Title Page

Kaleidoscope: A New Bioinformatics Pipeline Web Application for In Silico Hypothesis Exploration of Omics Signatures

Khaled Alganem[1], Rammohan Shukla[1], Hunter Eby[1], Mackenzie Abel[1], Xiaolu Zhang[1], William Brett McIntyre[2-3], Jiwon Lee[2], Christy Au-Yeung[2], Roshanak Asgariroozbehani[2-3], Roshni Panda[2], Sinead M O'Donovan[1], Adam Funk[1], Margaret Hahn[2-6], Jarek Meller[7-11], Robert McCullumsmith[1,7*]

[1]Department of Neurosciences, University of Toledo College of Medicine, Toledo, Ohio, USA
[2]Centre for Addiction and Mental Health, Toronto, Ontario, Canada
[3]Institute of Medical Sciences, University of Toronto, Toronto, Ontario, Canada
[4]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada
[5]Banting and Best Diabetes Centre, Toronto, Ontario, Canada
[6]Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada
[7]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA.
[8]Department of Cancer Biology, University of Cincinnati College of Medicine, Cincinnati, OH, USA.
[9]Department of Environmental Health, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA.
[10]Department of Electrical Engineering and Computing Systems, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA.
[11]Department of Informatics, Nicolaus Copernicus University, Torun, Poland.
[12]Neurosciences institute, ProMedica, Toledo, Ohio, USA.

* Corresponding author:
Robert McCullumsmith, MD, PhD
Dept. of Neurosciences, University of Toledo College of Medicine and Life Sciences
3000 Arlington Ave.
Mail Stop 1007
Block Health Science Building
Room 140
Toledo, OH 43614
Phone: 205-789-0841
Email: robert.mccullumsmith@utoledo.edu

Keywords:
Kaleidoscope, application, in silico, transcriptome, exploration, database, replication studies, integration, pipeline, R Shiny

Abstract

Background

In silico data exploration is a key first step of exploring a research question. There are many publicly available databases and tools that offer appealing features to help with such a task. However, many applications lack exposure or are constrained with unfriendly or outdated user interfaces. Thus, it follows that there are many resources that are relevant to investigation of medical disorders that are underutilized.

Results

We developed an R Shiny web application, called Kaleidoscope, to address this challenge. The application offers access to several omics databases and tools to let users explore research questions *in silico*. The application is designed to be user-friendly with a unified user interface, while also scalable by offering the option of uploading user-defined datasets. We demonstrate the application features with a starting query of a single gene (Disrupted in schizophrenia 1, DISC1) to assess its protein-protein interactions network. We then explore expression levels of the gene network across tissues and cell types in the brain, as well as across 34 schizophrenia versus control differential gene expression datasets.

Conclusion

Kaleidoscope provides easy access to several databases and tools under a unified user interface to explore research questions *in silico.* The web application is open-source and freely available at https://kalganem.shinyapps.io/Kaleidoscope/. This application streamlines the process of *in silico* data exploration for users and expands the efficient use of these tools to stakeholders without specific bioinformatics expertise.

Background

Increasing numbers of large biological datasets are being deposited into publicly available repositories (1). In conjunction, an ever-increasing number of bioinformatic tools are being developed to process, analyze and view a wide spectrum of biological datasets (2). However, the rapidly growing availability of databases and bioinformatic tools can be an impediment for scientists, often hampering discovery of these tools, especially for users who are not well versed in the bioinformatics field (3, 4).

We have observed that a large number of bioinformatics tools and databases that are relevant to psychiatric disorders are still relatively untapped. Even though some of these tools are well known, some researchers avoid utilizing them mainly due the sheer scale or the complexity of the user interface which can be overwhelming to novice users.

We addressed this issue by developing an interactive R Shiny web application, called Kaleidoscope. Kaleidoscope provides a platform for easy access to these resources via a user-friendly interface. Integrating multiple databases and tools in a single platform facilitates the users interactive exploration of research questions *in silico*. This approach to data exploration can lead to exciting observations that supplement existing hypotheses, generate new ones and possibly direct future studies (**Fig 1**) (5-8). This interactive exploratory data analysis platform is particularly targeted to a broader range of investigators who are not familiar with these tools. The platform solves the issue of outdated or complex user interfaces that impede many bioinformatics tools by presenting a simple and standardized user interface across the whole platform. Our web application utilizes application programing interfaces (APIs) to

access databases to extract, harmonize, and present expression data using meaningful visualizations. An easy and fast process of examining datasets is vital to the concept of data exploration. The platform is also designed to accommodate user-provided datasets to better suit the user's research interests. This in particular is a huge asset for investigators, as the process of finding, curating, formatting, and analyzing datasets is time consuming and requires trained individuals. Being aware of these challenges, we designed our application to minimize the process of data acquisition and formatting. This intended design allows researchers to primarily focus on *in silico* hypothesis exploration without spending unneeded time on the preceding steps. Moreover, as our user group grows and users upload their own curated datasets, we expect a sizable expansion of the number of datasets hosted in our platform. This promising feature will help grow the curated datasets found in the application in a user defined manner.

Implementation

Kaleidoscope is an R Shiny web application that integrates multiple databases that are relevant to psychiatric disorders; Brain RNA-Seq, BrainCloud, BrainAtlas, BrainSpan, STRING, iLINCS, Enrichr, GeneShot, and GWAS Catalog, as well as over 200 (and counting) disease-related differential gene expression datasets.

Brain RNA-Seq

The Brain RNA-Seq provides insight on gene expression of isolated and purified cells from human and mouse cortical tissue (9, 10). Our application utilizes the Brain RNA-Seq database to incorporate a searchable database of cell-specific mRNA expression in rodent and human brain tissues. The Brain RNA-Seq database has its own web interface, but it's limited to only searching one gene at a time. Our platform

permits inquiry of multiple genes simultaneously, while also displaying figures that represent the ratio of expression across the different cell types that are present in the database. This module in our platform can serve as a first step for assessing expression levels of target genes in different brain cells to understand their cell-specific function and role, while also highlighting the differences of profiles between the two species.

Search Tool for the Retrieval of Interacting Genes (STRING)

STRING is a popular database of known and predicted protein-protein interaction (PPI) networks (11). These interactions represent either physical or functional associations between proteins. This module in our application can be used to grow a gene of interest into a network by querying for PPI networks. Through a connection to the STRING API, the user has the option to specify the stringency of the predicated associations (by choosing the appropriate score cutoff), choosing the desired number of connected proteins, and the desired organism. Our application efficiently communicates the complex results from STRING by displaying the PPI network together with a figure legend, providing context for the different edges in the network. Moreover, a table is displayed that lists all of the proteins in the network with a brief description to each corresponding protein, as well as the values of each scoring method that STRING uses to calculate the combined scores (12).

BrainCloud

The BrainCloud database was developed through a collaboration between the Lieber Institute and The National Institute of Mental Health (NIMH) to give a global insight of the role of the human transcriptome in cortical development and aging (13). Using the data generated by BrainCloud, we can observe the patterns of expression in

the brain of our target genes across the lifespan of healthy humans. This data is used to explore the genetic control of transcription of our targets during development and aging using samples from the prefrontal cortex.

