Universitat Autònoma de Barcelona

Development of a bioinformatics platform for analysing big data from Omics analyses-Omnalysis

Punit Tyagi

THE UNIVERSITY OF VETERINARY MEDICINE AND PHARMACY IN KOŠICE
AND
THE AUTONOMOUS UNIVERSITY OF BARCELONA

DEVELOPMENT OF A BIOINFORMATICS PLATFORM FOR ANALYSING BIG DATA FROM OMICS ANALYSES-OMNALYSIS

Dissertation work

Program of study: Biology

Field of study: neuroscience

Workplace: Laboratory of biomedical microbiology and immunology, University of veterinary medicine and pharmacy in Košice, Komenského 73, Košice, Slovakia

Adviser: doc. MVDr. Mangesh Bhide, PhD

Co-adviser: doc. Armand Sánchez Bonastre, PhD

Industrial adviser: Milan Šamaj, MBA

Košice 2021

Punit Tyagi, MSc.

Univerzita veterinárskeho lekárstva a farmácie v Košiciach

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko autora: Mgr. Punit Tyagi
Študijný program: neurovedy (Jednoodborové štúdium, doktorandské III. st., denná forma)
Typ záverečnej práce: Dizertačná práca
Jazyk záverečnej práce: anglický

Názov záverečnej práce:
Development of a bioinformatics platform for analysing big data from Omics analyses – Omnalysis

Názov v sekundárnom jazyku:
Vývoj bioinformačnej platformy pre analýzu veľkých objemov dát z Omics analýz – Omnalysis

Anotácia:
The aim of this project is to develop an integrated bioinformatics platform to understand the biological relevance of the molecules (putative biomarkers) identified from the omics analyses. Next-generation sequencing (NGS) and mass spectrometry techniques are extensively being used to identify the key genes and proteins in the host-pathogen interactions. The big data generated from these machines required multiple steps of computational intensive processing that is a bottleneck for the biologist with minimum programming skills. Developing a bioinformatics web application is the most appropriate approach to fill the gap between big data analysis and biological insight. To this, we developed a bioinformatics web application OMnalysis, that will help researchers to achieve the most accurate biological knowledge in less time and effort.

Školiteľ: Doc. MVDr. Mangesh Ramesh Bhide, PhD.

Školiace pracovisko: ÚIMUNO - Ústav imunológie

Vedúci pracoviska:

Dátum schválenia: 01.01.2019

podpis školiteľa záverečnej práce

podpis študenta

Acknowledgement

First, I would like to express my heartfelt appreciation to my advisors Dr Mangesh Bhide, Dr Armand Sánchez Bonastre and Mr Milan Šamaj for providing me with the opportunity to pursue my career in bioinformatics tool development. I am earnestly thankful to Dr Mangesh Bhide for his consistent guidance and spending endless hours proofreading my manuscripts and dissertation.

I would like to thank entire LBMI team, especially MSc. Amod Kulkarni, PhD., Mgr. Ľuboš Čomor, PhD., RNDr. Zuzana Tkáčová, RNDr. Patrícia Petroušková, PhD., MVDr. Evelína Mochnáčová, PhD., and Ing. Viera Kopčáková for their assistance during the dissertation.

I would like to acknowledge European Union's Horizon 2020 research and innovation programme H2020-MSCA- ITN-2017- EJD: Marie Skłodowska-Curie Innovative Training Networks (European Joint Doctorate) – Grant agreement nº: 765423 – MANNA for providing financial and technical support during my entire doctoral studies.

I would like to thank my parents for their encouragement, support and patience. I want to dedicate this PhD to my mother late Mrs Seema Tyagi, without her support, I would not be where I am today. On a personal note, I would like to thank my wife and my daughters for their continuous assistance and sacrifices, without them this dissertation would not be a reality.

Abstract

In the past decade, RNA sequencing and mass spectrometry-based quantitative approaches are being used commonly to identify the differentially expressed biomarkers in different biological conditions. Data generated from these approaches come in different sizes (e.g., count matrix, normalized list of differentially expressed biomarkers, etc.) and shapes (e.g., sequences, spectral data, etc.). The list of differentially expressed biomarkers is used for functional interpretation and retrieve biological meaning, however, it requires moderate computational skills. Thus, researchers with no programming expertise find difficulty in data interpretation. Several bioinformatics tools are available to analyze such data, however, they are less flexible to perform the multiple steps of visualization and functional interpretation. We developed an easy-to-use shiny based web application (named as OMnalysis) that provides users with a single platform to analyze and visualize the differentially expressed data. The OMnalysis accepts the data in tabular form from *edgeR*, *DESeq2*, *MaxQuant Perseus*, *R packages*, and other similar software, which typically contains the list of differentially expressed genes or proteins, log of the fold change, log of the count per million, the *P* value, q-value, etc. The key features of the OMnalysis are multiple image type visualization and their dimension customization options, seven multiple hypothesis testing correction methods to get more significant gene ontology, network topology-based pathway analysis, and multiple databases support (*KEGG*, *Reactome*, *PANTHER*, *biocarta*, nature pathway interaction database - *NCI*, *PharmGKB* and *STRINGdb*) for extensive pathway enrichment analysis. OMnalysis also fetches the literature information from PubMed to provide supportive evidence to the biomarkers identified in the analysis. In nutshell, we present the OMnalysis as a well-organized user interface, supported by peer-reviewed R packages with updated databases for quick interpretation of the differential transcriptomics and proteomics data to biological meaning.

We believe that the web application OMnalysis developed here allow researchers to extract all possible biological knowledge from the quantitative differential Omics data generated using high-throughput technologies. The OMnalysis web application is freely available at **http://lbmi.uvlf.sk/omnalysis.html** or **https://omnalysis.shinyapps.io/OMnalysis/**.

Contents

List of figures

List of tables

List of abbreviations

ddNTPs	Dideoxynucleoside triphosphate
DNA	Deoxyribonucleic acid
FAIR	Findable, Accessible, Interoperable and Re-usable
MOPED	Multi-Omics Profiling Expression Database
CCD	Coupled-charge diode
SOLiD	Supported oligonucleotides ligation and detection
SMRT	Single-molecule real-time sequencing
HGP	Human Genome project
SNVs	Single nucleotide variations
GWAS	Genome-wide association studies
ELAND	Efficient large-scale Alignment of Nucleotide Databases
BWA	Burrows-Wheeler-Aligner
SNP	Single nucleotide polymorphisms
VEP	Variant Effect Predictor
IGV	Integrative Genomics Viewer
FWER	Familywise Error Rate
FDR	False Discovery rate
PCA	Principal Component Analysis
COA	Correspondence analysis
MCIA	Multiple coinertia analysis
PTA	Partial triadic analysis
SNPs	Single nucleotide polymorphism
ABI	Applied Biosystems
SAGE	Serial analysis of gene expression
CAPE	Cap analysis of gene expression
PE	Paired-end
SE	Single-end
RSEM	RNA-Seq by expectation and maximization
MS	Mass spectrometry
SRM	Selected Reaction Monitoring
MRM	Multiple Reaction Monitoring

PRM	Parallel Reaction Monitoring
DIA	Data-Independent Acquisition
SC	Spectral count
PPA	Peptide peak attribute
LC	Label-free quantification of complex
RIBAR	Robust intensity-based averaged ratio
HPRD	Human protein reference database
NGS	Next-generation sequencing
HGNC	HUGO gene nomenclature committee
NTA	Network topology analysis
SPIA	Signaling pathway impact analysis
GO	Gene ontology
KEGG	Kyoto encyclopedia of genes and genome
NCI	Nature pathway interaction database
EMBL-EBI	European Molecular Biology Laboratory's European Bioinformatics Institute
NCBI	National Center for Biotechnology Information
SBML	Systems Biology Markup Language
DEGs	Differentially expressed genes

1 Introduction

Next-generation sequencing (NGS) and mass spectrometry technologies allow the comprehensive profiling of the regulation of genes, proteins and metabolites. In particular, RNA-Seq uses NGS technology to estimate the number of RNA transcripts (WANG et al., 2009a). In contrast to the microarray, it is hybridization independent and no prior knowledge of probe is required. Due to this *de-novo* approach and higher accuracy, RNA-Seq is widely used in allele-specific expression identification, cancer studies, expression profiling in host-pathogen interaction and many more (LAGARRIGUE et al., 2013; MILANEZ-ALMEIDA et al., 2020; WANG et al., 2009b). The proteomics and metabolomics study requires multiple steps, in which, peptides and metabolites are separated from samples mixture, using liquid chromatography (LC) or gas chromatography (GC). Then, ionized particles are separated based on charge to a mass ratio in mass spectrometry. On the other hand, transcriptomics requires sample preparation, mRNA isolation, library preparation and sequencing. Data generated from these techniques come in all sizes and shapes, including spectral counts and raw read sequences. To shortlist important genes, protein or metabolites from the large data sets require rigorous data analysis approach. Consequently, research laboratories are dependent on the core facility to make sense of the huge wealth of data.

The core facility often provides a differential list of genes, proteins or metabolites. A typical workflow to analyze and visualize differential omics data requires multiple steps and separate tools. Many researchers with limited programming skills face difficulty to learn, tune and link these tools correctly. Another hurdle in the data analysis is scattered and outdated databases with different accession IDs. The lack of user-friendly, updated and scattered bioinformatics tools hinders the processing of differential data into actionable insight. The main aim of this research is to address these issues and develop an integrated bioinformatics platform named OMnalysis.

Data generated from high-throughput techniques require multiple steps analysis approach and the list of differentially expressed genes (DEGs) or proteins often contains the non-informative accession IDs (Ensemble or Entrez ID) that are required to be converted to more familiar IDs (gene name or protein name). Therefore, real-time ID conversion is necessary and it preferably works by fetching the information from the ensemble database (level 1 in data analysis). The *biomaRt* package version 2.46.3 can be used to convert the IDs with the help of an ensemble database (DURINCK et al., 2009).

In level 2 the dimension of the expression data has to be reduced and the best method to perform it is principal component analysis (PCA). Performing PCA helps to understand the cause and effect of treatments on the regulation of genes or proteins in a biological study (MA et al., 2011). Differentially expressed data from quantitative transcriptomics and proteomics have multiple variables (treatments) and observations (genes or proteins) that are required to be analyzed in reduced form, hence variable plots and biplot are generally used (VERHOECKX et al., 2004). The expression relationship between the multiple treatments and genes or proteins is sometimes difficult to understand, therefore PCA computes the variability in each treatment and displays the first few principal components (PCs) without losing much information on variable plot and biplot (KOHLER et al., 2005). The variable plot represents the measure of variables (contribution) that help to understand the association between the treatments and direction of dispersion of the expression data, whereas, biplot PCA displays the characteristics of a variable plot as well as observation (gene or proteins). Also, in biplot, each treatment principal component visualization and comparison are possible (KOHLER and LUNIAK, 2005). However, PC1 and PC2 are capable enough to represent most of the information in terms of relation and variability.

It is necessary to segregate and visualize a huge amount of quantitative transcriptomics and proteomics data based on level of expression (level 3 in data analysis), and that can be performed with scatter and volcano plots. Scatter and volcano plots are able to visualize thousands of differentially expressed genes (LI, 2012), which help to segregate up and down-regulated candidates, and outliers by distributing them on a plot using log fold change value against log count per million value. Scatter plot is also being used in quantitative proteomics to identify the dispersion of abundant proteins using log fold change value of control against treatment, whereas the volcano plot is a type of scatter plot used to identify the significant up and down-regulated genes using log fold change against *P* values (MCDERMAID et al., 2019). This provides quick identification of genes or proteins that have large fold changes.

In the next step, it is necessary to present the differentially expressed common and unique significantly differentially expressed genes or proteins identified among the treatments. The best tool that can be used is the Venn diagram (level 4). In the Venn diagram, the logical relationship between the set of genes or proteins explains their involvement in different treatments, providing the relationship in the expression data.

Functional interpretation is the most important aspect of downstream data analysis (level 5) that depends on the databases and servers, however, frequent updating of these

databases at a certain time interval is recommended. R package *clusterProfiler* version 3.18.1 (YU et al., 2012) supported by *AnnotationDbi* version 1.52.0 (PAGÈS et al., 2020) and species genome-wide annotation databases are updated biannually. These packages identify the set of genes involved in certain gene ontology classes (Biological processes (BP), Molecular function (MF) and Cellular component (CC)). *ClusterProfiler* version 3.18.1 has two algorithms ORA (over-represented analysis) and GSEA (Gene Set Enrichment Analysis) to facilitate the enrichment analysis using the set of genes with the supported biological information (annotation databases) as a background. ORA uses common genes provided in gene set and annotated gene set and then apply the hypergeometric test to reveal the significance of overlaps (HUANG DA et al., 2009). In contrast, GSEA ranked the genes according to the differential expression (logFC) and test whether the annotated gene set is over or under hyped using sum statistics (SUBRAMANIAN et al., 2005). Unfortunately, above mentioned R packages are standalone and require R environment and computational skills to use them.

In the next level, the visualization of differentially expressed molecules identified in gene ontology classes (BP, MF and CC) is required to be compared with their log fold change value among the treatments and the tool that helps in making a biological decision is Heatmap. This heatmap display in a pseudo-colored tabular format that is based on numerical data (ZHAO et al., 2014). Besides functional interpretation, actionable insight of expression data is recommended at the next level of data analysis and the best method in hand is the pathway enrichment analysis (TIPNEY et al., 2010). Pathway databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) (KANEHISA et al., 2000), Reactome (CROFT et al., 2011) and wikipathways (PICO et al., 2008) can be used to perform pathway analysis to identify the evoked pathways by the set of genes present in the treatments. Another strategy to enhance the functional insight is signaling pathway impact analysis (SPIA) that not only consider the enrichment of the pathway but also use the edges and nodes of network topology (TARCA et al., 2009). Adding to this, it also uses genes and their fold change value to identify the significantly affected pathway in the treatment (YAN et al., 2018).

In this study, we integrated the well-established annotation, visualization and statistical open-source R packages from the Bioconductor repository. OMnalysis is a web application with multiples options of statistical methods, visualization and functional profiling. OMnalysis has advantages such as one place to analyze and visualize the

differential expression data, publication-ready images (with multiple image format), availability to visualize up to four treatments gene expression value on heatmap and various pathways (KEGG and Reactome). It also supports R package *graphite* (GRAPH Interaction from pathway Topological Environment) version 1.36.0 (SALES et al., 2012) that uses a more powerful topological and multivariate approach by deriving the pathway topology from Biocarta, NCI, Panther, PharmGKB databases. But these databases are limited to human, drosophila and mouse species. OMnalysis is also supported by *STRINGdb* (Search Tool for the Retrieval of Interacting proteins database) version 2.2.2 (SZKLARCZYK et al., 2019) for the protein-protein interaction networks and functional enrichment analysis using the genes or proteins set in treatments. Moreover, OMnalysis also supports the fetching of published literature in Europe PubMed using Europe PubMed Central RESTful Web Service (*europepmc*) version 0.4 (LEVCHENKO et al., 2018). The above-mentioned collective features make OMnalysis an easy to use integrated web tool over scattered bioinformatics platforms such as Heatmapper (heatmap web-application) (VERHAAK et al., 2006), gene ontology resources (ASHBURNER et al., 2000) (Gene ontology repository) and PaintOmics3 (HERNANDEZ-DE-DIEGO et al., 2018). Additional details about OMnalysis and its applicability using a model example that elucidates the holistic picture of molecular activities involves in host-pathogen interaction are in the thesis.

