Glia Open Access Database (GOAD)

Holtman, Inge R.; Noback, Michiel; Bijlsma, Marieke; Duong, Kim N.; van der Geest, Marije A.; Ketelaars, Peer T.; Brouwer, Nieske; Vainchtein, Ilia D.; Eggen, Bart J. L.; Boddeke, Hendrikus W. G. M.

Published in:
Glia

DOI:
10.1002/glia.22810

REVIEW ARTICLE

Glia Open Access Database (GOAD)

A Comprehensive Gene Expression Encyclopedia of Glia Cells in Health and Disease

Inge R. Holtman,[1] Michiel Noback,[2] Marieke Bijlsma,[2] Kim N. Duong,[2] Marije A. van der Geest,[2] Peer T. Ketelaars,[2] Nieske Brouwer,[1] Ilia D. Vainchtein,[1] Bart J. L. Eggen,[1] and Hendrikus W. G. M. Boddeke[1]

Recently, the number of genome-wide transcriptome profiles of pure populations of glia cells has drastically increased, resulting in an unprecedented amount of data that offer opportunities to study glia phenotypes and functions in health and disease. To make genome-wide transcriptome data easily accessible, we developed the Glia Open Access Database (GOAD), available via www.goad.edu*cation*. GOAD contains a collection of previously published and unpublished transcriptome data, including datasets from isolated microglia, astrocytes and oligodendrocytes both at homeostatic and pathological conditions. It contains an intuitive web-based interface that consists of three features that enable searching, browsing, analyzing, and downloading of the data. The first feature is differential gene expression (DE) analysis that provides genes that are significantly up and downregulated with the associated fold changes and *p*-values between two conditions of interest. In addition, an interactive Venn diagram is generated to illustrate the overlap and differences between several DE gene lists. The second feature is quantitative gene expression (QE) analysis, to investigate which genes are expressed in a particular glial cell type and to what degree. The third feature is a search utility, which can be used to find a gene of interest and depict its expression in all available expression data sets by generating a gene card. In addition, quality guidelines and relevant concepts for transcriptome analysis are discussed. Finally, GOAD is discussed in relation to several online transcriptome tools developed in neuroscience and immunology. In conclusion, GOAD is a unique platform to facilitate integration of bioinformatics in glia biology.

GLIA 2015;63:1495–1506

Key words: transcriptome analysis, bioinformatics, RNA-seq, microglia, astrocytes, oligodendrocytes

The Aim of GOAD

Public databases for the storage and retrieval of genomic data have become an integral component of biomedical research. Such databases are often developed by large consortia that generated extensive datasets. Currently no platform is provided that integrates datasets from different studies in a comprehensive, easily accessible way for glia researchers. In this review, the Glia Open Access Database (GOAD) is presented, which is available via www.goad.education. Usage of this tool requires no programmatic or advanced bioinformatics skills, and this review additionally provides a general introduction to transcriptome analysis. We strived to develop a platform to facilitate further integration of bioinformatics in glia biology. GOAD will be updated at a regular basis to make newest datasets rapidly available. Future plans to include more organisms and other types of glia related genome-wide sequencing data are presented.

View this article online at wileyonlinelibrary.com. DOI: 10.1002/glia.22810

Published online March 25, 2015 in Wiley Online Library (wileyonlinelibrary.com). Received Oct 10, 2014, Accepted for publication Feb 13, 2015.

Address correspondence to H.W.G.M. Boddeke, Medical Physiology, University of Groningen, University Medical Center Groningen, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands. E-mail: h.w.g.m.boddeke@umcg.nl

From the [1]Medical Physiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [2]School for Life Science and Technology, Hanze University of Applied Sciences, Groningen, The Netherlands

Bart J. L. Eggen and Erik H. W. G. M. Boddeke contributed equally to this work.

Introduction to Transcriptomics

Transcriptome

The transcriptome is the set of all RNA molecules, including messenger RNA (mRNA) and microRNA (miRNA) in a cell population or tissue (Tuck and Tollervey, 2011). Most genome-wide expression studies enrich for mRNA, because the expression of protein-coding mRNA is most clearly associated with cell identity and function. In addition, genome-wide miRNA expression profiles were generated because miRNA can regulate mRNA stability and/or translation (Weichenhan and Plas, 2013). A recent article (Zhang et al., 2014) also focused on long noncoding RNA (lncRNA) in glia subtypes. In the context of glia biology, the majority of published articles focused on mRNA; accordingly here the term transcriptome refers to all mRNA expressed by a particular cell type or tissue.

Measuring the Transcriptome

The main aim of most transcriptome studies is to quantify (differential) gene expression, but several steps have to be taken before quantification can be done. These preprocessing procedures are very different for RNA sequencing (RNA-seq) and microarray hybridization studies.

Microarray Preprocessing and Quantification. For microarrays, the process from initial measurement to quantification is relatively straightforward. First, the transcriptome is converted to fluorescently labeled cDNA that binds to predesigned probes. The light intensity is converted to arbitrary units, and after a few preprocessing steps such as quality control, normalization and background removal, the data can be used for further analysis. It is important to mention that microarrays contain predesigned probes that measure the expression of a selected group of genes, and are therefore not truly genome-wide. Microarray preprocessing and analysis are most commonly performed using Limma Bioconductor-package (Smyth, 2005), and readers that are interested in learning more about these procedures are recommended to read the Limma user manual. See Table 1 for a list of the recommended transcriptome analysis tools and their URLs.

RNA-Seq Preprocessing and Quantification. In contrast to microarrays, RNA-seq data is computationally far more intensive and needs more rigorous and time-consuming analysis. In RNA-seq, cDNA or RNA is fragmented and the nucleotide sequence at the end(s) of these fragments is determined. This results in extensive lists containing ATCG-values for each position, with an associated quality statistic for each base pair. There are two options for sequencing. Each fragment can be sequenced from one end only (single-end sequencing) or from both ends (paired-end sequencing). Generally, paired-end sequencing results in more reliable alignment and is better suited to detect previously unidentified splice variants, transcripts or genes (McGettigan, 2013). However, it is also more expensive. Often several samples are simultaneously sequenced (multiplexed) in the same lane of a sequencer using bar-coding. The first step is to demultiplex the files, generating individual files for each sample containing all sequencing reads. The second step is quality control for which the FASTQC-software is often used (Andrews, 2010). If systematic errors in sequencing have occurred (for example, low quality sequencing at the end of many reads), these parts need to be trimmed (or removed). Next, the sequences per sample are aligned to a reference genome. Alignment refers to the process of determining where each RNA-seq read is located on the genome. This is a complicated process, especially for exon-spanning sequence reads. Alignment procedures for RNA-seq data were developed and Tophat (Trapnell et al., 2010) and GSNAP (Wu and Nacu, 2010) are among the most commonly used. In RNA-seq experiments, typically millions of reads have to be aligned to the genome and the percentage of alignment of unique reads to all generated reads is often used as a measure for the efficiency of the sequencing.

