

MousiPLIER: A Mouse Pathway-Level Information Extractor Model

Shuo Zhang,^{1,2*} Benjamin J. Heil,^{3*} Weiguang Mao,⁴ Maria Chikina,⁴ Casey S. Greene,^{1,5} and Elizabeth A. Heller^{1,2}

¹Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, ²Penn Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, ³Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, ⁴Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, and ⁵Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Denver, Colorado 80045

Abstract

High-throughput gene expression profiling measures individual gene expression across conditions. However, genes are regulated in complex networks, not as individual entities, limiting the interpretability of gene expression data. Machine learning models that incorporate prior biological knowledge are a powerful tool to extract meaningful biology from gene expression data. Pathway-level information extractor (PLIER) is an unsupervised machine learning method that defines biological pathways by leveraging the vast amount of published transcriptomic data. PLIER converts gene expression data into known pathway gene sets, termed latent variables (LVs), to substantially reduce data dimensionality and improve interpretability. In the current study, we trained the first mouse PLIER model on 190,111 mouse brain RNA-sequencing samples, the greatest amount of training data ever used by PLIER. We then validated the mousiPLIER approach in a study of microglia and astrocyte gene expression across mouse brain aging. mousiPLIER identified biological pathways that are significantly associated with aging, including one latent variable (LV41) corresponding to striatal signal. To gain further insight into the genes contained in LV41, we performed *k*-means clustering on the training data to identify studies that respond strongly to LV41. We found that the variable was relevant to striatum and aging across the scientific literature. Finally, we built a Web server (<http://mousiplier.greenelab.com/>) for users to easily explore the learned latent variables. Taken together, this study defines mousiPLIER as a method to uncover meaningful biological processes in mouse brain transcriptomic studies.

Received Aug. 23, 2023; revised May 10, 2024; accepted May 16, 2024.

The authors declare no competing financial interests.

Author contributions: S.Z., B.J.H., W.M., M.C., C.S.G., and E.A.H. designed research; S.Z., B.J.H., W.M., C.S.G., and E.A.H. performed research; S.Z., B.J.H., W.M., and C.S.G. analyzed data; S.Z., B.J.H., C.S.G., and E.A.H. wrote the paper.

This work was supported by the Gordon and Betty Moore Foundation (GBMF4552 to C.S.G.), the National Human Genome Research Institute (R01HG10067 to C.S.G.), National Institute of Drug Abuse (R01DA052465 to E.A.H.), and SynGAP Research Fund Postdoctoral Fellowship (S.Z.).

Continued on next page.

Key words: aging; gene expression; machine learning

Significance Statement

RNA-sequencing studies define differential expression of individual genes across conditions. However, genes are regulated in complex networks, not as individual entities. Machine learning models that incorporate biological pathway information are a powerful tool to analyze human gene expression. However, such models are lacking for mouse, despite the vast number of mouse RNA-seq datasets. We trained a mouse pathway-level information extractor model (mousiPLIER) to reduce data dimensionality from over 10,000 genes to 196 “latent variables” that map to known biological pathways. To validate this approach, we applied mousiPLIER to differential expression across mouse brain aging. We identified 26 functional pathways (latent variables) that varied across aging. Finally, we developed a Web server to facilitate use of mousiPLIER by the scientific community.

Introduction

Over the last decade, scientists have generated an astronomical amount of brain gene expression data (Carulli et al., 1998; Anders and Huber, 2010; Costa-Silva et al., 2017; Keil et al., 2018; Y. Zhang et al., 2021). Differential gene expression analysis of high-throughput RNA-sequencing data is commonly applied to interrogate the relative enrichment of a single transcript across samples. However, genes are regulated in complex networks, rather than as individual entities. Furthermore, gene expression profiling studies are limited in statistical power, as they tend to examine relatively few samples compared with the number of expressed transcripts and increasing the number of samples can be prohibitively expensive.

Machine learning models that incorporate prior pathway information have shown great power in analyzing human gene expression. To this end, we apply an unsupervised learning method that (1) reduces the dimensionality and/or (2) incorporates additional published gene expression datasets. Unsupervised machine learning is a method that defines the structure of “unlabeled data”, for which information on the biological context and experimental conditions is removed. Such methods are well suited for gene expression data and are often used for tasks such as reducing the dimensionality of expression datasets (Hotelling, 1933; der Maaten and Hinton, 2008; McInnes et al., 2018), clustering samples (Oyelade et al., 2016; Chen et al., 2020), or learning shared expression patterns across experiments (Tan et al., 2016; Handl et al., 2019). While unsupervised machine learning models are capable of analyzing large amounts of unlabeled expression data, many of them do not explicitly encode prior biological knowledge to encourage the model to learn biologically meaningful patterns of gene expression over technical ones.

A novel approach, the modeling framework pathway-level information extractor (PLIER; Mao et al., 2019), is built explicitly to work on expression data and uses matrix factorization to incorporate prior biological knowledge in the form of sets of genes corresponding to biological pathways or cell type markers. This approach converts gene expression data into a series of values called “latent variables” (LVs) that correspond to potentially biologically relevant combinations of differentially expressed genes. PLIER learns diverse biological pathways from entire compendia of expression data and can transfer that knowledge to smaller studies, such as MultiPLIER (Taroni et al., 2019). However, PLIER models are largely trained on a single dataset rather than a compendium (Rubenstein et al., 2020; Stogsdill et al., 2022; Z. Zhang et al., 2022), and past MultiPLIER runs have only trained models with up to tens of thousands of samples (Taroni et al., 2019; Banerjee et al., 2020).

