# Plan de gestion de données de structure de la plateforme GenomiqueENS

Laurent Jourdren

# DMP DU PROJET "GENOMIQUEENS PGD DE STRUCTURE"

*Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "INRAE - Modèle Structure" 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 "GenomiqueENS PGD de structure" |
| Version | Version initiale |
| Domaines de recherche (selon classification de l'OCDE) | Biological sciences (Natural sciences) |
| Langue | fra |
| Date de création | 2023-01-25 |
| Date de dernière modification | 2025-01-02 |
| Licence | Creative Commons Attribution Non Commercial Share Alike 4.0 International |

## RENSEIGNEMENTS SUR LE PROJET

| | |
|---|---|
| Titre du projet | GenomiqueENS PGD de structure |
| Acronyme | GenomiqueENS |

**Produits de recherche :**

1. Séquencage (Jeu de données)
2. Controle qualité des échantillons (Jeu de données)
3. Analyse secondaire bioinformatique RNA-seq (Workflow)

**Contributeurs**

| Nom | Affiliation | Rôles |
|---|---|---|
| Jourdren Laurent | IBENS Institut de biologie de l'Ecole Normale Supérieure | • Coordinateur du projet<br>• Personne contact pour les données (Analyse RNA-seq, Séquençage, QC échantillon)<br>• Responsable du plan de gestion de données |

# DMP DU PROJET "GENOMIQUEENS PGD DE STRUCTURE"

## INFORMATIONS SUR LA STRUCTURE

**Nom de la structure**
GenomiqueENS

**Type de structure**

- Infrastructure de recherche

Membre de France Génomique

*IDENTIFIANT DE LA STRUCTURE PRÉCISER LE FOURNISSEUR DE L'IDENTIFIANT (ISNI, VIAF, FUNDREF, DATACITE...).*
Question sans réponse.

*RESPONSABILITÉS DANS LA STRUCTURE*

| Nom, Prénom | Courriel | Rôle |
|---|---|---|
| Thomas-Cholier, Morgane | | Responsable scientifique |
| Blugeon, Corinne | | Responsable pôle expérimental |
| Jourdren, Laurent | | Responsable pôle bioinformatique |

**Etablissement(s) tutelle(s)**

- ENS/PSL
- CNRS (UMR 8197)
- Inserm (U1024)

**Financeur(s)** (*permettant l'acquisition des jeux de données – hors projet)*

- IBENS
- ENS
- CNRS
- Inserm
- France Génomique
- IBiSA
- Région Île-de-France

## INFORMATIONS SUR LE PLAN DE GESTION

*DOI (VERSION PUBLIÉE DU PLAN DE GESTION)*
https://doi.org/xxxxxxxx/xxxxxxx

**Historique des versions**

| Date | n° de version | Status | Auteur | Affiliation de l'auteur *(se reporter à l'annuaire INRAE)* | Validé par | Validé le |
|---|---|---|---|---|---|---|
| 08/02/2023 | 1 | Final | Laurent Jourdren | IBENS | Morgane Thomas-Chollier | 14/02/2023 |
| 10/01/2024 | 2 | Final | Laurent Jourdren | IBENS | Morgane Thomas-Chollier | 10/01/2024 |
| 7/01/2025 | 3 | Final | Laurent Jourdren | IBENS | Morgane Thomas-Chollier | 08/01/2025 |

## PRÉSENTATION GÉNÉRALE DES DONNÉES

### SÉQUENÇAGE

*MODE D'OBTENTION DES DONNÉES*

- Données générées par la structure
- Données de séquençage à haut débit

**Origine**

- Expérimentation

**Type de données**

- Dataset

*NATURE DES DONNÉES*
Données brutes de séquençage à haut débit.