There are many procedures to quantify aligned RNA-seq data and there is no consensus yet about the optimal procedure. Many procedures start by counting the number of fragments (or reads) per gene and standardize to the whole number of aligned reads (counts per million, CPM). The disadvantage of CPM is that gene length is not corrected for, and as a consequence, longer genes are on average more likely to have more reads aligned. To correct for this, a commonly used procedure is the fragments per kilobase of exon per million fragments mapped reads (FPKM)-metric. In FPKM, the sum of the reads that are aligned to a specific gene is calculated, and this number is subsequently normalized for the length of the gene and the total number of reads of that particular sample. The FPKM metric offers an indication about how each gene is expressed in relation to other genes. Readers interested in RNA-seq analysis using a combined Tophat and Cufflinks pipeline from the Linux terminal are advised to read the instruction article (Trapnell et al., 2012). Besides, these freeware programs, there are also commercial sequencing analysis tools that perform quality control, alignment, and data analysis, such as CLC Genomics Workbench (www.clcbio.com) and NextGene® (www.softgenetics.com).

RNA-Seq Differences Between Platforms

Several manufacturers offer equipment for RNA sequencing, including Illumina, 454 Life Sciences, and Helicos. Illumina sequencing represents the most commonly used RNA-seq approach. Most quality check, data alignment and analysis

TABLE 1: Overview of Databases and Tools that can be used for Transcriptome Analysis

Tools	Full-name	Category	Description	Reference	URL
EdgeR	Differential expression analysis of digital gene expression data	Transcriptome analysis	RNA-seq differential gene expression analysis - R package	Robinson, 2010	http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf
LIMMA	Linear Models for Microarray and RNA-Seq Data	Transcriptome analysis	Microarray and RNA-seq differential gene expression analysis - R package	Smyth. 2005	http://www.bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf
Cufflinks	Transcriptome assembly and differential expression analysis for RNA-Seq.	Transcriptome analysis	RNA-seq transcriptome assembly, quantitative and differential gene expression analysis - Linux Terminal software	Trapnell et al., 2010	http://cole-trapnell-lab.github.io/cufflinks/
DeSeq	Differential gene expression analysis based on the negative binomial distribution	Transcriptome analysis	RNA-seq differential gene expression analysis - R package	Anders and Huber, 2010	http://bioconductor.org/packages/release/bioc/vignettes/DESeq/inst/doc/DESeq.pdf
IMMGEN	Immunological Genome project	Expression database	Consortium generating high quality expression data from a large number of immune cells. Contains many bioinformatic tools.	Kim and Lanier, 2013	http://www.immgen.org/
HBT	Human Brain Transcriptome	Expression database	Database to access human brain spatial and temporal gene expression profiles	Kang et al., 2011	http://hbatlas.org/
Allen Brain Atlas	Allen Brain Atlas	Expression database	Database that contains maps of the mouse, macaque and human brains	Sunkin et al., 2013	http://www.brain-map.org/

TABLE 1: Continued

Tools	Full-name	Category	Description	Reference	URL
Expression Atlas	European Bioinformatic Institute: Gene Expression Atlas	Expression database	Repository database that contains expression data from an enormous amount of datasets. Makes the data easily accessible for analysis	Kolesnikov et al., 2014	http://www.ebi.ac.uk/gxa/home
CNS cell-type expression- Tool	No official name.	Expression database	Database for brain cell type splice variants and gene expression	Zhang et al., 2014	http://web.stanford.edu/group/barres_lab/brain_rnaseq.html
DAVID	Database for Annotation, Visualization and Integrated Discovery	Functional annotation	GO and KEGG - pathway enrichment analysis.	Huang, 2009	http://david.abcc.ncifcrf.gov/
Webgestalt	Web-based gene set analysis tools	Functional annotation	GO, KEGG-pathway, PPI, and TFBS - enrichment analysis	Wang et al, 2013	http://bioifo.vanderbilt.edu/webgestalt/
InnateDB	A knowledge resource for innate immunity interactions & pathways	Functional annotation	GO, KEGG-pathway, PPI, and TFBS - enrichment analysis, particularly focused on innate immune signaling	Breuer et al, 2013	http://www.innatedb.com/
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	Functional annotation	Known and predicted protein-protein interaction – tool	Jensen et al., 2009	http://string-db.org/

pipelines have been developed primarily for Illumina (McGettigan, 2013). In contrast to most other sequencing approaches, Helicos DGE directly sequences RNA. It does not depend on conversion of RNA to cDNA and subsequent PCR amplifications (Raz et al., 2011). The manufacturer claims that the initial broad PCR amplification, used by other manufacturers, results in a bias in favor of long transcripts and that their direct RNA-sequencing procedure is more reliable and accurate (Sam et al., 2011). This approach, which is also referred to as direct RNA sequencing, was used in the recent characterization of the microglia sensome by Hickman et al. (2013). In 454 Life Sciences RNA-seq, the length of the sequenced fragment is much longer than with other platforms (in the range of 1,000 bp), which is well suited for transcriptome analysis of a model organisms for which a high quality reference genome is not (yet) available (Hook et al., 2014). Readers that are interested to learn more about transcriptomics using RNA-seq are advised to read a current review by McGettigan (2013).

Differential Gene Expression Analysis

The aim of most transcriptome studies is to identify genes that are differentially expressed after a treatment or between conditions. The output of this type of analysis are gene lists with a fold-change and *p*-values per gene. Microarrays are primarily equipped for differential gene expression analysis, but RNA-seq is generally more sensitive in finding differentially expressed genes (Wang et al., 2014; Zhao et al., 2014). In addition, RNA-seq estimated fold changes more closely resemble quantitative RT-PCR data (Wang et al., 2014). For microarrays, differential gene expression analysis is most commonly performed using Limma. While for RNA-seq, many different procedures are used such as: EdgeR (Robinson et al., 2010), DeSeq (Anders and Huber, 2010), and Cuffdiff (Trapnell et al., 2010). These transcriptome analysis tools are listed in Table 1.

