Université Mohamed Seddik Ben Yahia, Jijel

Faculté des Sciences de la Nature ET de la Vie

Département des sciences de l'environnement ET des

Module : Logiciel libre and open source

Rappot TP :

Assignment

Réaliser par :

Touhami djannat el khold

Lamia Hameurlaine

Foul Inès

- # PART 1: COMPREHENSIVE THEORETICAL STUDY OF BIOPYTHON

1. . General Presentation of the Tool :

Biopython is a global, open-source collaborative project that provides a comprehensive collection of Python libraries dedicated to computational biology and bioinformatics. First released in 1999, Biopython was developed to reduce duplication of effort among researchers by offering standardized, well-tested, and reusable code for biological data analysis.

As a free and open-source software distributed under the Biopython License Agreement, it ensures transparency of algorithms and reproducibility of scientific analyses, which are fundamental principles in modern research. Its open nature allows continuous improvement by an international community of developers, researchers, and educators.

Biopython is particularly valued in life sciences and natural sciences because it enables researchers to automate complex data analysis workflows, interact directly with biological databases, and process large-scale datasets efficiently. Within the context of a Master's program in Phytopharmacie, it represents a bridge between biological sciences and computational methods, facilitating advanced research in genomics, molecular biology, and plant protection.

2. Main Functionalities (Detailed Analysis)

Biopython offers a wide range of advanced functionalities that support multiple domains of bioinformatics:

2.1 Sequence Analysis and Manipulation

The core of Biopython lies in its ability to handle biological sequences. Using modules such as Bio.Seq and Bio.SeqRecord, sequences are treated as structured objects rather than simple strings. This object-oriented design allows:

DNA transcription (DNA → RNA)

Translation (DNA/RNA → Protein)

Reverse complement generation

Sequence slicing and annotation

This approach ensures both biological accuracy and computational efficiency.

2.2 Complex File Parsing and Format Conversion (Bio.SeqIO)

The Bio.SeqIO module functions as a universal parser and converter for biological file formats. It supports numerous formats including:

FASTA

GenBank

EMBL

GFF

PDB

For example, a GenBank file rich in annotations can be converted into FASTA format using only a few lines of Python code. This capability is essential when integrating data from heterogeneous sources in large-scale bioinformatics projects.

2.3 Programmatic Access to Biological Databases (Bio.Entrez)

Biopython provides direct access to NCBI databases via the Bio.Entrez module. Researchers can:

Query GenBank

Retrieve PubMed articles

Download nucleotide or protein sequences

Access taxonomy data

This automated access is crucial for Big Data projects where thousands of records must be retrieved and processed without manual intervention.

2.4 Sequence Alignment and Comparative Analysis

Biopython supports both pairwise and multiple sequence alignments. It includes:

Internal alignment tools

Interfaces to external programs such as BLAST and Clustal

These tools are essential for identifying conserved regions, studying mutations, and performing comparative genomics.

2.5 Phylogenetic Analysis (Bio.Phylo)

The Bio.Phylo module enables:

Reading phylogenetic trees (Newick, Nexus formats)

Tree visualization

Evolutionary relationship analysis

This is particularly important in evolutionary biology and plant pathology studies where understanding common ancestry and species divergence is required.

2.6 Structural Bioinformatics (Bio.PDB)

The Bio.PDB module allows manipulation and analysis of 3D macromolecular structures obtained from Protein Data Bank (PDB) files. It can:

Calculate distances between atoms

Identify residue contacts

Analyze protein domains

Assist in protein folding studies

Such functionalities are essential in molecular modeling and drug-target interaction studies.

2.7 Motif and Pattern Analysis

Biopython also supports the identification of conserved motifs and biological patterns, contributing to functional genomics and regulatory sequence analysis.

3. Technical Aspects

Biopython is written in Python, a high-level, readable, and widely adopted programming language. It is compatible with major operating systems including Windows, Linux, and macOS.

Key Technical Characteristics:

Installation via pip or conda

Modular architecture (Bio.Align, Bio.Blast, Bio.Data, etc.)

