

**REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE**

**MINISTERE DE L'ENSEIGNEMENT SUPERIEUR**

**ET DE LA RECHERCHE SCIENTIFIQUE**

**UNIVERSITE DE JIJEL**

**Faculté des Sciences de la Nature et de la Vie**

**Département Biologie Moléculaire et Cellulaire**

**Toxicologie Fondamentale et Appliquée**



**TP**

## **Logiciels Libres et Open Source**

**Prépare par :**

**responsable :**

➤ **Djini rahma heiballah**

**Dr .Bensalem.A**

➤ **Boulamdaoud Dikra**

**Année universitaire : 2025/ 2026**

# Introduction

Les logiciels libres et open source ainsi que la science ouverte sont devenus des éléments essentiels de la recherche scientifique moderne, notamment dans les domaines des sciences de la nature et de la vie. Ils permettent aux chercheurs d'utiliser des outils scientifiques ouverts, de partager les résultats et les données, et de garantir la transparence ainsi que la reproductibilité des travaux.

Ce travail pratique a pour objectif de découvrir certains outils et plateformes utilisés dans le cadre de la science ouverte, à travers une étude théorique d'un logiciel libre employé en recherche scientifique et une étude appliquée visant à explorer une plateforme de dépôt de données scientifiques ouvertes. Il met également l'accent sur la compréhension des étapes de recherche, de sélection et d'analyse des jeux de données, ainsi que sur l'exploitation des métadonnées selon des normes reconnues.

Ce type de travaux contribue au renforcement des compétences des étudiants dans l'utilisation des outils numériques ouverts et à la compréhension de leur rôle dans le développement de la recherche scientifique et la facilitation de l'accès à la connaissance.(1)

# Partie 1 : Étude théorique

## 1. Présentation générale de Biopython :

Biopython est une bibliothèque open source écrite principalement en Python, destinée à faciliter le développement d'applications en bioinformatique et en biologie computationnelle.

Elle fournit un ensemble riche de modules permettant la manipulation, l'analyse et le traitement de données biologiques telles que les séquences ADN/ARN, les séquences protéiques, les structures moléculaires, ainsi que l'accès à des bases de données biologiques publiques.

Le projet Biopython est développé et maintenu par une communauté internationale de chercheurs et de développeurs, et s'inscrit pleinement dans la philosophie des logiciels libres et de la science ouverte. Il est largement utilisé dans les domaines de la génomique, de la protéomique, de l'évolution moléculaire et de l'analyse de données biologiques à grande échelle.(2)

**2. Fonctionnalités principales :** Biopython propose un large éventail de fonctionnalités essentielles pour l'analyse des données biologiques :

**Manipulation des séquences biologiques :** Lecture, écriture et analyse de séquences d'ADN, d'ARN et de protéines, avec la prise en charge de formats standards tels que FASTA, GenBank et EMBL.

**Alignement de séquences :** Interface permettant l'utilisation d'outils d'alignement largement répandus, notamment BLAST, ClustalW et MUSCLE.

**Accès aux bases de données biologiques :** Connexion aux bases de données du NCBI, telles que GenBank, PubMed et BLAST, via des services en ligne intégrés.

**Analyse phylogénétique :** Construction, analyse et visualisation d'arbres phylogénétiques afin d'étudier les relations évolutives entre les organismes.

**Analyse des structures biomoléculaires :** Lecture et exploitation des fichiers PDB pour l'étude des structures tridimensionnelles des protéines.

**Automatisation du traitement des données :** Outil particulièrement adapté à la mise en place de pipelines d'analyse et au traitement de grands volumes de données biologiques.(3)

## 3. Aspects techniques :

**Langage de programmation :** Python

**Type de logiciel :** Bibliothèque (library)

**Licence :** Biopython License (compatible open source)

**Systèmes d'exploitation :** Windows, Linux, macOS

**Interface :** Ligne de commande / scripts Python

**Dépendances :** Python ( $\geq 3.x$ ), bibliothèques scientifiques (NumPy, etc.)

Biopython est conçu pour être facilement intégrable dans des scripts Python, ce qui le rend très flexible et adapté à l'automatisation des analyses bioinformatiques. (4)

#### 4. Points forts :

- Logiciel libre et gratuit
- Forte communauté scientifique
- Documentation riche et bien structurée
- Compatible avec de nombreuses bases de données biologiques
- Très utilisé dans la recherche académique
- Facilite l'automatisation des analyses complexes
- Intégration facile avec d'autres bibliothèques Python (NumPy, Pandas, Matplotlib).(5)