Batch Effects, Design, and Number of Replicates

In order to generate high quality expression data, several factors need to be taken into account including batch effects, design and the number of replicates (see reviews from Auer and Doerge, 2010 and Leek et al., 2010 for more information). Batch effects are an underestimated source of variation, possibly resulting in erroneous findings. Therefore, the requirement of a good design before performing the experiment should not be underestimated. The only way to minimize batch effects is by standardizing and randomizing all procedures from the beginning (cell isolation) until the end (library formation and sample to lane assignments). Ideally, all samples should be isolated simultaneously and in randomized order. Practically, this is often not feasible, therefore it is important to prevent systematic biases, such as isolating treatment and control samples on different occasions. To minimize batch effects, one should randomize, preferably several times, throughout the sample collection procedures and during preparations. Moreover, as discussed before, in RNA-seq it is possible to bar-code individual samples, such that several samples can be sequenced in the same lane avoiding batch effects related to sequencing (Auer and Doerge, 2010).

Another important issue concerns the number of replicates per group. There are several factors that should be kept in mind. First, technical replicates (hybridizing or sequencing the same RNA-sample multiple times) have limited value and independent samples should be used. Second, the golden rule is the more the better. From a statistical point of view it is important to have a reliable measure for biological variation between samples. It is important to note that pooling several samples artificially decreases the variation between samples and should preferably be avoided. The third factor that should be taken into account is heterogeneity in genetic and environmental background. For example, a study of human post mortem brain tissues that originated from heterogeneous cell types (namely glia and neurons) from individuals of different ages, with different ethnical backgrounds and varying medical histories, require many more samples than acutely isolated pure inbred mouse microglia. The last factor that should be taken into account is the strength and consistency of the effect. It is often difficult to estimate the effect size, but it can be helpful to run a quantitative RT-PCR for a few well-known response genes on part of the sample, before generating costly, genome wide transcriptome data.

Construction and Content of GOAD

Programming Language and Structural Setup

The GOAD web application was developed with the Java programming language and its web technologies (version 1.7, see www.java.com) and is hosted on the Tomcat web application container (version7.0.47 see http://tomcat.apache.org/). jQuery version 10.1.2 and the JQuery plug-ins DataTables (version 1.10.0), jvenn (version 1.5), validator (version 1.11.1), jspdf (version 0.9.0rc2), and jQuery-ui (version 1.10.4) were used. The data have been stored using MySQL database management system version 5.5.37.

Criteria for Inclusion of Studies

In GOAD, expression datasets were included based on the following criteria. First, the dataset should be generated with pure populations of murine glia cells, rapidly *ex vivo* isolated, for example, by fluorescence-activated cell sorting (FACS) isolation or laser micro-dissection. The second criterion relates to the transcriptome analysis technique; only genome wide gene expression analyses were included, such as RNA-sequencing and microarrays. The third criterion for the DE

analysis is that at least three biological replicates per condition should be available. The last criterion is the availability of annotation files for the platform used. Using these criteria, we searched on GEO, ArrayExpress, and Pubmed with keywords "microglia," "astrocyte," "oligodendrocyte," and "transcriptome," and contacted individual researchers to obtain unprocessed expression data.

Generation of Recent Microglia Transcriptome Datasets

Two unpublished microglia datasets were included in the GOAD-database. A gene expression profile of mouse microglia that were isolated from different brain regions and a gene expression dataset of microglia that were treated with lipopolysaccharide (LPS). For both datasets the isolations and animal experiments were done in accordance with Dutch law and European animal regulations and were approved by the University of Groningen animal welfare committee.

Gene Expression Profile of Microglia from Different Brain Regions. Microglia were isolated from the hippocampus, cortical grey matter, corpus callosum white matter, and cerebellum using *ex vivo* isolation procedures as described previously and FACS sorting with CD11b and CD45 antibodies (Olah et al., 2012). Eight gene expression replicates per brain region were obtained and each replicate consists of a pool of RNA from three mice. RNA was converted to cDNA and hybridized to Illumina MouseRef8 microarrays. Gene expression values were obtained using Illumina Genome Studio.

*Gene Expression Profile of Microglia after LPS Treatment.*Young adult mice (2–4) months were injected with either PBS or LPS (0.25 mg/kg) and microglia were isolated 3 hr postinjection. Three biological replicates per group were obtained. RNA was converted to cDNA, prepped using Illumina TruSeq and 100 bp paired-end sequenced using a Illumina laneHiSeq2500. Sequencing depth was in the range of 11.6 to 23.5 million aligned high quality reads per sample.

Preprocessing of Transcriptome Datasets in GOAD

Preprocessing of the Microarray Datasets. Raw microarray expression values were preprocessed using R and the Bioconductor package Limma (Smyth, 2005). As a quality control, samples with an average inter-sample correlation three standard deviations below the mean intersample correlation after normalization were filtered out and this procedure was repeated until all samples in the study met the inclusion criteria. Quantile normalization was applied to the Illumina microarrays. For Agilent array preprocessing, background correction was performed with an offset of 50 followed by Lowess within-array normalization and Quantile between-array normalization. Relative intensities were converted into expression values. The Affymetrix microarrays were preprocessed using the Expresso-function of R package Affy (Gautier et al., 2004). The parameters were set to RMA background correction and quantile normalization, with pm correct pmonly and a medianpolish. Datasets from different platforms were made comparable at the level of gene symbols. The WGCNA collapseRows function was applied to calculate the representative gene expression for several probes, by picking the highest expressed probe, associated with a single gene (Miller et al., 2011).

Preprocessing of the RNA-Seq Datasets. Fastq reads values were quality checked and trimmed using FASTQC (Andrews, 2010) and aligned using Tophat (Trapnell et al., 2010), with Illumina Igenome build UCSC mm10 (http://support.illumina.com/sequencing/sequencing_software/igenome.html).