Allen Brain Map

We utilized the cell type database of the Allen Brain Map which examines the transcriptional profile of thousands of single cells with RNA-Seq (14). Currently our platform has access to RNA-Seq data, generated from intact nuclei derived from frozen human brain specimens, to survey cell type diversity in the human middle temporal gyrus (MTG). In total, 15,928 nuclei from 8 human tissue donors ranging in age from 24-66 years were analyzed (14). Analysis of these transcriptional profiles reveals approximately 75 transcriptionally distinct cell types, subdivided into 45 inhibitory neuron types, 24 excitatory neuron types, and 6 non-neuronal types (14). A list of target genes can be queried across all of these clusters of cell subtypes to examine their expression levels or enrichments. A heatmap with unsupervised hierarchical clustering is displayed to visualize the differences of gene expression across the target genes and also across the different cell subtype clusters. Additionally, the user may get results about whether genes are enriched in one cell subtype cluster versus another by comparing the proportions of expression across all cell subtypes, with an adjustable difference cutoff.

In addition, the BrainSpan atlas is integrated into the platform which complements the BrainCloud database by exploring the transcriptional mechanisms involved in human brain development, but with specific transcriptional profiling of different brain regions (15). Currently, 10 brain regions are included in Kaleidoscope. Data are displayed as Reads Per Kilobase Million (RPKM) expression values.

GTEx

The Genotype-Tissue Expression (GTEx) database contains data of tissue-specific gene expression and regulation. The database is constructed based on samples that were collected from 54 non-diseased tissues from almost 1000 individuals, primarily for molecular assays including whole genome and transcriptome sequencing (16). Our application allows the user to input a list of genes to query their expression levels across all of the tissues that were analyzed in the GTEx database. The results display the median TPM (Transcripts Per Kilobase Million) of each gene in the list presented as heatmap, with options of unsupervised hierarchical clustering or log transformation. This module is very helpful in capturing the tissue-specific regulation of expression of the user's genes of interest.

Moreover, the GTEx database contains eQTL (expression quantitative trait loci) mapping for almost all studied tissues to identify variant-gene expression associations. Our platform has access to eQTL mapping of tissues relative to our field, including samples from 5 different brain regions.

Lookup Replication Studies

Previously published, peer-reviewed, and publicly available transcriptomics and proteomics datasets were carefully curated and selected to probe for patterns of differential gene expression between healthy subjects and diseased/perturbed samples. These datasets were analyzed using well-established differential expression analysis R packages. Specifically, we used Limma for microarray datasets, and EdgeR/DESeq2 for RNASeq datasets (17-19).

The curated datasets cover several substrates, including stem cells, postmortem brain, and animal models. At present the application has over 200 datasets grouped under different modules. The modules currently available are schizophrenia, depression, antipsychotics, dopamine signaling, insulin signaling, bipolar disorder, Alzheimer's disease, aging, microcystin, and coronavirus. These modules are also being expanded as we integrate additional datasets. In addition, users have the ability to upload their own curated datasets into the application, thus expanding the different diseases hosted in the platform. Loaded datasets are automatically harmonized by calculating the empirical cumulative probabilities of the log2 fold change values of genes within each dataset (20).

Kaleidoscope displays the results for this section of the software across multiple tabs. The "Results" tab shows a table where each row represent a gene from the input list of genes and each column represent a dataset from the selected datasets. The values in the table are the log2 fold-change values and p-values. The "Lookup Graph" tab shows a figure to visualize the results from the table from the "Results" tab with additional information to help the user quickly navigate the results. The x-axis of the figure represents genes and the y-axis represents proportion values of number of "hits," log2 fold change values that passed a cutoff threshold which can be adjusted by the user. These values represent the number of hits over the number of datasets for each target gene, excluding datasets that have missing values. The colors represent the number of datasets which are found for each gene. The "Heatmap" tab displays an interactive heatmap representing log2 fold change values (Log2FC), fold change values (FC), and standardized scores based on the calculations of empirical cumulative

probabilities (ECDF). These heatmaps provide unsupervised hierarchical clustering for both genes and datasets. The clustering allows patterns of similar changes of expression across the list of genes and the datasets that were selected to be identified. The “Correlation” tab in this section shows a visualization of the concordance scores matrix calculated using either Pearson or Spearman correlation analysis based on the log2 fold change values. The user has the option to calculate the concordance scores based on the input list of genes or the full list of genes in the datasets. For each correlation analysis between two datasets, the test is applied using only the genes that were found in both datasets. The colors on the figure denotes the direction of the concordance scores, where red represents negative correlation (high discordance) and blue represents positive correlation (high concordance). The “References” tab shows a table with brief descriptions of each dataset and their references.

Alternatively, instead of inputting a list of genes, the user has the option to query the commonly differentially expressed genes across multiple datasets by sorting the list of genes based on their absolute log2 fold change values within each dataset and extracting the top list of genes. An additional tab is shown when the user selects this type of analysis; it displays a heatmap representing the overlapping hits across the queried datasets. The user can adjust the parameters of this analysis by selecting the desired number of extracted genes from the datasets and the final number of genes that have the most overlap across the selected datasets.

This section of the application helps the user to “lookup” expression patterns from publicly available and curated datasets to find any interesting patterns of gene expression changes. Seeing "hits" or above average gene expression differences

across multiple datasets is a strong indication that a gene or a panel of genes is involved in the disease process. Additionally, it can be used to complement a user's own transcriptome study that is related to any of the existing diseased/perturbed modules to conduct quick and easy lookup, replication, or confirmation studies (5).

The Library of Integrated Network-Based Cellular Signatures (LINCS)

The LINCS database is a large multi-omics profiling database. For transcriptional datasets, it utilizes the L1000. The L1000 is a gene-expression profiling assay based on the direct measurement of a reduced representation of the transcriptome (978 "landmark" genes) under different perturbations; gene knockdown, gene overexpression, and drug treatments (21). Kaleidoscope provides the option to generate L1000 signatures by extracting the 978 genes and averaging the log2 fold change values across the selected datasets. The user is then able to download the L1000 signature as a tab delimited file, and upload it to the integrative LINCS (iLINCS) portal to perform perturbagen connectivity analysis. iLINCS is a web platform developed to explore and analyze LINCS signatures and is mainly used to perform *in-silico* drug discovery analysis (22).

In our application, the user can also query gene knockdown signatures by inputting a list of genes and will be connected to iLINCS API. The results are displayed as a table with the total number of knockdown signatures found per gene. Also, another table is displayed with more information on these gene knockdown signatures by specifying the gene knockdown signatures ID, cell line, and a direct link to the signature web page on the iLINCS web portal.

Genome-Wide Association Study (GWAS) Catalog

The GWAS Catalog is supported by the National Human Genome Research Institute (NHGRI). The GWAS Catalog database is the largest curated collection of all published genome-wide association studies (23). In our application, the user can query the GWAS Catalog database by inputting a gene or a list of genes. The results are displayed as a table with all significant single-nucleotide polymorphisms (SNPs) mapped to the input genes. The table will show the gene name, SNP ID (rs ID), chromosome position, studied disease or phenotype, and a direct link to the study. In addition, the type of SNP will be shown as intergenic variant, 3 Prime UTR variant, regulatory region variant, etc. Two additional tabs display distinct figures: the first displaying the top overlapping disease/phenotype traits, the second an interactive sankey graph to represent the flow rate between the list of genes, SNP type, disease/phenotype trait.

