frontiers in Bioinformatics

TECHNOLOGY AND CODE
published: 25 November 2021
doi: 10.3389/fbinf.2021.693836

GeneCloudOmics: A Data Analytic Cloud Platform for High-Throughput Gene Expression Analysis

Mohamed Helmy[1,2], *Rahul Agrawal*[3], *Javed Ali*[3], *Mohamed Soudy*[4], *Thuy Tien Bui*[1] and *Kumar Selvarajoo*[1,5,6]*

[1]Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore, [2]Department of Computer Science, Lakehead University, Thunder Bay, ON, Canada, [3]Department of Geology and Geophysics, Indian Institute of Technology (IIT) Kharagpur, Kharagpur, India, [4]Proteomics and Metabolomics Unit, Children Cancer Hospital (CCHE-57357), Cairo, Egypt, [5]Singapore Institute of Food and Biotechnology Innovation (SIFBI), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore, [6]Synthetic Biology for Clinical and Technological Innovation (SynCTI), National University of Singapore (NUS), Singapore, Singapore

OPEN ACCESS

Edited by:
Stephen Taylor,
Weatherall Institute of Molecular Medicine (MRC), United Kingdom

Reviewed by:
Daniel Baum,
Zuse Institute Berlin, Germany
Bjorn Sommer,
Royal College of Art, United Kingdom

***Correspondence:**
Kumar Selvarajoo,
Kumar_Selvarajoo@bii.a-star.edu.sg

Specialty section:
This article was submitted to Data Visualization, a section of the journal Frontiers in Bioinformatics

Received: 07 May 2021
Accepted: 14 October 2021
Published: 25 November 2021

Citation:
Helmy M, Agrawal R, Ali J, Soudy M, Bui TT and Selvarajoo K (2021) GeneCloudOmics: A Data Analytic Cloud Platform for High-Throughput Gene Expression Analysis. Front. Bioinform. 1:693836. doi: 10.3389/fbinf.2021.693836

Gene expression profiling techniques, such as DNA microarray and RNA-Sequencing, have provided significant impact on our understanding of biological systems. They contribute to almost all aspects of biomedical research, including studying developmental biology, host-parasite relationships, disease progression and drug effects. However, the high-throughput data generations present challenges for many wet experimentalists to analyze and take full advantage of such rich and complex data. Here we present GeneCloudOmics, an easy-to-use web server for high-throughput gene expression analysis that extends the functionality of our previous ABioTrans with several new tools, including protein datasets analysis, and a web interface. GeneCloudOmics allows both microarray and RNA-Seq data analysis with a comprehensive range of data analytics tools in one package that no other current standalone software or web-based tool can do. In total, GeneCloudOmics provides the user access to 23 different data analytical and bioinformatics tasks including reads normalization, scatter plots, linear/non-linear correlations, PCA, clustering (hierarchical, k-means, t-SNE, SOM), differential expression analyses, pathway enrichments, evolutionary analyses, pathological analyses, and protein-protein interaction (PPI) identifications. Furthermore, GeneCloudOmics allows the direct import of gene expression data from the NCBI Gene Expression Omnibus database. The user can perform all tasks rapidly through an intuitive graphical user interface that overcomes the hassle of coding, installing tools/packages/libraries and dealing with operating systems compatibility and version issues, complications that make data analysis tasks challenging for biologists. Thus, GeneCloudOmics is a one-stop open-source tool for gene expression data analysis and visualization. It is freely available at http://combio-sifbi.org/GeneCloudOmics.

Keywords: OMICS data, gene expression analysis, bioinformatics, microarray, RNA-seq, transcriptomics, data analytics

INTRODUCTION

Multi-dimensional biological data is rapidly accumulating, and it is expected that the size of the data will exceed astronomical levels by 2025 (Stephens et al., 2015). This resulted in the development of computational tools that became vital in driving scientific discovery in recent times (Markowetz, 2017). A parallel increase in the development of online servers and databases has also been witnessed (Helmy et al., 2016), raising a new set of challenges related to the usability and maintenance of all these tools (Mangul et al., 2019). About half of the computational biology tools were found to be difficult to install, 28% of them are unavailable online in the provided URLs, and many others are missing adequate documentation and manuals (Mangul et al., 2019). The problem gets more complex with the limited computational and coding skills of two-thirds of the biologists who use these tools (Schultheiss, 2011). On the other hand, it was also noted that bioinformatics tools that are easy to install and use are highly cited, indicating wider usability by the community and a larger contribution to scientific discovery (Mangul et al., 2019). Thus, more web-based tools that avoid installation difficulties and operating system compatibility issues, simple point-and-click tools are required to tackle multi-dimensional omics datasets.

Gene expression profiling is widely used in biomedical research. They enable the investigation of expressed genes and their relevant pathways and cellular processes in a given time point or condition (Stark et al., 2019). Gene expression profiling is usually performed using RNA-Seq or microarray data since they detect the presence and quantify an RNA, the output indicator of an activated or deactivated gene (Yang et al., 2020). It also provides a deeper understanding of the biological system dynamics, growth or developmental process, drug effects or disease mechanisms through the differential gene expression (DGE) analysis (Piras et al., 2014, 2019; Simeoni et al., 2015; Hodgson et al., 2019; Bui et al., 2020; Wang et al., 2020). The DGE analysis determines genes with different expression levels between two or more conditions and statistically confirmed as differentially expressed (Pertea et al., 2016; Bui et al., 2020).

The analysis of gene expressions or transcriptomics data faces several challenges related to data size, quality, statistical analysis, visualization and interpretation of the results using current bioinformatics approaches (Mantione et al., 2014; Zou et al., 2019). Several bioinformatics or data science tools are available for addressing each of these challenges in the form of stand-alone software tools, web-server or R packages/Python libraries (Russo and Angelini, 2014; Poplawski et al., 2016; Velmeshev et al., 2016; McDermaid et al., 2019; Zou et al., 2019) (**Table 1**). However, most of the tools only provide a subset of analytics and require some level of programming skills. Often, the users need to move from one tool to another and this could lead to data compatibility issues (Chowdhury et al., 2019).

The analysis of gene expression data remains a burden for many biologists due to its intensive requirement of computational, statistical and programming skills that are lacking in two-thirds of biologists who use online biological resources (Schultheiss, 2011). Moreover, as mentioned above, most of the tools are individually scattered. Thus, there is a need to put the tools together in an easy-to-use manner with an intuitive GUI that will allow users to perform bioinformatics analyses with minimum computational skills and resources. In other words, a one-stop online server for transcriptomic data analysis that performs all essential steps of data import, pre-processing, statistical analyses, DGE identifications and functional interpretations of the results, through a friendly and simple user interface, is much needed.

Previously, we had developed ABioTrans as a stand-alone biostatistical tool for transcriptomics data analysis, including data pre-processing, statistical analyses, DGE and gene ontology (GO) classification (Zou et al., 2019). It is a downloadable executable that runs on any web browser with an interactive GUI (**Table 2**). However, as it is a stand-alone application written in R, the user needs to download it, install R or RStudio then run an installation script that installs all the required and up-to-date packages and dependencies. This was found to be challenging for some users as it requires a minimum level of programming familiarity, and several packages became incompatible with the new release of R (v4.0.0) in spring 2020. This is one common problem for most bioinformatics tools (Mangul et al., 2019). ABioTrans also needs approx. 10 min to download all packages before running. Hence, to provide users with a quick, ready-to-use that does not require regular system updates, a web server version is imminent.

To overcome the above-mentioned challenges, here we rebuilt ABioTrans as a new webserver and expanded its functionality to include several new analysis tools such as SOM, t-SNE, random forest clustering, and added further tools for bioinformatics functional analysis of gene and protein sets that includes PPI, protein complex analysis, evolutionary analysis, pathological analysis, physicochemical analysis, and more. We named this new revamped tool GeneCloudOmics, a web server for transcriptomics data analysis and gene/protein bioinformatics that is equipped with publication-ready plotting capabilities. GeneCloudOmics allows 12 biostatistical and data analytics tests and 11 bioinformatics tools for gene/protein datasets analysis and annotation (see Methods and Program Description). In addition, it provides direct data import from NCBI's GEO databases through GEO accession numbers. GeneCloudOmics webserver, thus, relieves the burdens of installation and version compatibilities and is designed to be a quick one-stop transcriptomics (RNASeq and microarray) data analysis tool that provides the user with all the required steps for their analysis (**Figure 1**). Overall, the web server targets users without any computational or programming skills and provides them with a wide spectrum of hassle-free analytic tools.

METHODS AND PROGRAM DESCRIPTION

The Gene Expression Profiling Workflow

The gene expression analysis aims to identify genes expressed under a particular condition, treatment, developmental stage, or disease. This requires assessing thousands of gene expressions of multiple conditions in raw format, pre-processing and

normalizing the expression levels, statistically analysing the data, identify DGEs between conditions and perform a functional analysis to elucidate the pathways and cellular functions of the DGEs (McDermaid et al., 2019) (**Figure 1**). GeneCloudOmics performs this workflow easily and smoothly on a web server as will be described below.

