sera jointe aux analyses.
Liste des métabolites de référence.
Les scripts R de traitement par ASICS seront sauvegardés dans l'espace collaboratif GitLab de la Forge MIA qui gère les versions de code source et l'environnement R.

Quels types de métadonnées seront produites pour accompagner les données ? Quels sont les standards et les vocabulaires ou taxonomies qui seront utilisés pour décrire les données ?
1- Métadonnées associées à l'origine de l'échantillon: format csv (fichier samples FAANG data base (EBI))
2- Métadonnées associées à l'extraction de la phase aqueuse: format csv
3- Métadonnées associées à la génération des données brutes: format texte
4- Workflow et métadonnées de W4M
5- R markdown et métadonnées générées par ASICS.
Utilisation des standards de MetaboLighhts (EBI)

Comment les métadonnées seront elles produites ?
Les métadonnées sont générées par le pipeline W4M et renseignées par l'utilisateur en fonction du standard MetaboLights (EBI).

Comment les fichiers de données sont-ils gérés et organisés au cours du projet : contrôle des versions, conventions de nommage des fichiers, organisation des fichiers…
Même convention et organisation que pour le produit de recherche 2.

Quelle est la procédure de contrôle qualité des données ? joindre éventuellement le plan d'assurance qualité
Un pool d'échantillons contrôle sera rajouté tous les 30 échantillons pour le contrôle qualité de l'acquisition des données.
La plateforme AXIOM est certifiée ISO9001 et NFX50-900 et intégrée dans l'infrastructure MetaboHUB.

H- STOCKAGE ET SECURITE DES DONNEES

Stockage : Quels seront les supports utilisés pour les données au cours du projet ?
Sauvegarde à la plateforme MetaToul pendant 5 ans.
Sauvegarde dans l'espace projet Genotoul Bioinfo.
Les données brutes et finales seront déposées dans l'entrepôt MetaboLights (EBI) à la fin de l'acquisition et de traitement (responsables du dépôt : Nathalie Marty-Gasset, Laurence Liaubet) avec un embargo jusqu'aux publications des résultats de l'analyse de ces données dans un article scientifique au moins.

Stockage : Quels seront les types de flux empruntés par les données au cours du projet ?
Voir produit de recherche 2.

Stockage : Quelle est la volumétrie prévisionnelle ?
448 spectres RMN 1D : 750 Mo
2 spectres RMN 2D : 50 Mo

Stockage : Où sont hébergées physiquement les données, sur quel type d'hébergement ?
Arche des données Toulouse-Occitanie

Stockage : Où sont localisées géographiquement les données ?
Auzeville, centre INRAE Toulouse-Occitanie

Sécurité : L'entité hébergeant physiquement les données a-t-elle une politique de sécurité pour son système d'information ?
Identique au produit de recherche 2.

Sécurité - Confidentialité : les données feront-elles l'objet d'échange ou de partage avec de tiers acteurs ?
Identique au produit de recherche 2.

Sécurité - Confidentialité : comment sont déterminés les droits d'accès aux données pendant les recherches ?
Identique au produit de recherche 2.

Sécurité - Confidentialité : De quelle manière l'ensemble des chercheurs partenaires du projet auront-ils accès aux données pendant la recherche ?
Identique au produit de recherche 2.

Sécurité - Intégrité - Tracabilité : Quelles sont les mesures de protection mises en œuvre pour suivre la production et l'analyse des données au cours du projet ?
Traçabilité des échantillons et des extractions par e-SiToul
Cahier de laboratoire
Gestion des codes et des outils sur la Forge MIA (GitLAB)

I- ARCHIVAGE ET CONSERVATION DES DONNEES APRES LA FIN DU PROJET

Quelles sont les données à conserver sur le moyen ou le long terme et quelles sont les données à détruire ?
Les données spectrales ainsi que les données de quantification de métabolites et Buckets (csv) seront à conserver sur le moyen et long terme pour une ré-exploitation ultérieure. La conservation à long terme des spectres sera assurée par un dépôt sur un entrepôt de l'EBI. Les données finales seront conservées à long terme dans l'entrepôt Dataverse INRAE et associées aux publications.

Sur quelle plateforme d'archivage pérenne seront archivées les données à conserver sur le long terme ? Sinon, quelles procédures seront mises en place pour la conservation à long terme ?

Les données brutes seront archivées dans l'entrepôt Metabolights (EBI).
Les données de quantification seront conservées à long terme dans l'entrepôt Data INRAE.