2 Literature of review

2.1 Origin and generations of sequencing

One of the most important events of the 1950s was the discovery of Watson and Crick three-dimensional structure of DNA that conceptualized the replication and encoding of protein (WATSON et al., 1953). However, the sequence of the four nucleotides (Adenine, Thymine, Guanine and Cytosine) of DNA that may code for proteins was slowed down due to the polynucleotide structure and structural similarities with the RNA molecule (ZALLEN, 2003). In 1965, Holley and Madison targeted pure species of RNA (t-RNA and 5S rRNA) to sequence as they are more stable and abundant in the cell. Soon after, in 1970 Ray Wu determined the sequence of the 186 DNA and λ DNA. Then, Sanger and Coulson sequenced the whole genome of phi X 174 bacteriophage using the plus-minus rapid method. Followed by progressive achievements the first rapid sequencing method was given by Gilbert and Sanger with chemical cleavage and chain termination, respectively (MAXAM et al., 1977; SANGER et al., 1977). In Sanger's method, the di-deoxynucleotide (ddNTPs) lack hydroxyl group (OH) at 3' prime required for extension of the DNA chain, thus resulting in chain termination (SANGER et al., 1977). Contrastingly, the Maxam and Gilbert method involve the incorporation of radioactive phosphate at the 5` ends, adding a combination of chemicals that leads to random chemical degradation at adenine, cytosine, guanine or thymine position.

The advantage of less toxic chemical and less complex procedure Sanger method was further adopted. For more detailed information on the evolution of sequencing methods and sequencers, please refer to the non-current content article published in Sciendo (TYAGI et al., 2020).

2.2 Genomics and data analysis

2.2.1 Application of NGS in genomics

Genomics term is used to study the complete set of organism's genes (complete DNA) using high-end bioinformatics and mathematical models. The first draft of the human genome was sequenced using the automated high throughput ABI prism 3700 DNA analyzer (VENTER et al., 2001). To reduce the time and cost of sequencing shotgun sequencing method was introduced, in which, the genome or random DNA strand first fragmented into pieces and then sequenced individually. During the advancement of the technology more rapid, cost-effective and parallel sequencing techniques was introduced and named next-generation sequencing (NGS). In genomics, the NGS technique can be applied to whole-

exome (coding region) sequencing, detection of variants, disease-causing mutation, parental testing and others (DILLIOTT et al., 2018; LOHMANN et al., 2014). However, applying NGS short-read method in genomics study is sometimes challenging as repetitive sequences in the genome are longer than the reads and intense GC regions cause miss-assemblies, biases and gaps (DIJK et al., 2018).

NGS technologies also contributed to an entire genome scanning to identify the genetic loci associated with disease or trait, technically called Genome-wide association studies (GWAS) (MACARTHUR et al., 2017). NGS approach can be applied to large populations, ethnicity and race to reveal both common and rare variants associated with the disease. Before NGS, array technology was used in GWAS to map the genomic loci and to identify the level of variation in gene regulation knows as expression quantitative trait loci (eQTL) and single nucleotide polymorphism (SNP) or copy number variants (HASIN et al., 2017).

2.2.2 Coverage and depth in whole-genome sequencing

The two important terms in whole-genome sequencing are coverage and depth. The term coverage refers to the average number of times each nucleotide is expected to be sequenced within a given length. On the other hand, depth in sequencing is redundancy of coverage that play an important role in mappability and genome alignment (SIMS et al., 2014). Both the term increases the confidence and reliability in data generated using whole-genome sequencing.

2.3 Bioinformatics tools for genomic data analysis

The analysis of genomics data requires specific bioinformatics tools and pipelines to work with each design of genomic data (KOBOLDT et al., 2013). Genomics data analysis is broadly categorized into four steps. First, nucleotide sequences are associated with quality score and genome assembly. Second, mapping reads to the reference genome. Third, variant calling and annotation. Fourth, visualization of the variants (MICHELE ARAÚJO PEREIRA et al., 2017). The above-mentioned steps are supported by commercially available integrated tools such as *CLC Genomics Workbench* version 21.0 allows to analyze of multiple sequencing platforms, species and workflow of NGS data (**https://digitalinsights.qiagen.com**), *StrandNGS* version 4.0 is powered with NGS data analysis from transcriptomics, epigenomics and genomics data on the genome browser (**https://www.strand-ngs.com/**), whereas, *Partek Genomics suite* is suitable for the analysis of qPCR, microarray and pre-processed NGS data (**https://www.partek.com/partek-**

genomics-suite/). An alternative open-source option to analyze most of the above-mentioned data is the cloud-based bioinformatics platform *Galaxy* (AFGAN et al., 2016). *Galaxy* server provides a comprehensive list of pipelines and their codes to execute the analysis.

To access the second, third, and fourth layer, an ample amount of offline, online open-source tools was developed. The details of the genomics tools are listed in table 1. Several algorithms were developed by the computational biologist to explore each layer of genomics data analysis. These algorithms were slightly different in terms of consuming time and computational power to execute the analysis of the same type of data.

Table 1. List of bioinformatics tools to perform a specific task in genomics data analysis

Genome assembler	**Algorithm**	**Reference**	**Genome assembly**	**Sequencing platform**
SSAKE	Seed and extend	(WARREN et al., 2007)	*de novo*	Sanger, Illumina,
VCAKE	Seed and extend	(JECK et al., 2007)	*de novo*	Illumina
SHARCGS	Seed and extend	(JECK et al., 2007)	*de novo*	Sanger, Illumina,
Edena	Overlap layout consensus (OLC)	(HERNANDEZ et al., 2008)	*de novo*	Illumina (very short reads, e.g., 35bp)
Velvet	Eulerian/*De Bruijin* graph (DBG)	(ZERBINO, 2010)	*de novo*	Sanger, 454, SOLiD, Illumina
ABySS	*De Bruijin* graph (DBG)	(SIMPSON et al., 2009)	*de novo*	SOLiD, Illumina
EULER	Eulerian/*De Bruijin* graph (DBG)	(PEVZNER et al., 2001)	*de novo*	Sanger, 454, SOLiD, Illumina
SOAPdenovo	*De Bruijn* graph (DBG)	(LUO et al., 2012)	*de novo*	Illumina
PE-Assembler	*De Bruijin* graph (DBG) (parallelization)	(ARIYARATNE et al., 2010)	*de novo*	Illumina
Layer-2 (aligner)				
Aligner	**Method**	**Reference**	**Aligner type**	**Input type**
RAGOO (reference guided contig ordering and orienting tool)	K-mer size	(ALONGE et al., 2019)	Reference-based	Fasta reads
BLAT-BLAST like alignment tool	Database indexing	(KENT et al., 2002)	Reference-based	Fasta reads (mRNA, ESTs, etc.)
MAQ-mapping and assembly with qualities	Indexing and measuring probability quality of mapping	(HERNANDEZ et al., 2008)	Reference-based	Paired-End (PE) reads, fasta and fastq files
Bowtie2	FM indexing (Burrows-Wheeler transform based)	(LANGMEAD et al., 2012)	Reference-based	Fastq and fasta file, Paired-End and Single End
BWA-Burrows-	Burrows-Wheeler transform based	(CHANDRAMOULI et al., 2009)	Reference-based	Fastq and fasta file, Paired-End and Single End

Wheeler-Aligner				
Layer-3 (Variant identification and interaction with sequencing data)				
Program	**Method**	**Reference**	**Type**	**Input type**
Samtools (Sequence Alignment/Map (SAM) format	Bayesian approach	(LI et al., 2009)	Alignment interacting and manipulating	Sequence Alignment Map file (SAM), Binary alignment map (BAM file) or CRAM (compressed columnar file format)
VarScan	Heuristic/statistics approach	(KOBOLDT et al., 2009)	Different type of variant detection	Mpileup file
GATK (Genome Analysis ToolKit)	Locus walker strategy and Tree-Reducible interface	(DEPRISTO et al., 2011)	Integrated framework for handling high throughput NGS data	Binary alignment map (BAM file)
ANNOVAR (Annotate variation)	Weighted majority algorithm	(WANG et al., 2010)	Functional annotation of genetic variants	VCF file format and text file
SnpEff	Interval forest algorithm	(CINGOLANI et al., 2012)	Annotation of genetic variants and functional prediction	VCF format file
Ensembl VEP-Variant Effect Predictor	Normalization based allele matching algorithm and SIFT (Sorting Intolerant from Tolerant)	(MCLAREN et al., 2016)	Effect of variants on the gene, transcripts, and protein sequences	VCF format file
Layer- 4 (Visualization)				
Tool	**Online/Offline**	**Reference**	**Type**	**Input type**
IGV-Integrative Genomics Viewer	Both	(THORVALDSDÓTTIR et al., 2013)	Visual exploration of genomics data	Multiple file format options (BAM, Fasta, VCF, etc.)
CIRCOS	Standalone (offline)	(KRZYWINSKI et al., 2009)	Visualization of genomics data and relationships	Multiple plain text files
ViVar	Standalone (offline)	(SANTE et al., 2014)	Processing, analysis and visualization of structural variants data	Multiple file formats (VCF, BAM, etc.)
Geneious (Paid)	Standalone (offline)	**https://www.geneious.com/resources/**	Molecular cloning and primer design, NGS, etc.	Multiple input file format (according to required analysis)

Source: self-made table

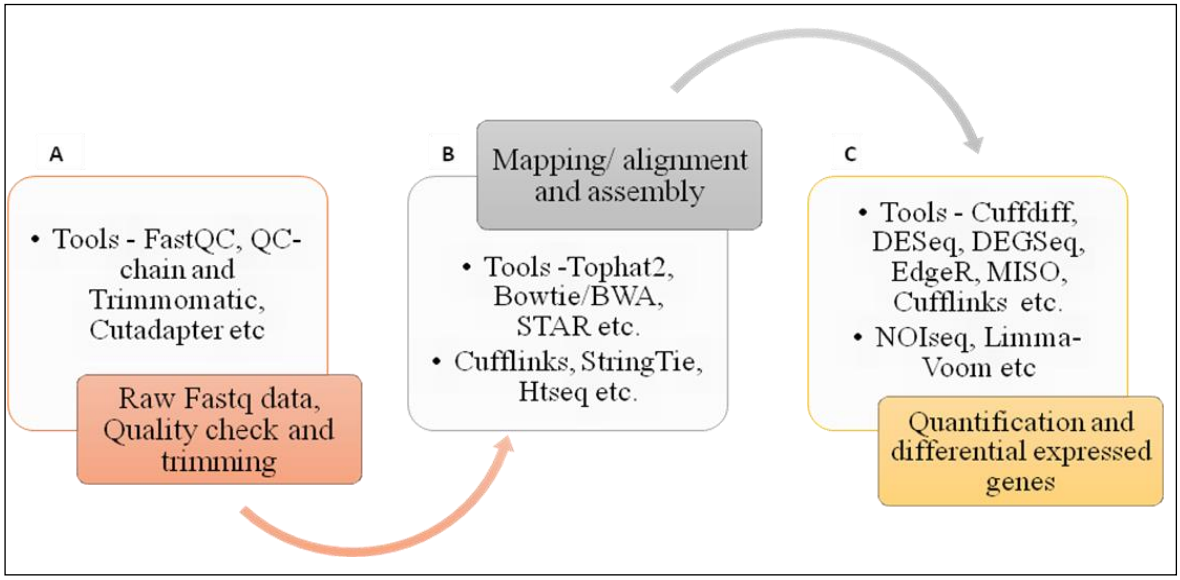
2.4 Transcriptomic data analysis

A Transcriptomics study is used to capture and measure the total transcripts present in a cell in different conditions and time points. Before NGS, microarray technology was used

for expression profiling of genes (FRESE et al., 2013). Microarray was limited in the discovery of novel isoforms, less dynamic range and no *de novo* transcriptomics study. As a result, the RNA-Seq technique took over and was aggressively used in single nucleotide polymorphism (SNPs), fusion genes, splice variants and differentially expression studies (RAPLEE et al., 2019). Also, comparing to other methods such as serial analysis of gene expression (SAGE) and cap analysis of gene expression (CAPE), RNA-Seq provides the quantitative view of gene expression data (KUKURBA et al., 2015).

In RNA-Seq, the common procedure is library preparation and one common step of cDNA reverse transcription from protein-coding mRNA and non-coding RNAs. The RNA library preparation steps may vary according to the sequencing platform and type of samples. To facilitate the alternative spliced isoforms and allele-specific expression sequencers may be used individually or in combination to reveal the novel gene structure (JAZAYERI et al., 2015). The data generated from the sequencers are in the form of fastq files, which can be processed using different pipelines (Figure 1).

Figure 1. Flow chart explaining the workflow of raw sequence analysis to DEGs



A-raw reads undergo quality check, on this basis, the adapters and low-quality reads were removed using above mention tools. B-these reads are mapped to the reference genome and assemble into transcripts using various bioinformatics tools. C-Transcripts quantified according to their statistical significance and differentially expressed genes using given tools

Source: self-made figure

2.4.1 Quality check of the data

The intrinsic limitations in RNA-Seq technology are ribosomal RNA (rRNA) residues, RNA degradation and different read coverage (YANG et al., 2018). For these reasons, it is mandatory to check and process the raw sequence data through a quality control procedure. The raw sequence file is compressed and ranges from approximately 500 megabytes to 3 giga bytes per sample.

To perform the quality check tools such as *FastQC* (provide a quick view about the problematic area in the NGS data) (ANDREWS, 2010), *NGS QC toolkit* (for low quality reads filtering) (PATEL et al., 2012), *FASTX-toolkit* (clipping the reads and separating barcode using the given script in it (**http://hannonlab.cshl.edu/fastx_toolkit/index.html/**). On the other hand, tools *QC-chain* (ZHOU et al., 2013) and *RseQC* (WANG et al., 2012) provides a better view of the identification of contamination and biases in the sequences, respectively. Table 2 shows the other quality control tools for NGS data quality visualization. The most widely used tool among the listed is a *FastQC*, which provides an overview of basic quality control and easy to use graphical interface. *FastQC* is benefitted from multiple input format options Binary Alignment/Map format (BAM), Sequence Alignment/Map format-SAM and FastQ zipped file. It also displays all information regarding reads, adapters used, quality score, overrepresented sequences and GC content etc.

Table 2. List of quality control tools used in the NGS data analysis with descriptions

Tools for quality check	Input formats	Tool description	Link	Citation
FASTQC	FASTQ, GZip compressed FastQ, SAM, BAM	A quality control tool for high throughput sequence data.	**https://www.bioinformatics.babraham.ac.uk/projects/fastqc/**	363 - source Google Scholar
FASTX-Toolkit	FASTA, FASTQ	The FASTX-Toolkit is a collection of command-line tools for Short-Reads FASTA/FASTQ files preprocessing.	**http://hannonlab.cshl.edu/fastx_toolkit/**	282 - source Google Scholar
QC-chain	FASTA, FASTQ	Quality control of next-generation sequencing (NGS) data	**https://omictools.com/qc-chain-tool**	49 - source Google Scholar
NGS QC-toolkit	FASTA, FASTQ	A toolkit for the quality control (QC) of next-generation sequencing (NGS) data	**http://www.nipgr.res.in/ngsqctoolkit.html**	1215 - source Google Scholar
RseQC	FASTA, FASTQ	An RNA-Seq Quality Control Package	**http://rseqc.sourceforge.net/**	600 - source Google scholar

RNASeQc	BAM	RNA-SeqC is a java program that computes a series of quality control metrics for RNA-Seq data	**https://software.broadinstitute.org/cancer/cga/RNA-Seqc**	359 - source Google Scholar

Source: self-made table

2.4.2 Identification of contamination and errors in reads

The second step in RNA-Seq data analysis is the trimming of unwanted data. Another sequencing platform such as ThermoFisher, PacBio, Oxford Nanopore has some constraints and in-depth discussion of each platform is beyond the scope of this thesis. Therefore, this section mainly focuses on highly used Illumina platform works on sequencing by synthesis technique. The need for contamination removal arises due to the use of adapter ligation and low-quality bases which may be called incorrectly by the sequencing machine (MARTIN, 2011). Other contamination includes primer dimers, very tiny DNA and rRNA (KUMAR et al., 2012). Also, in library preparation, random hexamer (primer in the construction of double-stranded cDNA) is used that causes biases in nucleotide arrangement in the initial of the reads (WILLIAMS et al., 2016).

