Computational and Structural Biotechnology Journal 23 (2024) 2488–2496

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

GeTeSEPdb: A comprehensive database and online tool for the identification and analysis of gene profiles with temporal-specific expression patterns

Ni Kuang [a], Qinfeng Ma [a], Xiao Zheng [a], Xuehang Meng [a], Zhaoyu Zhai [a], Qiang Li [a], Jianbo Pan [a,b,*]

[a] *Basic Medicine Research and Innovation Center for Novel Target and Therapeutic Intervention, Ministry of Education, Institute of Life Sciences, Chongqing Medical University, Chongqing 400016, China*
[b] *Precision Medicine Center, The Second Affiliated Hospital of Chongqing Medical University, Chongqing 400010, China*

ARTICLE INFO

Keywords:
Gene profiles with temporal-specific expression patterns
Database
Time series
Online tool
Regression analysis

ABSTRACT

Gene expression is dynamic and varies at different stages of processes. The identification of gene profiles with temporal-specific expression patterns can provide valuable insights into ongoing biological processes, such as the cell cycle, cell development, circadian rhythms, or responses to external stimuli such as drug treatments or viral infections. However, currently, no database defines, identifies or archives gene profiles with temporal-specific expression patterns. Here, using a high-throughput regression analysis approach, eight linear and nonlinear parametric models were fitted to gene expression profiles from time-series experiments to identify eight types of gene profiles with temporal-specific expression patterns. We curated 2684 time-series transcriptome datasets and identified 2644,370 gene profiles exhibiting temporal-specific expression patterns. The results were stored in the database GeTeSEPdb (gene profiles with temporal-specific expression patterns database, http://www.inbirg.com/GeTeSEPdb/). Moreover, we implemented an online tool to identify gene profiles with temporal-specific expression patterns from user-submitted data. In summary, GeTeSEPdb is a comprehensive web service that can be used to identify and analyse gene profiles with temporal-specific expression patterns. This approach facilitates the exploration of transcriptional changes and temporal patterns of responses. We firmly believe that GeTeSEPdb will become a valuable resource for biologists and bioinformaticians.

1. Introduction

In the field of gene research, the analysis of time-series data plays a crucial role in revealing the dynamics of gene expression. With advancements in research technologies, an increasing amount of time-series data for various biological processes are being meticulously collected and recorded, providing a rare opportunity to gain deeper insights into gene variations during different biological processes. Researchers, aided by microarray chips and RNA-seq technology, can effectively monitor gene expression changes in processes such as the cell cycle, developmental stages, response to drug treatments, and reactions to external stimuli [1,2]. Importantly, time-series gene expression data are being increasingly used for monitoring patient responses in clinical studies, focusing on human responses to injuries and diseases, as well as responses to treatments and preventive measures [3–6]. Patient heterogeneity renders the analysis of absolute expression levels meaningless [5]. Thus, assessing and measuring the dynamics of patient gene expression changes is particularly beneficial. For example, in a study of multiple sclerosis patients receiving recombinant interferon-β treatment, Baranzini et al. analysed gene expression levels before treatment (time point 0) and complete time-series gene expression response data to identify genes associated with treatment outcomes, thus predicting patient responses to treatment [7]. Subsequent studies of the same multiple sclerosis dataset showed that for certain genes, patients with two subtypes of good responses exhibited similar characteristic expression patterns, while those with adverse reactions showed significantly different gene expression patterns. Therefore, using the complete time response of selected genes can yield better classification results for

* Corresponding author at: Basic Medicine Research and Innovation Center for Novel Target and Therapeutic Intervention, Ministry of Education, Institute of Life Sciences, Chongqing Medical University, Chongqing 400016, China.
E-mail address: panjianbo@cqmu.edu.cn (J. Pan).

https://doi.org/10.1016/j.csbj.2024.06.003
Received 15 January 2024; Received in revised form 29 May 2024; Accepted 4 June 2024
Available online 5 June 2024

predicting patient responses to treatment [8,9]. Indeed, in recent years, the number of time-series datasets stored in major public databases has been growing exponentially.

However, the significant increase in sequencing capacity has mainly been used to generate static datasets, making time-series expression data particularly attractive as a complementary data form for understanding dynamic systems. Nevertheless, with ongoing technological advancements, the high dimensionality and complexity of data have presented challenges in data analysis. Therefore, researchers are working diligently to develop new methods and tools to more accurately analyse time series of gene expression data. These methods include approaches for identifying differentially expressed time-series genes, such as SAM [10], BETR [11], and ImpulseDE2 [12]; clustering temporally expressed genes, such as STEM [13], Mfuzz [14], and TimeClust [15]; utilizing hidden Markov models for aligning and classifying time-series gene expression in clinical studies, such as TRAM [8], and reconstructing dynamic regulatory networks, such as DREM [16]. To better understand the temporal and dynamic aspects of gene expression, researchers have developed various databases for time-series genes. Examples include the Temporal Expression Database during Development (TEDD [17]), which provides an overview of temporal-specific gene expression and chromatin accessibility during embryonic, foetal, neonatal, childhood, and adult development. Genes with spatially specific expression patterns, such as tissue-specific genes, have been widely identified and stored in databases such as the Tissue-specific Gene Expression and Regulation (TiGER [18]) database and the Tissue-Specific Gene Database (TiSGed [19]). However, there is currently no database that defines, identifies, or archives gene profiles with temporal-specific expression patterns. In this context, such genes are defined as those exhibiting specific patterns of expression over time.

Here, we introduce GeTeSEPdb, which aims to collect and organize a wide range of time-series gene expression data and provide powerful analysis tools and resources to enable researchers to explore the importance of gene expression dynamics in detail. Given that the temporal aspects of gene expression are often challenging to interpret and exhibit complex characteristics involving irregular time intervals, we employed both parametric and nonparametric regression models [20–22] to address this challenge. Specifically, we used *clust* [23] for gene set clustering and performed regression analyses to identify potential patterns. Ultimately, we identified 2644,370 gene profiles with temporal-specific expression patterns and stored them in the GeTeSEPdb database. Through the interface of GeTeSEPdb, users can further construct gene regulatory networks using gene profiles with similar expression patterns from the database. Users can also submit data for online analysis, which aids in identifying the interactions and regulatory relationships among genes participating in the same biological events within specific environments.