Analysis Tools: DE and QE Analysis

DE analysis was performed for microarray data using Limma (Smyth, 2005) and for RNA-seq data using EdgeR (Robinson et al., 2010). QE analysis was done using FPKM values that were calculated by Cufflinks and a 95% confidence interval was used to determine whether a gene was reliably expressed or not. Genes that were reliably expressed (95% lower confidence interval >0) were subsequently divided according to the percentile of expression and assigned an arbitrary categorization ranging from "very high expression" to "very low expression" (Fig. 4).

Applications of the GOAD Database

Glial Open Access Database

The Glia Open Access Database (GOAD; www.goad.education) (Fig. 1) contains three features: differential gene expression (DE) analysis, quantitative gene expression (QE) analysis, and a search utility. An online tutorial is provided that gives additional information about the application and output of each feature.

Differential Expression (DE) Analysis

Currently, 16 studies are available in GOAD and 37 comparisons can be generated (suppl. table 1). Studies included are (Beckervordersandforth et al., 2010; Beutner et al., 2013; Cahoy et al., 2008; Chiu et al, 2013; Doyle et al., 2008; Gautier et al., 2012; Hickman et al., 2013; Lovatt et al., 2007; Olah et al., 2012; Orre et al., 2014a, 2014b; Parakalan et al., 2012; Raj et al., 2014; Szulzewsky et al., 2015). In many instances, glia cells were compared with non-glia cells,

whole brain tissues or a FACS-negative (gated against selected markers) population. The non-glia cell types that are included in the database are neurons, neural stem cells, macrophages, and dendritic cells. The included datasets contain cell type- and disease-associated expression data and can roughly be divided into cell-type specific profiles related to (1) normal function or (2) disease state or neurodegenerative condition.

In GOAD, individual DE gene lists can be retrieved as sortable (by gene name, *p*-value, or fold-change) spread sheets (Fig. 2). It is possible to perform a DE analysis for several comparisons simultaneously and an interactive Venn diagram is generated that shows the overlap and differences between the data sets (Fig. 3). Moreover, the Venn diagram can be downloaded as a png file.

DE analysis is the most commonly used type of transcriptome analysis and is based on the difference between two or more samples, but the effect of control samples is often neglected in later interpretations. For example, in some studies the differences in gene expression profiles between CNS cell-types in healthy tissues were studied (Cahoy et al., 2007; Doyle et al., 2008; Zhang et al., 2014). These studies included astrocytes, oligodendrocytes and neurons, but among several other aspects, they differed in the choice of control samples used for their analyses. Cahoy et al. (2007) compared each cell-type to the other isolated cell types, while Doyle et al. compared each cell type to the FACS negative population. Zhang et al. (2014) contained a few cell populations that were not isolated in the other studies such as newly formed oligodendrocytes and pericytes. Both studies can result in astrocyte-specific gene expression profiles that are overlapping, but they do contain substantial differences. Intersecting such profiles, for example, using the Venn diagram will generate a gene list containing markers that are more reliably astrocyte-specific.

Recent transcriptome studies have focused on characterization of changes in gene expression related to disease and neuropathology (Chiu et al., 2013; Hickman et al., 2014; Olah et al., 2012). The first studies reporting on

FIGURE 1: Screenshot of the home page of the GOAD website. The GOAD website home page (www.goad.education) displaying the three primary features of the database: Differential gene Expression analysis, Quantitative gene Expression analysis and the Search utility.

ACCELERATED AGING (ERCC1) MICROGLIA vs. CONTROL MICROGLIA

Priming of Microglia in a DNA-Repair Deficient Model of Accelerated Aging

Raj et al (2014)

Show 10 entries Search:

Gene symbol	Fold Change	Log fold change	Adjusted p-value	Description
IFITM3	9.94892411877446	3.31454052038676	9.642E-19	
SPP1	9.92457832505251	3.31100580640883	2.798E-16	
OASL2	8.90090625074101	3.15395223235865	3.441E-18	2'-5' oligoadenylate synthetase-like 2
CCL5	8.46947802752917	3.08227305925054	5.587E-21	
IFI202B	7.55388778268619	2.91721935187914	5.126E-20	Interferon activated gene 202B
CLEC7A	6.83293240637815	2.77250485526779	3.949E-14	C-type lectin domain family 7, member a
AXL	6.74005540899495	2.75276045163765	7.664E-19	
APOE	6.66615750829118	2.73685540591891	3.056E-14	Transcribed locus, moderately similar to XP_001104482.1 apolipoprotein E [Macaca mulatta]
RSAD2	6.61000227460977	2.72465076818798	7.664E-19	Transcribed locus
IFIT2	5.8865802780554	2.55742976501177	2.486E-15	Interferon-induced protein with tetratricopeptide repeats 2

Showing 1 to 10 of 529 entries (filtered from 14,740 total entries) Previous 1 2 3 4 5 .. 53 Next

FIGURE 2: Screenshot of the output spreadsheet of the DE feature of GOAD. Individual DE gene lists can be extracted from GOAD as sortable (by *P* value, fold-change, or gene name) data in a spreadsheet containing the gene symbols, the fold changes in expression, the statistical significance and a brief description of gene functions.

gene expression profiles of acutely isolated pure samples of astrocytes related to aging and neurodegenerative disease have recently been published (Orre et al., 2014a,b). By combining these datasets, genes that are up or down-regulated in disease models and aging can be identified. These types of analyses cannot be performed with the individual published studies and illustrate the value of the GOAD-tool.

Quantitative Expression (QE) Analysis

Currently, four studies are available for QE analysis and 10 pure cell type expression profiles can be generated (Suppl. table 1).

Differential Gene Expression

Cut-off values:
Adjusted p-value: α < 0.1
Log fold change: up regulated

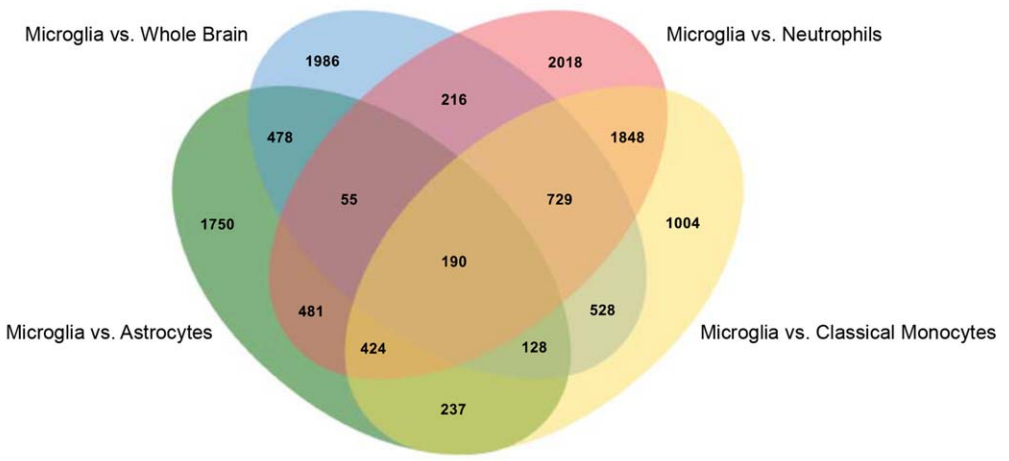


FIGURE 3: Screenshot of the output Venn diagram of the DE feature of GOAD. A Venn diagram generated in GOAD, depicting the overlap and differences between four DE gene lists.