Enrichr and GeneShot

Enrichr and GeneShot are two tools that were developed by The Ma'ayan Laboratory, Icahn School of Medicine at Mount Sinai (24, 25). Enrichr is a tool for performing gene set enrichment analysis across many databases (24). GeneShot is a search engine for search terms and genes mentions based on arbitrary text queries of published literature (25). Kaleidoscope utilizes these tools' APIs to integrate them across Kaleidoscope for easy access to gene set enrichment analyses.

Discussion

To demonstrate the application's features, we ran an example session starting with a single gene of interest. The purpose of this demonstration is to highlight some of the rich information and data that can be extracted using Kaleidoscope. We picked the

Disrupted in schizophrenia 1 (DISC1) gene as the starting gene of interest. DISC1 has emerged as a strong candidate gene underlying the risk for major mental disorders (26, 27). We expanded our single gene into a network using the STRING tab, by selecting human as the desired species and using 500 as the score cutoff as well as limiting the desired number of nodes to 25 (**Fig 2A**). A table is generated that displays a brief description of each gene, and association scores under the different criteria (experimental database, co-expression, text mining etc.) (**Supplementary Table 1**). Using this list of genes from the PPI network, we data mined the other available databases in Kaleidoscope. We used the GeneShot tab to search the relevance of schizophrenia and our PPI network across previous publications. In **Fig 2B** the total publications that mention each gene and the proportions of these publications that also mention schizophrenia is shown. As expected, DISC1 has a high proportion of publications that were linked to schizophrenia. Interestingly, some genes in the network appeared to be understudied with regards to their association to schizophrenia (Serine racemase (SSR), Oligodendrocyte transcription factor 2 (OLIG2), Glycogen synthase kinase-3 beta (GSK3B)).

Next, the Brain RNA-Seq tab was used to explore DISC1 cell type specific gene expression. Fragments Per Kilobase of transcript per Million mapped reads (FPKM) for DISC1 in both human and mouse are shown for multiple cell types (neurons, astrocytes, endothelial cells, oligodendrocyte progenitor cells, etc.) (**Fig S1**). Additionally, we queried the Brain RNA-Seq tab to look at cell-type specific gene expression, using the whole network and selecting the option of multiple targets input. The gene expression figures for each gene are displayed. Boxplot and proportion figures are also displayed

when using the multiple target option. We observed higher levels of expression for this network in neurons and oligodendrocytes (**Fig 3A-B**). Oligodendrocytes have been linked to schizophrenia in regards to myelin dysfunction and neurocircuitry abnormalities (28).

We continue to explore the expression levels of our target genes in different cell types, this time using the BrainAtlas tab. Inputting our list of genes in that tab yields a table and a gene expression heatmap. The table displays the range of count per million (CPM) values for each gene across all of the different cell types, using all the sub-clusters within each cell type. A heatmap is also generated with unsupervised hierarchical clustering for both genes and cell type sub clusters (**Fig 3C**). GTEx was also queried to observe tissue specific gene expression levels. Samples from brain tissues clustered together, suggesting a similar pattern of expression of gene networks in the brain compared to other tissue types. Also, an enrichment of gene expression in brain tissues has been observed for many of the genes in our network, including OLIG2, NRXN1 and DLG4 (**Fig 3D**). We then used the lookup tab to search relative gene expression changes of the network of genes across different schizophrenia transcriptional datasets (29-42). Several figures and tables are displayed, most importantly a heatmap of log2 fold change values and table of gene expression changes across the different datasets (**Fig 4A, Supplementary Table 2**). The results indicate a clear dysregulation in expression for our network of genes, consistent with previous findings of their association with schizophrenia (43-45). Also, a correlation matrix figure is shown based on the list of the genes from the DISC1 network and paired based on the selected datasets (**Fig 4B**). Finally, the iLINCS tab was utilized to search for

knockdown signatures of the list of genes from the network and generate a table of signature IDs, cell line, and direct link to the signature iLINCS webpage for further exploration (**Supplementary Table 3**).

Conclusion

As demonstrated here, Kaleidoscope provides an integrated platform to perform *in silico* data exploration of transcriptional signatures. The ability to efficiently explore DISC1 related signatures across all of these databases and tools offers insights on its protein-protein interactors and their regulation patterns across schizophrenia transcriptional datasets. Kaleidoscope has been recently utilized to perform *in silico* replication analyses of publicly available datasets, highlighting its utility (5-8). Our platform was used to supplement findings of bioenergetic gene expression dysregulation, and to explore adenosine system dysregulation in schizophrenia (5, 6). It was also used to explore abnormal regulators of protein prenylation in schizophrenia (7). Most recently, Kaleidoscope was utilized to investigate alteration of glutamate transporter interacting proteins in schizophrenia, major depression, and amyotrophic lateral sclerosis (8). Taken together, these examples show the advantage of having a platform that streamlines the process of *in silico* data exploration, making bioinformatics tools accessible to a wider range of users to test and investigate scientific questions and findings *in silico*.

Availability and requirements

Project name: Kaleidoscope
Webpage: https://kalganem.shinyapps.io/Kaleidoscope/
Project home page: https://github.com/kalganem/kaleidoscope
Operating system(s): Platform independent
Programming language: R
Other requirements: e.g. Dependent R packages
License: GNU GPL.
Any restrictions to use by non-academics: none

List of abbreviations

API: Application programing interface
PPI: Protein-protein interaction
STRING: Search Tool for the Retrieval of Interacting Genes
MTG: middle temporal gyrus
RPKM: Reads Per Kilobase Million
TPM: Transcripts Per Kilobase Million
eQTL: Expression quantitative trait loci
SNP: Single nucleotide polymorphism
LINCS: The Library of Integrated Network-Based Cellular Signatures
iLINCS: Integrative LINCS
GWAS: genome-wide association study
UTR: untranslated region
FPKM: Fragments Per Kilobase of transcript per Million mapped reads
CPM: count per million

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the GitHub repository, https://github.com/kalganem/kaleidoscope/tree/master/data
The application is available at https://kalganem.shinyapps.io/Kaleidoscope/

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by NIMH R01 MH107487 and MH121102

Authors' contributions

KA developed and designed the software. KA and RM wrote the manuscript. RS, SO, AF, MH, JM and RM provided feedback on the application. KA, HE, MA, XZ, WM, JL, CA, RA, and RP curated the datasets. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

