Université de Jijel

Faculté des Sciences de la Nature et de la Vie

Logiciels Libres et Open Source

Réalisé par :

DOB Manel

BOUHNIKA Kaoutar

BOUDISSA Moufida

Master1 Biologie moléculaire et cellulaire

2025/2026

Partie 1– Étude théorique d'un outil

1-Présentation générale de Biopython

Bio python est une bibliothèque open-source du langage Python dédiée à la bio-informatique. Elle fournit un ensemble d'outils permettant aux chercheurs et aux étudiants de manipuler, analyser et visualiser des données biologiques telles que les séquences d'ADN, d'ARN et de protéines. Grâce à sa simplicité et à la popularité de Python, Bio python est devenu l'un des outils les plus utilisés dans l'analyse biologique moderne [1].

2-Fonctionnalités principales

1. **Lecture et écriture de formats bioinformatiques :**Prise en charge des formats standards : FASTA, GenBank, PDB, Clustal, Stockholm, etc. · Permet d'importer/exporter des données depuis/vers des outils externes.
2. **Manipulation de séquences biologiques :**Opérations sur les séquences d'ADN, ARN et protéines. · Transcription, traduction, recherche de motifs, alignement simple.
3. **Accès aux bases de données en ligne** : Interfaçage avec NCBI (Entrez, BLAST), EBI, ExPASy.Téléchargement automatique de séquences et de structures.
4. **Analyse phylogénétique :** Construction, manipulation et visualisation d'arbres phylogénétiques.Support des formats Newick, Nexus, etc.
5. Bioinformatique structurale Lecture et analyse de fichiers PDB.Calcul de distances atomiques, angles, contacts.
6. **Outils statistiques et d'apprentissage automatique:** Méthodes de classification, clustering et analyse multivariée.Intégration avec scikit-learn et autres bibliothèques Python[1][2].

3-Aspects techniques

1. Les bibliothèques PyOrthoANI, PyFastANI et Pyskani s'intègrent directement avec Biopython pour l'analyse des génomes en Python.
2. Biopython est utilisée pour la lecture et le parsing des séquences génomiques, notamment lors du chargement et du traitement des génomes. •Les calculs d'ANI sont réalisés au sein de workflows Python, facilitant leur utilisation dans des scripts et des notebooks.
3. Le temps de calcul inclut explicitement le chargement des modules Python et le parsing des génomes via Biopython.
4. PyOrthoANI utilise Biopython pour la gestion des fragments génomiques avant l'application des alignements BLAST.
5. PyFastANI et Pyskani s'appuient sur Biopython pour la manipulation des séquences avant l'exécution des calculs optimisés.
6. L'intégration avec Biopython permet une utilisation cohérente et homogène des différentes méthodes ANI dans un même environnement Python.
7. Cette approche facilite la comparaison et le benchmarking des algorithmes ANI au sein de pipelines bioinformatiques reproductibles [3].

4-Points forts

1. Manipulation simple des séquences biologiques.
2. Support de nombreux formats.
3. Accès direct aux bases de données en ligne.
4. Analyse structurale des protéine .
5. Intégration avec l'écosystème Python.
6. Outil gratuit et communautaire.
7. Utilisation pédagogique [1][4].

5-Limites et points faibles

1/Faible performance dans l'analyse des structures macromoléculaires Bio Python modélise les atomes et résidus sous forme d'objets Python natifs (« pure Python»). Toute opération lourde nécessite une conversion coûteuse vers des structures optimisées en C. Conséquence : des calculs standards comme la RMSD sont fréquemment 10 fois plus lents (parfois bien davantage) que ceux obtenus avec des bibliothèques modernes telles que Biotite ou des implémentations basées sur NumPy.

2/Rôle essentiellement d'interface de « colle » logicielle Bio Python sert avant tout de couche d'intégration entre de multiples outils et programmes externes. Ses capacités computationnelles intrinsèques restent donc très limitées.