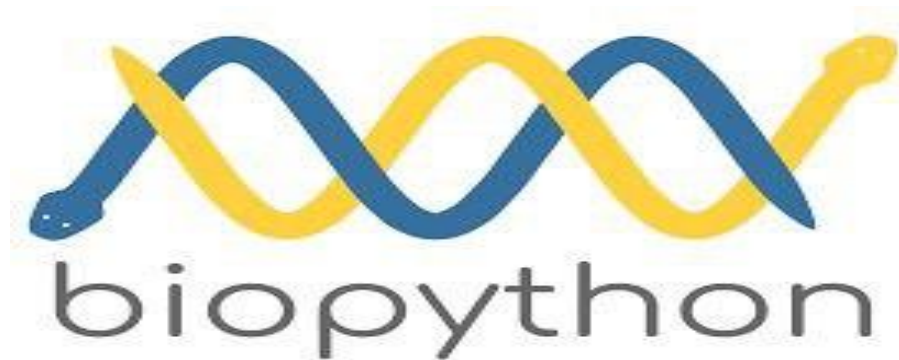
#### 5. Limites et points faibles :

- Nécessite des connaissances en programmation Python
- Pas d'interface graphique native pour les utilisateurs débutants
- Certaines analyses complexes dépendent d'outils externes
- Peut être moins intuitif pour les biologistes sans formation en informatique(6)

#### 6. Conclusion

Biopython est un outil puissant et flexible, largement adopté dans le domaine de la bioinformatique. Grâce à son caractère open source, sa richesse fonctionnelle et sa compatibilité avec les bases de données biologiques majeures, il constitue une solution de choix pour l'analyse et le traitement des données en sciences de la vie. Malgré la nécessité de compétences en programmation, Biopython reste un outil incontournable pour la recherche moderne en biologie computationnelle.(7)

## Illustrations et exemples d'utilisation :



**Figure 1 :** Logo officiel du projet Biopython, bibliothèque open source pour la bioinformatique en Python.

The main function is `Bio.SeqIO.parse(...)` which takes an input file handle (or in recent versions of Biopython alternatively a filename as a string), and format string. This returns an iterator giving `SeqRecord` objects:

```
>>> from Bio import SeqIO
>>> for record in SeqIO.parse("Fasta/f002
...     print("%s %i" % (record.id, len(r
gi|1348912|gb|G26680|G26680 633
gi|1348917|gb|G26685|G26685 413
gi|1592936|gb|G29385|G29385 471
```

**Figure2 :** Exemple d'utilisation de Biopython pour la lecture et l'analyse d'une séquence biologique.

La fonction `SeqIO.parse()` de la bibliothèque Biopython est utilisée pour lire des fichiers de séquences biologiques (comme FASTA).

Elle prend le nom du fichier et le format, puis retourne un itérateur de type `SeqRecord`.

Chaque `SeqRecord` contient l'identifiant de la séquence et sa longueur.

Dans l'exemple, on parcourt chaque séquence et on affiche son ID et sa longueur.

## **PARTIE 2 : Étude pratique : exploration de Zenodo**

### **1. Présentation de Zenodo :**

#### **a) Qu'est-ce que Zenodo ?**

Zenodo est une plateforme en ligne d'archive et de diffusion en libre accès, créée en 2013 par le CERN (Organisation européenne pour la recherche nucléaire) en collaboration avec le programme européen OpenAIRE.

Elle permet aux chercheurs, enseignants et institutions de déposer, partager et préserver différents types de productions scientifiques.

Zenodo attribue à chaque dépôt un DOI (Digital Object Identifier), ce qui garantit une citation stable et une conservation à long terme des données.(8)

#### **b) Objectifs de Zenodo :**

Les principaux objectifs de la plateforme Zenodo sont :

- Favoriser l'accès libre (Open Access) aux résultats de la recherche scientifique.
- Permettre le partage et la conservation à long terme des données de recherche.
- Améliorer la visibilité, la réutilisation et la citation des productions scientifiques.
- Soutenir les principes de la science ouverte (Open Science).
- Offrir une plateforme gratuite et fiable, indépendante des éditeurs commerciaux.(9)

#### **c) Types de contenus hébergés sur Zenodo :**

Zenodo accepte et héberge une grande variété de contenus scientifiques, notamment :

- Datasets (données expérimentales, biologiques, génomiques, environnementales, etc.)

