Data in Brief 35 (2021) 106863

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Differential gene expression data from the human central nervous system across Alzheimer's disease, Lewy body diseases, and the amyotrophic lateral sclerosis and frontotemporal dementia spectrum

Ayush Noori[a,b,c,d], Aziz M. Mezlini[b,c,d,e], Bradley T. Hyman[b,d,e], Alberto Serrano-Pozo[b,d,e,*], Sudeshna Das[b,c,d,e,*]

[a] Harvard College, Cambridge, MA 02138, United States of America
[b] Department of Neurology, Massachusetts General Hospital, Boston, MA 02114, United States of America
[c] MIND Data Science Lab, Cambridge, MA 02139, United States of America
[d] MassGeneral Institute for Neurodegenerative Disease, Charlestown, MA 02129, United States of America
[e] Harvard Medical School, Boston, MA 02115, United States of America

ARTICLE INFO

Article history:
Received 18 December 2020
Revised 5 February 2021
Accepted 8 February 2021
Available online 11 February 2021

Keywords:
Alzheimer's disease
Amyotrophic lateral sclerosis
Differential expression
Frontotemporal dementia
Lewy body diseases
Meta-analysis
Neurodegeneration
Transcriptomics

ABSTRACT

In Noori et al. [1], we hypothesized that there is a shared gene expression signature underlying neurodegenerative proteinopathies including Alzheimer's disease (AD), Lewy body diseases (LBD), and the amyotrophic lateral sclerosis and frontotemporal dementia (ALS-FTD) spectrum. To test this hypothesis, we performed a systematic review and meta-analysis of 60 human central nervous system transcriptomic datasets in the public Gene Expression Omnibus and ArrayExpress repositories, comprising a total of 2,600 AD, LBD, and ALS-FTD patients and age-matched controls which passed our stringent quality control pipeline. Here, we provide the results of differential expression analyses with data quality reports for each of these 60 datasets. This atlas of differential expression across AD, LBD, and ALS-FTD may guide future work to elucidate the pathophysiological drivers of these

DOI of original article: 10.1016/j.nbd.2020.105225
* Corresponding authors.
E-mail addresses: ASERRANO1@mgh.harvard.edu (A. Serrano-Pozo), SDAS5@mgh.harvard.edu (S. Das).
Social media: (A. Noori), (A. Serrano-Pozo), (S. Das)

https://doi.org/10.1016/j.dib.2021.106863

individual diseases as well as the common substrate of neurodegeneration.

Specifications Table

Subject	Medical Sciences, Bioinformatics
Specific subject area	Neurodegeneration, Transcriptomics
Type of data	Differential expression analyses, data quality reports, and visualizations.
How data were acquired	Systematic review of human central nervous system (CNS) transcriptomics datasets from patients with neurodegenerative diseases and healthy controls.
Data format	Analyzed
Parameters for data collection	Datasets were selected based on prespecified inclusion and exclusion criteria.
Description of data collection	Systematic review of publicly available human gene expression datasets from neuropathologically relevant CNS regions of patients with Alzheimer's disease (AD), Lewy body diseases (LBD), and the amyotrophic lateral sclerosis and frontotemporal dementia (ALS-FTD) spectrum, as well as healthy controls. Datasets were retrieved from the NCBI Gene Expression Omnibus (GEO) and EMBL-EBI ArrayExpress repositories, followed by rigorous data pre-processing and differential expression analysis.
Data source location	Massachusetts General Hospital Boston, Massachusetts United States Primary data sources are described below. AD Primary Data Sources: GEO: GSE109887, GSE139384, GSE118553, GSE132903, GSE131617, GSE122063, GSE106241, GSE84422, GSE33000, GSE48350, GSE29378, GSE44770, GSE36980, GSE13214, GSE26972, GSE37263, GSE32645, GSE28146, GSE26927, GSE16759, GSE15222, GSE12685, GSE6834, GSE5281, GSE1297; ArrayExpress: E-MEXP-2280 LBD Primary Data Sources: GEO: GSE77666, GSE49036, GSE43490, GSE54282, GSE34516, GSE23290, GSE28894, GSE26927, GSE24378, GSE20164, GSE20163, GSE20159, GSE19587, GSE20292, GSE20291, GSE20333, GSE20146, GSE20141, GSE8397, GSE7621, GSE7307 ALS-FTD Primary Data Sources: GEO: GSE139384, GSE68605, GSE56500, GSE26927, GSE20589, GSE19332, GSE18920, GSE13162, GSE4595, GSE833; ArrayExpress: E-MTAB-6189, E-MTAB-1925, E-MEXP-2280
Data accessibility	Results of differential expression analyses with data quality reports are available at our Mendeley Data repository. https://doi.org/10.17632/752nd4w7pd Source code is available on GitHub and has been archived in the Zenodo open-access repository. https://doi.org/10.5281/zenodo.4501047
Related research article	A. Noori, A.M. Mezlini, B.T. Hyman, A. Serrano-Pozo and S. Das. Systematic review and meta-analysis of human transcriptomics reveals neuroinflammation, deficient energy metabolism, and proteostasis failure across neurodegeneration. *Neurobiol. Dis.* **149**, 2021, 105225, https://doi.org/10.1016/j.nbd.2020.105225.