Overview of GeneCloudOmics Web Server

GeneCloudOmics provides users with a complete pipeline for analysing and interpreting their transcriptome data (**Figure 2** and **Table 2**):

1) Data types: users input microarray (.cel files) or RNA-Seq data (raw or normalized read count table in. csv format). In addition, users can provide NCBI GEO database accession and GeneCloudOmics automatically imports the data from the database.
2) Pre-processing raw data using four different normalization techniques (RPKM, FPKM, TPM, RUV), then plotting the normalized data versus the raw data inbox and/or with violin plots. The pre-processed data can be downloaded into a CSV file.
3) Analyse the pre-processed data using nine different statistical tests (read normalization, scatter plots, linear/non-linear correlations, PCA, hierarchical clustering, k-means clustering, t-SNE clustering and SOM clustering) then plot the results of each test in a publication-ready quality.
4) Perform DGE analysis using three of the most commonly used methods DESeq2 (Love et al., 2014), NOISeq (Tarazona et al., 2015) and EdgeR (Robinson et al., 2009) with a single

TABLE 1 | Comparison between 30 different gene expression analysis tools.

Tool Name	Interface		Data		Transcriptomics Data Analysis											Gene/Protein Set Analysis													References
	Active*	Web server	RNA-Seq	Microarray	DEA	PCA 2D/3D	Scatter Plot	Heatmap	Correlation Analysis	Entropy	Noise Analysis	Volcano Plot	Dispersion Plot	Gene Clustering	Sample Clustering	Gene ontology	Pathway Enrichment	Protein-Protein Interaction	Protein Domains	Co-expression	Complex enrichment	Protein Function	Subcellular Localization	Tissue Expression	Protein Physicochemical	Protein Evolutionary analysis	Protein Phylogenetic Analysis	Protein Pathological Analysis	
GeneCloudOmics	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	
AbioTrans	■		■		■	■	■	■	■	■	■	■	■	■		■													(Zou et al., 2019)
AltAnalyze	■		■	■	■											■	■	■	■										(D et al., 2010)
ASAP	■	■	■		■	■		■						■		■	■			■									(Gardeux et al., 2017)
CANEapp	■		■		■																								(Velmeshev et al., 2016)
compcodeR	■		■		■																								(Soneson, 2014)
DEBrowser	■		■		■	■	■	■				■	■	■	■	■	■												(Kucukural et al., 2019)
DEIVA	■		■									■	■	■															(Harshbarger et al., 2017)
derfinder	■		■		■	■																							(Collado-Torres et al., 2017)
Degust	■	■	■			■		■				■	■				■												(David Powell, 2015)
EXPath Tool	■	■	■	■	■				■							■	■			■									(Zheng et al., 2017)
expVIP	■	■	■		■									■	■														(Borrill et al., 2016)
GENAVi	■	■	■	■	■	■		■				■		■		■													(Reyes et al., 2019)
GENE-Counter	■		■	■	■				■							■													(Cumbie et al., 2011)
IDEAMEX	■	■	■		■	■	■	■	■																				(Jiménez-Jacinto et al., 2019)
iDEP	■	■	■	■	■	■		■				■	■	■	■	■	■												(Ge et al., 2018)
IRIS-EDA	■	■	■		■	■	■	■	■			■	■	■	■														(Monier et al., 2019)
MeV	■	■	■		■	■								■	■	■													(Howe et al., 2011)
MyRNA	■		■		■																								(Langmead et al., 2010)
NetworkAnalyst	■	■	■	■				■				■				■	■	■		■									(Zhou et al., 2019)
omicplotR	■		■		■	■																							(Giguere et al., 2021)
RSEB**		■	■													■			■										(Nussbaumer et al., 2014)
RNASeqGUI			■		■	■	■	■								■	■												(Russo and Angelini, 2014)
RobiNA	■	■	■	■	■																								(Lohse et al., 2012)
RseqFlow	■		■		■																								(Wang et al., 2011)
RSEQREP	■		■		■	■		■						■		■	■			■									(Jensen et al., 2018)
ShinyNGS	■		■		■	■	■	■				■			■		■												(Manning, 2017)
sleuth	■		■		■			■				■																	(Pimentel et al., 2017)
SPARTA	■		■		■		■																						(Johnson et al., 2016)
ScatLay	■	■	■	■	■		■																						(Bui et al., 2020)
START App		■	■		■	■		■				■																	(Nelson et al., 2017)

Blue and red cells indicate the features presence and absence, respectively. *Active: it is currently available. **RNASeq Expression Browser.

TABLE 2 | Features comparison between ABioTrans and GeneCloudOmics.

Architecture	ABioTrans	GeneCloudOmics
Application Type	Stand alone	Web-based
Requirements	R, RStudio, Web browser	Web browser
Gene expression data		
Supported transcriptome data	RNA-seq	RNA-Seq, Microarray
Supported transcriptome data formats	Gene expression table	Gene expression table microarray cel files
Preprocessing and normalization		
Low count filtering	Yes	Yes
Sequencing depth correction methods	TPM, RPKM, FPKM	TPM, RPKM, FPKM
Batch effect correction	UQ, RUV, TMM	UQ, RUV, TMM
Biostatistics and Analytics		
Dimension reduction	PCA	PCA Sparse PCA Self-organizing map (SOM)
Distribution fitting	Yes	Yes
Scatter plot	Yes	Yes
Correlation analysis	Pearson, spearman	Pearson, spearman
Entropy	Yes	Yes
Noise	Yes	Yes
Differential expression analysis		
Supporting plots	volcano plot, dispersion plot	volcano plot, dispersion plot
Gene-based Clustering		
Clustering algorithm	K-means clustering hierarchical clustering	K-means clustering hierarchical clustering
Visualization	Gene expression heatmap	Gene expression heatmap
Sample-based Clustering		
Clustering algorithm	K-means clustering on PC space	K-means clustering on PC space Random Forest clustering Self-organizing map (SOM)
Visualization	RF plot	RF plot, property plot, count plot, codes plot, distance plot and cluster plot
Gene set analysis		
Gene ontology	Yes	Yes
Gene ontology data	Local databases (NIH)	Online databases (UniProt)
Gene ontology visualization	Pie chart hairball graph	Bar chart Table
Pathway enrichment analysis	No	g:Profiler, Cytoscape (V)
Protein set analysis		
Protein-protein interaction	No	UniProt, Cytoscape (V)
Complex enrichment	No	CORUM
Protein function	No	UniProt
Subcellular localization	No	UniProt
Protein domains	No	UniProt
Tissue Expression	No	UniProt
Co-expression	No	GeneMANIA
Protein sequences	No	UniProt
Protein physicochemical analysis	No	Charge, GRAVY
Protein phylogenetic analysis	No	MAS, PGT., Chrom. Loc., G.Tree
Protein pathological analysis	No	UniProt

interface for choosing the parameters for each of the methods and in a similar way to plot the results in volcano or dispersion plots. The user can then download the results as a CSV file and the plots as or PNG or PDF.

5) Functionally interpret the DGEs or proteins set using 11 different bioinformatics tools (listed in detail below and in **Table 2**) that help the user perform essential enrichments and annotations to the gene/protein sets such as pathway enrichment analysis, gene ontology (GO) enrichment, PPI, and protein function enrichment. All the tests are performed through the same interface which allows the user to upload or paste a list of genes or proteins, choose the test parameters, run the analysis, and plot the results, using the standard visualization provided, or download them. The gene/protein set interpretation features are independent from the DGE analysis and can be used separately with any gene/protein set as a stand-alone feature (see demonstration sections below).
6) Creating an analysis report that summarizes and gathers all analyses of the user. In each test or analysis, the user can choose "Add to Report" option which will add the plot and the analysis title to the analysis report. When the user clicks "Analysis report" link in the main menu, the system generates

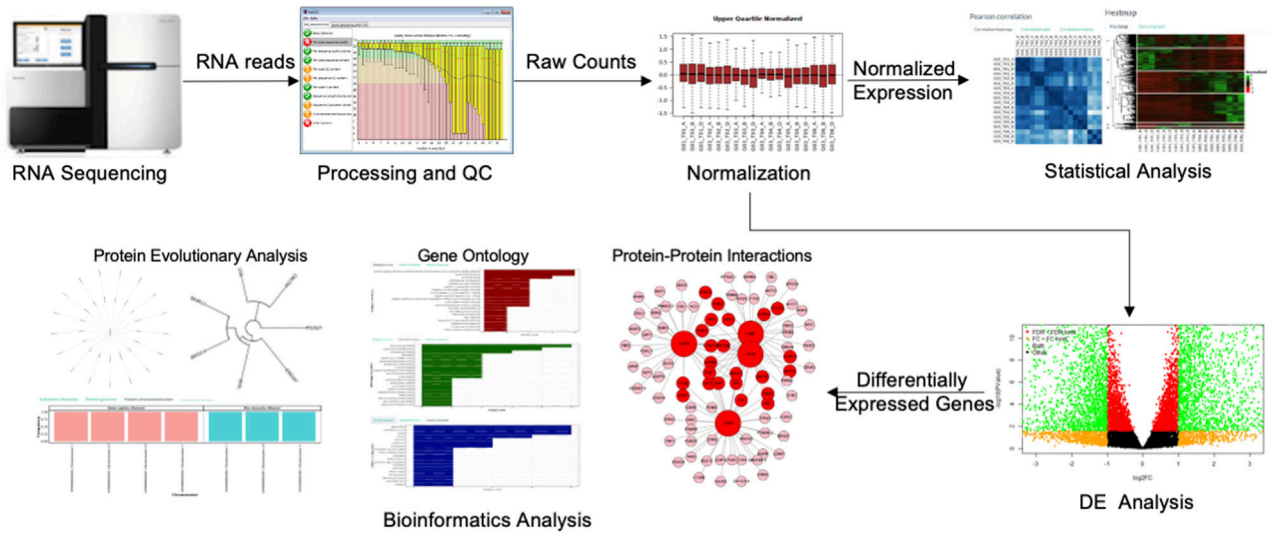

FIGURE 1 | The gene expression profiling workflow. The RNA sequencer produces raw RNA read counts that are aligned on the cell's genome and processed through the quality control (QC) steps. The raw read counts result from QC are next normalized and analyzed statistically to infer the differential gene expressions (DGEs) or other analyses such as Shannon Entropy, Correlations or PCA. Several bioinformatics analyses can also be performed on the list of DEG for functional inference.