- Articles scientifiques (préprints, postprints)
- Logiciels et codes sources
- Rapports techniques
- Présentations et posters scientifiques
- Images, vidéos et fichiers multimédias
- Thèses et mémoires
- Matériel pédagogique
- La plateforme accepte pratiquement tous les formats de fichiers et permet des dépôts jusqu'à plusieurs dizaines de Go.(10)

#### **d) Importance de Zenodo pour la science ouverte :**

Zenodo joue un rôle essentiel dans le développement de la science ouverte (Open Science), car elle :

- Garantit un accès libre et gratuit aux données et publications scientifiques.
- Facilite la reproductibilité des recherches grâce au partage des données et du code
- Encourage la transparence scientifique.
- Permet la réutilisation des données par d'autres chercheurs.
- Respecte les principes FAIR (Findable, Accessible, Interoperable, Reusable).
- Grâce à Zenodo, les chercheurs peuvent diffuser leurs travaux sans barrière financière, contribuant ainsi à une science plus collaborative et plus équitable.(11)

## **2. Description des étapes réalisées : (12)**

### **a. Recherche effectuée :**

Une recherche a été réalisée sur la plateforme Zenodo en utilisant la barre de recherche disponible sur la page d'accueil.

Le mot-clé « genome » a été employé afin d'identifier des jeux de données en lien avec le domaine des sciences de la vie.

Cette recherche a permis d'obtenir plusieurs résultats correspondant à différents datasets accessibles sur la plateforme.

### **b. Critères de sélection du dataset :**

Le choix du dataset s'est basé sur plusieurs critères, notamment sa pertinence scientifique par rapport au thème étudié, la clarté et la précision de son titre et de sa description, la disponibilité et la complétude des métadonnées associées, ainsi que la possibilité de télécharger les fichiers.

Le dataset retenu contient des données génomiques bien documentées, ce qui le rend adapté à une exploitation scientifique.

Après la sélection du jeu de données jugé pertinent, le dataset a été téléchargé directement depuis la plateforme Zenodo à l'aide de l'option de téléchargement disponible sur la page du dépôt. Les fichiers associés ont ensuite été consultés afin d'identifier leur nature, leur format et leur intérêt scientifique.

### **c. Navigation sur la plateforme Zenodo :**

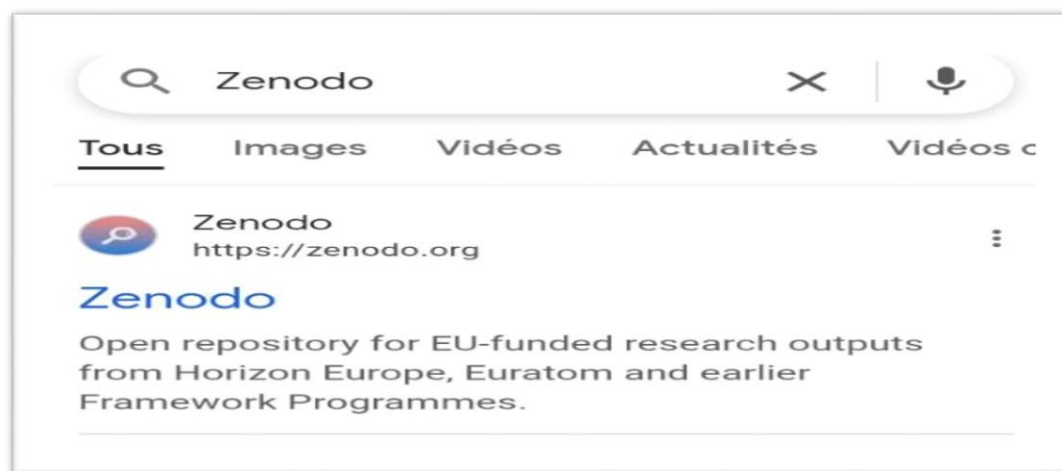
La navigation sur la plateforme Zenodo est simple et intuitive. À partir de la page d'accueil, l'utilisateur peut accéder rapidement à la barre de recherche.

Chaque résultat de recherche renvoie vers une page détaillée présentant la description du dataset, les métadonnées, les fichiers téléchargeables ainsi que le DOI.

Cette organisation facilite l'accès aux informations et permet une compréhension claire et structurée du contenu scientifique.