*FORMAT DES DONNÉES*

- Illumina
    - BCL (et autres données produites par un run Illumina)

        - FASTQ
        - HTML (Contrôle Qualité des runs et échantillons)
- Nanopore
    - Fast5
    - POD5
    - TSV (Nanopore, sequencing summary)
    - JSON (Nanopore, sequencing telemetry)
    - FASTQ
    - BAM (données non alignées)
    - HTML(Contrôle Qualité des runs)

*PÉRIMÈTRE THÉMATIQUE DES DONNÉES*

- Omics

**CONTRÔLE QUALITÉ DES ÉCHANTILLONS**

*MODE D'OBTENTION DES DONNÉES*

- Données générées par la structure
- Données de contrôle qualité des échantillons biologiques

**Origine**

- Expérimentation

**Type de données**

- Other (à préciser dans la zone "Informations supplémentaires")

Graphiques et données produits par les appareils utilisés pour contrôler la qualité des échantillons biologiques.

*NATURE DES DONNÉES*
Contrôle qualité d'échantillons biologiques.

*FORMAT DES DONNÉES*

- Rapports au format PDF
- Images au format PNG pouvant être copiées-collées dans d'autres documents (ex : fiche de suivi de projet au format XLSX ou dans un cahier de laboratoire)

*PÉRIMÈTRE THÉMATIQUE DES DONNÉES*

- Omics

**ANALYSE SECONDAIRE BIOINFORMATIQUE RNA-SEQ**

*MODE D'OBTENTION DES DONNÉES*

- Données produites par un tiers
- Données générées par la structure
- Données produites par un tiers
    - Séquences et annotations des génomes de référence
    - Données de séquençage à haut débit
- Données générées par la structure
    - Données de séquençage à haut débit
    - Données de contrôle qualité des échantillons biologique
    - Séquences et annotations des génomes de référence

**Origine**

- Analyse

**Type de données**

- Workflow
- Dataset

*NATURE DES DONNÉES*
Données d'analyse secondaire provenant de l'analyse de données brutes de séquençage à haut débit.

*FORMAT DES DONNÉES*

- TSV (fichiers d'expression, matrices de comptages, résultats d'analyse différentielle)
- XLSX (fichiers d'expression, matrices de comptages, résultats d'analyse différentielle)
- BAM (fichiers d'alignements sur le génome)
- BED (positions des alignements sur les génomes)
- BedGraph (fichier de couverture)

- FASTQ (fichiers de séquences filtrés)
- HTML (rapport qualité des séquences par échantillon)

*PÉRIMÈTRE THÉMATIQUE DES DONNÉES*

- Omics

## DROITS DE PROPRIÉTÉ INTELLECTUELLE

### SÉQUENÇAGE

*QUI DÉTIENDRA LES DROITS SUR LES DONNÉES ET LES AUTRES INFORMATIONS CRÉÉES ?*
Les institutions et laboratoires qui ont payé pour les prestations.

### CONTRÔLE QUALITÉ DES ÉCHANTILLONS

*QUI DÉTIENDRA LES DROITS SUR LES DONNÉES ET LES AUTRES INFORMATIONS CRÉÉES ?*
Les institutions et laboratoires qui ont payé pour les prestations.

### ANALYSE SECONDAIRE BIOINFORMATIQUE RNA-SEQ

*QUI DÉTIENDRA LES DROITS SUR LES DONNÉES ET LES AUTRES INFORMATIONS CRÉÉES ?*
Les institutions et laboratoires qui ont payé pour les prestations.

## SENSIBILITÉ DES DONNÉES

**Identification du niveau de sensibilité des jeux de données**

- Confidentiel

La diffusion des données dans des banques publiques peut-être réalisée sur demande des utilisateurs.

**Quelles sont les mesures prises et les normes auxquelles il est nécessaire de se conformer pour garantir la sécurité des données sensibles ?**
La plateforme GenomiqueENS ne traite pas des données sensibles.