2. Materials and methods

2.1. Data collection and processing

We conducted comprehensive searches of the Gene Expression Omnibus (GEO) [24] and ArrayExpress [25] databases using the keywords “temporal,” “time series,” and “time course” to retrieve transcriptomic datasets across various species. Among these datasets, we selected the 13 species with the greatest number of datasets, which included *Homo sapiens, Mus musculus*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Danio rerio*, *Escherichia coli*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Bos taurus*, *Oryza sativa*, *Sus scrofa*, and *Macaca mulatta*. Subsequently, we manually reviewed the search results to confirm whether they contained data for at least 5 time points. We also collected relevant information, such as experimental factors that influence temporal changes, disease information, gene editing information, and drug-related data. In cases in which time-series samples were influenced by the same factor in a study, we organized them into a single dataset. For each organized dataset, different time points will be sorted in ascending order and then converted into a numerical pseudo-time array from 1, 2,. N (N is the number of time points). The converted numerical data points will be labeled on the samples in the dataset for further modeling.

We further downloaded the majority of the raw-format data, including expression profiles annotated with probe IDs and raw RNA sequencing data. For microarray data, we used the R package GEOquery [26] to download expression profiles from GEO. Notably, the microarray data downloaded using GEOquery were preprocessed. For ArrayExpress, we obtained processed expression profiles via FTP links. For RNA sequencing data, FASTQ-compressed files were primarily obtained from the European Nucleotide Archive [27] and the DNA Data Bank of Japan [28]. We employed a customized workflow tailored to different experimental research libraries, as described below:

i. *Microarray data*: To facilitate downstream analysis, probe identifiers for each transcriptome profile were converted to gene symbols. Multiple probes matching the same gene symbol were merged using average expression values. Subsequently, log2 transformation was applied to the nontransformed gene expression profiles.
ii. *RNA-seq data*: Reference genomes and annotation information for the 13 organisms were acquired from the RefSeq database [29]. The following md5 checksums were computed for the FASTQ raw data and verified to ensure data accuracy during transmission. After undergoing quality control with fastp [30] (v0.23.1), TPM values were obtained using Salmon [31] quantification.

A diagram of the process of data acquisition and preprocessing is shown in Fig. 1.

2.2. Data analysis pipeline

We first employed the *clust* software to analyse the expression profiles of each dataset, aiming to identify gene clusters exhibiting high correlations among expression patterns in a given time span (supplementary additional information). In addition to identifying co-expressed gene clusters, *clust* performs a dimensionality reduction process on the identified co-expressed gene clusters, similar to the singular value decomposition of a matrix in linear algebra, to retain its important features. Therefore, *clust* also outputs a list of cluster eigengene [32] as a representative of the gene expression profiles for each cluster. To be noted, cluster eigengene is not one gene within the cluster chosen to represent the entire cluster, but instead a newly created representative. Through clustering analysis, gene profiles within the identified clusters are supposed to be highly correlated and informative. And eigengene could represent the potential variation trend of the cluster. Since the genes within each cluster are highly correlated, using the cluster eigengene for fitting can serve as a standard to help determine the gene expression pattern of the cluster. Furthermore, by fitting the expression profiles of each gene in the cluster based on this pattern, we can more accurately identify a group of genes that perform the same function in a particular biological process.

Following the identification of gene clusters, the cluster eigengene from each identified cluster were fitted to two linear models—a linear function and a constant function—as well as six nonlinear regression models, which included logistic regression, logarithmic, exponential, power, trigonometric, and quadratic functions. The constant function was firstly evaluated based on its coefficient of variation (CV), defined as the ratio of the standard deviation to the mean, with the threshold to signify a constant trend in gene expression set at less than 0.1. Those genes that fail to meet the criteria of constant function will be further fitted to other seven types of models. The linear function and trigonometric function were fitted using the lm function and the nlsLM function from the R package “nlme” (v3.1–163), respectively. The remaining five functions were fitted using the R package “aomisc” (v0.650), which

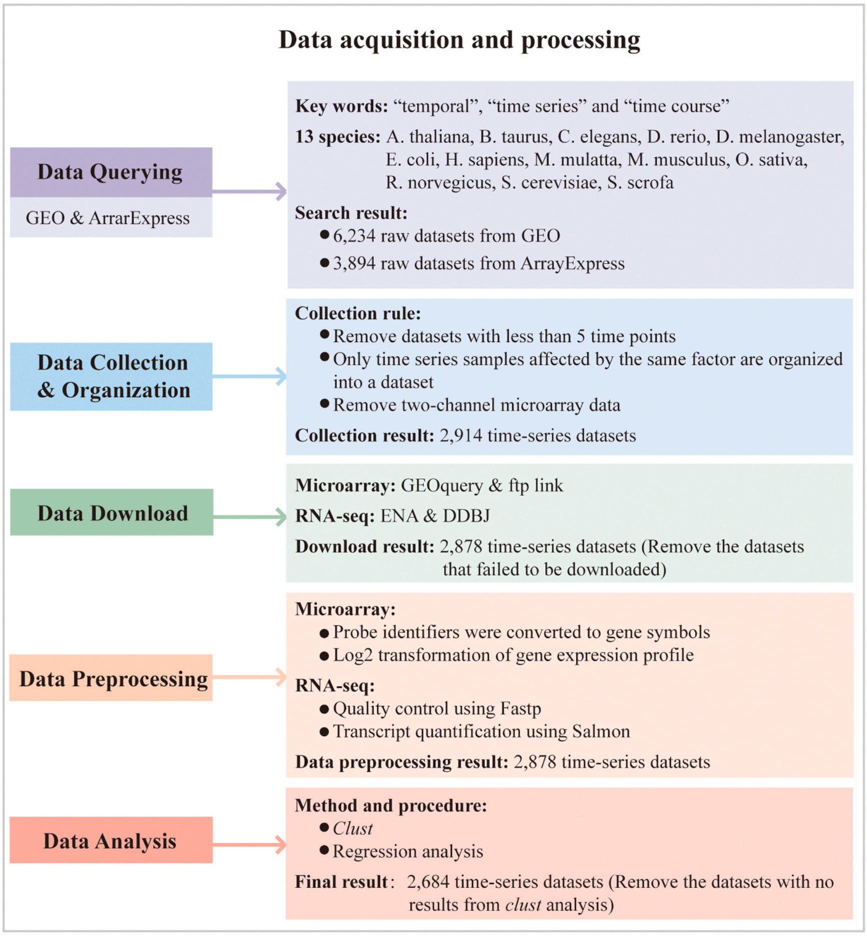

Data acquisition and processing
Data Querying
GEO & ArrarExpress
Key words: "temporal", "time series" and "time course"
13 species: A. thaliana, B. taurus, C. elegans, D. rerio, D. melanogaster, E. coli, H. sapiens, M. mulatta, M. musculus, O. sativa, R. norvegicus, S. cerevisiae, S. scrofa
Search result:
● 6,234 raw datasets from GEO
● 3,894 raw datasets from ArrayExpress
Data Collection & Organization
Collection rule:
● Remove datasets with less than 5 time points
● Only time series samples affected by the same factor are organized into a dataset
● Remove two-channel microarray data
Collection result: 2,914 time-series datasets
Data Download
Microarray: GEOquery & ftp link
RNA-seq: ENA & DDBJ
Download result: 2,878 time-series datasets (Remove the datasets that failed to be downloaded)
Data Preprocessing
Microarray:
● Probe identifiers were converted to gene symbols
● Log2 transformation of gene expression profile
RNA-seq:
● Quality control using Fastp
● Transcript quantification using Salmon
Data preprocessing result: 2,878 time-series datasets
Data Analysis
Method and procedure:
● Clust
● Regression analysis
Final result: 2,684 time-series datasets (Remove the datasets with no results from clust analysis)