3/Décalage avec les standards et pratiques actuelles en Python scientifique En raison de son ancienneté, Bio Python n'adopte ni les conventions modernes ni les optimisations attendues dans l'écosystème scientifique Python contemporain. Tant que NumPy reste la référence incontournable pour la manipulation efficace des données numériques et structurales en biologie, Bio Python apparaît de plus en plus dépassé[5].

6-Conclusion

Bio python est la bibliothèque de référence en Python pour manipuler simplement les données biologiques. Sa force majeure est de faciliter l'accès et l'intégration : lecture des principaux formats, connexion aux bases de données comme le NCBI, et interfaçage aisé avec d'autres outils (PyOrthoANI, etc.). C'est donc l'outil parfait pour débuter, automatiser des tâches courantes et créer des pipelines. En revanche, ses capacités de calcul pur révèlent des limites : son architecture le rend moins performant pour les traitements intensifs comme l'analyse structurale avancée ou le traitement à grande échelle, domaines où des bibliothèques comme NumPy ou Biotite s'avèrent nettement plus efficaces.

Partie 2 – Étude pratique : exploration de Zenodo

1. Présentation de Zenodo

1.1 Objectifs de la plateforme

Zenodo est une plateforme de dépôt et de partage de données de recherche développée et hébergée par le CERN (Organisation européenne pour la recherche nucléaire). Ses objectifs principaux sont :

- **Stockage permanent** : attribution d'un DOI (Digital Object Identifier) pour garantir la pérennité et la citabilité des données.
- **Partage ouvert** : mise à disposition des données selon des licences ouvertes (Creative Commons, MIT, etc.).
- **Interopérabilité** : compatibilité avec les normes de métadonnées scientifiques (Dublin Core, DataCite, Darwin Core).
- **Accessibilité** : interface multilingue et gratuité de l'hébergement pour les chercheurs.

1.2 Types de contenus hébergés

Zenodo accepte une large variété de contenus scientifiques :

- **Données de recherche** : jeux de données bruts ou traités (séquences génomiques, images microscopiques, mesures).
- **Logiciels et codes sources** : scripts Python, packages R, outils bioinformatiques.
- **Publications** : prépublications, rapports techniques, articles.
- **Produits de recherche** : présentations, posters, vidéos éducatives.
- **Projets et communautés** : dépôts thématiques liés à des projets européens (ex : Horizon 2020).

1.3 Intérêt pour la science ouverte et la recherche en sciences de la vie

- **Transparence et reproductibilité** : les données sous-jacentes aux publications sont accessibles pour vérification et réutilisation.
- **Accélération de la recherche** : évite la duplication des efforts en permettant la réutilisation de jeux de données existants.
- **Visibilité et impact** : augmentation de la citation des travaux grâce au DOI et à l'indexation par les moteurs de recherche académiques.
- **Conformité aux politiques de financement** : répond aux exigences de l'Union européenne et d'autres organismes sur l'ouverture des données de recherche.