**S'il y a des données à caractère personnel, quelles sont les mesures envisagées pour les protéger au cours du projet ou dans le cadre d'une réutilisation ?**
Les données à caractère personnel (nom, prénom, courriel et employeur) des utilisateurs ne sont pas utilisées dans les traitements des données des projets. Seul l'acronyme du projet est utilisé. Lors de la soumission des données dans des banques publiques (GEO), nous demandons au porteur du projet la liste des participants, car celle-ci doit être impérativement déclarée lors de la soumission.
L'accès aux données à caractère personnel (nom, prénom, courriel et employeur) des utilisateurs sont dans des fichiers (formulaire de contact, fiche de contact, devis…). L' accès restreint aux personnels de la plateforme. Ces informations sont partagées avec France Génomique (Unité Mixte de Service, CNRS 3628, INRA 1396, Inserm 026) dans le cadre de la remontée automatisée des indicateurs d'activité de notre structure.

## PARTAGE DES DONNÉES

### SÉQUENÇAGE

*Y A T'IL UNE OBLIGATION DE PARTAGE (OU À L'INVERSE UNE INTERDICTION OU UNE RESTRICTION) ?*
Non. Les données produites sont la propriété des porteurs de projets. Nous proposons et incitons à nos utilisateurs de déposer leurs données dans une banque publique (GEO).

*QUELLES SONT LES RÉUTILISATIONS POTENTIELLES DE CES DONNÉES ?*

- Réanalyse des résultats avec d'autres outils/pipeline d'analyse.
- Agrégation des données avec d'autres expériences.

*LA LECTURE DES DONNÉES NÉCESSITE-T-ELLE LE RECOURS À UN LOGICIEL OU UN OUTIL SPÉCIFIQUE ? SI OUI, LEQUEL ?*
Oui. Les données issues de l'analyse primaire de nos utilisateurs sont déposées dans la banque publique GEO. Pour pouvoir lire ces séquences, il faut utiliser le logiciel SRA toolkit. https://www.ncbi.nlm.nih.gov/sra/docs/sradownload/.
Le site NCBI permet toutefois de télécharger les données au format FASTQ. Cependant les données issues de l'analyse primaire et secondaire conservées sur la plateforme ne nécessitent pas de logiciel particulier (mis à part des outils classiques de création d'archive et de compression comme tar, zip, gzip…).

**Comment les données seront-elles partagées ?**
Les données à rendre publiques sont partagées via un dépôt des données sur la banque publique GEO.
À la demande du porteur de projet, les données (même si elles n'ont pas été déposées sur GEO) être mise à disposition d'un collaborateur via un serveur SFTP hébergé par le service informatique de notre institut.

*AVEC QUI ?*

- Tous (open acces)

Une fois les données rendues publiques sur GEO, les données sont accessibles par tous.

*SOUS QUELLE LICENCE ?*

- Autre (à préciser dans la zone d'Informations supplémentaires
- GEO Attribution required
- Open Data Commons (ODC) Public Domain Dedication and Licence (PDDL) 1.0
- HHS Vulnerability Disclosure Policy

### CONTRÔLE QUALITÉ DES ÉCHANTILLONS

*Y A T'IL UNE OBLIGATION DE PARTAGE (OU À L'INVERSE UNE INTERDICTION OU UNE RESTRICTION) ?*
Non. Ces données de contrôle qualité des échantillons en elles-mêmes n'ont pas vocation à être partagées. Elles sont à déposer dans un cahier de laboratoire qui lui peut-être partagé.

*QUELLES SONT LES RÉUTILISATIONS POTENTIELLES DE CES DONNÉES ?*

- Cahier de laboratoire

- Métadonnées de qualité d'échantillons

*LA LECTURE DES DONNÉES NÉCESSITE-T-ELLE LE RECOURS À UN LOGICIEL OU UN OUTIL SPÉCIFIQUE ? SI OUI, LEQUEL ?*
Non. Les rapports sont au format PDF et les images dans un format standard (PNG).

