NoRCE: Non-coding RNA Sets Cis Enrichment Tool

Gulden Olgun[1], Afshan Nabi[2], Oznur Tastan[2*]
[1] Department of Computer Engineering, Bilkent University, Ankara, Turkey
[2] Faculty of Engineering and Natural Sciences, Sabanci University, Tuzla, Istanbul, Turkey
* otastan@sabanciuniv.edu

Abstract

Summary: While some non-coding RNAs (ncRNAs) have been found to play critical regulatory roles in biological processes, most remain functionally uncharacterized. This presents a challenge whenever an interesting set of ncRNAs set needs to be analyzed in a functional context. Transcripts located close-by on the genome are often regulated together, and this spatial proximity hints at a functional association. Based on this idea, we present an R package, NoRCE, that performs cis enrichment analysis for a given set of ncRNAs. Enrichment is carried out by using the functional annotations of the coding genes located proximally to the input ncRNAs. NoRCE allows incorporating other biological information such as the topologically associating domain (TAD) regions, co-expression patterns, and miRNA target information. NoRCE repository includes several data files, such as cell line specific TAD regions, functional gene sets, and cancer expression data. Additionally, users can input custom data files. Results can be retrieved in a tabular format or viewed as graphs. NoRCE is currently available for the following species: human, mouse, rat, zebrafish, fruit fly, worm and yeast.

Availability and Implementation: NoRCE R package is platform independent, available at `https://github.com/guldenolgun/NoRCE` and Bioconductor.

1 Introduction

The ncRNAs represent the largest class of transcripts in the human genome [2, 7]. They have been reported to have diverse regulatory roles in gene expression [3]. However, only a small fraction of all ncRNAs have been functionally interrogated and the functions played by the majority of ncRNAs are still unknown. This presents a challenge when an ncRNA set of interest is available and needs to functionally investigated. In this work, we introduce an R package, NoRCE, which performs functional enrichment of a set of ncRNAs by transferring functional annotation available from the coding genes nearby. Although not all functionally related genes are located nearby on the genome, previous work has shown that functionally related genes show substantial spatial concentration within chromosomes and across chromosomes [9]. There is also evidence suggesting that both ncRNAs and coding genes can influence the expression of neighboring genomics loci [5].

For each ncRNA in the input set, NoRCE first determines the set of coding genes located within a user-defined distance range (in base-pairs) of this ncRNA and use the union of annotations of all coding genes assigned to the ncRNA set to perform enrichment. NoRCE provides options to perform enrichment analysis by using Fisher's exact, hypergeometric, binomial, or chi-squared tests.

Several useful options are available for the curation of the input ncRNA set. Users can filter ncRNAs with respect to their biotypes (sense, antisense, lincRNA, etc.). The input set's neighboring coding gene set can be filtered or expanded with coding genes found to be co-expressed with the input ncRNAs. The co-expression analysis can be performed by using the data available in NoRCE repository, such as The Cancer Genome Atlas (TCGA) project expression data or by using custom expression data. Since the interruption in TAD boundaries affects the expression of neighboring genes, filtering option with TAD boundaries allows the user to limit the genes to be enriched to a set that falls into TAD regions[8]. For miRNA specific input, the neighbor coding set can be filtered or expanded with predicted targets of the input miRNAs or with co-expressed genes. Finally, the outcome of the above analysis can be summarized in a tabular format or visualized using relevant graphs, including KEGG pathway map, Reactome diagram, and directed acyclic graph (DAG) of the enriched GO-term (See Figure 1 for the

detailed workflow of the package). The analysis is available for several species including human, mouse, rat, zebrafish, fruit fly, worm and yeast (For details on the assemblies supported, see Supplementary Table S1).

2 Implementation and Features

NoRCE accepts a set of ncRNAs, $S = \{r_1, \ldots, r_n\}$. If the user wants to conduct a biotype specific analysis (e.g., lincRNAs), NoRCE can select the elements in S with the selected biotype and use this subset in the subsequent steps. For each ncRNA in the input list, NoRCE identifies the nearby protein-coding genes that are located upstream and downstream in a given user-specified base pair limit. Thus, for each $r_i \in S$, NoRCE creates a coding gene list of C_i and conducts the analysis with the union of these $C = \cup_{i=1}^{n} C_i$. This set of coding genes, C, can be further filtered or expanded based on additional biological evidence:

Co-expressed coding genes: If the filtering option is set, for every input ncRNA r_i, only the nearby genes in C_i that are co-expressed with r_i are placed into the final coding gene set C. If the expansion option is set, a coding gene is co-expressed with any of the $r_i \in S$, and each coding gene is included in C. For cancer analysis, we provide pre-computed results based on expression analysis of the patient data obtained from The Cancer Genome Atlas (TCGA) portal [1]. Users can also load custom expression data and perform correlation based expression analysis.

TAD regions: Chromosomes are organized into distinct, evolutionarily conserved, cell-type invariant secondary structures called TADs. TADs are known to have functional importance since the subsets of genes within a TAD are co-regulated and some gene clusters, such as cytochrome genes and olfactory receptors are organized into individual TADs [4]. Since the interruption in TAD boundaries affects the expression of neighboring genes [8], we allow a filtering option with the TAD regions. If this option is on, enrichment is conducted only with ncRNAs that fall into a TAD region. Also, in the coding gene assignment, only those that are in the same TAD region are included in the neighborhood coding gene set. NoRCE provides several cell-line specific TAD regions in its data repository. Alternatively, the user can input a custom TAD file.

miRNA target information: If the user list is a set of miRNAs, the neighborhood of miRNA genes can be subsetted with the potential miRNA targets. For this step, NoRCE utilizes several available target prediction algorithms (See Supplementary Table S3).

2.1 Enrichment Analysis of the Coding Genes

Once the coding gene list, C, is curated, NoRCE can conduct enrichment analysis based on gene sets derived GO-terms, KEGG, Reactome, WikiPathways, coding genes, or integrated pathway dataset. Supplementary Table S2 reports the size of gene ontologies available for each species in NoRCE. Users can input custom gene sets as well. Enrichment is performed according to the standard candidate versus background sets using Fisher's exact, hypergeometric, binomial and chi-squared tests. Several multi-test hypothesis correction methods and GO enrichment methods are also provided.

2.2 Visualization of the Results

Finally, coding genes found in enriched pathways can be visualized within the context of KEGG or Reactome maps marked with color (Supplementary Fig S4).