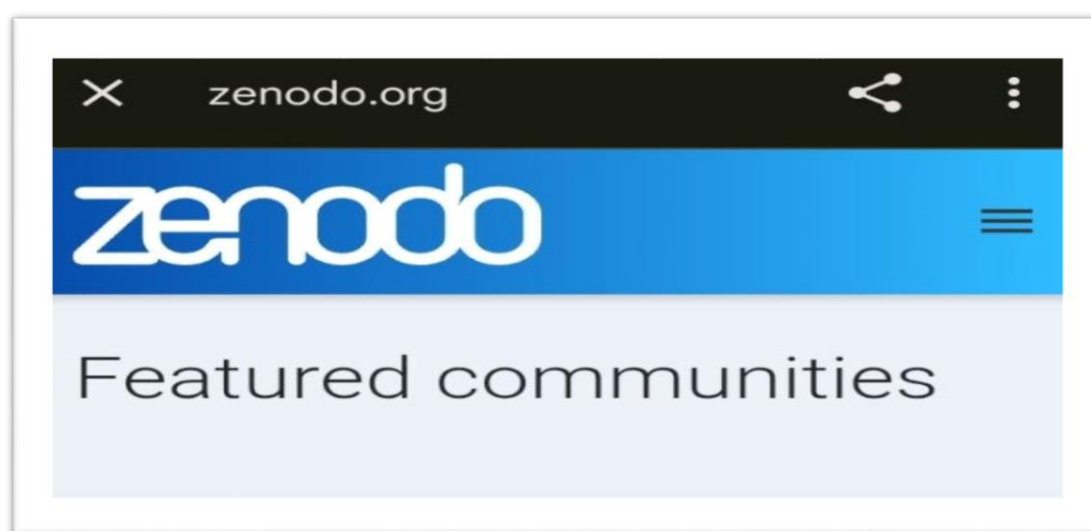


- Pour illustrer la navigation sur la plateforme Zenodo, la figure suivante est présentée :



**Figure 3 :** Page d'accueil de la plateforme Zenodo.

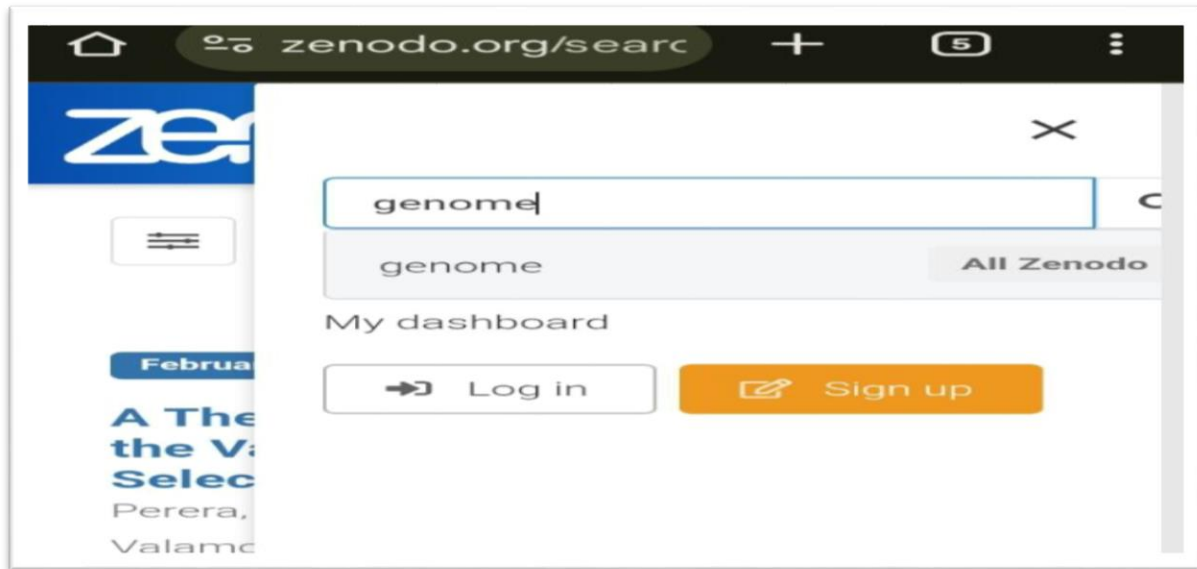
**Description :** Zenodo est une plateforme en libre accès permettant le partage et l'archivage des productions scientifiques telles que les jeux de données, les articles, les logiciels et les rapports de recherche. Elle facilite la diffusion des travaux scientifiques et l'attribution d'un identifiant DOI.



**Figure 4 :** Section « Featured communities » de la plateforme Zenodo.

**Description :** Cette figure présente la section Featured communities de Zenodo, qui regroupe des communautés scientifiques sélectionnées organisant les productions de

recherche par domaine ou par projet. Cette section facilite l'accès aux contenus scientifiques et permet aux chercheurs de partager leurs travaux au sein de communautés spécialisées.



**Figure 5 :** Exemple de recherche d'un jeu de données sur la plateforme Zenodo.

**Description :** Cette figure illustre l'utilisation de la barre de recherche de Zenodo avec le mot-clé « genome » afin de rechercher des jeux de données et des ressources scientifiques liés à la génomique. La plateforme permet de filtrer et d'accéder facilement aux résultats correspondant au domaine de recherche choisi.