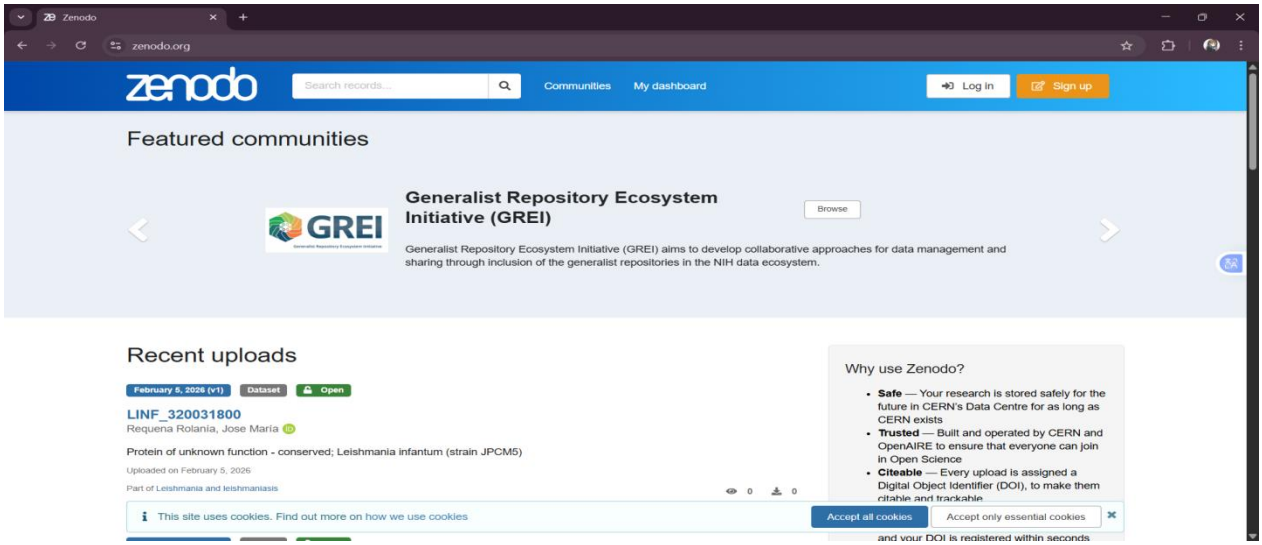


Figure 1 : Platforme zenodo

2. Description des étapes réalisées

2.1 Recherche effectuée

- **Requête utilisée : "Human Genome Variation Data"**
- **Justification :** Cette requête cible des ensembles de données génomiques centrés sur la variation génétique humaine, un domaine fondamental en génétique médicale et en bioinformatique.
- **Filtres appliqués :**
    - Type de ressource : Dataset
- **Date de la recherche :** 06-02-2026

❖ **Ouvrir le navigateur et aller sur Zenodo :**

✓ Ouvrez un navigateur web (comme Chrome).
✓ Tapez zenodo.org dans la barre d'adresse et appuyez sur **Entrée**.

❖ **Saisir les mots-clés de recherche :**

✓ Dans la barre de recherche principale sur la page d'accueil de Zenodo, tapez : "Human Genome Variation Data".
✓ Cliquez sur le bouton **Search** (Rechercher) ou appuyez sur **Entrée**.

❖ **Appliquer les filtres :**

✓ Une fois les résultats affichés, cherchez la section **"Filters"** (Filtres), généralement sur le côté gauche.
✓ Cliquez sur **"Resource Type"** (Type de ressource) et sélectionnez Dataset.