3 Results and Discussion

To demonstrate how NoRCE could be used to gain insight into the functional importance of a list of interesting ncRNAs, we showcase NoRCE with different scenarios (Supplementary Text). As demonstrated in Case Study 1 (Supplementary Section 2.1), a list of ncRNAs differentially expressed in three psychiatric disorders, including autism spectrum disorder (ASD), schizophrenia (SCZ), and bipolar disorder (BD) generated by Gandal et al. [6], are analyzed with NoRCE. We find that brain-related terms, as well as more general terms like transcription, transmembrane transport, and RNA processing, are enriched. We repeat this analysis with a TAD filter where the TAD regions of the dorsolateral prefrontal

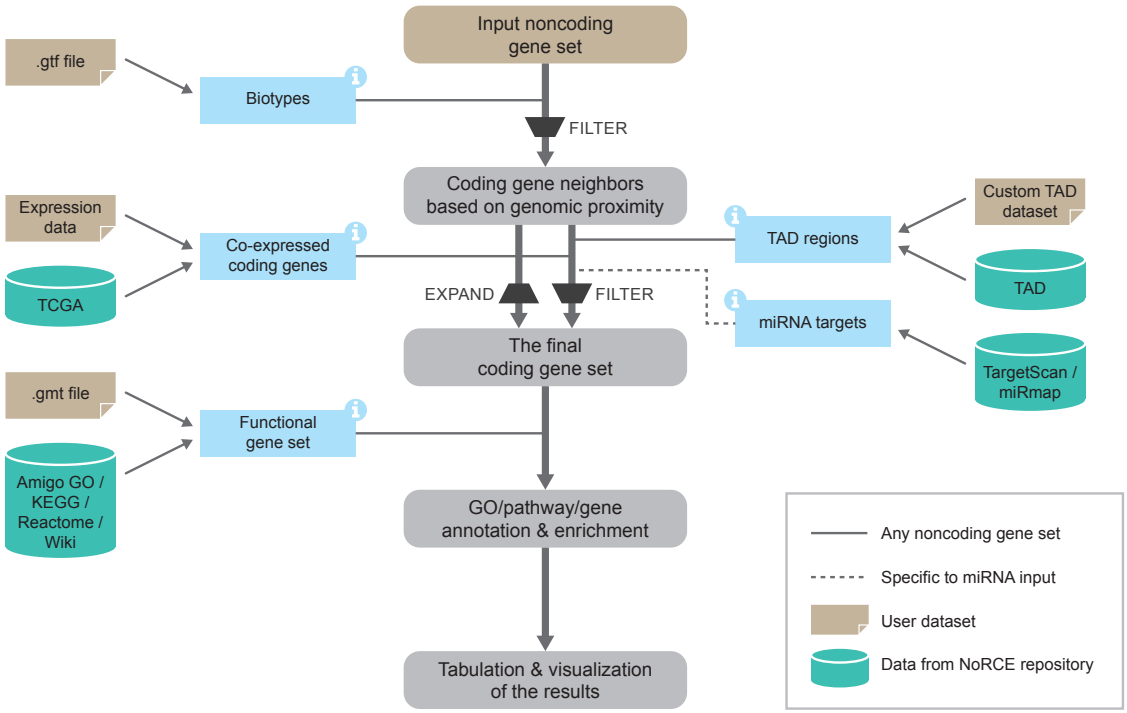

Figure 1: The NoRCE workflow

cortex are used (Supplementary Section 2.1.2). Thirdly, on this dataset, we show the ability of NoRCE to perform pathway enrichment with pathways in the data repository and also with user-defined gene sets. In each case study, we observe that brain related functions are highlighted. However, we also observe the potentially irrelevant terms such as bone development could be returned. Next, to demonstrate miRNA specific functionalities, we conduct a functional enrichment analysis on a set of differentially expressed miRNAs obtained from [10] (Supplementary Section 2.2, Case Study 2). In this study, we filter the neighboring coding genes with target and custom co-expression analyses. Also, we perform miRNA pathway enrichment based on non-coding gene neighbors.

Funding

We thank Sabanci and Bilkent Universities for internal funding support.

References

[1] Mahmoud Ahmed, Huynh Nguyen, Trang Lai, and Deok Ryong Kim. mircancerdb: a database for correlation analysis between microrna and gene expression in cancer. *BMC research notes*, 11(1):103, 2018.

[2] Pedro J Batista and Howard Y Chang. Long noncoding rnas: cellular address codes in development and disease. *Cell*, 152(6):1298–1307, 2013.

[3] Thomas R Cech and Joan A Steitz. The noncoding rna revolution-trashing old rules to forge new ones. *Cell*, 157(1):77–94, 2014.

[4] Jesse R Dixon, David U Gorkin, and Bing Ren. Chromatin domains: the unit of chromosome organization. *Molecular cell*, 62(5):668–680, 2016.

[5] Jesse M Engreitz, Jenna E Haines, Elizabeth M Perez, Glen Munson, et al. Local regulation of gene expression by lncrna promoters, transcription and splicing. *Nature*, 539(7629):452, 2016.

[6] Michael J Gandal, Pan Zhang, Evi Hadjimichael, Rebecca L Walker, et al. Transcriptome-wide isoform-level dysregulation in asd, schizophrenia, and bipolar disorder. *Science*, 362(6420):eaat8127, 2018.

[7] Matthew K Iyer, Yashar S Niknafs, Rohit Malik, Udit Singhal, et al. The landscape of long noncoding rnas in the human transcriptome. *Nature genetics*, 47(3):199, 2015.

[8] Darío G Lupiáñez, Malte Spielmann, and Stefan Mundlos. Breaking tads:how alterations of chromatin domains result in disease. *Trends in Genet.*, 32(4):225–237, 2016.

[9] Annelyse Thévenin, Liat Ein-Dor, Michal Ozery-Flato, and Ron Shamir. Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. *Nucl. Acids Res.*, 42(15):9854–9861, 2014.

[10] Zhen Yang, Liangcai Wu, Anqiang Wang, Wei Tang, et al. dbdemc 2.0: updated database of differentially expressed mirnas in human cancers. *Nucl. Acids Res.*, 45(D1):D812–D818, 2016.

Appendix

Table of Contents

A Supplementary Methods **5**

A.1 Data sources 5

B Supplementary Results **7**

B.1 Case Study 1: Functional enrichment analysis of ncRNAs differentially expressed in psychiatric disorders 7

B.2 Case Study 2: Functional enrichment analysis of variably expressed miRNAs in brain cancer 10

B.3 Data in NoRCE repository for the vignette and supplementary results 16

B.4 Other features 17

A Supplementary Methods

A.1 Data sources

NoRCE repository contains various datasets. Below, we detail these sources.

A.1.1 Gene annotations

Gene locations and annotations are retrieved from three different databases: ENSEMBL, UCSC, and GENCODE since no single source contain information on all the transcripts. We collect the ENSEMBL data via biomaRt package [6]. Genes were retrieved from UCSC on June 20th 2018, and annotations of the GENCODE v30 was downloaded from GENCODE on April 28th 2019. Supported assemblies for different species are provided in Table S1.

Table S1: The supported assemblies for different species in NoRCE.

Species	Supported Assembly	
	UCSC	NCBI
Homo sapiens	hg19	GRCh37, GCA_000001405.1, Feb. 2009
Homo sapiens	hg38	GRCh38, GCA_000001405.15, Dec. 2013
Mus musculus	mm10	GRCm38.p6, INSDC Assembly GCA_000001635.8, Jan 2012
Rattus norvegicus (brown rat)	rn6	Rnor_6.0, INSDC Assembly GCA_000001895.4, Jul 2014
Drosophila melanogaster (fruit fly)	dm6	BDGP6, INSDC Assembly GCA_000001215.4, Jul 2014
Danio rerio (zebrafish)	danRer10	GRCz10, GCA_000002035.3, Sep. 2014
Caenorhabditis elegans (worm)	ce11	WBcel235, GCA_000002985.3, Feb. 2013
Saccharomyces cerevisiae (yeast)	sacCer3	UCSC version sacCer3, 2011

A.1.2 Gene ontology annotations

We retrieve the gene ontologies for each species from the EMBL-EBI GOA database(`https://www.ebi.ac.uk/GOA/downloads`) and the GO slim annotations from the Gene Ontology (GO) Consortium [2]. The sizes of different GO ontologies for each species are provided in Table S2.