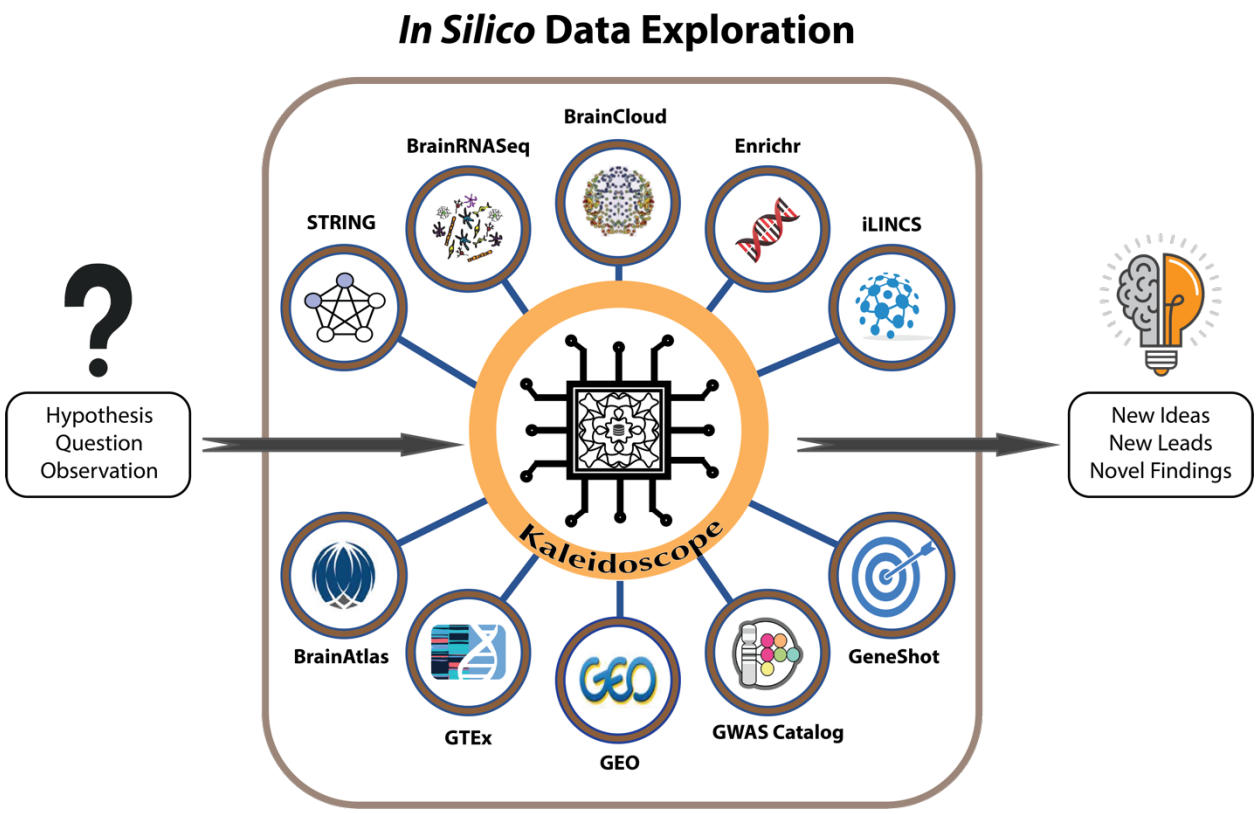

Figure 1. A workflow model for using Kaleidoscope to perform *in silico* exploratory data analyses. Starting with hypotheses and questions, Kaleidoscope integrates multiple platforms and tools to streamline the process of *in silico* data exploration. STRING (Search Tool for the Retrieval of Interacting Genes) provides protein-protein interaction networks. The Brain RNASeq database is used for cell specific gene expression levels in human and mouse brain tissues. BrainCloud is used to extract gene expression patterns in the brain across the lifespan of healthy humans. Enrichr is integrated to perform gene set enrichment analyses. iLINCS (Integrative Library of Integrated Network-Based Cellular Signatures) is queried to search for gene knockdown transcriptional signatures and generate L1000 signatures that can be further explored using the iLINCS platform. BrainAtlas is used to assess gene expression levels in different brain cell types. GTEx (Genotype-Tissue Expression) is used to explore tissue specific gene expression patterns. GEO (Gene Expression Omnibus) is utilized to extract and curate previously published differential gene expression datasets. GWAS Catalog (Genome-Wide Association Study Catalog) is searched for previously published genome-wide association studies. The GeneShot tool is used to text mine published studies to associate search terms and genes.

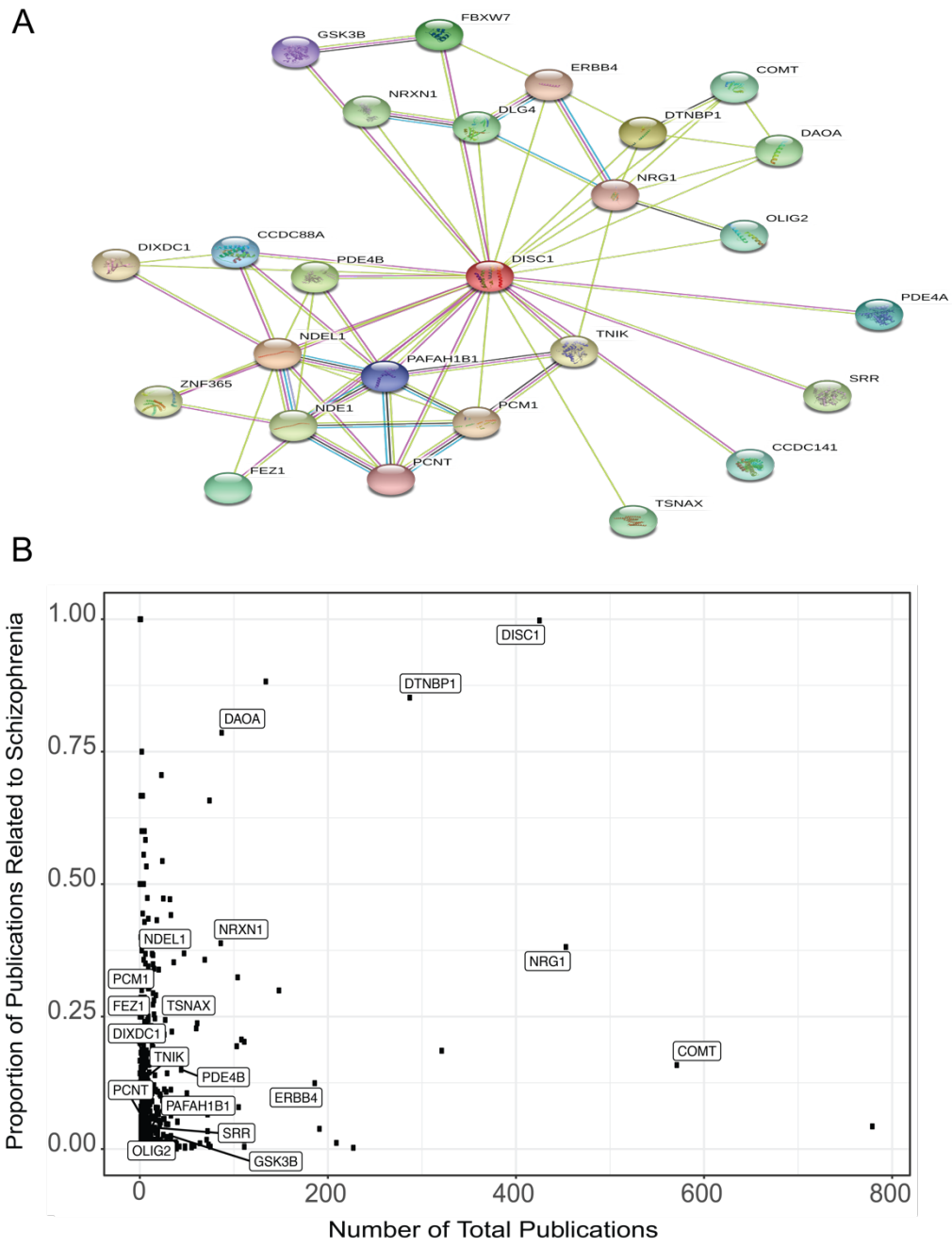

Figure 2. Protein-protein interaction (PPI) network for the DISC1 protein and mentions in the schizophrenia literature. A) The PPI network generated from the STRING (Search Tool for the Retrieval of Interacting Genes) database for DISC1 and its 25 close interactors (0.500 was used as the minimum required interaction score). B) A scatterplot to represent the associations of the DISC1 gene set from the PPI network with schizophrenia (number of papers that mention both the gene name and schizophrenia over the number of total number of papers that mention that gene).