Quelle est la durée de conservation des données ?
Les données seront conservées de manière pérennes aussi longtemps que les entrepôts de données de l'EBI existeront.

Qui sera responsable de la conservation à long terme ? nommer un contact individuel.
Le responsable de l'archivage à long terme sera Laurence Liaubet (GenPhySE).

Quel sera le volume de ces données ?
Spectres RMN 1D : 750 Mo
Spectres RMN 2D : 50 Mo

Quelles garanties de financements couvriront les coûts associés à la conservation à long terme ?
Le stockage dans les entrepôts de données de l'EBI est gratuit.

4- LIPIDOMIQUE : LIPIDES NEUTRES GC/FID

A- INFORMATIONS SUR LE PLAN DE GESTION

Voir produit de recherche 1.

B- INFORMATIONS SUR LE PROJET

Voir produit de recherche 1.

C- CONFIDENTIALITE

Voir produit de recherche 1.

D- PRESENTATION GENERALE DES DONNEES DU PROJET

Brève présentation des données générées, collectées ou réutilisées :
La phase lipidique de chaque échantillon sera dérivée et analysé en chromatographie gazeuse et détection FID (colonne Zebron ZB-5MS (Phenomenex), longueur 5m ID 0,25 mm, film 0,25 et détecteur FOCUS Finnigan (Thermo Scientific) par la plateforme I2MC (MetaToul) afin d'obtenir une identification des pics et une quantification relative. Les lipides neutres seront quantifiés par rapport à des standards internes et exprimés par mg de tissu ou par mg de protéines.
Les données générées sont des quantifications relatives des lipides neutres et sont disponibles sous forme de tableur excel (Dataset de quantification) et sous format CSV.

E- DROITS DE PROPRIETE INTELLECTUELLE

Voir produit de recherche 1.

F- PARTAGE DES DONNEES A L'ISSUE DU PROJET

Voir produit de recherche 1.

G- DESCRIPTION ET ORGANISATION DES DONNEES

Quels méthodes et outils sont utilisés pour acquérir et traiter les données ? Précisez les différents formats dans lesquels les données seront disponibles aux différentes phases de la recherche
La phase lipidique de chaque extraction des échantillons sera utilisée pour l'analyse lipidomique par chromatographie en phase gazeuse et détection FID (colonne Zebron ZB-5MS (Phenomenex), longueur 5m ID 0,25 mm, film 0,25 et détecteur FOCUS Finnigan (Thermo Scientific). Des chromatogrammes seront obtenus.
La quantification relative des lipides neutres (Cholestérol libre; Cholestérol estérifie C16, C18 et C20 :4 ; Triacylglycérides C49, C51, C53, C55, C57, C59) seront déterminées par mesure d'aires sous la courbe de chaque pic correspondant aux molécules d'intérêt (X) comparé à l'aire sous la courbe des standards internes (ISTD) de chaque famille étudiée (Stigmasterol, Cholesterol ester C17, TG 19) soit aires (X) / aire (ISTD). Ce rapport sera ensuite normalisé par mg de tissu et de protéines.
Les données seront disponibles sous forme de fichiers csv.

Documentation associée aux données

Origine des échantillons
"datas-calculs" regroupant les données brutes sorties de spectromètres (colonnes Surface)
Calculs utilisés pour la quantification des molécules d'intérêts.

Quels types de métadonnées seront produites pour accompagner les données ? Quels sont les standards et les vocabulaires ou taxonomies qui seront utilisés pour décrire les données ?

1- Métadonnées associées à l'origine de l'échantillon: format csv (fichier samples FAANG data base (EBI))
2- Métadonnées associées à l'extraction de la phase lipidique: format csv
3- Les lipides seront nommés suivant la nomenclature CHEBI.
4- Utilisation des standards de MetaboLighhts pour générer les métadonnées (EBI)

Comment les métadonnées seront elles produites ?
Les métadonnées sont générées par l'utilisateur en fonction du standard MetaboLights (EBI)

Comment les fichiers de données sont-ils gérés et organisés au cours du projet : contrôle des versions, conventions de nommage des fichiers, organisation des fichiers…
Même convention et organisation que pour le produit de recherche 2.