Strong integration with scientific libraries such as NumPy, SciPy, and Matplotlib

Optimization of computationally intensive tasks using NumPy arrays

Compatibility with interactive environments like Jupyter Notebook

Its modular architecture ensures that only required components are loaded, optimizing memory usage.

4. Strengths

Biopython presents several major advantages:

Open-source and cost-free, making it accessible to students and universities with limited budgets.

Extensive documentation and community support, facilitating learning and troubleshooting.

Automation capabilities, allowing the creation of bioinformatics pipelines where data is downloaded, cleaned, analyzed, and visualized automatically.

Reproducibility, thanks to transparent algorithms and script-based workflows.

Flexibility and extensibility, enabling integration with other computational tools.

These strengths explain why Biopython is one of the most widely used bioinformatics libraries worldwide.

5. Limitations and Weaknesses

Despite its many advantages, Biopython has certain limitations:

Programming requirement: It requires knowledge of Python, which may represent a barrier for students without coding background.

Performance limitations: For very large-scale genomic datasets (e.g., Next-Generation Sequencing with billions of reads), lower-level languages like C++ or Rust may offer better performance.

Dependence on external tools: Some advanced analyses require integration with external software (e.g., BLAST executables).

Limited graphical interface: It is primarily script-based and does not provide extensive graphical user interfaces.

However, these limitations can often be mitigated by combining Biopython with other specialized software tools.

6. Conclusion

Biopython is a powerful and versatile bioinformatics toolkit that plays a central role in modern computational biology. Its object-oriented design, wide range of functionalities, and integration with scientific Python libraries make it particularly suitable for sequence analysis, database interaction, phylogenetics, and structural biology.

In the context of a Master I program in Phytopharmacie at the University of Jijel, Biopython is not merely a software library but a gateway to high-level research. It equips students with the computational skills necessary to analyze biological data rigorously, automate workflows, and participate in reproducible scientific research.

- Although it requires initial investment in learning programming concepts, its long-term benefits in research efficiency and analytical power make it an indispensable tool for modern biologists

- **PART 2** – PRACTICAL STUDY: EXPLORATION OF ZENODO

1. Presentation of Zenodo

Zenodo is an open-access research repository developed under the European OpenAIRE program and operated by the CERN (European Organization for Nuclear Research). It was launched in 2013 to provide researchers from all scientific disciplines with a free and sustainable platform for sharing and preserving research outputs.

Zenodo allows researchers to upload, store, share, and cite a wide range of research materials, ensuring long-term preservation and global accessibility. The platform supports the principles of Open Science by removing access barriers and promoting transparency and collaboration in scientific research.

Reference:

CERN & OpenAIRE. Zenodo – Research. Shared. https://www.zenodo.org

2. Objectives of the Platform

The main objectives of Zenodo are:

To facilitate open sharing of research outputs by providing a free repository accessible to researchers worldwide, regardless of discipline or institution.
To ensure citability and recognition of research products by assigning a Digital Object Identifier (DOI) to each deposited item, making them permanently identifiable and citable in scientific literature.
To support Open Science infrastructures, particularly by integrating with OpenAIRE and other research indexing systems, thus increasing the visibility and impact of scientific work.
To guarantee long-term preservation of research outputs, using CERN's robust digital infrastructure to ensure data durability and accessibility over time.
These objectives align with modern research policies that encourage openness, reproducibility, and responsible data management.
References:
OpenAIRE. Zenodo Guide. https://www.openaire.eu/zenodo-guide
European Commission (2016). Open Science Policy Platform.

3. Types of Content Hosted on Zenodo

Zenodo supports a wide variety of research outputs, including:

3.1 Scientific Publications

Research articles
Technical reports
Preprints

3.2 Research Data

Experimental and observational datasets
Tables, statistics, and raw data files

3.3 Software and Source Code

Research software
Analysis scripts and computational tools

3.4 Other Research Outputs

Conference presentations and posters
Images, videos, and audio files
Theses, educational material, and documentation
Each deposited item is described using standardized metadata, which enhances discoverability and interoperability according to the FAIR principles (Findable, Accessible, Interoperable, Reusable).
Reference:
Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018.