Fig. 1. Data processing workflow. The workflow of data querying, collection, organization, downloading, preprocessing and analysis is shown, including the number of datasets from the initial collection to the final analysis, along with the reasons for exclusion.

incorporates self-starting functions, and the drm function. The good model for each gene cluster eigengene was determined by calculating the coefficient of determination (R^2), the adjusted coefficient of determination (adjusted R^2), Akaike information criterion (AIC), and root mean square error (RMSE) values. However, due to the biological complexity of gene expression, we cannot rely solely on these criteria to determine the good model that fits this biological process. Therefore, all models with adjusted $R^2 > 0.7$ were selected as fitted models. It is necessary to combine AIC, RMSE, prior knowledge, and experimental information to confirm the appropriate model for the cluster. In addition, if this eigengene does not explain a relevant amount of the variation, we will get rid of this cluster. Although we believe some individual gene profiles within this cluster may be fitted to a model, it can not be classified into a group of genes for downstream analysis. That's why we used the whole process we described but not gene profile analysis one by one.

Moreover, the "clusterProfiler" package (v4.0) in R was used to perform GO and KEGG enrichment analyses on the gene subsets, facilitating the exploration of key genes involved in different biological processes. Enriched functional terms with a adjusted P value < 0.05 were selected.

2.3. *Database construction*

GeTeSEPdb can be accessed for free at http://www.inbirg.com/GeTeSEPdb/. The online database framework was constructed using Django (v2.2.5) and was deployed on NGINX and uWSGI in a centOS environment. All datasets within GeTeSEPdb were managed and recorded using MySQL (v8.0.26) servers and the file system. For the front end, packages such as jQuery (v1.10.2), DataTable (v1.10.21), and Plotly (v2.8.3) were used to visualize and display the results. Statistical analysis was conducted using Python packages such as pandas (v1.4.1) and numpy (v1.23.1).

3. Results

3.1. *Overview of GeTeSEPdb*

GeTeSEPdb is a comprehensive gene database and mining tool with temporal-specific expression patterns designed to facilitate the linkage of transcriptional changes and temporal patterns of responses between genes. We employed *clust* for gene clustering and conducted regression analyses to determine the associated patterns. In the current version of the GeTeSEPdb database, we manually collected and organized the raw data of 2684 time-series transcriptome datasets from the GEO and ArrayExpress repositories. These datasets are derived from well-structured studies of temporal processes (comprising at least five time points) involving 13 different species, including humans. Consequently, we identified and stored 2644,370 gene profiles exhibiting temporal-specific expression patterns in the user-friendly GeTeSEPdb database. Furthermore, through the tools available in GeTeSEPdb, users have the ability to submit single-gene or omics data for online analysis, enabling the extraction of genes or gene subsets with specific expression patterns. The workflow of GeTeSEPdb is presented in Fig. 2.

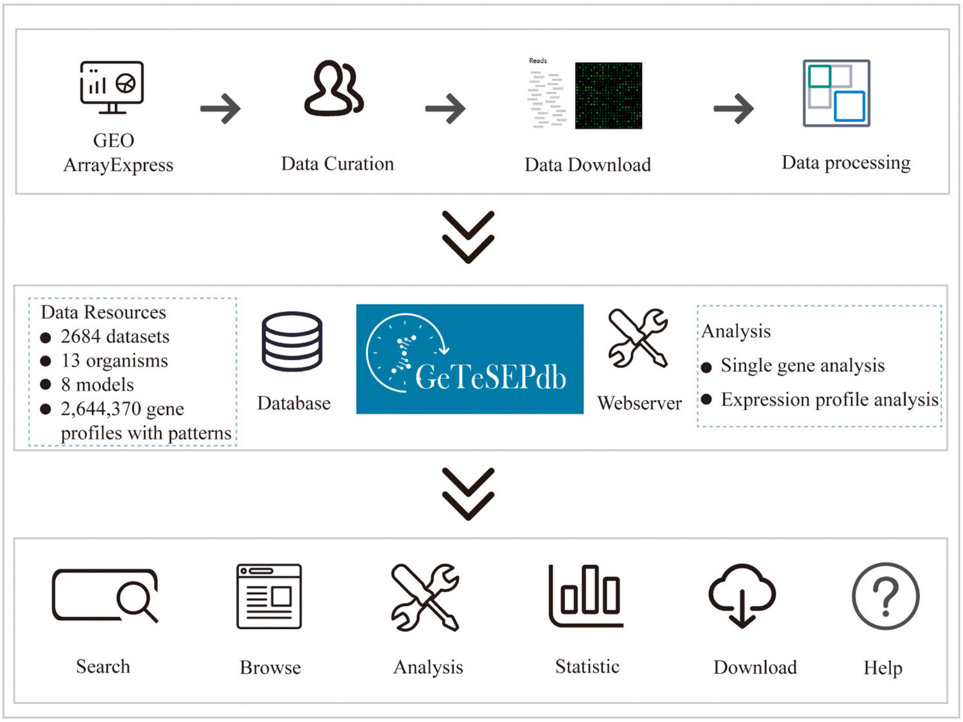


Fig. 2. Workflow of GeTeSEPdb. First, we manually collected and organized the raw data from time series of transcriptome datasets from the GEO and ArrayExpress repositories. Next, we identified and stored genes exhibiting temporal-specific expression patterns in GeTeSEPdb database, and tables and plots were generated for users. Furthermore, through the tools available in GeTeSEPdb, users have the ability to submit data for online analysis, enabling the extraction of gene subsets with specific expression patterns.

3.2. Features and utilities of GeTeSEPdb

GeTeSEPdb offers a web service divided into seven main pages: Home, Search, Browse, Analysis, Statistics, Download, and Help (Fig. 3). The Home page provides a brief overview of GeTeSEPdb's functionalities and features. On the Search page, users can explore gene profiles with specific expression patterns under various experimental conditions according to their research needs. This page includes six selection boxes: species, experimental type, specification of the expression pattern, tissue/cell type, sequencing type, and gene symbol. Users can randomly select a filter to select genes of interest. An autocomplete widget is implemented for gene query which provides suggestions while users type a gene symbol into the field. Additionally, a Help button is provided on this page, offering a concise summary of its functions and usage. After selecting the desired criteria, users can obtain analysis results by clicking the submit button.