Quelle est la procédure de contrôle qualité des données ? joindre éventuellement le plan d'assurance qualité
Des blancs extraits et non extraits sont systématiquement analysés pour vérifier les effets mémoire de l'instrument ainsi que l'absence de contamination des solvants et du matériel d'extraction. Les ISTD sont injectés systématiquement avant et après extraction pour vérifier l'état de l'instrument et les rendements d'extractions.
Un pool d'échantillon contrôle sera rajouté tous les 35 échantillons pour s'assurer de la reproductibilité de l'analyse.
La plateforme est certifiée ISO9001 2015 et NFX 50900 : le périmètre de la certification recouvre tous les instruments et protocoles utilisés sur la plateforme.

H- STOCKAGE ET SECURITE DES DONNEES

Stockage : Quels seront les supports utilisés pour les données au cours du projet ?
Toutes les données brutes produites sur MetaToul sont enregistrées une fois par semaine et sont stockées à court terme (6 mois) dans des serveurs gérés par la plateforme BioInformatique de Toulouse (GénoToul), les archivages longs (5 ans) sont gérés par Agrodataring (entité d'archivage INRAE).
Les données de quantification seront sauvegardées sur l'espace projet de Genotoul Bioinfo et sur le sharepoint du projet.
Les données brutes et finales seront déposées dans l'entrepôt MetaboLights (EBI) à la fin de l'acquisition et du traitement (responsables du dépôt : Nathalie Marty-Gasset, Cécile Bonnefont) avec un embargo jusqu'aux publications des résultats de l'analyse de ces données dans un article scientifique au moins.

Stockage : Quels seront les types de flux empruntés par les données au cours du projet ?
Voir produit de recherche 2.

Stockage : Quelle est la volumétrie prévisionnelle ?
Question sans réponse.

Stockage : Où sont hébergées physiquement les données, sur quel type d'hébergement ?
Arche de données INRAE Toulouse-Occitanie.

Stockage : Où sont localisées géographiquement les données ?
Auzeville, centre INRAE Toulouse-Occitanie

Sécurité : L'entité hébergeant physiquement les données a-t-elle une politique de sécurité pour son système d'information ?
Voir produit de recherche 2.

Sécurité - Confidentialité : les données feront-elles l'objet d'échange ou de partage avec de tiers acteurs
Voir produit de recherche 2.

Sécurité - Confidentialité : comment sont déterminés les droits d'accès aux données pendant les recherches ?
Voir produit de recherche 2.

Sécurité - Confidentialité : De quelle manière l'ensemble des chercheurs partenaires du projet auront-ils accès aux données pendant la recherche ?
Voir produit de recherche 2.

Sécurité - Intégrité - Tracabilité : Quelles sont les mesures de protection mises en œuvre pour suivre la production et l'analyse des données au cours du projet ?
Traçabilité des échantillons et des extractions par e-SiToul
Cahier de laboratoire

I- ARCHIVAGE ET CONSERVATION DES DONNEES APRES LA FIN DU PROJET

Quelles sont les données à conserver sur le moyen ou le long terme et quelles sont les données à détruire ?
Les données chromatographiques ainsi que les données de quantification relative des lipides neutres (csv) seront à conserver sur le moyen et long terme pour une pré-exploitation ultérieure. La conservation à long terme des chromatogrammes sera assurée par un dépôt sur un entrepôt de l'EBI. Les données de quantification seront conservées à long terme dans l'entrepôt Dataverse INRAE et associées aux publications.

Sur quelle plateforme d'archivage pérenne seront archivées les données à conserver sur le long terme ? Sinon, quelles procédures seront mises en place pour la conservation à long terme ?
Les données brutes seront archivées dans l'entrepôt Metabolights (EBI).
Les données de quantification seront conservées à long terme dans l'entrepôt Data INRAE.

Quelle est la durée de conservation des données ?
Les données seront conservées de manière pérennes aussi longtemps que les entrepôts de données de l'EBI existeront.

Qui sera responsable de la conservation à long terme ? nommer un contact individuel.
Le responsable de l'archivage à long terme sera Laurence Liaubet (GenPhySE).

Quel sera le volume de ces données ?
Volume très faible (~50 Mo)

Quelles garanties de financements couvriront les coûts associés à la conservation à long terme ?
Le stockage dans les entrepôts de données de l'EBI est gratuit.

5- STATISTIQUE : ANALYSES STATISTIQUES

A- INFORMATIONS SUR LE PLAN DE GESTION

Voir produit de recherche 1.

B- INFORMATIONS SUR LE PROJET

Voir produit de recherche 1.

