Bioinformatic Analysis of Differentially Expressed Genes (DEGs) Detected from RNA-Sequence Profiles of Mouse Striatum

Bandhan Sarker, Md. Matiur Rahaman, Suman Khan, Priyanka Bosu, and Md. Nurul Haque Mollah

Abstract Bioinformatic analysis is a powerful statistical analysis to investigate the significant genes and their biological information from RNA-sequence (RNA-Seq)-based gene expression profiles. The most differentially expressed genes (DEGs) of mouse striatum with their valuable information may be significantly contributed to the neuroscience research. Two inbred mouse strains, for instance, C57BL/6J (B6) and DBA/2J (D2), in neuroscience research are commonly used, and B6 strain sequences are mostly available. Our study's focus on the identification of significant DEGs of B6 and D2 samples, protein–protein interaction network, to identify their biological functions, molecular pathway analysis, miRNAs-target gene interactions, downstream analysis, and to find out driven genes. Two samples, 10 B6 and 11 D2, were deeply analyzed, which were retrieved from the Gene Expression Omnibus (GEO) database with accession number GSE26024. DESeq2, edgeR, and limma tools were utilized to screen the DEGs somewhere in the range of B6 and D2 samples. DESeq2, edgeR, and limma had identified a total of 736, 757, and 530 DEGs with 37, 48, and 31 up-regulated genes, respectively. Protein–protein interaction network analyses of those DEGs were visualized using a search tool for the Retrieval of Interacting Genes and Cytoscape software. We selected the top 50 high-degree hub DEGs for each of the three methods, and explored 21 common hub genes along with three up-regulated genes Bdkrb2, Aplnr, and Ccl28. To explore the biological insights of these 21 common hub DEGs, Gene Ontology (GO) and KEGG pathway analysis were executed. In downstream analysis, hierarchical and k-means clustering techniques were used, and both the methods clustered Bdkrb2, Aplnr, and Ccl28 genes into the same group. Furthermore, DEGs, specifically the genes Bdkrb2, Aplnr, and Ccl28, are probably the core genes in inbred mouse strains. In conclusion, these genes probably are the biomarkers for further neuroscience research.

B. Sarker · Md. M. Rahaman (✉) · S. Khan · P. Bosu
Department of Statistics, Faculty of Science, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj 8100, Bangladesh

Md. N. H. Mollah
Bioinformatics Laboratory, Department of Statistics, University of Rajshahi, Rajshahi 6205, Bangladesh
e-mail: mollah.stat.bio@ru.ac.bd

B. K. Sinha et al. (eds.), *Data Science and SDGs*,
https://doi.org/10.1007/978-981-16-1919-9_9

Keywords Mouse RNA-Seq profiles · Differential gene expressions · Functional analysis · Molecular pathway analysis · Bioinformatic approaches

1 Introduction

RNA-Seq may be a way to investigate the number and sequences of RNA in a sample. Over the past decade, the revolution of next-generation sequencing has exceedingly produced a greater yield of sequence data at an inferior cost (Van Dijk et al. 2014). Simultaneously, analysis techniques used for inspecting sequence data have emerged (Alioto et al. 2013; Anders et al. 2013; Huber et al. 2015). Among the widespread methods, RNA-Seq is the largest project for analyzing sequence data. Over the past decade, the genome-wide mRNA expression data derivation from cell population has been demonstrated to be useful in many more studies (Soneson and Delorenzi 2013; Bacher and Kendziorski 2016).

Although traditional expression methods existed for analyzing thousands of cells, they sometimes cover or even misinterpret ones of interest. Nowadays, advanced technologies allow us to induce transcriptome-wide large-scale information from cells. This improvement is not simply another progression to enhance expression profiling, yet rather a major development that will empower crucial experiences into biology (Bacher and Kendziorski 2016). The analysis of RNA-Seq data assumes a crucial part to understand the inherent and extraneous cell measures in biological and biomedical exploration (Wang et al. 2019a). To understand biological processes, a more precise understanding of the transcriptome in cells is needed for explicating their role in cellular functions and understanding how differentially expressed genes (DEGs) can promote advantageous or harmful design (Hwang et al. 2018). Appropriate analysis and utilization of the massive amounts of data generated from RNA-Seq experiments are challenging (Pop and Salzberg 2008; Shendure and Ji 2008). However, DEGs detection is one of the most significant efforts in RNA-Seq data analysis. Several methods have been used for identifying DEGs from count RNA-Seq data in bioinformatic analysis based on Poisson and negative binomial distribution. Poisson distribution faces an over-dispersion problem; therefore, the negative binomial distribution is more reliable. In this study, we used three familiar methods (DESeq2, edgeR, and limma) to follow negative binomial distribution for examining DEGs, and we are going to discuss the fundamental principles of bioinformatic techniques, focusing on concepts important in the analysis of RNA-Seq mouse striatum data.

Multiple brain regions based on different inbred mouse strains gene expression profiles have been established previously (Sandberg et al. 2000; Hovatta et al. 2005). The distinct opioid-related phenotype has been studied by gene expression profiling in the mouse striatum (Korostynski et al. 2006). Strain reviews exhibited that affectability to morphine is an unprecedented degree reliant on hereditary determinants. In our study, we performed bioinformatic analysis on gene expression profiles of mouse striatum and chose two samples, C57BL/6J and DBA/2J. DESeq2, edgeR, and limma detected the DEGs and took the top 50 DEGs from each. From these genes, we determined common hub DEGs, and performed GO annotation and KEGG pathway analysis. For common hub DEGs, miRNA–mRNA network is constructed. After that, downstream analysis is also carried out to find the driven genes. Therefore, the bioinformatic approach paved the way for the investigation of genes from RNA-Seq profiles of mouse striatum that can be contributed further to molecular research in neuroscience.

2 Materials and Methods

We analyzed RNA-Seq read count data of mouse striatum. The following flow chart shown in Fig. 1 describes the steps of bioinformatic analysis of the data set used in this study.

2.1 RNA-Seq Data Collection

We downloaded gene expression profile GSE26024 from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26024) (Bottomly et al. 2011). It is also available at http://bowtie-bio.sourceforge.net/recount/. GSE26024 dataset contains 21 samples, including two samples, 10 C57BL/6J (B6) and 11 DBA/2J (D2), and 36,536 genes (Korostynski et al. 2006; Wang et al. 2019b).