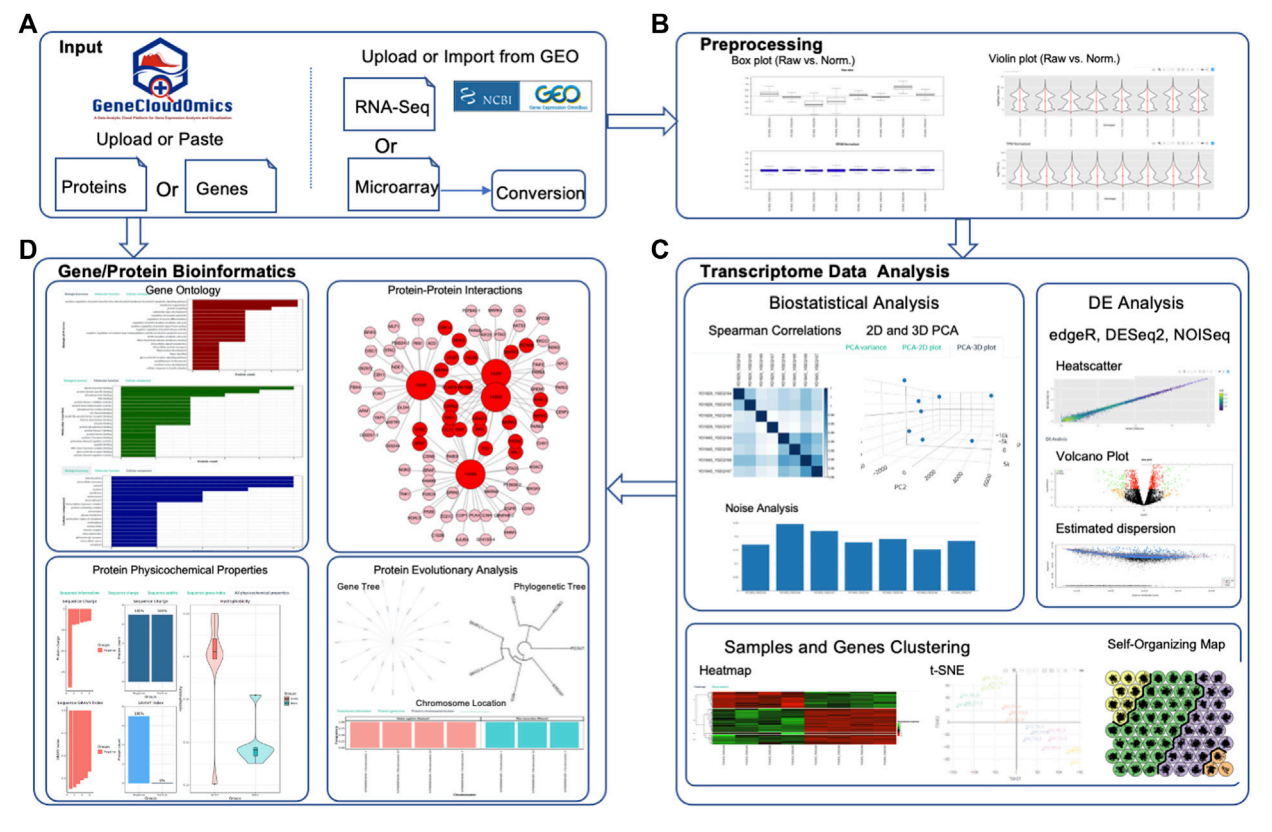

FIGURE 2 | Schematic Overview of GeneCloudOmics. **(A)** RNASeq and Microarray data uploading, **(B)** Data Pre-processing, **(C)** Transcriptome Data Analysis (e.g. Correlation, DGE analysis and heatmap clustering), and **(D)** Gene or Protein Bioinformatics Analysis.

an HTML report containing all the selected plots. The user can then download the report as a PDF.

Data Analytics Features

GeneCloudOmics accepts both gene expression matrix from RNA-Seq and raw microarray CEL file formats, either through data upload forms or via direct import from GEO database. Examples of valid input files are hyperlinked at each upload section to aid the user with the input files.

For RNA-Seq, two input files are required: 1) gene expression matrix, and 2) metadata table. The gene expression matrix should contain estimated abundance (either raw count or normalized) of all genes for all samples in the experiment; and the metadata table should specify experimental conditions (e.g., Control, Treated, etc.) for each sample listed in the expression matrix. Depending on target analysis, the user can upload supporting files including gene length and list of negative control genes to facilitate the pre-processing step.

For microarray, the user can upload CEL files to GeneCloudOmics, upon which matrix of gene expression level will be extracted and the user can proceed to subsequent analyses. The data obtained directly from GEO database will undergo an initial exploratory analysis that overviews the quality of data using several plots.

Next, the transcriptomics data is processed and analyzed using the following analytics:

1- Data preprocessing: Preprocessing includes two steps: 1) low-expression gene filtering, and 2) data normalization. Removal of lowly expressed genes is crucial to reduce the effects of measurement noise, and consequently improve the number of differentially expressed genes (Sha et al., 2015). GeneCloudOmics provides the option for the user to indicate the minimum expression value and the minimum number of samples that are required to exceed the threshold for each gene. If input data contain raw read counts, the user can choose one of the normalization options: Fragments Per Kilobase Million (FPKM), Reads Per Kilobase Million (RPKM), Transcripts Per Kilobase Million (TPM) (Li et al., 2015), Remove Unwanted Variation (RUV) (Risso et al., 2014) or Upper Quartile (Bullard et al., 2010). FPKM, RPKM and TPM option perform normalization for sequencing depth and gene length, whereas RUV and upper quartile eliminate the unwanted variation between samples. To check for sample variation, Relative Log Expression (RLE) plots (Gandolfo and Speed, 2018) of input and processed data are displayed for comparison.

2- Transcriptome-wide distributions: Gene expressions are known to follow certain statistical distributions such as power-law or lognormal (Furusawa and Kaneko, 2003; Bengtsson et al., 2005; Beal, 2017), which has been applied to determine a suitable gene expression threshold for low signal-to-noise expression cut-off (Piras et al., 2014, 2019; Piras and Selvarajoo, 2015; Simeoni et al., 2015; Bui and Selvarajoo, 2020). GeneCloudOmics can compare the cumulative distribution function (CDF) of transcriptome-wide expression with six model distributions: Log-normal, Log-logistic, Pareto (or power law), Burr, Weibull, and Gamma. The goodness-of-fit for each distribution is measured by the Akaike information criterion (AIC), from which the user can choose the best-fitted distribution and select threshold for low-expression gene removal.

3- Scatter plot: Scatter plot compares any two samples (or two replicates) by displaying the respective expression of all genes in 2D space. As gene expression data is densely distributed in the low-expression region, making the scatter dots indistinguishable, GeneCloudOmics also overlays the estimated 2D kernel density on the scatter to better visualize the scatter dot density. The scatter plot also shows how variable the gene expressions are between any two samples. The wider the scatter, the less similar the global responses and vice-versa (Piras et al., 2014).

4- Pearson and Spearman correlations: GeneCloudOmics can evaluate the transcriptome-wide relationship between any two samples by linear (Pearson) and monotonic non-linear (Spearman) correlations, displayed in 1) actual values in a table or 2) as a heat map.

5- Principal components analysis and sample clustering: Principal Components Analysis (PCA) is used for simplifying the high-dimensional gene expression data into two or more dimensions, termed the principal components. Doing so, the whole transcriptome data can be visualized on a 2D or 3D plot. Each principal component is a linear combination of the original variables, hence, we can ascribe meaning to what the components represent. From the principal components, GeneCloudOmics can cluster the samples into groups based on their similarity by *K*-means clustering.

6- t-distributed stochastic neighbour embedding (t-SNE): t-SNE is another dimensionality-reduction approach that reduces the complexity of transcriptomic data (Cieslak et al., 2020). GeneCloudOmics introduces an intuitive interface that allows performing t-SNE analysis on the processed untransformed transcriptomic. The user can also choose to log transform the data before submission. Sample clustering by *K*-means is also applied on the t-SNE transformed dataset upon user selection.