To expand the application and utility of PLIER for identifying meaningful biological pathways from gene expression data, we trained a PLIER model on a compendium of mouse gene expression data. In doing so, we trained the first mouse compendium PLIER model (mousiPLIER), on the greatest amount of training data (190,111 samples) ever used by this model. We demonstrated successful optimization of the model training, which generated hypotheses on regulation of mouse brain aging. A further innovation applied *k*-means clustering in the latent variable space to identify the microglia-associated latent variables that corresponded to aging-related changes in the training data. Finally, to maximize widespread usability of mousiPLIER, we built a Web server that allows others to visualize the results and find patterns in the data based on their own latent variables of interest. Going forward, this model and its associated Web server will be a useful tool for better understanding mouse gene expression.

Materials and Methods

Data. We began by downloading all the mouse gene expression data in Recount3, along with its corresponding metadata (Wilks et al., 2021). We then removed the single-cell RNA-seq data from the dataset to ensure our data sources were consistent across samples and studies. A total of 190,111 samples from mice of either sex were left for downstream processing. Next, we filtered the expression data, keeping only genes that overlapped between Recount3 and our prior knowledge gene sets. Then, we normalized the expression into TPM (transcripts per million) using gene lengths from the Ensembl BioMart database (Howe et al., 2021). Finally, we Z-scored the expression data to ensure a consistent range for the downstream PLIER model.

For our prior-knowledge gene sets, we used cell type marker genes from CellMarker (X. Zhang et al., 2019), pathway gene sets from Reactome (Gillespie et al., 2022), and

*S.Z. and B.J.H. contributed equally to this work.

W.M.'s present address: Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, New York 10010

Correspondence should be addressed to Casey S. Greene at casey.s.greene@cunhschutz.edu or Elizabeth A. Heller at eheller@pennmedicine.upenn.edu.

Copyright © 2024 Zhang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license, which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

manually curated brain marker genes from Allen Mouse Brain Atlas (<https://mouse.brain-map.org>; Lein et al., 2007). We selected cell type marker genes corresponding to all available mouse cell types within the CellMarker database. For mouse biological pathways, we downloaded pathway information from the Reactome database. More specifically, we processed the files “Ensembl2Reactome_All_Levels.txt,” “ReactomePathways.txt,” and “ReactomePathwaysRelation.txt,” selecting only pathways using mouse genes, filtering out all pathways with fewer than five genes present, and keeping only pathways that were leaf nodes on the pathway hierarchy. Because we were interested in mouse brains in particular, we rounded out our set of prior information by manually selecting marker genes for the striatum, midbrain, and cerebral cortex. In total, we used 1,003 prior knowledge pathways when training our model.

PLIER. The mousiPLIER is built on PLIER, which transforms gene expression data into latent variable space with prior biological pathways incorporated using matrix factorization (Mao et al., 2019). The inputs for PLIER are gene expression matrix (Y , genes as row and samples as columns), and prior knowledge matrix (C , genes as rows and gene sets as columns). For a given Y and C , PLIER tries to find loadings for LVs (Z , genes as rows and LVs as columns), representation of the original data in latent variable space (B , LVs as rows and samples as columns), and an assignment of gene sets to LVs (U , gene sets as rows and LVs as columns) by minimizing the following formula:

$$\|Y - ZB\|_F^2 + \lambda_1 \|Z - CU\|_F^2 + \lambda_2 \|B\|_F^2 + \lambda_3 \|U\|_{L^1},$$

where λ_1 , λ_2 , and λ_3 are the parameters. The first term represents reconstruction error when converting expression data from gene space to latent variable space. The second term forces latent variables to align with prior knowledge gene sets. The third one is L^2 penalty on B to ensure no single LV explained too much. The final term is L^1 penalty on U to ensure that a LV is only associated with a few gene sets.

Due to the large size (~40 GB) of preprocessed Recount3 expression data, we began the PLIER pipeline by precomputing the initialization for PLIER with incremental principal component analysis (PCA) in scikit-learn (Pedregosa et al., 2011). We then used the expression compendium, prior knowledge gene sets, and PCA initializations to train a PLIER model with default parameters. The resulting task took 2 d to run and yielded 196 latent variables.

RNA-seq processing. RNA-seq reads from male mouse microglia and astrocytes (Pan et al., 2020) were mapped to mm10 reference genome using STAR (v2.7.1a; Dobin et al., 2013) with parameters: `-outFilterMismatchNmax 3 -outFilterMultimapNmax 1 -alignSJoverhangMin 8`. Gene level read counts were prepared using featureCounts (subread v1.6.1; Liao et al., 2014). The gene annotation file used in featureCounts was downloaded from Recount3 (<https://rna.recount.bio/docs/raw-files.html#annotation-files>). The gene expression data were TPM normalized and Z-scored in the same way as Recount3 training dataset.

LV significance for mouse aging RNA-seq data. We first transformed the mouse aging expression data from gene space (Y_{target}) to latent variable space (B_{target}) using a custom Python script based on this equation: $B_{\text{target}} = (Z^T Z + \lambda_2 I)^{-1} Z^T Y_{\text{target}}$ (as in Taroni et al., 2019), where I is an identity matrix. To determine which latent variables were associated with aging in each disease and cell type (WT microglia, WT astrocyte, AD microglia, and AD astrocyte), we used a linear model. In the model, we look at LV expression as a function of mouse age for each LV by treating development stage (in month) as a numerical variable. To correct the p values for multiple testing, we used the Benjamini–Hochberg procedure (FDR; Benjamini and Hochberg, 1995). LVs with FDR < 0.05 were considered to be significantly associated. Overlap of significant LVs was plotted with nVennR (Pérez-Silva et al., 2018). Gene sets associated with each LV were visualized using pheatmap (Kolde, 2019).