2.2 Methods for Identification of DEGs

For identifying the DEGs from the RNA-Seq dataset, several methods such as DESeq, DESeq2, EBSeq, edgeR, baySeq, limma, NBPSeq, etc., have been developed. In our study, three popular methods, DESeq2 (Love et al. 2014), edgeR (Robinson et al. 2010), and limma (Smyth et al. 2005), were used from Bioconductor (www.bioconductor.org) project to examine the DEGs between B6 and D2 samples. The following subsections explain a summary of these three methods.

2.3 DESeq2

DESeq2 is described based on the negative binomial distribution model (Love et al. 2014). A generalized linear model is used for DESeq2 and the model form is:

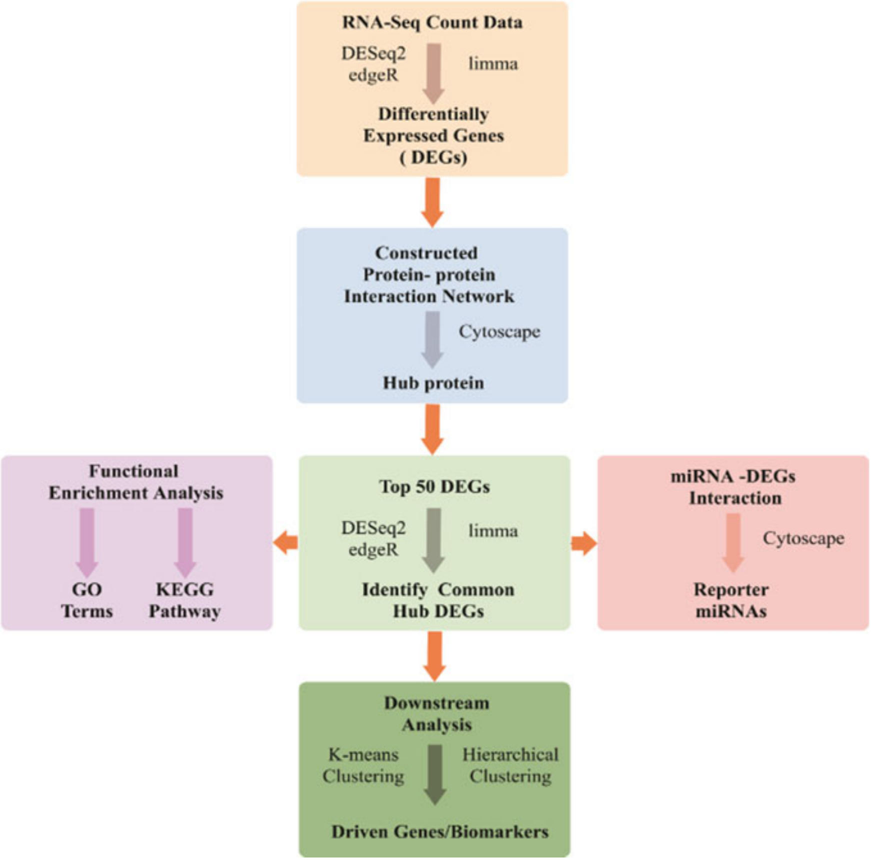


Fig. 1 RNA-Sequencing profiles of mouse striatum data analysis workflow (*Source* Created by the authors)

$$K_{ij} \sim NB(\mu_{ij,\alpha i})$$
$$\mu_{ij} = s_j q_{ij}$$
$$\log_2(q_{ij}) = x_j \beta_i$$

where, count K_{ij} is i-th gene and j-th sample model supported a negative binomial distribution; fitted mean and gene-specific dispersion parameters are denoted by μ_{ij} and α_i, respectively. The fitted mean is examined by a sample-specific size factor and a parameter, s_j and q_{ij}, respectively. The coefficients β_i calculated the log$_2$-fold changes of the model matrix (**X**) each column for gene i. Sample and gene-dependent normalization factors s_{ij} are used after generalization of the model and the variance of counts K_{ij},

$$Var(K_{ij}) = [(K_{ij} - \mu_{ij})^2] = \mu_{ij} + \alpha_i \mu_{ij}^2$$

Maximum a posterior estimation of the $\log_2$-fold changes in β_i after incorporating a zero-centered normal prior provided by DESeq2 (Love et al. 2014).

2.4 edgeR

edgeR model and software was developed by Robinson et al. (2010). edgeR considered the hypothesis,

$$H_0 : \lambda_{j1} = \lambda_{j2} \text{ (Equally expressed)}$$

$$(and) H_A : \lambda_{j1} \neq \lambda_{j2} \text{ (Differentially expressed)}$$

In edgeR, the proportion of total reads, $\lambda_{jk(i)} = \sum_{i=1}^{C_k} \lambda_{ji}$, where, $\lambda_{jk(i)}$ is the jth genes of the kth group, and λ_{ji} is defined as the proportion of reads in the jth gene in an ith sample. Then the moderate mean $\mu_{ji} = n_i \lambda_{jk(i)}$, where, n_i is the ith library. According to the gene-wise or pair-wise assumption, the dispersion parameter ϕ is estimated by a maximizing conditional weighted log-likelihood,

$$WL(\phi_j) = l_j(\phi_j) + \alpha l_c(\phi_j)$$

where α is the weight, l_c is the maximum estimator denoted by $\hat{\phi}_j^{WL}$ which is considered as an empirical Bayesian solution. To estimate dispersion parameter, Robinson et al. 2010 proposed quantile-adjusted conditional maximum likelihood (qCML) and CML as follows,

$$y_{ji} \sim NB(\mu_{ji}, \phi)$$

The maximum likelihood estimator (MLE) becomes $\frac{\sum_{i \in c_j} y_{ji}}{\sum_{i \in c_j} n_i}$ and the dispersion parameter is given as $Z_j = \sum_{i=1}^{m_j} y_{ki}$ and the common likelihood function $l_c(\phi)$,

$$l_c(\phi) = \sum_{j=1}^{G} l_j(\phi) = \sum_{j=1}^{G} \sum_{k=1}^{K} [\sum_{g=1}^{m_k} \log \Gamma(y_{ki} + \phi^{-1}) +$$
$$\log \Gamma(n_k \phi^{-1}) - \log \Gamma(Z_k + n_k \phi^{-1}) - n_k \log \Gamma(\phi^{-1})]$$

To assess the perfect dispersion parameter ϕ, the common likelihood $l_c(\phi)$ is used and the MLE of $\lambda_{jk(i)}$ depending on ϕ. After estimating MLE, the hypothesis is tested, and the alternative hypothesis H_A is used for identifying differentially expressed genes.

