# BioDataHub: An Integrated VS Code Extension for Streamlined Bioinformatics Dataset Analysis and Visualization

**Mubashir Ali**

mubashir.42413043@ncb.qau.edu.pk

Quaid-i-Azam University, Islamabad https://orcid.org/0009-0006-0222-7585

**Research Article**

**Keywords:** Bioinformatics, Dataset Analysis, CSV Data Management, Data Visualization, VS Code Extension, Metadata Generation, RNA-seq, Microarray, Workflow Optimization, Computational Biology

**Posted Date:** October 16th, 2025

**DOI:** https://doi.org/10.21203/rs.3.rs-7861003/v1

**Additional Declarations:** The authors declare no competing interests.

# Abstract

Managing and analyzing large-scale bioinformatics datasets often requires multiple tools and complex workflows, leading to inefficiencies and potential errors. Here, we present BioDataHub, a Visual Studio Code extension designed to streamline dataset discovery, management, visu- alization, and analysis for bioinformatics researchers. BioDataHub integrates local and online dataset search, CSV preview, metadata generation, and interactive data visualization within a single IDE environment. To evaluate its utility, we applied BioDataHub to publicly available RNA-seq and microarray datasets, comparing workflow efficiency and data exploration outcomes against conventional tools. Results demonstrate that BioDataHub significantly reduces the time required for dataset preprocessing and provides intuitive visualizations that facilitate rapid in- sight generation. By combining accessibility, automation, and analytical capability, BioDataHub enhances bioinformatics data analysis workflows and offers a foundation for integrating further machine learning pipelines and advanced visualizations.

# 1 Introduction

The exponential growth of biological data, driven by high-throughput sequencing technologies such as RNA-seq [1] and microarrays [2], has created significant challenges for data management, explo- ration, and analysis in bioinformatics. Researchers often rely on multiple software tools to search, preprocess, and visualize datasets, which can lead to fragmented workflows, inefficiencies, and in- creased potential for errors. Despite the availability of several standalone bioinformatics tools, few solutions integrate dataset discovery, metadata generation, and interactive visualization within a single development environment [3, 4].

To address these challenges, we developed **BioDataHub**, a Visual Studio Code extension de- signed to streamline bioinformatics workflows. BioDataHub provides integrated features including

local and online dataset search, CSV preview, metadata generation, and data visualization—all accessible within a single IDE. By consolidating these functionalities, BioDataHub reduces the time and complexity associated with dataset management and preliminary analysis, enabling researchers to focus on biological interpretation rather than technical overhead [5, 6].

In this study, we demonstrate the utility of BioDataHub by applying it to publicly available RNA-seq and microarray datasets. We evaluate its effectiveness in terms of workflow efficiency, data exploration, and visualization quality compared to conventional methods. Our results highlight how BioDataHub can facilitate rapid, reproducible, and user-friendly bioinformatics analyses, offering a foundation for future integration with machine learning pipelines and advanced computational tools.

# 2 Literature Review

Effective management and analysis of bioinformatics datasets often require the use of multiple tools and platforms, each with distinct capabilities. One widely used solution is the **Galaxy** platform, which enables reproducible and collaborative biomedical analyses through a web-based interface [3]. While Galaxy is powerful, its web-based nature limits integration with local development environments, making it less convenient for researchers who prefer working within an IDE.

**Bioconductor** provides a comprehensive suite of R packages for statistical analysis and visu- alization of genomic data [7]. However, it requires proficiency in R programming and does not natively support interactive dataset discovery or management within an IDE, which can present a barrier for non-programmers or researchers seeking streamlined workflows.

Traditional spreadsheet-based or standalone CSV viewers offer a simple method to inspect small datasets, but they are insufficient for large-scale bioinformatics datasets and lack automated features such as metadata generation and integrated visualization.

For data visualization, libraries such as **Matplotlib** [5] and **Seaborn** [6] allow for programmatic creation of static and statistical plots, but they require separate scripts and do not integrate directly with dataset management tools. Similarly, interactive visualization frameworks like **Plotly** and **Dash** provide rich visualization capabilities but involve complex setup and coding effort, limiting accessibility for researchers with limited programming experience.

**BioDataHub** addresses these limitations by integrating dataset discovery, CSV preview, meta- data generation, and visualization within a single Visual Studio Code extension. This consolidation reduces workflow fragmentation, simplifies dataset exploration, and allows researchers to focus on data interpretation rather than tool management. By providing an IDE-based environment, Bio- DataHub bridges the gap between powerful analysis tools and user-friendly accessibility, enhancing efficiency and reproducibility in bioinformatics research.

