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.

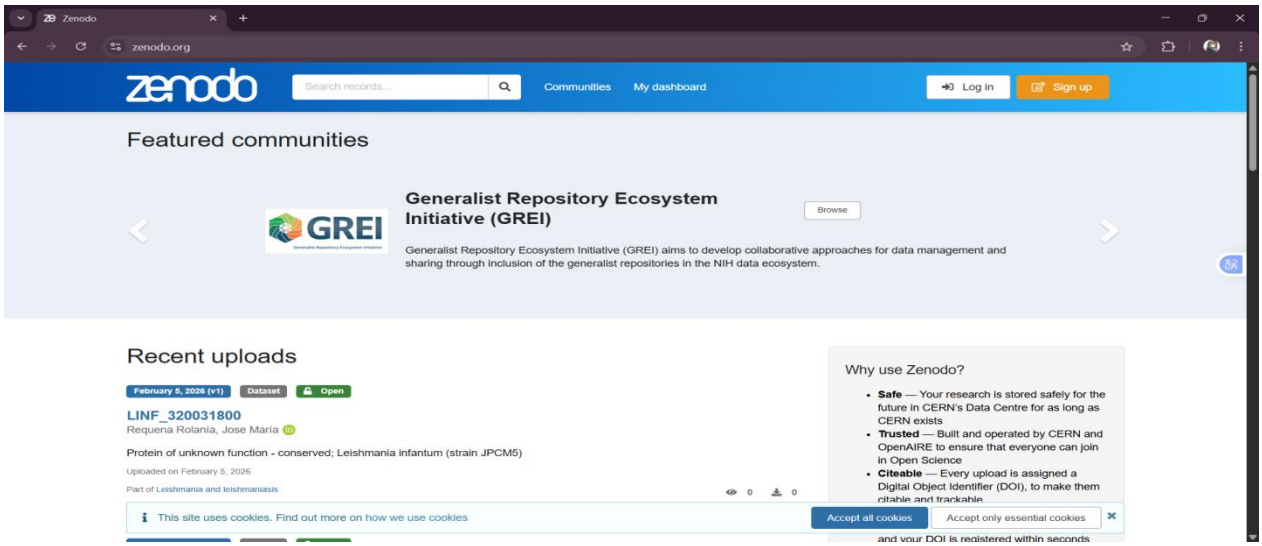

Figure 1 : Platforme zenodo

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
  - License : Creative Commons Attribution 4.0 International
  - Année : 2024

