

# NeuroVar: An Open-source Tool for Gene Expression and Variation Data Visualization for Biomarkers of Neurological Diseases

Hiba Ben Aribi<sup>1\*</sup>, Najla Abassi<sup>2</sup> and Olaitan I. Awe<sup>3,4</sup>

<sup>1</sup> Faculty of Sciences of Tunis, University of Tunis El Manar, Tunis, Tunisia

<sup>2</sup> Laboratory of Biomedical Genomics and Oncogenetics, Institut Pasteur de Tunis, University of Tunis El Manar, Tunisia

<sup>3</sup> Department of Computer Science, Faculty of Science, University of Ibadan, Ibadan, Oyo State, Nigeria

<sup>4</sup> African Society for Bioinformatics and Computational Biology, Cape Town, South Africa

\* Corresponding Author: Hiba Ben Aribi, [benaribi.hiba@gmail.com](mailto:benaribi.hiba@gmail.com)

## Abstract

**Background:** The expanding availability of large-scale genomic data and the growing interest in uncovering gene-disease associations call for efficient tools to visualize and evaluate gene expression and genetic variation data.

**Methodology:** Data collection involved filtering biomarkers related to multiple neurological diseases from the ClinGen database. We developed a comprehensive pipeline that was implemented as an interactive Shiny application and a standalone desktop application.

**Results:** NeuroVar is a tool for visualizing genetic variation (single nucleotide polymorphisms and insertions/deletions) and gene expression profiles of biomarkers of neurological diseases.

**Conclusion:** The tool provides a user-friendly graphical user interface to visualize genomic data and is freely accessible on the project's [GitHub repository](https://github.com/omicscodeathon/neurovar) (<https://github.com/omicscodeathon/neurovar>).

**Keywords:** Software and Workflows, Bioinformatics, Genetics

## **I. Statement of Need:**

Disease biomarkers are genes or molecules that indicate the presence or severity of a disease. Their identification provides important insights into disease etiology and can facilitate the development of new treatments and therapies [1]. Integrating multi-omics data, such as gene expression and genetic variations, has emerged as a powerful approach for biomarker discovery.

Several genomics studies have discovered multiple genetic variations linked to numerous neurological conditions which are complex diseases with a significant level of heterogeneity, such as Alzheimer's disease [2] and Parkinson's disease [3]. Some studies have also used genetic variants to detect the presence of human disorders [4].

The discovered biomarkers are extensively documented in various scientific publications and are accessible through databases like the Clinical Genome (ClinGen) database. ClinGen stores a vast amount of genomic data, including a comprehensive dataset of biomarkers associated with multiple diseases, including various neurological disorders [5].

Multiple computational tools have been developed in recent years to analyze genomic data including gene expression data analysis [6,7], identification of potential inhibitors for therapeutic targets [8], and comparative analysis of molecular and genetic evolution [9]. However, there is still a need for a specialized tool that focuses on filtering critical disease biomarkers as this will help in studies that work on finding genes that are involved in diseases using transcriptomic data generated from sequencing experiments [10,11,12,13]. Such a tool would help users identify phenotypic subtypes of diseases in their patients, thereby facilitating more accurate diagnoses and personalized treatment plans.

In this study, we developed a novel tool, named “NeuroVar”, to analyze biomarker data specifically for neurological diseases, including gene expression profiles and genetic variations particularly single nucleotide polymorphisms (SNPs) and nucleotide insertion and/or deletion (Indels).

## **II. Implementation**

### **● Data collection**

The ClinGen database [5] provides a dataset of biomarkers of multiple diseases from which we

filtered data of eleven neurological syndromes and seven non-neurological diseases with neurological manifestations.

### ● **Software development**

Two versions of the tool were developed: an R shiny and a desktop application.

The shiny application was developed using multiple R packages including shiny [14] and shinydashboard [15]. Other R packages are used for data manipulation including dplyr [16], readr [17], tidyverse [18], purrr [19], vcfR [20], bslib [21], stringr [22], data.table [23], fs [24], DT [25], sqldf [26], and ggplot2 [27].

For the stand-alone desktop application, wxPython framework [28] was used to build a similar GUI. A variety of Python libraries were employed including pandas [29], matplotlib [30], numpy [31]. After testing, the application was packaged as an installer using cx\_Freeze [32]. Finally, it was distributed as a zip file to be downloaded.