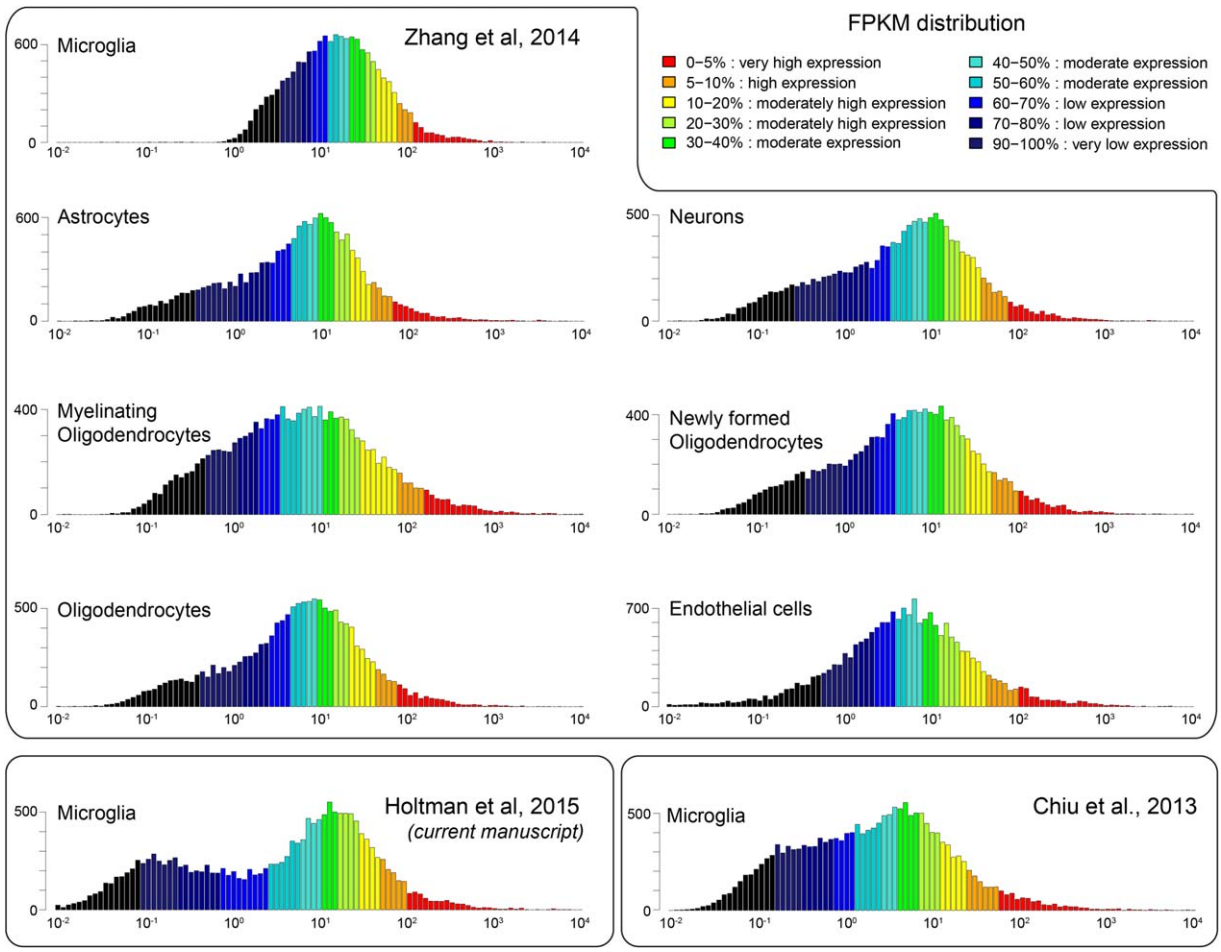


FIGURE 4: Distribution of genes across FPKM values for datasets present in the QE feature. FPKM values for the expressed genes are approximately normally distributed across a logarithmic scale. Colors correspond to percentiles of expression as indicated.