C- PRESENTATION GENERALE DES DONNEES DU PROJET

Brève présentation des données générées, collectées ou réutilisées :
Les analyses différentielles, d'intégration des données multi-omiques (RNAseq, métabolomique et lipidomique) et multi-tissus seront effectués à partir des produits de recherche précédents et à l'aide du logiciel libre R.
L'interprétation biologique sera effectuée par enrichissement fonctionnel à l'aide des logiciels sous licence Webgestalt, Genecodis4, String et Ingenuity.
Les données générées sont les résultats d'analyses différentielles, des représentations graphiques, des graphes de corrélations et des tableaux d'enrichissement fonctionnels.

D- DROITS DE PROPRIETE INTELLECTUELLE

Qui détiendra les droits sur les données et les autres informations créées lors du projet ?
Le détenteur de la propriété intellectuelle des données est INRAE dont relève le porteur de projet.

Du matériel protégé par des droits spécifiques sera t-il utilisé au cours du projet ? Dans ce cas, qui s'occupe des formalités à accomplir, obtient les autorisations d'utilisation et de diffusion éventuelle …

Voir produit de recherche 1.

E- CONFIDENTIALITE

Voir produit de recherche 1.

F- PARTAGE DES DONNEES A L'ISSUE DU PROJET

Voir produit de recherche 1.

G- DESCRIPTION ET ORGANISATION DES DONNEES

Quels méthodes et outils sont utilisés pour acquérir et traiter les données ? Précisez les différents formats dans lesquels les données seront disponibles aux différentes phases de la recherche
Les analyses statistiques différentielles seront effectuées sur les jeux de données traitées issus des autres produits de recherche et en utilisant des modèles mixtes généralisés ou linéaires (packages R : DESeq2/edgeR et lme4) en fonction de la spécificité des jeux de données. Les analyses seront complétées par des comparaisons deux à deux et par une correction pour les tests multiples.
L'intégration des données multi-omiques et multi-tissus sera effectuée par des méthodes non supervisées (packages R: Partial Least Square (PLS),DIABLO (mixOmics)) ainsi que par des méthodes d'inférence de réseaux.
Des méthodes de prédiction, classification et machine learning seront utilisées pour définir une signature moléculaire du statut de maturité (class prediction, class discovery, package CARET*e.g.,*Random Forest , logistic regression, Support Vector Machines).
Les résultats seront disponibles sous forme de tableur (csv), de graphes de classification et de réseaux.

Documentation associée aux données
Origine des échantillons
Description des méthodes
Codes source
Notebooks générées

Quels types de métadonnées seront produites pour accompagner les données ? Quels sont les standards et les vocabulaires ou taxonomies qui seront utilisés pour décrire les données ?
Les métadonnées incluront la description des méthodes et les procédures utilisées ainsi que les versions des outils et de l'environnement R.

Comment les métadonnées seront elles produites ?
Pour l'homogénéisation et l'extraction des paramètres d'environnement des analyses statistiques produites sous R, nous utiliserons le package R renv.

Comment les fichiers de données sont-ils gérés et organisés au cours du projet : contrôle des versions, conventions de nommage des fichiers, organisation des fichiers…
Les scripts réalisant les analyses statistiques et leurs versions seront gérés sur la ForgeMIA (https://forgemia.inra.fr/adminforgemia/doc-public/-/wikis/home) avec un environnement R commun défini par le package renv. Les scripts consisteront en des fichiers Rmarkdown + HTML pour les analyses les plus simples et en des scripts R simples pour les analyses les plus lourdes. Autant que possible, les scripts seront exécutés sur le cluster GenoToul Bioinfo.
Les scripts d'analyse produiront des fichiers résultats (fichiers de gènes différentiellement exprimés, fichiers de métabolites différentiellement exprimés, etc). Les différents fichiers de résultats issus des analyses statistiques ainsi que les métadonnées attachées seront gérés dans des dossiers situés dans l'espace collaboratif du cluster GenoToul Bioinfo de l'arche de données Occitanie-Toulouse selon la même organisation que les produits de recherche précédents.

Les différentes analyses seront réparties à raison d'un dossier par délivrable dans l'espace collaboratif et chaque dossier sera accompagné d'un fichier README décrivant le contenu de chaque fichier et les modifications accompagnant les versions.

Quelle est la procédure de contrôle qualité des données ? joindre éventuellement le plan d'assurance qualité
Question sans réponse.