7- Shannon entropy: GeneCloudOmics adopts the formula of Shannon entropy (Shannon, 1948) from information theory to measure the disorder of a high-dimensional gene expression sample, where a higher value indicates higher disorder. As the original formula for entropy is restricted to discrete variables, GeneCloudOmics has to discretize gene expression data (which is a continuous variable) by histogram-based binning; the number of bins are determined by Doane's rule (Doane, 1976; Piras et al., 2014).

8- Averaged transcriptome-wide noise: Averaged transcriptome-wide noise quantifies the variability between gene expression scatters of all replicates in one experimental condition (Piras et al., 2014). The noise is defined as the average of variance (σ^2) of expression divided by the square mean expression (μ^2), for all genes between all possible pairs of replicates (Piras et al., 2014).

9- Differential Expression (DE) Analysis: DE analysis identifies genes that are statistically different in expression levels between any two selected conditions. GeneCloudOmics

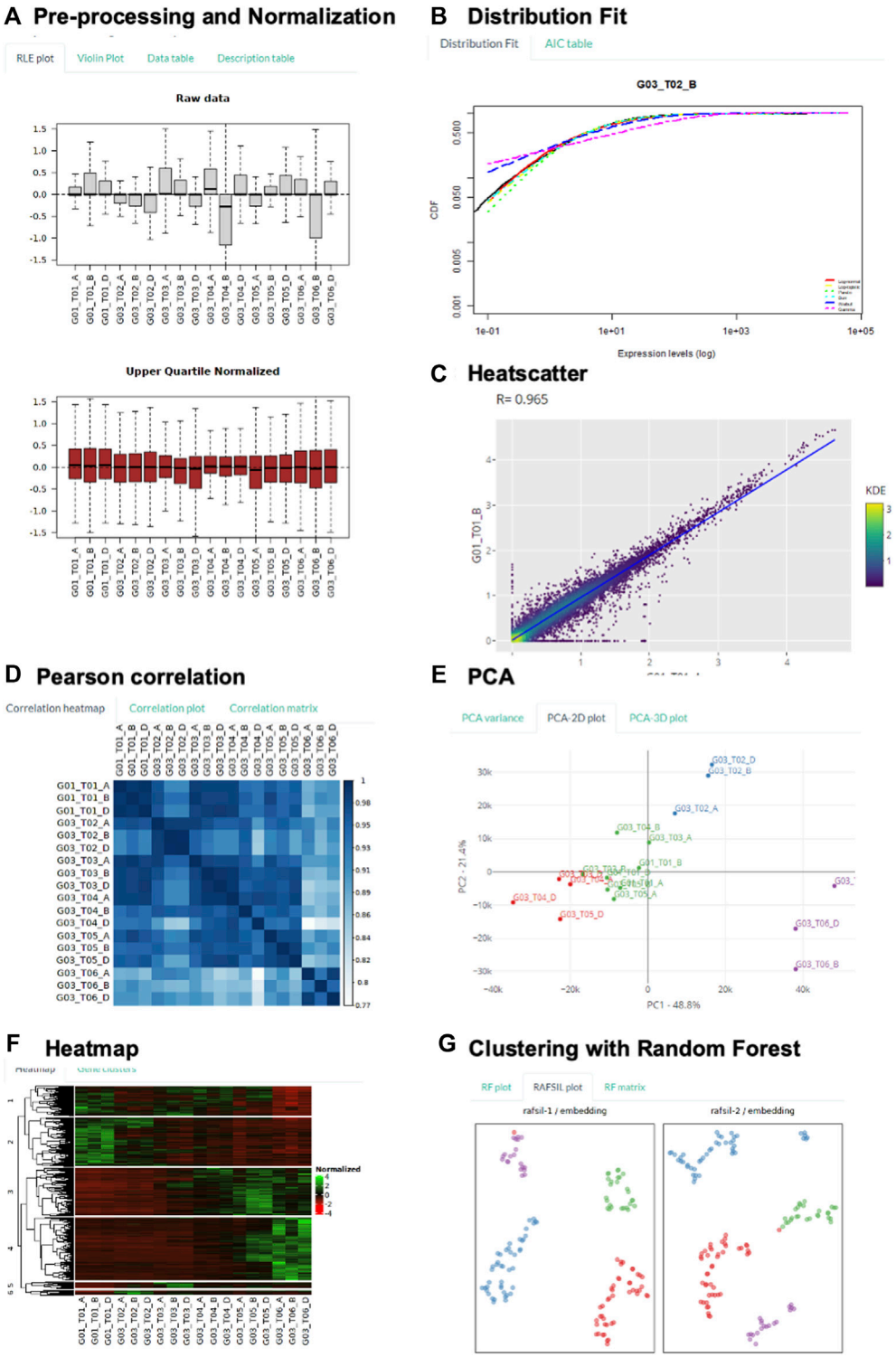

FIGURE 3 | Demonstration of Key Transcriptomic Analysis using GeneCloudOmics. **(A–F)** using bulk-RNASeq human T-regulatory cell differentiation data, and **(G)** using single-cell RNASeq mouse distal lung epithelium data. **(A)**: RLE plot of raw and normalized data, showing sample variation reduced after normalization. **(B)**: Comparing transcriptome-wide distribution with six model distributions to select suitable expression cut-off threshold. **(C)**: Between-replicate transcriptome-wide variation visualized by scatter plot. **(D)**: Pairwise Pearson correlation between all samples. **(E)**: Principal component analysis visualizes all sample data points in 2 dimensions. **(F)**: Hierarchical clustering reveals common expression patterns throughout the T cell differentiation process, visualized by heat map of expression level. **(G)**: Random Forest clustering divides single cells according to their developmental stages.

implements three popular DE methods: edgeR, DESeq2 and NOISeq. In case there are no replicates available for any of the experimental condition, technical replicates can be simulated by NOISeq. To better visualize differentially expressed genes among the others, a volcano plot (plot of $\log_{10}$-p-value and $\log_2$-fold change for all genes) distinguishing the DE and non-DE genes is displayed. Plot of dispersion estimation, which correlates to gene variation, is also available in DESeq2 and EdgeR method.

10- Heatmap and gene clustering: This function clusters differentially expressed genes (result from previous step) into groups of co-varying genes. Expression levels of DE genes first undergo scaling defined by $z_j(p_i) = (x_j(p_i) - \overline{x_j})/\sigma_{x_j}$ where $z_j(p_i)$ is the scaled expression of the j^{th} gene, $x_j(p_i)$ is an expression of the j^{th} gene in sample p_i, $\overline{x_j}$ is the mean expression across all samples and σ_{x_j} is the standard deviation (Simeoni et al., 2015). Subsequently, Ward hierarchical clustering is applied on the scaled expression.

11- Random forest-based clustering: GeneCloudOmics uses RAFSIL (Pouyan and Kostka, 2018), which is a random forest based similarities learning method between single cells from RNA sequencing experiments. RAFSIL utilizes random forest algorithm to learn the pairwise dissimilarity among cells/samples, which in turn is used as an input to the K-means clustering algorithm. The resultant data is subsequently enhanced using t-SNE-reduced dimensions, to reveal clearer clusters of cells/samples.

12- Self-Organizing Map (SOM): SOM is a dimensionality reduction technique that produces a two-dimensional, discretized representation of the high-dimensional gene expression matrix (Yin, 2008). GeneCloudOmics provides a SOM function that outputs five different plots: property plot, count plot, codes plot, distance plot and cluster plot.

Bioinformatics Tools

DGE analysis usually outputs a list of genes that are statistically determined as differentially expressed genes (DEGs). Next, the list of DEGs is analyzed, interpreted, and annotated to learn more about the functions, pathways, and cellular processes where these genes are involved, for example, diseases they are associated with or perform other investigations on the properties of those genes/proteins (such as phylogenetic or physiochemical analyses). Most of the currently available DGE analysis tools do not include bioinformatics features for gene set analysis or include only a few basic analyses such as GO and pathways enrichment (**Table 1**). Even our previous tool, ABioTrans, only provides one GO tool for interpreting the DEGs. In GeneCloudOmics, we redesigned the GO feature to be dynamic by reading the GO terms associated with the genes/proteins directly from UniProt Knowledgebase (Bateman, 2019) then visualize each of the three GO domains (cellular component, molecular function and biological process) in independent tabs. Furthermore, we have introduced 11 new bioinformatics tools that can be performed on a given gene/protein dataset.