2.4.2.1 Sequencing machine errors

Sequencing errors such as phasing, cluster density and optical detection are responsible for the incorrect base call resulted in incorrect read alignment (FULLER et al., 2009). The phasing error arises in sequencing by synthesis technique when nucleotide with a blocker (terminal cap) did not remove correctly after signal detection, resulting in no new nucleotide binding on this DNA fragment. Cluster density is different from phasing error as it occurs due to the amount of DNA loaded onto the flow cell causing under and over clustering (KUMAR et al., 2012). In contrast to clustering density, signal detection (optical detection) of excited fluorophore attached to the nucleotides affected due to the limitation of photobleaching, photons per sec, camera size (CCD) and read rate (FULLER et al., 2009).

2.4.3 Library preparation in RNA-Seq

RNA-Seq library preparation varies according to the experimental need (small RNA, miRNA, mRNA, etc.). mRNA library preparation is followed by a selection of mRNA or depletion of rRNA, mRNA enrichment, mRNA fragmentation, cDNA synthesis and adapter ligation. Whereas, in miRNA and sRNA library preparation 3` and 5` adapter ligation is performed first (CHAO et al., 2019; DARD-DASCOT et al., 2018). Before sequencing, researchers preferred fragmentation or sizing of the targeted sequences (insert size)

according to the sequencing platform and read length. The resultant fragmented sequence reverse transcribed to double-stranded cDNA, followed by cDNA repair and end polishing. The final step in mRNA library preparation is platform-specific adapter ligation and amplification (HRDLICKOVA et al., 2017; SYED et al., 2009).

2.4.4 Single-end and paired-end RNA-Seq

The experimental design and biological question influenced the option to perform paired-end (PE) or single-end (SE) sequencing. In paired-end sequencing the fragment is sequenced from both ends, increasing the coverage and quality of data. This type of sequencing can be useful in variants detection, repeats in the genomic DNA, gene fusion, and identification of novel transcripts (ANDERS et al., 2015). Contrastingly, in SE sequencing fragment is sequenced from only one end with less cost per base. SE sequencing is widely accepted for the estimation of transcript quantified in different samples (CORLEY et al., 2017).

2.4.5 Contamination trimming in RNA-Seq

To perform the downstream analysis raw reads, need to be trimmed from the above-mentioned contaminations and adapter (Universal adapters of Illumina machine). Table 3 lists the bioinformatics tools and programming scripts used to obtain the sequence of interest for the mapping to reference. Most of the NGS reads pre-processing tools were not designed to work on paired-end data efficiently, therefore, a tool was developed *Trimmomatic* which is widely used for both PE and SE read trimming (BOLGER et al., 2014). This tool is more flexible in trimming the adapter, Illumina specific sequences and converting quality scores to phred 33 or 64 accordingly (DEL FABBRO et al., 2013).

Table 3. Enlisted the popular trimming tools used in NGS data analysis and their elaborated description

Tool for Trimming	Input	Tool description	Link	Cons	Citation
Cutadapt 1.1	Zip file	Python and C based program	**https://cutadapt.readthedocs.io/en/stable/guide.html**	Remove adapter, multi-threaded but not enable by default	4966 - source Google Scholar
Trimmomatic	FASTQ zip file	Trim and crop fastq data, remove the adapter, while not capable to merge the overlap reads to longer single	**http://www.usadellab.org/cms/index.php?page=trimmomatic**	Two modes paired-end and single-end, multithreading	9124 - source Google Scholar

		end reads directly			
AftreQC	Fastq data from HiSeq, Nextseq, Miniseq, Illumina 1.8	Automatic filtering, trimming, error removing	**https://github.com/OpenGene/AfterQC**	Produces three folders, good, bad and QC.	30 - source Google Scholar
Skewer	NGS paired-end sequences	The bit-masked k-difference matching algorithm and able to process Nextera long mate-pair reads (LMP)	**https://sourceforge.net/projects/skewer**	Detection and removal of adapter sequences, multi-threading, Single, paired and long mate-pair reads	281 - source Google Scholar
Trimgalore	FastQ file and compressed as well	Remove biased methylation positions for RRBS sequences file and a wrapper tool for the Cutadapt and FASTQC to increase some functionality, programmed in Perl language	**https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/**	Accept other adapter and 13bp of Illumina standard adapter, after completion directly run FASTQC.	183 - source Google Scholar
BBDuk	Fasta or fastq compressed and uncompressed file	Compare reads to the kmers in a reference dataset and a part of BBTools	**https://github.com/BioInfoTools/BBMap/blob/master/sh/bbduk.sh**	Also, trim the matching parts of the reads rather than binning the reads, histogram regeneration	53 - source Google scholar (BBMAP)

Source: self-made table

2.4.6 Tools for mapping/alignment of NGS data

After rigorous quality check and trimming the next step is mapping or alignment to the reference genome using the appropriate aligner or mapper (FONSECA et al., 2012). The reads are often mapped to the well-annotated genome or transcriptome and the percentage of reads mapped to the genome is termed as mapping quality (CONESA et al., 2016). The higher the mapping quality higher the chance to get significant information.

In genomics, short reads, splicing, paralogous sequences and alignment bias affect the detection of variants and genome alignment, respectively (THANKASWAMY-KOSALAI et al., 2017). Several aligners or mappers (Table 4) were developed that works on hash table algorithms or index-based algorithms and heuristic-based dynamic algorithms. *BWA* and *Bowtie2* are index-based aligners used for mapping long gaps, however, these are slow and

consume more memory compared to heuristic aligners (BLAST-short-read mapping) (THANKASWAMY-KOSALAI et al., 2017).

2.4.6.1 Aligner for transcriptomics data

The three most dominated spliced aligners in transcriptomics data analysis are *Tophat2* (KIM et al., 2013), *STAR* version 2.7 (DOBIN et al., 2013) and *HISAT2* (KIM et al., 2015). Table 4 shows the list of aligners available for genomics and transcriptomics data. Comparing *tophat2* in case of sensitive spliced alignment is much slower than *STAR* aligner (ENGSTROM et al., 2013), however, only *tophat2* provide XS tag (genomic strand), which helps in the evaluation of transcripts reconstruction.

Table 4. Describing the tools for alignment and mapping of NGS data

Tool for aligning/ mapping	Input	Tool description	Link	Cons	Citation
RSeQC	BED, SAM, BAM, Fasta	Check sequence quality, nucleotide composition bias, PCR bias and GC bias	**http://rseqc.sourceforge.net/**	Specific for RNA-Seq, evaluate sequencing saturation, coverage uniformity etc.	600 - source Google Scholar
Qualimap	SAM/BAM	Platform independent application, Java and R based	**http://qualimap.bioinfo.cipf.es/**	Quality control of alignment sequencing data, multi comparison of alignment and count data	301 - source Google Scholar
STAR 2.6.0a	Fastq paired-end	Merging and mapping of overlapping paired-end reads, personal variants overlapping alignments	**https://github.com/alexdobin/STAR**	Can flush ambiguous insertion positions, control the number of sorting bins	6267 - source Google Scholar
Tophat2	Fasta or Fastq	Fast splice junction mapper RNA-Seq reads	**https://github.com/infphilo/tophat**	Ultra-high throughput short read aligner bowtie and align with RNA-Seq, however, it is now under low maintenance	6214 - source Google Scholar
GSNAP	Paired and single-end reads fastq	Can align 14 nucleotides and arbitrarily long length	**https://bio.tools/gsnap**	Detect short and long-distance splicing, including intrachromosomal splicing using the database and probabilistic	1443 - source Google Scholar
PALmapper	Short mRNA sequences	Fast and accurate spliced alignment of sequence reads	**http://ftp.raetschlab.org/software/palmapper**	Combine genomeMapper and QPALMA, Alignment with	70 - source Google Scholar

				mismatches and indels	
Readsmap	Fastq files and de-compressed	Use samtools and TMAP aligner to map single end reads against a reference	**http://www.softberry.com/berry.phtml?topic=readsmap-i&group=help&subgroup=pipelines**	Implement under python, generates input for SHAPEMapper program for single-end reads.	Not traceable - source Google Scholar
HISAT2 2.1.0	Fastq files	Fast and sensitive alignment program for both DNA and RNA with the low memory requirement	**https://ccb.jhu.edu/software/hisat2/index.shtml**	Uses small GFM indexes, accurate alignment of sequencing reads, new hierarchical graph FM index	1724 - source Google Scholar

Source: self-made table

Once the mapping to the reference is completed, the further step is to assemble the mapped reads to the transcript (KUKURBA and MONTGOMERY, 2015). Tools such as *Cufflinks* version 2.2.1 (TRAPNELL et al., 2012), *StringTie* version 2.1.4 (PERTEA et al., 2015) and *HTseq* version 0.13.5 (ANDERS et al., 2015) were used to construct the transcript assembly and calculating their abundance. To perform reference-independent transcriptome assembly *Trinity* version 2.12.0 (GRABHERR et al., 2011), *Oases* version 0.2.08 (very short reads) (SCHULZ et al., 2012) and *transABySS* version 2.0 (ROBERTSON et al., 2010) tools preferably used. *Trinity* is famous for the non-model organism, whereas, *Oases* was developed to deal with the alternative splicing events and repetitive regions using the *velvet* algorithm (YANG et al., 2015; ZERBINO et al., 2008).

2.4.7 Quantification of Transcripts

RNA-Seq quantification is categorized into gene-level quantification and transcript/Isoform level quantification (YANG and KIM, 2015). At the gene level, count summarizes over gene, whereas, transcript level summarizes over transcripts. Tools such as *HT-seq* (ANDERS et al., 2015) or *feature count* can be used for gene-level quantification (LIAO et al., 2013). Also, both tools are supported by a gene transfer format (GTF or GFF) file containing the genome coordinates of exome and genes. Table 5 provides the details of the tools available for quantification.

To compare expression level values among the sample's raw read counts are not sufficient as a total number of reads, whereas, transcript length and sequencing biases normalization is recommended. RPKM (Read per kilobase of exon model per million mapped reads) and FPKM (Fragment per kilobase of exon model per million mapped reads)

methods can be used for the reduction of feature-length and library size effect (WAGNER et al., 2012).

The transcript level quantification detects the changes in the expression of transcript isoforms from the same gene, whereas the alternative splicing algorithm first estimates isoform expression then compared the difference (*Cuffdiff2*) (TRAPNELL et al., 2012). In contrast, the exon-based approach considers the distribution of the reads on the exon and the junction of the gene between the compared samples (*DSGSeq*) (WANG et al., 2013).

The bioinformatics tool *Cufflinks* (TRAPNELL et al., 2012) perform the expression by counting the number of reads that mapped to full-length transcripts at a different time, whereas, *HTseq* quantify the expression and directly count the number of reads that are mapped to an exon without assembling transcripts (ANDERS et al., 2015). Also, the tool *StringTie* compiles transcripts and computes their expression level simultaneously (PERTEA et al., 2015).

Algorithms such as *Sailfish* or *Salmon* works on an alignment-free quantification approach to speed up the process by counting the k-mer and then using only the unique k-mer for the expression analysis (mapping to the transcriptome to find the transcript) (PATRO et al., 2014).

Table 5. Tools available for the quantification of differentially expressed genes

Tools for assembling transcripts and abundance	**Input**	**Tool description**	**Link**	**pros**	**Citation**
Cufflinks	SAM format	Transcriptome assembly and differential expression analysis	**http://cole-trapnell-lab.github.io/cufflinks/releases/v2.2.1/**	It produces three output files, gene expression with FPKM value	6339 - source Google Scholar
RSEM	Single-end and paired-end read data	Maximum likelihood algorithm, bias correction, multithread support	**https://github.com/deweylab/RSEM**	Fast and no indel alignment, estimation of gene and isoform expression level from RNA-Seq data	4988 - source Google Scholar
eXpress	Targeted sequences and a set of sequenced fragments	Quantify the abundance of a set of targeted sequences, based on the online EM algorithm	**https://pachterlab.github.io/eXpress/overview.html#**	Accurately quantify much larger samples than other tools, abundance transcripts in multi-isoforms genes.	534 - source Google Scholar

Kallisto	RNA-Seq data, single-cell RNA-Seq data	The novel idea of pseudo-alignment with standard RNA-Seq data can quantify 30 million human reads in less than 3 minutes	**http://pachterlab.github.io/kallisto//releases/2018/11/17/v0.45.0**	Accurate and fast as existing quantification tools.	1216 - source Google Scholar
Salmon	RNA-Seq read data	Work with both VM/EM algorithm	**https://combine-lab.github.io/salmon/getting_started/#obtaining-salmon**	Free software tool, fast, accurate and bias-aware transcripts quantification.	522 - source Google Scholar
Htseq	Raw Reads, Fasta and Fastq	Linux, Mac, Window, Work with Python of Ht Data, Bed and Vcf Output	**Https://Htseq.Readthedocs.Io/En/Release_0.10.0/**	Handle big data as well, open-source, based on python scripts	5332 - source Google Scholar
Cuffdiff2	Transcript gtf file with SAM file	find expression in each condition, splicing, promoter use	**https://github.com/cole-trapnell-lab/cufflinks**	open-source and accurate for RNA- seq as well, part of tuxedo pipeline,	6339 - source Google scholar (Part of Cufflink package)
DESeq2	Un-normalized matrix or integers value from high throughput experiments, BAM file	Import transcript abundance estimates from various external software, R package	**https://www.bioconductor.org/packages/release/bioc/html/DESeq2.html**	Freely available, uses negative binomial generalized linear models, also work for ChIP-Seq, HiC, shRNA screening and mass spectrometry	8669 - source Google Scholar
edgeR	Tab-delimited file, read count file	Applied to differential expression at the gene, exon, transcript or tag level, R package	**http://bioconductor.org/packages/release/bioc/html/edgeR.html**	Use classic and generalized linear model (glms) empirical Bayes method, estimate gene-specific biological variation, open-source	10374 - source Google Scholar

Source: self-made table

2.4.8 Normalization and differentially expressed genes

The detection and analysis of differentially expressed transcripts is the primary objective of RNA-Seq data analysis. The significant result can be attained using the correct statistical approach and strict error rate control methods. In early RNA-Seq, the technical

replicates studies feature count distribution fitted well to Poisson distribution, however, in biological replicates it underestimates the problem of overdispersion and Poisson distribution does not control the type –I error (HOWE et al., 2011; ZHANG et al., 2014).

In differential expression analysis, the logFC values were compared among the samples and can be layered into three processes. First, estimate dispersion (variance) to determine whether the treatment causes a significant change in the gene expression. Second, test differential expression (log (base2) fold change) and p-value. Third, FDR (multiple hypothesis testing corrections), (LORAINE et al., 2015). The normalization strategies (FPKM, TPM and so on) are performed poorly when the samples have heterogeneous transcript distributions (highly expressed feature). This can be handled using the PoissonSeq, UpperQuartile, TMM (Trimmed mean of M-values) and DESeq normalization methods. Tools to correct the biases between the samples or within the samples are *Cuffdiff2* (KIM et al., 2013), *DESeq* (use negative binomial distribution) (ANDERS et al., 2010), *EdgeR* (ROBINSON et al., 2010), *NOISeq* (TARAZONA et al., 2015), *EBSeq* (AMUNUGAMA et al., 2013), *SAMSeq* (KOTOKA et al., 2017) and other.