2.5 Limma

Linear models for microarray data, *i.e.*, the limma tool is broadly used for the analysis of RNA-Seq data (Law et al. 2014). Different steps of limma for analyzing DEGs are described as follows:

(a) 1st step is the normalizing of the data. Suppose, data matrix r_{gi} defined the RNA-Seq read count, where row and column defined the genes and samples, respectively ($g = 1, 2 \dots G$; $i = 1, 2, \dots, nk$). Voom method is used to transform the read count data matrix to log-counts per million (log-cpm) as follows:

$$y_{gi} = \log\left(\frac{c_{gi} + 0.5}{C_i + 1} \times 10^6\right)$$

where C_i denotes the mapped reads for sample i,

$$C_i = \sum_{g=1}^{G} C_{gi}.$$

(b) 2nd step is the searching of low expression genes and filter them.

(c) 3rd step is the introduction of a linear model for analyzing DEGs that describes the treatment factors assigned to different RNA samples. The model used here is:

$$E(y_{gi}) = \mu_{gi} = x_i^T \beta_g.$$

Here, covariate vector x_i and an unknown coefficient β_g represent $\log_2$-fold changes with the range of conditions of the experiment. In matrix form,

$$E(y_g) = X\beta_g$$

Here, a log-cpm value of vector is y_g for gene g and the design matrix is X with column x_i. The fitted model is

$$\hat{\mu}_{gi} = x_i^T \hat{\beta}_g$$

The mean log-cpm is transformed to mean log count value by:

$$\tilde{c} = \bar{y}_g + \log_2(\tilde{C}) - \log_2(10^6)$$

Here, the geometric mean is $\tilde{C}$. The log-cpm fitted values $\hat{\mu}_{gi}$ are transformed into fitted counts by

$$\hat{\lambda}_{gi} = \hat{\mu}_{gi} + \log_2(C_i + 1) - \log_2(10^6)$$

(d) Calculated voom weights using LOWESS curve (Cleveland, 1979) that is statistically robust and used to describe a piecewise function lo () which is linear. After that, the voom weight is $w_{gi} = lo(\hat{\lambda}_{gi})^{-4}$ called voom precision.

(e) Then fitted the contrasts of coefficient. The contrast is given by $\beta_g = M^T \alpha_g$, where, M defined the contrasts matrix, $\hat{\beta}_{gi} | \beta_{gi}, \sigma_g^2 \sim N(\beta_{gi}, v_{gi}\sigma_g^2)$.

(f) Empirical Bayes is used for getting better estimates, and it assumes the inverse of Chi-square prior σ_g^2 with mean s_0^2, f_0 is the degrees of freedom, and f_g is the residual degree. The posterior values for the residual variances are

$$\tilde{s}_g^2 = \frac{f_0 s_0^2 + f_g s_g^2}{f_0 + f_g}$$

Then the moderate t-statistic is

$$\tilde{t}_{gi} = \frac{\hat{\beta}_{gi}}{u_{gi}\tilde{s}_g}$$

(g) Adjust p-values for false discovery rate, and access the results that make sense for identifying differentially expressed genes.

2.6 Methods for Functional Analysis of DEGs

Functional analysis is carried out for annotations of DEGs and to explain their biological insights.

2.7 PPI Analysis of DEGs

PPI network represents the interaction of proteins, where nodes and edges represent the proteins and their interaction. Search tool for the Retrieval of Interacting Genes (STRING) database (http://www.string-db.org/) was used to collect information for DEGs (Szklarczyk et al. 2015), and an interaction network was considered where combined score > 0.4. Cytoscape software version 3.7.1 was used to visualize the regulatory network of their corresponding genes (Su et al. 2014). For the analysis of core genes, Network Analyzer in Cytoscape software was used for the interaction network.

2.8 GO Enrichment and KEGG Pathway Analysis of DEGs

Normally, high-throughput genomics or transcriptomics data is annotated by the GO enrichment analysis (Ashburner et al. 2000). Additionally, KEGG is a knowledge-based database used to manage natural pathways and infections. A significant genes list was submitted to the Gene Ontology (http://www.geneontology.org/) and KEGG pathway (http://www.genome.jp/kegg/) for inspecting over-represented GO and pathway classes. GO is studied to predict the possible elements of the DEGs in BP, biological process or GO process; MF, molecular function or GO function; and CC, cellular component or GO component. KEGG pathway analysis is performed for gene functions investigation (Altermann and Klaenhammer 2005), connecting genomic information with higher-level systemic functions, etc. In addition, we have considered statistically significant over-represented pathway categories in KEGG pathway enrichment analysis.

2.9 miRNAs-Target Gene Interactions of DEGs

miRNAs molecules are involved with numerous physiological and disease processes. Each miRNA is assumed to control manifold genes to select probable miRNA–mRNA interaction within the hub genes network (Lim et al. 2003). We used miRDB (http://mirdb.org/) for miRNAs-target gene interactions (Wong and Wang 2015). Cytoscape software was used to develop a regulatory miRNA–mRNA network.

2.10 Downstream Analysis of DEGs

Clustering is crucial for understanding gene expression data. Clusters are obtained by the similarity of genes in a gene expression profile. The popular k-means clustering algorithm is used for clustering DEGs. We also used hierarchical clustering that is also known as hierarchical cluster analysis. It attempts to group genes into small clusters and to group clusters into higher-level systems (Eisen et al. 1998; Kuklin et al. 2001). A common method for visualization of gene expression data using hierarchical clustering is the *heatmap*. The *heatmap* may also be combined with hierarchical clustering methods, which may split genes into groups and/or samples together, and support to display DEGs expression pattern. This may also be useful for identifying genes that are commonly regulated, or biological signatures related to a selected condition.

Table 1 Number of DEGs with *p*-value < 0.01

Methods	DEGs	Up-regulated DEGs	Down-regulated DEGs
DESeq2	736	37	699
edgeR	757	48	709
limma	530	31	499

Source Created by the authors

3 Results

3.1 Identified DEGs

DESeq2, edgeR, and limma methods identified DEGs summarized in Table 1. We identified DEGs by considering p-value < 0.01 and discriminate up-regulated and down-regulated genes based on the cut-off criteria, log FC > 2.0 and log FC < −2.0, respectively.

3.2 PPI Analysis of DEGs

According to the information in the STRING database, the gene interaction network contained many nodes and edges. Nodes and edges are listed in Table 2. DEGs are demonstrated by the nodes, and interactions between DEGs are showed by the edges. Predicted scores (degree) are used to rank core genes.

We selected the top 50 high-degree hub DEGs for each method, and the distribution of the top 50 DEGs in the interaction network is shown in Fig. 2. The relationship between the data points and comparing points on the line are roughly 0.821, 0.844, and 0.842, and the R^2 values are 0.912, 0.907, and 0.897 for DESeq2, edgeR, and limma, respectively.