### ● **Pipeline validation and case study**

To validate the pipeline, a case study was performed on the public dataset SRP149638 [33] available on the SRA database [34]. The dataset corresponds to RNA sequencing data from the peripheral blood mononuclear cells from healthy donors and Amyotrophic Lateral Sclerosis (ALS) patients. The ALS patients involved in the study have mutations in the FUS, TARDBP, SOD1, and VCP genes.

The file's preprocessing, genetic expression analysis, and variant calling were performed using the Exvar R package [35]. The Exvar package uses the rfastp package [36] and the gmapR package [37] for preprocessing fastq files, the GenomicAlignments package [38], and the DESeq2 packages [39] for gene expression data analysis, as well as the VariantTools [40] and VariantAnnotation [41] packages for variant calling.

## **III. Results**

### ● **Supported Disease**

NeuroVar integrates biomarkers of multiple neurological diseases including epilepsy, amyotrophic lateral sclerosis, intellectual disability, autism spectrum disorder, brain malformation syndrome,

syndromic disorders, cerebral palsy, RASopathy, aminoacidopathy, craniofacial malformations, Parkinson's disease, PHARC syndrome. It also integrates seven non-neurological diseases with neurological manifestations, including peroxisomal disorders, hereditary cancer, mitochondrial disease, retina-related disorders, general gene curation, hearing loss, and fatty acid oxidation disorders. Each disease syndrome includes multiple disease types, for example, sixteen types of amyotrophic lateral sclerosis disorder are integrated.

### ● **Operation and Implementation**

The desktop and Shiny applications have the same user interface, however, the implementation is different.

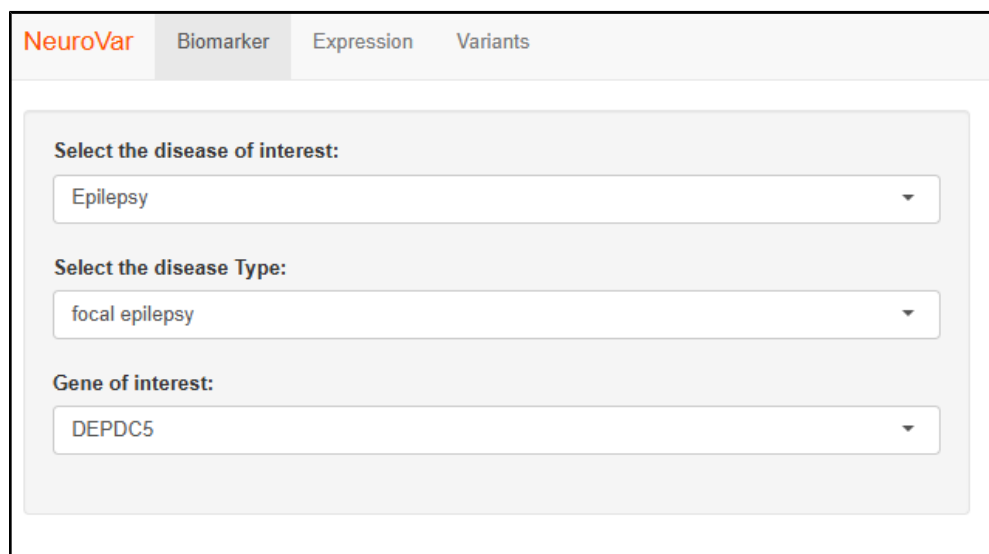
The Shiny application is platform-independent, however, the desktop application is optimized for the Windows operating system. The necessary library requirements for the tool are automatically installed in both versions. The amount of RAM used depends on the servers or the machine being used, and the only prerequisites for using the tool are having R installed for the shiny application and having Python installed for the desktop application.

The tool is compatible with RNA sequencing data. The input data files should be in CSV format for gene expression data and VCF format for genetic variants data. Guidance of the files' organization is available in the tool's Github repository in detail (path: [omicscodeathon/neurovar/demonstration\\_data](#))

Detailed guidelines for installing and using both versions of the application are provided in the project's GitHub repository.