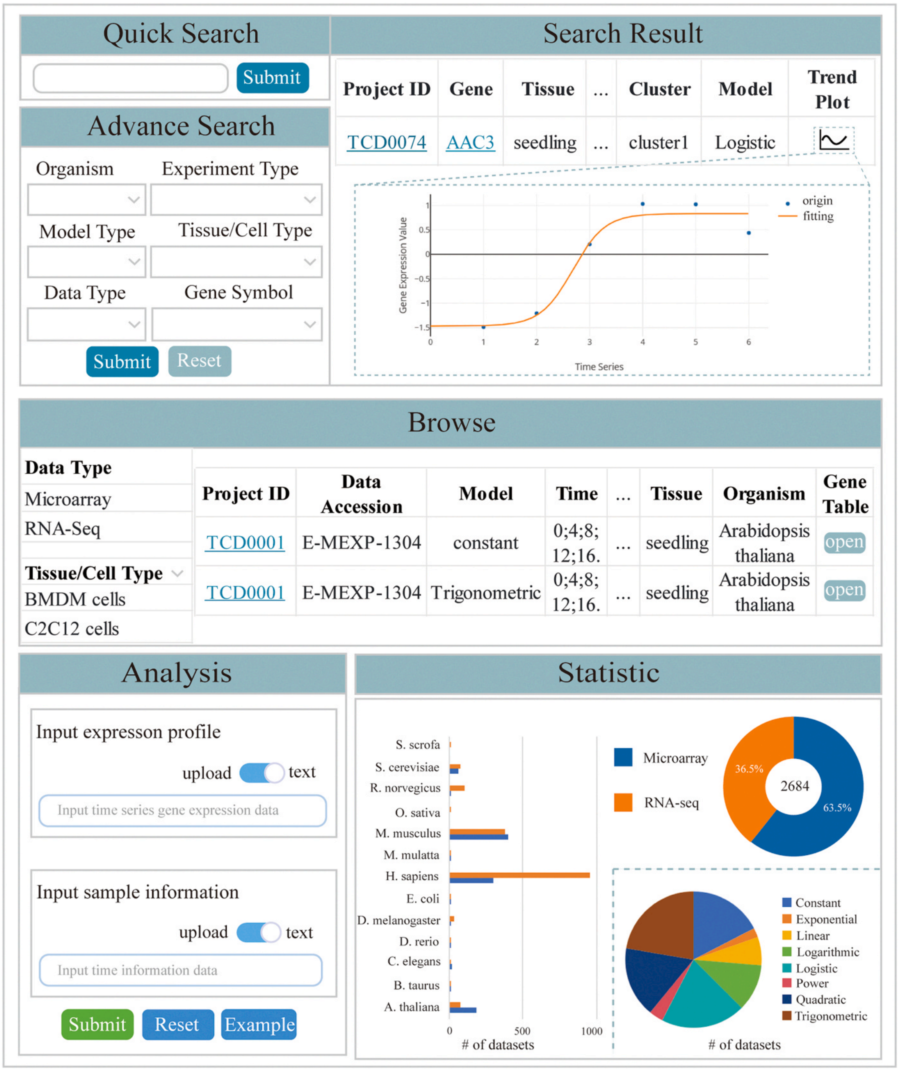
The Browse page allows users to access basic information about all collected datasets. By clicking on the project ID, users can access detailed analysis results for each dataset. After clicking Browse by pattern, a table displaying the data information corresponding to different patterns is shown. In the last column, the open entries are highlighted in colour. Clicking it enables access to genes for each expression pattern, and clicking on gene names provides detailed information and results for that gene.

The Analysis page offers two services: users can analyse the expression patterns of individual genes or explore gene profiles with different expression pattern specification within the entire expression profile dataset. Due to varying file sizes, this feature might require a long analysis time. Therefore, users are recommended to provide an email address, as the result page link will be sent upon completion. Additionally, a retrospective function is available, allowing users to view historical result pages based on analysis task IDs. Notably, historical results are stored for 30 days.

The Statistics page provides statistical information about available data within GeTeSEPdb. The Download page offers download links for users to access basic dataset information and analysis results. The Help page provides a comprehensive user guide to assist users in quickly starting GeTeSEPdb.

In summary, our GeTeSEPdb interface is primarily divided into two key components:

i. *Database*: Clusters of genes exhibiting similar dynamic patterns over a broad spectrum of response types can be thoroughly explored and analysed to assess their diverse functionalities. We then examined the reference literature for some of the datasets, observed the gene expression patterns they produced, and consulted mathematical and biological-related literature and books [33–42]. We employed eight regression models to discern and classify these gene clusters. Linear models are ideal for capturing gene profiles with gradual linear expression changes over time, thus aiding research on developmental processes and responses to external stimuli. The constant models are designed to identify gene profiles that maintain stable expression levels over time, serving as dependable internal controls or reference genes. Logistic models are effective at recognizing gene profiles that exhibit gradual or switch-like behaviours, which are particularly relevant in the context of processes such as cell differentiation or fate determination. The logarithmic models are tailored to detect gene profiles with linear growth or decay patterns, related to ageing, proliferation, and metabolic processes. Exponential models are ideal for identifying gene profiles with exponential growth or decay patterns and are particularly applicable to cases with cell growth and viral replication. Power models are designed to capture gene profiles with power law distributions, indicating their involvement in complex regulatory networks, gene family evolution, and other genomic quantities. Trigonometric models are proficient at identifying gene profiles with periodic or rhythmic expression patterns commonly associated with circadian rhythms and cell cycle progression. Quadratic function models are designed to detect gene profiles with



Project ID	Gene	Tissue	...	Cluster	Model	Trend Plot
TCD0074	AAC3	seedling	...	cluster1	Logistic	

Data Type
Microarray
RNA-Seq
Tissue/Cell Type
BMDM cells
C2C12 cells

Project ID	Data Accession	Model	Time	...	Tissue	Organism	Gene Table
TCD0001	E-MEXP-1304	constant	0;4;8; 12;16.	...	seedling	Arabidopsis thaliana	open
TCD0001	E-MEXP-1304	Trigonometric	0;4;8; 12;16.	...	seedling	Arabidopsis thaliana	open

Input expresson profile

upload text

Input time series gene expression data

Input sample information

upload text

Input time information data

Fig. 3. GeTeSEPdb content. On the Home page, users can conduct quick searches via the keyword of the gene symbol. On the Advanced search page, users can perform more advanced searches by filtering one or more features of the obtained dataset. On the "Results" page, datasets meeting the criteria are listed in a table. On the Browse page, users can browse sample datasets or filter gene datasets by experiment type, model type, data type, tissue type or organism. On the Analysis page, users can submit gene expression data and time series of sample information to analyse the expression patterns of individual genes or to identify gene profiles with different expression patterns in the entire expression profile dataset. Statistical graphics are also presented to visualize the distribution of datasets and model types on the Statistics page.