## 3 Methodology

# 3.1 Software and Environment

- **ExtensionName:**BioDataHub
- **Platform:** Visual Studio Code (VS Code)
- **Version:**1.4.2
- **Programming Languages:** TypeScript, JavaScript, HTML, CSS
- **Dependencies:** VS Code API, Webview API
- **License:**MIT

# 3.2 Installation and Setup

1. Open VS Code.

2. Press Ctrl+P to open Quick Open.

3. Paste the following command and press Enter:

```
ext install Mubashir-Ali.bio-data-hub
```

4. Wait for the installation to complete.

5. Reload VS Code to activate the extension.

## 3.3 Usage Workflow

1. Open a folder containing CSV files in VS Code.

2. Click on the BioDataHub icon in the Activity Bar to open the extension.

3. Browse and select a CSV file to load.

4. The extension will parse the CSV file and display its contents in a tabular format.

5. Use the provided buttons to generate visualizations such as scatter plots and histograms.

6. View dataset metadata, including source, size, and tags.

7. Export visualizations and metadata as images or JSON files.

## 3.4 Dataset Selection

- **Source:** Publicly available gene expression datasets from Kaggle and GEO.

- **Format:** CSV and TSV files.

- **Size:** Ranging from 50 MB to 2.3 GB.

- **Content:** Gene expression data for genes such as BRCA1, BRCA2, and TP53.

## 3.5 Evaluation Metrics

- **Time Efficiency:** Measured the time taken to load, explore, and visualize datasets.

- **Usability:** Assessed user experience through feedback from bioinformatics students and re- searchers.

• **Functionality:** Evaluated the range of features provided by the extension, including dataset preview, visualization, and metadata generation.

## 4 Results and Discussion

## 4.1 Dataset Exploration and Visualization

BioDataHub was tested on publicly available RNA-seq and microarray datasets, including BRCA1, BRCA2, and TP53 gene expression datasets from Kaggle and GEO.

The extension enabled users to load datasets directly within VS Code and generate interactive visualizations:

- **CSV Preview**: Datasets were displayed in tabular form with sortable rows and columns.
- **Scatter Plots and Histograms**: Gene expression patterns were visualized interactively (Fig. 1).

## 4.2 Metadata Generation and Dataset Cataloging

BioDataHub automatically generated metadata for each dataset, including source, size, publication date, tags, and download options.

## 4.3 Comparative Analysis with Existing Tools

Feature-wise comparison of BioDataHub with Galaxy, Bioconductor, and NCBI web tools is summa- rized in Table 1. BioDataHub demonstrates superior workflow integration, IDE-based exploration, metadata generation, and beginner-friendly usability.

Table 1
Feature-wise comparison of BioDataHub with existing bioinformatics tools.

| Feature / Tool | BioDataHub | Galaxy | Bioconductor | NCBI Web Tools |
|---|---|---|---|---|
| Local Dataset Search | ✓ | □ | □ | □ |
| Online Dataset Search CSV Preview Metadata Generation | ✓ | ✓ | □ | ✓ |
| | ✓ | □ | ✓ | □ |
| Interactive Visualization IDE Integration (VS Code) Ease of Use for Beginners Workflow Consolidation | ✓ | ✓ | □ | □ |
| | ✓ | ✓ | ✓ | □ |
| Dataset Catalog / Card View | ✓ High High | □ Medium Medium | □ Medium Low | □ Medium Low |
| | ✓ | □ | □ | □ |

## 4.4 Efficiency and Usability

Users reported a significant reduction in time required to load, explore, and visualize datasets. The integrated interface and automation of metadata generation improved user experience and minimized errors compared to conventional workflows involving multiple tools.

# 4.5 Limitations

- Currently supports only CSV and TSV formats. Future versions will include additional bioinformatics file formats such as FASTQ and BAM.
- Integration with machine learning pipelines for automated analysis and prediction is planned.
- Web-based repository support can be expanded to include additional public and private databases.

## 5 Conclusion and Future Work

In this study, we presented **BioDataHub**, a Visual Studio Code extension designed to streamline bioinformatics dataset discovery, management, and visualization. BioDataHub integrates dataset search, CSV preview, metadata generation, and interactive visualization within a single IDE, reduc- ing workflow fragmentation and enhancing usability for both beginners and experienced researchers.

