## Plan DE GESTION DES DONNEES communes de l’equipe PNDS sur psilo

### 1. Description des données et collecte ou réutilisation de données existantes

**1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?**

* Concernant les données produites :

Elles consistent en majeure partie de données de séquençage ADN et ARN de type Illumina® et BGI DNBseqTM et en moindre partie de d’images de microscopie.

* Concernant l’utilisation de données préexistantes :

Elles consistent en des données de séquençage ADN et ARN de type Illumina® et BGI DNBseqTM.

**1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?**

Les données sont au format le plus répandu dans notre communauté de recherche : fastq et stockées sous forme compressée (fastq.gz) pour les données de séquençage et .lif (format Leica), .czi (format Zeiss), et .tiff pour les données de microscopie confocale. Le volume de l’ensemble de ces données brutes est anticipé autour de 10 To.

Des fichiers découlant de l’analyse de ces données sont également stockés (format bam, sam, hic, bw, bedgraph, txt, png, pdf, tiff).

Le volume total de tout ce qui est décrit ci-dessus est de 30 To. Le volume dans les 5 ans à venir est estimé à 50 To.

### 2. Documentation et qualité des données

**2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?**

Les métadonnées associées aux données de séquençage ou de microscopie produites sont spécifiées par l’expérimentateur et s’organisent dans un fichier texte *Lisez-moi* ou *Readme* à la racine de l’arborescence du dossier contenant les données. Le dépôt des fichiers sur l’espace de stockage PSILO est conditionné au dépôt de ce fichier de métadonnées. Ces fichiers se basent sur les champs recommandés par le standard « Minimum Information about a high-throughput SEQuencing Experiment (MINSEQE) » et contiennent systématiquement :

1. la description du système biologique,
2. des échantillons et des variables expérimentales étudiées,
3. les données de lecture de la séquence pour chaque essai,
4. les données "finales" traitées (ou résumées) pour l'ensemble des essais de l'étude,
5. des informations générales sur l'expérience et les relations échantillon-données,
6. les protocoles essentiels d'expérimentation et de traitement des données.

Concernant les fichiers découlant de l’analyse des données brutes, les scripts ayant permis de les produire sont consignés avec la version des outils utilisés.

**2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?**

Le fichier de métadonnées référence également le lien vers la page correspondante du cahier de laboratoire en ligne de l’expérimentateur (https://www.elabftw.net/).

### 3. Stockage et sauvegarde pendant le processus de recherche

**3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?**

Pour l'hébergement physique des données, les données de séquençage sont stockées sur l’espace dit "kingdoms".

L’ensemble de ces environnements sont hébergés sur des serveurs de la plateforme informatique de l’IBENS et font l’objet de sauvegardes quotidiennes et le système est nativement protégé́ contre les ransomwares.

Stockage et redondance des données sur les serveurs de l’IBENS :

Les données informatiques et les résultats sont stockés sur un système de stockage Ceph dédié au projet PLASTONUC sur la plateforme informatique d'IBENS. Le stockage Ceph comprend une infrastructure triplée pour permettre la redondance.

Sauvegarde des données :

Les sauvegardes incrémentales sont effectuées quotidiennement, les sauvegardes différentielles sont effectuées hebdomadairement et les sauvegardes complètes sont effectuées une fois par mois sur des bandes LTO. Les bandes LTO sont stockées dans un autre bâtiment de l'ENS.

**3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?**

Ici une phrase sur l’organisation de PSILO (copies sur combien de disques, processus de recuperation)

Concernant le partage des données : les données sont partagées en lecture avec les membres de l’équipe PNDS et les droits d’écriture sont réservés à une personne de l’équipe PNDS en charge de cette activité et expérimentée (Adrien Vidal pour 2024).

Pour les échanges de données dans le cadre de collaborations, le partage se fait via le système renater du CNRS ou dropSU de Sorbonne Université.

### 4. Exigences légales et éthiques, codes de conduite

**4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?**

Aucune donnée à caractère personnel ou commercial n’est stockée.

**4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?**

Les données de ce projet seront rendues publiques et librement accessibles après publication des résultats dans des articles de recherche (via Gene Expression Omnibus du NCBI, <https://www.ncbi.nlm.nih.gov/geo/> ou Image Data Resource <https://idr.openmicroscopy.org/>).

Les articles de recherche et les programmes développés seront publiés sous une licence de type Creative Commons CC-BY dans la mesure du possible. Cela permet une utilisation sans restriction, une distribution et reproduction sans restriction sur tout support, à condition que l'œuvre originale soit correctement citée.

**4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?**

Pas de question éthique à signaler sur l’ensemble de nos données.

### 5. Partage des données et conservation à long terme

**5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?**

Les données publiées de ce projet seront rendues publiques et librement accessibles via Gene Expression Omnibus du NCBI ou Image Data Resource seulement après accord de Fredy Barneche (FB) et de Clara Richet-Bourbousse (CRB) ou publication des résultats dans des articles de recherche.

**5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?**

Conservation à long terme :

Après publication, et si elles sont retirées des disques, les données sont archivées par la plateforme informatique avec l'application d'archivage StorIQ One sur 2 bandes LTO différentes. Une bande est stockée à IBENS, l'autre dans un autre bâtiment de l'ENS.

**5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?**

Les données de séquençage sont dans des formats standards et seront facilement réutilisables sur le long terme.

**5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?**

Les données publiées sur Gene Expression Omnibus auront un identifiant unique généré par le NCBI.

### 6. Responsabilités et ressources en matière de gestion des données

**6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?**

La saisie des données est effectuée par la personne en charge au sein de l’équipe PNDS (Adrien Vidal en 2024) après accord de Fredy Barneche et Clara Richet-Bourbousse. Il s’assure que la production des métadonnées a été réalisée en suivant le guide établi par l’équipe PNDS. FB et CRB prennent en charge l’archivage et le partage. Fredy Barneche et Clara Richet-Bourbousse sont responsables de la mise en œuvre du PGD et s'assurent avec la direction de ArtBio, si nécessaire, de sa révision.

**6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?**

Le temps pour la saisie et la maintenance des données est prévu dans le temps de travail d’Adrien Vidal (en 2024) et/ou des membres de l’équipe PNDS impliqués dans l’encadrement ou le développement du projet concerné. Les ressources allouées proviennent des ressources de l’équipe PNDS.