### ● **The application's usage**

The application dashboard includes three pages. The first page, named "Biomarker" provides data on the disease's biomarkers. Initially, the user should select the target disease syndrome and the specific disease subtypes from the provided list (Figure 1).



The screenshot shows the 'Biomarker' tab of the NeuroVar interface. It contains three dropdown menus for user input:

- Select the disease of interest:** A dropdown menu with 'Epilepsy' selected.
- Select the disease Type:** A dropdown menu with 'focal epilepsy' selected.
- Gene of interest:** A dropdown menu with 'DEPDC5' selected.

**Figure 1. The layout of the “Biomarker” page. The user is requested to define the target disease, disease type, and gene of interest.**

Next, a list of biomarkers is provided associated with additional data including the gene’s mode of inheritance, description, type, and transcripts. Also, a link for the official online report validating the gene's association with the disease is provided (Figure 2).

About the gene

Show10▼entries

Search:

gene	GENE ID (HGNC)	DISEASE ID (MONDO)	MOI	SOP	CLASSIFICATION	ONLINE REPORT	CLASSIFICATION DATE	
1	DEPDC5	HGNC:18423	MONDO:0005384	AD	SOP9	Definitive	<a href="https://search.clinicalgenome.org/kb/gene-validity/CGGV.assertion_82a82c75-f9a5-4f51-a15c-6513c13df57c-2018-08-07T040000.000Z">https://search.clinicalgenome.org/kb/gene-validity/CGGV.assertion_82a82c75-f9a5-4f51-a15c-6513c13df57c-2018-08-07T040000.000Z</a>	2018-08-07T04:00:00Z

Showing 1 to 1 of 1 entries

Previous1Next

gene's transcript

Show10▼entries

Search:

gene	Transcript name	Transcript type	Transcription start site (TSS)	Transcript end (bp)	Transcript start (bp)
1	DEPDC5	DEPDC5-221	protein_coding	31753867	31906992
2	DEPDC5	DEPDC5-247	protein_coding	31753898	31837279
3	DEPDC5	DEPDC5-230	protein_coding	31753898	31837324
4	DEPDC5	DEPDC5-248	protein_coding	31753900	31907005
5	DEPDC5	DEPDC5-239	protein_coding	31753912	31906989
6	DEPDC5	DEPDC5-245	protein_coding	31753932	31907001
7	DEPDC5	DEPDC5-208	protein_coding	31753944	31784788
8	DEPDC5	DEPDC5-237	retained_intron	31753955	31769772
9	DEPDC5	DEPDC5-249	nonsense_mediated_decay	31753955	31906989
10	DEPDC5	DEPDC5-207	protein_coding	31753955	31906998

Showing 1 to 10 of 64 entries

Previous1234567Next

**Figure 2. The output of the “Biomarker” page. The output includes two tables detailing key information about the selected gene.**

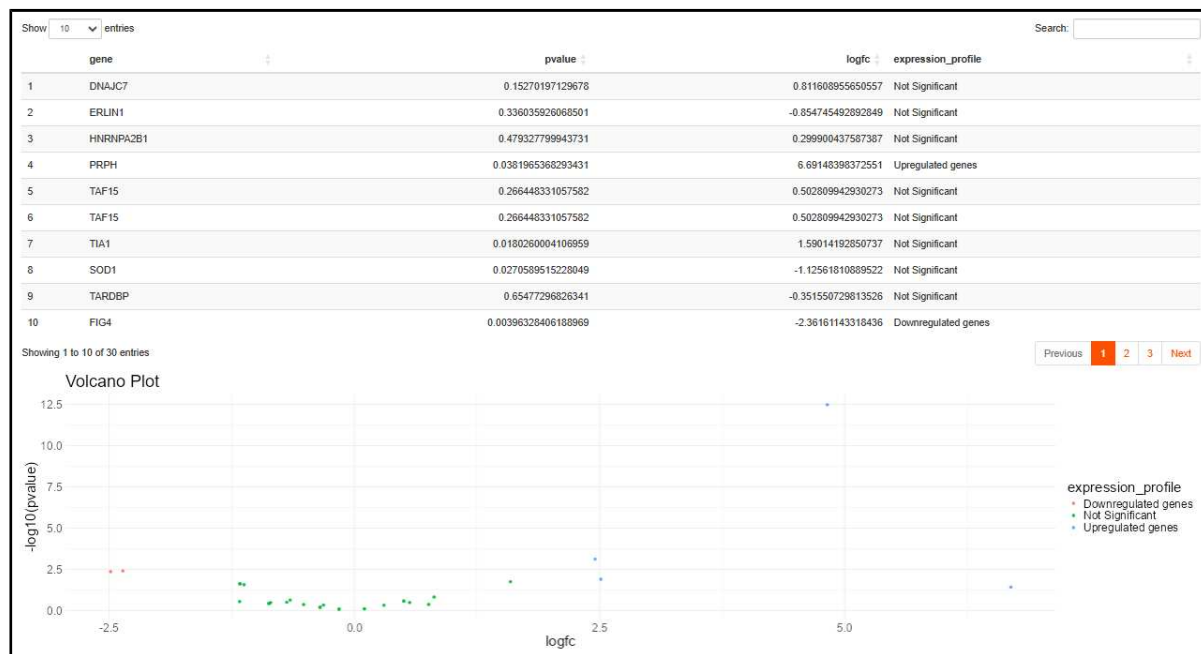
The second page, named "Expression", is used to visualize the biomarkers expression profile.