**Comment les données seront-elles partagées ?**
Rapport PDF envoyé via un courriel.

*AVEC QUI ?*

- Partenaire·s identifié·s

*SOUS QUELLE LICENCE ?*

- Autre (à préciser dans la zone d'Informations supplémentaires

Aucune licence définie, car ces données n'ont pas vocation à être rendue publique en tant tel. Elles sont à déposer dans un cahier de laboratoire.

**ANALYSE SECONDAIRE BIOINFORMATIQUE RNA-SEQ**

*Y A T'IL UNE OBLIGATION DE PARTAGE (OU À L'INVERSE UNE INTERDICTION OU UNE RESTRICTION) ?*
Non. Les données produites sont la propriété des porteurs de projets. Nous proposons et incitons à nos utilisateurs de déposer leurs données dans une banque publique (GEO).

*QUELLES SONT LES RÉUTILISATIONS POTENTIELLES DE CES DONNÉES ?*

- Analyse bioinformatique tertiaire des résultats expérimentaux.
- Agrégation des données avec d'autres expériences.

*LA LECTURE DES DONNÉES NÉCESSITE-T-ELLE LE RECOURS À UN LOGICIEL OU UN OUTIL SPÉCIFIQUE ? SI OUI, LEQUEL ?*
Les données issues de l'analyse secondaire conservées sur la plateforme ne nécessitent pas de logiciel particulier (mis à part des outils classiques de création d'archive et de compression comme tar, zip, gzip…). La plupart des formats sont des fichiers textes et les formats binaires (BAM, BedGraph) sont bien documentés et les logiciels nécessaires pour les lires sont des logiciels libres (ex : samtools).

**Comment les données seront-elles partagées ?**
Les données à rendre publiques sont partagées via un dépôt des données sur la banque publique GEO.
À la demande du porteur de projet, les données (même si elles n'ont pas été déposées sur GEO) peuvent être mise à disposition d'un collaborateur via un serveur SFTP hébergé par le service informatique de notre institut.

*AVEC QUI ?*

- Tous (open acces)

*SOUS QUELLE LICENCE ?*

- Autre (à préciser dans la zone d'Informations supplémentaires)
- GEO Attribution required
- Open Data Commons (ODC) Public Domain Dedication and Licence (PDDL) 1.0
- HHS Vulnerability Disclosure Policy

## ORGANISATION ET DOCUMENTATION DES DONNÉES

### SÉQUENÇAGE

*QUELS MÉTHODES ET OUTILS SONT UTILISÉS POUR ACQUÉRIR ET TRAITER LES DONNÉES, DEPUIS LEUR ACQUISITION JUSQU'À LEUR MISE À DISPOSITION, LEUR ARCHIVAGE OU LEUR DESTRUCTION ? UTILISER ÉVENTUELLEMENT UN LIEN VERS UN SCHÉMA ILLUSTRANT LES PROCESSUS*

- Données de séquençage à haut débit Illumina (NextSeq 2000)
    - Le détail du mode opératoire pour traiter les données du séquenceur NextSeq 2000 est défini dans un mode d'emploi sur le wiki en accès restreint de la plateforme.
    - Le logiciel NextSeq Control Software Suite réalise l'acquisition des données sur le séquenceur.
    - Le séquenceur écrit les données de sortie à la fois dans un espace de stockage sur le séquenceur et dans un partage hébergé et sauvegardé par le service informatique de l'institut.
    - Le démultiplexage et le contrôle qualité est ensuite effectué par le logiciel Aozan (développé par la plateforme) sur un serveur hébergé par le service informatique de l'institut.
    - Les données ultra-brutes (BCL) de séquençage Illumina peuvent être **détruites 6 mois après le séquençage**.
    - Les données démultiplexées sont archivées sur bande LTO au bout d'un an.
    - Le détail du mode opératoire de la mise à disposition des données (partages sur GEO et SFTP) est défini dans des modes d'emploi sur le wiki en accès restreint de la plateforme.
- Données de séquençage à haut débit Nanopore (MinION, PromethION)
    - Le détail du mode opératoire pour traiter les données des séquenceurs Nanopore est défini dans un mode d'emploi sur le wiki en accès restreint de la plateforme.
    - Le logiciel MinKNOW Control Software Suite réalise l'acquisition des données sur le séquenceur (Mk1C) ou le PC d'acquisition (Mk1B/P2 solo…).
    - Le séquenceur écrit les données de sortie sur le séquenceur (Mk1C) ou le PC d'acquisition (Mk1B/P2 solo…). À la fin du run, les données sont automatiquement transférées sur un partage hébergé au service informatique.
    - L'appel de base et le contrôle qualité est ensuite lancé manuellement.
    - Les données ultra-brutes (Fast5, Pod5) et appelées sont conservées **que deux mois après le séquençage** au cas où les utilisateurs souhaiteraient réaliser un appel de base ultérieur.
    - Les données appelées et démultiplexées sont archivées sur bande LTO au bout d'un an.