Table S2: The size of gene ontologies available for each species in NoRCE

	Biological Process	Cellular Component	Molecular Function
Homo sapiens	137,575	80,584	66,106
Mus musculus	161,638	103,261	91,633
Rattus norvegicus	205,265	107,124	87,949
Drosophila melanogaster	50,525	27,156	27,894
Danio rerio	64,866	46,429	55,383
Caenorhabditis elegans	36,699	30,189	32,716
Saccharomyces cerevisiae	35,136	33,417	26,526

A.1.3 miRNA target information

We obtain computationally predicted conserved miRNA-target interactions from the TargetScan [16] for the species except rat. Rat data was obtained from the miRmap [19]. More information is provided in Table S3.

Table S3: List of miRNA target prediction algorithms and the corresponding number of target pairs for each species. To the best of our knowledge, no miRNA target predictions have been reported for *Saccharomyces cerevisiae*.

Species	miRNA-Target Prediction Database	Version/Date	No of Pairs	Reference
Homo sapiens	TargetScan	v7.2	235,128	[16]
Mus musculus	TargetScan	v7.1	179,430	[10]
Rattus norvegicus	miRmap	2013	2,258,219	[19]
Drosophila melanogaster	TargetScan	v6.2	15,641	[15]
Danio rerio	TargetScan	v6.2	1,147,028	[10]
Caenorhabditis elegans	TargetScan	v6.2	208,146	[8]

A.1.4 TAD regions

We also include bed formatted topological associated domain (TAD) regions for different cell types in NoRCE repository. These regions are not available for all species. We list the available ones in Table S4.

Table S4: Topological associating domain data that included in the NoRCE repository

Species	Name	Version	# of cell lines	# of TAD regions	Source
Homo sapiens	tad_hg19	hg19	37	74,424	3D Genome Browser[20]
Homo sapiens	tad_hg38	hg38	42	96,526	3D Genome Browser[20]
Mus musculus	tad_mm10	mm10	5	16,866	3D Genome Browser [20]
D. melanogaster	tad_dmel	dm6	1	2,846	HiCBrowser [14]

B Supplementary Results

B.1 Case Study 1: Functional enrichment analysis of ncRNAs differentially expressed in psychiatric disorders

To demonstrate how NoRCE could be used to gain functional insights given a set of ncRNAs, we retrieved a set of ncRNAs exhibiting gene- or isoform-level differential expression data on brain disorders. These ncRNAs are differentially expressed in at least one of the following disorders: autism spectrum disorder (ASD), schizophrenia (SCZ), and bipolar disorder (BD) and pertinent data were generated by Gandal et al. [7]. In total, the ncRNA gene set contains 1,363 differentially expressed human ncRNAs. In this case study, we also showcase the ability to constrain protein-coding genes with the TAD regions. For this, we use the dorsolateral prefrontal cortex TAD regions data obtained from the same study [7]. Finally, to demonstrate the capability of NoRCE working with custom gene sets, we use the Bader lab pathway gene set [12]. This dataset contains 3,758 pathways compiled from various databases (see Table S5). This data set is accessible through the NoRCE data repository and can be loaded by the users, see Section B.4 for a summary and data frame name.

Table S5: The number of pathways and the different pathway sources included in the in the March 2019 Bader Lab pathway data set.

Number of Pathways	Pathway Database
NetPath	25
IOB	33
Panther	173
NCI	223
Cyc	240
WikiPathways	490
MSigdb	520
Reactome	2,054

B.1.1 Functional enrichment results and example outputs

We use the default parameter setting in all the analyses. In NoRCE, by default, the neighborhood of each ncRNA is taken as the 10 kb extensions towards the transcription start site (TSS) and transcription end site (TES). The protein-coding genes that fall into this neighborhood region of the ncRNAs are input to the enrichment analysis. To increase the power of the statistical analysis, only the GO terms with at least 5 annotated protein-coding genes are considered as suggested by [9].

Table S6 lists a subset of the enriched GO terms and shows what the tabular format outputs should include. The full list of enriched GO terms (p-adj < 0.05) will be provided upon request. Interestingly, this list contains terms which are related to neurological functions as long-term memory, sensory perception of pain, neuronal action potential, and positive regulation of dendritic spine morphogenesis. There are also more general terms that are related to the cell cycle arrest, RNA processing, and signaling. On the downside, we observe some irrelevant terms, which are associated with bone development. However, taken as a whole, relevant terms are more pervasive, suggesting that NoRCE can indeed provide insights into ncRNA functions.

Fig 2 shows an alternative representation of the results. The dot plot views the top 35 enriched BP GO terms, sorted based on the significance of enrichment. Fig 2 shows the network that can be generated by NoRCE. This network includes the top 7 BP GO terms - enriched gene pairs interaction. The node coloring highlights the clusters with optimal modularity based on a modularity measure maximized over all possible partitions [3]. NoRCE deploys igraph package [5] for this analysis.

B.1.2 Filtering with TAD regions

We repeat the analysis in the previous section; but this time, applying a filter based on TAD regions. When this filter is applied, the assignment of the neighboring coding genes is altered; only the protein-coding genes that are nearby to the ncRNA and at the same time reside within the same TAD regions

Table S6: Top 10 enrichment results of brain related biological process GO terms that are obtained from the neighbourhood coding genes of ncRNA set of the [7]. The GeneRatio is computed by dividing the number of genes that are overlapping with the coding ones to the number of all protein-coding genes within the input set neighbourhood. The BGRatio column represents the ratio of the number of genes found in the enriched GO term set to the size of the background gene set. The EGNo refers to the size of the overlap between the corresponding GO term gene set and the neighboring coding gene set. GeneList columns contain these genes.

GO-ID	GO term	Pvalue	GeneRatio	BGRatio	EGNo	GeneList
GO:0007616	long-term memory	1.92e−4	8/1,260	33/17,523	8	PRKCZ GRIA1 ADCY1 RELN ARC DRD2 CCND2 NPAS4
GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	2.02e−4	15/1,260	91/17,523	15	COPA TMEM115 GOLPH3 KIF2A KDELR2 RINT1 BICD2 KIF26A KIF23 COPZ2 ATP9A PITPNB ARFGAP3 SEC22B KIF5A
GO:0006349	regulation of gene expression by genetic imprinting	3.11e−4	5/1,260	15/17,523	5	IGF2 EED CTCF CTCFL ARID4A
GO:0071850	mitotic cell cycle arrest	3.11e−4	5/1,260	15/17,523	5	FAP MAGI2 MCPH1 CDC14B E4F1
GO:0034765	regulation of ion transmembrane transport	3.48e−4	17/1,260	115/17,523	17	CLCNKA CLCNKB KCNK12 SCN3A SCN2A SCN1A SCN9A CACNA2D2 KCNQ5 KCNB2 KCNV1 TMEM109 KCNK4 SCN8A ASIC2 CACNA1G KCNC3
GO:0072659	protein localization to plasma membrane	3.66e−4	19/1,260	135/17,523	19	PRKCZ KCNIP3 ABCA12 DLG1 TSPAN5 ANK2 BAG4 ANK1 KCNB2 DENND4C PTCH1 RAB40A TSPAN15 GOLGA7B RAB40C RAB37 SLC9A3R1 MRAP EPB41L3
GO:0030032	lamellipodium assembly	3.75e−4	7/1,260	29/17,523	7	NCK2 GOLPH3 VAV2 ABLIM1 SPATA13 CDH13 NCK1
GO:0007050	cell cycle arrest	4.32e−4	20/1,260	147/17,523	20	LAMTOR5 TGFB2 MSH2 BARD1 INHA RASSF1 IL12A SOX2 IL12B CDK6 PRKAG2 MYC GAS1 PRKAB1 WHAMMP3 MLST8 TP53 MAP2K6 RPTOR PRKAG1
GO:0032008	positive regulation of TOR signaling	4.59e−4	7/1,260	30/17,523	7	LAMTOR5 GOLPH3 MIOS RELN WDR24 MLST8 RPTOR