1) Pathways Enrichment Analysis: For a given gene or protein set, GeneCloudOmics uses g:Profiler (Raudvere et al., 2019) to perform a pathway enrichment analysis and displays the results as a network where the nodes are the pathways and the edges are the overlap between the pathways (**Figure 3A**). We use Cytoscape. JS for the network visualization (Franz et al., 2015) and through this, the network properties such as colour and layout can be changed and the final network can be downloaded.
2) Protein-Protein Interaction: GeneCloudOmics provides the user with an interface where they can upload a set of proteins (UniProt accessions) and get all the interactions associated with them. The interactions are visualized as a network where the nodes are the proteins, and the edges are the interactions, and the node size corresponds to the number of interactors of the protein. This feature uses Cytoscape. JS for the network visualization (Franz et al., 2015).
3) Complex Enrichment: The identification of the subunits of the protein complexes is important to understand the protein functions and the formation of these macromolecular machines. GeneCloudOmics provides the user with a complex enrichment feature that allows identification of proteins in the provided dataset that are part of a known protein complex using CORUM databases (Giurgiu et al., 2019).
4) Protein Function: UniProt provides a detailed function for thousands of protein sequences. The protein function feature retrieves protein function information from UniProt of a given protein set.
5) Protein Subcellular Localization: Protein localization critically affects a protein function. The protein subcellular localization feature provides the user with an interface to UniProt to get the subcellular localization information for a given list of proteins.
6) Protein Domains: The protein domains are functional subunits of the proteins that contribute to their overall function. GeneCloudOmics provides the user with a protein domain feature that retrieves the domain information from UniProt for a given list of proteins.
7) Tissue Expression: The distinct expression profile of genes and proteins per tissue is what gives different tissues the suitability for their functions. The tissue expression feature in GeneCloudOmics provides the user with tissue expression information from UniProt for each protein in a given list.
8) Gene Co-expression: The co-expression analysis is a common analysis that assesses the expression level of different genes to identify simultaneously expressed genes, which indicates that they are controlled by the same transcriptional mechanism (Vella et al., 2017). GeneCloudOmics provides the user with an interface where they can submit a co-expression query to GeneMANIA (Franz et al., 2018).
9) Protein Physicochemical Properties: For a given set of proteins (UniProt accessions), this feature provides the user with complete sequences of them in a single FASTA file and allows the user to investigate their physicochemical properties, sequence charge, GRAVY index (Kyte and Doolittle, 1982) and hydrophobicity. The full sequences

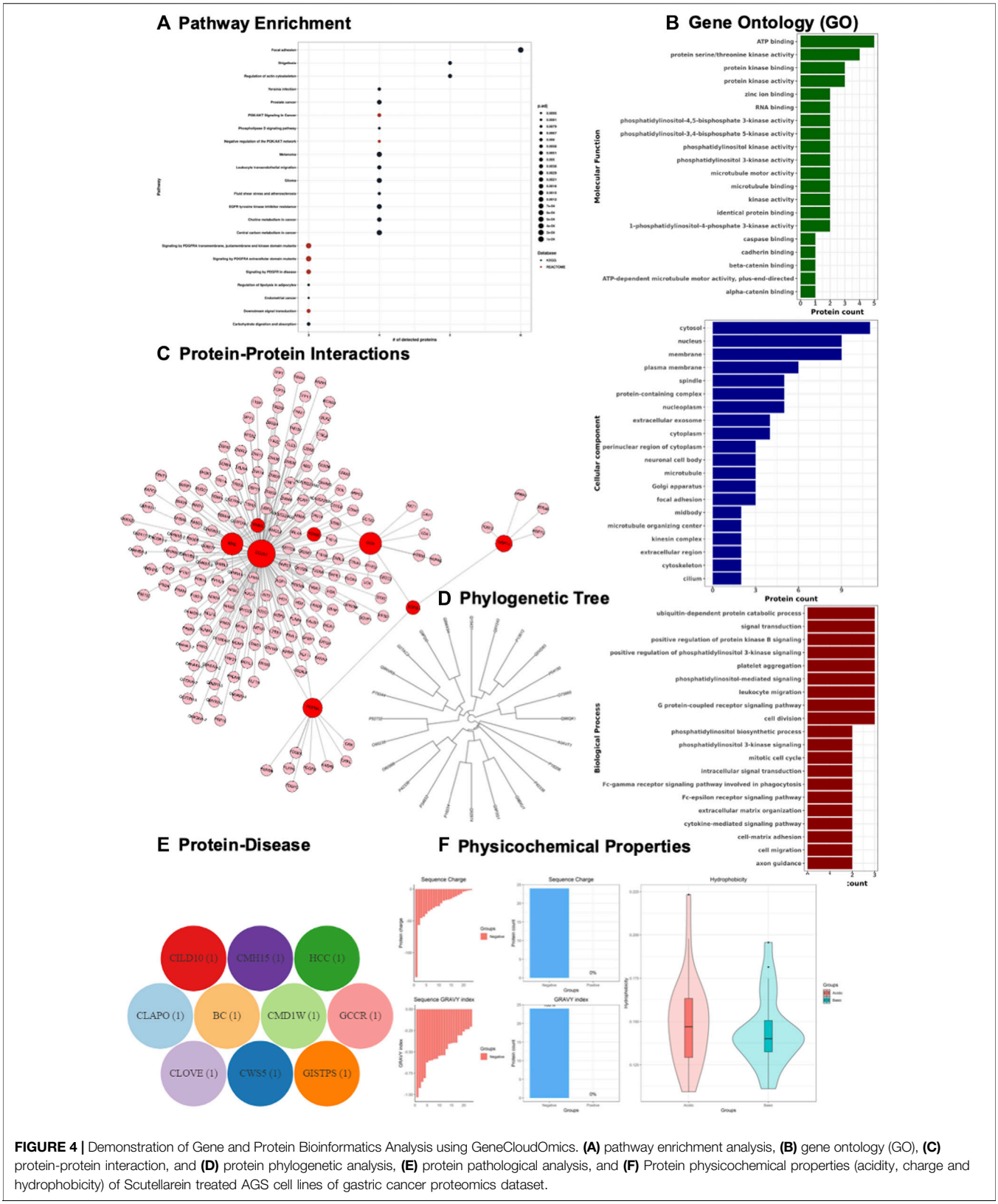

FIGURE 4 | Demonstration of Gene and Protein Bioinformatics Analysis using GeneCloudOmics. **(A)** pathway enrichment analysis, **(B)** gene ontology (GO), **(C)** protein-protein interaction, and **(D)** protein phylogenetic analysis, **(E)** protein pathological analysis, and **(F)** Protein physicochemical properties (acidity, charge and hydrophobicity) of Scutellarein treated AGS cell lines of gastric cancer proteomics dataset.

of the proteins are automatically obtained from UniProt Knowledgebase while the physicochemical properties are investigated and plotted using the UniProtR package (Bateman, 2019; Soudy et al., 2020)

10) Protein Evolutionary Analysis: For a given set of proteins, this feature provides the user with a phylogenetic and evolutionary analysis that includes multiple sequence alignment (MSA) of the protein sequences, clustering based on the amino acid sequences, chromosomal locations, or gene trees.
11) Protein Pathological Analysis: Several diseases are associated with the malfunction of certain genes or proteins. The disease-protein association is collected in different online resources such as OMIM databases (Amberger et al., 2019), DisProt (Hatos et al., 2020) and DisGeNET (Piñero et al., 2020). GeneCloudOmics provides the user with an interface that retrieves the disease-protein association from online databases for a given list of proteins and visualizes disease-protein association as a bubble.

The features that communicate with UniProt use UniProtR, an R package for data retrieval and visualization from UniProt (Soudy et al., 2020). Since all the bioinformatics features only accept gene names (gene symbol) or UniProt accessions, we provide the user on each page with links to two ID converters UniProt ID mapping (Bateman, 2019) and g:Convert (Raudvere et al., 2019) to convert their identifiers to gene names or UniProt accessions. All the analyses are either performed on the uploaded data or involve connecting to a remote server such as UniProt Knowledgebase. GeneCloudOmics does not store any uploaded data and does not contain any databases.

DEMONSTRATION OF GENECLOUDOMICS UTILITY

Transcriptome Analysis Features

We performed a demonstration of transcriptomic analysis with a recent study on the time-resolved bulk cell RNA-Seq profile of human T regulatory cell differentiation (Schmidt et al., 2018). In the study, human T regulatory cells were isolated from peripheral blood; upon which differentiation was induced by adding TGF-ß factor, in comparison to naïve (unstimulated) T regulatory cells as the control group. At the indicated time points (0, 2, 6, 24, 48 h, 6 days), the cells were collected for RNA extraction and sequencing. Here, we illustrate how GeneCloudOmics was used for data pre-processing (normalization and filter low count), differential analysis, and data clustering.

Firstly, unwanted variation among samples was removed by Upper Quartile normalization. The RLE plot clearly illustrates the normalization effects: high between-sample variation in raw data versus low variation after normalizing (**Figure 3A**). We also utilized the transcriptome-wide distribution fitting feature to determine the expression threshold for low count filtering (**Figure 3B**) (Simeoni et al., 2015; Bui et al., 2020). The threshold of five counts was selected because from this expression level onwards, transcriptome-wide expression was observed to follow most of the model statistical distributions.

Next, pairwise scatter, pair-wise sample correlation, and PCA were used to visualize the global relationship of all data samples, through which initial assessment on data quality can be gauged. For example, the low between-replicate variation in contrast with high between-condition variation could be shown by the width of scatter plots (**Figure 3C**, **Supplementary Figure S1A**). It is further illustrated by the correlation heatmap, in which the replicates of the same condition all show close-to-unity Pearson correlation value along the diagonal axis (**Figure 3D**); whereas decreasing correlation value with time was observed along the edge of the heatmap. This information is of high importance because low correlation or high variance across replicates will negatively impact the power to detect differentially expressed genes. Clustering of replicates of similar time points was further illustrated in PCA and t-SNE plots, in which the last time point (T06 - when the T cells were fully differentiated) formed a distinct cluster from the transitioning time points (**Figure 3E**, **Supplementary Figure S1B**). From these analyses, we knew that the data show low variation between replicates and that gene expression globally changed along the differentiation time.