concave or convex expression patterns, facilitating the study of processes characterized by different stages or transitions. These eight types of regression models constitute the foundation for identifying and characterizing gene clusters with distinct expression dynamics, providing valuable insights into various biological processes and mechanisms. On the Search and Browse pages, we provide detailed information on the project IDs and genes. For the project ID detail pages, users can access comprehensive dataset information, evaluation metric results obtained with each of the eight models, expressions and parameter values of the fitting model as determined by these metrics, intuitive heatmaps and curve plots illustrating the specific expression patterns of different gene clusters, and GO and KEGG enrichment results; this information can be used to assess the diverse biological functions of different genes. Similarly, on the detailed gene pages, users can access complete information about genes, expression and parameter results for a respective model and view curve plots depicting expression changes. The biological significance of these genes can be considered when identifying potential biomarkers.

ii. *Web server*: Users can analyse the expression patterns of single genes or explore gene profiles with different specific expression patterns contained in the entire expression profile dataset. Initially, the biological significance of these genes can be determined based on functional annotation results to identify potential biomarkers. Subsequently, biomarkers in clinical samples can be used to assess their relevance to disease status, treatment outcomes, etc. Understanding gene expression patterns over time can help researchers identify drug targets suitable for different stages of drug development. Furthermore, genes profiles with similar expression patterns can be used to construct gene regulatory networks, facilitating the understanding of the interactions and regulatory relationships among genes involved in the same biological events within specific experimental contexts.

3.3. Application of GeTeSEPdb

In GeTeSEPdb, temporal genes are classified into one of eight specific expression patterns. Among these, a common pattern is rhythmic expression, characterized by gene expression conforming to sinusoidal and cosine function models. This pattern has been observed in various cell processes [43,44]. For instance, the cyclic behaviour of the yeast cell cycle is controlled by a series of transcription factors, with the expression level of each factor peaking at different stages of the cell cycle, thereby regulating the expression of the next stage transcription factors [45]. In our study, through the analysis of the TCD0105 dataset, we identified similar specific expression patterns. The TCD0105 dataset is primarily used to investigate the effects of sugar and diurnal rhythms on the regulation of diurnal gene expression in *Arabidopsis thaliana*. After preprocessing, the dataset contained a total of 20,817 genes. Our analysis revealed seven gene clusters with specific expression patterns. Based on model evaluation metrics such as R^2, adjusted R^2, RMSE, and AIC, we found that the adjusted R^2 of cluster 2 was less than 0.7 and thus excluded. Trigonometric function models were found to be fitted to the expression profiles of clusters 3 and 6, while the expression pattern of clusters 5 aligned with a quadratic function model. Additionally, both quadratic function and trigonometric function models were found to be fitted to the expression profiles of clusters 1 and 4 (Table S1). Observations revealed differences in the expression trends of these five clusters. For cluster 1, gene expression peaked at the end of the night and the beginning of the day (Fig. 4A). Enrichment analysis indicated those genes are related to diurnal rhythms, light responses in cells, and fatty acid and nitrate metabolism (Fig. 4C and D). In contrast, cluster 3, cluster 4, and cluster 6 reached their peak expression levels during the transition from the end of the day to the beginning of the night, after which their expression gradually decreased (Figs. 4B and S1). Enrichment analysis suggested those pattern genes are associated with ribosome biogenesis, amino acid, and protein biosynthesis processes (Fig. S2). These observations indicate that genes with specific metabolic