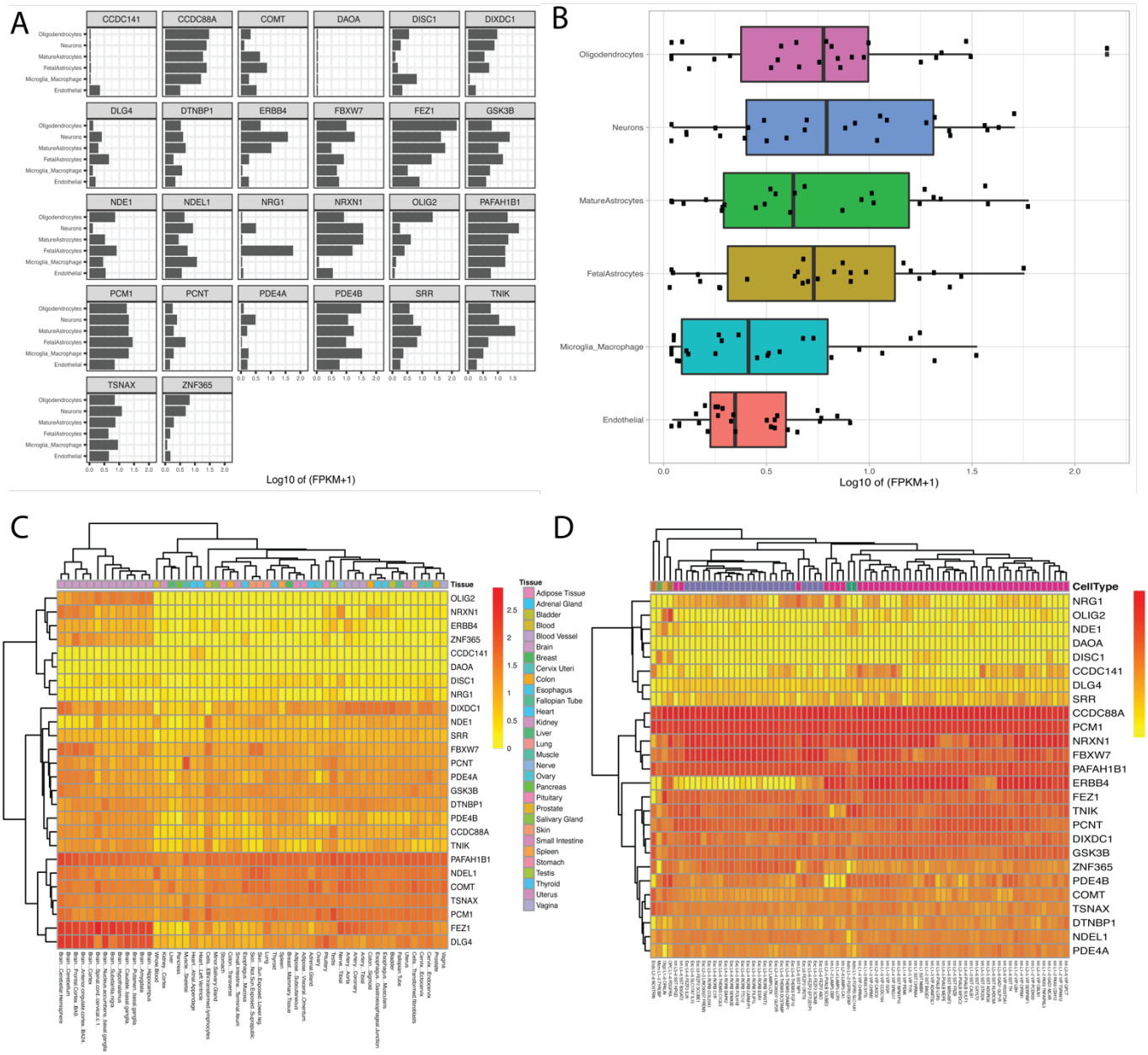

Figure 3. Plots and heatmaps to represent transcription enrichments across different cell types and tissues of the DISC1 protein interactors. A) Results from the Brain RNA-Seq. Each facet represents a gene, and the values show expression levels (log2(FPKM+1)) across the 5 different cell types in human brain tissue. B) Boxplots to show the distribution of gene expression levels for the full list of genes. C) A heatmap to highlight the difference of gene expression levels in different tissues from the GTEx database. D) Heatmap to show the gene expression levels in the different cell types from the BrainAtlas with unsupervised hierarchical clustering