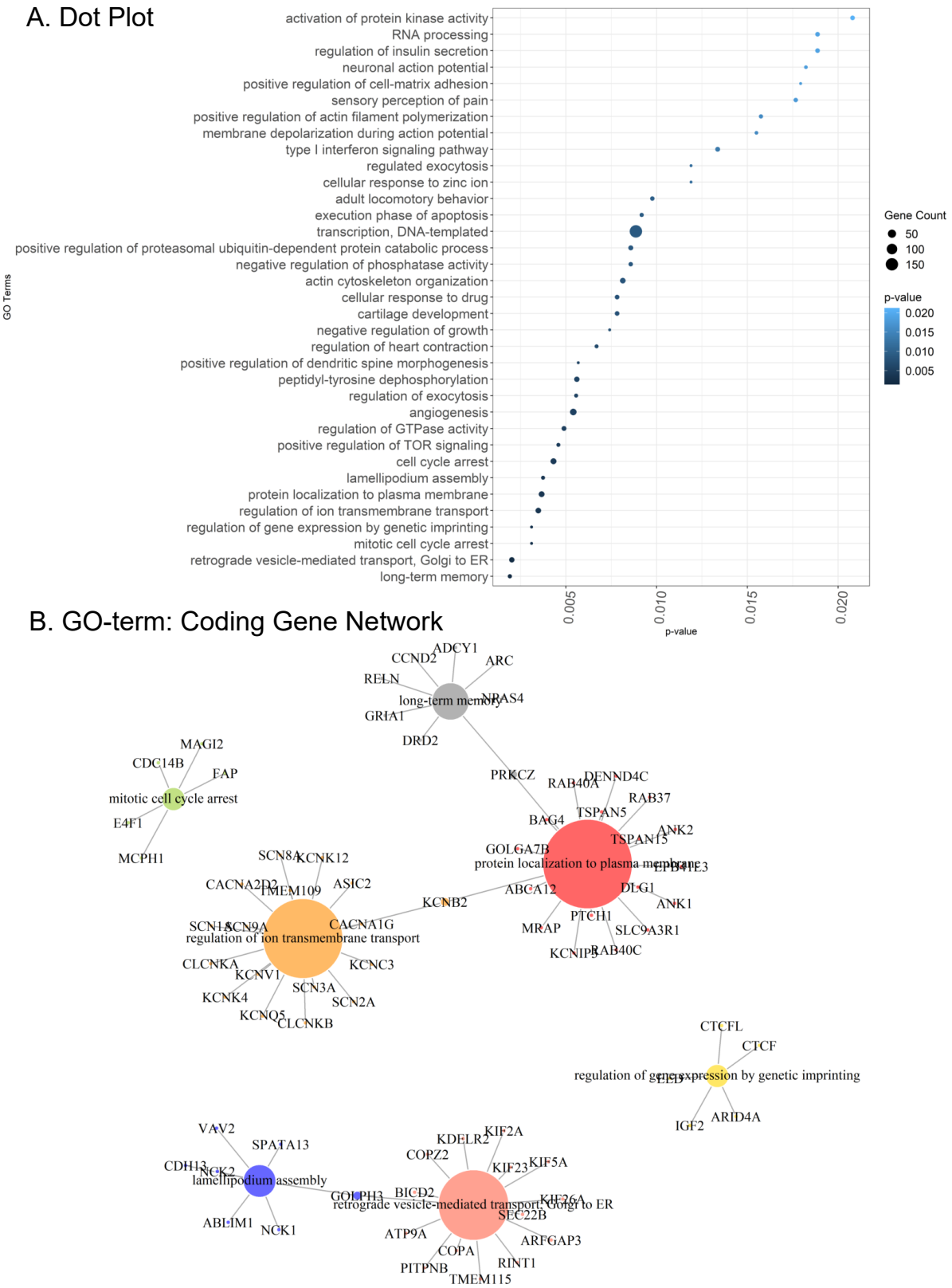

Figure 2: GO biological process enrichment results pertaining to differentially expressed brain ncRNAs. A) Top 35 GO terms are shown in the dot plot. The x-axis represents the p-values; the y-axis represents the GO terms. Dot area is proportional to the size of the overlapping gene set, and the color signifies the p-value of the enrichment test for the corresponding GO term. B) The top 7 enriched GO terms and the coding genes which are associated with them are depicted as a network. The size of the node represents the number of the coding genes which are associated with the GO term, and each color represents a different GO term.

are included. In this analysis, we use custom defined TAD regions for the adult dorsolateral prefrontal cortex that are provided by the [7] study.

Top 10 GO term enrichment results produced by this TAD based analysis are provided in Table S7 and the statistically significant results will be provided upon request. The enrichment results are visualized in Fig 3 (A). The networks of the enriched GO terms. Coding genes are provided in Fig 3 (B).

In this use case, we use a custom TAD region file. Additionally, NoRCE repository includes TAD regions for the *human*, *mouse* and *fruit fly* and can be readily used for analysis (See Section B.3).

B.1.3 Pathway enrichment using predefined pathway gene sets and custom user-defined gene sets

For pathway enrichment analysis, the latest versions of the Reactome (Reactome database in R [11], Wiki Pathways, and KEGG (KEGG database in R, [4]) databases are available in NoRCE. These databases will be updated as new versions are made available. Also, NoRCE supports pathway enrichment using custom pathway databases such as MSigDb, or other user curated data. To demonstrate this, we use the Bader Lab dataset as the user-defined pathway gene set analysis.

We provide the statistically significant enrichment results based on custom Bader Lab dataset annotation. In this analysis, given ncRNA set are expanded within 10kb, and the neighborhood coding genes are enriched with custom Bader Lab gene sets.

Table S8 lists the pathways that are found to be enriched. Many gene sets that are involved in diseases are related to brain functions such as axon guidance. This observation suggests that it is possible to gain insights into the functions of ncRNAs by studying the functions of the protein-coding genes located proximally to the ncRNAs.

B.2 Case Study 2: Functional enrichment analysis of variably expressed miRNAs in brain cancer

Integrating biological information from different sources such as co-expression data is likely to enhance the quality of functional analysis for ncRNAs. Therefore, NoRCE also supports filtering of input ncRNA gene set based on co-expression analysis. The user can input an expression data that includes coding and non-coding RNA expression. NoRCE will compute the co-expressed pairs. When defining coding gene neighborhoods for an ncRNA, the user can choose the option to include a coding gene only if it is co-expressed. Or alternatively, the user can choose to augment the list of coding genes with the co-expressed coding genes.

Exclusively for the miRNA input, NoRCE also allows filtering with miRNA targets. User can choose to filter ncRNA neighbors so that only those that are among the targets of the miRNA are included. To demonstrate these options, we use a set of miRNAs differentially expressed in brain cancer obtained from dbDEMC 2.0[21]. This set contains 407 miRNAs and is provided in NoRCE with the name *brain_ miRNA*. For the co-expression filter, we use the brain cancer patient data obtained from TCGA. The data include expression levels for mRNA and miRNA for 527 tumor patients. We apply the same pre-processing step as in our previous method [13]. Genes and miRNAs that have very low expression, (RPKM < 0.05) in many patients (more than %20 of the samples) are filtered. The gene expression values are log2 transformed, and those that have high variability are retained for co-expression analysis. For this, only the genes with median absolute deviation (MAD) above 0.5 are used.