Venn diagram discovered 21 common hub DEGs among the top 50 high-degree hub DEGs as shown in Fig. 3. These 21 DEGs are Bdkrb2, C5ar1, C3ar1, Fpr1, Ccr6, Ptgs2, Mki67, Tas1r2, Sstr5, Ccl28, Aplnr, Apln, Gpr55, B2m, H2-K1, F2r, Dnajc3, Trhr, Polr1a, Adcy4 and Mog. Venn diagram is drawn using the R package "VennDiagram." Again the interaction network of the 21 common hub DEGs is made

Table 2 Nodes and edges were identified based on *p*-value < 0.01

Methods	Nodes	Edges
DESeq2	725	1441
edgeR	744	1713
limma	520	678

Source Created by the authors

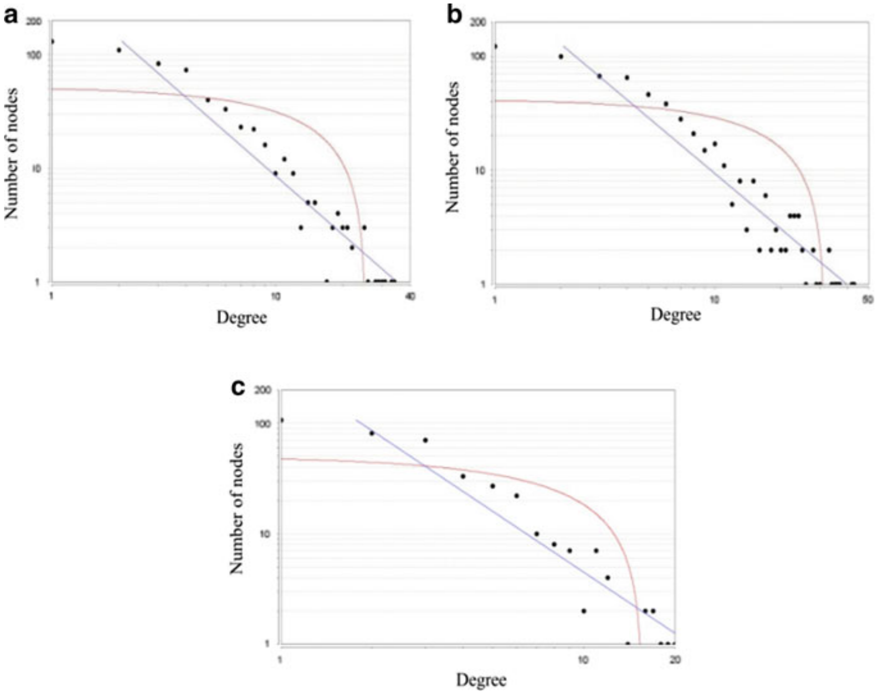
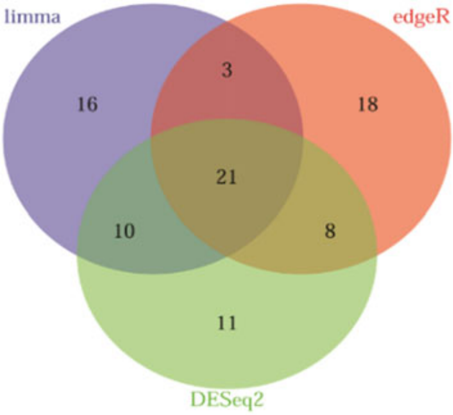


Fig. 2 Nodes-degree relationship where (**a**) DEGs found through DEseq2, (**b**) DEGs found through edgeR, and (**c**) DEGs found through limma. The dot (black) node indicates the core genes, the curve (red) indicates the fitted line, and the straight (blue) line indicates the power law. (*Source* Created by the authors)

Fig. 3 Venn diagram of the DEGs detected by the three methods (*Source* Created by the authors)



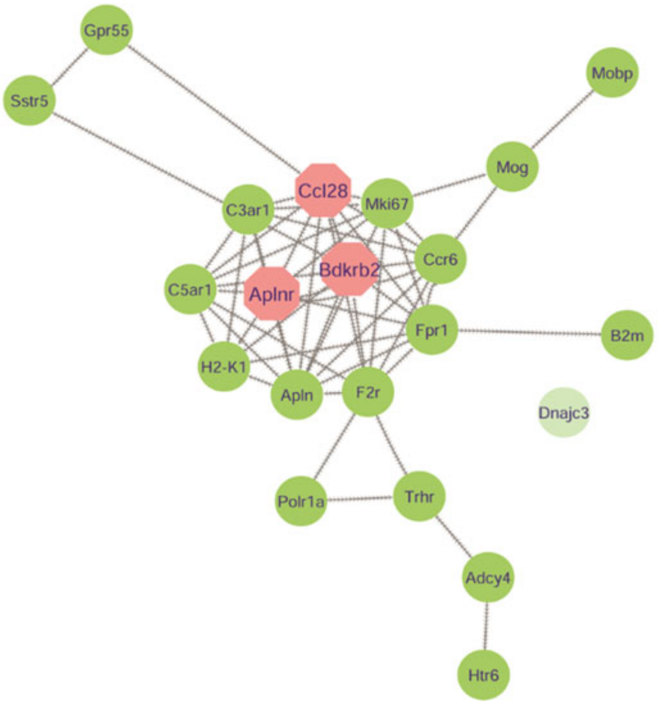


Fig. 4 The interaction network of the common 21 hub DEGs. The hub genes are indicated by the nodes, and the interactions between the hub genes are indicated by the edges. (*Source* Created by the authors)

by the STRING database containing 21 nodes and 67 edges (Fig. 4). Up-regulated hub genes Bdkrb2, Aplnr, and Ccl28 were highlighted by a different color from other down-regulated genes.

3.3 GO Enrichment Analysis of DEGs

Functional analysis of the common 21 hub DEGs is clarified through GO analysis. GO function indicates that these 21 hub DEGs are enriched in G-protein coupled peptide receptor activity, peptide binding, signaling receptor binding, etc. GO process indicates that these 21 hub DEGs are enriched in cell death, response to stimulus, signaling, homeostatic process, immune system process, response to stimulus, blood vessel development, cAMP-mediated signaling, heart development, and other biological processes. For the GO component, the 21 hub DEGs were enriched in the plasma membrane, an integral component of the plasma membrane, cytoplasmic vesicle, and so on. GO analysis results of these DEGs are explained in Table 3.