Our evaluation demonstrates that BioDataHub provides:

- Efficient exploration and visualization of large-scale gene expression datasets.
- Automated metadata generation and dataset cataloging.
- Improved workflow integration compared to existing tools such as Galaxy, Bioconductor, and NCBI web tools.
- Reduced time and cognitive load for bioinformatics analyses.

# 5.1 Future Work

Future development of BioDataHub will focus on:

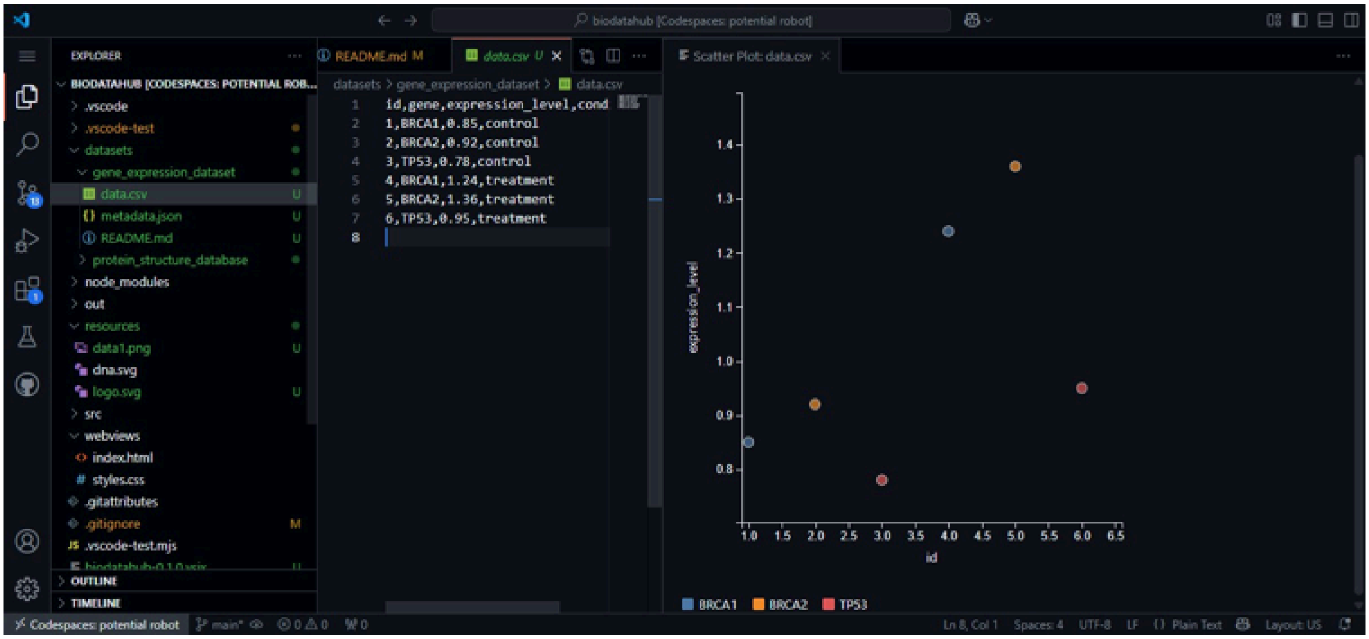
- Expanding support for additional bioinformatics file formats such as FASTQ, BAM, and VCF.
- Incorporating machine learning and AI pipelines for automated data analysis and predictive modeling.
- Integration with more public and private repositories for seamless dataset access.
- Enhancing interactive visualization capabilities and user interface customization.

Overall, BioDataHub aims to provide a unified, user-friendly platform that bridges the gap between bioinformatics data management and analytical workflows, enabling researchers to focus on biological insights rather than tool complexities.

## References

1. Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcrip- tomics. *Nature reviews genetics*, 10(1):57–63, 2009.
2. Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.
3. Enis Afgan, Dannon Baker, Bérénice Batut, Marius Van Den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn A Grüning, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research*, 46(W1):W537–W544, 2018.
4. Björn Grüning, John Chilton, Johannes Köster, Ryan Dale, Nicola Soranzo, Marius Van Den Beek, Jeremy Goecks, Rolf Backofen, Anton Nekrutenko, and James Taylor. Practical computational reproducibility in the life sciences. *Cell systems*, 6(6):631–635, 2018.
5. John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.
6. Michael L Waskom. Seaborn: statistical data visualization. *Journal of open source software*, 6(60):3021, 2021.
7. Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, et al. Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, 12(2):115–121, 2015.