Clustering. We selected the latent variables significantly associated with aging in mouse microglia as a biological starting point. We then used these latent variables to query the training data and see which studies seemed associated with the same biological signals. To do so, we used k -means clustering with a k of 2, to look for experiments where there was some experimental condition that affected the latent variable. We then ranked the top 10 studies based on their silhouette scores and looked to see which conditions were associated with relevant experimental variables.

Hardware and software. The PLIER model training was performed on the Penn Medicine high-performance computing cluster running CentOS v7.8. We used R v4.1.0 and PLIER v0.1.6 for the pipeline. The full pipeline takes ~2 weeks to run, with the main bottlenecks being the Recount3 data download, which takes 1 week to run, and training the PLIER model, which takes 2 d on a compute node (Dell R940 big memory system) with 250 GB of random access memory (RAM). Transforming mouse aging expression on to LV space was performed on Dell C6420 Quad node systems. This step can also be easily accomplished on a personal computer.

Web server. The Web server for visualizing the results was built on top of the ADAGE (Analysis using Denoising Autoencoders of Gene Expression) Web app framework (Tan et al., 2017). The main changes we made were to substitute

the latent variables and gene sets from our trained PLIER model and to forgo uploading the input expression data as the mouse compendium we used was much larger than the input expression for ADAGE.

Data and code availability. All data and code used in this study can be found at <https://github.com/greenelab/mousiplier>.

Results

MousiPLIER learned latent variables with ideal pathway-level and gene-level sparsity

We trained mousiPLIER using on-disk PCA implementation to initialize PLIER, modified the pipeline to work with mouse data, and used a high-memory compute node to manage the size of the matrix decomposition (see Materials and Methods). The resulting model had 196 latent variables with ideal pathway-level and gene-level sparsity. The per latent variable distribution had an average of 65% sparsity, such that the latent variables tended to use only ~35% of the genes in the training data (Fig. 1A). While many of the latent variables corresponded to no pathways, indicating signals in the training data not passed in as prior knowledge, those that remained corresponded to few pathways (Fig. 1B). This optimal pathway-level and gene-level sparsity allowed us to interrogate individual latent variables that corresponded to a small number of biological functions.

MousiPLIER identified LVs were associated with aging

To validate the utility of mousiPLIER, we next interrogated brain-relevant latent variables that our mousiPLIER learned from the training data. To this end, we analyzed an individual study on mouse brain aging (Pan et al., 2020). This study measures wild-type and Alzheimer's disease (APP-PS1) mouse gene expression in microglia and astrocytes at five ages across adulthood. We first projected the RNA-seq data from this study (gene space) to mousiPLIER (LV space). Then, we used a linear model to identify the latent variables that changed significantly across developmental aging. mousiPLIER identified a specific set of significantly changed LVs in each condition in the study (Fig. 2A, Table 1). These mousiPLIER-learned LVs are aligned to diverse, prior-knowledge gene sets (Fig. 2B). In particular, latent variable 41 uniquely corresponds to striatal signal. This latent variable decreased throughout aging in wild-type microglia. Top-weighted genes of latent variable 41 were functionally associated with striatal cell type specificity (Fig. 2C,D). Previous studies show that STEP (encoded by *Ptpn5*) and PDE10A (encoded by *Pde10a*) protein levels decline in striatum during aging (Fazio et al., 2017; Cases et al., 2018). The identification of these genes in LV41 indicate that microglia might exhibit molecular processes that occur in aging. As latent variable 41 is mapped to a single gene set and has a potential role in aging, we focus on this latent variable in the rest of the study.

Latent variable 41 demonstrated the biological relevance of mousiPLIER latent variables

Having identified microglia-associated latent variables of interest, we next sought to validate the relevance of this gene set by finding which studies in the mousiPLIER training data responded strongly to them. To do so, we developed a novel method to rank studies based on their latent variable weights. More precisely, we performed k -means clustering with a k of two on each study in each latent variable space and ranked studies by their silhouette scores, a metric measuring the degree to which clusters are separated from each other. Using this approach, we identified studies that contained samples distinguishable by their values for our latent variables of interest.

We focused this approach on latent variable 41. We found that many of the studies with the highest silhouette scores for latent variable 41 indicated processes occurring in the brain (Fig. 3A). We dug deeper into which specific samples were

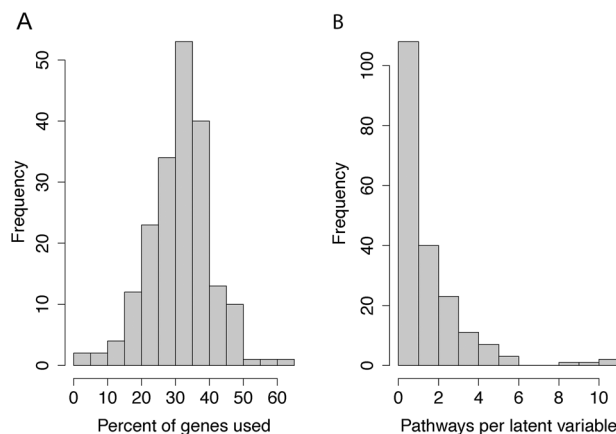


Figure 1. mousiPLIER learned latent variables with ideal pathway-level and gene-level sparsity. **A**, The distribution of the percentage of genes from the training set used per latent variable. **B**, The distribution of the number of prior knowledge gene sets used per latent variable.