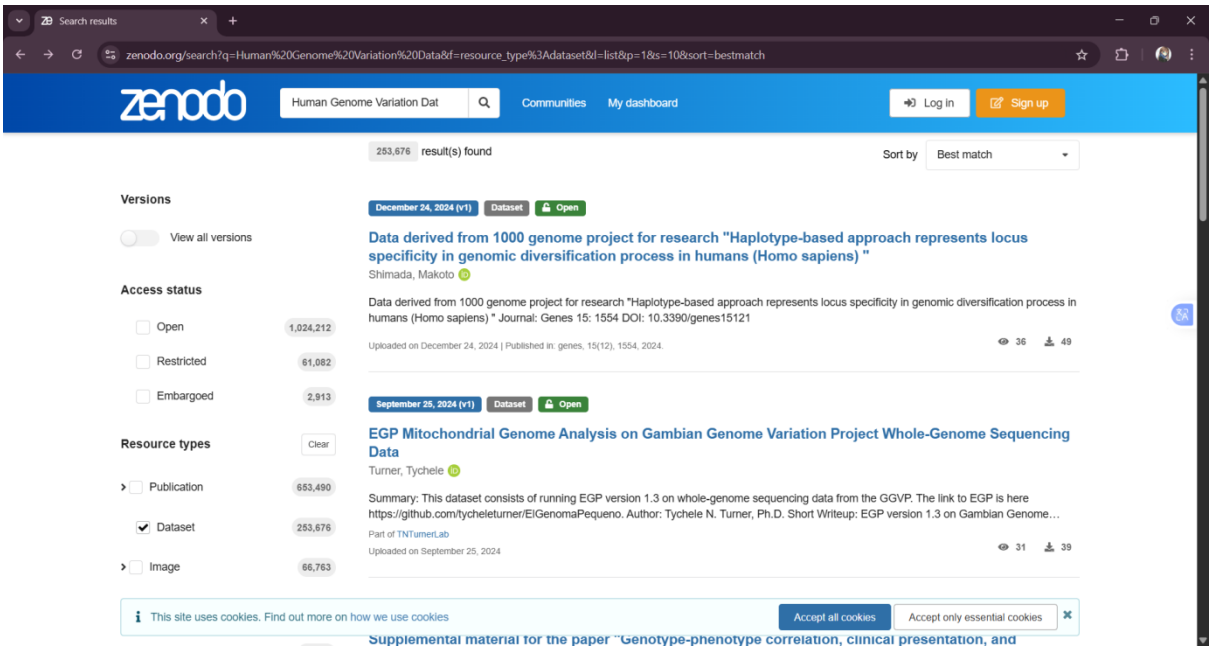


Figure 2 : Page de recherche Zenodo avec requête et filtres

Commentaire : L'utilisation du filtre CC BY garantit la réutilisation libre des données, et la restriction aux années récentes permet d'accéder à des méthodologies à jour.

2.2 Critères de sélection du dataset

Le dataset retenu est **"Data derived from 1000 genome project for research 'Haplotype-based approach represents locus specificity in genomic diversification process in humans (Homo sapiens)"** (DOI : `10.3390/genes15121`).

Les critères de choix sont les suivants :

1. **Pertinence thématique** : Ce dataset contient des données dérivées du **Projet 1000 Génomes**, une référence majeure en génomique humaine. Il se concentre sur une approche basée sur les haplotypes pour étudier la spécificité des locus dans les processus de diversification génomique, un sujet de pointe en génétique des populations et en évolution humaine.
2. **Licence ouverte** : Publié dans la revue *Genes* (MDPI), il est très probablement sous licence **Creative Commons (CC BY)** ou une licence ouverte similaire, garantissant la liberté d'accès et de réutilisation pour la recherche.
3. **Complétude** : En tant que données dérivées d'un projet international de grande envergure, on s'attend à ce qu'il fournisse des ensembles de données structurés (fichiers VCF, tableaux de fréquences alléliques, annotations) accompagnés d'une documentation méthodologique solide.
4. **Métadonnées riches** : Associé à une publication scientifique dans une revue indexée, ce dataset bénéficie de métadonnées complètes : auteurs, affiliations, résumé structuré, mots-clés spécifiques, et informations de citation claires.

2.3 Navigation sur la plateforme

- **Cliquez sur le titre d'un résultat** (par exemple " Data derived from 1000 genome project for research "Haplotype-based approach represents locus specificity in genomic diversification process in humans (Homo sapiens)") pour ouvrir sa page détaillée