4. Importance of Zenodo for Open Science and Life Sciences Research

Zenodo plays a crucial role in promoting Open Science, particularly in the life sciences, where data sharing and reproducibility are essential.

4.1 Open and Free Access

Zenodo provides free access to research outputs, allowing scientists, students, and institutions worldwide to consult and reuse scientific results without financial or legal barriers.

4.2 Improved Visibility and Citability

By assigning DOIs, Zenodo enables datasets, software, and supplementary materials to be cited like traditional publications, increasing academic recognition and research impact.

4.3 Long-Term Preservation

The platform ensures secure, long-term storage of scientific data, which is especially important for biological datasets that may be reused for future analyses or comparative studies.

4.4 Support for FAIR Principles

Zenodo helps researchers comply with FAIR data management requirements, which are increasingly mandatory in life sciences projects funded by international agencies.

4.5 Reproducibility and Collaboration

In life sciences research, Zenodo facilitates the sharing of datasets and analytical workflows, supporting reproducibility and enabling collaboration across institutions and countries.

References:

Mons et al. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use.

ResearchGate (2024). Zenodo: A platform for open access and sustainable digital research repository.

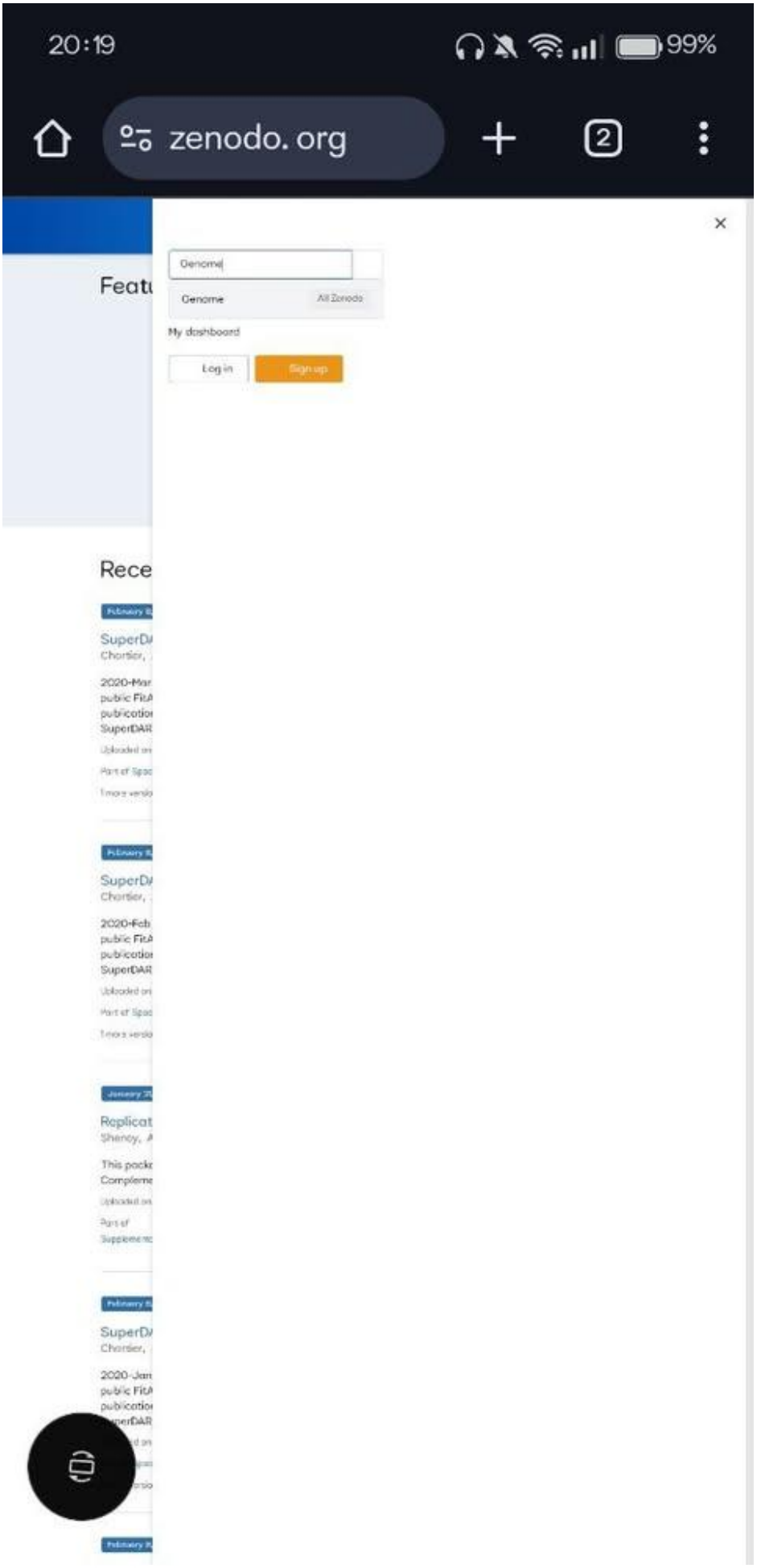
5. Conclusion

Zenodo is a key infrastructure for Open Science, providing researchers with a reliable platform to share, preserve, and cite diverse research outputs. Its role is particularly significant in the life sciences, where open data, transparency, and reproducibility are essential for scientific progress. By supporting FAIR principles and offering long-term preservation through CERN, Zenodo contributes substantially to the modernization and democratization of scientific research.

➢ Step-by-Step Guide for Finding Plant Genome Datasets on Zenodo

Step 1: Initiation of Search

Begin by accessing the Zenodo repository and typing the keyword "Genome" into the search bar to explore available scientific