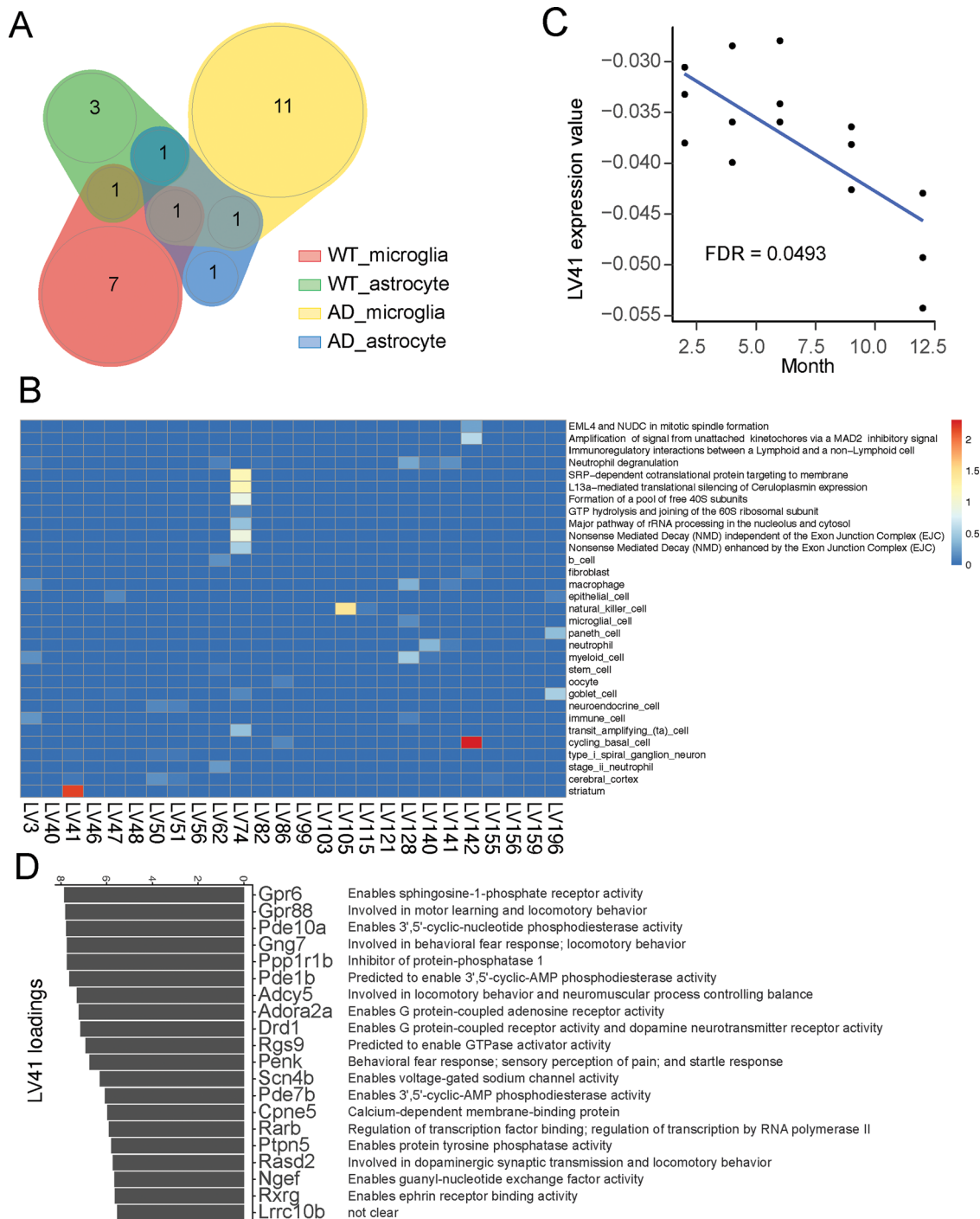


Figure 2. mousiPLIER identified LVs associated with aging. **A**, Venn diagram showing the number of significant LVs and their overlap across cell types and experimental conditions. **B**, LV41 is decreased significantly in wild-type microglia during aging. **C**, A heatmap showing significant LV-associated biological pathways or cell type markers. A linear model is used to test the effect of aging on each LV. *p* values are adjusted for multiple comparisons using Benjamini-Hochberg method. An LV is differentially expressed if FDR < 0.05. LV, latent variable; AD, Alzheimer’s disease; WT, wild-type. **D**, Top 20 genes with highest weight associated with LV41 and their potential annotation. LV, latent variable.

present in each cluster and found that latent variable 41 was in fact learning something brain (and more specifically striatum) related (Fig. 3).