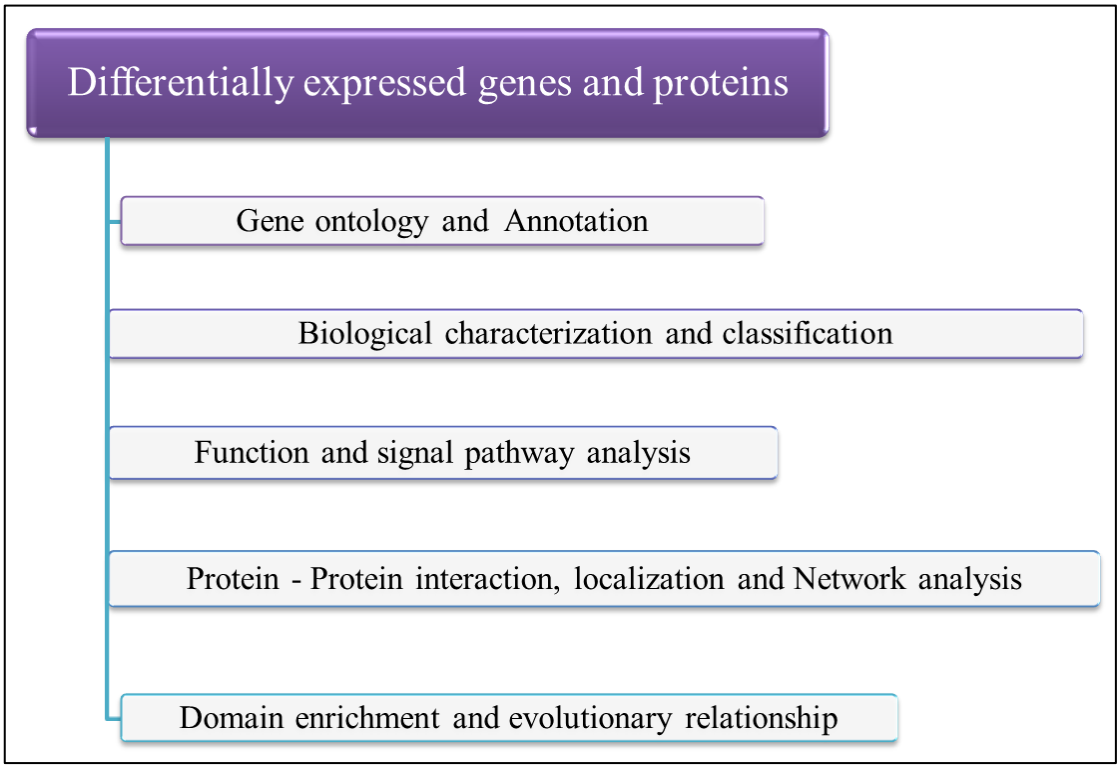
MISO (mixture of isoforms) tool is available for finding differentially regulated isoforms or exons (LEE et al., 2013). A study was performed on some differential expression analysis software and concluded that in terms of precision, accuracy and sensitivity *NOIseq* (TARAZONA et al., 2012), *Limma* (RITCHIE et al., 2015), *Voom* (LAW et al., 2014) and *DESeq2* (LOVE et al., 2014) were more balanced software than others (COSTA-SILVA et al., 2017).

2.5 Gene ontology and pathway analysis

The functional interpretation and annotation of significantly differentially expressed genes involved in the molecular function or pathways is the actionable step in the transcriptomics study (Figure 2). The resources, such as gene ontology databases (HARRIS et al., 2004), DAVID (DENNIS et al., 2003) and Bioconductor and so on, are the rich source of functional annotation. The gene ontology contains the dictionary of the annotation terms in Molecular function (MC), Biological process (BP) and cellular components (CC). DEGs that occur frequently to the specific GO term in the list are termed as over-represented or enriched. Tools that can be used to perform enrichment analysis are the *R Bioconductor* package, *GSEA* (SUBRAMANIAN et al., 2005), *topGO* (ALEXA et al., 2016) and others listed in Table 6. *KEGG* (Kyoto Encyclopedia of gene and genomes) that is rich in curated

molecular pathways and disease signature is another rich source for biological insight (YANG et al., 2018).

Figure 2. Representation of the DEGs/proteins into their functional interpretation



Workflow showing the annotation approaches after getting the DEGs. Pathway analysis of genes and proteins involved in specific biological functions. Finally, the multiple putative genes and proteins are classified according to their interaction and evolutionary relationship.

Source: self-made figure

The protein-coding transcripts can be annotated orthogonally using sequences similarity against the swiss-prot (BAIROCH et al., 2000) database or to the protein domain repositories such as *P-fam* version 33.1 (BATEMAN et al., 2004) and *InterProscan 5* (JONES et al., 2014). Furthermore, the interaction analysis can be performed using the online protein-protein interaction (PPI) tool *STRING* version 11.0 (SZKLARCZYK et al., 2019) that cover 5090 organisms with more than 9,643,763 proteins and 2000 million interactions. Using only experimental evidence interaction from *stringdb* the PPI network between the DEGs can be constructed using *Cytoscape* (SHANNON et al., 2003).

Table 6. Lists of bioinformatics tools are mostly used by the research community for the annotation of the differentially expressed data.

Stage in analysis	**Gene ontology- Biological, Cellular and Molecular process**				
Bioinformatics tools	***Reactome***	***Panther***	***Go Consortium***	***Blast2go***	***Interproscan***
Pros	Highly informative, hierarchy view of biological function, multiple analysis function links	Based on the evolutionary relationship, searched by gene and sequence	Rich knowledge of experimental literature	Standalone, annotation of novel high throughput sequences	Functional analysis of protein and nucleotide sequences, standalone
Cons	Multiple windows give a hazy view at first	Registration for backup	Not standalone, supported by a panther	Limited access, paid	Online having limited input, skills required to integrate other data sets
Operating System	Online tool	Online tool	Online tool	Mac, window, Linux	Linux
Active or Inactive	Active	Active	Active	Active	Active
Format of Input	Gene list	Gene list	Gene ids, and gene product	List of sequences	Protein and nucleotide sequence
Format of Output	Pathway screenshot and	Diagram and annotation in CSV	Diagram and annotation in CSV	Fasta sequence file	Five output formats, XML, tsv, etc.
Peer-Reviewed	Yes	Yes	Yes	Yes	Yes
Background Algorithm	Mapping to a database, java, shell etc.	Text mining on integrated data	Java, data integration, text mining	Smith-waterman algorithms, text mining	Integrated database and multiple layers of interface
Basic and Advanced User	Basic	Basic	Basic	Advance	Advance
Open Source or Paid	Open-source	Open-source	Open-source	No	Open-source
Link	**Https://reactome.org/**	**Http://www.pantherdb.org/**	**Http://www.geneontology.org/**	**Https://www.blast2go.com/**	**Http://www.ebi.ac.uk/interpro/**
Accuracy	Moderate	Above Moderate	Above Moderate	Above Moderate	Above Moderate
Computational Requirement	Internet	Internet	Internet	High End for Big Data Set	High End for Big Data Set
Final Output Form	Screenshot and integrated output for plugins	CSV, excel file, etc.	CSV, excel file, etc.	Graphical and statistical views, CSV	Gtf, Html view, signature description
Cited by Scientific Community	Yes	Yes	Yes	Yes	Yes
Integration of Tool in	No	No	No	No	No

Different Pipeline					
Search Type (DNA Or Protein)	Genes, molecule and process	Gene list, protein and transcript	Gene ids and gene product	Fasta sequences of gene or protein	Fasta sequences of gene or protein
Machine Interface	Human-machine interface	Human-machine interface	Human-machine interface	Human-machine interface	Human-machine interface
Organisms	Eukaryotes	Eukaryotes	Eukaryotes	Eukaryotes	Eukaryotes
Reference Genome Required or Not	Directly no	Directly no	Directly no	Directly no	No
Citation	7597 (ALL)- source Google Scholar	1967 - source Google Scholar	23206 - source Google Scholar	7088 - source Google Scholar	1850 - source Google Scholar

Table continues below -

Stage in analysis	**Gene ontology- Biological, Cellular and Molecular process**				
Bioinformatics tools	***David***	***Gorilla***	***Genemania***	***Great***	***Predictprotein***
Pros	Functional Annotation and Gene Classification, Fast	Visualization According to Biological Process	Function Prediction of Genes and Gene Sets	Biological Meaning to Non-Coding Genomics Region and Chip-Seq, the functioning of the cis-regulatory system	Structure and Functional Annotation of Fasta Protein Sequence
Cons	Lack of Frequent Update, sometime website not functional	No Update From 2013 And Listed for Some Species	Only Network View	Limited Functioning and genome assemblies support	Registration Required
Operating system	Mac, window	Online tool	Online and standalone plugins	Online tool	Linux, centos online
Active	Yes	Yes	Yes	Yes	Yes
Format of input	Gene list	Gene, protein refined list	Gene name list	BED file, genomic regions file	Fasta protein sequence
Format of output	The bar chart and lists of gene	Flowchart, excel, revigo plugin	Networks	Table, Html, enriched terms	Pictorial view with details
Peer-reviewed	Yes	Yes	Yes	Yes	Yes

Background algorithm	Integrated data mining	Data mining	Keyword mapping	Probabilistic approach	Smith-waterman algorithms, mining
User-level	Basic	Basic	Basic	Basic	Advance
Open-source	Yes	Yes	Yes	Yes	Yes
Link	**Https://David.Ncifcrf.Gov**	**http://cbl-gorilla.cs.technion.ac.il/**	**Https://Genemania.Org/**	**http://great.stanford.edu/public/html/**	**Https://Www.Predictprotein.Org**
Accuracy	Above Moderate	Moderate	Moderate	Moderate	Moderate
Computational requirement	High end for standalone big data	Internet	Internet and moderate processor	Internet	High end for big data set
Final output	Text view, 2d-heat map	The flowchart and excel format of ontology	Network, excel, network image	Table of go, a histogram of tss sites	Html, text
Cited by the scientific community	Yes	Yes	Yes	Yes	Yes
Integration of tool in different pipeline	No	No	No	No	No
Search type (DNA or protein)	Gene list, protein and transcript	Gene, protein, ensemble ids, etc.	Stable gene name	Bed format file from the chip-seq output	Fasta sequences of protein
Machine interface	Human-machine interface	Human-machine interface	Human-machine interface	Human-machine interface	Human-machine interface
Organisms	Prokaryotes, Eukaryotes	Eukaryotes	Eukaryotes	Eukaryotes	Eukaryotes
Reference genome required or not	No	No	No	No	No
Citation	1092 - source Google Scholar	1763 - source Google Scholar	1346 - source Google Scholar	1733 - source Google Scholar	303 - source Google Scholar

Table continues below –

Stage in analysis	**Gene ontology- Biological, Cellular and Molecular process**				
Bioinformatics Tools	***Navigo***	***Metascape***	***Genes2go***	***Quickgo***	***Cluepedia***
Pros	Interactive Visualization tool, Go Retrieval, written in Ruby, Python and Perl	Human, Mouse Etc. DEGs And Metabolomics Data Visualization and Interpretation	Gene Ontology with Matrix and R-based web application	Ontology and Annotation searching,	Finding Pathways with Experimental and *in-Silico* Data, Cytoscape plugin

Cons	Input only in go terms	Not independent	Linked only to one data	Online having limited input, skills required to integrate other data sets	depends on another plugin, cluego
Operating System	Online tool	Mac, window, Linux, but java needed	Online tool	Online tool	Mac, window, Linux
Active	Yes	Yes	No	Yes	Yes
Format of Input	Go term	DEGs text file	Gene accession id and go term	Function keyword, go term	Pathway or network
Format of Output	Network	Network, Pdf Other format Images	Text Format	Charts with Option	Pdf Image, Network with Details
Peer-Reviewed	Yes	Yes	Yes	Yes	Yes
Background Algorithm	Database Mapping Using Statistics	Python Libraries, Java and Minning, High Network Algorithms	Keyword Searching and Building Matrix	Keyword Mapping	Calculate Linear and Non-Linear Statistical Dependencies
User-level	Basic	Medium	Basic	Basic	Basic
Open Source	Open-source	Open-source	Open-source	Open-source	Open-source
Link	**Http://kiharalab.org/web/navigo**	**Https://cytoscape.org/**	**Http://norstore-trd-bio0.hpc.ntnu.no:8080/genes2go/continue.do**	**Https://www.ebi.ac.uk/quickgo/**	**Http://apps.cytoscape.org/apps/cluepedia**
Accuracy	Moderate	Above Moderate	Moderate	Moderate	Moderate
Computational Requirement	Internet	High End for Large Networks with Multiple Nodes	Internet	Internet	Good and Updated Computer
Final Output Form	Network with association between GO terms and genes	Interactive and modified network pdf, jpeg etc.	Text only with data	JSON only	Network with edges and nodes
Cited by Scientific Community	Yes	Yes	Yes	Yes	Yes
Integration of Tool in Different Pipeline	No	Yes, But Limited	Connect to Other Database	Connect to Other Database	Connect to Other Database
Search Type (DNA Or Protein)	Protein set And Go Term	Network Integration, Kgml	Go Terms and Gene Id	Process, Go Term etc.	Network or Stored Session
Machine Interface	Human-machine interface	Human-machine interface	Human-machine interface	Human-machine interface	Human-machine interface
Organisms	Eukaryotes	Prokaryotes, Eukaryotes	Eukaryotes	Eukaryotes	Prokaryotes, Eukaryotes

Reference Genome Required	No	No	No	No	No
Citation	07- source Google Scholar	319 - source Google Scholar	00 - source Google Scholar	436- source Google Scholar	319 - source Google Scholar

Table continues below –

Stage in analysis	**Gene ontology- Biological, Cellular and Molecular process**				
Bioinformatics Tools	***Pingo***	***Gprofiler***	***Autoannotate***	***CateGOrizer***	***Gopet***
Pros	Find Gene Candidate in Pathway And concise the functions that belong to particular functional classes	Ht- Gene Annotation, 80+ Species, Multiple Option etc.	Finding Cluster and Visually Annotation automatically, Cytoscape app	Est And Ht RNA-Seq Data Analysis, Perform Step by Step Classification with Go Database, formally known as GO term classification counter	Sequence Annotation Tool, Gives Molecular Function, and automatic prediction of GO terms
Cons	Dependent on Other Tool as a Cytoscape Plug-in	Online Having Limited Input, Skills Required to Integrate Other Data Sets	Dependent on Cytoscape	Online Having Limited Input, Skills Required to Integrate Other Data Sets	Web Interface, not maintained and stable
Operating System	Mac, window, Linux	Any online tool	Any, plugin of Cytoscape	Any online tool	Online tool
Active	Active	Active	Active	Active	Inactive
Format of Input	Network of genes	Gene name list	Cluster network	Go ids	Protein and cDNA sequences
Format of Output	Tab-delimited text file	Tab-delimited text file, pdf and png image	Customized pathways and network	Tab-delimited text file	Go-term associated
Peer-Reviewed	Yes	Yes	Yes	Yes	Yes
Background Algorithm	Hypergeometric and binomial test	Data integration and mining	Database mapping using statistics	Database mapping using statistics	Homology and keyword mapping
User-level	Basic	Advance When Using In R	Basic	Basic	Basic
Open Source	Open-source	Open-source	Open-source	Open-source	Open-source
Link	**https://www.psb.ugent.be/esb/PiNGO/Home.ht**	**Https://biit.cs.ut.ee/gprofiler**	**Http://apps.cytoscape.org/apps/autoannotate**	**https://www.animalgenome.org/tools/catego/**	**Https://omictools.com/gopet-tool**

	ml go/home.html				
Accuracy	Moderate	Moderate	Moderate	Moderate	Basic
Computational Requirement	Good and updated computer	Internet and r package	Java, Cytoscape	Perl cgi programs	Internet
Final Output Form	Interactive network, and tsv file	Png, pdf, text	Network with nodes	Tabular output	Tabular file
Cited by Scientific Community	Yes	Yes	Yes	Yes	Yes
Integration of Tool in Different Pipeline	Yes, but limited	Connect to another database	Yes, but limited	No	No
Search Type (DNA Or Protein)	Network	Gene Ids	Network	Go Ids	Sequences
Machine Interface	Human-machine interface	Human-machine interface	Human-machine interface	Human-machine interface	Human-machine interface
Organisms	Prokaryote, eukaryotes	Eukaryotes	Prokaryotes, eukaryotes	Eukaryotes	Eukaryotes
Reference Genome Required	No	No	No	No	No
Citation	33 - source Google Scholar	517 - source Google Scholar	22 - source Google Scholar	209 - source Google Scholar	68 - source Google Scholar