After importing a CSV file and identifying the key columns, the log2FC value and adjusted p-value are requested to define the differential expression profile. By default, the adjusted p-value is set to less than 0.01 and the logFC value is set to less or more than 2 ( Figure 3).

The screenshot shows the 'Expression' tab of the NeuroVar application. At the top, there are four tabs: 'NeuroVar', 'Biomarker', 'Expression' (selected), and 'Variants'. Below the tabs, there is a section titled 'Choose file to upload' with a 'Browse...' button and a 'No file selected' status. Underneath, there is a 'Separator:' section with three radio buttons: the first is selected, followed by a semicolon, a comma, and a colon. Below this are three dropdown menus labeled 'Select Gene Column', 'Select Adjusted P-value Column', and 'Select LogFC Column'. A note below these menus reads: 'Note: Make Sure the disease is selected in the first tab ! Define the p-value and LogFC value to identify the differentially expressed genes'. At the bottom, there are two sliders. The first is labeled 'P\_value:' and has a range from 0 to 0.05 with a slider knob at 0.01. The second is labeled 'Log value:' and has a range from 0 to 5 with a slider knob at 2.

**Figure 3.** The layout of the “Expression” page. the user is requested to upload the data file and select the p-value, and the log-FC value required to construct the differential expression profile.

As a result, the expression profiles of the target disease biomarkers (previously selected) are summarized in a table and represented in a volcano plot (Figure 4).



**Figure 4. The output of the "Expression" page. As output, a summary of the genes' expression profiles is displayed in a table and a volcano plot.**

The third page, named "Variants", allows the visualization of SNPs and Indels data. The user is requested to define the path to the directory containing the VCF files. The files are expected to be divided into two folders, named "controls" and "patients", containing the VCF files of the controls and patients groups respectively. The user needs to define the variant type as SNPs or Indels (Figure 5).

Figure 5 shows the layout of the "Variant" page. The page has a header with tabs: NeuroVar, Biomarker, Expression, and Variants. The main content area contains a form with the following elements:

- Enter folder path:** A text input field.
- Note:** Make sure the path contain two folders named 'control' and 'patient'
- variant type:**
  - ☒ SNP
  - ☐ Indels
- Submit** button

**Figure 5. The layout of the "Variant" page. The user is prompted to specify the path to the data-containing folder and the data type.**

The VCF files are processed and annotated, and then the variants in the target disease biomarkers are filtered and resumed in a table comparing the reference genome, the control group, and the patients' group (Figures 6 and 7).