Table 3 GO enrichment analysis of common 21 hub DEGs

Gene	Gene title	GO: function	GO: process	GO: component
Bdkrb2	Bradykinin receptor, beta 2	G-protein coupled peptide receptor activity, peptide binding	Cell death, response to stimulus, signaling	Plasma membrane
C5ar1	C5a anaphylatoxin chemotactic receptor 1	G-protein coupled receptor activity, phospholipase C activity	Activation of phospholipase C activity, immune response, immune response-activating cell surface receptor signaling pathway, inflammatory response	Intracellular, cytoplasmic vesicle
C3ar1	C3a anaphylatoxin chemotactic receptor	G-protein coupled receptor activity, phospholipase C activity	Activation of phospholipase C activity, immune response, immune response-activating cell surface receptor signaling pathway, inflammatory response, inositol phosphate-mediated signaling	Intracellular, plasma membrane
Fpr1	fMet-Leu-Phe receptor	G-protein coupled peptide receptor activity, phospholipase C activity	Activation of phospholipase C activity, immune response, immune response-activating cell surface receptor signaling pathway, inflammatory response, inositol phosphate-mediated signaling	Intracellular, plasma membrane
Ccr6	C–C chemokine receptor type 6	G-protein coupled peptide receptor activity	Calcium-mediated signaling, cell chemotaxis, immune response, positive regulation of cytosolic calcium ion concentration	External side of the plasma membrane, intracellular
Ptgs2	Prostaglandin G/H synthase 2	Oxidoreductase activity, cell death, response to stimulus	Immune system process, system development, cell differentiation	Endoplasmic reticulum, plasma membrane

(continued)

Table 3 (continued)

Gene	Gene title	GO: function	GO: process	GO: component
Mki67	Proliferation marker protein Ki-67	DNA binding	Cell population proliferation, system development	Non-membrane-bounded organelle, nucleus
Tas1r2	Taste receptor type 1 member 2	G-protein coupled receptor activity, taste receptor activity	Sensory perception of sweet taste response to stimulus, signaling	Integral component of plasma membrane
Sstr5	Somatostatin receptor type 5	G-protein coupled receptor activity, neuropeptide binding	Neuropeptide signaling pathway, response to stimulus, signaling, homeostatic process	Integral component of plasma membrane
Ccl28	C–C motif chemokine 28	Signaling receptor binding	Homeostatic process, immune system process, response to stimulus	Cytoplasmic vesicle
Aplnr	Apelin receptor	G-protein coupled peptide receptor activity	Blood vessel development, cAMP-mediated signaling, heart development	Intracellular, plasma membrane
Apln	Apelin	Signaling receptor binding, extracellular region	Cell population proliferation, establishment of localization, signaling	Extracellular region, signaling receptor binding
Gpr55	G-protein coupled receptor 55	G-protein coupled receptor activity, phospholipase C activity	Rho protein signal transduction activation of phospholipase C activity, inositol phosphate-mediated signaling, positive regulation of cytosolic calcium ion concentration	Integral component of plasma membrane intracellular
B2m	Beta-2-microglobulin		Homeostatic process, cell differentiation, system development, immune system process	Cytosol, Golgi apparatus, plasma membrane

(continued)

Table 3 (continued)

Gene	Gene title	GO: function	GO: process	GO: component
H2-K1	H-2 class I histocompatibility antigen	Peptide binding signaling receptor binding	Adaptive immune response, immune effector process, positive regulation of adaptive immune response	
F2r	Proteinase-activated receptor 1	G-protein coupled receptor activity, phospholipase C activity	Rho protein signal transduction, activation of phospholipase C activity, inositol phosphate-mediated signaling, positive regulation of cytosolic calcium ion concentration	Integral component of plasma membrane, intracellular
Dnajc3	DnaJ homolog subfamily C member 3	Chaperone binding, unfolded protein binding, signaling receptor binding	Protein folding in endoplasmic reticulum, cell differentiation, cellular component organization, system development, immune system process	Endoplasmic reticulum, plasma membrane, Golgi apparatus
Trhr	Thyrotropin-releasing hormone receptor	Signaling receptor activity	Muscle contraction, sensory perception, homeostatic process, response to stimulus, signaling	Plasma membrane
Polr1a	DNA-directed RNA polymerase subunit RPA1	RNA polymerase I activity, transferase	Nucleic acid-templated transcription	DNA-directed RNA polymerase I complex, nuclear chromatin
Adcy4	Adenylate cyclase type 4	G-protein coupled receptor activity, adenylate cyclase activity	Activation of adenylate cyclase activity, adenylate cyclase-activating G-protein coupled receptor signaling pathway, regulation of adenylate cyclase activity	Integral component of plasma membrane, intracellular

(continued)

Table 3 (continued)

Gene	Gene title	GO: function	GO: process	GO: component
Mog	Myelin-oligodendrocyte glycoprotein	Signaling receptor binding, carbohydrate derivative binding	T cell receptor signaling pathway, immune response, immune system process, response to stimulus	External side of plasma membrane, leaflet of membrane bilayer

Source Created by the authors

3.4 KEGG Pathway Analysis of DEGs

In the analysis of the KEGG pathway, we have considered a false discovery rate (FDR) less than 0.05 and found out significant genes. KEGG pathway analysis exposed and targeted pathways enriched in neuroactive ligand–receptor interaction, pathways in cancer, ovarian steroidogenesis, and other significant pathways described in Table 4.

Pathway ranking associated with genes is displayed in Fig. 5. The first-ranked staphylococcus aureus infection pathway had the 6% genes that involved C5ar1, C3ar1, and Fpr1. The second is the complement and coagulation cascades pathway with 4.5% related genes that are Bdkrb2, C5ar1, C3ar1, and F2r. The third, regulation of lipolysis in adipocytes pathway, had the 3.65% related genes that included Ptgs2, Adcy4. The fourth, the ovarian steroidogenesis pathway, had 3.51% related genes that are Adcy4, Ptgs2. And, the fifth, neuroactive ligand–receptor interaction pathway, had the 3% related genes that are Bdkrb2, C5ar1, C3ar1, Fpr1, Sstr5, Aplnr, Apln, and F2r.

3.5 miRNA–mRNA Network Construction for DEGs

The common 21 hub DEGs were closely associated with related miRNA and predicted potential miRNAs. The prediction scores were likewise gathered from the miRDB database, and therefore the miRNA–mRNA with a high score implied near-potential function of miRNA inside the guideline of the objective mRNA. The miRNA–mRNA network appeared in Fig. 6 with cutoff > 70.