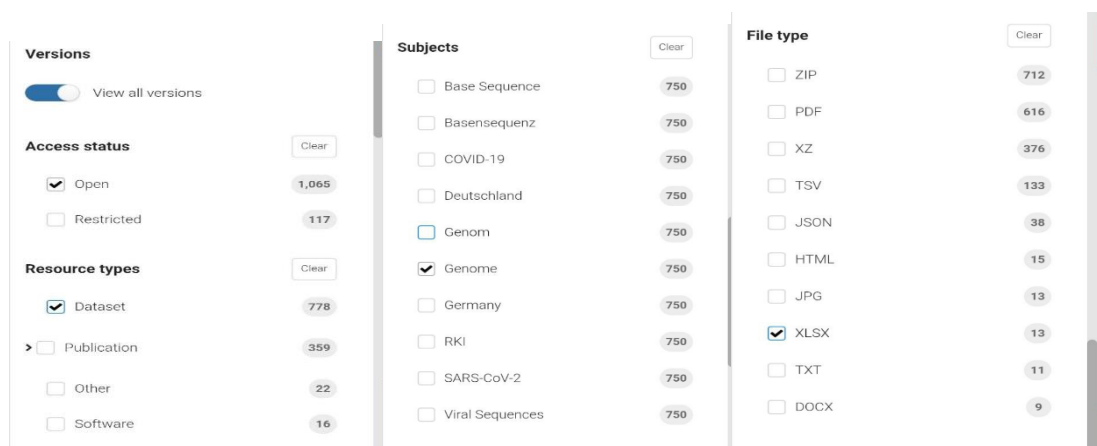
**Figure 6 :** Exemple de résultat de recherche sur la plateforme Zenodo.

**Description :** Cette figure présente un exemple de jeu de données scientifique trouvé sur la plateforme Zenodo à la suite d'une recherche par mot-clé. Le résultat affiché correspond à un dataset en accès libre, accompagné de ses métadonnées principales telles que le titre, les auteurs, la date de publication et le type de ressource. Cela illustre la facilité d'accès aux données scientifiques partagées via Zenodo.



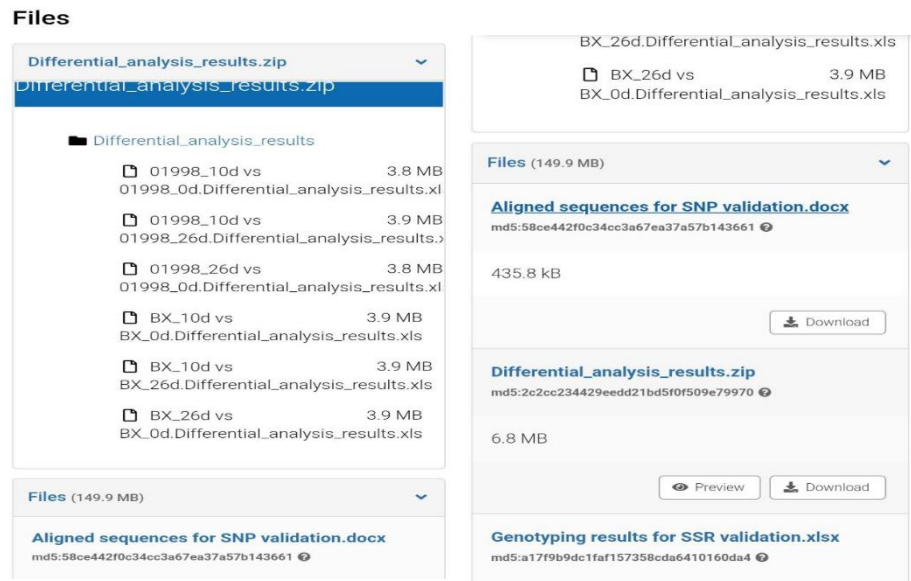
**Figure07 :** Page d'un jeu de données scientifique sur la plateforme Zenodo.

**Description :** Cette figure montre la page d'un jeu de données publié sur Zenodo, comprenant le titre de l'étude, la date de publication, le type de ressource (Dataset) et le statut d'accès libre. Elle illustre la manière dont les données scientifiques sont mises à disposition des chercheurs pour être réutilisées dans la recherche.