Show

10

entries

Search:

	Gene	Chromosome	Start	End	SNP Position	SNP ID	Reference Genome Allele	Control's Allele	Patient's Allele	Comparison	TiTv control	TiTv Patient
24	ARX	X	25003694	25016420	25012156	-	A	G	-	deletion	-	Ti
72	SRPX2	X	100644195	100675788	100667950	-	-	-	A	addition	Ti	-
73	SRPX2	X	100644195	100675788	100668080	-	-	-	C	addition	Ti	-
110	WDR45	X	49074433	49101170	49090523	-	-	-	A	addition	Ti	-
111	WDR45	X	49074433	49101170	49090565	-	-	-	C	addition	Tv	-
112	WDR45	X	49074433	49101170	49090574	rs113439843	-	-	C	addition	Ti	-
113	WDR45	X	49074433	49101170	49090618	-	-	-	C	addition	Ti	-
114	WDR45	X	49074433	49101170	49090633	rs111423624	-	-	C	addition	Ti	-
42	STXP1	9	127579370	127696027	127619922	rs944952	-	-	C	addition	Ti	-
43	STXP1	9	127579370	127696027	127620170	-	-	-	G	addition	Ti	-

Showing 1 to 10 of 114 entries

Previous

1

2

3

4

5

...

12

Next

**Figure 6. The output of the “Variant” page. Table summarizing the SNPs in the target disease’s biomarkers**

Show

10

entries

Search:

	Gene	Chromosome	Start	End	Position	Reference Genome Allele	Control's Allele	Patient's Allele	Comparison
1	CNTNAP2	7	146116002	148420998	148334398	ATGAC	A	-	deletion
2	CNTNAP2	7	146116002	148420998	148334459	T	TC	-	deletion
3	SCN8A	12	51590266	51812864	51714418	CTTT	C	-	deletion
4	DOCK7	1	62454298	62688386	62504590	C	CACAAGG	-	deletion
5	GABRB3	15	26543552	26939539	26890642	CA	C	-	deletion
6	NECAP1	12	8076939	8097881	8080154	AATC	A	-	deletion
7	SZT2	1	43389882	43454247	43403902	G	GC	-	deletion
8	PRICKLE1	12	42456757	42590355	42478264	C	CA	-	deletion
9	KCNMA1	10	76869601	77638369	77539716	C	CA	-	deletion
10	KCNMA1	10	76869601	77638369	77539716	C	CA	-	deletion

Showing 1 to 10 of 12 entries

Previous

1

2

Next

**Figure 7. The output of the “Variant” page. Table summarizing the INDELs in the target disease’s biomarkers**

## ● Case study results

To validate the pipeline, we conducted a case study using the public dataset that provides RNA sequencing data of ALS patients who are declared to have mutations in the FUS, TARDBP, SOD1, and VCP genes [33].

Initially, we used NeuroVar to explore the roles of the genes FUS, TARDBP, SOD1, and VCP in ALS. Our findings confirmed that FUS, TARDBP, and SOD1 are recognized ALS biomarkers, while VCP is not. ALS has 26 subtypes, with FUS being a biomarker for type 6, SOD1 for type 1, and TARDBP for type 10, suggesting that the patients in the study may represent a mixture of these ALS subtypes.

Next, we investigated whether mutations in these genes impacted their expression profiles. Using an adjusted p-value threshold of 0.05 and a log fold change (logFC) cutoff of 2, we found that out of 21 known ALS biomarkers, only one gene—TUBA4A—was differentially expressed. Notably, none of the four genes (FUS, TARDBP, SOD1, and VCP) showed differential expression.

Finally, we examined the types of mutations present in these genes. We detected 23 SNPs across 7 biomarkers: DAO (all ALS types), FIG4 (ALS type 11), ERBB4 (ALS type 19), TUBA4A (ALS type 22), KIF5A (ALS type 25), C9orf72 (ALS type 1), and TBK1 (ALS type 4). No indels were detected in any of the biomarkers. Interestingly, the biomarkers FUS, TARDBP, and SOD1 exhibited neither SNPs nor indels, suggesting that the mutations in these genes may be due to other types of genomic changes.

#### **IV. Discussion and conclusion**

NeuroVar is a novel tool for visualizing genetic variation and gene expression data related to neurological diseases. The tool is designed to visualize genetic variation and gene expression data, with a particular emphasis on neurological disorders. This specialization makes it an invaluable resource for researchers and clinicians focused on these conditions. It offers features to filter biomarkers by specific diseases, which aids in confirming gene-disease associations and prioritizing genes for further investigation.

The tool supports eleven neurological syndromes and seven non-neurological diseases with neurological manifestations. While the supported diseases list is currently limited to data from the ClinGen database, it will be frequently updated, and data sources will be expanded to include other databases in the future.