2.6 Proteomics analysis using tandem mass spectrometry (MS/MS)

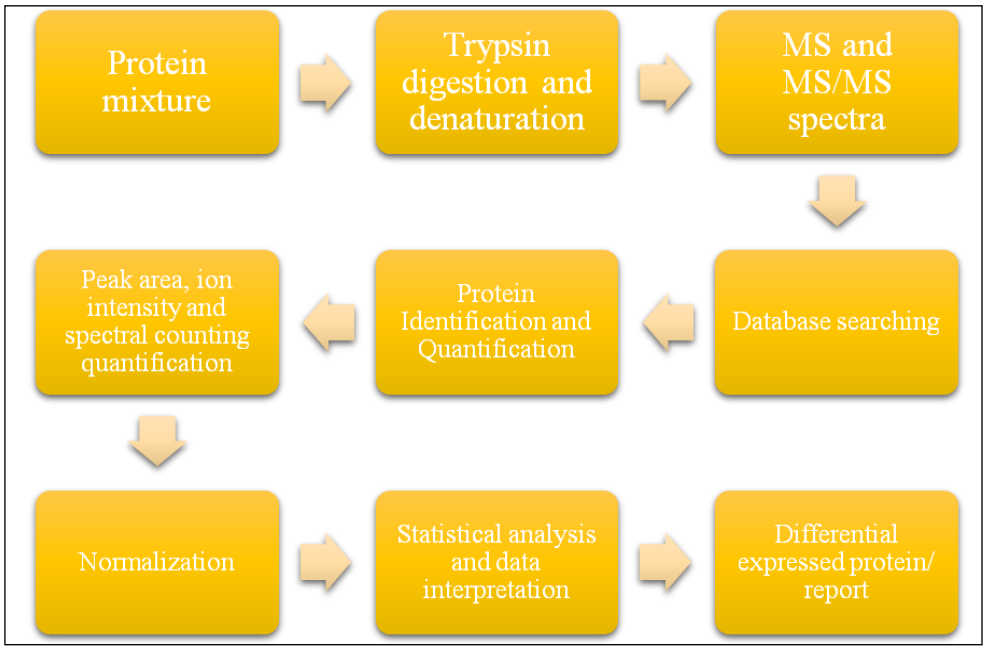
Proteomics is defined as the study and characterization of the complete set of proteins at a particular time, present in a cell, organ or organism (CHANDRAMOULI and QIAN, 2009). The function of an individual protein in a complex cellular process is determined by correlating, cell-cycle stage, disease state and growth condition (GULCICEK et al., 2005). The most promising tool for protein identification is the mass spectrometer, regardless of choices of proteomic separation techniques such as gel-based or gel-free. To study the protein mixture, a combination of electron spray ionization and tandem mass spectrometry is being used (PITT, 2009). The protein identification and measuring of protein expression from tandem mass spectrometry (MS/MS) can be termed quantitative mass spectrometry (KOLKER et al., 2006).

Quantitative mass spectrometry approaches are categorized into label-free, labelled, and targeted proteomics. Label-free proteomics performs with spectral counting and ion intensity analysis. While, in labelled proteomics, metabolic SILAC-Stable isotope labelling, chemical isotope (^{18}O), chemical isobaric labelling iTRAQ (isobaric tag for relative and

absolute quantification) and TMT-Tandem mass tags can be used (H. R. FULLER, 2012). To study the sensitive, specific, absolute protein quantification, "targeted proteomics" is preferred. It carries out with selected reaction monitoring (SRM) or multiple reaction monitoring (MRM) (Protein analyzed on triple quadrupole mass spectrometer) well established among other workflows. Another method is parallel reaction monitoring (PRM) that is more sensitive than SRM in term of model system study (RONSEIN et al., 2015).

The (MRM or SRM) can be considered as a gold standard data acquisition technique to perform quantitative proteomics studies. Although the new emerging data-independent acquisition (DIA) technique also known as SWATH (Sequential Windowed Acquisition of All Theoretical Fragment Ion Mass Spectra) can quantify a large number of proteins in a short time. It also supports fewer missing values and a lower coefficient of variation across the replicates (GULCICEK et al., 2005). The DIA works with the identification and quantification of predefined fragmented m/z range of ions regardless of the sample and is further categorized into targeted and untargeted acquisition. Comparing to SRM or MRM the label-free DIA techniques is much high throughput (KOCKMANN et al., 2016). Figure 3 depicting the workflow of quantitative proteomics mass spectrometry.

Figure 3. Workflow depicting the production and analysis of quantitative proteomics data



From protein mixture, digestion with a proteolytic enzyme, using Mass spectrometry produces spectra of peptides. Then these follow database searching, quantification and statistical filtration to obtain the quantitative differentially expressed data.

Source: Self-made figure

2.7 Quantitative proteomics

With the rapid advancement in mass spectrometry instruments and analysis strategies. The computational tool can help in deciding which proteins are differentially expressed and pathways reformed due to the specific stimulus.

The MS-based proteomics is divided into a top-down and bottom-up approach in which intact protein and surrogate peptide for the protein of interest are measured, respectively (SCHUBERT et al., 2017). The most widely accepted approach is bottom-up mass spectrometry-based proteomics (AMUNUGAMA et al., 2013). The bottom-up approach is further categorized into peptide mass fingerprinting and tandem mass spectrometry (MS-MS) for the identification of the protein.

This section focuses on label-free quantification and the bottom-up proteomics approach also known as "shotgun proteomics" in which the biological sample is directly digested with a proteolytic enzyme (ex. trypsin). This enzyme cleaves at the well-defined site to create a complex peptide mixture. Further, these digested peptides undergo liquid chromatography for separation and are directly sprayed into the mass spectrometer. In MS-MS two level of Mass, Spectra measurement takes place. The first level leads to sampling ionization and produces a mixture of ions. From these mixture ions, a precursor ion with a specific mass to charge ratio (m/z) is selected and considered as MS1. MS1 is further fragmented to MS2, which generates a product ion for the detection, followed by m/z separation. Collectively, this technique empowering the confident identification of the peptide in the sample and can be used for further proteomics analysis and functional interpretation (SCHUBERT et al., 2017).

2.7.1 Database searching, protein identification and post-processing

The method by John Yates and Jimmy Eng uses database searching instead of the interpretation or pre-processing step of MS/MS data. This works with a cross-correlation algorithm that compares an experimental MS/MS spectrum against spectra predicted from the candidate peptide sequence (COTTRELL, 2011). There are a sufficient number of online and offline tools explained in table 7 available to perform this task. *SEQUEST* tool works with a cross-correlation function-based algorithm, while, *Mascot* (PERKINS et al., 1999) uses a probability modelling algorithm (GENTZEL et al., 2003). *X! Tandem* consider semi-tryptic peptides and B/Y-type Ions.

Identified proteins are required to be filtered because they do not only correspond to a single protein. Therefore, even a small error rate in peptide classification can conclude in a larger number of protein identification errors. To overcome two methods can be used first, filter the peptides based upon database scores and the second properties of the assigned peptides. Although, these methods overlook potentially valid lower-scoring peptides matches and eliminate a large portion of the false positive peptide identified (SCHUBERT et al., 2017).

Table 7. A detailed description of tools for Database searching and protein identification

Tools for proteomics study	**Pros**	**Cons**	**Format of Input**	**Format of Output**	**Algorithm**	**Link**	**Citation**
Sequest	Protein identification, ms/ms, Linux based	Linux only	Ms/ms spectra	Peptide sequence	Cross-correlation approach	**http://fields.scripps.edu/yates**	100 - source google scholar
Mascot	Ms/ms database searching and peptide mass fingerprinting	Limited to smaller data, online and Linux and windows for a paid version	Raw data to the peak list	Peptide sequence	C++, Java and Perl were used for different plug-ins	**http://www.matrixscience.com/server.html**	301 - source google scholar
X!Tandem	Matching ms/ms spectra, with peptide sequences	Window, Linux and mac	XML file	XML file further search engine	Works with API	**https://www.thegpm.org/tandem/**	2173 - source google scholar
Proteome Discoverer	Identify and quantify proteins in complex biological samples, support multiple databases searching	Commercial, no open source	Raw mass spectra peak list	DTA archive file, MZDATA file	X-link algorithm	**Https://www.thermofisher.com/order/catalog/product/OPTON-30795**	2188 - source omicx
Scaffold	Classify proteins by molecular function or organelle, compare samples to find the biological relevance	High dpi/scaling and font issue in previous windows operating system	Peak list, mgf, .dat, .xml file .rov file, etc.	User-friendly interface with a tabular form	Protein grouping algorithm, heuristic rule embedded in the algorithm	**http://www.proteomesoftware.com/products/free-trial/**	389 - source google scholar

OMMSA	Robust, graphical interface and can integrate into the existing protein identification pipeline	Now the project is not maintained due to fund crisis, command-line interface	The fasta database file and spectra file	Omssa result file	Spectra algorithm, C++ language	**https://github.com/dbaileychess/compass**	1420 - source google scholar

Source: self-made table

Quantitative proteomics provides more information, rather than only browsing the list of recognized proteins. A label-free quantitative approach can be categorized into absolute and relative quantification. The relative label-free protein quantitation can be measured using spectral counting, ion intensity and peak area, among them spectral counting is convenient fast and easy to implement. While the limitation of this method is the quality of MS/MS peptide identification and inaccuracy for the protein identified.

2.7.2 Tools for quantitative proteomics analysis

To understand the correlation between the protein tools such as *ProPCA*, which is related to principal components analysis (PCA), combines spectral count (SC) and peptide peak attribute (PPA) data to obtain estimates of relative protein abundance (DICKER et al., 2010). *MaxQuant* (TYANOVA et al., 2016a), *Progenesis* (**http://www.nonlinear.com/progenesis**), *MSQuant* (MORTENSEN et al., 2010), and *Protmax* (EGELHOFER et al., 2013) holds trait for 'ion intensity count' and 'objective searching' peptides of a specific set.

OpenMS (ROST et al., 2016) and *SuperHirn* (MUELLER et al., 2007) tool comprises a novel platform for label-free quantification of complex LC-MS data and consolidate various functionalities for the mass spectrometric data processing (MUELLER et al., 2007). Another tool is a robust intensity-based averaged ratio (COLAERT et al., 2011) (*RIBAR*) introduced to overcome the issues that arise across replicates of protein ration obtained from label-free approaches. *RIBAR* estimates the intensity of the corresponding spectra fragment in two observations and exponentially altered protein abundance index. As a result, normalized spectral abundance factors outperformed other spectral counting-based approaches (COLAERT et al., 2011). However, reaching high quantification accuracy is difficult because of data-independent ions sampling and dynamics exclusion list setting (NAHNSEN et al., 2013). Another tool *StatQuant* (VAN BREUKELEN et al., 2009) provides a graphical interface and a range of advanced statistical procedures that involves

P-value and standard deviation for performing post quantification analysis of protein abundance ratio.

2.7.3 Functional interpretation of abundant proteins/genes

This section explains the bioinformatics tools and databases used for the functional annotation and interpretation of the differentially expressed genes or proteins. Figure 4 depict the pipeline and tools available for the downstream analysis. The first step of analysis is to connect protein names to their corresponding genes that can be performed using the databases UniProt knowledgebase for protein sequence and functional information manually or automatically annotated (MAGRANE et al., 2011), Ensembl for genome browsing and Ensembl gene IDs (YATES et al., 2020), peptide atlas project (DEUTSCH, 2010) for peptide identification in MS/MS proteomics experiments and the global proteome machine for protein identification (BEAVIS, 2006).

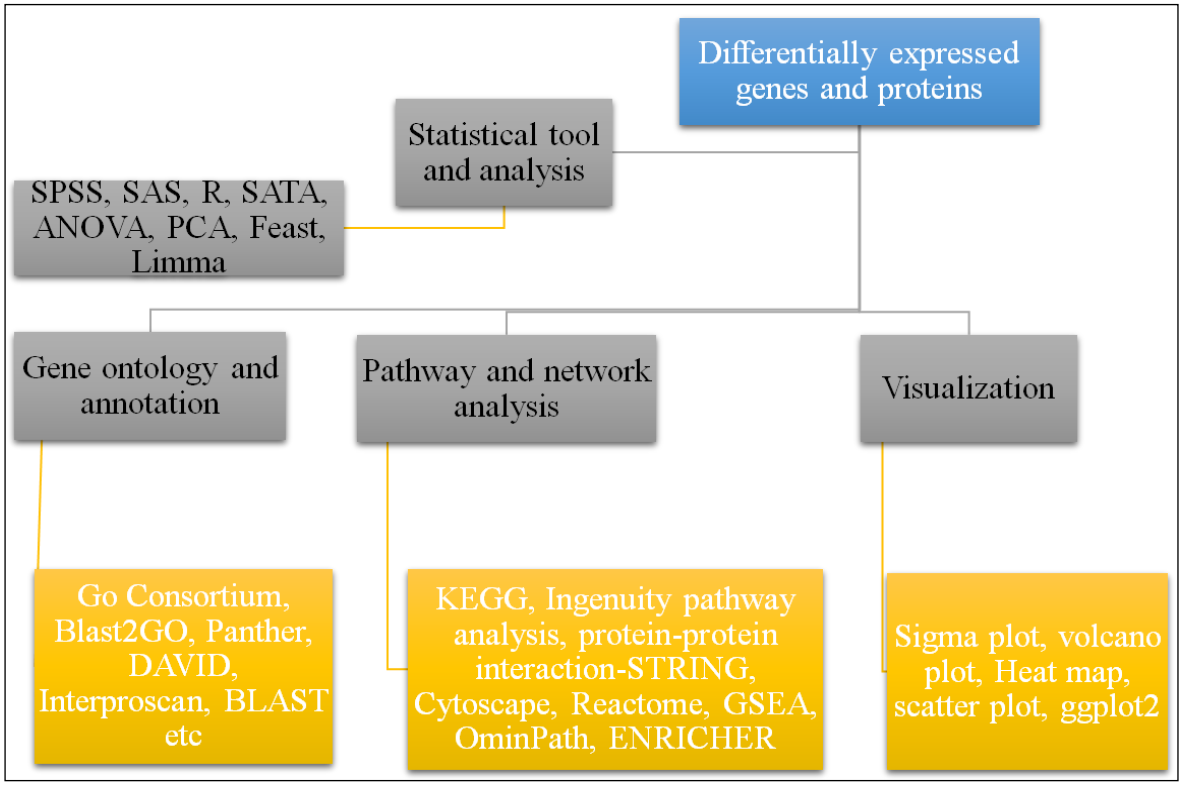
Once the differentially expressed genes or proteins are statistically filtered and quantified they can be subjected to functional interpretation using gene ontology analysis databases, tools such as *g: Profiler* (RAUDVERE et al., 2019), *clusterprofiler* (YU et al., 2012), *DAVID* (DENNIS et al., 2003) and *Blast2Go* (CONESA et al., 2005).

There are 120 tools available for various gene ontology purposes (source omicX). For mechanical insight pathways analysis can be performed using various online tools namely *KEGG* (KANEHISA and GOTO, 2000), *Reactome* (CROFT et al., 2011), I*ngenuity pathways* (QIAGEN, Inc., **https://targetexplorer.ingenuity.com/**) and *Biocarta* pathway datasets (ROUILLARD et al., 2016). Whereas interaction analysis of these genes or proteins can be performed using databases namely *BioGRID* an interaction repository (OUGHTRED et al., 2021), *IntAct* for analysis of molecular interaction by modelling and storing (KERRIEN et al., 2007) and *MINT* – molecular interaction database is a protein-protein public repository (LICATA et al., 2012).

The *HPRD*-human protein reference database is available to browse the interactions, motifs and domains of the human proteome (KESHAVA PRASAD et al., 2009). *STRING* database is capable to predict the association between proteins for a large number of organisms (SZKLARCZYK et al., 2017). The data from the above-listed tool can be further extended to the *ChEMBL* database that works on the integration of bioactivity from medical chemistry literature, disease screening, crop protection data, drug metabolism and data from patents (GAULTON et al., 2017). Among others, *STRING* and *BioGRID* are more user-

friendly tools to perform protein interaction studies and extending the visualization of enriched categories using *Cytoscape*.