Value of the Data

- This dataset is a comprehensive atlas of differential gene expression data from Alzheimer's disease (AD), Lewy body diseases (LBD), and the amyotrophic lateral sclerosis and frontotemporal dementia (ALS-FTD) spectrum, providing insight into the genes and functional pathways which may drive pathogenesis uniquely in each disease, as well as collectively across all three neurodegenerative proteinopathies.

- Researchers may leverage this dataset to identify specific genes and pathways underlying neurodegeneration for future research endeavors such as pathophysiological studies and biomarker and therapeutic development.
- Future directions of this work may include comparison with single-cell and single-nuclei RNA-seq studies in postmortem specimens from patients with various neurodegenerative diseases as well as healthy subjects across the lifespan.

1. Data Description

Each gene expression dataset from the NCBI Gene Expression Omnibus (GEO) or EMBL-EBI ArrayExpress database, contained within individual directories in our Mendeley Data repository, was categorized by both disease and brain region and studied individually, entailing 89 separate analyses. For each analysis, the following data files are provided:

1. Data quality report generated via the *arrayQualityMetrics* package [2,3]. Data quality reports include dataset metadata from GEO or ArrayExpress, inter-array distance comparisons, principal component analyses, array intensity distributions, variance mean dependence, and MA plots for individual array quality. Each report can be examined by opening the index.html file included within the appropriate subdirectory.
2. Boxplot of normalized expression data. Boxplots indicate sample label, disease label, and outlier detection from the data quality report, as well as the 20^{th} percentile of expression.
3. Differential gene expression analysis. Each table lists differentially expressed genes (DEGs) along with their statistical significance, effect size, and accompanying metadata.
4. Volcano plot of DEGs. Volcano plots represent the statistical significance ($-\log_{10}$ of *p*-value) against the effect size ($\log_2$ of fold-change).

2. Experimental design, Materials and Methods

All data analyses were performed in the R programming language and statistical computing environment (version 4.0.2).

2.1. Systematic review

The methodology of our systematic review of publicly available human central nervous system (CNS) gene expression datasets from AD, LBD, and ALS-FTD patients in the GEO and ArrayExpress repositories (following PRISMA guidelines [4]) is described in detail in Noori et al. [1]. Datasets were selected based on prespecified eligibility criteria. Briefly, inclusion criteria were: (1) original datasets, and (2) human microarray datasets from neuropathologically relevant CNS regions in AD, LBD, and ALS-FTD patients as well as healthy controls. This systematic review yielded 1648 control and 1586 disease samples from 60 datasets: 26 AD, 21 LBD, and 13 ALS-FTD. After the data pre-processing and quality control steps described below, a total of 2600 samples were analyzed.

2.2. Data pre-processing and analysis

Data were pre-processed and analyzed as described in Noori et al. [1]. Briefly, for each analysis, we used the Robust Multichip Average approach from the *oligo* package [5,6] followed by the *arrayQualityMetrics* package [2,3] to normalize expression data as needed, generate data quality reports, and detect outliers. Outliers were identified via boxplots, MA plots, and inter-array

distance comparison. Samples which failed to pass any of these three outlier detection steps were discarded. The full data quality reports along with boxplots of the normalized expression data [7] are available in our Mendeley Data repository [8]. Outliers were represented by dashed lines in the boxplots. Next, probes were capped to filter for low signal, followed by surrogate variable analysis [9–11]. Finally, we performed differential expression analysis using the *limma* package [12–16] and created volcano plots of the DEGs [17]. The results of our differential expression analyses and the accompanying volcano plots are also available within our Mendeley Data repository [8].

2.3. Code availability

Source code is available on GitHub (https://github.com/ayushnoori/nd-diff-expr) and Zenodo [18].

Ethics Statement

Transcriptomics datasets included in our study were obtained from public repositories and there was no interaction with living human subjects. The downloaded transcriptomics datasets were generated from deidentified postmortem CNS tissue samples.

CRediT Author Statement

Ayush Noori: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft; **Aziz M. Mezlini:** Methodology, Writing - review & editing; **Bradley T. Hyman:** Writing - review & editing, Funding acquisition; **Alberto Serrano-Pozo:** Conceptualization, Writing - original draft, Funding acquisition; **Sudeshna Das:** Supervision, Methodology, Writing - review & editing, Funding acquisition.