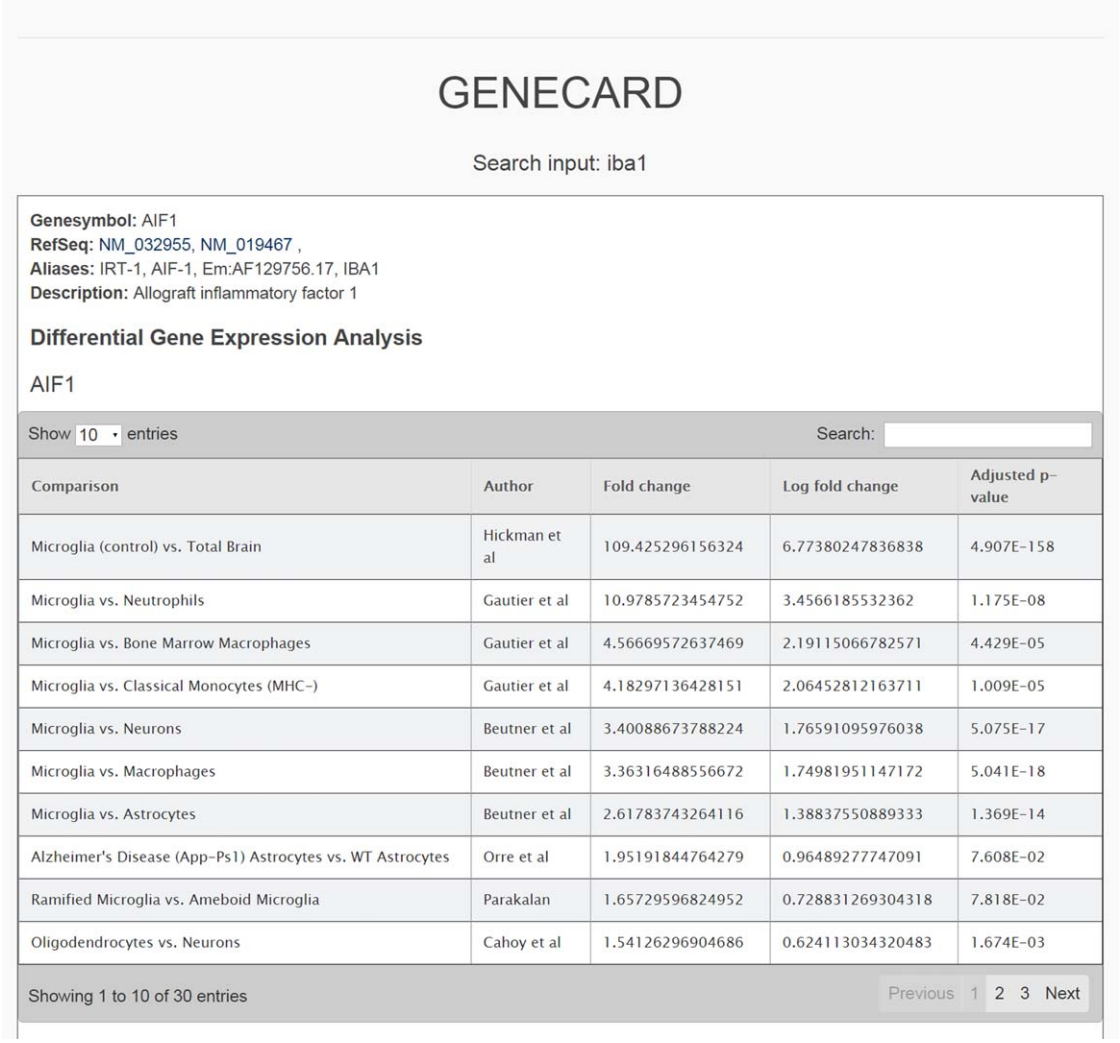
The first dataset was generated using microglia isolated from mouse spinal cord (Chiu et al., 2013). The second dataset was generated using whole brain mouse microglia (see Gene expression profile of microglia after LPS treatment section for more information). The third dataset (Zhang et al., 2014) was obtained from microglia, astrocytes, myelinating oligodendrocytes, oligodendrocytes precursor cells, newly formed oligodendrocytes, and pericytes from the cortex. The fourth dataset was generated from different tissue macrophages and other innate immune cells such as Kuppfer cells, spleen macrophages, small and large intestine macrophages, monocytes, neutrophils, and microglia (Lavin et al., 2014). For each of these datasets, the FPKM values were generated and expressed genes were subsequently subdivided according to the percentile of expression (Fig. 4). The FPKM values are approximately normally distributed across a logarithmic scale. To facilitate interpretation of the percentiles, an arbitrary categorization system is provided, with for example values from zero to the fifth percentile being considered as very highly expressed.

Search Utility and GeneCard

The third feature of the database is a Search Utility. Genes of interest can be searched in GOAD using gene symbols and accession numbers, resulting in a GeneCard. The GeneCard contains information about the gene and will show the results of the gene in the DE and QE analyses. The Search Utility is capable of detecting both the official gene symbols as well as gene synonyms. For example, the official name of microglia marker Iba1 is Aif1. The Search Utility is able to find Aif1 when searching for Iba1 (Fig. 5). This gene card can be downloaded as a pdf file.

GOAD in Relation to Other Databases and Additional Tools

Several other transcriptome database tools, similar to GOAD, have been developed over the last years including Immunological Genome Project (IMMGEN), Human Brain Transcriptome (HBT), Allen Brain Atlas (Kang et al., 2011; Kim and Lanier, 2013; Sunkin et al., 2013). GOAD is unique in



GENECARD

Search input: iba1

Genesymbol: AIF1
RefSeq: NM_032955, NM_019467 ,
Aliases: IRT-1, AIF-1, Em:AF129756.17, IBA1
Description: Allograft inflammatory factor 1

Differential Gene Expression Analysis

AIF1

Show 10 entries Search:

Comparison	Author	Fold change	Log fold change	Adjusted p-value
Microglia (control) vs. Total Brain	Hickman et al	109.425296156324	6.77380247836838	4.907E-158
Microglia vs. Neutrophils	Gautier et al	10.9785723454752	3.4566185532362	1.175E-08
Microglia vs. Bone Marrow Macrophages	Gautier et al	4.56669572637469	2.19115066782571	4.429E-05
Microglia vs. Classical Monocytes (MHC-)	Gautier et al	4.18297136428151	2.06452812163711	1.009E-05
Microglia vs. Neurons	Beutner et al	3.40088673788224	1.76591095976038	5.075E-17
Microglia vs. Macrophages	Beutner et al	3.36316488556672	1.74981951147172	5.041E-18
Microglia vs. Astrocytes	Beutner et al	2.61783743264116	1.38837550889333	1.369E-14
Alzheimer's Disease (App-Ps1) Astrocytes vs. WT Astrocytes	Orre et al	1.95191844764279	0.96489277747091	7.608E-02
Ramified Microglia vs. Ameboid Microglia	Parakalan	1.65729596824952	0.728831269304318	7.818E-02
Oligodendrocytes vs. Neurons	Cahoy et al	1.54126296904686	0.624113034320483	1.674E-03

Showing 1 to 10 of 30 entries Previous 1 2 3 Next

FIGURE 5: Screenshot of the GeneCard of the Search Utility of GOAD. The Search Utility of GOAD depicting the GeneCard of Aif1 (Iba1) containing DE expression data from selected studies.

the sense that it collects all current, publicly available glia gene expression datasets from different studies and backgrounds. Some databases provide comprehensive information on all publicly available transcriptome data such as the EMBL-EBI expression atlas (Kapushesky, 2010). This database provides differential gene expression analysis for all publicly available datasets. This can be helpful to consult if one is interested in the comprehensive information about a particular gene. Due to the enormous amount of data it is difficult to find datasets that provide information on specific cell types and to compare such datasets to each other.