- Le détail du mode opératoire de la mise à disposition des données (partages sur GEO et SFTP) est défini dans des modes d'emploi sur le wiki en accès restreint de la plateforme.

**Quelles métadonnées seront utilisées pour accompagner le jeu de données ? Quels seront les standards, vocabulaires, taxonomies... utilisés pour décrire et représenter les données et éléments de métadonnées ? Comment les métadonnées seront-elles produites et mises à jour ?**

| **Métadonnées** | **Origine, mode de production des métadonnées (ex : saisie manuelle, annotation atomatique...)** | **Standard, Vocabulaires associés** | **Conditions ou fréquence de la mise à jour (si applicable)** *(ex : changement de l'accessibilité* |
|---|---|---|---|
| Title | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Summary | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Experimental design | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Contributors | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Sample names | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Sample organisms | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Sample molecules | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Other sample descriptions | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Single or paired-end | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Instrument model | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Library construction protocol | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Library strategy | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| FASTQ file names | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| FASTQ file check sums | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |

*UNE DOCUMENTATION COMPLÉMENTAIRE AUX MÉTADONNÉES EST-ELLE NÉCESSAIRE POUR DÉCRIRE LES DONNÉES ET ASSURER LEUR RÉUTILISABILITÉ SUR LE LONG TERME ?*
Tous les formats ainsi l'organisation des données sont expliquées dans les fichiers fournis aux utilisateurs.

*COMMENT LES FICHIERS DE DONNÉES SONT-ILS GÉRÉS ET ORGANISÉS : CONTRÔLE DES VERSIONS, CONVENTIONS DE NOMMAGE DES FICHIERS, ORGANISATION DES FICHIERS*
Pour les données brutes, appelées et démultiplexées des runs, le traitement se fait par numéro de run. Pour les séquençages Illumina, le numéro de run est imposé par le fabriquant. Pour les runs Nanopore, la convention est la suivante : AAAAMMJJ_NomProjet.

**Quel est le processus de contrôle qualité des données ?**

- Pour les runs Illumina, nous utilisons le logiciel Aozan (https://www.outils.genomique.biologie.ens.fr/aozan/) développé sur la plateforme pour réaliser le contrôle qualité des runs Illumina.

- Pour les runs Nanopore, nous utilisons le logiciel ToulligQC (https://github.com/GenomiqueENS/toulligQC) développé sur la plateforme pour réaliser le contrôle qualité des runs Nanopore.

**CONTRÔLE QUALITÉ DES ÉCHANTILLONS**

*QUELLES MÉTHODES ET OUTILS SONT UTILISÉS POUR ACQUÉRIR ET TRAITER LES DONNÉES, DEPUIS LEUR ACQUISITION JUSQU'À LEUR MISE À DISPOSITION, LEUR ARCHIVAGE OU LEUR DESTRUCTION ? UTILISER ÉVENTUELLEMENT UN LIEN VERS UN SCHÉMA ILLUSTRANT LES PROCESSUS*
Le logiciel d'acquisition des données produit les rapports au format PDF ainsi que les images. Les fichiers sont enregistrés en utilisant le système de dossiers partagés de la plateforme, puis sont envoyés par courriel.