20:19
99%
zenodo. org
Genome
Genome
All Zenodo
My dashboard
Log in
Sign up

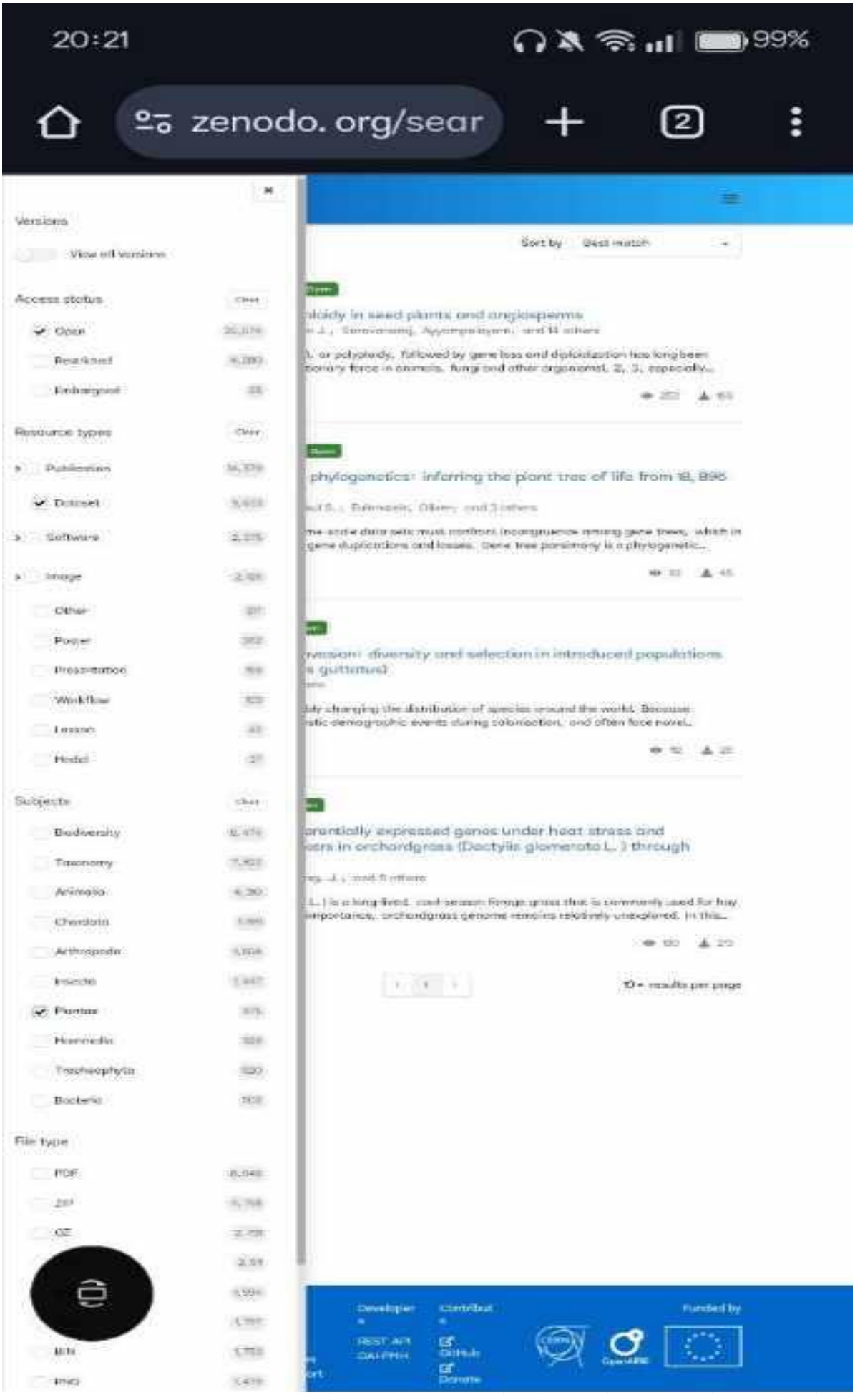
STEP 2: APPLYING SEARCH FILTERS

Refine the results by using the sidebar filters to ensure high-quality and relevant data:

Set Access Status to "Open" for unrestricted data.

Set Resource Types to "Dataset" to find raw genomic data rather than just articles.

Set Subjects to "Plantae" to narrow the focus specifically to plant biology.


20:21
99%
zenodo. org/sear
2
Versions
View all versions
Access status
Open
Restricted
Embargoed
Resource types
Publication
Dataset
Software
Image
Other
Poster
Presentation
Workflow
Lesson
Model
Subjects
Biodiversity
Taxonomy
Animalia
Chordata
Arthropoda
Insecta
Plantae
Mammalia
Tracheophyta
Bacteria
File type
PDF
ZIP
GZ
BIN
PNG
Sort by
Best match
Open
loidy in seed plants and angiosperms
L. or polyploidy, followed by gene loss and diploidization has long been
ionary force in animals, fungi and other organisms. 2, 3, especially...
Open
phylogenetics: inferring the plant tree of life from 18, 896
ne-scale data sets must confront incongruence among gene trees, which in
gene duplications and losses. Gene tree parsimony is a phylogenetic...
ression: diversity and selection in introduced populations
s guttatus)
dly changing the distribution of species around the world. Because
atic demographic events during colonization, and often face novel...
rentially expressed genes under heat stress and
ers in orchardgrass (Dactylis glomerata L.) through
L.) is a long-lived, cool-season forage grass that is commonly used for hay
mportance, orchardgrass genome resources relatively unexplored. In this...
10 results per page
Developer
Contribut
Funded by
REST API
OAI-PMH
GitHub
Donate