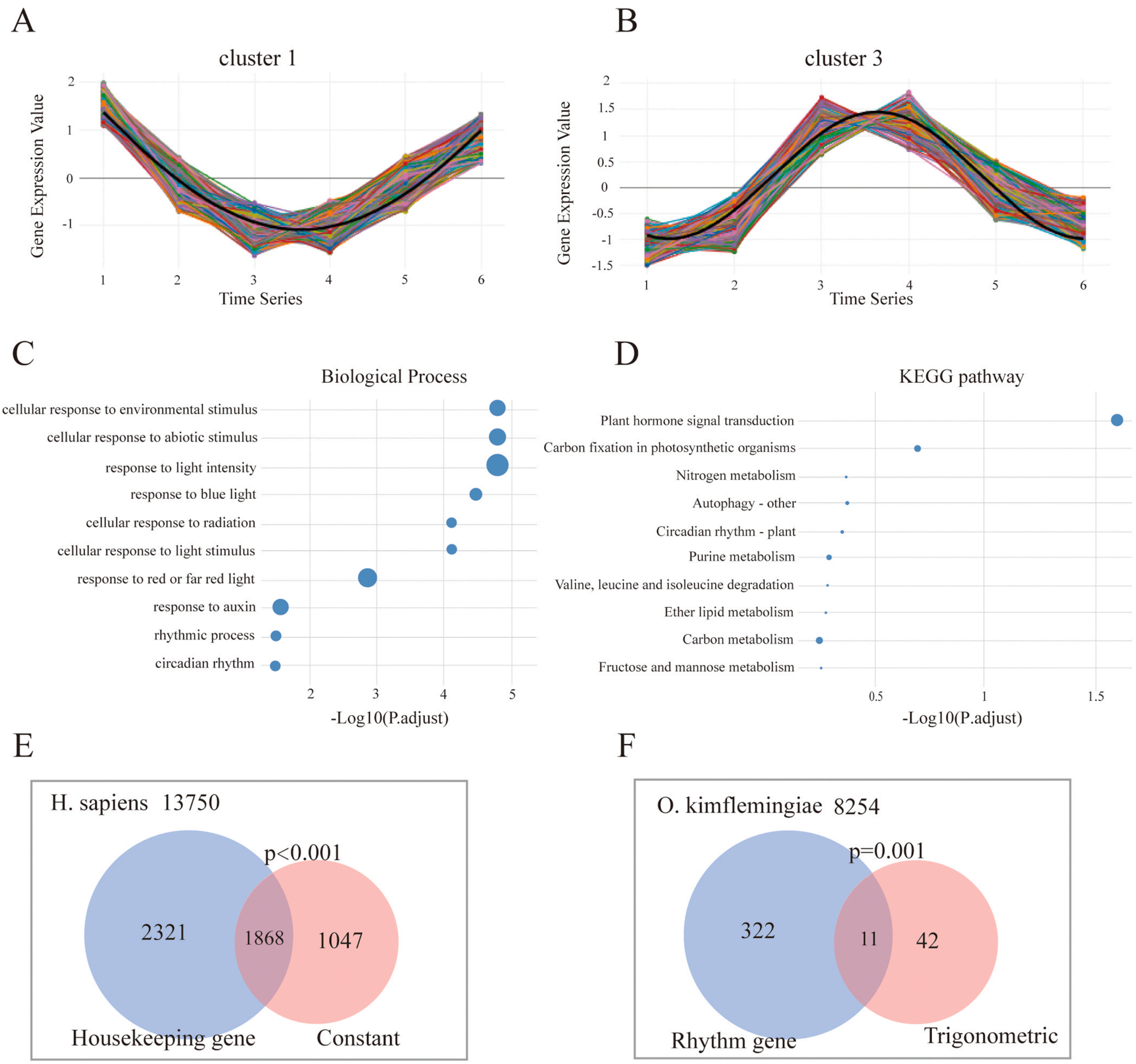


Fig. 4. Main application of GeTeSEPdb. (A) Temporal trends of genes in clusters 1. The coloured line is the gene expression trend, and the smooth black curve is the trigonometric function model fitting curve. (B) Temporal trends of genes in clusters 3. The coloured line is the gene expression trend, and the smooth black curve is the trigonometric function model fitting curve. (C) GO annotation results for cluster 1. (D) KEGG enrichment results for cluster 1. (E) Venn diagram showing the results of a hypergeometric examination of the human housekeeping genes. (F) Venn diagram showing the results of the hypergeometric examination of rhythm-related genes.

functions often exhibit different diurnal responses. The circadian clock and sugars play major roles in the diurnal regulation of gene expression, which is in line with previous literature findings [46].

Moreover, previous studies have indicated the significance of Bmal1 as a core circadian gene. This gene encodes a protein that typically forms a complex with the Clock protein, participating in the rhythmic regulation of the biological clock. This complex plays a pivotal role in transcriptional regulation by governing the expression of a series of downstream genes, maintaining the normal rhythm of the biological clock. Bmal1 has multiple functions in various physiological and pathological contexts, including the DNA damage response and insulin secretion [47,48]. In our database, we identified three datasets related to Bmal1. In the TCD1960, TCD2086, and TCD2436 datasets, Bmal1 demonstrated a trigonometric pattern (Table S2).

Another commonality is the constant function model, and recognized housekeeping genes adhere to this modelling pattern. A total of 76,093 gene profiles conforming to the constant-function model were identified, 22,064 of which were housekeeping gene profiles in humans. Subsequently, we selected 18,986 human genes as background genes and obtained 4189 genes known to be housekeeping genes [49,50] as background genes. We extracted 2915 genes from the 22,064 gene profiles that appeared at least 40 times and performed a hypergeometric distribution test to verify the intersection of these genes with the 4189 genes, resulting in a p value of less than 0.001 (Fig. 4E). These results underscore the statistical significance of our specific identified expression pattern genes.

Additionally, our analysis tools employed data from other species excluded during data collection, such as the GSE101312 dataset for *Ophiocordyceps kimflemingiae*. Through this analysis, we identified six gene clusters. Clusters 1, 2, 3, and 4 were excluded because they had adjusted R^2 values less than 0.7. Clusters 0 exhibited a constant function model, while clusters 5 demonstrated trigonometric and quadratic function pattern (Table S3). Furthermore, from articles related to this dataset, we identified a validated list of 333 rhythmic genes [51]. Based on the genome of *Ophiocordyceps kimflemingiae*, which contains 8629 genes as background genes, and a hypergeometric distribution test to verify the significance of their intersection with the 53 trigonometric model genes we identified, we obtained a p value of 0.001 (Fig. 4F and Table S4). Thus, we believe our method to be effective and capable of identifying potential biomarkers, thereby contributing to research in development, drug studies, and disease treatments. We aspire for GeTeSEPdb to become a valuable resource for both biologists and bioinformaticians.

4. Discussion and future research directions

To date, an increasing number of studies have explored gene expression across multiple species, tissues, and cell types under diverse experimental conditions, such as development, drug action, and pathogen infection scenarios, employing state-of-the-art sequencing technologies. In response, we systematically curated and analysed datasets to design and establish a reference database, aiming to decipher essential biological mechanisms within distinct experimental contexts. We believe that this resource, through the identification of coexpression patterns, holds the potential to reveal key genes and their dynamic biological networks involved in processes such as development or disease progression.

Herein, we introduce GeTeSEPdb, an interactive user-friendly database that exclusively focuses on gene profiles with specific expression patterns across 13 significant species, derived from the GEO and ArrayExpress repositories. Within GeTeSEPdb, users have the liberty to explore gene expression trends under various experimental conditions to gain deeper insights into the functions of genes exhibiting different expression patterns in biological processes. Such comprehensive analytical results are often challenging to find in existing databases.

Notably, traditional clustering methods, such as k-means [52], hierarchical clustering [53], and self-organizing maps [52], have been widely applied to gene expression data with the expectation that they can group observed outcomes together based on shared behaviour. However, these methods may not always identify genes that coregulate each other within biological systems. Additionally, numerous techniques, including Bayesian hierarchical clustering algorithms [54], hidden Markov model (HMM) algorithms [55], and curve fitting using spline clustering models [56], have been developed to model unknown shapes relatively effortlessly without requiring prior knowledge of the data structure. Nonetheless, these methods may lack biological relevance. Linear and nonlinear regression models represent vital statistical methods for defining relationships between variables of interest within biological systems [57]. Specific models fit certain datasets better than others. Regression analysis can be employed to effectively explain gene function and operating modes and to directly infer biological information from the shape of expression profiles [20,21]. Thus, we utilized *clust* to extract coexpressed gene clusters, aligned them with biological expectations for coexpressed gene clusters, and employed regression analysis to characterize the expression trend of each gene cluster, thereby elucidating the biological changes reflected by these genes.

However, there is room for improvement in this approach. We plan to further improve GeTeSEPdb in the following areas to make it more useful. Firstly, as a limitation in the current method, although temporal information is input, *clust* does not fully utilize it and may fail to capture some valuable temporal patterns. Therefore, we will keep optimizing the method of identification of temporal-specific expression patterns. Secondly, different types of spatial information in time-series data are equally important [58,59], and network analysis aids in exploring the interactions between genes exhibiting specific expression patterns in detail. Thirdly, integrating single-cell transcriptomics data to investigate temporal-specific expression patterns among cells will be one of our future research directions. In the future, we will strive to collect and integrate multiomics data and incorporate more models and network analysis methods into our database to provide more comprehensive information to assess biological phenomena, advance drug development, and gain deeper insights into the mechanisms behind disease progression.