Human Brain Transcriptome (HBT) is a public database containing transcriptome data from 16 different regions of the developing and adult human brain including associated genotyping data per sample. With HBT, it is possible to study gene expression profiles across developmental time points and across regions for individual genes of interest. Moreover, it is possible to study spatiotemporal expression profiles of 90 different neurobiological processes, such as cell-type, neurodevelopment, and neurotransmission related categories. It depicts the first principal component of the genes clustered to such a biological category across regions and developmental phase.

The IMMGEN Project is a consortium aimed at generating gene expression profiles of innate and adaptive immune cells through different states of development and maturation, activation responses, effector stages, tissue localization, age, and genetic variation of mouse immune cells. This is done in a highly standardized way, resulting in high quality data. The IMMGEN data are accessible using one of the nine different data browsers such as "Gene Skyline," to study the expression profile of individual genes across cell types, which is similar to the QE analysis, and the "Population Comparison" to compare different cell populations with each other directly.

Other interesting features of IMMGEN are: "Differential Splicing" searches, to find unique splice variants for a particular cell type or class and "Modules and Regulators" to find co-expressed genes and predicted transcription factors that could regulate the co-expression modules. This consortium has generated major advances in the gene regulatory mechanisms of different classes of immune cells.

A recent publication contains RNA-seq data of the most common cell types of the CNS (Zhang et al., 2014). In this article, an online database tool was presented to allow easy access to the data. This tool contains several functions such as "gene search," "interactive splicing browser," "cell type enrichment," and "cell type specific splicing."

The Allen Brain Atlas from the Allen Institute for Brain Science contains transcriptome maps from the (developing) mouse, rhesus macaque and human brain, with high regional specificity. Using these atlases, it is possible to study expression profiles across brain regions over time. Overall there are three main features: "Gene Search," "Differential Search," and "Gene Classification." The Gene Search shows the expression of a gene of interest across time, individuals and regions using a heatmap. The Differential Search can be used to compare different regions to each other, to find genes differentially expressed between brain regions. The Gene Classification shows the gene expression profiles of biologically relevant categories across the dataset. The gene expression data are not only depicted as a heatmap but are also integrated with the three-dimensional structure of the brain, resulting in depiction of the gene expression profiles across brain regions. The tools and the generated data have been used in many studies and resulted in great insight in brain development. These databases and a list of frequently used and useful annotation tools for transcriptome data are depicted in Table 1.

Conclusions and Future Perspectives

GOAD is an online tool that is generated with the aim to facilitate access and analysis of genome-wide expression profiles from glia transcriptome datasets. With a continuing decrease in sequencing costs, a rapid increase in the number of transcriptome datasets is to be expected. This unprecedented amount of data will give new insight in the role of glia cells in health and disease. GOAD aims to provide the glia community easy access to these data, but is dependent on input from the scientific glia community.

Having established the first release of GOAD, several aims are envisaged for the short- and long-term. First, transcriptome data will be added to GOAD on a regular basis to remain up to date. Second, the search utility will be expanded to graphically depict the QE expression values. Third, the GOAD-database currently contains transcriptome data from pure murine glia samples. Currently, glia gene expression profiles of other organisms including human, macaque, zebrafish, and fruit fly are collected. Fourth, GOAD only contains transcriptome data. It is expected that in the near future genome-wide epigenetic data (histone modifications and DNA methylation profiles) as well as miRNA expression profiles will be generated from glia cells. An aim of GOAD is to incorporate such datasets in future updates of the website.

Acknowledgment

The authors gratefully acknowledge Prof. Jon Laman for reading and providing constructive feedback on this article.

References

Andrews S. FastQC a quality control tool for high throughput sequence data.Available at http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/.

Auer PL, Doerge RW. 2010. Statistical design and analysis of RNA sequencing data. Genetics 185:405–416.

Beckervordersandforth R, Tripathi P, Ninkovic J, Bayam E, Lepier A, Stempfhuber B, Kirchhoff F, Hirrlinger J, Haslinger A, Lie DC, Beckers J, Yoder B, Irmler M, Götz M. 2010. In vivo fate mapping and expression analysis reveals molecular hallmarks of prospectively isolated adult neural stem cells. Cell Stem Cell 7:744–758.

Beutner C, Linnartz-Gerlach B, Schmidt SV, Beyer M, Mallmann MR, Staratschek-Jox A, Schultze JL, Neumann H. 2013. Unique transcriptome signature of mouse microglia. Glia 61: 1429–1442.

Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, Winsor GL, Hancock REW, Brinkman FSL, Lynn DJ. 2013. InnateDB: systems biology of innate immunity and beyond–Recent updates and continuing curation. Nucleic Acids Res 41:D1228–D1233.

Butovsky O, Jedrychowski MP, Moore CS, Cialic R, Lanser AJ, Gabriely G, Koeglsperger T, Dake B, Wu PM, Doykan CE, Fanek Z, Liu L, Chen Z, Rothstein JD, Ransohoff RM, Gygi SP, Antel JP, Weiner HL. 2014. Identification of a unique TGF-β-dependent molecular and functional signature in microglia. Nat Neurosci 17: 131–143.

Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, Xing Y, Lubischer JL, Krieg PA, Krupenko SA, Thompson WJ, Barres BA. 2008. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J Neurosci 28:264–278.

Chiu IM, Morimoto ET, Goodarzi H, Liao JT, O'Keeffe S, Phatnani HP, Muratet M, Carroll MC, Levy S, Tavazoie S, Myers RM, Maniatis T. 2013. A neurodegeneration-specific gene-expression signature of acutely isolated microglia from an amyotrophic lateral sclerosis mouse model. Cell reports 4: 385–401.

Doyle JP, Dougherty JD, Heiman M, Schmidt EF, Stevens TR, Ma G, Bupp S, Shrestha P, Shah RD, Doughty ML, Gong S, Greengard P, Heintz N. 2008. Application of a translational profiling approach for the comparative analysis of CNS cell types. Cell 135:749–762.

Gautier EL, Shay T, Miller J, Greter M, Jakubzick C, Ivanov S, Helft J, Chow A, Elpek KG, Gordonov S, Mazloom AR, Ma'ayan A, Chua WJ, Hansen TH, Turley SJ, Merad M, Randolph GJ. 2012. Immunological Genome Consortium. Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. Nat immunol 13:1118–1128.