Final expression data set contains 444 miRNA and $12,643$ mRNA genes on 527 tumor patients on which we perform Pearson correlation analysis. The mRNAs that have more than 0.3 correlation with a miRNA are retained. Additionally, we apply the miRNA target filter. NoRCE only allows the coding genes that are highly correlated with the mRNA and intersected with the nearby coding genes. Enrichment analysis is conducted on this list.

B.2.1 Functional annotation of miRNA based on custom expression

We perform enrichment analysis for the miRNAs that are differentially expressed in brain cancer. We apply default parameter settings for this analysis. The results obtained after performing enrichment analysis using a gene set that is both targeted by and co-expressed with the input miRNAs are shown in Table S8. We observe that most of the results are related to brain functionality, suggesting that miRNAs

Table S7: The Top 10 biological process GO term enrichment results. These are obtained by performing TAD based analysis using *brain_disorder_ncRNA* data. The GeneRatio column represents the ratio of the number of coding genes overlapping with the input coding genes within the GO term set to the number of all protein-coding geneswithin the neighbourhood of the input set. The BGRatio column represents the ratio of the number of genes found in the enriched GO term set to the size of the background gene set. The EGNo denotes the size of the overlap with the corresponding GO term gene set and the neighboring coding gene set. GeneList columns contain these genes.

GO-ID	GO term	Pvalue	GeneRatio	BGRatio	EGNo	GeneList
GO:0007616	long-term memory	8.86e−5	8/1,118	33/17,523	8	PRKCZ GRIA1 ADCY1 RELN ARC DRD2 CCND2 NPAS4
GO:0001525	angiogenesis	1.02e−4	27/1,118	223/17,523	27	TAL1 PTGS2 TGFB2 CYP1B1 FAP SETD2 COL8A1 PLXND1 GAB1 NUS1 EPHA1 ANGPT2 FGFR1 VAV2 UNC5B MMRN2 JAM3 PTPRB PGF WARS CIB1 MMP2 BCAS3 PRKCA RNF213 ANGPTL4 HIF3A
GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	1.81e−4	14/1,118	91/17,523	14	COPA TMEM115 GOLPH3 KIF2A KDELR2 BICD2 KIF26A KIF23 COPZ2 ATP9A PITPNB ARFGAP3 SEC22B KIF5A
GO:0071850	mitotic cell cycle arrest	1.84e−4	5/1,118	15/17,523	5	FAP MAGI2 MCPH1 CDC14B E4F1
GO:0030032	lamellipodium assembly	1.90e−4	7/1,118	29/17,523	7	NCK2 GOLPH3 VAV2 ABLIM1 SPATA13 CDH13 NCK1
GO:0072659	protein localization to plasma membrane	2.34e−4	18/1,118	135/17,523	18	PRKCZ KCNIP3 ABCA12 DLG1 TSPAN5 ANK2 BAG4 ANK1 KCNB2 DENND4C PTCH1 RAB40A TSPAN15 GOLGA7B RAB40C RAB37 SLC9A3R1 EPB41L3
GO:0007050	cell cycle arrest	2.58e−4	19/1,118	147/17,523	19	LAMTOR5 TGFB2 MSH2 BARD1 INHA RASSF1 IL12A SOX2 IL12B CDK6 PRKAG2 MYC GAS1 PRKAB1 WHAMMP3 MLST8 MAP2K6 RPTOR PRKAG1
GO:0034765	regulation of ion transmembrane transport	2.59e−4	16/1,118	115/17,523	16	CLCNKA CLCNKB KCNK12 SCN3A SCN2A SCN1A SCN9A CACNA2D2 KCNQ5 KCNB2 KCNV1 TMEM109 SCN8A ASIC2 CACNA1G KCNC3
GO:0017157	regulation of exocytosis	2.897e−4	7/1,118	31/17,523	7	STXBP5L STXBP5 RALA RAB40A RAB40C RAB37 LLGL2
GO:0051216	cartilage development	3.39e−4	10/1,118	58/17,523	10	HOXD3 SATB2 HYAL2 TAPT1 SMAD1 ZEB1 ESRRA SOX9 GNAS COL11A2

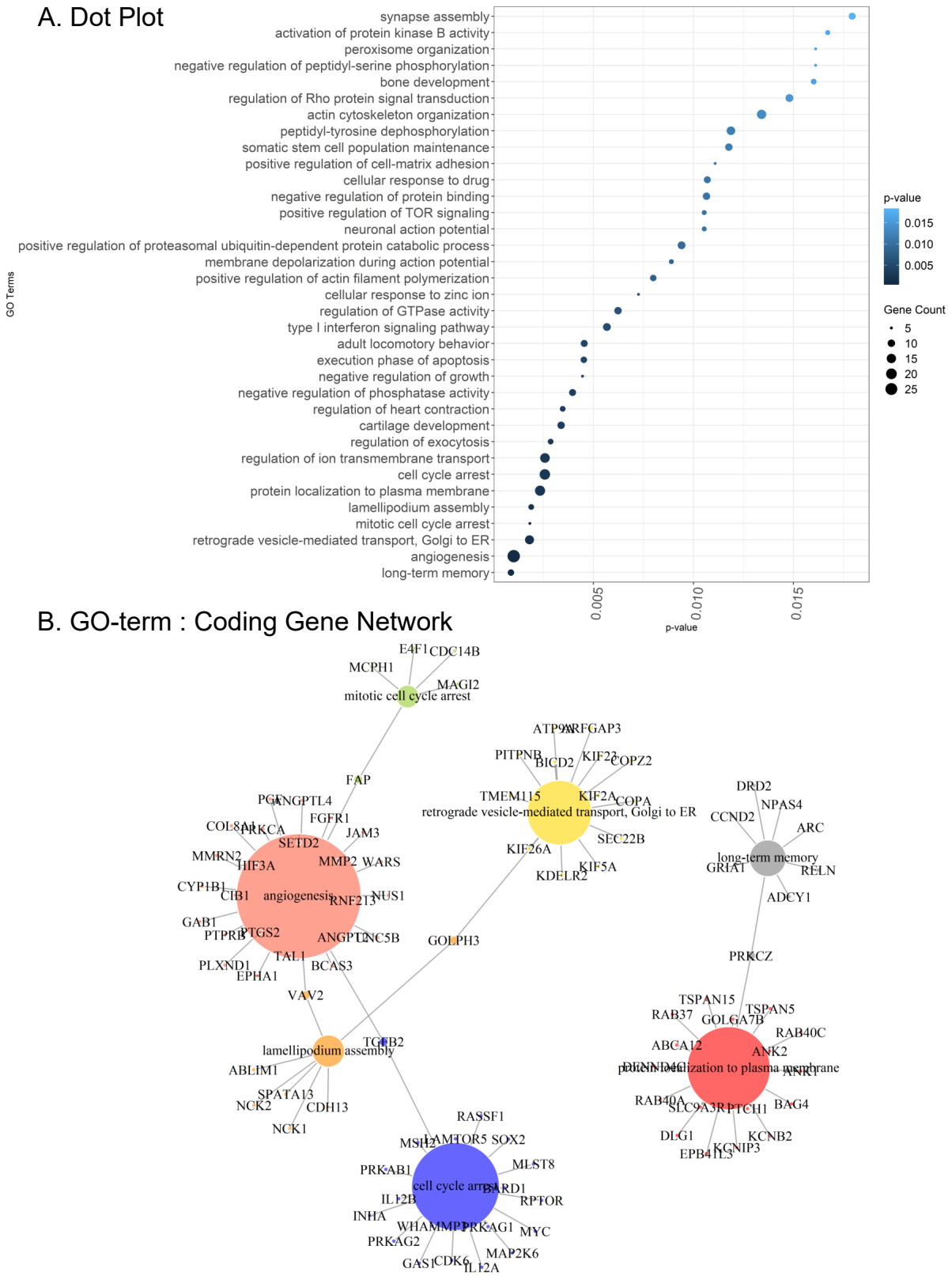

Figure 3: The enrichment results obtained by performing TAD based analysis on ncRNAs obtained from [7]. (A) The dot plot where the Y-axis denotes p-values and the x-axis denotes GO terms. The size and the color of dots signify the size of the enriched gene set and the p-value for the corresponding GO term, respectively. B) The network of GO term:protein coding gene. Size of the nodes is proportional to the degree of the node. Different colors are used to differentiate the cluster of proteins.