We performed differential expression (DE) analysis with all three supported DE methods: EdgeR, DESeq2 and NOISeq; and presented the analysis conducted with DESeq2 (**Supplementary Figures S1D,E**). The last time point (T06) was compared against the control group (T01) to extract DE genes in the differentiation process (with 0.05 *p*-value and 2-fold expression threshold). Two important steps in DESeq 2 were visualized: 1) the estimation of gene-wise dispersion and empirical shrinkage of these estimates to produce a more accurate dispersion estimate for actual gene count modelling (**Supplementary Figure S1E**); and 2) the volcano plot that summarizes DESeq2 p-value and expression fold difference for every gene (**Supplementary Figure S1E**). The list of all 5,033 differentially expressed genes (3,017 up, 2,016 down) was also listed in a separate table. Finally, the DE genes were channelled into heatmap gene clustering feature, from which DE genes sharing similar patterns of gene expression change throughout the differentiation process were identified (**Figure 3F**). Four common expression patterns were observed: 1) gradual decrease (Group 2), 2) gradual increase (Group 3 and 4), 3) initial increase followed by decrease (Group 5 and 6), and 4) sharp decrease, followed by a gradual increase, and finally decrease (Group 1).

To further illustrate the dimension-reduced visualization features t-SNE and random forest clustering, we used another single-cell RNA-Seq dataset of distal lung epithelium (Treutlein et al., 2014). The study measured gene expression of a total 198 individual mouse lung epithelial cells at four different stages (E14.5, E16.5, E18.5, adult) throughout development. Sample clustering by k-means on t-SNE1 and t-SNE2 space divided the cells into clusters that are aligned with their respective development stages (**Supplementary Figure S1C** and **Additional File S1**): Cluster 1 contains mostly E18 cells, Cluster 2 and 3 contain mostly AT2 cells, Cluster four contains mostly E16 cells, and Cluster 5 contains mostly E14 cells. Finally, clustering by random forest approach (Pouyan and Kostka, 2018) determined the number of clusters as the number

of types of cells provided in the input metadata table and subsequently grouped the cells according to their developmental stages (**Figure 3G**).

New Bioinformatics Features

To demonstrate the utility of the bioinformatics section, we used data from a differential proteomics analysis that was conducted using the AGS cell lines of gastric cancer (GC) (Saralamma et al., 2020). The AGS cells were treated with Scutellarein, a flavone known for its anticancer effect. The study identified 41 proteins that are differentially expressed in AGS when treated with Scutellarein, 24 of them were downregulated and 17 were upregulated.

Pathway analysis shows that the down-regulated proteins are associated with movement of cellular or subcellular components and platelet activation (**Figure 4A**), while that pathways enrichment for the up-regulated proteins did not result in any significantly enriched pathways. Functional analysis is retrieved, visualized, and represented as Gene Ontology (GO) terms (Biological process; Molecular function; Cellular component). The down-regulated profile shows cell processing components including cell cycle, cell division, and cell migration (**Figure 4B**), while the up-regulated profile shows a regulation of apoptotic process including positive and negative regulation associated with cytokine-mediated signalling pathway (**Supplementary Figure S2A**). Protein-protein interaction (PPI) network of both down-regulated and up-regulated proteins retrieved from UniProt (Bateman, 2019) and visualized using Cytoscape. JS (Franz et al., 2015) and GeneCloudOmics protein interaction feature (**Figure 4C**, **Supplementary Figure S2B**).

GeneCloudOmics internally uses ClustalOmega (Sievers and Higgins, 2021) to perform a multiple sequence alignment (MSA) which was used to investigate and visualise the homogeneity among protein sequences (**Figure 4D** and **Supplementary Figure S2C**). Pathological analysis of the protein list is a crucial step in data interpretation for connecting computational output with biological data, so the protein accession list is mapped to OMIM database disease IDs for providing information about diseases associated with proteins (**Figures 4E** and **Supplementary Figure S2D**). Physicochemical analysis of the two sets of proteins shows that sequence charge of 100% of the down-regulated proteins is negative while in the up-regulated proteins it is 94% negative and 6% positive (**Figures 4F** and **Supplementary Figure S2E**).

SUMMARY AND FUTURE DEVELOPMENTS

In this paper, we have introduced a new webserver, GeneCloudOmics, for gene expression data analysis using a simple easy-to-use GUI that contains 23 data analytic and bioinformatics tools. This is the largest number of tools in any current webserver to our knowledge (**Table 1**). We have demonstrated the utility of key functions using recently published human T regulatory cell differentiation and mouse distal lung epithelium RNA-Seq dataset (Risso et al., 2014; Schmidt et al., 2018) and Scutellarein treated AGS cell lines of gastric cancer proteomics dataset (Saralamma et al., 2020).

In the next few years, GeneCloudOmics could be extended to support additional types of high throughput data, on top of RNA-Seq or microarrays. The plan includes supporting the analysis of proteomics, metabolomics, chromatin immunoprecipitation sequencing (ChIP-Seq) and cross-linking immunoprecipitation (CLIP-Seq) data. In addition, we hope to continue improving the transcriptome data analysis by adding new features such as other DGE methods [e.g. Limma (Dias-Audibert et al., 2020) and ScatLay (Bui et al., 2020)], sample overlap analysis (Venn diagram), additional data plots (e.g. density plot) and support for Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005). The gene and protein IDs could also be extended to support different IDs, so the user is not restricted to use gene names and UniProt accessions only.

DATA AVAILABILITY STATEMENT

The GeneCloudOmics web server can be freely accessed at http://combio-sifbi.org/GeneCloudOmics. The software is written using the open-source R programming language (R: a language and environment for statistical computing) and the Shiny framework (Web Application Framework for R [R package shiny version 1.6.0], 2021). A Docker container image is also available (docker pull jaktab/GeneCloudOmics-webserver:latest). GeneCloudOmics is optimized for Google Chrome. Details on the R packages used in GeneCloudOmics, their versions and sources are available in **Supplementary Table S1** and in the tool documentation on GitHub (https://github.com/cbio-astar-tools/GeneCloudOmics).

AUTHOR CONTRIBUTIONS

MH: led development of the software and drafted the manuscript RA: software development MS: software development TB: software development and writing a section of the manuscript KS: conceptualize and led the whole project, wrote the manuscript. All authors read and approved the manuscript.

ACKNOWLEDGMENTS

The authors thank Derek Smith for comments. The web interface was partially funded by BII core budget, IAF-PP grant to SIFBI, A*STAR and partially through the Google Summer of Code (GSoC′20 and ‘21) programs to RA, MS, and JA for the National Resources for Network Biology (NRNB), United States.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2021.693836/full#supplementary-material

Supplementary Figure S1 | Demonstration of GeneCloudOmics transcriptomic analysis features on **(A–E)**: bulk-RNASeq human T regulatory cell differentiation data, and **(C)**: single-cell RNASeq mouse distal lung epithelium data: **(A)**: Between-replicate and between-condition transcriptome wide variation visualized by scatter plot. **(B, C)**: t-SNE (right) visualize all sample data points in 2 dimensions for **(B)**: bulk-cell RNA-Seq human T cell data, and **(C)**: single-cell RNASeq distal lung epithelium data. **(D, E)**: Estimated dispersion **(D)** and resulting volcano plot **(E)** from DESeq2 differential expression analysis with p-value threshold at 0.05 and expression fold threshold at 2.

Supplementary Figure S2 | Protein bioinformatics analysis for the upregulated proteins set. **(A)** gene ontology (GO), **(B)** protein-protein interaction, and **(C)** protein phylogenetic tree, **(D)** protein pathological analysis, and **(E)** Protein physicochemical properties (acidity, charge and hydrophobicity).

Supplementary Table S1 | The list of the R packages used in GeneCloudOmics.

REFERENCES

Amberger, J. S., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2019). OMIM.org: Leveraging Knowledge across Phenotype-Gene Relationships. *Nucleic Acids Res.* 47, D1038–D1043. doi:10.1093/nar/gky1151

Bateman, A. (2019). UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* 47, D506–D515. doi:10.1093/nar/gky1049

Beal, J. (2017). Biochemical Complexity Drives Log-normal Variation in Genetic Expression. *Eng. Biol.* 1, 55–60. doi:10.1049/enb.2017.0004

Bengtsson, M., Ståhlberg, A., Rorsman, P., and Kubista, M. (2005). Gene Expression Profiling in Single Cells from the Pancreatic Islets of Langerhans Reveals Lognormal Distribution of mRNA Levels. *Genome Res.* 15, 1388–1392. doi:10.1101/gr.3820805

Borrill, P., Ramirez-Gonzalez, R., and Uauy, C. (2016). expVIP: a Customizable RNA-Seq Data Analysis and Visualization Platform. *Plant Physiol.* 170, 2172–2186. doi:10.1104/PP.15.01667

Bui, T. T., and Selvarajoo, K. (2020). Attractor Concepts to Evaluate the Transcriptome-wide Dynamics Guiding Anaerobic to Aerobic State Transition in *Escherichia coli*. *Sci. Rep.* 10, 5878–5914. doi:10.1038/s41598-020-62804-3