For example, in a study to delineate tissue-specific transcriptional consequences of copy number variant within 16p11.2, a common cause of autism spectrum disorder, gene expression data were profiled from three genotypes (wild type, deletion,

Table 1. Significantly changed latent variables in each condition

LV_ID	Genotype	Region	Adjusted <i>p</i> value
LV3	WT	Microglia	0.0385
LV40	AD	Astrocyte	0.0408
LV40	AD	Microglia	0.0159
LV41	WT	Microglia	0.0493
LV46	WT	Astrocyte	0.0416
LV46	WT	Microglia	0.0159
LV47	WT	Astrocyte	0.0165
LV48	WT	Microglia	0.0385
LV50	AD	Astrocyte	0.0385
LV51	WT	Microglia	0.0377
LV56	AD	Microglia	0.0165
LV62	AD	Microglia	0.0398
LV74	AD	Microglia	0.0422
LV82	AD	Microglia	0.0061
LV86	WT	Microglia	0.0398
LV99	AD	Astrocyte	0.0385
LV99	AD	Microglia	0.0159
LV99	WT	Microglia	0.0086
LV103	AD	Microglia	0.0385
LV105	AD	Microglia	0.0398
LV115	WT	Microglia	0.0007
LV121	AD	Microglia	0.0398
LV128	AD	Microglia	0.0086
LV140	AD	Microglia	0.0385
LV141	AD	Microglia	0.0214
LV142	WT	Astrocyte	0.0057
LV155	AD	Astrocyte	0.0477
LV155	WT	Astrocyte	0.0165
LV156	WT	Astrocyte	0.0497
LV159	WT	Microglia	0.0165
LV196	AD	Microglia	0.0165

Gene expression data were converted into latent space and tested for differential expression during aging. WT, wild type; AD, Alzheimer disease; LV, latent variable.

and duplication of the 16p11.2 region) and six tissues (brown fat, liver, white fat, cerebellum, cortex, and striatum; <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76872>). The LV experimental values in the striatal samples, irrespective of the genotype, clearly stand apart from the other tissues (Fig. 3B). Additionally, in a study to investigate transcriptional effects of selected long noncoding RNAs (lncRNAs), mRNA expression is generated from embryonic and adult whole brains of wild-type and lncRNA knock-out mouse (Goff et al., 2015). LV41 expression is higher in adult (7.6–14.1 weeks) samples compared with embryonic day 14.5 timepoint regardless of knock-out status (Fig. 3C), supporting the association between latent variable 41 and aging found in the study (Pan et al., 2020) we used to derive the latent variables.

Web server

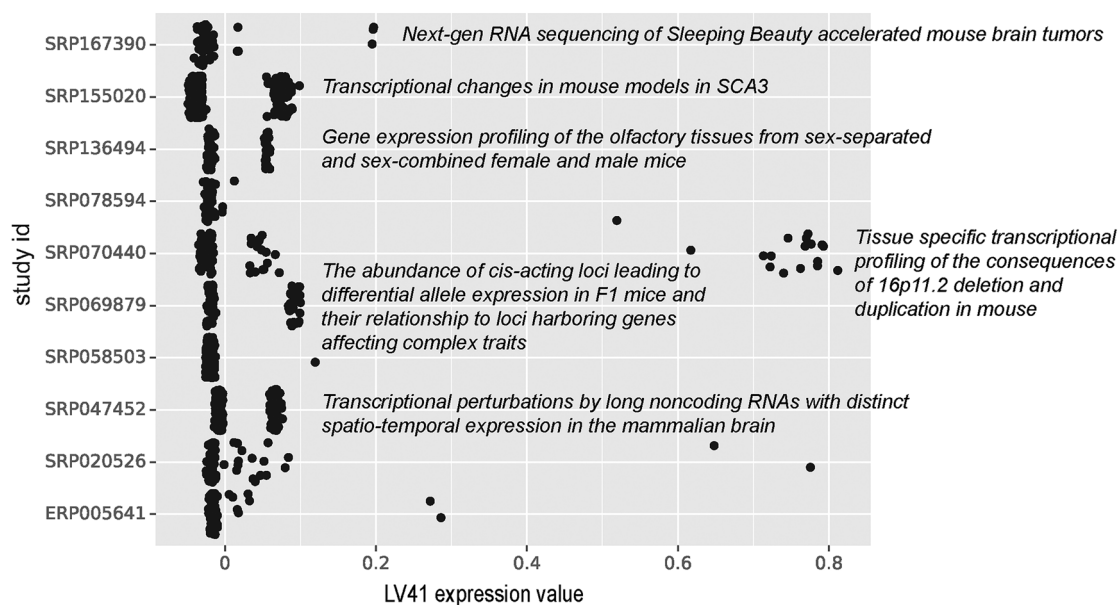
To allow others to independently examine mousiPLIER learned latent variables, we developed a Web server at <http://mousiplier.greenelab.com/>. This server allows users to list the genes present in, visualize which experiments had high cluster scores for, and see which biological pathways participate in each latent variable (Fig. 4).

Discussion

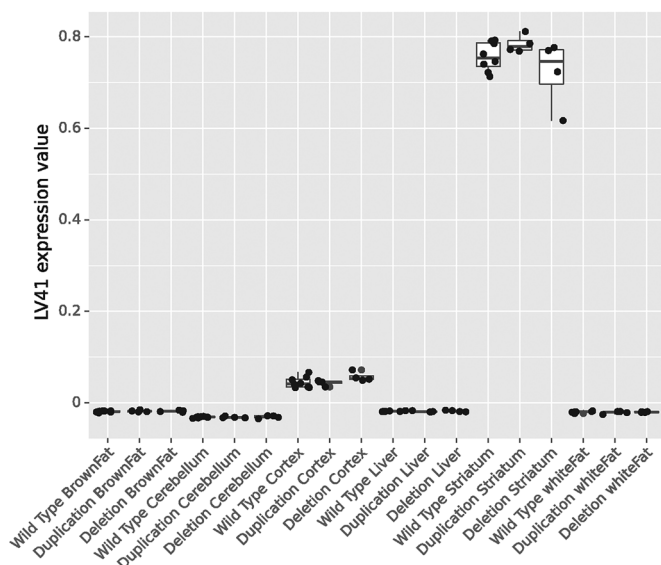
In this paper we developed mousiPLIER and established proof of concept for training extremely large PLIER models on mouse data. The learned latent variables mapped to various biological processes and cell types. Further, we applied a novel approach for surfacing latent-variable relevant experiments from an expression compendium. Specifically, we clustered training experiments based on latent variable values, allowing us to query a large compendium for experiments pertaining to mouse striatal aging. Finally, we created a Web server to make the model's results more easily accessible to other scientists.