Figure 4. Workflow showing the tools and steps for the classification of differentially expressed proteins and genes



Depicting the pipeline for the analysis of differentially expressed genes and differentially abundant protein with their possible analysis level, bioinformatics and statistical tools.

Source: self-made figure.

2.8 The statistical and false-positive control approach

In a comparative study, the data from the machine either in transcriptomics or proteomics is in the form of signal intensities, read counts or mass spectra. Expression data undergoes pre-processing, statistical analysis, validation and functional prediction explained in the above sections. The major role of statistical approaches is the extraction of unbiased and significant results from the huge amount of data.

Transcriptomics and proteomics data are high-dimensional in nature, in which the number of measurement p is always surpassed the number of observations n (genes or proteins). When the quantitative differential data pass through the significant expression test between the treatments, it is difficult to decide whether the result is false positive, false negative or by chance. To counter this confusion, the Bonferroni correction method for

testing multiple hypotheses was introduced. Another method, Familywise Error Rate (FWER) proposed by R. J Simes (DIZ et al., 2011) minimizes or controls the probability of shaping more than one false positive errors in a set of tests.

The alternative approach to control the FWER is the False Discovery rate (FDR), which can accept the false-positive results among the declared set of significant results (LEEK et al., 2012). Another method introduced by the Benjamini-Hochberg (B-H) to control the FDR, in which the p-value sorted in descending order, ranked from smallest to largest and then the B-H critical value (i/m) Q is calculated, where i is the rank of a p-value, m is the total number of tests and Q is selected FDR. From the result, the largest p-value is less than the critical value considered as significant and all p-values lower than it. The alternative tools *limma R* package and statistical testing such as Storey-Tibshirani and Bayes are also capable of controlling the false discoveries numbers and can be corrected for the multiple testing problems (PEPKE et al., 2009).

3 Aims of the study

1. To generate complex repositories generated for immunoinformatics software platform generated – OMnalysis.
2. To develop algorithms to process big data derived from transcriptomic and proteomic analysis – OMnalysis
3. To use OMnalysis on transcriptomic and proteomics data derived from challenging with pathogens.

4 Materials and Methods

4.1 Omics data gathering and repository generation

Data generated using Omics approaches were gathered from various online and offline resources. The main sources were NCBI (National Center for Biotechnology Information) (**https://www.ncbi.nlm.nih.gov/home/download/**) and EMBL-EBI (European Molecular Biology Laboratory's European Bioinformatics Institute) ArrayExpress ((**https://www.ebi.ac.uk/arrayexpress/browse.html**) for sequenced data, WikiPathways (**https://www.wikipathways.org/index.php/Download_Pathways**) and SBML (Systems Biology Markup Language) pathway repository for pathways (**http://www.systems-biology.org/001/001.html**), the GENE ONTOLOGY RESOURCES (**http://geneontology.org/**) for gene ontology annotation. We focused on the omics data of humans, pigs, cattle and chicken to make a tool for researchers who are working in the veterinary field. Unstructured data in the form of a full-text paper was downloaded using the NCBI PubMed (**https://www.ncbi.nlm.nih.gov/public/**) file transfer protocol (FTP).

The collected data was stored in an in-house hard drive (Synology, Taiwan) by assigning each data type a separate repository (Table 8).

Table 8. Structured and unstructured data repositories

Structured data	Data in Gigabytes	Link
Human (airway, intestinal) epithelium transcriptome	2.3 Gb zipped (approximately)	**http://gofile.me/6DOhe/97a86OOc6**
Pig epithelium transcriptome	12.7 Gb zipped (approximately)	**http://gofile.me/6DOhe/75hzXSIw6**
Chicken epithelium transcriptome	90.4 Mb zipped (approximately)	**http://gofile.me/6DOhe/ShVW8kIz3**
Cattle epithelium transcriptome	14.3 Gb zipped (approximately)	**http://gofile.me/6DOhe/sCb8cZB4X**
Human lymphocytes transcriptome	16.5 Gb zipped (approximately)	**http://gofile.me/6DOhe/QY7pERRUC**
Pig lymphocytes transcriptome	73.2 Mb zipped (approximately)	**http://gofile.me/6DOhe/omZoY4V7h**
Chicken lymphocytes transcriptome	130 Mb zipped (approximately)	**http://gofile.me/6DOhe/8TbAVZo5A**

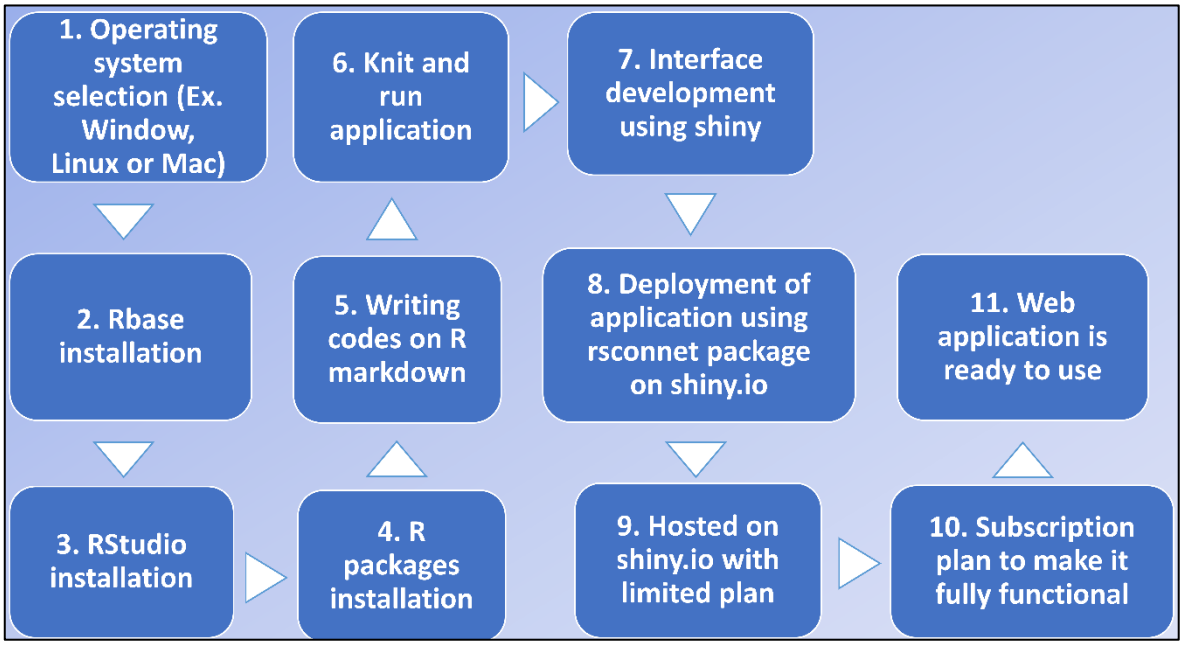
Cattle lymphocytes transcriptome	887 Mb zipped (approximately)	**http://gofile.me/6DOhe/xwvmRAmGi**
Human monocytes transcriptome	5.37 Gb zipped (approximately)	**http://gofile.me/6DOhe/aJdCeb9pQ**
Pig monocytes transcriptome	949 Mb zipped (approximately)	**http://gofile.me/6DOhe/T5FzasXq9**
Chicken monocytes transcriptome	61.9 Mb zipped (approximately)	**http://gofile.me/6DOhe/mQPX35H7R**
Cattle monocytes transcriptome	179 Mb zipped (approximately)	**http://gofile.me/6DOhe/zeEH8gops**
Human microbiota	115 Mb zipped (approximately)	**http://gofile.me/6DOhe/IxsMKKqsI**
Pig microbiota	113 Mb zipped (approximately)	**http://gofile.me/6DOhe/LkMrmn3L7**
Chicken microbiota	98.4 Mb zipped (approximately)	**http://gofile.me/6DOhe/iDx2MlCP9**
Cattle microbiota	32.6 Mb zipped (approximately)	**http://gofile.me/6DOhe/MfgYqf06p**
Un-structured data		
Research articles from NCBI	More than 300 Gb (approximately)	**http://gofile.me/6DOhe/QcJa0BRrF**

Data collected from various sources can be accessed using the provided link.

4.2 Construction of graphical interface using R and Shiny

Figure 5 outlines the development and integration of the components required to build an interactive Shiny application. The developed web application is sectioned into eleven user interfaces (UI); each part performs a specific function and produces results. We focused on the development of an easy to use, robust and intuitive web application so that researchers with less or no bioinformatics experience can routinely translate the expression data to biological insight.

Figure 5. Diagram depicting the workflow of Shiny based web application development



The workflow shows the steps to develop a Shiny based web application. From step 1 to 3 is a selection of an operating system and installation of R and RStudio. Steps 4 to 7 is package installation, writing codes and testing using R Shiny. Steps 8 to 11 is app deployment, hosting on the server, testing and updating.

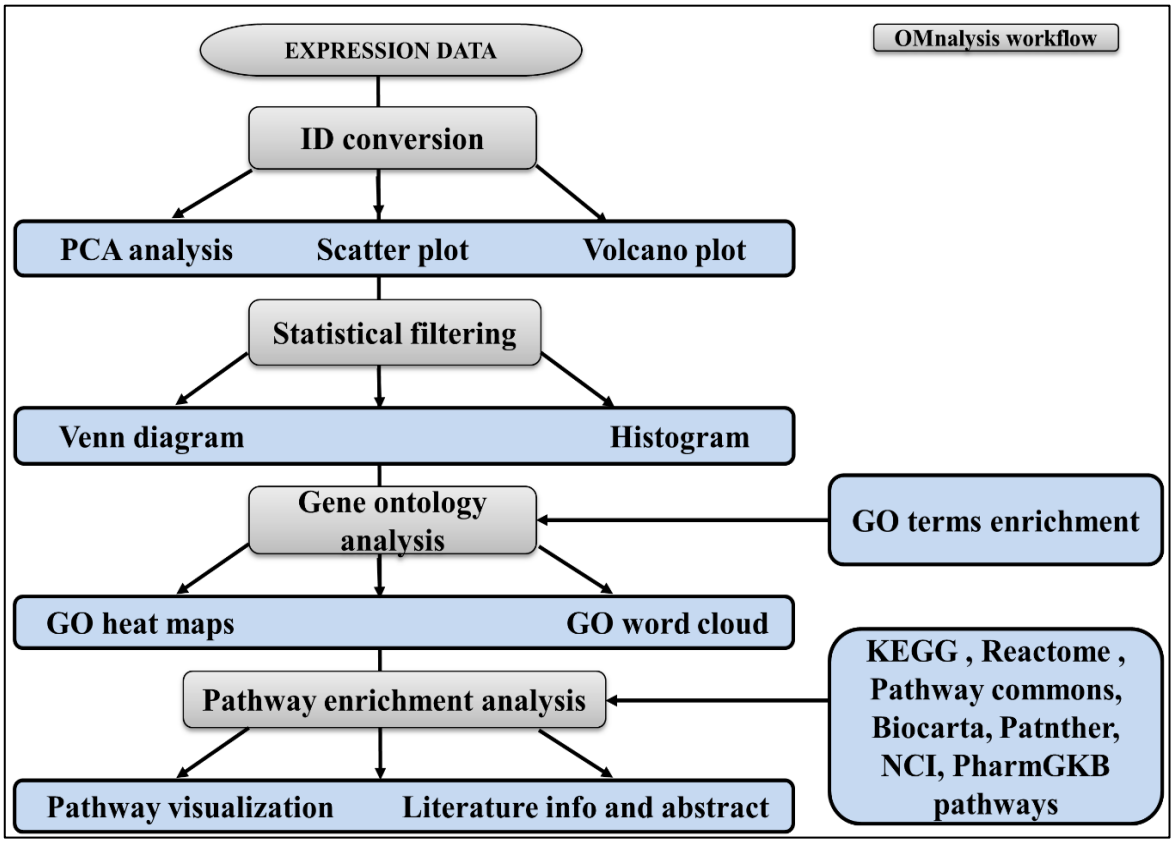
4.2.1 R packages used to establish OMnalysis

OMnalysis is an interactive R shiny based web application composed of multiple sectioned user interfaces (UI) in the form of tabs. It is built to perform the exploration of differential expression data efficiently and iteratively. R packages used for the development of the UI and its components are as follows: *R Shiny* version 1.6.0 (CHANG et al., 2017), *flexdashboard* version 0.5.2 (IANNONE et al., 2018), *Shiny Themes* version 1.2.0 (CHANG et al., 2018), *rmarkdown* version 2.8 (ALLAIRE et al., 2020), *knitr* version 1.33 (XIE, 2019) and *shiny dashboard* version 0.7.1 (CHANG et al., 2019). Each sectioned UI is further divided into interactive and display panels. The interactive panel works on Shiny's reactivity property, which automatically updates the values in the output panel when the user interacts or changes the input components (plots, tables, actions, etc.).

The biological analysis is supported by the following R packages: *biomaRt* version 2.46.3 (DURINCK et al., 2009), *clusterProfiler* version 3.18.1 (YU et al., 2012), *ReactomePA* version 1.34.0 (YU et al., 2016), *reactome.db* version 1.74.0 (LIGTENBERG, 2019), *pathview* version 1.30.1 (LUO et al., 2013), *SPIA* version 2.42.0 (TARCA et al., 2009), *SBGNview* version 1.4.1 (DONG et al., 2021), *STRINGdb* version 2.2.2 (SZKLARCZYK et al., 2019), *org.Hs.eg.db*, *org.Gg.eg.db*, *org.Ss.eg.db*, *org.Bt.eg.db*

version 3.12.0 (CARLSON, 2019). Visualization of the results from the analysis is backed by the following packages: *EnhancedVolcano* version 1.8.0 (BLIGHE et al., 2020), *gplots* version 3.1.1 (WARNES et al., 2016), *ggbiplot* version 0.55 (VU, 2016), *ggplot2* version 3.3.3 (WICKHAM et al., 2016), *VennDiagram* version 1.6.20 (CHEN et al., 2011), *wordcloud* version 2.6 (FELLOWS, 2018), *dplyr* version 1.0.5 (WICKHAM et al., 2015) and *DT* version 0.18 (XIE et al., 2018).

Figure 6. Workflow depicting the pipeline of OMnalysis tool



The flow of expression data from uploading to biological insight.

4.2.2 Knitting of R packages on R Markdown

We tested the selected R packages using in *RStudio* desktop free subscription version 1.4.1103 (RSTUDIO TEAM, 2015). Codes were written on R Markdown with the .rmd file extension. Initial lines of codes represent the title of the application, output type (flexdashboard, theme type, orientation of the application), and runtime Shiny. We added the path of supporting R data files required to execute pathway analysis.

4.3 Deployment of Shiny application

To connect a Shiny application to the *shinyapps.io* cloud server we used *rsconnect* version 0.8.18 (MCPHERSON et al., 2021). After that, we created an account in *shinyapps.io* and used the secret token provided with it to deploy the Shiny application to the *shinyapps.io* server. Before deployment, we provided setRepositories and devtools install GitHub function in RStudio console to allow easy installation of packages and their dependencies from a comprehensive repository of archive network (CRAN), *Bioconductor* version 3.12 (GENTLEMAN et al., 2004) and GitHub repository (**https://github.com/**) on the shinyapps.io server. We used the paid subscription plan of the shinyapps.io server with a memory size of 8192 MB.