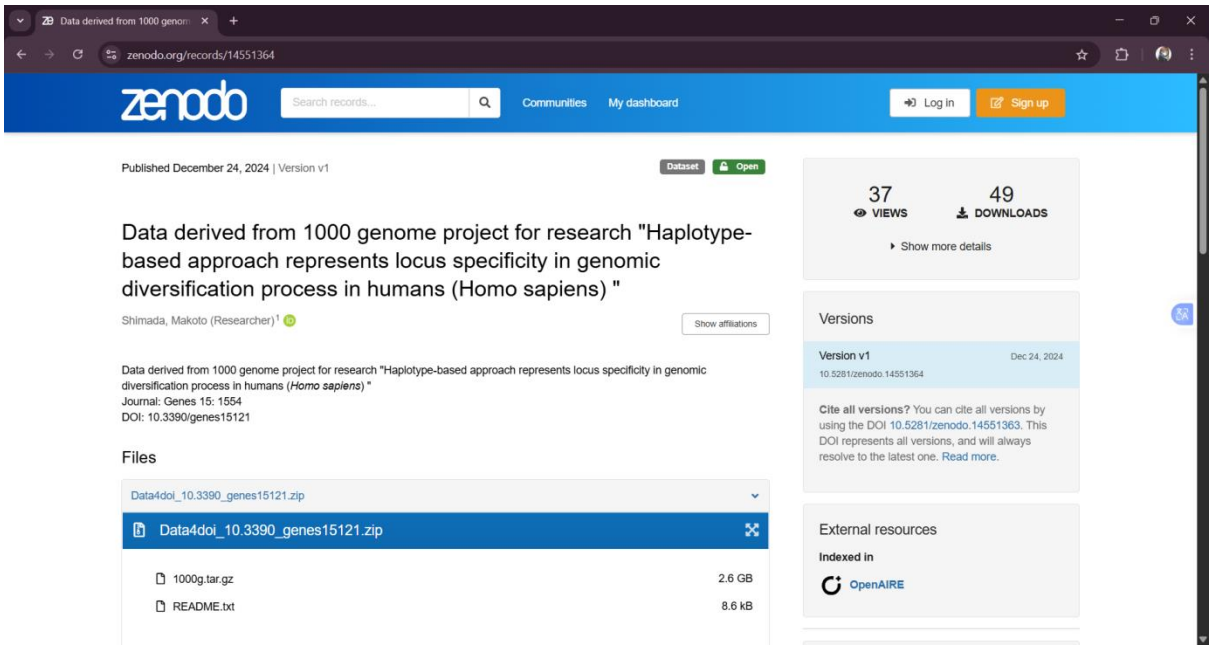
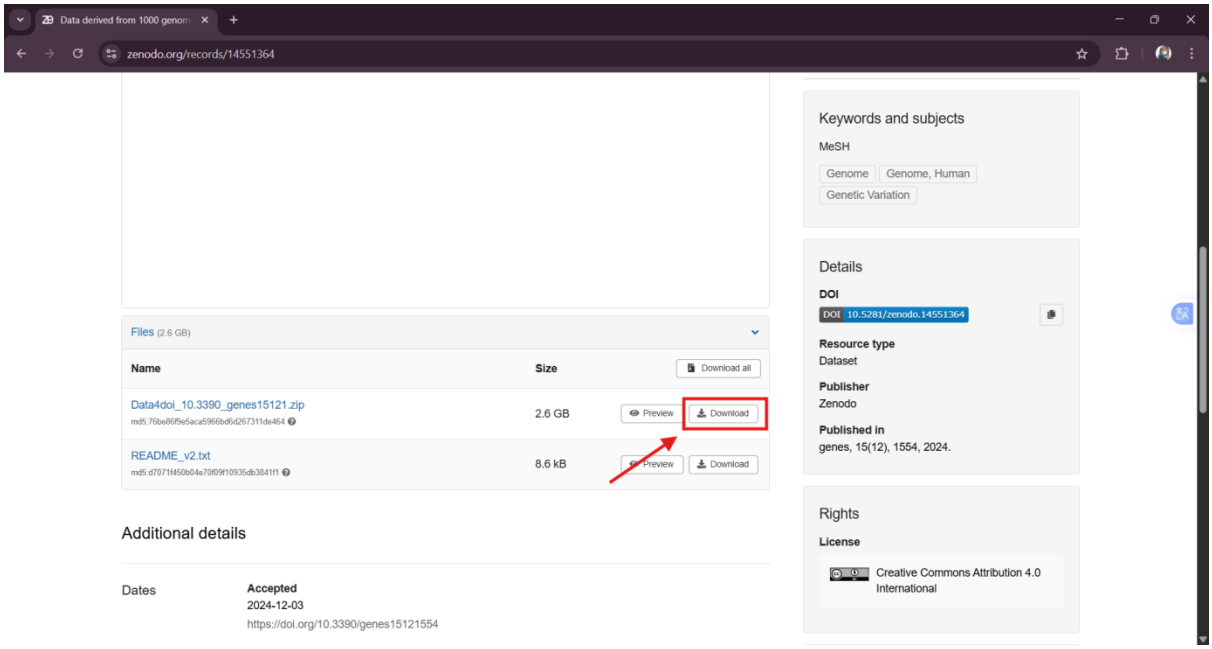