Bui, T. T., Lee, D., and Selvarajoo, K. (2020). ScatLay: Utilizing Transcriptome-wide Noise for Identifying and Visualizing Differentially Expressed Genes. *Sci. Rep.* 10, 17483–17511. doi:10.1038/s41598-020-74564-1

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments. *BMC Bioinformatics* 11, 94. doi:10.1186/1471-2105-11-94

Chowdhury, H. A., Bhattacharyya, D. K., and Kalita, J. K. (2019). (Differential) Co-expression Analysis of Gene Expression: A Survey of Best Practices. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 17, 1. doi:10.1109/TCBB.2019.2893170

Cieslak, M. C., Castelfranco, A. M., Roncalli, V., Lenz, P. H., and Hartline, D. K. (2020). t-Distributed Stochastic Neighbor Embedding (T-SNE): A Tool for Eco-Physiological Transcriptomic Analysis. *Mar. Genomics* 51, 100723. doi:10.1016/j.margen.2019.100723

Collado-Torres, L., Nellore, A., Frazee, A. C., Wilks, C., Love, M. I., Langmead, B., et al. (2017). Flexible Expressed Region Analysis for RNA-Seq with Derfinder. *Nucleic Acids Res.* 45, e9. doi:10.1093/NAR/GKW852

CRAN (2021). Web Application Framework for R [R Package Shiny Version 1.6.0]. Available at: https://cran.r-project.org/package=shiny (Accessed March 10, 2021).

Cumbie, J. S., Kimbrel, J. A., Di, Y., Schafer, D. W., Wilhelm, L. J., Fox, S. E., et al. (2011). GENE-Counter: A Computational Pipeline for the Analysis of RNA-Seq Data for Gene Expression Differences. *PLoS One* 6, e25279. doi:10.1371/JOURNAL.PONE.0025279

Dias-Audibert, F. L., Navarro, L. C., de Oliveira, D. N., Delafiori, J., Melo, C. F. O. R., Guerreiro, T. M., et al. (2020). Combining Machine Learning and Metabolomics to Identify Weight Gain Biomarkers. *Front. Bioeng. Biotechnol.* 8, 6. doi:10.3389/fbioe.2020.00006

Doane, D. P. (1976). Aesthetic Frequency Classifications. *Am. Statistician* 30, 181. doi:10.2307/2683757

Emig, D., Salomonis, N., Baumbach, J., Lengauer, T., Conklin, B. R., and Albrecht, M. (2010). AltAnalyze and DomainGraph: Analyzing and Visualizing Exon Expression Data. *Nucleic Acids Res.* 38, W755–W762. doi:10.1093/NAR/GKQ405

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2015). Cytoscape.js: a Graph Theory Library for Visualisation and Analysis. *Bioinformatics* 32, 309–311. doi:10.1093/bioinformatics/btv557

Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G. D., et al. (2018). GeneMANIA Update 2018. *Nucleic Acids Res.* 46, W60–W64. doi:10.1093/nar/gky311

Furusawa, C., and Kaneko, K. (2003). Zipf's Law in Gene Expression. *Phys. Rev. Lett.* 90, 088102. doi:10.1103/PhysRevLett.90.088102

Gandolfo, L. C., and Speed, T. P. (2018). RLE Plots: Visualizing Unwanted Variation in High Dimensional Data. *PLoS One* 13, e0191629. doi:10.1371/journal.pone.0191629

Gardeux, V., David, F. P. A., Shajkofci, A., Schwalie, P. C., and Deplancke, B. (2017). ASAP: A Web-Based Platform for the Analysis and Interactive Visualization of Single-Cell RNA-Seq Data. *Bioinformatics* 33, 3123–3125. doi:10.1093/BIOINFORMATICS/BTX337

GBIF (2021). R: A Language and Environment for Statistical Computing. Available at: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing (Accessed March 10, 2021).

Ge, S. X., Son, E. W., and Yao, R. (2018). iDEP: an Integrated Web Application for Differential Expression and Pathway Analysis of RNA-Seq Data. *BMC Bioinformatics* 19, 534–624. doi:10.1186/S12859-018-2486-6

Giguere, D., Macklaim, J., and Gloor, G. (2021). omicplotR: Visual Exploration of Omic Datasets Using a Shiny App. Available at: https://bioconductor.org/packages/release/bioc/html/omicplotR.html (Accessed September 26, 2021).

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., et al. (2019). CORUM: the Comprehensive Resource of Mammalian Protein Complexes-2019. *Nucleic Acids Res.* 47, D559–D563. doi:10.1093/nar/gky973

Harshbarger, J., Kratz, A., and Carninci, P. (2017). DEIVA: a Web Application for Interactive Visual Analysis of Differential Gene Expression Profiles. *BMC Genomics* 18, 47–55. doi:10.1186/S12864-016-3396-5

Hatos, A., Hajdu-Soltész, B., Monzon, A. M., Palopoli, N., Álvarez, L., Aykac-Fas, B., et al. (2020). DisProt: Intrinsic Protein Disorder Annotation in 2020. *Nucleic Acids Res.* 48, D269–D276. doi:10.1093/nar/gkz975

Helmy, M., Crits-Christoph, A., and Bader, G. D. (2016). Ten Simple Rules for Developing Public Biological Databases. *PLOS Comput. Biol.* 12, e1005128. doi:10.1371/journal.pcbi.1005128

Hodgson, S. H., Muller, J., Lockstone, H. E., Hill, A. V. S., Marsh, K., Draper, S. J., et al. (2019). Use of Gene Expression Studies to Investigate the Human Immunological Response to Malaria Infection. *Malar. J.* 18, 418. doi:10.1186/s12936-019-3035-0

Howe, E. A., Sinha, R., Schlauch, D., and Quackenbush, J. (2011). RNA-Seq Analysis in MeV. *Bioinformatics* 27, 3209–3210. doi:10.1093/BIOINFORMATICS/BTR490

Jensen, T. L., Frasketi, M., Conway, K., Villarroel, L., Hill, H., Krampis, K., et al. (2018). RSEQREP: RNA-Seq Reports, an Open-Source Cloud-Enabled Framework for Reproducible RNA-Seq Data Processing, Analysis, and Result Reporting. *F1000Res* 6, 2162. doi:10.12688/f1000research.13049.2

Jiménez-Jacinto, V., Sanchez-Flores, A., and Vega-Alvarado, L. (2019). Integrative Differential Expression Analysis for Multiple EXperiments (IDEAMEX): A Web Server Tool for Integrated RNA-Seq Data Analysis. *Front. Genet.* 10, 279. doi:10.3389/FGENE.2019.00279

Johnson, B. K., Scholz, M. B., Teal, T. K., and Abramovitch, R. B. (2016). SPARTA: Simple Program for Automated Reference-Based Bacterial RNA-Seq Transcriptome Analysis. *BMC Bioinformatics* 17, 66–74. doi:10.1186/S12859-016-0923-Y

Kucukural, A., Yukselen, O., Ozata, D. M., Moore, M. J., and Garber, M. (2019). DEBrowser: Interactive Differential Expression Analysis and Visualization Tool for Count Data. *BMC Genomics* 20, 6. doi:10.1186/S12864-018-5362-X

Kyte, J., and Doolittle, R. F. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* 157, 105–132. doi:10.1016/0022-2836(82)90515-0

Langmead, B., Hansen, K. D., and Leek, J. T. (2010). Cloud-scale RNA-Sequencing Differential Expression Analysis with Myrna. *Genome Biol.* 11 (11), R83–R11. doi:10.1186/GB-2010-11-8-R83

Li, P., Piao, Y., Shon, H. S., and Ryu, K. H. (2015). Comparing the Normalization Methods for the Differential Analysis of Illumina High-Throughput RNA-Seq Data. *BMC Bioinformatics* 16, 347. doi:10.1186/s12859-015-0778-7

Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., et al. (2012). RobiNA: a User-Friendly, Integrated Software Solution for RNA-Seq-Based Transcriptomics. *Nucleic Acids Res.* 40, W622–W627. doi:10.1093/NAR/GKS540

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8

Mangul, S., Mosqueiro, T., Abdill, R. J., Duong, D., Mitchell, K., Sarwal, V., et al. (2019). Challenges and Recommendations to Improve the Installability and Archival Stability of Omics Computational Tools. *PLOS Biol.* 17, e3000333. doi:10.1371/journal.pbio.3000333

Manning, J. (2017). Interactive Downstream Analysis with ShinyNGS. Available at: https://rawgit.com/pinin4fjords/shinyngs/master/inst/doc/shinyngs.html (Accessed September 26, 2021).