Table S8: Disease related pathway enrichment results for the *brain_disorder_ncRNA* dataset using customized pathway databases. The The GeneRatio column represents the ratio of the number of coding genes overlapping with the input coding genes within the GO term set to the number of all protein-coding genes within the neighbourhood of the input set. The BGRatio column represents the ratio of the number of genes found in the enriched GO term set to the size of the background gene set. The EGNo denotes the size of the overlap with the corresponding GO term gene set and the neighboring coding gene set. GeneList columns contain these genes.

GO term	Database	GO ID	Pvalue	GeneRatio	BGRatio	EGNo	GeneList	
MEIOSIS	REACTOME	R-HSA-1500620.2	5.49e−5	3.25e−4	22/967	118/12,930	9	HIST2H4A HIST2H2BE ATR TOP3A POT1 TEX12 FKBP6 SUN1 HIST2H2AC
AXON GUIDANCE	REACTOME	R-HSA-422475.5	5.60e−5	3.25e−4	65/967	533/12,930	63	RELN DAB1 LAMB1 RPL3L TRPC1 NCK1 PSMD14 PSMD7 ABLIM1 ABLIM2 ITGA10 GAB1 PRKCA ANK2 ANK1 ACTG1 SEMA4D DPYSL5 COL4A3 PLXND1 RPS2 TIAM1 RPS6 KALRN RPL11 EPHA1 RPL38 RPL14 RPS4Y1 VAV2 RPLP1 RPS28 RPS24 UNC5B DSCAM AGRN LDB1 DLG1 RASA1 PFN2 SCN9A SCN8A SCN3A SCN2A SCN1A MMP2 DOK6 NCK2 TRPC3 PSMB11 PSMA3 CXCL12 PSMD1 PSMB5 RBM8A NCAM1 CACNA1G MAGOH RNPS1 PRKCQ EPHA4 VASP ARHGEF28
Amyotrophic lateral sclerosis (ALS)	WikiPathways	WP2447	4.40e−5	1.77e−3	10/967	39/12,930	10	GRIA1 BAD NEFM PPP3CC CASP1 NEFH CASP12 TP53 MAP2K6 CASP9
Parkinsons Disease Pathway	WikiPathways	WP2371	4.93e−5	1.88e−3	16/967	85/12,930	10	MIR17HG ATXN2 GPR37 UBE2L3 UBE2J2 CASP9 SNCAIP MIR136 MIR16-2 CCNE2

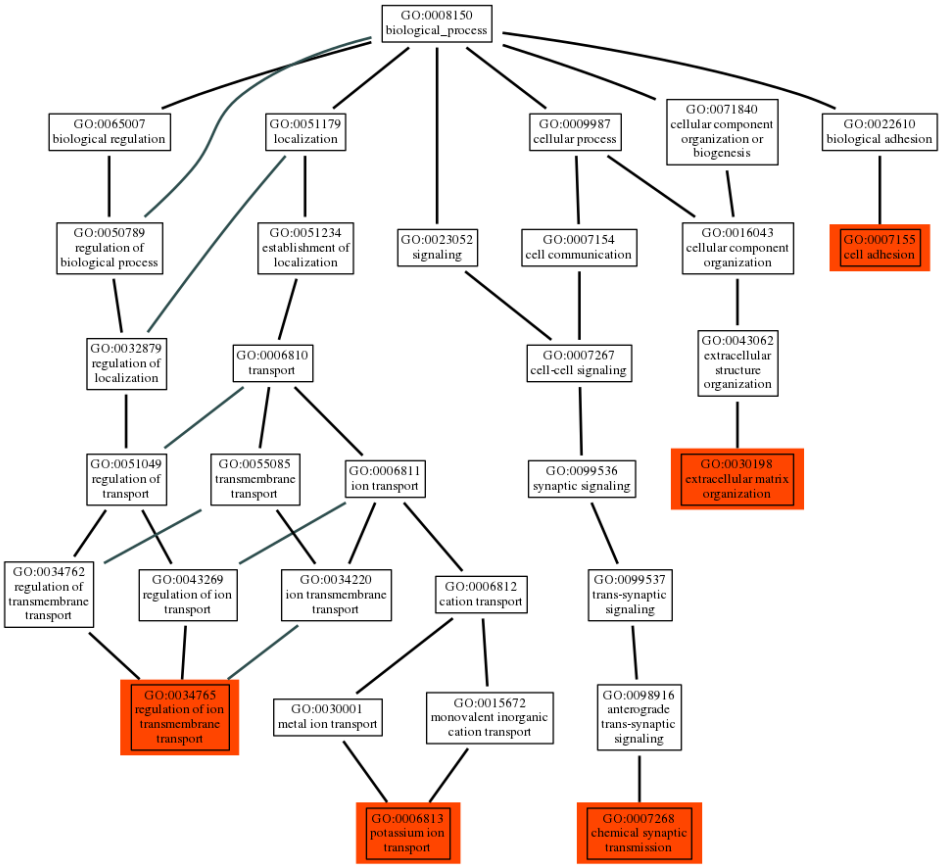

Figure 4: Directed acyclic graph of the enriched top 5 GO terms based on a gene set that are co-expressioned and targeted by the brain cancer miRNA. GO terms are colored based on their p-values

differentially expressed in brain cancer might be involved in these processes (S9). Here, Fig 4 displays the hierarchical view of the enriched GO terms, which is an alternative visualization available under NoRCE.

B.2.2 Reactome pathway enrichment of miRNA set based on neighboring protein-coding genes and miRNA target analysis

In this section, the neighboring protein-coding genes targeted by the input miRNA list (used in the previous section) are used for Reactome pathway enrichment. Results show that cellular transport based processes are enriched processes. Table S10 lists a term which is found to be enriched in the previous analysis, as well. Moreover, previous studies have shown that transport mechanisms play a role in metastatic brain tumors [18]. Together these observations suggest that NoRCE can facilitate the identification of important pathways in which ncRNAs participate. 5 also shows an alternative visualization provided by NoRCE. The genes that overlap with the ncRNA's coding neighbors are highlighted in the pathway map.