CRediT authorship contribution statement

Jianbo Pan: Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Conceptualization. **Qiang Li:** Visualization. **Qinfeng Ma:** Methodology. **Ni Kuang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Zhaoyu Zhai:** Visualization. **Xuehang Meng:** Data curation. **Xiao Zheng:** Data curation.

Declaration of Competing Interest

The authors declare no competing interests.

Data availability statement

Any additional information required to reanalyze the data reported in this paper is available in the supplementary file. The source code for the GeTeSEPdb is available on GitHub (https://github.com/knlgd/GeTeSEPdb).

Acknowledgements

This work was supported by the Research Startup Funds of Chongqing Medical University, the Natural Science Foundation of Chongqing, China (CSTB2023NSCQ-MSX0289) and the University Innovation Research Group Project of Chongqing (CXQT21016).

Funding for open access charge: Chongqing Medical University.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.06.003.

References

[1] Bar-Joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes using time-series gene expression data. Nat Rev Genet 2012;13:552–64. https://doi.org/10.1038/nrg3244.

[2] Androulakis IP, Yang E, Almon RR. Analysis of time-series gene expression data: methods, challenges, and opportunities. Annu Rev Biomed Eng 2007;9:205–28. https://doi.org/10.1146/annurev.bioeng.9.060906.151904.

[3] Huang Y, Zaas AK, Rao A, Dobigeon N, Woolf PJ, et al. Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. PLoS Genet 2011;7:e1002234. https://doi.org/10.1371/journal.pgen.1002234.

[4] Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, et al. A network-based analysis of systemic inflammation in humans. Nature 2005;437:1032–7. https://doi.org/10.1038/nature03985.

[5] Desai KH, Tan CS, Leek JT, Maier RV, Tompkins RG, et al. Dissecting inflammatory complications in critically injured patients by within-patient gene expression changes: a longitudinal clinical genomics study. PLoS Med 2011;8:e1001093. https://doi.org/10.1371/journal.pmed.1001093.

[6] Taylor MW, Tsukahara T, Brodsky L, Schaley J, Sanda C, et al. Changes in gene expression during pegylated interferon and ribavirin therapy of chronic hepatitis C virus distinguish responders from nonresponders to antiviral therapy. J Virol 2007; 81:3391–401. https://doi.org/10.1128/JVI.02640-06.

[7] Baranzini SE, Mousavi P, Rio J, Caillier SJ, Stillman A, et al. Transcription-based prediction of response to IFNbeta using supervised computational methods. PLoS Biol 2005;3:e2. https://doi.org/10.1371/journal.pbio.0030002.

[8] Lin TH, Kaminski N, Bar-Joseph Z. Alignment and classification of time series gene expression in clinical studies. Bioinformatics 2008;24:i147–55. https://doi.org/10.1093/bioinformatics/btn152.

[9] Costa IG, Schonhuth A, Hafemeister C, Schliep A. Constrained mixture estimation for analysis and robust classification of clinical time series. Bioinformatics 2009;25: i6–14. https://doi.org/10.1093/bioinformatics/btp222.

[10] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98:5116–21. https://doi.org/10.1073/pnas.091062498.

[11] Aryee MJ, Gutierrez-Pabello JA, Kramnik I, Maiti T, Quackenbush J. An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). BMC Bioinforma 2009;10:409. https://doi.org/10.1186/1471-2105-10-409.

[12] Fischer DS, Theis FJ, Yosef N. Impulse model-based differential expression analysis of time course sequencing data. Nucleic Acids Res 2018;46:e119. https://doi.org/10.1093/nar/gky675.

[13] Ernst J, Bar-Joseph Z. STEM: a tool for the analysis of short time series gene expression data. BMC Bioinforma 2006;7:191. https://doi.org/10.1186/1471-2105-7-191.

[14] Kumar L, M EF. Mfuzz: a software package for soft clustering of microarray data. Bioinformation 2007;2:5–7. https://doi.org/10.6026/97320630002005.

[15] Magni P, Ferrazzi F, Sacchi L, Bellazzi R. TimeClust: a clustering tool for gene expression time series. Bioinformatics 2008;24:430–2. https://doi.org/10.1093/bioinformatics/btm605.

[16] Schulz MH, Devanny WE, Gitter A, Zhong S, Ernst J, et al. DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. BMC Syst Biol 2012;6:104. https://doi.org/10.1186/1752-0509-6-104.

[17] Zhou Z, Tan C, Chau MHK, Jiang X, Ke Z, et al. TEDD: a database of temporal gene expression patterns during multiple developmental periods in human and model organisms. Nucleic Acids Res 2023;51:D1168–78. https://doi.org/10.1093/nar/gkac978.

[18] Liu X, Yu X, Zack DJ, Zhu H, Qian J. TiGER: a database for tissue-specific gene expression and regulation. BMC Bioinforma 2008;9:271. https://doi.org/10.1186/1471-2105-9-271.

[19] Xiao SJ, Zhang C, Zou Q, Ji ZL. TiSGeD: a database for tissue-specific genes. Bioinformatics 2010;26:1273–5. https://doi.org/10.1093/bioinformatics/btq109.

[20] Chechik G, Koller D. Timing of gene expression responses to environmental changes. J Comput Biol 2009;16:279–90. https://doi.org/10.1089/cmb.2008.13TT.

[21] Kalaitzis AA, Lawrence ND. A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. BMC Bioinforma 2011;12:180. https://doi.org/10.1186/1471-2105-12-180.

[22] Ahdesmaki M, Lahdesmaki H, Gracey A, Shmulevich L, Yli-Harja O. Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data. BMC Bioinforma 2007;8:233. https://doi.org/10.1186/1471-2105-8-233.

[23] Abu-Jamous B, Kelly S. Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. Genome Biol 2018;19:172. https://doi.org/10.1186/s13059-018-1536-8.

[24] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res 2013;41:D991–5. https://doi.org/10.1093/nar/gks1193.

[25] Sarkans U, Fullgrabe A, Ali A, Athar A, Behrangi E, et al. From ArrayExpress to BioStudies. Nucleic Acids Res 2021;49:D1502–6. https://doi.org/10.1093/nar/gkaa1062.

[26] Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 2007;23:1846–7. https://doi.org/10.1093/bioinformatics/btm254.

[27] Harrison PW, Ahamed A, Aslam R, Alako BTF, Burgin J, et al. The European Nucleotide Archive in 2020. Nucleic Acids Res 2021;49:D82–5. https://doi.org/10.1093/nar/gkaa1028.

[28] Ogasawara O, Kodama Y, Mashima J, Kosuge T, Fujisawa T. DDBJ Database updates and computational infrastructure enhancement. Nucleic Acids Res *48*, D45-D50 2020. https://doi.org/10.1093/nar/gkz982.

[29] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44:D733–45. https://doi.org/10.1093/nar/gkv1189.