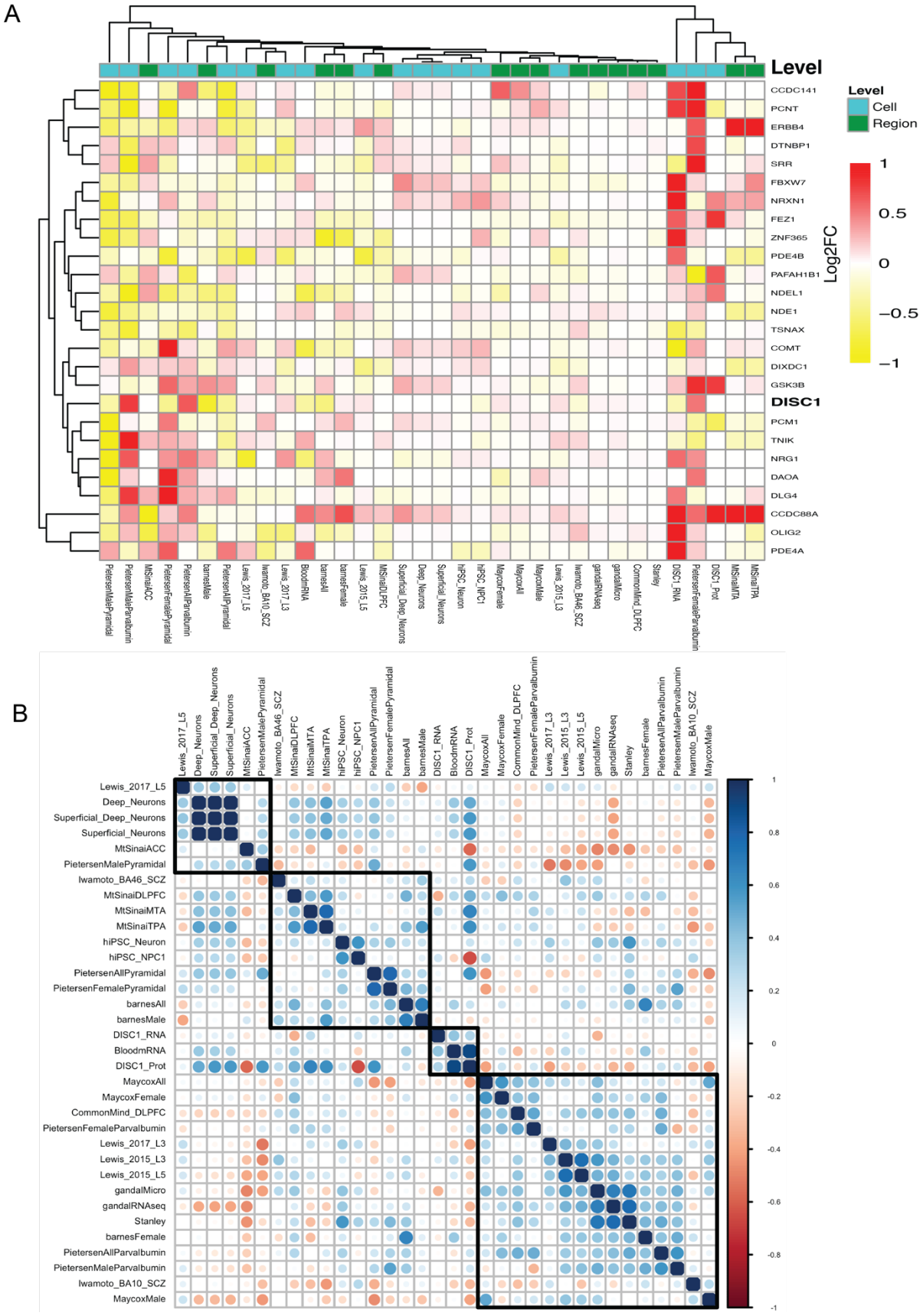

Figure 4. Differential gene expression of DISC1 protein interactors across 34 schizophrenia datasets. A) Heatmap with log2 fold change values between cases and controls across curated schizophrenia datasets. The datasets are grouped by their sample level (Region; samples taken from tissues, Cell; samples taken from pools of isolated cells). B) Correlation analysis of log2 fold change values of the DISC1 protein interactors between the schizophrenia datasets (Spearman correlation). Blue represents high concordance and red represents high discordance.

Reference:

1. Clough E, Barrett T. The Gene Expression Omnibus Database. Methods Mol Biol. 2016;1418:93-110.
2. Duck G, Nenadic G, Filannino M, Brass A, Robertson DL, Stevens R. A Survey of Bioinformatics Database and Software Usage through Mining the Literature. PloS one. 2016;11(6):e0157989.
3. Cannata N, Merelli E, Altman RB. Time to organize the bioinformatics resourceome. PLoS Comput Biol. 2005;1(7):e76.
4. Wren JD, Bateman A. Databases, data tombs and dust in the wind. Bioinformatics. 2008;24(19):2127-8.
5. Sullivan CR, Mielnik CA, O'Donovan SM, Funk AJ, Bentea E, DePasquale EA, et al. Connectivity Analyses of Bioenergetic Changes in Schizophrenia: Identification of Novel Treatments. Mol Neurobiol. 2019;56(6):4492-517.
6. Moody CL, Funk AJ, Devine E, Devore Homan RC, Boison D, McCullumsmith RE, et al. Adenosine Kinase Expression in the Frontal Cortex in Schizophrenia. Schizophrenia bulletin. 2019.
7. Pinner AL, Mueller TM, Alganem K, McCullumsmith R, Meador-Woodruff JH. Protein expression of prenyltransferase subunits in postmortem schizophrenia dorsolateral prefrontal cortex. Translational psychiatry. 2020;10(1):3.
8. Asah S, Alganem K, McCullumsmith RE, O'Donovan SM. A bioinformatic inquiry of the EAAT2 interactome in postmortem and neuropsychiatric datasets. Schizophrenia research. 2020.
9. Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, et al. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2014;34(36):11929-47.
10. Zhang Y, Sloan SA, Clarke LE, Caneda C, Plaza CA, Blumenthal PD, et al. Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. Neuron. 2016;89(1):37-53.
11. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607-D13.
12. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005;33(Database issue):D433-7.
13. Colantuoni C, Lipska BK, Ye T, Hyde TM, Tao R, Leek JT, et al. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature. 2011;478(7370):519-23.
14. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012;489(7416):391-9.
15. Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, et al. Transcriptional landscape of the prenatal human brain. Nature. 2014;508(7495):199-206.

16. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.
17. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
18. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.
19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
20. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford). 2016;2016.
21. Keenan AB, Jenkins SL, Jagodnik KM, Koplev S, He E, Torre D, et al. The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. Cell Syst. 2018;6(1):13-24.
22. Pilarczyk M, Najafabadi MF, Kouril M, Vasiliauskas J, Niu W, Shamsaei B, et al. Connecting omics signatures of diseases, drugs, and mechanisms of actions with iLINCS. bioRxiv. 2019:826271.
23. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005-D12.
24. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14:128.
25. Lachmann A, Schilder BM, Wojciechowicz ML, Torre D, Kuleshov MV, Keenan AB, et al. Geneshot: search engine for ranking genes from arbitrary text queries. Nucleic Acids Res. 2019;47(W1):W571-W7.
26. Blackwood DH, Fordyce A, Walker MT, St Clair DM, Porteous DJ, Muir WJ. Schizophrenia and affective disorders--cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family. Am J Hum Genet. 2001;69(2):428-33.
27. Bentea E, Depasquale EAK, O'Donovan SM, Sullivan CR, Simmons M, Meador-Woodruff JH, et al. Kinase network dysregulation in a human induced pluripotent stem cell model of DISC1 schizophrenia. Mol Omics. 2019;15(3):173-88.
28. Takahashi N, Sakurai T, Davis KL, Buxbaum JD. Linking oligodendrocyte and myelin dysfunction to neurocircuitry abnormalities in schizophrenia. Prog Neurobiol. 2011;93(1):13-24.
29. Xu Y, Yao Shugart Y, Wang G, Cheng Z, Jin C, Zhang K, et al. Altered expression of mRNA profiles in blood of early-onset schizophrenia. Sci Rep. 2016;6:16767.
30. Wen Z, Nguyen HN, Guo Z, Lalli MA, Wang X, Su Y, et al. Synaptic dysregulation in a human iPS cell model of mental disorders. Nature. 2014;515(7527):414-8.
31. Gandal MJ, Haney JR, Parikshak NN, Leppa V, Ramaswami G, Hartl C, et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. Science (New York, NY). 2018;359(6376):693-7.