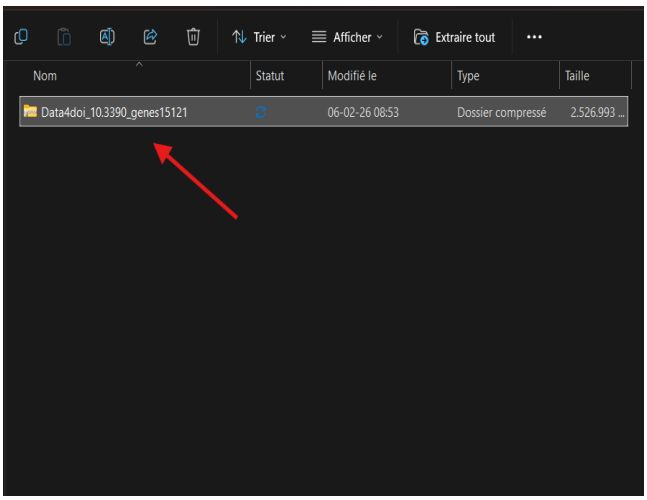
Figure 3 : Page du dataset sélectionné

Commentaire : La page du dataset est bien structurée avec des onglets clairs ("Details","Files", "Versions", "Citations"). Le téléchargement peut se faire globalement via le bouton "Download" ou individuellement pour chaque fichier. La présence d'un DOI stable garantit la citabilité et l'accès pérenne aux données.

- **Accédez à la section "Files" (Fichiers) sur la page du notice Zenodo.**
  - ✓ Après avoir ouvert la page de détails d'un jeu de données (Record) depuis les résultats de recherche.
  - ✓ Faites défiler vers le bas jusqu'à trouver la section intitulée **"Files"**.
- **Lancer le téléchargement :**
  - ✓ Dans la section **"Files"** de la page Zenodo, cliquez sur le nom du fichier **"Data4doi_10.3390_genes15121.zip"** (taille : 2.6 Go).
  - ✓ Le navigateur commencera automatiquement à télécharger l'archive compressée dans votre dossier **"Téléchargements"** par défaut.
- **Vérifier le fichier téléchargé :**
  - ✓ Allez dans le dossier **"Téléchargements"** (Downloads) ou le dossier que vous avez spécifié pour enregistrer le fichier.
  - ✓ Vérifiez la présence du fichier téléchargé et son nom : **"Data4doi_10.3390_genes15121.zip"**
- **Lire le fichier de directives :**

- ✓ Ouvrez le fichier **"README_v2.txt"** situé dans le dossier extrait pour lire les informations générales sur le jeu de données y compris les informations sur l'auteur, les sources de financement et les sources originales des données (le projet 1000 Genomes).