**Figure 08 :** Options de filtrage des résultats sur la plateforme Zenodo.

**Description :** Cette figure présente les filtres de recherche disponibles sur la plateforme Zenodo. Elle permet de sélectionner les résultats selon le type de fichier (ZIP, PDF, XLSX, etc.), les thématiques scientifiques (génomique, COVID-19, séquences virales), le statut d'accès (ouvert ou restreint), ainsi que le type de ressource (jeu de données, publication, logiciel). Ces filtres facilitent l'identification rapide et précise des ressources scientifiques pertinentes selon les besoins du chercheur.



**Figure 09 :** Fichiers associés aux résultats d'analyses génomiques sur la plateforme Zenodo .

**Description :** Cette figure présente l'ensemble des fichiers mis à disposition dans une ressource scientifique déposée sur la plateforme Zenodo, relatifs aux résultats d'analyses différentielles de l'expression génique et à la validation de marqueurs moléculaires.

### ➤ Les fichiers comprennent :

#### ✓ Résultats d'analyses différentielles (Differential analysis results) :

Des fichiers au format XLS, regroupés dans une archive ZIP, contenant des comparaisons entre différents échantillons et différents temps d'échantillonnage (par exemple 0 jour, 10 jours et 26 jours) ou entre différentes lignées (telles que BX et

01998). Ces données permettent d'identifier les gènes dont l'expression varie significativement entre les conditions étudiées.

✓ **Séquences alignées pour la validation des SNP (Aligned sequences for SNP validation) :**

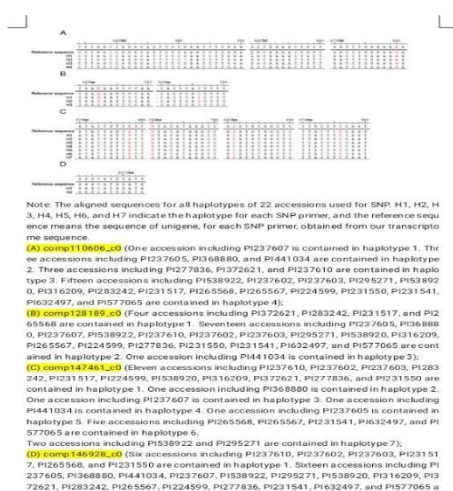
Un fichier au format DOCX présentant l'alignement des séquences nucléotidiques, utilisé pour confirmer la présence et la fiabilité des polymorphismes nucléotidiques simples (SNP).

✓ **Résultats de génotypage pour la validation des SSR (Genotyping results for SSR validation) :**

Un fichier au format XLSX contenant les données de génotypage des marqueurs SSR, permettant l'évaluation de la diversité génétique et la confirmation des variations observées entre les échantillons.

Les tailles des fichiers ainsi que les valeurs de contrôle MD5 sont indiquées afin de garantir l'intégrité des données mises à disposition.

✓ **Intérêt scientifique :** L'ensemble de ces fichiers fournit des données structurées et exploitables favorisant la reproductibilité des analyses, la vérification des résultats et la réutilisation des données dans de futurs travaux en génomique, conformément aux principes de la science ouverte.



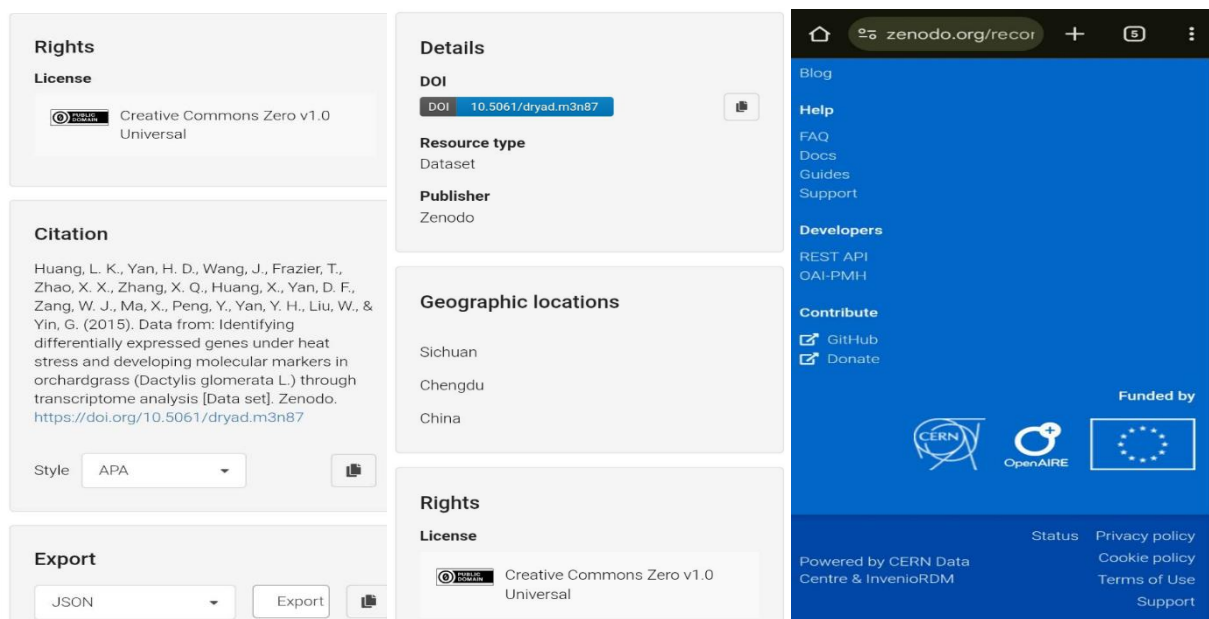
**Figure 10 :** Alignement des séquences nucléotidiques illustrant la validation et la distribution des haplotypes SNPs chez 22 accessions de *Dactylis glomerata*.