B.2.3 Brain miRNA gene list enrichment based on driver cancer gene set

We apply gene enrichment analysis to identify the regions that are statistically similar to the given gene set. In this case study, we use differentially expressed brain cancer miRNA named as *brain_mirna* in NoRCE to identify whether the neighborhood of the miRNA set is statistically similar to a brain cancer driver gene set obtained from Intogen [17]. In this analysis, only the neighborhood of the miRNAs is considered, and we apply the default parameter settings for non-coding gene enrichment. We find that brain miRNA gene list is statistically enriched for the driver gene list($p - value = 3.33\text{e}{-132}$ and

Table S9: Some brain related biological process functions that are enriched based on target and custom co-expression analysis in brain cancer specific miRNAs. The The GeneRatio column represents the ratio of the number of coding genes overlapping with the input coding genes within the GO term set to the number of all protein-coding genes that are within the neighbourhood of the input set. The BGRatio column represents the ratio of the number of genes within the enriched GO term set to the size of the background gene set. The EGNo denotes the size of the overlap with the corresponding GO term gene set and the neighboring coding gene set while GeneList columns list these genes.

GO-ID	GO term	Pval	PAdj	GeneRatio	BGRatio	GeneList
GO:0007411	axon guidance	4.66e−15	5.62e−12	88/3,617	193/17,523	ARX BDNF CDH4 CDK5R1 CHL1 CNTN2 CNTN4 CSF1R CXCL12 CXCR4 DAB1 DOK1 DOK4 DOK5 DOK6 DPYSL5 DRAXIN DRGX EFNA1 EFNA2 EFNA3 EFNA4 EFNA5 EFNB2 EFNB3 EPHA5 EPHA8 EPHB2 EPHB3 ETV1 EZR FAM129B FEZF1 FLRT2 FLRT3 FOXD1 GAP43 GATA3 GFRA2 GFRA3 GLI2 GLI3 GRB10 GRB7 KIF5A KIF5C L1CAM LAMA2 LGI1 LHX2 MATN2 NFASC NGFR NR4A3 NRP1 NRXN1 NRXN3 PAX6 PDLIM7 PIK3CD PRKCQ RAP1GAP RELN ROBO2 RPS6KA5 SCN1B SEMA3A SEMA3C SEMA3F SEMA4F SEMA6C SHANK3 SHC3 SHH SLIT1 SLIT2 SLIT3 SPTB SPTBN2 SPTBN4 TNFRSF8 UNC5A UNC5C UNC5D VASP VAX1 VLDLR WNT5A
GO:0007399	nervous system development	1.39e−12	1.01e−9	121/3,617	322/17,523	ADORA1 APBA1 APOB ARHGAP26 BARX2 BDNF BZW2 CABLES1 CAMK2A CAMK2G CBLN1 CHRDL1 CHRM1 CHRM2 CHRM3 CNTFR CNTN3 CNTN4 CPLX2 CRIM1 CSGALNACT1 CXCL1 CYP46A1 DCLK1 DCX DLG2 DLG3 DLG4 DLX5 DLX6 DOC2A DPF1 DPF3 DPYSL4 DPYSL5 EDNRB EFNA5 EFNB3 ELAVL3 ENC1 EPHB2 ERBB4 FGF12 FGF13 FGF14 FGF2 FGF5 FOS FUT9 GAS7 GDA GFRA3 GJB1 GLRB HES1 INTU JAG1 KALRN L1CAM LGI1 MAFK MDK MEF2C MPPED2 MYT1 MYT1L N4BP3 NBL1 NDN NDP NEUROD2 NINJ1 NOTCH2 NR2E1 NRG1 NRGN NRN1 NRSN1 NUMBL OLFM1 P2RX5 PCDH1 PCDH10 PCDH18 PCDH19 PCDH8 PCDHA11 PCDHA13 PCDHA5 PCDHA6 PCDHAC2 PCDHGA12 PCDHGA3 PCDHGA9 PCDHGC5 PCSK2 PTN PURA RAPGEF5 RAPGEFL1 RBFOX1 RBFOX2 RBFOX3 SCN2B SCN3B SCN8A SEMA3E SEMA4F SEMA6B SHOX2 SLC4A10 SPOCK1 SRRM4 ST8SIA4 TMOD2 TRPC5 TTLL7 VEGFA VLDLR ZC3H12A ZIC5
GO:0007409	axonogenesis	3.16e−9	1.13e−6	46/3,617	97/17,523	AMIGO1 ANK3 ATL1 ATP8A2 BCL2 BRSK1 BRSK2 CNP CNTN4 CTNNA2 DOCK7 DSCAML1 ECM2 FGFR2 FMOD LINGO1 LINGO2 LRFN2 LRFN5 LRRN3 LUM MAP2 NEFH NKX2-8 NRCAM NTNG1 NTNG2 NUMBL OGN OMD OMG PAK3 PAX2 POU4F1 PRELP RTN4R SEMA4A SLITRK1 SLITRK2 SLITRK3 SLITRK4 SLITRK5 SPTBN4 STMN1 VAX2 WNT7A
GO:0007612	learning	3.81e−8	7.87e−6	32/3,617	61/17,523	ARC ATP8A1 CHRNB2 CNTN2 CNTNAP2 DLG4 DRD1 EPHB2 FGF13 FOSL1 GRM5 JPH3 JPH4 JUN NLGN3 NPAS4 NRXN1 NRXN3 NTRK2 PRKAR2B PTGS2 PTN RGS14 SHANK2 SHANK3 SLC12A5 SLC24A2 SLC6A1 SLC8A2 SLC8A3 SORCS3 SYNJ1

Table S10: Pathway enrichment of miRNA gene set based on neighbour targets.The GeneRatio column represents the ratio of the number of coding genes overlapping with the input coding genes within the GO term set to the number of all protein-coding genes within the input set neighbourhood. The BGRatio column represents the ratio of the number of genes found in the enriched GO term set to the size of all of the genes in the background gene set. The EGNo denotes the size of the overlap with the corresponding GO term gene set and the neighboring coding gene set. GeneList columns list these genes.

GO term	GO ID	Pvalue	PAdjust	GeneRatio	BGRatio	EGNo	GeneList
R-HSA-199992	trans-Golgi Network Vesicle Budding	7.16e−5	1.00e−4	5/64	72/10,554	5	AP1S1 CLVS1 DNM2 BLOC1S3 ARRB1
R-HSA-421837	Clathrin derived vesicle budding	7.16e−5	1.00e−4	5/64	72/10,554	5	AP1S1 CLVS1 DNM2 BLOC1S3 ARRB1