**Quelles métadonnées seront utilisées pour accompagner le jeu de données ? Quels seront les standards, vocabulaires, taxonomies... utilisés pour décrire et représenter les données et éléments de métadonnées ? Comment les métadonnées seront-elles produites et mises à jour ?**
Aucune métadonnée n'est utilisée pour accompagner les jeux de données à part l'identifiant de l'échantillon.

*UNE DOCUMENTATION COMPLÉMENTAIRE AUX MÉTADONNÉES EST-ELLE NÉCESSAIRE POUR DÉCRIRE LES DONNÉES ET ASSURER LEUR RÉUTILISABILITÉ SUR LE LONG TERME ?*
Non. Les données étant basiques, cela ne se justifie pas.

*COMMENT LES FICHIERS DE DONNÉES SONT-ILS GÉRÉS ET ORGANISÉS : CONTRÔLE DES VERSIONS, CONVENTIONS DE NOMMAGE DES FICHIERS, ORGANISATION DES FICHIERS*
Les noms des rapports PDF sont définis de la manière suivante : NOMPROJET_TYPEMATERIEL.pdf (ex : MonProjet_A2022_cDNA.pdf)

**Quel est le processus de contrôle qualité des données ?**
Il n'y a aucun processus de contrôle de qualité des données puisqu'elles sont extrêmement simples. De plus les données produites sont déjà des données de contrôle qualité.

**ANALYSE SECONDAIRE BIOINFORMATIQUE RNA-SEQ**

*QUELLES MÉTHODES ET OUTILS SONT UTILISÉS POUR ACQUÉRIR ET TRAITER LES DONNÉES, DEPUIS LEUR ACQUISITION JUSQU'À LEUR MISE À DISPOSITION, LEUR ARCHIVAGE OU LEUR DESTRUCTION ? UTILISER ÉVENTUELLEMENT UN LIEN VERS UN SCHÉMA ILLUSTRANT LES PROCESSUS*

- L'analyse secondaire est réalisée à l'aide du logiciel Eoulsan développé sur la plateforme.

- Le détail des outils utilisés ainsi que les traitements à réaliser lors l'analyse RNA-seq est défini dans un mode d'emploi sur le wiki en accès restreint de la plateforme.

- Chaque projet analysé (hors projets avec une analyse « automatisée ») dispose d'une page dédié sur le wiki en accès restreint de notre plateforme.

- L'ensemble des traitements est réalisé sur les partages hébergés au service informatique du département.
- Le détail du mode opératoire de la mise à disposition des données (partages sur GEO et SFTP) est défini dans des modes d'emploi sur le wiki en accès restreint de la plateforme.
- Les analyses secondaires sont **archivées sur bande au bout de 6 mois**. Les données d'analyse des projets réalisés en mode **analyse « automatisée » sont détruites 3 mois** après la remise des résultats.

**Quelles métadonnées seront utilisées pour accompagner le jeu de données ? Quels seront les standards, vocabulaires, taxonomies... utilisés pour décrire et représenter les données et éléments de métadonnées ? Comment les métadonnées seront-elles produites et mises à jour ?**

| **Métadonnées** | **Origine, mode de production des métadonnées (ex : saisie manuelle, annotation atomatique...)** | **Standard, Vocabulaires associés** | **Conditions ou fréquence de la mise à jour (si applicable)** *(ex : changement de l'accessibilité* |
|---|---|---|---|
| Title | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Summary | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Experimental design | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Contributors | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Sample names | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Sample organisms | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Sample molecules | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Other sample descriptions | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Single or paired-end | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Instrument model | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Library construction protocol | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Library strategy | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Data processing steps | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| Genome build/assembly | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |

| | | | |
|---|---|---|---|
| Processed data files and content | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| FASTQ file names | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |
| FASTQ file check sums | Saisie manuelle | Vocabulaire défini dans le formulaire de soumission GEO | Saisie au moment de la soumission dans les banques publiques |

*UNE DOCUMENTATION COMPLÉMENTAIRE AUX MÉTADONNÉES EST-ELLE NÉCESSAIRE POUR DÉCRIRE LES DONNÉES ET ASSURER LEUR RÉUTILISABILITÉ SUR LE LONG TERME ?*
Tous les formats ainsi l'organisation des données sont expliquées dans les fichiers fournis aux utilisateurs.