Gautier L, Cope L, Bolstad BM, Irizarry RA. 2004. affy–Analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20:307–315.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57.

Hook SE, Twine NA, Simpson SL, Spadaro DA, Moncuquet P, Wilkins MR. 2014. 454 pyrosequencing-based analysis of gene expression profiles in the amphipod *Melita plumulosa*: Transcriptome assembly and toxicant induced changes. Aquat Toxicol 153:73–88.

Hickman SE, Kingery ND, Ohsumi TK, Borowsky ML, Wang LC, Means TK, El Khoury J. 2013. The microglial sensome revealed by direct RNA sequencing. Nat Neurosci 16:1896–1905.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. 2009. STRING 8–A global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37:D412–D416.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AMM, Pletikos M, Meyer KA, Sedmak G, Guennel T, Shin Y, Johnson MB, Krsnik Z, Mayer S, Fertuzinhos S, Umlauf S, Lisgo SN, Vortmeyer A, Weinberger DR, Mane S, Hyde TM, Huttner A, Reimers M, Kleinman JE, Sestan N. 2011. Spatio-temporal transcriptome of the human brain. Nature 478:483–489.

Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A. 2010. Gene expression atlas at the European bioinformatics institute. Nucleic Acids Res 38:D690–D698.

Kim CC, Lanier LL. 2013. Beyond the transcriptome: Completion of act one of the Immunological Genome Project. Curr Opin Immunol 25:593–597.

Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A. 2014. ArrayExpress update-simplifying data submissions. Nucleic Acids Res 28:43:D1113–D1116.

Lavin Y, Winter D, Blecher-Gonen R, David E, Keren-Shaul H, Merad M, Jung S, Amit I. 2014. Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. Cell 159:1312–1326.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11:733–739.

Lovatt D, Sonnewald U, Waagepetersen HS, Schousboe A, He W, Lin JH, Han X, Takano T, Wang S, Sim FJ, Goldman SA, Nedergaard M. 2007. The transcriptome and metabolic gene signature of protoplasmic astrocytes in the adult murine cortex. J Neurosci 27:12255–12266.

McGettigan PA. 2013. Transcriptomics in the RNA-seq era. Curr Opin Chem Biol 17:4–11.

Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, Horvath S. 2011. Strategies for aggregating gene expression data: The collapseRows R function. BMC Bioinformatics 12:322.

Olah M, Amor S, Brouwer N, Vinet J, Eggen B, Biber K, Boddeke HW. 2012. Identification of a microglia phenotype supportive of remyelination. Glia 60: 306–321.

Orre M, Kamphuis W, Osborn LM, Jansen AH, Kooijman L, Bossers K, Hol EM. 2014a. Isolation of glia from Alzheimer's mice reveals inflammation and dysfunction. Neurobiol Aging 35:2746–2760.

Orre M, Kamphuis W, Osborn LM, Melief J, Kooijman L, Huitinga I, Klooster J, Bossers K, Hol EM. 2014b. Acute isolation and transcriptome characterization of cortical astrocytes and microglia from young and aged mice. Neurobiol Aging 35:1–14.

Parakalan R, Jiang B, Nimmi B, Janani M, Jayapal M, Lu J, Tay SS, Ling EA, Dheen ST. 2012. Transcriptome analysis of amoeboid and ramified microglia isolated from the corpus callosum of rat brain. BMC Neurosci 14:64.

Raj DD, Jaarsma D, Holtman IR, Olah M, Ferreira FM, Schaafsma W, Brouwer N, Meijer MM, de Waard MC, van der Pluijm I, Brandt R, Kreft KL, Laman JD, de Haan G, Biber KP, Hoeijmakers JH, Eggen BJ, Boddeke HW. 2014. Priming of microglia in a DNA-repair deficient model of accelerated aging. Neurobiol Aging 35:2147–2160.

Raz T, Causey M, Jones DR, Kieu A, Letovsky S, Lipson D, Thayer E, Thompson JF, Milos PM. 2011. RNA sequencing and quantitation using the Helicos Genetic Analysis System. Methods Mol Biol 733:37–49.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.

Sam LT, Lipson D, Raz T, Cao X, Thompson J, Milos PM, Robinson D, Chinnaiyan AM, Kumar-Sinha C, Maher CA. 2011. A comparison of single molecule and amplification based sequencing of cancer transcriptomes. PLoS One 6: e17305. Available from: http://dx.plos.org/10.1371/journal.pone.0017305.

Smyth GK. 2005. Limma: Linear models for microarray data. Bioinformatics: 397–420.

Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, Thompson CL, Hawrylycz M, Dang C. 2013 Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res 41:D996–D1008.

Szulzewsky F, Pelz A, Feng X, Synowitz M, Markovic D, Langmann T, Holtman IR, Wang X, Eggen BJL, Boddeke HWGM, Hambardzumyan D, Wolf SA, Kettenmann H. Glioma-associated microglia/macrophages display an expression profile different from M1 and M2 polarization and highly express Gpnmb and Spp1. PLoS One 10:e0116644.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578.

Tuck AC, Tollervey D. 2011. RNA in pieces. Trends Genet 27:422–432.

Wang J, Duncan D, Shi Z, Zhang B. 2013. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): Update 2013. Nucleic Acids Res 41:W77–W83.

Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, Fang H, Hong H, Shen J, Su Z, Meehan J, Li X, Yang L, Li H, Labaj PP, Kreil DP, Megherbi D, Gaj S, Caiment F, van Delft J, Kleinjans J, Scherer A, Devanarayan V, Wang J, Yang Y, Qian HR, Lancashire LJ, Bessarabova M, Nikolsky Y, Furlanello C, Chierici M, Albanese D, Jurman G, Riccadonna S, Filosi M, Visintainer R, Zhang KK, Li J, Hsieh JH, Svoboda DL, Fuscoe JC, Deng Y, Shi L, Paules RS, Auerbach SS, Tong W. 2014. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotechnol 32:926–932.

Weichenhan D, Plass C. 2013. The evolving epigenome. Hum Mol Genet 15: R1–R6.

Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873–881.

Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, Phatnani HP, Guarnieri P, Caneda C, Ruderisch N, Deng S, Liddelow SA, Zhang C, Daneman R, Maniatis T, Barres BA, Wu JQ. 2014. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. J Neurosci 34:11929–11947.

Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. 2014. Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells. PLoS One 16:e78644.