Mantione, K. J., Kream, R. M., Kuzelova, H., Ptacek, R., Raboch, J., Samuel, J. M., et al. (2014). Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. *Med. Sci. Monit. Basic Res.* 20, 138–142. doi:10.12659/MSMBR.892101

Markowetz, F. (2017). All Biology Is Computational Biology. *Plos Biol.* 15, e2002050. doi:10.1371/journal.pbio.2002050

McDermaid, A., Monier, B., Zhao, J., Liu, B., and Ma, Q. (2019). Interpretation of Differential Gene Expression Results of RNA-Seq Data: Review and Integration. *Brief. Bioinform.* 20, 2044–2054. doi:10.1093/bib/bby067

Monier, B., McDermaid, A., Wang, C., Zhao, J., Miller, A., Fennell, A., et al. (2019). IRIS-EDA: An Integrated RNA-Seq Interpretation System for Gene Expression Data Analysis. *PLOS Comput. Biol.* 15, e1006792. doi:10.1371/JOURNAL.PCBI.1006792

Nelson, J. W., Sklenar, J., Barnes, A. P., and Minnier, J. (2017). The START App: a Web-Based RNAseq Analysis and Visualization Resource. *Bioinformatics* 33, 447–449. doi:10.1093/BIOINFORMATICS/BTW624

Nussbaumer, T., Kugler, K. G., Bader, K. C., Sharma, S., Seidel, M., and Mayer, K. F. (2014). RNASeqExpressionBrowser--a Web Interface to Browse and Visualize High-Throughput Expression Data. *Bioinformatics* 30, 2519–2520. doi:10.1093/BIOINFORMATICS/BTU334

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level Expression Analysis of RNA-Seq Experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* 11, 1650–1667. doi:10.1038/nprot.2016.095

Pimentel, H., Bray, N. L., Puente, S., Melsted, P., and Pachter, L. (2017). Differential Analysis of RNA-Seq Incorporating Quantification Uncertainty. *Nat. Methods* 14 (14), 687–690. doi:10.1038/nmeth.4324

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update. *Nucleic Acids Res.* 48, D845–D855. doi:10.1093/nar/gkz1021

Piras, V., and Selvarajoo, K. (2015). The Reduction of Gene Expression Variability from Single Cells to Populations Follows Simple Statistical Laws. *Genomics* 105, 137–144. doi:10.1016/j.ygeno.2014.12.007

Piras, V., Tomita, M., and Selvarajoo, K. (2014). Transcriptome-wide Variability in Single Embryonic Development Cells. *Sci. Rep.* 4, 7137–7139. doi:10.1038/srep07137

Piras, V., Chiow, A., and Selvarajoo, K. (2019). Long-range Order and Short-range Disorder in *Saccharomyces cerevisiae* Biofilm. *Eng. Biol.* 3, 12–19. doi:10.1049/enb.2018.5008

Poplawski, A., Marini, F., Hess, M., Zeller, T., Mazur, J., and Binder, H. (2016). Systematically Evaluating Interfaces for RNA-Seq Analysis from a Life Scientist Perspective. *Brief. Bioinform.* 17, 213–223. doi:10.1093/bib/bbv036

Pouyan, M. B., and Kostka, D. (2018). Random forest Based Similarity Learning for Single Cell RNA Sequencing Data. *Bioinformatics* 34, i79–i88. doi:10.1093/bioinformatics/bty260

Powell, D. (2015). An Interactive Web-Tool for RNA-Seq Analysis (v3.2.0). *GitHub Repository*. Available at https://github.com/drpowell/degust/tree/v3.2.0.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). G:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update). *Nucleic Acids Res.* 47, W191–W198. doi:10.1093/nar/gkz369

Reyes, A. L. P., Silva, T. C., Coetzee, S. G., Plummer, J. T., Davis, B. D., Chen, S., et al. (2019). GENAVi: a Shiny Web Application for Gene Expression Normalization, Analysis and Visualization. *BMC Genomics* 20, 745. doi:10.1186/S12864-019-6073-7

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples. *Nat. Biotechnol.* 32, 896–902. doi:10.1038/nbt.2931

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616

Russo, F., and Angelini, C. (2014). RNASeqGUI: a GUI for Analysing RNA-Seq Data. *Bioinformatics* 30, 2514–2516. doi:10.1093/bioinformatics/btu308

Saralamma, V. V. G., Vetrivel, P., Lee, H. J., Kim, S. M., Ha, S. E., Murugesan, R., et al. (2020). Comparative Proteomic Analysis Uncovers Potential Biomarkers Involved in the Anticancer Effect of Scutellarein in Human Gastric Cancer Cells. *Oncol. Rep.* 44, 939–958. doi:10.3892/or.2020.7677

Schmidt, A., Marabita, F., Kiani, N. A., Gross, C. C., Johansson, H. J., Éliás, S., et al. (2018). Time-resolved Transcriptome and Proteome Landscape of Human Regulatory T Cell (Treg) Differentiation Reveals Novel Regulators of FOXP3. *BMC Biol.* 16, 47. doi:10.1186/s12915-018-0518-3

Schultheiss, S. J. (2011). Ten Simple Rules for Providing a Scientific Web Resource. *Plos Comput. Biol.* 7, e1001126. doi:10.1371/journal.pcbi.1001126

Sha, Y., Phan, J. H., and Wang, M. D. (2015). "Effect of Low-Expression Gene Filtering on Detection of Differentially Expressed Genes in RNA-Seq Data," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, Milan, Italy, March 31, 2015 (Institute of Electrical and Electronics Engineers Inc.), 6461–6464. doi:10.1109/EMBC.2015.7319872

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 27, 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Sievers, F., and Higgins, D. G. (2021). The Clustal Omega Multiple Alignment Package. *Methods Mol. Biol.* 2231, 3–16. doi:10.1007/978-1-0716-1036-7_1

Simeoni, O., Piras, V., Tomita, M., and Selvarajoo, K. (2015). Tracking Global Gene Expression Responses in T Cell Differentiation. *Gene* 569, 259–266. doi:10.1016/j.gene.2015.05.061

Soneson, C. (2014). compcodeR--an R Package for Benchmarking Differential Expression Methods for RNA-Seq Data. *Bioinformatics* 30, 2517–2518. doi:10.1093/BIOINFORMATICS/BTU324

Soudy, M., Anwar, A. M., Ahmed, E. A., Osama, A., Ezzeldin, S., Mahgoub, S., et al. (2020). UniprotR: Retrieving and Visualizing Protein Sequence and Functional Information from Universal Protein Resource (UniProt Knowledgebase). *J. Proteomics* 213, 103613. doi:10.1016/j.jprot.2019.103613

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA Sequencing: the Teenage Years. *Nat. Rev. Genet.* 20, 631–656. doi:10.1038/s41576-019-0150-2

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., et al. (2015). Big Data: Astronomical or Genomical? *PLOS Biol.* 13, e1002195. doi:10.1371/journal.pbio.1002195

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., et al. (2015). Data Quality Aware Analysis of Differential Expression in RNA-Seq with NOISeq R/Bioc Package. *Nucleic Acids Res.* 43, e140. doi:10.1093/nar/gkv711

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., et al. (2014). Reconstructing Lineage Hierarchies of the Distal Lung Epithelium Using Single-Cell RNA-Seq. *Nature* 509, 371–375. doi:10.1038/nature13173

Vella, D., Zoppis, I., Mauri, G., Mauri, P., and Di Silvestre, D. (2017). From Protein-Protein Interactions to Protein Co-expression Networks: a New Perspective to Evaluate Large-Scale Proteomic Data. *EURASIP J. Bioinform Syst. Biol.* 2017, 6. doi:10.1186/s13637-017-0059-z

Velmeshev, D., Lally, P., Magistri, M., and Faghihi, M. A. (2016). CANEapp: A User-Friendly Application for Automated Next Generation Transcriptomic Data Analysis. *BMC Genomics* 17, 49. doi:10.1186/s12864-015-2346-y

Wang, Y., Mehta, G., Mayani, R., Lu, J., Souaiaia, T., Chen, Y., et al. (2011). RseqFlow: Workflows for RNA-Seq Data Analysis. *Bioinformatics* 27, 2598–2600. doi:10.1093/BIOINFORMATICS/BTR441

Wang, Y., Mashock, M., Tong, Z., Mu, X., Chen, H., Zhou, X., et al. (2020). Changing Technologies of RNA Sequencing and Their Applications in Clinical Oncology. *Front. Oncol.* 10, 447. doi:10.3389/fonc.2020.00447

Yang, X., Kui, L., Tang, M., Li, D., Wei, K., Chen, W., et al. (2020). High-Throughput Transcriptome Profiling in Drug and Biomarker Discovery. *Front. Genet.* 11, 19. doi:10.3389/fgene.2020.00019

Yin, H. (2008). The Self-Organizing Maps: Background, Theories, Extensions and Applications. *Stud. Comput. Intell.* 115, 715–762. doi:10.1007/978-3-540-78293-3_17

Zheng, H. Q., Wu, N. Y., Chow, C. N., Tseng, K. C., Chien, C. H., Hung, Y. C., et al. (2017). EXPath Tool-A System for Comprehensively Analyzing Regulatory Pathways and Coexpression Networks from High-Throughput Transcriptome Data. *DNA Res.* 24, 371–375. doi:10.1093/DNARES/DSX009

Zhou, G., Soufan, O., Ewald, J., Hancock, R. E. W., Basu, N., and Xia, J. (2019). NetworkAnalyst 3.0: a Visual Analytics Platform for Comprehensive Gene Expression Profiling and Meta-Analysis. *Nucleic Acids Res.* 47, W234–W241. doi:10.1093/NAR/GKZ240

Zou, Y., Bui, T. T., and Selvarajoo, K. (2019). ABioTrans: A Biostatistical Tool for Transcriptomics Analysis. *Front. Genet.* 10, 499. doi:10.3389/fgene.2019.00499

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.