*COMMENT LES FICHIERS DE DONNÉES SONT-ILS GÉRÉS ET ORGANISÉS : CONTRÔLE DES VERSIONS, CONVENTIONS DE NOMMAGE DES FICHIERS, ORGANISATION DES FICHIERS*
Les données analysées sont organisées par projet. Chaque projet dispose d'un acronyme. Toutes les données produites lors de l'analyse secondaire sont dans un unique dossier nommé avec l'acronyme du projet. Les versions des logiciels utilisées lors de l'analyse sont définies dans le fichier workflow utilisé par Eoulsan en entrée. Les conventions de nommage des fichiers produits sont celles d'Eoulsan. Les résultats de chaque étape de l'analyse sont dans un sous dossier dédié.

**Quel est le processus de contrôle qualité des données ?**
Le mode d'emploi de l'analyse et le modèle de fiche projet d'analyse sur le wiki en accès restreint de la plateforme indique les différents contrôles à effectuer afin de détecter d'éventuelles erreurs lors de l'analyse.

## STOCKAGE ET SÉCURITÉ DES DONNÉES

**Les systèmes d'information de la structure ont-ils fait l'objet d'une analyse de risques ou d'une homologation ?**

- Non

**Quels types de supports physiques sont utilisés pour stocker les données ?**
Les séquenceurs disposent de disques durs SSD. Les séquenceurs sont branchés sur un onduleur et sont disposés dans une pièce dédiée (climatisée avec accès restreint) sur la plateforme génomique.
Nous utilisons pour les calculs et le stockage, les partages et serveurs hébergés par le service informatique de notre institut.
L'archivage sur bandes LTO est également géré par le service informatique de notre institut.

**Quelles sont les mesures de sécurité mises en place lors des étapes de transfert des données ?**
Les transferts de données sont exclusivement effectués vers nos utilisateurs via le protocole chiffré SFTP.
Lors du dépôt des données dans la banque publique GEO, les données sont envoyées via le protocole non chiffré FTP, mais il s'agit de données allant être rapidement rendues publiques.

**Quelle est la volumétrie actuelle et prévisionnelle ?**

- Activité actuelle
    - Pour Illumina nous produisons 1,5 To de données de FASTQ (et un volume similaire de données ultra-brutes) par an.
    - Pour Nanopore nous produisons 600 Go de données de FASTQ par an ainsi que 9 To de données ultra-brutes.
- Activité prévisionnelle

- La volumétrie pour le séquençage Illumina devraient rester relativement stables dans les prochaines années.

- La volumétrie pour le séquençage Nanopore devrait augmenter avec du nombre croissant de séquençages réalisés sur un séquenceur très haut débit PromethION P2 solo. Le volume maximal de données ultra-brutes produites lors d'un run avec une flowcell PromethION étant supérieur à 2 To.

**L'entité hébergeant physiquement les données a-t-elle une politique de sécurité de l'information et a-t-elle un plan d'assurance sécurité ?**
L'entité hébergeante (ENS) dispose d'une PSSI. En plus de cette PSSI, l'IBENS suit également la PSSI du CNRS et de l'Inserm, cependant la PSSI de l'ENS est la PSSI de référence.

**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. Le stockage Ceph comprend une infrastructure tripliquée pour permettre la redondance et éviter la perte de données.