3.6 Downstream Analysis for DEGs

Cluster analysis of 21 hub DEGs is shown in Fig. 7. Two popular clustering methods, hierarchical clustering and k-means, were applied for finding the similarity of DEGs. We divided DEGs into three clusters for both methods. In the k-means algorithm, it observed that Ptgs2, Mog, and Dnajc3 are clustered together in Group 1; Polr1a,

Table 4 KEGG pathway analysis of common 21 hub DEGs

Pathway	Description	Genes count	Associated genes	FDR
mmu04080	Neuroactive ligand–receptor interaction	8 of 284	Bdkrb2, C5ar1, C3ar1, Fpr1, Sstr5, Aplnr, Apln, F2r	1.3E−08
mmu04610	Complement and coagulation cascades	4 of 88	Bdkrb2, C5ar1, C3ar1, F2r	6.9E−05
mmu05150	Staphylococcus aureus infection	3 of 50	C5ar1, C3ar1, Fpr1	0.0005
mmu04020	Calcium signaling pathway	4 of 180	Bdkrb2, F2r, Trhr, Adcy4	0.0005
mmu04371	Apelin signaling pathway	3 of 134	Aplnr, Apln, H2-K1	0.0050
mmu04062	Chemokine signaling pathway	3 of 179	Ccr6, Ccl28, Adcy4	0.0095
mmu04024	cAMP signaling pathway	3 of 194	Sstr5, F2r, Adcy4	0.0103
mmu04015	Rap1 signaling pathway	3 of 207	Fpr1, F2r, Adcy4	0.0108
mmu05200	Pathways in cancer	4 of 522	Adcy4, F2r, Ptgs2, Bdkrb2	0.0129
mmu04923	Regulation of lipolysis in adipocytes	2 of 55	Ptgs2, Adcy4	0.0129
mmu04913	Ovarian steroidogenesis	2 of 57	Adcy4, Ptgs2	0.0129
mmu04612	Antigen processing and presentation	2 of 78	B2m, H2-K1	0.0189
mmu04742	Taste transduction	2 of 86	Tas1r2, Adcy4	0.0210
mmu04750	Inflammatory mediator regulation of TRP channels	2 of 119	Adcy4, F2r	0.0363
mmu04611	Platelet activation	2 of 122	F2r, Adcy4	0.0363
mmu04921	Oxytocin signaling pathway	2 of 149	Ptgs2, Adcy4	0.0463
mmu04723	Retrograde endocannabinoid signaling	2 of 145	Ptgs2, Adcy4	0.0463
mmu04072	Phospholipase D signaling pathway	2 of 145	Adcy4, F2r	0.0463
mmu04022	cGMP-PKG signaling pathway	2 of 164	Bdkrb2, Adcy4	0.0492
mmu04141	Protein processing in endoplasmic reticulum	1 of 164	Dnajc3	0.0496

Source Created by the authors

Apln, and B2m belong to Group 3; and remaining DEGs are contained in Group 2. Hierarchical clustering using heatmap presentation of DEGs observed that Ptgs2, Mog, Dnajc3, Apln, and B2m are clustered together in Group 1; Polr1a, Htr6, F2r, Sstr5, and Trhr belong to Group 3; and the remaining DEGs are clustered together in Group 2.

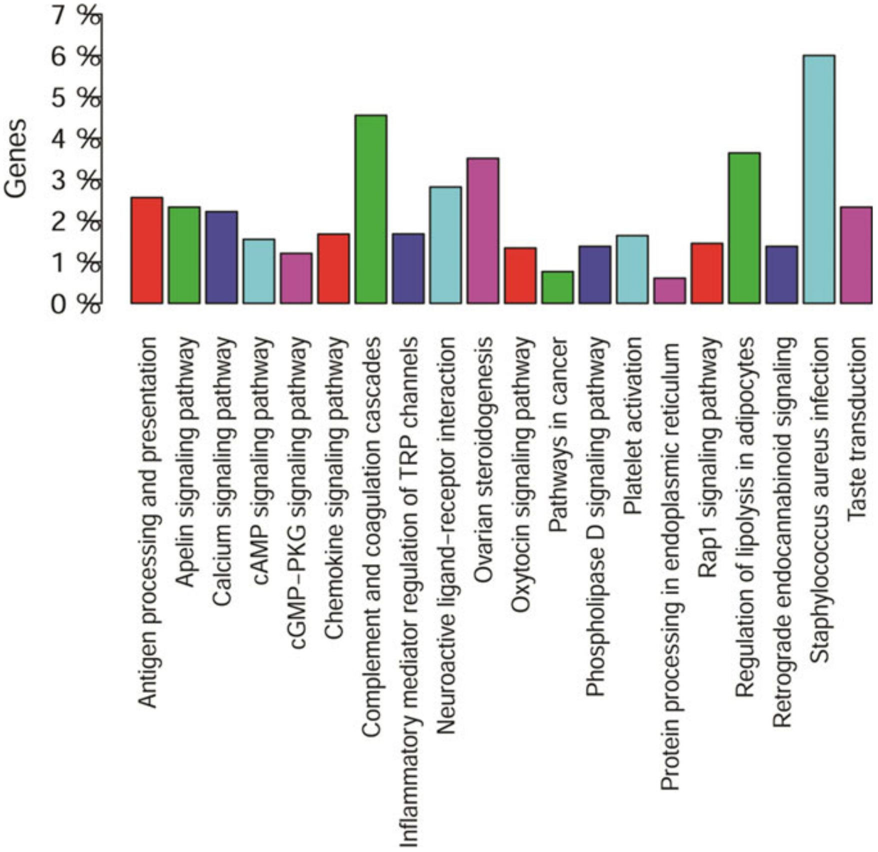


Fig. 5 KEGG analysis of common 21 hub DEGs. The different color means different pathways. (*Source* Created by the authors)

4 Discussion

The most recognized mouse strains such as C57BL/6J (B6) and DBA/2J (D2) samples are widely used in neuroscience research (Sandberg et al. 2000). In the current study, the mouse striatum gene expression profile of GSE26024 was downloaded, and to identify core genes bioinformatic analysis was performed. These investigations confirmed that 736, 757, and 530 DEGs are identified using DESeq2, edgeR, and limma with 37, 48, and 31 up-regulated genes, respectively (Table 1). Furthermore, protein–protein interaction network analysis, GO, KEGG pathway, construction of miRNA–mRNA network, and downstream analysis were executed to access the biomarkers or the core genes.