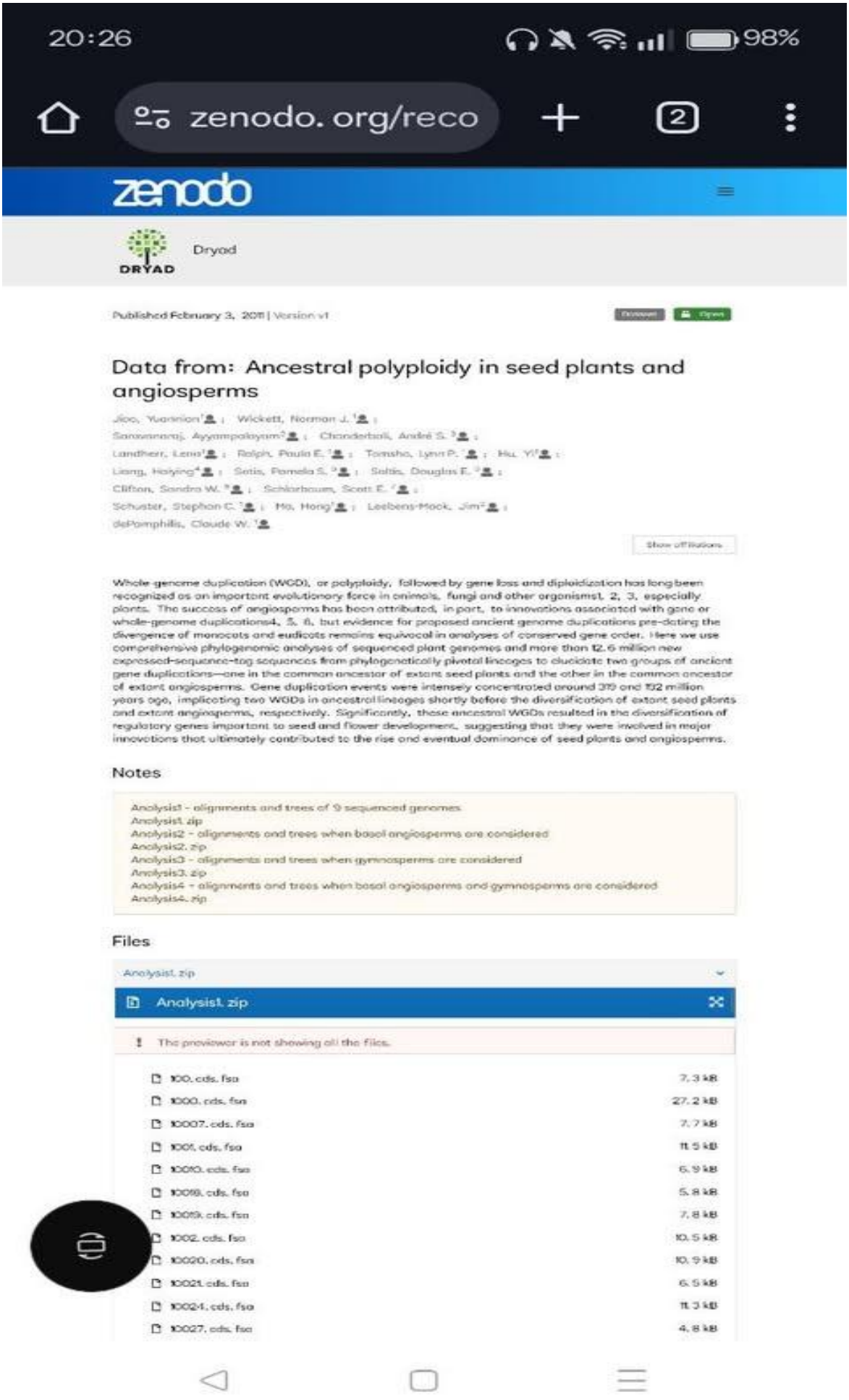
STEP 3: REVIEWING THE RESULTS

Analyze the filtered list, which includes specialized data on ancestral polyploidy, phylogenetics, and transcriptome analysis under environmental stress.


20:21
99%
zenodo.org/sear
zenodo
Sort by
Data from: Ancestral polyploidy in seed plants and angiosperms
Whole-genome duplication (WGD), or polyploidy, followed by gene loss and diploidization has long been recognized as an important evolutionary force in animals, fungi and other organisms. It, 2, especially...
Data from: Genomic-scale phylogenetics: inferring the plant tree of life from 18, 896 gene trees
Phylogenetic analyses using genome-scale data sets must confront incongruence among gene trees, which in plants is exacerbated by frequent gene duplications and losses. Gene tree parsimony is a phylogenetic...
Data from: Genomics of invasion: diversity and selection in introduced populations of monkeyflowers (Mimulus guttatus)
Global trade and travel is irreversibly changing the distribution of species around the world. Because introduced species experience drastic demographic events during colonization, and often face novel...
Data from: Identifying differentially expressed genes under heat stress and developing molecular markers in orchardgrass (Dactylis glomerata L.) through transcriptome analysis
Orchardgrass (Dactylis glomerata L.) is a long-lived, cool-season forage grass that is commonly used for hay production. Despite its economic importance, orchardgrass genome remains relatively unexplored. In this...
About
Blog
Help
Developer
Contribut
Funded by
FAQ
Donate