**Sauvegardes et archivages**
Des sauvegardes incrémentales sont effectuées quotidiennement, les sauvegardes différentielles hebdomadairement et les sauvegardes complètes tous les deux mois sur des bandes LTO.
Les bandes LTO sont stockées dans un autre bâtiment de l'ENS situé à 500 m de l'institut. En cas d'incident, les données sont récupérées par la plateforme informatique de l'IBENS grâce aux sauvegardes incrémentales. Les durées de rétention des sauvegardes sont de 5 à 6 mois.

**Sécurité de l'exploitation de serveurs de calcul et de stockages**
L'accès physique aux serveurs et consoles d'administration de ceux-ci est limité aux seules personnes autorisées par système de badge. Les serveurs de stockage et les serveurs de calcul sont sous contrat de garantie ou de maintenance.
Aucune sous-traitance n'est utilisée. L'intégrité des services (mise à jour critique, surveillance intrusion, etc.) est réalisée par la plateforme informatique de l'IBENS.
Lorsqu'une pièce est défectueuse, cette dernière est remplacée dans les plus brefs délais si elle est encore sous garantie ou jetée dans les déchets électroniques. S'il s'agit d'un disque dur, il est stocké dans un lieu sécurisé avant destruction.
Enfin les salles serveurs sont équipées de plusieurs circuits électriques indépendants et protégées contre de petites coupures de courant grâce à des onduleurs.

**Sécurité - Confidentialité : les données font-elles l'objet d'échange ou de partage avec de tiers acteurs et selon quelles modalités ? comment sont déterminés les droits d'accès aux données avant leur publication ?**

- Les données ne font pas l'objet d'échange ou de partage avec de tiers acteurs.

- Le droit d'accès aux données avant publication se fait sur demande du porteur de projet.

- Seuls les membres de la plateforme ont accès aux données.

**Sécurité - Intégrité - Tracabilité : Quelles sont les mesures de protection mises en œuvre pour suivre la production et l'analyse des données ?**

- Base de données des projets

- Base de données de suivi qualité

- Fichier de suivi de projet

- Page wiki pour l'analyse d'un projet

- Horodatage des données et résultats

**Les agents de la structure ont-ils bénéficié d'une sensibilisation aux bonnes pratiques d'hygiène numérique ?**

- Oui

## ARCHIVAGE ET CONSERVATION DES DONNÉES

*QUELLES SONT LES DONNÉES À CONSERVER SUR LE MOYEN OU LE LONG TERME ET QUELLES SONT LES DONNÉES À DÉTRUIRE ?*
Les données à conserver sont :

- Les données ultra-brutes des séquenceurs Nanopore (Fast5, Pod5) car les logiciels d'appel de base s'améliorent avec le temps. Cette conservation ne concerne que les projets de recherche et développement. Les données ultra-brutes des projets de prestation ne sont conservées **que deux mois** au cas où les utilisateurs souhaiteraient réaliser un appel de base ultérieur.

- Les données appelées et démultiplexées.

- Les données de contrôle qualité.

- Les données d'analyse secondaire.

- Base de données des projets

- Fichier de suivi projet

- Pages wiki pour l'analyse d'un projet

Les données à détruire à moyen terme :

- Les données ultra-brutes de sortie de séquenceur Illumina (BCL)

*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 pérennes sont archivées sur bandes par la plateforme informatique et conservées à 500 mètres de l'institut.

- Les données partagées avec le public sont hébergées par GEO.

- La plateforme GenomiqueENS n'a pas vocation à archiver sur le long terme les données de ses utilisateurs.

*QUELLE EST LA DURÉE DE CONSERVATION DES DONNÉES ?*
Les données restent récupérables par les porteurs de projet au moins 2 mois après la fin du projet.

*QUELLES GARANTIES DE FINANCEMENTS COUVRIRONT LES COÛTS ASSOCIÉS À LA CONSERVATION À LONG TERME ?*
Le coût de conservation des données archivées sur bande LTO est faible et est pris en charge par notre institut.