H- STOCKAGE ET SECURITE DES DONNEES

Stockage : Quels seront les supports utilisés pour les données au cours du projet ?
L'environnement de travail R sera positionné dans l'espace de calcul du projet (work) de la plateforme Genotoul Bioinfo, labellisée IBISA.
Les analyses se feront dans l'espace projet du cluster GenoToul Bioinfo ou sur le disque réseau P du laboratoire.

Les résultats issus de l'analyse statistique seront également stockés dans le wiki de la Forge DGA.
Les scripts seront sauvegardés sur la Forge MIA.

Stockage : Quels seront les types de flux empruntés par les données au cours du projet ?
Les jeux de données traités finaux provenant des précédents produits de recherche seront disponibles sur l'espace projet de la plateforme Genotoul Bioinfo et sur la forge DGA. Les analyses statistiques seront effectuées dans l'espace projet de la plateforme ou à défaut dans un espace sauvegardé des ordinateurs (disque réseau P) avec un environnement synchronisé et contrôlé par le package renv. Les versions de scripts pré-finales et finales des analyses seront versionnées dans un projet dédié de la forgeMIA.

Stockage : Quelle est la volumétrie prévisionnelle ?
La volumétrie des scripts (code source) est probablement négligeable. La volumétrie des notebooks issue des sources est plus importante mais de l'ordre de quelques Mo par fichier donc relativement négligeable aussi.

Stockage : Où sont hébergées physiquement les données, sur quel type d'hébergement ?
Les données seront hébergées sur l'arche de données INRAE Toulouse-Occitanie

Stockage : Où sont localisées géographiquement les données ?
Auzeville INRAE Toulouse-Occitanie.

Sécurité : L'entité hébergeant physiquement les données a-t-elle une politique de sécurité pour son système d'information ?.
Voir produit de recherche 2.

Sécurité - Confidentialité : les données feront-elles l'objet d'échange ou de partage avec de tiers acteurs ?
Les données ne feront pas l'objet d'échanges ou de partage avec un tiers acteur.

Sécurité - Confidentialité : comment sont déterminés les droits d'accès aux données pendant les recherches ?

Voir produit de recherche 2.

Sécurité - Confidentialité : De quelle manière l'ensemble des chercheurs partenaires du projet auront-ils accès aux données pendant la recherche ?
Voir produit de recherche 1.

Sécurité - Intégrité - Tracabilité : Quelles sont les mesures de protection mises en œuvre pour suivre la production et l'analyse des données au cours du projet ?
La version de R et de l'environnement associé (version des packages utilisés) sera homogénéisée sur les postes individuels et l'espace projet Genotoul Bioinfo, notamment avec l'aide du package renv. Les analyses en version finale ou pré-finale seront systématiquement relancées sur le cluster avant dépôt des codes sources sur l'espace projet de la Forge MIA et publication éventuelle du notebook sur le Dataverse INRAe. La ForgeMIA assurera la traçabilité de l'évolution des analyses par versionnage git.

I- ARCHIVAGE ET CONSERVATION DES DONNEES APRES LA FIN DU PROJET

Quelles sont les données à conserver sur le moyen ou le long terme et quelles sont les données à détruire ?
Les scripts (code source et notebooks associés) correspondant aux analyses finales et pré-finales liées à des publications ainsi que les données intermédiaires issues de ces analyses seront à conserver à moyen et/ou long termes.

Sur quelle plateforme d'archivage pérenne seront archivées les données à conserver sur le long terme ? Sinon, quelles procédures seront mises en place pour la conservation à long terme ?
Les scripts bruts seront conservés sur la Forge MIA.
Les données intermédiaires issues des analyses statistiques seront associées aux publications comme matériel supplémentaire et pourront, par exemple, être déposés dans le Dataverse INRAe.

Quelle est la durée de conservation des données ?
Ces données seront conservées aussi longtemps que les entrepôts de dépôt (Forge MIA et Dataverse INRAe) seront conservés.

Qui sera responsable de la conservation à long terme ? nommer un contact individuel.
Le responsable de l'archivage à long terme sera Laurence Liaubet (GenPhySE).

Quel sera le volume de ces données ?
La volumétrie des scripts (code source) est probablement négligeable. La volumétrie des notebooks issue des sources est plus importante mais de l'ordre de quelques Mo par fichier donc relativement négligeable aussi.

Quelles garanties de financements couvriront les coûts associés à la conservation à long terme ?
La conservation des données sur la Forge MIA et sur le Data INRAe sont gratuits.