Table 2 displayed the nodes and edges of the DEGs assessed by the three different methods. The protein–protein interaction network investigation recognized the top 50 highest-degree hub genes of DEGs selected from each DEGs set. Figure 2 explained

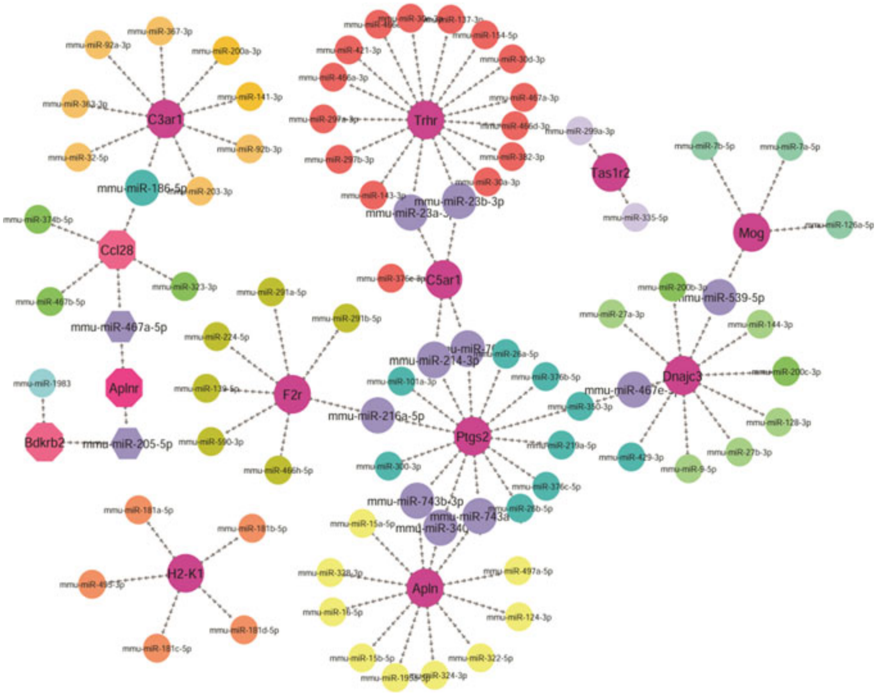


Fig. 6. miRNA–mRNA interaction network of DEGs (*Source* Created by the authors)

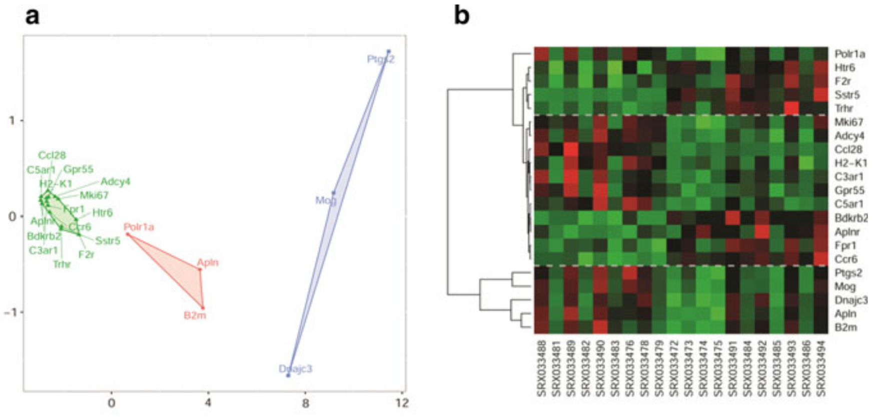


Fig. 7 Cluster analysis of common 21 hub DEGs. (**a**) K-means clustering and (**b**) Hierarchical clustering of DEGs with three clusters (*Source* Created by the authors)

nodes-degree relationship of the top 50 hub DEGs. It describes core gene distribution by giving generally high certainty that the basic model is linear in the interaction network. Figure 3 is a Venn diagram displaying and identifying 21 common hub genes based on the top 50 hub DEGs. Among the common 21 hub DEGs, DESeq2 and limma found Bdkrb2 and Ccl28, and edgeR found Aplnr up-regulated genes. The other genes are observed to be down-regulated. We also performed gene interaction network analysis for these 21 common hub genes and observed that the gene "Dnajc3" had no interaction with other genes (Fig. 4).

To disclose the underlying molecular mechanisms, we have characterized the possible GO terms and biological pathways of common hub genes. GO enrichment analysis is displayed in Table 3. The up-regulated DEGs, Bdkrb2 and Aplnr, are mainly involved in the same functional terms such as plasma membrane, and Ccl28 is associated with cytoplasmic vesicle; contrariwise, down-regulated DEGs are observed to be rich in biological intracellular, plasma membrane, endo plasmic reticulum, DNA-directed RNA polymerase I complex, non-membrane-bounded organelle, and so on.

Besides, KEGG pathway analysis is used for identifying the functional analysis of DEGs. According to KEGG pathway analysis, multiple genes are associated with same pathway as well as a same gene is associated with several pathways. Bdkrb2 is enriched in the Neuroactive ligand–receptor interaction pathway, Complement and coagulation cascades, Calcium signaling pathway, and Pathways in cancer. C5ar1 and C3ar1 regulate the Neuroactive ligand–receptor interaction, Complement and coagulation cascades, and Staphylococcus aureus infection. Fpr1 is associated with Neuroactive ligand–receptor interaction, Staphylococcus aureus infection, and Rap1 signaling pathway. Adcy4 is associated with several pathways such as Calcium signaling pathway, Chemokine signaling pathway, cAMP signaling pathway, Rap1 signaling pathway, Pathways in cancer, Regulation of lipolysis in adipocytes, Ovarian steroidogenesis, Taste transduction, Inflammatory mediator regulation of TRP channels, Platelet activation, Oxytocin signaling pathway, Retrograde endocannabinoid signaling and Phospholipase D signaling pathway, and so on (Table 4). The up-regulated DEGs, Bdkrb2 and Aplnr, are significantly enriched in Neuroactive ligand-receptor interaction pathway, while Ccl28 is enriched in Chemokine signaling pathway. Figure 5 describes the percentage of genes which are involved with different pathways.