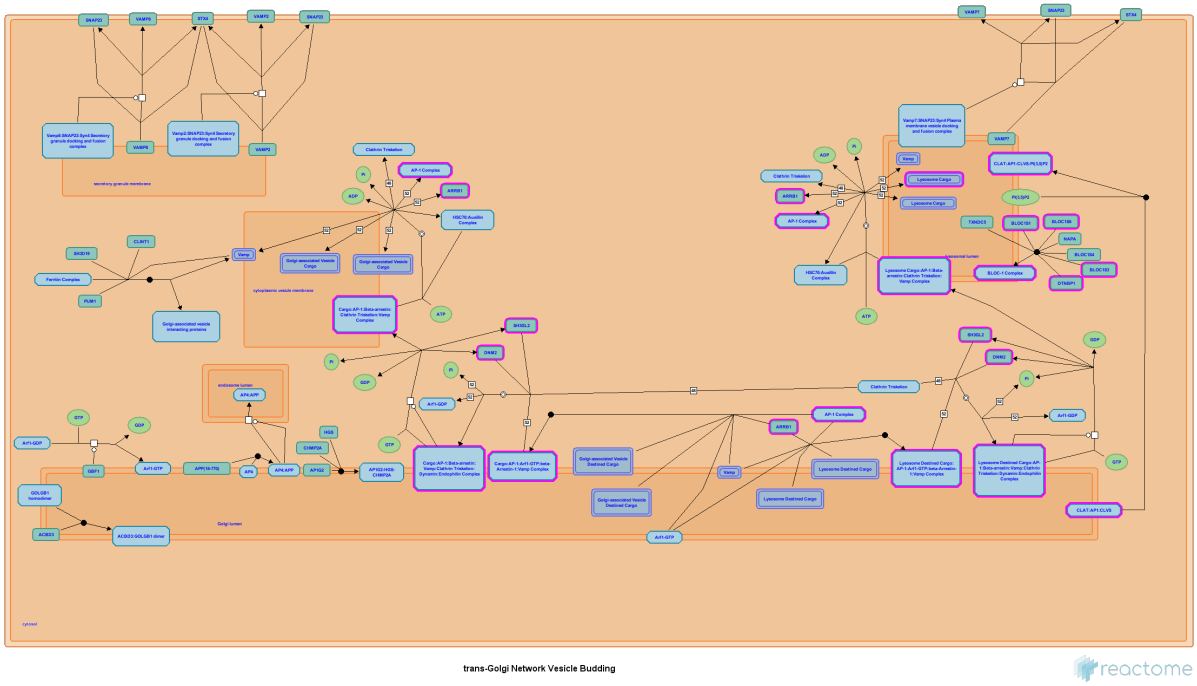

Figure 5: Reactome pathway map for the enriched pathway, R-HSA-199992(trans-Golgi Network Vesicle Budding). Genes that are enriched for this pathway are marked with purple.

$p-Adjust = 3.33e{-}132$). This analysis showcases the use of a custom gene-set for functional enrichment analysis.

B.3 Data in NoRCE repository for the vignette and supplementary results

- brain_disorder_ncRNA : A list of ncRNAs differentially expressed in three psychiatric disorders including autism spectrum disorder (ASD), schizophrenia (SCZ), and bipolar disorder (BD) generated by Gandal et al. [7]. The set comprises1,363 ncRNAs. (Case study 1, Section 3.1).
- tad_custom : The TAD regions for dorsolateral prefrontal cortex obtained from [7]. The dataset contains 2,735 TAD regions. (Case study 1, Section 3.1).

- miRCancerdb: The pre-computed Pearson correlation values of the expressions of pairs of 18,069,409 miRNAs and genes for 34 cancer types. The data incorporates 1,015 miRNAs, which are curated based on patient expression profiles that are provided by the TCGA. miRCancerdb is available at `https://figshare.com/articles/miRCancer_db_gz/5576329` [1]. (Case study 2, Section 3.3).
- brain_mirna : 407 differentially expressed miRNAs in human brain obtained from dbDEMC 2.0. (Case study 2, Section 3.4). (Case study 2, Section 3.4).
- ncRegion : The relevant regions of the differentially expressed human ncRNA genes (except pseudogenes) for three psychiatric disorders. The diseases include autism spectrum disorder (ASD), schizophrenia (SCZ), and bipolar disorder (BD), and the data were generated by Gandal et al. [7]. The final set contains 930 gene regions.

B.4 Other features

A list of features and the related questions that can be answered using NoRCE.

- Filter the ncRNAs based on biotypes and perform enrichment analysis on the curated ncRNA gene list
- Perform GO annotation and enrichment on the GO-slim dataset
- Neighbourhood gene set can be limited to the protein coding genes that are exons or introns and fall into the user-defined neighborhood region

References

[1] Mahmoud Ahmed, Huynh Nguyen, Trang Lai, and Deok Ryong Kim. mircancerdb: a database for correlation analysis between microrna and gene expression in cancer. *BMC research notes*, 11(1):103, 2018.

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

[3] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On modularity clustering. *IEEE transactions on knowledge and data engineering*, 20(2):172–188, 2008.

[4] M. Carlson. kegg.db, 2016. KEGG.db: A set of annotation maps for KEGG. R package version 3.2.3.

[5] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[6] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005.

[7] Michael J Gandal, Pan Zhang, Evi Hadjimichael, Rebecca L Walker, Chao Chen, Shuang Liu, Hyejung Won, Harm van Bakel, Merina Varghese, Yongjun Wang, et al. Transcriptome-wide isoform-level dysregulation in asd, schizophrenia, and bipolar disorder. *Science*, 362(6420):eaat8127, 2018.

[8] Calvin H Jan, Robin C Friedman, J Graham Ruby, and David P Bartel. Formation, regulation and evolution of caenorhabditis elegans 3âĂš utrs. *Nature*, 469(7328):97, 2011.

[9] Qinghua Jiang, Rui Ma, Jixuan Wang, Xiaoliang Wu, Shuilin Jin, Jiajie Peng, Renjie Tan, Tianjiao Zhang, Yu Li, and Yadong Wang. Lncrna2function: a comprehensive resource for functional investigation of human lncrnas based on rna-seq data. In *BMC genomics*, volume 16, page S2. BioMed Central, 2015.

[10] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *cell*, 120(1):15–20, 2005.

[11] W Ligtenberg. reactome.db, 2019. reactome.db: A set of annotation maps for reactome. R package version 1.68.0.

[12] Daniele Merico, Ruth Isserlin, Oliver Stueker, Andrew Emili, and Gary D Bader. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one*, 5(11):e13984, 2010.

[13] Gulden Olgun, Ozgur Sahin, and Oznur Tastan. Discovering lncrna mediated sponge interactions in breast cancer molecular subtypes. *BMC genomics*, 19(1):650, 2018.

[14] Fidel Ramírez, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke. High-resolution tads reveal dna sequences underlying genome organization in flies. *Nature communications*, 9(1):189, 2018.

[15] J Graham Ruby, Alexander Stark, Wendy K Johnston, Manolis Kellis, David P Bartel, and Eric C Lai. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of drosophila micrornas. *Genome research*, 17(12):000–000, 2007.

[16] Chanseok Shin, Jin-Wu Nam, Kyle Kai-How Farh, H Rosaria Chiang, Alena Shkumatava, and David P Bartel. Expanding the microrna targeting code: functional sites with centered pairing. *Molecular cell*, 38(6):789–802, 2010.

[17] David Tamborero, Carlota Rubio-Perez, Jordi Deu-Pons, Michael P Schroeder, Ana Vivancos, Ana Rovira, Ignasi Tusquets, Joan Albanell, Jordi Rodon, Josep Tabernero, et al. Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome medicine*, 10(1):25, 2018.

[18] Shweta Tiwary, John E Morales, Sam C Kwiatkowski, Frederick F Lang, Ganesh Rao, and Joseph H McCarty. Metastatic brain tumors disrupt the blood-brain barrier and alter lipid metabolism by inhibiting expression of the endothelial cell fatty acid transporter mfsd2a. *Scientific reports*, 8(1):8267, 2018.

[19] Charles E Vejnar and Evgeny M Zdobnov. Mirmap: comprehensive prediction of microrna target repression strength. *Nucleic acids research*, 40(22):11673–11683, 2012.

[20] Yanli Wang, Fan Song, Bo Zhang, Lijun Zhang, Jie Xu, Da Kuang, Daofeng Li, Mayank NK Choudhary, Yun Li, Ming Hu, et al. The 3d genome browser: a web-based browser for visualizing 3d genome organization and long-range chromatin interactions. *Genome biology*, 19(1):151, 2018.

[21] Zhen Yang, Liangcai Wu, Anqiang Wang, Wei Tang, Yi Zhao, Haitao Zhao, and Andrew E Teschendorff. dbdemc 2.0: updated database of differentially expressed mirnas in human cancers. *Nucleic acids research*, 45(D1):D812–D818, 2016.