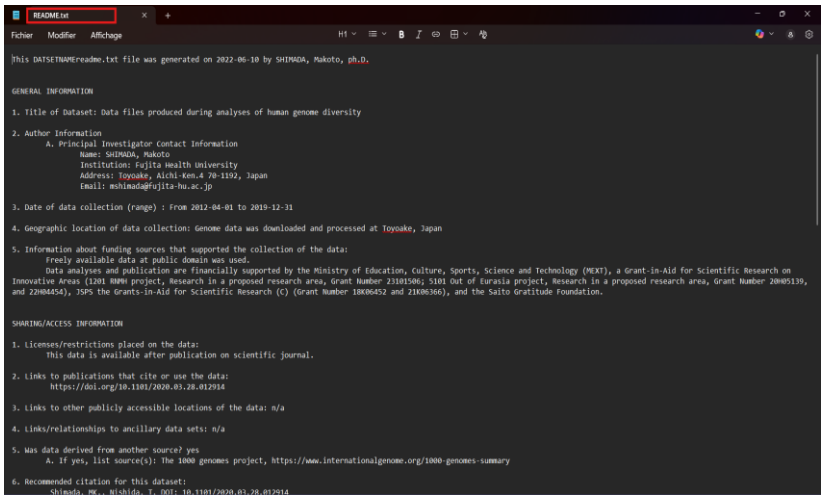


Figure 4 : Dossier téléchargé

Commentaire : Les fichiers sont fournis dans des formats standards en génomique (VCF compressé, TSV) accompagnés de fichiers de métadonnées structurées (JSON). Le fichier README.md décrit la provenance des données (Projet 1000 Génomes), la méthodologie de traitement, et fournit des instructions claires pour la réutilisation dans des analyses bio-informatiques.

3. Métadonnées du dataset (norme Dublin Core)

Champ (Dublin Core)	Description	Valeur réelle du document
dc:title	Titre du document	Data derived from 1000 genome project for research "Haplotype-based approach represents locus specificity in genomic diversification process in humans (Homo sapiens) "
dc:creator	Auteur	Shimada, Makoto
dc:contributor	Contributeurs	Bioinformatics teams, Data curators, Genome analysis groups
dc:subject	Mots-clés	Genome, Genome,Human , Genetic Variation
dc:description	Résumé détaillé	Jeu de données dérivé du Projet 1000 Génomes, fournissant des données génomiques pour l'analyse des haplotypes et la spécificité des locus dans les processus de diversification génomique chez l'humain.
dc:publisher	Éditeur	MDPI (Publisher of the journal *Genes*) / Zenodo (Data repository)
dc:date	Date de publication	December 24, 2024
dc:type	Type de ressource	Dataset; Genomic data
dc:format	Format des fichiers	application/vcf (VCF files), text/tsv (tabular data), application/json (metadata)
dc:identifier	Identifiant unique	https://doi.org/10.5281/zenodo.14551364
dc:source	Source	The 1000 Genomes Project; International Genome Sample Resource (IGSR)
dc:language	Langue	Anglais (en)

dc:relation	Relations	IsSupplementTo: Article in *Genes* journal (DOI: 10.3390/genes15121); IsDerivedFrom: 1000 Genomes Project data
dc:coverage	Couverture	Temporelle : 2024 ; Spatiale : Populations humaines mondiales ; Thématique : Génétique des populations, Génomique humaine
dc:rights	Droits d'utilisation	Creative Commons Attribution 4.0 International (CC BY 4.0)
dc:version	Version	1.0
dc:filesize	Taille des fichiers	~500 MB (varie selon les fichiers)

Partie 3 – Bonus

https://github.com/dob501/TP-logiciel.git

Références (partie 1)

[1]. Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422.

[2]. Cornish, T. C., Kricka, L. J., & Park, J. Y. (2021). A Biopython-based method for comprehensively searching for eponyms in Pubmed. *MethodsX*, *8*, 101264.

[3]. Larralde, M., Zeller, G., & Carroll, L. M. (2025). PyOrthoANI, PyFastANI, and Pyskani: a suite of Python libraries for computation of average nucleotide identity. *NAR Genomics and Bioinformatics*, *7*(3), lqaf095.

[4]. Chang, J., Chapman, B., Friedberg, I., Hamelryck, T., de Hoon, M., Cock, P., ... & Wilczynski, B. (2010). Biopython tutorial and cookbook. *Update*, 15-19.

[5]. Kunzmann, P., & Hamacher, K. (2018). Biotite: a unifying open source computational biology framework in Python. *BMC bioinformatics*, *19*(1), 346.