NeuroVar is available as a desktop application and a Shiny application. Both versions are user-friendly with no computational skills needed for operation. Additionally, all necessary

dependencies are automatically installed with the tools. This dual accessibility of NeuroVar caters to users with varying preferences and technical backgrounds which makes it more accessible and easier to use than other genetic variant data visualization tools such as the command line tool VIVA [42] to analyze VCF files and the “Transcriptomics oSPARC” web tool for gene expression data visualization hosted on the o<sup>2</sup>S<sup>2</sup>PARC platform [6].

In addition to its user-friendly design, NeuroVar streamlines the research workflow by eliminating the need for multiple filtering steps across different platforms. By integrating essential functions within a single interface, it allows users to conduct comprehensive analyses without leaving the application, thereby enhancing efficiency and reducing errors. The inclusion of a quick-access library on the first page further aids in referencing important data, making it easier to revisit and validate findings. This centralization of tasks, coupled with a focus on neurological diseases and extensive biomarker information, makes NeuroVar a highly useful tool for advancing research in the field.

## **Data Availability**

Data came from ClinVar, and the presented case study was performed on the public dataset SRP149638 from the SRA database.

The Open Source code of the Shiny application and the desktop application are available in the project’s GitHub Repository: <https://github.com/omicscodeathon/neurovar>

Installation Guide, demonstration data and video demonstration are also available in the project’s GitHub Repository: <https://github.com/omicscodeathon/neurovar>

## **Availability of supporting source code and requirements**

Project name: NeuroVar

Project home page: <https://github.com/omicscodeathon/neurovar>

Operating system: Platform independent

Programming language: Python and R

Other requirements: None

License: Artistic license 2.0

RRID: SCR\_025640

## Abbreviations

SNP: Single Nucleotide Polymorphism

Indel: Insertion Deletion

CSV: Comma Separated Values

VCF: Variant Call Format

## References

1. Hirschhorn, J. N., Lohmueller, K. E., Byrne, E., & Hirschhorn, K. A comprehensive review of genetic association studies. *Genetics in Medicine*, 4(2), 45–61. 2002; doi: [10.1097/00125817-200203000-00002](https://doi.org/10.1097/00125817-200203000-00002).
2. Kunkle, B.W., Grenier-Boley, B., Sims, R. et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A $\beta$ , tau, immunity and lipid processing. *Nat Genet* 51, 414–430. 2019; doi: [10.1038/s41588-019-0358-2](https://doi.org/10.1038/s41588-019-0358-2).
3. Li, W., Fu, Y., Halliday, G. M., & Sue, C. M. PARK genes link mitochondrial dysfunction and Alpha-Synuclein pathology in sporadic Parkinson's disease. *Frontiers in Cell and Developmental Biology*, 9. 2021; doi: [10.3389/fcell.2021.612476](https://doi.org/10.3389/fcell.2021.612476).
4. Wesonga, R.M. and Awe, O.I. (2022). An Assessment of Traditional and Genomic Screening in Newborns and their Applicability for Africa. *Informatics in Medicine Unlocked*, 32:101050. <https://doi.org/10.1016/j.imu.2022.101050>
5. Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L. and Plon, S.E.. ClinGen—the clinical genome resource. *New England Journal of Medicine*, 372(23), 2235-2242. 2015.
6. Aribi, H. B., Ding, M., & Kiran, A. Gene expression data visualization tool on the o<sup>3</sup>S<sup>2</sup>PARC platform. *F1000Research*, 11. 2022; doi: [10.12688/f1000research.126840.2](https://doi.org/10.12688/f1000research.126840.2).
7. Die, J.V., Elmassry, M.M., LeBlanc, K.H., Awe, O.I., Dillman, A., and Busby, B. (2019). geneHummus: an R package to define gene families and their expression in legumes and beyond. *BMC Genomics* 20, 591. <https://doi.org/10.1186/s12864-019-5952-2>
8. Ogbodo U.C., Enejoh O.A., Okonkwo C.H., Gnanasekar P., Gachanja P.W., Osata S.,