3. DATASET METADATA

The metadata for the selected record, "Data from: Ancestral polyploidy in seed plants and angiosperms", has been extracted following international standards like Dublin Core. This metadata provides essential information about the resource to ensure its findability and reuse


20:26
98%
zenodo.org/reco
2
zenodo
Dryad
DRYAD
Published February 3, 2011 | Version v1
Data from: Ancestral polyploidy in seed plants and angiosperms
Whole-genome duplication (WGD), or polyploidy, followed by gene loss and diploidization has long been recognized as an important evolutionary force in animals, fungi and other organisms1, 2, 3, especially plants. The success of angiosperms has been attributed, in part, to innovations associated with gene or whole-genome duplications4, 5, 6, but evidence for proposed ancient genome duplications pre-dating the divergence of monocots and eudicots remains equivocal in analyses of conserved gene order. Here we use comprehensive phylogenomic analyses of sequenced plant genomes and more than 12.6 million new expressed-sequence-tag sequences from phylogenetically pivotal lineages to elucidate two groups of ancient gene duplications—one in the common ancestor of extant seed plants and the other in the common ancestor of extant angiosperms. Gene duplication events were intensely concentrated around 319 and 192 million years ago, implicating two WGDs in ancestral lineages shortly before the diversification of extant seed plants and extant angiosperms, respectively. Significantly, these ancestral WGDs resulted in the diversification of regulatory genes important to seed and flower development, suggesting that they were involved in major innovations that ultimately contributed to the rise and eventual dominance of seed plants and angiosperms.
Notes
Analysis1 - alignments and trees of 9 sequenced genomes
Analysis1.zip
Analysis2 - alignments and trees when basal angiosperms are considered
Analysis2.zip
Analysis3 - alignments and trees when gymnosperms are considered
Analysis3.zip
Analysis4 - alignments and trees when basal angiosperms and gymnosperms are considered
Analysis4.zip
Files
Analysis1.zip
The previewer is not showing all the files.
100.cds.fsa 7.3 kB
1000.cds.fsa 27.2 kB
10007.cds.fsa 7.7 kB
1001.cds.fsa 11.5 kB
10010.cds.fsa 6.9 kB
10018.cds.fsa 5.8 kB
10019.cds.fsa 7.8 kB
1002.cds.fsa 10.5 kB
10020.cds.fsa 10.9 kB
10021.cds.fsa 6.5 kB
10024.cds.fsa 11.3 kB
10027.cds.fsa 4.8 kB

- **DATASET METADATA DESCRIPTION**
- **The dataset is titled "Data from: Ancestral polyploidy in seed plants and angiosperms".**
- CREATORS: It was authored by Yuannian Jiao, Norman J. Wickett, Ayyampalayam Saravanaraj, and 18 other contributors.
- Publication Date: The resource was published on February 3, 2011, as Version v1.
- REPOSITORY: The data is hosted on Zenodo via the Dryad Digital Repository.
- RESOURCE TYPE: This record is classified as a Dataset.
- ACCESS RIGHT: The dataset is Open Access under the Creative Commons Zero v1.0 Universal license.
- DIGITAL OBJECT IDENTIFIER (DOI): The unique identifier for this dataset is 10.5061/dryad.7995.
- DESCRIPTION (ABSTRACT): This dataset explores ancestral genome duplication (polyploidy) in seed plants and angiosperms using phylogenomic analyses of sequenced genomes.

- **KEYWORDS:** Key terms associated with this data include genome duplication, polyploidy, angiosperms, phylogenomics, and seed plants.
- **FILES CONTAINED:** The dataset includes multiple ZIP archives, such as Analysis1.zip and Analysis2.zip, which contain sequence files like .cds.fsa.

EXPLANATION OF METADATA STANDARD (DUBLIN CORE):

The metadata presented above follows the Dublin Core standard, which is a set of 15 core elements used to describe digital and physical resources. In this case:

Identifier: Represented by the DOI.

Rights: Indicated as "Open Access".

Format: Shown as ZIP files containing genomic sequence data.