We also have constructed miRNA–mRNA network for the common hub genes (Fig. 6). Multiple hub genes are observed to be connected with miRNAs. Trhr and C5ar1 hub genes related to mmu-miR-23a-3p and mmu-miR-23b-3p. MiR-23a downregulation is the following experiment of traumatic brain injury (Sabirzhanov et al. 2014) and MiR-23b is involved in cancer aggressive (Grossi et al. 2018). Ptgs2 and C5ar1 genes are connected with mmu-miR-761 and mmu-miR-214-3p. MiR-761 is involved in suppressing the remodeling of nasal mucosa (Cheng et al. 2020). F2r and Ptgs2 are observed to be connected with mmu-miR-216a-5p while Dnajc3 and Ptgs2 are connected with mmu-miR-467e-5p, Dnajc3 and Mog are connected with mmu-miR-539-5p, Apln and Ptgs2 are connected with mmu-miR-743a-3p, mmu-miR-743b-3p, mmu-miR-340-5p, and C3ar1and Ccl28 are connected with mmu-miR-186-5p. It is more interesting that up-regulated hub genes, Ccl28 and Aplnr, are associated with mmu-miR-467a-5p while Aplnr and Bdkrb2 are interconnected with mmu-miR-205-5p. MiR-467a is highly expressed in tumor suppressors (Inoue et al. 2017) and MiR-205 upregulation determines the aggressiveness and metastatic activity of malignant tumors (Dahmke et al. 2013).

The downstream analysis (Fig. 7) explained the cluster of 21 hub DEGs, in which maximum DEGs clustered in group 2 and a small number of DEGs clustered in group 1 and 3. We observed that the up-regulated DEGs, Ccl28, Aplnr, and Bdkrb2, belong to the same cluster (group 2) of both k-means and hierarchical clustering methods. From the above discussions, we may highlight that the genes Ccl28, Aplnr, and Bdkrb2 are crucial genes and might be the driven genes. More importantly, they might be the biomarkers for further neuroscience research.

5 Conclusions

In summary, DEGs are identified from RNA-Seq profiles of mouse striatum using the three popular DEGs calculation methods, and applied PPI network on DEGs. Then, the 21 common hub DEGs were recognized including the up-regulated genes Bdkrb2, Aplnr, and Ccl28. Analysis of GO and KEGG pathway identified significant genes to explore the biological insights of the DEGs. The downstream analysis explained that Bdkrb2, Aplnr, and Ccl28 genes belong to the same group. Finally, we have concluded that the hub genes, Bdkrb2, Aplnr, and Ccl28, might be the driven genes in inbred mouse strains. These identified driven genes might be promising candidates or biomarkers for further neuroscience research. Furthermore, experimental validation is needed and should be made in future studies.

Acknowledgements We are thankful to the reviewers for their many valuable suggestions to improve the manuscript.

References

Alioto, T., Behr, J., Bohnert, R., Campagna, D., Davis, C. A., Dobin, A., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods, 10*, 1185–1191.

Altermann, E., & Klaenhammer, T. R. (2005). PathwayVoyager: Pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. *BMC Genomics, 6*, 60.

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., et al. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols, 8*, 1765.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics, 25*, 25–29.

Bacher, R., & Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology, 17*, 63.

Bottomly, D., Walter, N. A., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., et al. (2011). Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PloS One*, *6*(3), e17820.

Cheng, J., Chen, J., Zhao, Y., Yang, J., Xue, K., & Wang, Z. (2020). MicroRNA-761 suppresses remodeling of nasal mucosa and epithelial–mesenchymal transition in mice with chronic rhinosinusitis through LCN2. *Stem Cell Research and Therapy, 11*, 1–11.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829–836.

Dahmke, I. N., Backes, C., Rudzitis-Auth, J., Laschke, M. W., Leidinger, P., Menger, M. D., et al. (2013). Curcumin intake affects miRNA signature in murine melanoma with mmu-miR-205-5p most significantly altered. *PLoS One*, *8*, e81122.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences. National Acad Sciences*, *95*, 14863–14868.

Grossi, I., Salvi, A., Baiocchi, G., Portolani, N., & De Petro, G. (2018). Functional role of microRNA-23b-3p in cancer biology. *MicroRNA, 7*, 156–166.

Hovatta, I., Tennant, R. S., Helton, R., Marr, R. A., Singer, O., Redwine, J. M., et al. (2005). Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. *Nature, 438*, 662–666.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*. Nature Publishing Group, *12*, 115.

Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine, 50*, 1–14.

Inoue, K., Hirose, M., Inoue, H., Hatanaka, Y., Honda, A., Hasegawa, A., et al. (2017). The rodent-specific microRNA cluster within the Sfmbt2 gene is imprinted and essential for placental development. *Cell Reports, 19*, 949–956.

Korostynski, M., Kaminska-Chowaniec, D., Piechota, M., & Przewlocki, R. (2006). Gene expression profiling in the striatum of inbred mouse strains with distinct opioid-related phenotypes. *BMC Genomics, 7*, 146.

Kuklin, A., Shah, S., Hoff, B., & Shams, S. (2001). *Data management in microarray fabrication, image processing, and data mining* (p. 115). Technologies and Experimental Strategies. CRC Press.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology, 15*, R29.

Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B., & Bartel, D. P. (2003). Vertebrate microRNA genes. *Science. American Association for the Advancement of Science*, *299*, 1540–1540.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*, 550.

Pop, M., & Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends in Genetics, 24*, 142–149.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*, 139–140.

Sabirzhanov, B., Zhao, Z., Stoica, B. A., Loane, D. J., Wu, J., Borroto, C., et al. (2014). Downregulation of miR-23a and miR-27a following experimental traumatic brain injury induces neuronal cell death through activation of proapoptotic Bcl-2 proteins. *Journal of Neuroscience, 34*, 10055–10071.

Sandberg, R., Yasuda, R., Pankratz, D. G., Carter, T. A., Del Rio, J. A., Wodicka, L., et al. (2000). Regional and strain-specific gene expression mapping in the adult mouse brain. *Proceedings of the National Academy of Sciences, 97*, 11038–11043.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology, 26*, 1135.

Smyth, G. K., Ritchie, M., Thorne, N., & Wettenhall, J. (2005). LIMMA: Linear models for microarray data. Bioinformatics and computational biology solutions using R and bioconductor. *Statistics for Biology and Health*, 397–420.

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics, 14*, 91.

Su, G., Morris, J. H., Demchak, B., & Bader, G. D. (2014). Biological network exploration with Cytoscape 3. *Current Protocols in Bioinformatics, 47*, 8–13.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research, 43*, D447–D452.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics., 30*, 418–426.

Wang, T., Li, B., Nelson, C. E., & Nabavi, S. (2019a). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics, 20*, 40.

Wang, J., Geisert, E. E., & Struebing, F. L. (2019b). RNA sequencing profiling of the retina in C57BL/6J and DBA/2J mice: Enhancing the retinal microarray data sets from GeneNetwork. *Molecular Vision, 25*, 345.

Wong, N., & Wang, X. (2015). miRDB: An online resource for microRNA target prediction and functional annotations. *Nucleic Acids Research, 43*, D146–D152.