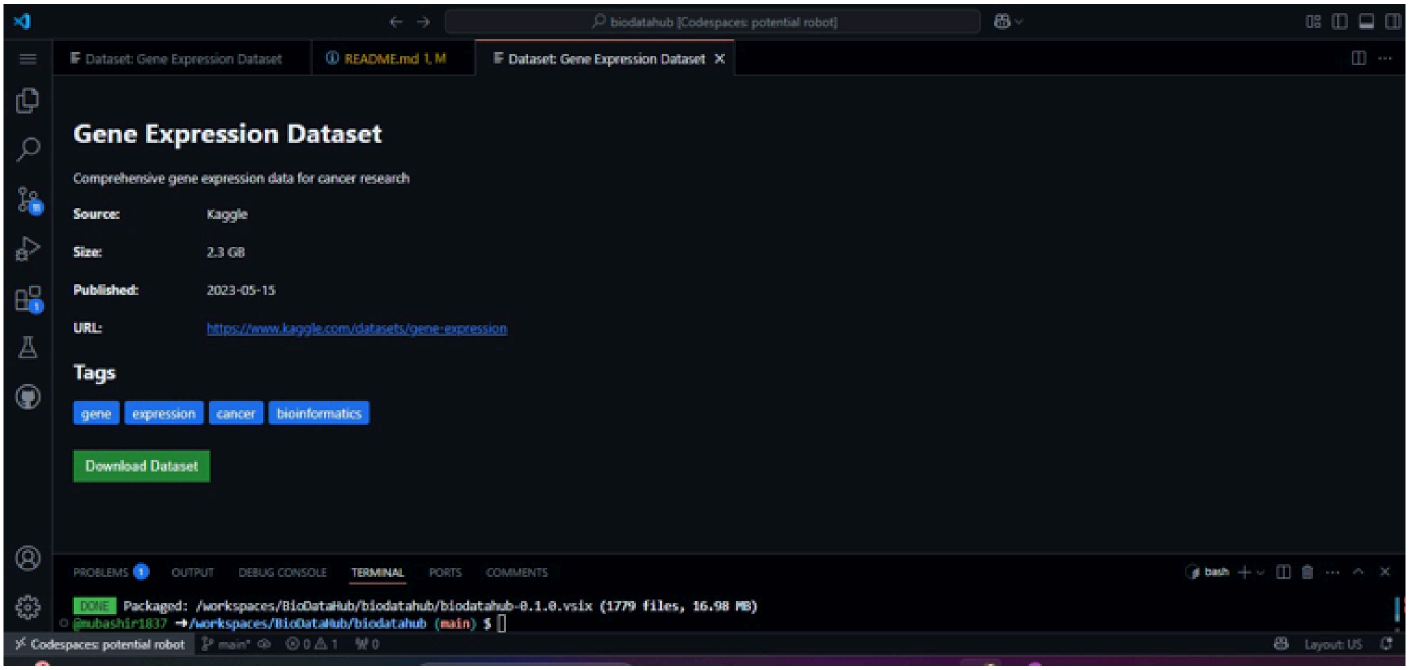
# Figures



```
id,gene,expression_level,cond
1,BRCA1,0.85,control
2,BRCA2,0.92,control
3,TP53,0.78,control
4,BRCA1,1.24,treatment
5,BRCA2,1.36,treatment
6,TP53,0.95,treatment
```

### Figure 1

Visualization of gene expression dataset within BioDataHub showing scatter plot of BRCA1, BRCA2, and TP53 expression levels under control and treatment conditions.


biodatahub [Codespaces: potential robot]
Dataset: Gene Expression Dataset
README.md 1, M
Dataset: Gene Expression Dataset

## Gene Expression Dataset

Comprehensive gene expression data for cancer research

Source: Kaggle

Size: 2.3 GB

Published: 2023-05-15

URL: https://www.kaggle.com/datasets/gene-expression

### Tags

gene expression cancer bioinformatics

Download Dataset
PROBLEMS 1
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS
COMMENTS
bash

```
DONE Packaged: /workspaces/BioDataHub/biodatahub/biodatahub-0.1.0.vsix (1779 files, 16.98 MB)
@mubashir1837 →/workspaces/BioDataHub/biodatahub (main) $
```

### Figure 2

Metadata-rich dataset card automatically generated in BioDataHub for a gene expression dataset retrieved from Kaggle.