**Description :** Cette figure illustre l'alignement de séquences nucléotidiques utilisé pour la validation de SNPs (Single Nucleotide Polymorphisms) chez 22 accessions de *Dactylis glomerata*.

La séquence de référence correspond à la séquence du unigène obtenue à partir de l'analyse du transcriptome, tandis que les séquences H1 à H7 représentent les différents haplotypes identifiés. Les nucléotides en rouge indiquent les positions polymorphes (SNPs).

Quatre loci SNP distincts sont analysés : comp110606\_c0 (A), comp128189\_c0 (B), comp147461\_c0 (C) et comp146928\_c0 (D). Pour chaque locus, les accessions sont regroupées en haplotypes selon leur profil de variation nucléotidique. Le nombre d'haplotypes et la distribution des accessions varient d'un locus à l'autre, révélant des niveaux différents de diversité génétique.

Ces résultats confirment la fiabilité des SNPs identifiés et leur pertinence comme marqueurs moléculaires, utiles pour l'étude de la variabilité génétique et pour le développement de marqueurs associés à la réponse au stress thermique.



**Figure 11 :** Présentation des métadonnées, du cadre institutionnel et des informations de diffusion du jeu de données sur la plateforme Zenodo.

**Description :** Ces sections montrent que la plateforme Zenodo est scientifiquement fiable, bénéficiant d'un soutien technique du CERN, étant liée à l'initiative OpenAIRE, financée par l'Union européenne, et reposant sur le système InvenioRDM pour la gestion des données de recherche.

Les informations du jeu de données comprennent des éléments essentiels tels que l'identifiant numérique permanent (DOI), le type de ressource (Dataset), l'éditeur, ainsi que la localisation géographique de la collecte des données, un facteur important pour l'interprétation scientifique des résultats.

La plateforme propose également une licence ouverte (CC0) permettant la réutilisation des données sans restriction, ainsi qu'un mode de citation scientifique clair et des options d'export des métadonnées, renforçant ainsi la transparence, la réutilisabilité et la crédibilité de la recherche scientifique.



### 3. Métadonnées du dataset :

Les métadonnées du jeu de données ont été récupérées et organisées conformément à la norme Dublin Core, permettant une description structurée et standardisée des informations associées au dataset. **(13)**

**Tableau01** : des métadonnées du dataset.

<b>Élément (Metadata)</b>	<b>Description</b>
<b>Titre</b>	Data from: Identifying differentially expressed genes under heat stress and developing molecular markers in orchardgrass ( <i>Dactylis glomerata</i> L.) through transcriptome analysis
<b>Auteur(s)</b>	Huang, Lin K.; Yan, Hai D.; Wang, J.; Frazier, T.; Zhao, X. X.; Zhang, X. Q.; et al.
<b>Date de publication</b>	21 avril 2015
<b>Type de ressource</b>	Jeu de données (Dataset)
<b>Version</b>	v1
<b>Éditeur / Dépôt</b>	Zenodo (Dryad)
<b>Accès</b>	Accès libre (Open access)
<b>Domaine / Sujet</b>	Analyse du transcriptome, stress thermique, gènes différentiellement exprimés, marqueurs moléculaires, génomique végétale .
<b>Description</b>	Jeu de données issu d'une analyse transcriptomique visant l'identification des gènes différentiellement exprimés sous stress thermique chez l'orchardgrass ( <i>Dactylis glomerata</i> L.), ainsi que le développement de marqueurs moléculaires.