- **Date de la recherche :** 06-02-2026

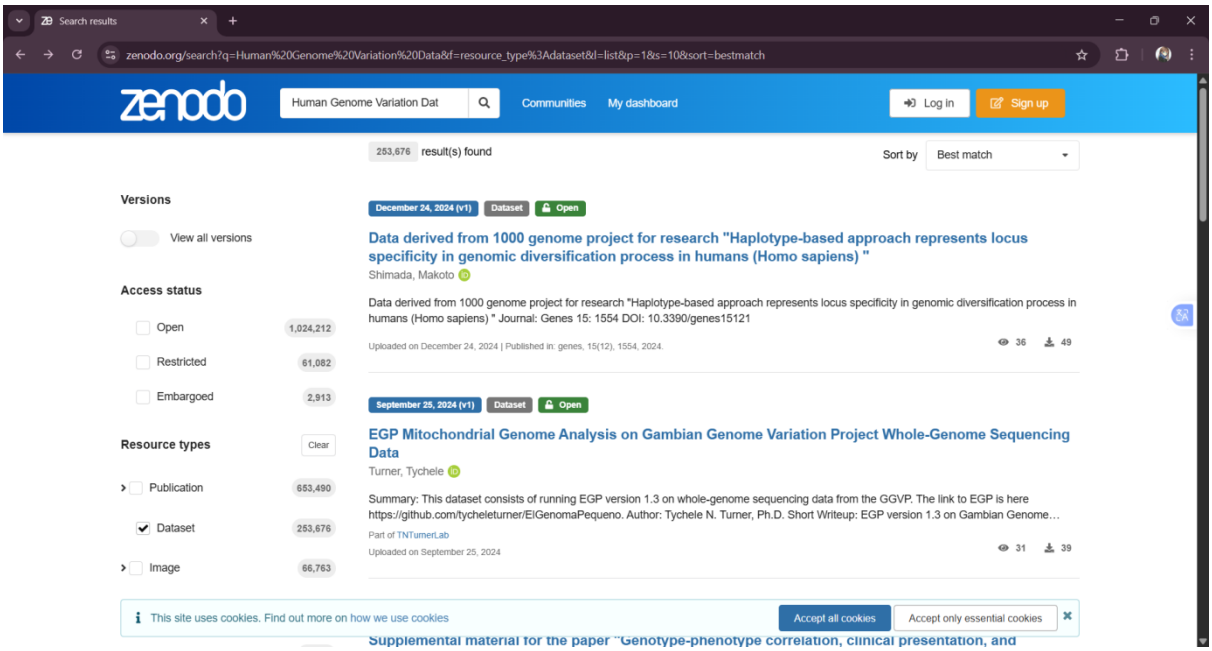

Figure 2 : Page de recherche Zenodo avec requête et filtres

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

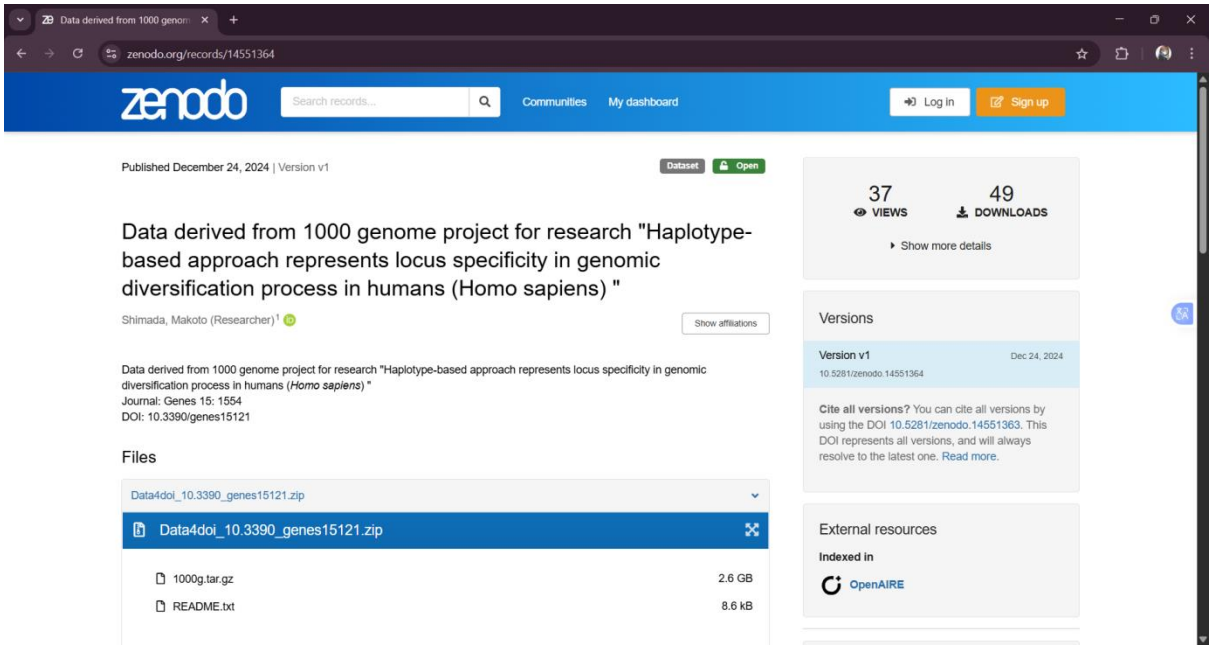

Figure 3 : Page du dataset sélectionné

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.

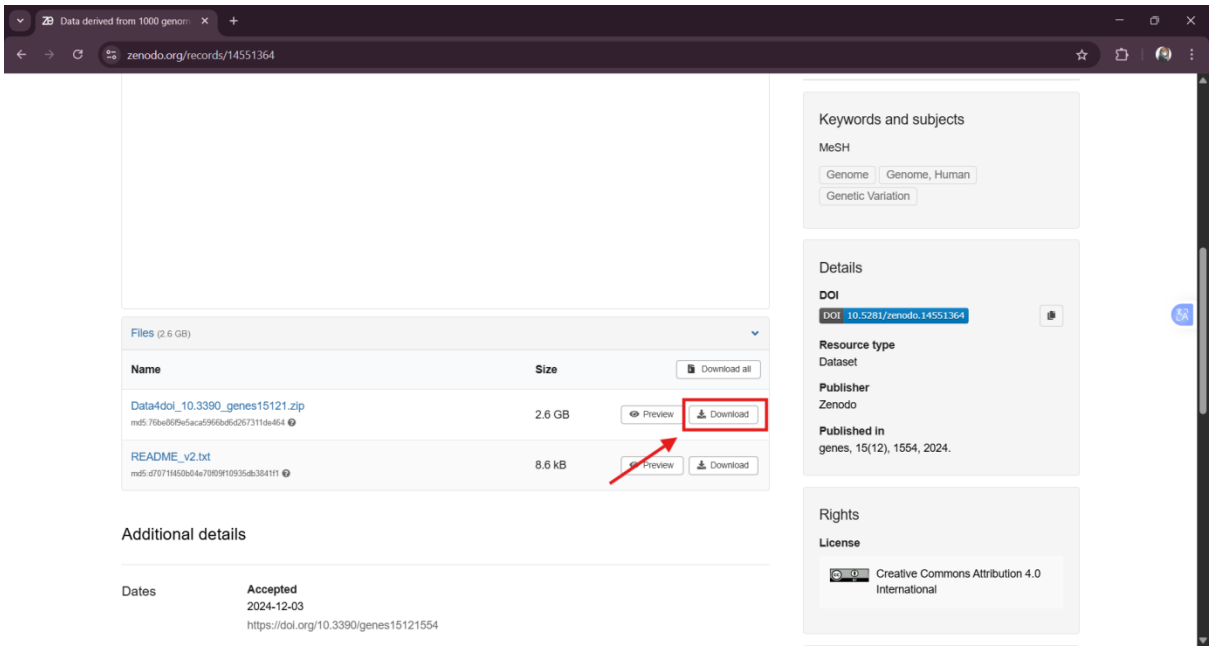

Figure 4 : Dossier téléchargé

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)