Although we focused our analysis on LV41 to validate the utility of mousiPLIER, we identified other significantly changed LVs that are associated with aging-relevant pathways. For example, LV142 is associated with cycling basal cells (Fig. 2B) and is significantly decreased in WT astrocytes. The top weighted genes of LV142 contain cell division genes, such as Cytoskeleton-associated protein 2-like (Ckap2l) and Nucleolar and spindle-associated protein 1 (Nusap1). Association of such cell cycle markers in basal cells is consistent with their high turnover rate. The observation that WT astrocytes

A



B



C

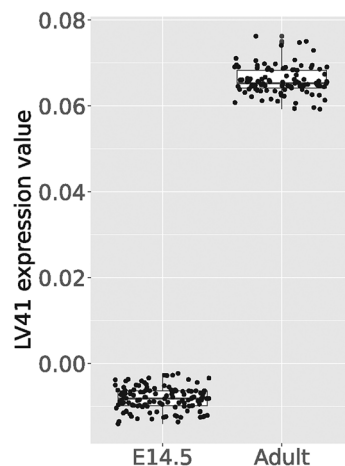


Figure 3. Latent variable 41 demonstrated the biological relevance of mousiPLIER latent variables. **A**, Studies with the 10 highest silhouette scores after clustering according to LV41 expression values. **B**, LV41 expression values are higher for striatal tissue than other tissues in SRP070440. **C**, Effects of development on LV41. Samples are collapsed based on developmental timepoints.

exhibit decreased LV142 in aged mice is likely due to reduced local proliferation of astrocytes during aging, as shown in mouse dentate gyrus (Schneider et al., 2022). In addition, LV74 is significantly increased in AD microglia. Top weighted genes in LV74 encode ribosomal proteins, and LV74 is associated with several pathways related to mRNA translation and protein translocation (Fig. 2B). One of the associated pathways is signal recognition particle (SRP)–dependent cotranslational protein targeting to membrane, which is the top enriched pathway for differentially expressed genes between AD microglia and healthy microglia (Wang and Li, 2021). More specifically, increased microglial expression of genes in this pathway is associated with more severe AD pathology (Patel et al., 2020). This is consistent with our result that the expression of LV74 is gradually increased in AD microglia during aging, highlighting the importance of this pathway in AD progression. Further study of these LVs may provide clues to biological pathways relevant to aging that are utilized for different functions across distinct cell types.



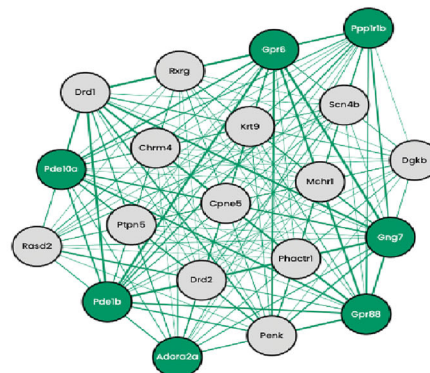
A

Explore Genes

Selected Genes

Standard Name	Systematic Na...	Entrez ID	Entrez Description
× Gpr6	Gpr6	140741	G protein-coupled receptor 6
× Gpr88	Gpr88	64378	G-protein coupled receptor 88
× Pdel0a	Pdel0a	23984	phosphodiesterase 10A
× Gng7	Gng7	14708	guanine nucleotide binding protein...
× Ppp1rb	Ppp1rb	19049	protein phosphatase 1, regulatory i...
× Pdelb	Pdelb	18574	phosphodiesterase 1B, Ca2+-calmo...
× Adora2a	Adora2a	11540	adenosine A2a receptor

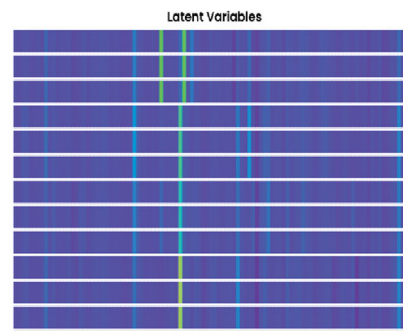
Deselect All × Download ↓



B

Explore Experiments

SRP136492	Group	Sample
SRP136492	12	266364
SRP136492	12	266365
SRP136492	12	266366
SRP136492	12	266367
SRP136492	12	266368
SRP136492	12	266369
SRP136492	12	266370
SRP136492	25	266371
SRP136492	12	266372
SRP136492	12	266373
SRP136492	12	266374
SRP136492	12	266375



C

Explore Latent Variables

Standard N...	Systematic ...	Entrez ID	Entrez Description	Weight ↑
Gpr6	Gpr6	140741	G protein-coupled receptor 6	7.854
Gpr88	Gpr88	64378	G-protein coupled receptor 88	7.809
Pdel0a	Pdel0a	23984	phosphodiesterase 10A	7.778
Gng7	Gng7	14708	guanine nucleotide binding protein (L...	7.745
Ppp1rb	Ppp1rb	19049	protein phosphatase 1, regulatory inh...	7.740
Pdelb	Pdelb	18574	phosphodiesterase 1B, Ca2+-calmod...	7.625
Adora2a	Adora2a	11540	adenosine A2a receptor	7.217