[30] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34:i884–90. https://doi.org/10.1093/bioinformatics/bty560.

[31] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 2017;14:417–9. https://doi.org/10.1038/nmeth.4197.

[32] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 2000;97: 10101–6. https://doi.org/10.1073/pnas.97.18.10101.

[33] Otto SP, Day T, De G. A Biologist's Guide to Mathematical Modeling in Ecology and Evolution. Princeton University Press,; 2011.

[34] Batschelet E. Introduction to Mathematics for Life Scientists. Springer-Verlag,; 1973.

[35] Novozhilov AS, Karev GP, Koonin EV. Biological applications of the theory of birth-and-death processes. Brief Bioinform 2006;7:70–85. https://doi.org/10.1093/bib/bbk006.

[36] Slowikowski K, Nguyen HN, Noss EH, Simmons DP, Mizoguchi F, et al. CUX1 and IkappaBzeta (NFKBIZ) mediate the synergistic inflammatory response to TNF and IL-17A in stromal fibroblasts. Proc Natl Acad Sci USA 2020;117:5532–41. https://doi.org/10.1073/pnas.1912702117.

[37] Ghandhi SA, Sima C, Weber WM, Melo DR, Rudqvist N, et al. Dose and Dose-Rate Effects in a Mouse Model of Internal Exposure to 137Cs. Part 1: Global Transcriptomic Responses in Blood. Radiat Res 2020;196:478–90. https://doi.org/10.1667/RADE-20-00041.

[38] Keller MA, Addya S, Vadigepalli R, Banini B, Delgrosso K, et al. Transcriptional regulatory network analysis of developing human erythroid progenitors reveals patterns of coregulation and potential transcriptional regulators. Physiol Genom 2006;28:114–28. https://doi.org/10.1152/physiolgenomics.00055.2006.

[39] Delic D, Wunderlich F, Al-Quraishy S, Abdel-Baki AS, Dkhil MA, et al. Vaccination accelerates hepatic erythroblastosis induced by blood-stage malaria. Malar J 2020; 19:49. https://doi.org/10.1186/s12936-020-3130-2.

[40] Starmans MH, Chu KC, Haider S, Nguyen F, Seigneuric R, et al. The prognostic value of temporal in vitro and in vivo derived hypoxia gene-expression signatures in breast cancer. Radio Oncol 2012;102:436–43. https://doi.org/10.1016/j.radonc.2012.02.002.

[41] Almon RR, Yang E, Lai W, Androulakis IP, DuBois DC, et al. Circadian variations in rat liver gene expression: relationships to drug actions. J Pharm Exp Ther 2008; 326:700–16. https://doi.org/10.1124/jpet.108.140186.

[42] Denolet E, De Gendt K, Allemeersch J, Engelen K, Marchal K, et al. The effect of a sertoli cell-selective knockout of the androgen receptor on testicular gene expression in prepubertal mice. Mol Endocrinol 2006;20:321–34. https://doi.org/10.1210/me.2005-0113.

[43] Ramsey SA, Klemm SL, Zak DE, Kennedy KA, Thorsson V, et al. Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. PLoS Comput Biol 2008;4:e1000021. https://doi.org/10.1371/journal.pcbi.1000021.

[44] Perrin L, Loizides-Mangold U, Chanon S, Gobet C, Hulo N, et al. Transcriptomic analyses reveal rhythmic and CLOCK-driven pathways in human skeletal muscle. Elife 2018;7. https://doi.org/10.7554/eLife.34114.

[45] Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, et al. Serial regulation of transcriptional regulators in the yeast cell cycle. Cell 2001;106:697–708. https://doi.org/10.1016/s0092-8674(01)00494-9.

[46] Michael TP, Mockler TC, Breton G, McEntee C, Byer A, et al. Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. PLoS Genet 2008;4:e14. https://doi.org/10.1371/journal.pgen.0040014.

[47] Sun Y, Wang P, Li H, Dai J. BMAL1 and CLOCK proteins in regulating UVB-induced apoptosis and DNA damage responses in human keratinocytes. J Cell Physiol 2018; 233:9563–74. https://doi.org/10.1002/jcp.26859.

[48] Kalsbeek A, Foppen E, Schalij I, Van Heijningen C, van der Vliet J, et al. Circadian control of the daily plasma glucose rhythm: an interplay of GABA and glutamate. PLoS One 2008;3:e3194. https://doi.org/10.1371/journal.pone.0003194.

[49] Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet 2013;29:569–74. https://doi.org/10.1016/j.tig.2013.05.010.

[50] Hounkpe BW, Chenou F, de Lima F, De Paula EV. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. Nucleic Acids Res 2021;49: D947–55. https://doi.org/10.1093/nar/gkaa609.

[51] de Bekker C, Will I, Hughes DP, Brachmann A, Merrow M. Daily rhythms and enrichment patterns in the transcriptome of the behavior-manipulating parasite Ophiocordyceps kimflemingiae. PLoS One 2017;12:e0187170. https://doi.org/10.1371/journal.pone.0187170.

[52] McLachlan GJ, Bean RW, Ng SK. Clustering. Methods Mol Biol *1526* 2017:345–62. https://doi.org/10.1007/978-1-4939-6613-4_19.

[53] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863–8. https://doi.org/10.1073/pnas.95.25.14863.

[54] Cooke EJ, Savage RS, Kirk PD, Darkins R, Wild DL. Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. BMC Bioinforma 2011;12:399. https://doi.org/10.1186/1471-2105-12-399.

[55] Oh S, Song S, Grabowski G, Zhao H, Noonan JP. Time series expression analyses using RNA-seq: a statistical approach. Biomed Res Int 2013;2013:203681. https://doi.org/10.1155/2013/203681.

[56] Ma P, Castillo-Davis CI, Zhong W, Liu JS. A data-driven clustering method for time course gene expression data. Nucleic Acids Res 2006;34:1261–9. https://doi.org/10.1093/nar/gkl013.

[57] Jarantow SW, Pisors ED, Chiu ML. Introduction to the Use of Linear and Nonlinear Regression Analysis in Quantitative Biological Assays. Curr Protoc 2023;3:e801. https://doi.org/10.1002/cpz1.801.

[58] Zepp JA, Zacharias WJ, Frank DB, Cavanaugh CA, Zhou S, et al. Distinct mesenchymal lineages and niches promote epithelial self-renewal and myofibrogenesis in the lung. e1110 Cell 2017;170:1134–48. https://doi.org/10.1016/j.cell.2017.07.034.

[59] Niethamer TK, Stabler CT, Leach JP, Zepp JA, Morley MP, et al. Defining the role of pulmonary endothelial cell heterogeneity in the response to acute lung injury. Elife 2020;9. https://doi.org/10.7554/eLife.53072.