4.4 Example data for OMnalysis development

The web application is designed to analyze two types of quantitative omics, transcriptomics and proteomics. For transcriptomics, RNA-Seq data generated previously by us and deposited in ArrayExpress was used (**www.ebi.ac.uk/arrayexpress**). Expression analysis was performed on human brain microvascular endothelial cells (hBMEC) induced with various pathogens: *Borrelia burgdorferi* (Treatment1, *retrieved from* ArrayExpress accession number E-MTAB-8053), *Neisseria meningitidis* (Treatment2, E-MTAB-8008), *Streptococcus pneumoniae* (Treatment3, E-MTAB-8054), and West Nile Virus (Treatment4, E-MTAB-8052). Treatment 1 to 4 (See Table 9 first five rows) are in text format processed from the TSV file generated from the *edgeR's glmTreat* function (ROBINSON et al., 2010). Three columns (logFC, logCPM, and Pvalue) from each of those treatments were copied to make a master file in CSV format (Table 9). In table 9, column (col.) value 1^{st} belongs to Ensembl ID, 2^{nd} to 4^{th} (treatment 1^{st}), 5^{th} to 7^{th} (treatment 2^{nd}), 8^{th} to 10^{th} (treatment 3^{rd}) and 11^{th} to 13^{th} (treatment 4^{th}). The expression data of each sample can be accessed using the link provided in table 9.

Table 9. DEGs input file for OMnalysis.

Sample	**Pathogen**						**Link**					
Treatment 1	*Borrelia burgdorferi*						**http://gofile.me/6DOhe/4uBkNkkq8**					
Treatment 2	*Neisseria meningitidis* (NM)						**http://gofile.me/6DOhe/3eYClkcZz**					
Treatment 3	*Streptococcus pneumoniae*						**http://gofile.me/6DOhe/Zsi0JhAvA**					
Treatment 4	West Nile Virus (WNV)						**http://gofile.me/6DOhe/zLbbBqdt8**					
DEGs input file for OMnalysis (http://gofile.me/6DOhe/R3nAq3o0T)												
Ensembl IDs	**Treatment one**			**Treatment two**			**Treatment three**			**Treatment four**		
col.[1] 1	**col. 2**	**col. 3**	**col. 4**	**col. 5**	**col. 6**	**col. 7**	**col. 8**	**col. 9**	**col. 10**	**col. 11**	**col. 12**	**col. 13**
ENSEMBLGENE	**logFC**	**logCPM**	**Pvalue**	**logFC**	**logCPM**	**Pvalue**	**logFC**	**logCPM**	**Pvalue**	**logFC**	**logCPM**	**Pvalue**
ENSG00000000003	-0.74375	4.557846	0.267133	-0.90616	4.557846	0.141092	0.417361	4.557846	0.581714	-0.35996	4.557846	0.640167
ENSG00000000419	0.108453	4.758842	0.890174	0.096469	4.758842	0.902259	-0.12866	4.758842	0.872534	-0.78139	4.758842	0.254466
ENSG00000000460	-0.45651	1.782839	0.559013	-0.13903	1.782839	0.8607	-1.43246	1.782839	0.043123	0.792963	1.782839	0.277646
ENSG00000000971	-0.45012	6.933399	0.553369	-0.32462	6.933399	0.676356	0.131856	6.933399	0.868554	-0.62059	6.933399	0.380078
ENSG00000001036	-0.06613	5.462408	0.934619	-0.35845	5.462408	0.642259	-0.29022	5.462408	0.710371	0.171774	5.462408	0.827248
ENSG00000001084	-0.18668	4.100634	0.813381	-0.62329	4.100634	0.381325	0.057774	4.100634	0.941401	-0.01379	4.100634	0.986338
ENSG00000001167	0.414895	3.653166	0.584106	0.34253	3.653166	0.656407	-0.38414	3.653166	0.615672	0.178518	3.653166	0.820939
ENSG00000001461	-0.73107	5.259942	0.275607	-1.37751	5.259942	0.002377	-0.39492	5.259942	0.606211	-1.1324	5.259942	0.0298
ENSG00000001497	0.018868	4.674989	0.980772	0.013774	4.674989	0.985861	-0.41221	4.674989	0.589126	-0.54674	4.674989	0.454072
ENSG00000001617	-0.74375	4.557846	0.267133	-0.90616	4.557846	0.141092	0.417361	4.557846	0.581714	-0.35996	4.557846	0.640167

Part of the example dataset was used to check the functionality of OMnalysis. The input file header ENSEMBLGENE, logFC, logCPM and Pvalue must be identical to the table above. The first column must be ENSEMBL Ids and then logFC, logCPM and Pvalue or FDR value for each treatment in series.

[1] col. is abbreviation of Column

In the case of proteomics, the data matrix was generated from one of the differential abundance analysis software *Perseus* version 1.6.15 (TYANOVA et al., 2016b). To check the functionality, we retrieved the .xlsx format file from the experiment performed to quantify protein abundance in the milk whey collected at different time points from the cow with *Streptococcus uberis* infection (MUDALIAR et al., 2016). The columns in this data matrix were arranged in the following order: UniProt ID, FDR-adjusted P-value, and Fold Change in an excel file for each experimental condition (4-time points, in this case, designated as Treatment1 (36 hours), Treatment2 (42 hours), Treatment3 (57 hours), and Treatment4 (81 hours); Figure 7).

Figure 7. Differentially abundance proteins input data for OMnalysis.

	A	B	C	D
1	**UniProt ID**	**FDR-adjusted P-value**	**Fold Change**	
2	Q8SPP7	4.5E-10	3304.8	
3	P54229	1.9E-08	1443.5	
4	P56425	1.6E-06	1217.0	
5	P22226	2.8E-08	1026.2	
6	Q2TBU0	3.8E-08	996.8	
7	F1N465	1.5E-03	527.2	
8	E1BCU6	1.5E-06	401.1	
9	Q9TU03	1.6E-04	312.8	
10	P52176	1.1E-04	219.2	

Treatment1 | Treatment2 | Treatment3 | Treatment4

Differentially expressed quantitative label-free proteomics data, with four treatments (Treatment1, 2, 3 and 4) in four different excel sheets. Each excels sheet presents one treatment data in which columns A, B and C shows the UniProt protein ID, FDR-adjusted P-value and ratio of the change to control (0 hours), (Fold Change).

4.5 Data modification and ID conversion

We used the *read.csv* function of *Utils* package version 3.6.2 (TEAM, 2013) to upload the CSV format file (Table 9) to the OMnalysis. Whereas, for proteomics, we used the *import_list* function of *rio* package version 0.5.26 (CHAN et al., 2018) to upload the data (Figure 7). For proteomics data, three functions were used to convert data (Figure 7) to make an input table for the OMnalysis. First, the *rio* package was used to convert Treatment sheets to the Treatment column. Second, the *log2* function of *base R* was used to transform the Fold Change column values to logFC (log Fold Change), and the third, Colnames function was used to change the treatment column name to Treatments. The duplicate proteins in the

treatments were identified using the *group_by* function of *dplyr* version 1.0.5 (WICKHAM et al., 2015) and the *mean* function of *base R* to obtain the mean of their log fold change value. Such conversion is not necessary in the case of transcriptomic data.

Transcriptomic data matrix contains Ensembl IDs, while proteomic data comes with UniProt IDs. To convert these IDs into five different ID types (Ensemble gene ID, gene name, HGNC symbol, gene description, and UniProtKB/Swiss-Prot ID) we used the *getBM* function of *biomaRt* package version 2.46.3 (DURINCK et al., 2009) to fetch the latest information from the Ensembl database (YATES et al., 2020). We have incorporated the possibility of ID conversion for 9 species (Human, Chicken, Pig, Cow, Mouse, Rat, Dog, *Drosophila melanogaster* and *C.elegans*) in OMnalysis.

4.6 Principal component analysis

For dimension reduction and identification of variation in the data set with many variables. First, the log fold change was taken as a variable and genes or proteins as individuals, then these values were scaled to avoid the domination of log fold change in the association among the variables (Treatments). The *fviz_pca_var* function of *factoextra* package version 1.0.7 (KASSAMBARA et al., 2017) and *prcomp* function of *stat* version 3.6.2 (TEAM, 2013) were used for Variable PCA and Biplot PCA, respectively. We used the *biplot* function of *ggbiplot* (VU, 2016) for the visualization of the relation between the samples and variable contribution in the principal component analysis.

4.7 Data visualization

Thousands of genes were plotted on a scatter plot using *ggplot2* version 3.3.3 R package (WICKHAM et al., 2016) to visualize the dispersion of data after applying specific thresholds values. *EnhancedVolcano* version 1.8.0 package was used to generate a volcano plot of differentially expressed data (BLIGHE et al., 2020). The *Venn diagram* R package version 1.6.20 (CHEN and BOUTROS, 2011) was used for the generation of the Venn diagram, showing common and non-common significant genes in the uploaded treatments. For the construction of the histogram, the filtered data was rearranged using the *rbind* function of *base R* (RSTUDIO TEAM, 2015). *Ggplot* function of the *ggplot2* R package was used to generate a single or all treatments histogram.

4.8 Statistical filtering

To perform statistical filtering we used the *dplyr* version 1.0.5 R package's *filter* function (WICKHAM et al., 2015). The expression data were filtered according to the column of each treatment and omics type. From the CSV or xlsx file columns were selected based on log fold change (logFC), log count per million (logCPM), *P* value (Pvalue)/FDR in transcriptomics or log fold change and FDR-adjusted P-value in proteomics. *Dplyr* package's *select* function was used to separate the treatments with their ensemble id (ENSEMBLGENE) or UniProt ID, log fold change (logFC), *P* value (Pvalue) or FDR-adjusted P-value and log count per million (logCPM). *Filter* function was used with the pipe function to separate up and down-regulated genes according to their log fold change. The list of up and down-regulated genes were extracted as unique.

4.9 Functional profiling of DEGs

Once the data matrix was statistically filtered, we used *clusterProfiler* version 3.18.1 of Bioconductor packages (YU et al., 2012) to obtain the functional interpretation of the significantly expressed genes or proteins. *Names* and *sort* functions of *base R* package version 4.0.3 were used to set the names of genes and arranging them in decreasing order using log fold change value, respectively. Two enrichment analysis functions of the *clusterProfiler* were used, the first, *EnrichGO* function on genes or proteins to perform over-representation analysis (ORA) and the second *gseGO* function on sorted genes or proteins with respect to logFC values to perform gene set enrichment analysis (GSEA). To support the enrichment analysis, *AnnotationDbi* version 1.52.0 of Bioconductor databases (PAGÈS et al., 2020) was used. Mark Carlson species-specific genome-wide annotation databases version 3.12.0 were used for human (*org.Hs.eg.db*), chicken (*org.Gg.eg.db*), pig (*org.Ss.eg.db*), and cattle (*org.Bt.eg.db*), mouse (*org.Mm.eg.db*), rat (*org.Rn.eg.db*), dog (*org.Cf.eg.db*), *Drosophila melanogaster* (*org.Dm.eg.db*) and *C.elegans* (*org.Ce.eg.db*) (CARLSON, 2019). P-value cutoff input and PAdjust function of *ClusterProfiler* with seven multiple hypotheses testing correction methods were used to avoid the influence of false-positive results on the overall enrichment analysis.

4.10 Visualization of GO terms

To visualize the enriched biological terms of analyses on heatmap we used *heatmap.2* function of *gplots* version 3.1.1 of R packages (WARNES et al., 2016). The output data in the form of multiple rows and columns were handled using *DT* R package version 0.18 (XIE

et al., 2018). To provide colours to the heatmaps we used the *rainbow* function of the R base version 4.0.3 (RSTUDIO TEAM, 2015). We assigned legends to each treatment in the generated heatmap with the help of the legend function of *gplots*. Hundreds to thousands of gene ontology terms were enriched and to visualize the maximum number of gene ontology terms with their frequency *word cloud* R package version 2.6 was used (FELLOWS, 2018).

4.11 Pathway analysis of DEGs

To get the mechanistic insight from the list of significantly differentially expressed genes or proteins we used pathway enrichment analysis. To run the *enrichKEGG* and *gseKEGG* function, we used the *cluster profiler's* version 3.18.1, which yields the enriched pathways terms with enrichment score and detailed statistical information. This analysis was supported by another three methods first, *network topology analysis* (*NTA*) (ALEXEYENKO et al., 2012), second *Reactome pathway analysis* (YU and HE, 2016) and third, *STRINGdb* version 2.2.2 (SZKLARCZYK et al., 2019).

To perform the *NTA*, we first manipulated and arranged the tabular data using the *merge, cbind* and *gsub* function of *R base* version 4.0.3 (RSTUDIO TEAM, 2015). The reference database such as *biocarta* (ROUILLARD et al., 2016), *panther* (THOMAS et al., 2003), *NCI- nature pathway interaction database* (ANTHONY et al., 2011), and *pharmGKB* (KLEIN et al., 2004) is required to perform *NTA*. Thus we used Entrez IDs and *graphite pathway* function of *graphite* (GRAPH Interaction from pathway Topological Environment) R package version 1.36.0 (SALES et al., 2012) to generate the reference pathway database. *Graphite runSPIA* function was used to perform network topology analysis.

An *enrichpathway* function of *ReactomePA* (YU and HE, 2016) R package version 1.34.0 was used to perform Reactome pathway analysis using the *Reactome pathway* database version 1.74.0 (CROFT et al., 2011). Enrichment analysis using *STRINGdb* was performed by adding new stringdb and assigning species, score threshold and input directory (SZKLARCZYK et al., 2019). The newly generated stringdb was used to map the statistically significantly differentially expressed gene or proteins with stringids against several databases (GO annotation, KEGG pathways, PubMed publications, Pfam domains, InterPro domains, UniProt Keywords SMART domains).

The output of ORA and GSEA enriched pathway analysis was visualized on the selected pathway using *pathview* R package version 1.30.1 (LUO and BROUWER, 2013). The *ReactomePA* output was visualized using an R Bioconductor package *SBGNview*

(overlay omics data onto sbgn pathway diagrams) version 1.4.1 (DONG et al., 2021). *SBGNview* depends on the pathways.info and sbgn.xmls files to generate an integrated pathway for *ReactomePA* analysis. The plot network function of *STRINGdb* to was used generate the protein-protein interaction network.

4.12 Literature retrieval

To retrieve the published scientific information for the identified biomarkers, we used *Europe PMC*- an R client for Europe PMC RESTful articles (LEVCHENKO et al., 2018). To fetch the scientific information, *europepmc* details function of R *europepmc* package version 0.4 was used.