32. Hoffman GE, Hartley BJ, Flaherty E, Ladran I, Gochman P, Ruderfer DM, et al. Transcriptional signatures of schizophrenia in hiPSC-derived NPCs and neurons are concordant with post-mortem adult brains. Nat Commun. 2017;8(1):2225.
33. Roussos P, Katsel P, Davis KL, Siever LJ, Haroutunian V. A system-level transcriptomic analysis of schizophrenia using postmortem brain tissue samples. Archives of general psychiatry. 2012;69(12):1205-13.
34. Torrey EF, Webster M, Knable M, Johnston N, Yolken RH. The stanley foundation brain collection and neuropathology consortium. Schizophrenia research. 2000;44(2):151-5.
35. Arion D, Corradi JP, Tang S, Datta D, Boothe F, He A, et al. Distinctive transcriptome alterations of prefrontal pyramidal neurons in schizophrenia and schizoaffective disorder. Mol Psychiatry. 2015;20(11):1397-405.
36. Pietersen CY, Mauney SA, Kim SS, Lim MP, Rooney RJ, Goldstein JM, et al. Molecular profiles of pyramidal neurons in the superior temporal cortex in schizophrenia. J Neurogenet. 2014;28(1-2):53-69.
37. Pietersen CY, Mauney SA, Kim SS, Passeri E, Lim MP, Rooney RJ, et al. Molecular profiles of parvalbumin-immunoreactive neurons in the superior temporal cortex in schizophrenia. J Neurogenet. 2014;28(1-2):70-85.
38. Barnes MR, Huxley-Jones J, Maycox PR, Lennon M, Thornber A, Kelly F, et al. Transcription and pathway analysis of the superior temporal cortex and anterior prefrontal cortex in schizophrenia. Journal of neuroscience research. 2011;89(8):1218-27.
39. Maycox PR, Kelly F, Taylor A, Bates S, Reid J, Logendra R, et al. Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. Mol Psychiatry. 2009;14(12):1083-94.
40. Iwamoto K, Bundo M, Kato T. Altered expression of mitochondria-related genes in postmortem brains of patients with bipolar disorder or schizophrenia, as revealed by large-scale DNA microarray analysis. Human molecular genetics. 2005;14(2):241-53.
41. Iwamoto K, Kakiuchi C, Bundo M, Ikeda K, Kato T. Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. Mol Psychiatry. 2004;9(4):406-16.
42. Wu X, Shukla R, Alganem K, Depasquale E, Reigle J, Simmons M, et al. Transcriptional profile of pyramidal neurons in chronic schizophrenia reveals lamina-specific dysfunction of neuronal immunity. bioRxiv. 2020:2020.01.14.906214.
43. Labrie V, Fukumura R, Rastogi A, Fick LJ, Wang W, Boutros PC, et al. Serine racemase is associated with schizophrenia susceptibility in humans and in a mouse model. Human molecular genetics. 2009;18(17):3227-43.
44. Harrison PJ, Law AJ. Neuregulin 1 and schizophrenia: genetics, gene expression, and neurobiology. Biological psychiatry. 2006;60(2):132-40.
45. Funk AJ, Mielnik CA, Koene R, Newburn E, Ramsey AJ, Lipska BK, et al. Postsynaptic Density-95 Isoform Abnormalities in Schizophrenia. Schizophrenia bulletin. 2017;43(4):891-9.