Download ↓

Experiment	Activities	Activiti...	Sampl...	Min	Max	Range	Silho... ↑
ERP005641	█	100	484	-0.022	0.286	0.308	0.976
SRP058503	█	205	205	-0.026	0.120	0.145	0.970
SRP078594	█	48	48	-0.029	0.519	0.548	0.967
SRP136494	█	72	72	-0.025	0.060	0.084	0.961
SRP070440	█	96	96	-0.035	0.812	0.847	0.959
SRP020526	█	60	61	-0.025	0.776	0.800	0.958
SRP069879	█	69	71	-0.024	0.100	0.124	0.957

Download ↓

Figure 4. Snapshot of mousiPLIER Web server showing its functions. **A**, An example to visualize the gene network of LV41 associated genes. **B**, An example to explore experiments and view samples' activities in mousiPLIER latent space. **C**, An example to explore latent variables. Gene weights and the clustering of experiments for each LV is displayed and can be easily downloaded.

Of note, LV115 and LV105 are significantly increased in WT microglia and AD microglia, respectively. This latent variable is associated with the natural killer (NK) cell marker (Fig. 2B). We suspect that the isolated microglia contain a small portion of NK cells as CD45^{low-to-intermediate} and CD11b (markers used in Pan et al., 2020) are also present in NK cells (Chiossone et al., 2009; He et al., 2016). Moreover, the expression of CD45 from microglia can be upregulated in aged brain (Honarpisheh et al., 2020), reducing the specificity of CD45 expression to select microglia during aging. Our speculation is further supported by the dramatic increase of NK cells in aged mouse brain (18 months old) compared with 3-month-old mice (Jin et al., 2021). Finally, high NK cell number, though lower than wild type, is also observed in 7–8 months 3xTa-AD mouse brain (Y. Zhang et al., 2020), consistent with the increase of NK signal during aging.

Traditional pathway analyses are a powerful tool to gain biological insights from RNA-seq data. However, there are limitations for such methods. First, for pathway analysis tools using a subset of differentially expressed genes (DEGs), there is

no standard cutoff for selecting DEGs (Khatri et al., 2012). Second, examining ~20,000 expressed genes in mouse, for example, imposes a high multiple testing burden. Finally, although a ranked gene list could be theoretically applied provided to existing pathway analysis tools, these packages cannot distinguish true pathway signal from batch effects due noise generated by the technical variability common in RNA-sequencing (Leek et al., 2010). mousiPLIER overcomes those issues by transforming gene expression data into latent variable space. The resulting space only has 196 dimensions, greatly reducing multiple testing burden. Moreover, mousiPLIER reduces technical noises by separating technical noises into pathway-irrelevant LVs, which is extremely valuable in comparing transcriptomic data from different studies or laboratories. Researchers can test decreased or increased pathway-associated LVs directly at the LV space. After finding LVs significantly changed among specific experimental conditions, users can interrogate the training dataset to see if the LVs demonstrate biological relevance as we did for the latent variable 41 in the manuscript. Users may further pick genes based on the loadings of the LVs for experimental examination.

Although mousiPLIER is trained on a compendium of bulk RNA-seq data, it could be potentially applied to single-cell transcriptomic datasets. The input for mousiPLIER is a read count table where rows are genes and columns are samples. But the result of a typical single-cell RNA-seq data is a table with columns being sequenced cells, and single-cell level count is too sparse to be used by mousiPLIER directly. Users can aggregate single-cell level count into sample level count for a specific cell type and then apply the pseudobulk samples to mousiPLIER.

While we successfully transformed a study from outside the training data into the latent space and identified study-specific latent variables, application of mousiPLIER was not universally successful across transcriptomic studies, such as RNA-seq data from drug addiction (Carpenter et al., 2020). This may be due to high variance across samples in training compendium, too few samples in the study of interest to generate sufficient statistical power, or other factors. In these cases, there is not yet a method to select meaningful latent variables to guide downstream analyses.

Finally, as a linear model, PLIER can only approximate nonlinear relationships between the genes used to train the model and the learned biological pathways. While we do not expect this to have a large impact (Heil et al., 2023), incorporating prior knowledge into nonlinear models such as neural networks is an exciting field of research and a potential improvement for the MultiPLIER and mousiPLIER frameworks. Going forward, our model and Web server will allow scientists to explore the latent space of their own experiments and learn about relevant biological pathways and cell types.