<b>Formats des fichiers</b>	ZIP, XLS/XLSX, DOCX
<b>Langue</b>	Anglais
<b>Espèce étudiée</b> <b>Licence</b>	Dactylis glomerata L. Accès libre (via Zenodo / Dryad)

## 4. Pourquoi ce dataset a été choisi ?

Ce dataset a été choisi car il fournit des données transcriptomiques complètes et fiables permettant d'étudier l'expression différentielle des gènes en réponse au stress thermique chez une espèce végétale d'intérêt agronomique, *Dactylis glomerata* L..

Il est en libre accès, bien documenté et contient des fichiers exploitables (résultats d'analyses différentielles, validation SNP et SSR), ce qui le rend particulièrement adapté à l'apprentissage de l'exploration et de l'analyse de données génomiques dans le cadre académique.(14)

## 5 .Conclusion :

Cette étude pratique a permis de comprendre le fonctionnement de la plateforme Zenodo et son rôle dans la science ouverte. L'exploration d'un jeu de données scientifique a mis en évidence l'importance des métadonnées, du DOI et du libre accès pour assurer la visibilité, la réutilisation et la reproductibilité des travaux de recherche. Ainsi, Zenodo constitue un outil fiable et essentiel pour le partage et la valorisation des données scientifiques. (15)

## Conclusion générale

Ce travail a permis de mettre en évidence l'importance des logiciels libres et des plateformes de science ouverte dans la recherche scientifique moderne. L'étude théorique de Biopython a montré le rôle essentiel des outils open source dans l'analyse et le traitement des données biologiques, ainsi que leur contribution à la reproductibilité et à l'automatisation des analyses en bioinformatique.

Par ailleurs, l'étude pratique de la plateforme Zenodo a permis de comprendre les principes du libre accès, l'importance des métadonnées et du DOI, ainsi que le rôle du partage des données dans la valorisation et la transparence des travaux scientifiques.

Ainsi, l'utilisation conjointe des logiciels libres et des plateformes ouvertes constitue un pilier fondamental pour le développement de la science ouverte et le progrès de la recherche scientifique. **(16)**

## Référence :

- (1) European Commission. Open Science. Publications Office of the European Union, 2016.
- (2) Cock, P. J. A., Antao, T., Chang, J. T., et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422–1423, 2009. DOI: 10.1093/bioinformatics/btp163.
- (3) Biopython Project. Biopython Tutorial and Cookbook. Official documentation. <https://biopython.org/wiki/Documentation>.
- (4) Cock, P. J. A., et al. The Biopython project: Python tools for bioinformatics. *Briefings in Bioinformatics*, 11(4), 435–439, 2010. DOI: 10.1093/bib/bbp010.
- (5) Hinsén, K. Computational science: shifting the focus from tools to models. *F1000Research*, 3:101, 2014.
- (6) Perez-Riverol, Y., et al. Ten Simple Rules for Taking Advantage of Git and GitHub. *PLoS Computational Biology*, 12(7), e1004947, 2016.
- (7) Chapman, B., & Chang, J. Biopython and the Python ecosystem for bioinformatics. *Briefings in Bioinformatics*, 11(4), 435–439, 2010.
- (8) Zenodo. About Zenodo. CERN & OpenAIRE, 2013. <https://zenodo.org>.
- (9) About Zenodo. CERN & OpenAIRE, 2013. Nielsen, L. H., et al. OpenAIRE Report, 2014.
- (9) <https://zenodo.org>.
- (10) OpenAIRE. Guidelines for Open Science. European Commission, 2018.
- (11) Zenodo Documentation. Supported content and formats. CERN, 2023.

**(12)** The FAIR Guiding Principles for scientific data management and stewardship.

Scientific Data, 3:160018, 2016.

**(13)** Tenopir, C., et al. Data sharing practices and perceptions among scientists.

PLoS ONE, 6(6), e21101, 2011.

**(14)** Huang, L. K., Yan, H. D., Wang, J., et al. Identifying differentially expressed genes under heat stress and developing molecular markers in orchardgrass (*Dactylis glomerata* L.) through transcriptome analysis. BMC Plant Biology, 15:131, 2015. DOI: 10.1186/s12870-015-0504-2.

**(15)** Wilkinson, M. D., et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3:160018, 2016.

DOI: 10.1038/sdata.2016.18.

**(16)** Logiciels libres & Biopython Cock, P. J. A., Antao, T., Chang, J. T., et al.

Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422–1423, 2009.

DOI: 10.1093/bioinformatics/btp163

**(16)** Science ouverte, Zenodo, métadonnées et DOI Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3:160018, 2016. DOI: 10.1038/sdata.2016.18.