- Atanda H.C., Iwuchukwu E.A., Achilonu, I. and Awe, O.I. (2023). Computational Identification of Potential Inhibitors Targeting cdk1 in Colorectal Cancer. *Frontiers in Chemistry*. 2023; doi:10.3389/fchem.2023.1264808.
9. Awe, O.I., En najih, N., Nyamari, M.N., and Mukanga, L.B. (2023). Comparative study between molecular and genetic evolutionary analysis tools using African SARS-CoV2 variants. *Informatics in Medicine Unlocked* 36, 101143. 2023; doi:10.1016/j.imu.2022.101143.
10. El Abed, F., Baraket, G., Nyamari, M.N., Naitore, C., and Awe, O.I. (2023). Differential Expression Analysis of miRNAs and mRNAs in Epilepsy Uncovers Potential Biomarkers. *bioRxiv*.
11. Chikwambi, Z., Hidjo, M., Chikondowa, P., Afolabi, L., Aketch, V., Jayeoba, G., Enoma, D.O. and Awe, O.I. (2023). Multi-omics data integration approach identifies potential biomarkers for Prostate cancer. *bioRxiv*.
12. Nyamari, M.N., Omar, K.M., Fayehun A.F., Dachi, O., Bwana, B.K. and Awe, O.I. (2023). Expression Level Analysis of ACE2 Receptor Gene in African-American and Non-African-American COVID-19 Patients. *BioRxiv*.
13. Nzungize, L., Kengne-Ouafo, J.A., Wesonga, M.R., Umuhoza, D., Murithi, K., Kimani, P., Awe, O.I. and Dillman, A. (2022). Transcriptional Profiles Analysis of COVID-19 and Malaria Patients Reveals Potential Biomarkers in Children. *bioRxiv*.
14. Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2023). shiny: Web Application Framework for R. R package version 1.7.4.9002, <https://shiny.rstudio.com/>.

15. Chang W., Borges Ribeiro B. (2023). Create dashboards with 'Shiny'.\_  
<http://rstudio.github.io/shinydashboard/>.
16. Wickham H, François R, Henry L, Müller K (2022). dplyr: A Grammar of Data Manipulation. <https://github.com/tidyverse/dplyr>.
17. Wickham H, Hester J, Bryan J (2022). readr: Read Rectangular Text Data.  
<https://readr.tidyverse.org>, <https://github.com/tidyverse/readr>.
18. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. 2019; doi:[10.21105/joss.01686](https://doi.org/10.21105/joss.01686).
19. Wickham H, Henry L. (2023) purrr: Functional Programming Tools.  
<https://purrr.tidyverse.org>.
20. Knaus BJ, Grünwald NJ. “VCFR: a package to manipulate and visualize variant call format data in R.” Molecular Ecology Resources, 17(1), 44–53. ISSN 757. 2017; doi:  
[10.1111/1755-0998.12549](https://doi.org/10.1111/1755-0998.12549).
21. Sievert C, Cheng J (2023). bslib: Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'. <https://rstudio.github.io/bslib/>, <https://github.com/rstudio/bslib>.
22. Wickham H (2022). stringr: Simple, Consistent Wrappers for Common String Operations.  
<https://stringr.tidyverse.org>, <https://github.com/tidyverse/stringr>.
23. Dowle M., Srinivasan A., Gorecki J., Chirico M., Stetsenko P., Short T., Lianoglou S., Antonyan E., et al. (2023) Data.table: Extension of 'data.frame'.\_  
<https://github.com/Rdatatable/data.table>.
24. Hester J., Wickham H., Csárdi G.. (2023) Fs: Cross-Platform File System Operations Based on 'libuv'. <https://fs.r-lib.org>.
25. Xie Y., Cheng J., Tan X., Allaire JJ., Girlich M., Freedman Ellis G., Rauh J., Reavis B. et al., (2023). DT: A Wrapper of the JavaScript Library 'DataTables'.\_  
<https://github.com/rstudio/DT>.
26. Grothendieck G (2017). sqldf: Manipulate R Data Frames Using SQL. Retrieved from  
<https://cran.r-project.org/web/packages/sqldf/index.html>.
27. Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, <https://ggplot2.tidyverse.org>.
28. Talbot, H. (2000). WxPython, a GUI toolkit. Linux Journal, 2000(74), 5.  
<https://dl.acm.org/citation.cfm?id=349312>.