Protein	Combined Score	Description	Experimental	Database	Gene Fusion	Textmining	CoExpression	Phylogenetic	Neighborhood
NDEL1	0.93	Nuclear distribution protein nudE-like 1; Required for organization of the cellular microtubule array and microtubule anchoring at the centrosome. May regulate microtubule organization at least in part by targeting the microtubule severing protein KATNA1 to the centrosome. Also positively regulates the activity of the minus-end directed microtubule motor protein dynein. May enhance dynein-mediated microtubule sliding by targeting dynein to the microtubule plus ends. Required for several dynein- and microtubule-dependent processes such as the maintenance of Golgi integrity, the centripe [...]	0.38	0	0	0.9	0	0	0
DTNBP1	0.93	Dysbindin; Component of the BLOC-1 complex, a complex that is required for normal biogenesis of lysosome-related organelles (LRO), such as platelet dense granules and melanosomes. In concert with the AP-3 complex, the BLOC-1 complex is required to target membrane protein cargos into vesicles assembled at cell bodies for delivery into neurites and nerve terminals. The BLOC-1 complex, in association with SNARE proteins, is also proposed to be involved in neurite extension. Associates with the BLOC-2 complex to facilitate the transport of TYRP1 independent of AP-3 function. Plays a role i [...]	0	0	0	0.93	0	0	0
PDE4B	0.9	cAMP-specific 3',5'-cyclic phosphodiesterase 4B; Hydrolyzes the second messenger cAMP, which is a key regulator of many important physiological processes. May be involved in mediating central nervous system effects of therapeutic agents ranging from antidepressants to antiasthmatic and anti-inflammatory agents; Phosphodiesterases	0.23	0	0	0.87	0	0	0
FBXW7	0.89	F-box/WD repeat-containing protein 7; Substrate recognition component of a SCF (SKP1-CUL1-F- box protein) E3 ubiquitin-protein ligase complex which mediates the ubiquitination and subsequent proteasomal degradation of target proteins. Recognizes and binds phosphorylated sites/phosphodegrons within target proteins and thereafter bring them to the SCF complex for ubiquitination. Identified substrates include cyclin-E (CCNE1 or CCNE2), JUN, MYC, NOTCH1 released notch intracellular domain (NICD), and probably PSEN1. Acts as a negative regulator of JNK signaling by binding to phosphorylated [...]	0.89	0	0	0.09	0	0	0
FEZ1	0.88	Fasciculation and elongation protein zeta-1; May be involved in axonal outgrowth as component of the network of molecules that regulate cellular morphology and axon guidance machinery. Able to restore partial locomotion and axonal fasciculation to C.elegans unc-76 mutants in germline transformation experiments. May participate in the transport of mitochondria and other cargos along microtubules; Belongs to the zygin family	0.38	0	0	0.81	0	0	0
PDE4A	0.82	cAMP-specific 3',5'-cyclic phosphodiesterase 4A; Hydrolyzes the second messenger cAMP, which is a key regulator of many important physiological processes; Belongs to the cyclic nucleotide phosphodiesterase family. PDE4 subfamily	0.26	0	0	0.77	0	0	0
CCDC88A	0.81	Girdin; Plays a role as a key modulator of the AKT-mTOR signaling pathway controlling the tempo of the process of newborn neurons integration during adult neurogenesis, including correct neuron positioning, dendritic development and synapse formation (By similarity). Enhances phosphoinositide 3-kinase (PI3K)- dependent phosphorylation and kinase activity of AKT1/PKB, but does not possess kinase activity itself (By similarity). Phosphorylation of AKT1/PKB thereby induces the phosphorylation of downstream effectors GSK3 and FOXO1/FKHR, and regulates DNA replication and cell proliferation [...]	0.12	0	0	0.8	0	0	0
PAFAH1B1	0.81	Platelet-activating factor acetylhydrolase IB subunit alpha; Required for proper activation of Rho GTPases and actin polymerization at the leading edge of locomoting cerebellar neurons and postmigratory hippocampal neurons in response to calcium influx triggered via NMDA receptors. Non-catalytic subunit of an acetylhydrolase complex which inactivates platelet- activating factor (PAF) by removing the acetyl group at the SN-2 position (By similarity). Positively regulates the activity of the minus-end directed microtubule motor protein dynein. May enhance dynein-mediated microtubule slid [...]	0.12	0	0	0.8	0	0	0
GSK3B	0.8	Glycogen synthase kinase-3 beta; Constitutively active protein kinase that acts as a negative regulator in the hormonal control of glucose homeostasis, Wnt signaling and regulation of transcription factors and microtubules, by phosphorylating and inactivating glycogen synthase (GYS1 or GYS2), EIF2B, CTNNB1/beta-catenin, APC, AXIN1, DPYSL2/CRMP2, JUN, NFATC1/NFATC, MAPT/TAU and MACF1. Requires primed phosphorylation of the majority of its substrates. In skeletal muscle, contributes to insulin regulation of glycogen synthesis by phosphorylating and inhibiting GYS1 activity and hence glyc [...]	0.47	0	0	0.64	0	0	0
PCNT	0.8	Pericentrin; Integral component of the filamentous matrix of the centrosome involved in the initial establishment of organized microtubule arrays in both mitosis and meiosis. Plays a role, together with DISC1, in the microtubule network formation. Is an integral component of the pericentriolar material (PCM). May play an important role in preventing premature centrosome splitting during interphase by inhibiting NEK2 kinase activity at the centrosome	0.38	0	0	0.69	0	0	0
NRG1	0.79	Pro-neuregulin-1, membrane-bound isoform; Direct ligand for ERBB3 and ERBB4 tyrosine kinase receptors. Concomitantly recruits ERBB1 and ERBB2 coreceptors, resulting in ligand-stimulated tyrosine phosphorylation and activation of the ERBB receptors. The multiple isoforms perform diverse functions such as inducing growth and differentiation of epithelial, glial, neuronal, and skeletal muscle cells; inducing expression of acetylcholine receptor in synaptic vesicles during the formation of the neuromuscular junction; stimulating lobuloalveolar budding and milk production in the mammary gla [...]	0	0	0	0.79	0	0	0
ERBB4	0.78	Receptor tyrosine-protein kinase erbB-4; Tyrosine-protein kinase that plays an essential role as cell surface receptor for neuregulins and EGF family members and regulates development of the heart, the central nervous system and the mammary gland, gene transcription, cell proliferation, differentiation, migration and apoptosis. Required for normal cardiac muscle differentiation during embryonic development, and for postnatal cardiomyocyte proliferation. Required for normal development of the embryonic central nervous system, especially for normal neural crest cell migration and normal [...]	0	0	0	0.78	0	0	0
PCM1	0.77	Pericentriolar material 1 protein; Required for centrosome assembly and function. Essential for the correct localization of several centrosomal proteins including CEP250, CETN3, PCNT and NEK2. Required to anchor microtubules to the centrosome. Involved in the biogenesis of cilia; Belongs to the PCM1 family	0	0	0	0.77	0	0	0
DIXDC1	0.75	Dixin; Positive effector of the Wnt signaling pathway; activates WNT3A signaling via DVL2. Regulates JNK activation by AXIN1 and DVL2; Belongs to the DIXDC1 family	0	0	0	0.75	0	0	0
TNIK	0.75	TRAF2 and NCK-interacting protein kinase; Serine/threonine kinase that acts as an essential activator of the Wnt signaling pathway. Recruited to promoters of Wnt target genes and required to activate their expression. May act by phosphorylating TCF4/TCF7L2. Appears to act upstream of the JUN N-terminal pathway. May play a role in the response to environmental stress. Part of a signaling complex composed of NEDD4, RAP2A and TNIK which regulates neuronal dendrite extension and arborization during development. More generally, it may play a role in cytoskeletal rearrangements and regulate [...]	0.12	0	0	0.73	0	0	0
ZNF365	0.75	Protein ZNF365; Involved in the regulation of neurogenesis. Negatively regulates neurite outgrowth. Involved in the morphogenesis of basket cells in the somatosensory cortex during embryogenesis. Involved in the positive regulation of oligodendrocyte differentiation during postnatal growth. Involved in dendritic arborization, morphogenesis of spine density dendrite, and establishment of postsynaptic dendrite density in cortical pyramidal neurons (By similarity). Involved in homologous recombination (HR) repair pathway. Required for proper resolution of DNA double-strand breaks (DSBs) b [...]	0.12	0	0	0.73	0	0	0
NDE1	0.73	Nuclear distribution protein nudE homolog 1; Required for centrosome duplication and formation and function of the mitotic spindle. Essential for the development of the cerebral cortex. May regulate the production of neurons by controlling the orientation of the mitotic spindle during division of cortical neuronal progenitors of the proliferative ventricular zone of the brain. Orientation of the division plane perpendicular to the layers of the cortex gives rise to two proliferative neuronal progenitors whereas parallel orientation of the division plane yields one proliferative neurona [...]	0.18	0	0	0.69	0	0	0
NRXN1	0.73	Neurexin-1; Cell surface protein involved in cell-cell-interactions, exocytosis of secretory granules and regulation of signal transmission. Function is isoform-specific. Alpha-type isoforms have a long N-terminus with six laminin G-like domains and play an important role in synaptic signal transmission. Alpha-type isoforms play a role in the regulation of calcium channel activity and Ca(2+)-triggered neurotransmitter release at synapses and at neuromuscular junctions. They play an important role in Ca(2+)- triggered exocytosis of secretory granules in pituitary gland. They may effect [...]	0	0	0	0.73	0	0	0
SRR	0.73	Serine racemase; Catalyzes the synthesis of D-serine from L-serine. D- serine is a key coagonist with glutamate at NMDA receptors. Has dehydratase activity towards both L-serine and D-serine	0.46	0	0	0.52	0	0	0
DAOA	0.73	D-amino acid oxidase activator; Seems to activate D-amino acid oxidase	0	0	0	0.73	0	0	0
DLG4	0.72	Disks large homolog 4; Interacts with the cytoplasmic tail of NMDA receptor subunits and shaker-type potassium channels. Required for synaptic plasticity associated with NMDA receptor signaling. Overexpression or depletion of DLG4 changes the ratio of excitatory to inhibitory synapses in hippocampal neurons. May reduce the amplitude of ASIC3 acid-evoked currents by retaining the channel intracellularly. May regulate the intracellular trafficking of ADR1B (By similarity); Belongs to the MAGUK family	0	0	0	0.72	0	0	0
TSNAX	0.72	Translin-associated protein X; Acts in combination with TSN as an endonuclease involved in the activation of the RNA-induced silencing complex (RISC). Possible role in spermatogenesis	0	0	0	0.72	0	0	0
COMT	0.71	Catechol O-methyltransferase; Catalyzes the O-methylation, and thereby the inactivation, of catecholamine neurotransmitters and catechol hormones. Also shortens the biological half-lives of certain neuroactive drugs, like L-DOPA, alpha-methyl DOPA and isoproterenol; Seven-beta-strand methyltransferase motif containing	0	0	0	0.71	0	0	0
OLIG2	0.71	Oligodendrocyte transcription factor 2; Required for oligodendrocyte and motor neuron specification in the spinal cord, as well as for the development of somatic motor neurons in the hindbrain. Cooperates with OLIG1 to establish the pMN domain of the embryonic neural tube. Antagonist of V2 interneuron and of NKX2-2-induced V3 interneuron development (By similarity); Basic helix-loop-helix proteins	0	0	0	0.7	0	0	0
CCDC141	0.69	Coiled-coil domain-containing protein 141; Plays a critical role in radial migration and centrosomal function; Immunoglobulin like domain containing	0.24	0	0	0.61	0	0	0
DISC1	NA	Disrupted in schizophrenia 1 protein; Involved in the regulation of multiple aspects of embryonic and adult neurogenesis. Required for neural progenitor proliferation in the ventrical/subventrical zone during embryonic brain development and in the adult dentate gyrus of the hippocampus. Participates in the Wnt-mediated neural progenitor proliferation as a positive regulator by modulating GSK3B activity and CTNNB1 abundance. Plays a role as a modulator of the AKT-mTOR signaling pathway controlling the tempo of the process of newborn neurons integration during adult neurogenesis, includi [...]	NA	NA	NA	NA			

Supplementary Table 1: STRING protein-protein association network for the DISC1 protein. The table displays a brief description of each protein, and association scores under the different criteria. The combined score is computed based on the assumption of independence of the other scores (12).

GENE	KNOCKDOWN SIGNATURES
GSK3B	15
COMT	11
ERBB4	11
FEZ1	11
PAFAH1B1	11
NRG1	10
PCM1	10
TNIK	10
ZNF365	8
DISC1	2
DIXDC1	2
NRXN1	2

Supplementary Table 3: iLINCS gene knockdown signatures table. A tabular summary displaying the number of gene knockdown signatures found in iLINCS for each gene in the protein-protein interaction (PPI) network for the DISC1 protein.