References

- Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11:R106.
- Banerjee J, et al. (2020) Integrative analysis identifies candidate tumor microenvironment and intracellular signaling pathways that define tumor heterogeneity in NF1. *Genes* 11:226.
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 57:289–300.
- Carpenter MD, Hu Q, Bond AM, Lombroso SI, Czarnecki KS, Lim CJ, Song H, Wimmer ME, Pierce RC, Heller EA (2020) Nr4a1 suppresses cocaine-induced behavior via epigenetic regulation of homeostatic target genes. *Nat Commun* 11:504.
- Carulli JP, et al. (1998) High throughput analysis of differential gene expression. *J Cell Biochem* 72:286–296.
- Cases S, Saavedra A, Tyebji S, Giralt A, Alberch J, Pérez-Navarro E (2018) Age-related changes in STriatal-enriched protein tyrosine phosphatase levels: regulation by BDNF. *Mol Cell Neurosci* 86:41–49.
- Chen R, Yang L, Goodison S, Sun Y (2020) Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. *Bioinformatics* 36:1476–1483.
- Chiossone L, Chaix J, Fuseri N, Roth C, Vivier E, Walzer T (2009) Maturation of mouse NK cells is a 4-stage developmental program. *Blood* 113:5488–5496.
- Costa-Silva J, Domingues D, Lopes FM (2017) RNA-seq differential expression analysis: an extended review and a software tool. *PLoS One* 12:e0190152.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.
- Fazio P, et al. (2017) Patterns of age related changes for phosphodiesterase type-10A in comparison with dopamine D2/3 receptors and sub-cortical volumes in the human basal ganglia: a PET study with 18F-MNI-659 and 11C-raclopride with correction for partial volume effect. *Neuroimage* 152:330–339.
- Gillespie M, et al. (2022) The reactome pathway knowledgebase 2022. *Nucleic Acids Res* 50:D687–D692.
- Goff LA, et al. (2015) Spatiotemporal expression and transcriptional perturbations by long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* 112:6855–6862.
- Handl L, Jalali A, Scherer M, Eggeling R, Pfeifer N (2019) Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data. *Bioinformatics* 35:i154–i163.
- He H, Geng T, Chen P, Wang M, Hu J, Kang L, Song W, Tang H (2016) NK cells promote neutrophil recruitment in the brain during sepsis-induced neuroinflammation. *Sci Rep* 6:27711.
- Heil BJ, Crawford J, Greene CS (2023) The effect of non-linear signal in classification problems using gene expression. *PLoS Comput Biol* 19:e1010984.
- Honarpisheh P, et al. (2020) Potential caveats of putative microglia-specific markers for assessment of age-related cerebrovascular neuroinflammation. *J Neuroinflammation* 17:366.
- Hotelling H (1933) Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 24:417.
- Howe KL, et al. (2021) Ensembl 2021. *Nucleic Acids Res* 49:D884–D891.
- Jin W-N, Shi K, He W, Sun J-H, Van Kaer L, Shi F-D, Liu Q (2021) Neuroblast senescence in the aged brain augments natural killer cell cytotoxicity leading to impaired neurogenesis and cognition. *Nat Neurosci* 24:61–73.
- Keil JM, Qalieh A, Kwan KY (2018) Brain transcriptome databases: a user's guide. *J Neurosci* 38:2399–2412.
- Khatri P, Sirota M, Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* 8:e1002375.
- Kolde R (2019) pheatmap: pretty heatmaps. R package version 1.0.12.
- Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11:733–739.

- Lein ES, et al. (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445:168–176.
- Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923–930.
- Mao W, Zaslavsky E, Hartmann BM, Sealfon SC, Chikina M (2019) Pathway-level information extractor (PLIER) for gene expression data. *Nat Methods* 16:607–610.
- McInnes L, Healy J, Melville J (2018) Umap: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426*.
- Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, Achas M, Adebisi E (2016) Clustering algorithms: their application to gene expression data. *Bioinform Biol Insights* 10:237–253.
- Pan J, Ma N, Yu B, Zhang W, Wan J (2020) Transcriptomic profiling of microglia and astrocytes throughout aging. *J Neuroinflammation* 17:97.
- Patel S, Howard D, Man A, Schwartz D, Jee J, Felsky D, Pausova Z, Paus T, French L (2020) Donor-specific transcriptomic analysis of Alzheimer's disease-associated hypometabolism highlights a unique donor, ribosomal proteins and microglia. *eNeuro* 7: ENEURO.0255-20.2020.
- Pedregosa F, et al. (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830.
- Pérez-Silva JG, Araujo-Voces M, Quesada V (2018) nVenn: generalized, quasi-proportional Venn and Euler diagrams. *Bioinformatics* 34:2322–2324.
- Rubenstein AB, et al. (2020) Single-cell transcriptional profiles in human skeletal muscle. *Sci Rep* 10:229.
- Schneider J, et al. (2022) Astrogenesis in the murine dentate gyrus is a life-long and dynamic process. *EMBO J* 41:e110409.
- Stogsdill JA, Kim K, Binan L, Farhi SL, Levin JZ, Arlotta P (2022) Pyramidal neuron subtype diversity governs microglia states in the neocortex. *Nature* 608:750–756.
- Tan J, Hammond JH, Hogan DA, Greene CS (2016) ADAGE-based integration of publicly available pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. *mSystems* 1:e00025-15.
- Tan J, Huyck M, Hu D, Zelaya RA, Hogan DA, Greene CS (2017) ADAGE signature analysis: differential expression analysis with data-defined gene sets. *BMC Bioinformatics* 18:512.
- Taroni JN, Grayson PC, Hu Q, Eddy S, Kretzler M, Merkel PA, Greene CS (2019) MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *Cell Syst* 8:380–394.
- Van Der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9:2579–2625.
- Wang X-L, Li L (2021) Cell type-specific potential pathogenic genes and functional pathways in Alzheimer's disease. *BMC Neurol* 21:381.
- Wilks C, et al. (2021) recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol* 22:323.
- Zhang X, et al. (2019) CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res* 47:D721–D728.
- Zhang Y, et al. (2020) Depletion of NK cells improves cognitive function in the Alzheimer disease mouse model. *J Immunol* 205:502–510.
- Zhang Z, et al. (2022) Single nucleus transcriptome and chromatin accessibility of postmortem human pituitaries reveal diverse stem cell regulatory mechanisms. *Cell Rep* 38:110467.
- Zhang Y, Thompson KN, Huttenhower C, Franzosa EA (2021) Statistical approaches for differential expression analysis in meta-transcriptomics. *Bioinformatics* 37:i34–i41.