Funding

This work was supported by the Alzheimer's Association (AACF-17-524184 to AS-P), the National Institute on Aging (K08AG064039 to AS-P and P30AG062421 to BTH and SD), the Rainwater Charitable Foundation (to BTH), and a MassLife Sciences MassCATS award (to BTH and SD). The funding sources had no role in study design; data collection, analysis, and interpretation; or manuscript preparation.

Declaration of Competing Interest

The authors declare no competing financial interests.

References

[1] A. Noori, A.M. Mezlini, B.T. Hyman, A. Serrano-Pozo, S. Das, Systematic review and meta-analysis of human transcriptomics reveals neuroinflammation, deficient energy metabolism, and proteostasis failure across neurodegeneration, Neurobiol. Dis. 149 (2021) 105225, doi:10.1016/j.nbd.2020.105225.
[2] A. Kauffmann, W. Huber, Microarray data quality control improves the detection of differentially expressed genes, Genomics 95 (2010) 138–142, doi:10.1016/j.ygeno.2010.01.003.
[3] A. Kauffmann, R. Gentleman, W. Huber, *arrayQualityMetrics* - a Bioconductor package for quality assessment of microarray data, Bioinformatics 25 (2009) 415–416, doi:10.1093/bioinformatics/btn647.

[4] A. Liberati, D.G. Altman, J. Tetzlaff, C. Mulrow, P.C. Gøtzsche, J.P.A. Ioannidis, M. Clarke, P.J. Devereaux, J. Kleijnen, D. Moher, The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration, The BMJ. 339 (2009), doi:10.1093/10.1136/bmj.b2700.
[5] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, T.P. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostat. Oxf. Engl. 4 (2003) 249–264, doi:10.1093/biostatistics/4.2.249.
[6] B.S. Carvalho, R.A. Irizarry, A framework for oligonucleotide microarray preprocessing, Bioinformatics 26 (2010) 2363–2367, doi:10.1093/bioinformatics/btq431.
[7] H. Wickham, *ggplot2*: Elegant Graphics For Data Analysis, 1st ed., Springer, New York, 2009 https://www.springer.com/gp/book/9780387981413.
[8] A. Noori, A.M. Mezlini, B.T. Hyman, A. Serrano-Pozo, S. Das, Differential gene expression data from the human central nervous system across Alzheimer's disease, Lewy body diseases, and the amyotrophic lateral sclerosis and frontotemporal dementia spectrum, Data in Brief, 2021, doi:10.17632/752nd4w7pd.
[9] J.T. Leek, Asymptotic conditional singular value decomposition for high-dimensional genomic data, Biometrics 67 (2011) 344–352, doi:10.1111/j.1541-0420.2010.01455.x.
[10] J.T. Leek, J.D. Storey, Capturing heterogeneity in gene expression studies by surrogate variable analysis, PLOS Genet 3 (2007) e161, doi:10.1371/journal.pgen.0030161.
[11] J.T. Leek, W.E. Johnson, H.S. Parker, A.E. Jaffe, J.D. Storey, The *sva* package for removing batch effects and other unwanted variation in high-throughput experiments, Bioinformatics 28 (2012) 882–883, doi:10.1093/bioinformatics/bts034.
[12] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, *limma* powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (2015) e47, doi:10.1093/nar/gkv007.
[13] B. Phipson, S. Lee, I.J. Majewski, W.S. Alexander, G.K. Smyth, Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression, Ann. Appl. Stat. 10 (2016) 946–963, doi:10.1214/16-AOAS920.
[14] D. Maglott, J. Ostell, K.D. Pruitt, T. Tatusova, Entrez Gene: gene-centered information at NCBI, Nucleic Acids Res 35 (2007) D26–D31, doi:10.1093/nar/gkl993.
[15] C.J. Walsh, P. Hu, J. Batt, C.C.D. Santos, Microarray meta-analysis and cross-platform normalization: integrative genomics for robust biomarker discovery, microarrays Basel Switz 4 (2015) 389–406, doi:10.3390/microarrays4030389.
[16] X. Wang, Y. Lin, C. Song, E. Sibille, G.C. Tseng, Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: with application to major depressive disorder, BMC Bioinformatics 13 (2012) 52, doi:10.1186/1471-2105-13-52.
[17] K. Blighe, S. Rana, M. Lewis, *EnhancedVolcano*: Publication-Ready Volcano Plots With Enhanced Colouring and Labeling, 2020. https://github.com/kevinblighe/EnhancedVolcano.
[18] A. Noori, ayushnoori/nd-diff-expr, Zenodo, (2021), doi:10.5281/zenodo.4501047.