29. McKinney, W. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, No. 1, pp. 51-56). 2010.
30. Hunter, J. D.. Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95. 2007.
31. Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J. and Kern, R.. Array programming with NumPy. Nature, 585(7825), 357-362. 2020; doi:10.1038/s41586-020-2649-2.
32. Duarte M.. GitHub - marcelotduarte/cx\_Freeze: Create standalone executables from Python scripts, with the same performance and is cross-platform.  
[https://github.com/marcelotduarte/cx\\_Freeze](https://github.com/marcelotduarte/cx_Freeze). 2023.
33. Zucca, S., Gagliardi, S., Pandini, C., Diamanti, L., Bordoni, M., Sproviero, D., Arigoni, M., Olivero, M., Pansarasa, O., Ceroni, M. and Calogero, R. RNA-Seq profiling in peripheral blood mononuclear cells of amyotrophic lateral sclerosis patients and controls. Scientific Data, 6(1), 1-8. 2019; doi:10.1038/sdata.2019.6.
34. Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011 Jan;39(Database issue):D19-21. doi:[10.1093/nar/gkq1019](https://doi.org/10.1093/nar/gkq1019).
35. Ben Aribi H., Dixon I., Abassi N., and Awe O. I. (2023). Exvar: An R Package for Gene Expression And Genetic Variation Data Analysis And Visualization.  
<https://github.com/omicscodeathon/Exvar>.
36. Wang W, Carroll T (2022). Rfastp: An Ultra-Fast and All-in-One Fastq Preprocessor (Quality Control, Adapter, low quality and polyX trimming) and UMI Sequence Parsing). R package version 1.6.0.
37. Barr C, Wu T, Lawrence M (2022). gmapR: An R interface to the GMAP/GSNAP/GSTRUCT suite (Version 1.40.0).

38. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. and Carey, V.J.. Software for computing and annotating genomic ranges. PLoS computational biology, 9(8), e1003118. 2013; doi:10.1371/journal.pcbi.1003118.
39. Love MI, Huber W, Anders S. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, 550. 2014; doi:10.1186/s13059-014-0550-8.
40. Lawrence M, Degenhardt J, Gentleman R (2022). VariantTools: Tools for Exploratory Analysis of Variant Calls (Version 1.40.0).
41. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M.  
“VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants.” Bioinformatics, 30(14), 2076-2078. 2014; doi:10.1093/bioinformatics/btu168.
42. Tollefson, G. A., Schuster, J., Gelin, F., Agudelo, A., Ragavendran, A., Restrepo, I., Stey, P., Padbury, J. F., & Uzun, A. VIVA (VIsualization of VArants): a VCF file visualization tool. Scientific Reports, 9(1). 2019; doi:[10.1038/s41598-019-49114-z](https://doi.org/10.1038/s41598-019-49114-z).

## Acknowledgments

The authors thank the National Institutes of Health (NIH) Office of Data Science Strategy (ODSS), and the National Center for Biotechnology Information (NCBI) for their immense support before and during the April 2023 Omics codeathon organized by the African Society for Bioinformatics and Computational Biology (ASBCB).

## Funding

The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

## Author Information

### Affiliations

**Faculty of Sciences of Tunis, University of Tunis El Manar (UTM), Tunis, Tunisia**

Hiba Ben Aribi

**Laboratory of Biomedical Genomics and Oncogenetics, Institut Pasteur de Tunis, University of Tunis El Manar, Tunisia**

Najla Abassi

**Department of Computer Science, Faculty of Science, University of Ibadan, Ibadan, Oyo State, Nigeria.**

**African Society for Bioinformatics and Computational Biology, Cape Town, South Africa.**

Olaitan I. Awe

#### **Author ORCID ID**

<b>Author Name</b>	<b>ORCID ID</b>
Hiba Ben Aribi	0000-0001-9547-8725
Najla Abassi	0000-0001-8357-0938
Olaitan I. Awe	0000-0002-4257-3611

#### **Author Contributions**

HBA: Conceptualization, Methodology, Validation, Writing

NA: Data Analysis, Methodology, Validation, Writing

OIA: Resources, Manuscript review, and Project Supervision

#### **Declarations**

##### **Ethics approval and consent to participate**

Not applicable.

##### **Consent for publication**

Not applicable.

##### **Competing interests**

The authors declare that they have no competing interests.