4.13 Output file types

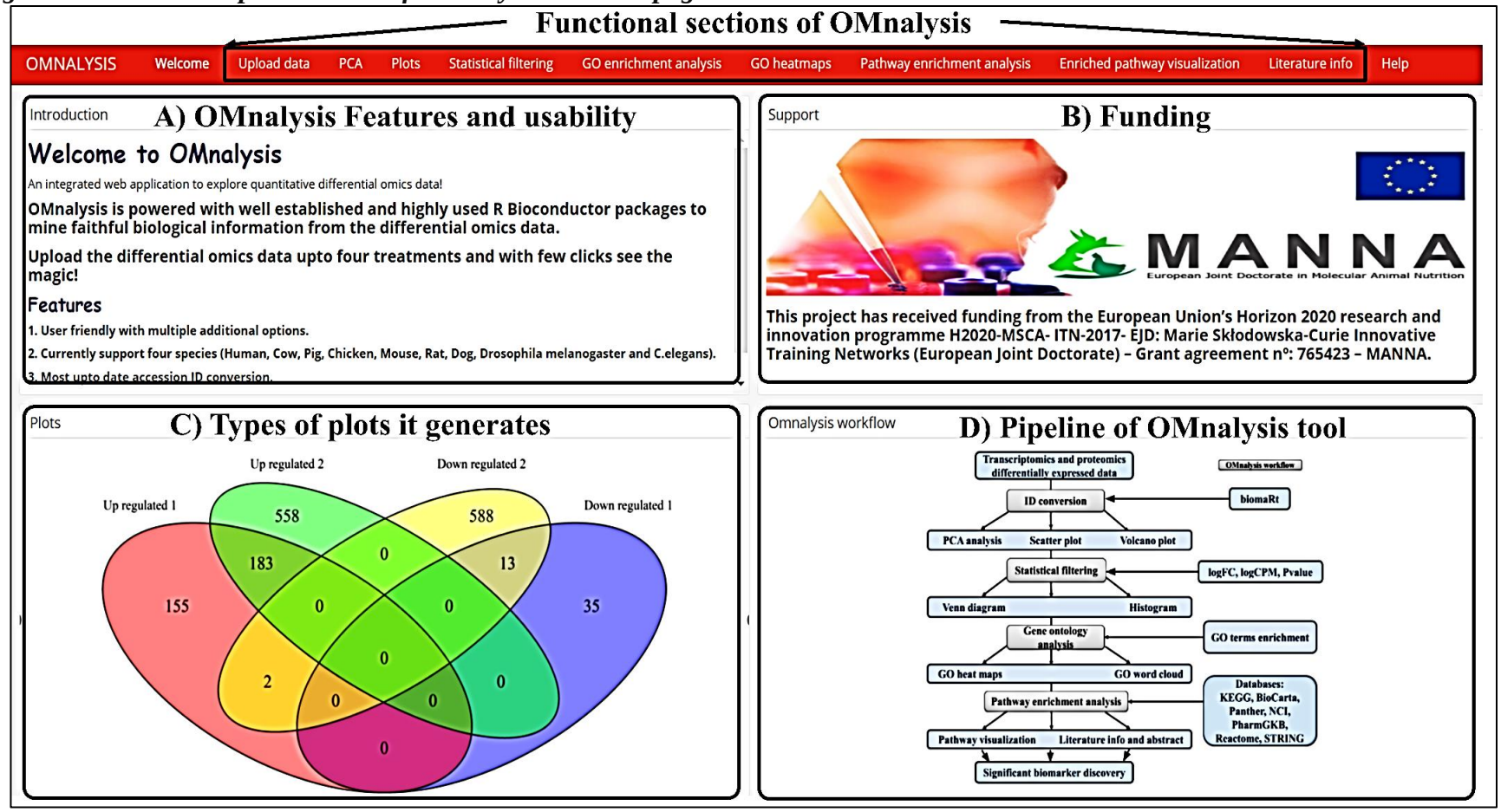
Write.csv function of utils version 2.10.1 (BENGTSSON et al., 2020) was used to write comma-separated-values, *ggsave* function of *ggplot2* version 3.3.3 (WICKHAM et al., 2016) to save plots, *grid.draw* function of *grid* version 3.6.2 (MURRELL, 2018) to produce graphical layouts of multiple plots, *print* function of *R base* version 4.0.3 to print image, *recordPlot* function of *grDevices* (RSTUDIO TEAM, 2015) to save and replay the current plot and *file.copy* function to copy temporary files. These functions were used in the downloadHandler of *Shiny R* package version 1.6.0 (CHANG et al., 2017) to download the processed file.

5 Results

5.1 Welcome page

Figure 8 present the first page of the OMnalysis web application and it is divided into four sections (A), (B), (C) and (D). Top left (A) details about important features and where they can be used, top-right (B), details of funding, bottom right (C) illustrate the workflow of the OMnalysis web application and at the bottom left (D) shows the types of visualization OMnalysis tool generates (Figure 8). The functional sections of OMnalysis shown in Figure 8, (Upload data, PCA, Plots, Statistical filtering, GO enrichment analysis, GO heatmaps, Pathway enrichment analysis, Enriched pathway visualization and Literature info) streamlined the expression data analysis, whereas, welcome and help sections provide information about advantages and instruction to use OMnalysis, respectively.

Figure 8. Schematic representation of OMnalysis welcome page



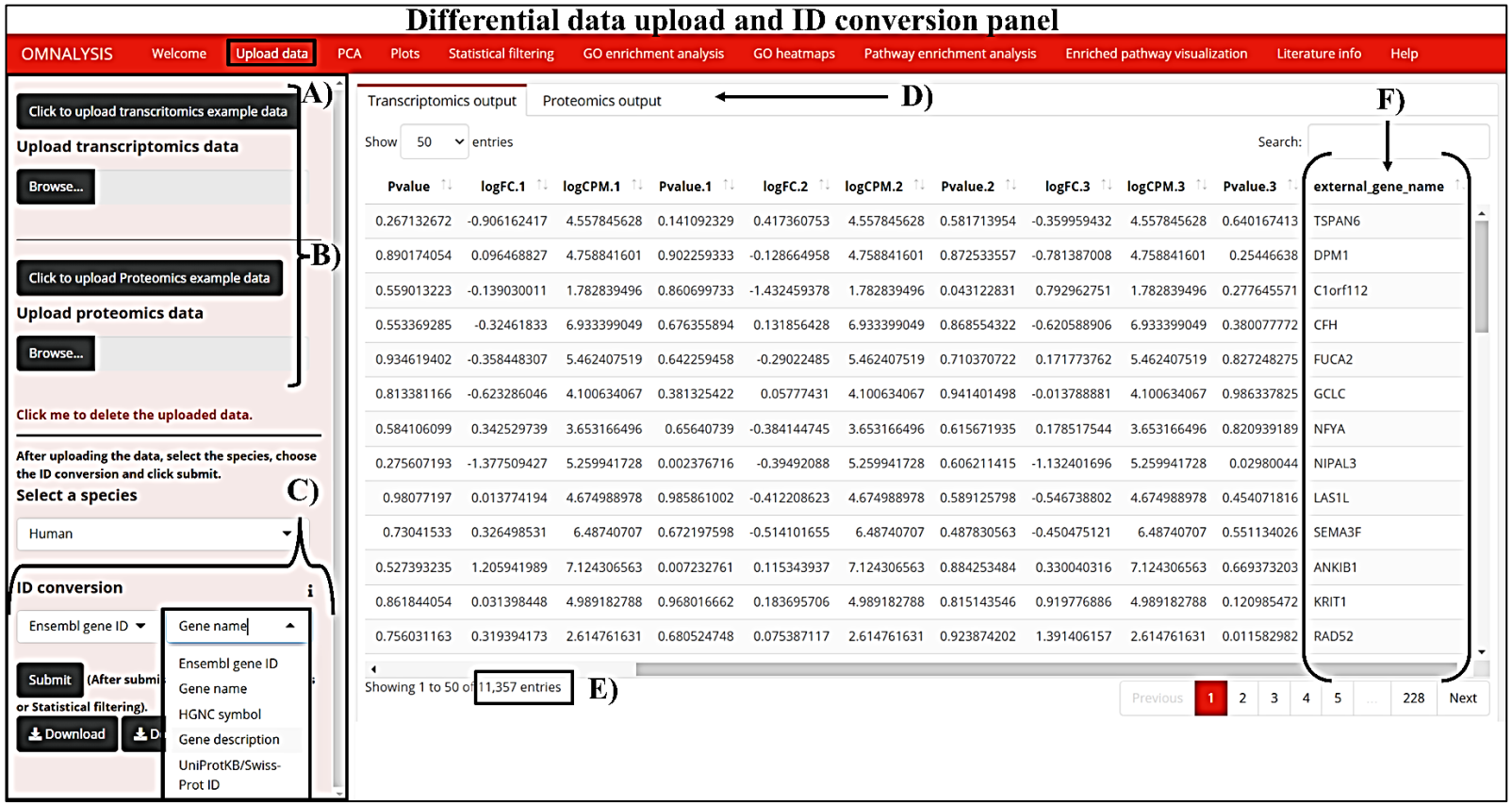
A-presents the features of OMnalysis application. B-financial support to develop OMnalysis. C-shows the types of visualization that can be generated using OMnalysis (carousal). D-systematic workflow to analyze the differential data, starting from ID conversion followed by biomarker discovery and finally scientific information.

5.2 Upload data and ID conversion

The tool was designed to handle up to four treatments at a time. To confirm this, four treatments were uploaded in a single run. The upload data section in OMnalysis is divided into two panels, left side interactive and right-side output. The interactive panel (Figure 9, A) is populated with tabs such as Click to upload transcriptomics or proteomics example data and upload transcriptomics or proteomics data (Figure 9, B), selection of species and others to explore the functionality of this section. OMnalysis supports four types of ID conversion, gene name, HGNC symbol, gene description and UniProtKB/Swiss-prot ID (Figure 9, C). By selecting the gene name and clicking on the submit button we performed ID conversion. OMnalysis is built to assign the most updated Ids (Figure 9, D). To the transcriptomic data matrix, the OMnalysis assigned 11,357 updated Ensembl Ids (Figure 9, E) from a total of 11,398 uploaded Ensembl ID. Next, to those updated IDs, OMnalysis was successful in assigning 10,951 human gene names (Figure 9, F), 10,932 HGNC symbols, 11,354 gene descriptions, and 7,463 UniProtKB/Swiss-prot ID. The result of transcriptomics ID conversion is presented in Supplementary information 1 (**http://gofile.me/6DOhe/WxhCFMzaU**).

In the case of proteomic data, 731 UniProt IDs were submitted to OMnalysis. Please note that these 731 IDs include several repeated UniProt IDs from 4 treatments. From this list, 281 UniProt IDs were mapped to Ensembl gene ID, 277 to the gene name, 0 to HGNC symbol, 273 to gene description and 273 to UniProtKB/Swiss-Prot ID. The result of proteomics ID conversion is presented in Supplementary information 1 (**http://gofile.me/6DOhe/WxhCFMzaU**).

Figure 9. Web interface to upload the differential expression data



A-interactive panel for users to provide input to OMnalysis. B-presents the input tabs to upload own or example expression data in CSV for transcriptomics or xlsx format for proteomics. C- shows the list of available IDs for conversion and Submit button to execute the ID conversion. D-presents the output window sectioned into transcriptomics and proteomics output to show the uploaded and converted data. E-shows the number of genes before and after ID conversion (before it was 11,398 and after 11,357). F-shows the results of ID conversion by adding a column of converted id at the last of the uploaded data.

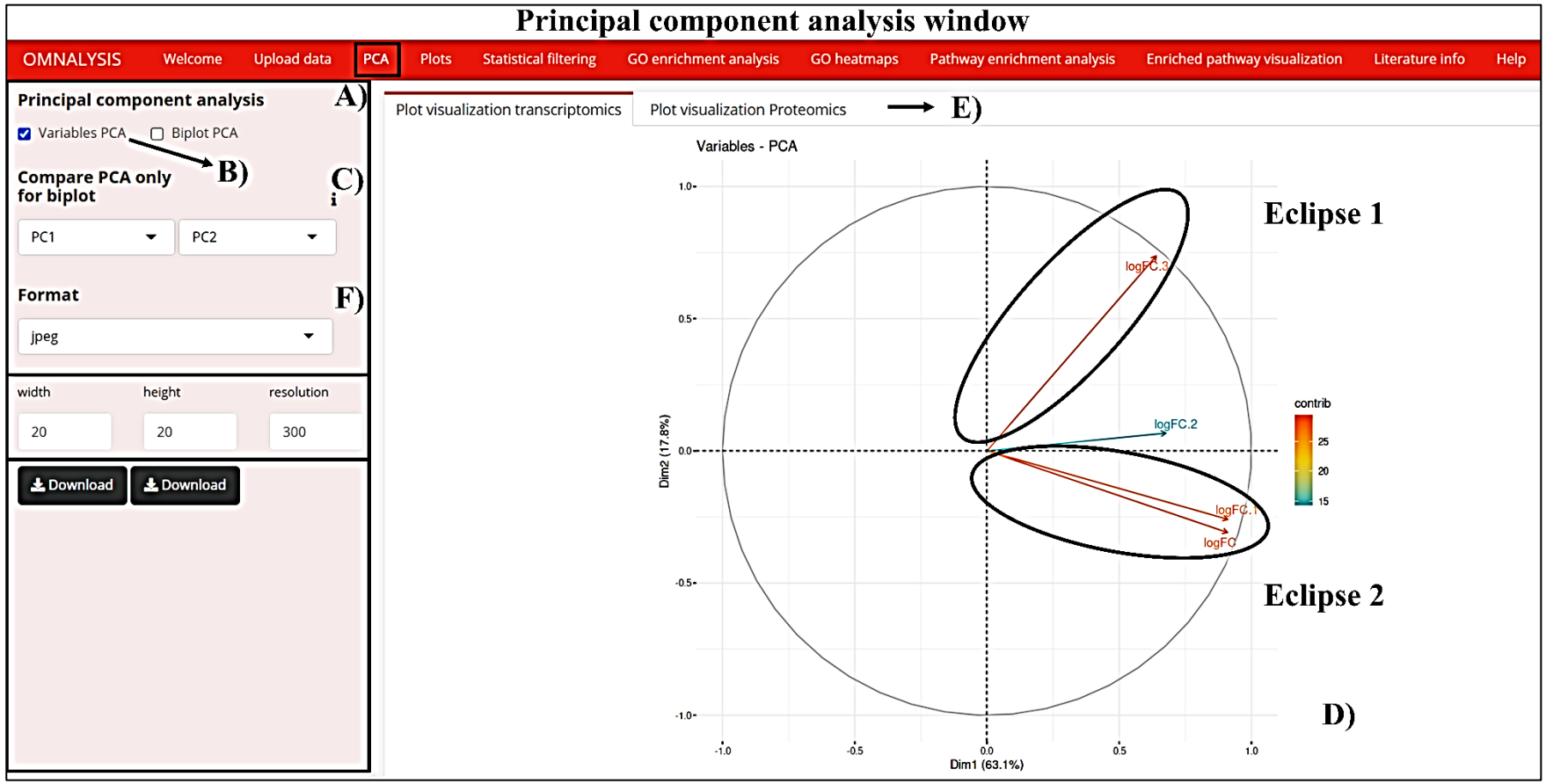
5.3 Principal component analysis (PCA)

After the ID conversion, the third section in OMnalysis is PCA, which is divided into an interactive and output panel. The interactive panel (Figure 10, A) is populated with Variable and Biplot PCA checkboxes (Figure 10, B) and principal components (PCs) drop-down menu tab (Figure 10, C) to compare the PCs. These checkboxes were employed to identify the variability and relationship between the genes in the treatments. The result from variable PCA transcriptomics (Figure 10 panel D) shows a high association between the variables (logFC and logFC.1) of treatment 1 and treatment 2 (Ellipse 2) that may be due to both pathogens are gram-positive having peptidoglycan cell wall (*Borrelia burgdorferi* and *Neisseria meningitidis*), whereas, logFC.3 distant. As shown in Figure 10, D (Ellipse 1 and 2), genes of treatment 1, 2 and 4 (logFC, logFC.1, logFC.3) contributed the most to the variability in the PCA, whereas, treatment 3 (logFC.2) contributed least among all.

Considering the importance of the association and cluster formation between the genes of treatments, we used the biplot PCA. Figure 11 presents the dimension reduction of the expression data by providing 80.9% of the variation in the first two principal components (PCs). The red, green, blue and purple colour in the biplot (Figure 11) presents each variable (logFC values and treatments). Out of 11,357 genes, the majority of them were cluster at 0 value in biplot PCA, showing similarities among the treatments in PC1 and PC2. Interestingly, gene CD38 and CLEC4E of treatment 2 and 4, respectively, show positive correlations in both PC1 and PC2 (Figure 11). For a better understanding, we performed a PCA comparison of (PC2 versus PC3) and (PC3 versus PC4). We observed 34.4% variability in biplot PC2 versus PC3, whereas, 19.1% in PC3 versus PC4. It was suggested that most of the information was shown in PC1 and PC2, and then its significance reduces gradually from PC1 to PC4. The complete result of the variable and biplot plot of transcriptomics is presented in Supplementary information 2 (**http://gofile.me/6DOhe/w5HfJ3qUa**).

In case of proteomics, the variable plot of proteomics treatment hour 36 (logFC) and 42 (logFC.1), 57 (logFC.2) and 81 (logFC.3) were highly correlated, respectively. From the variable plot, we identified 81-hour treatment contributed most, whereas, 42-hour treatment contributed least in the PC1 and PC2 variability.

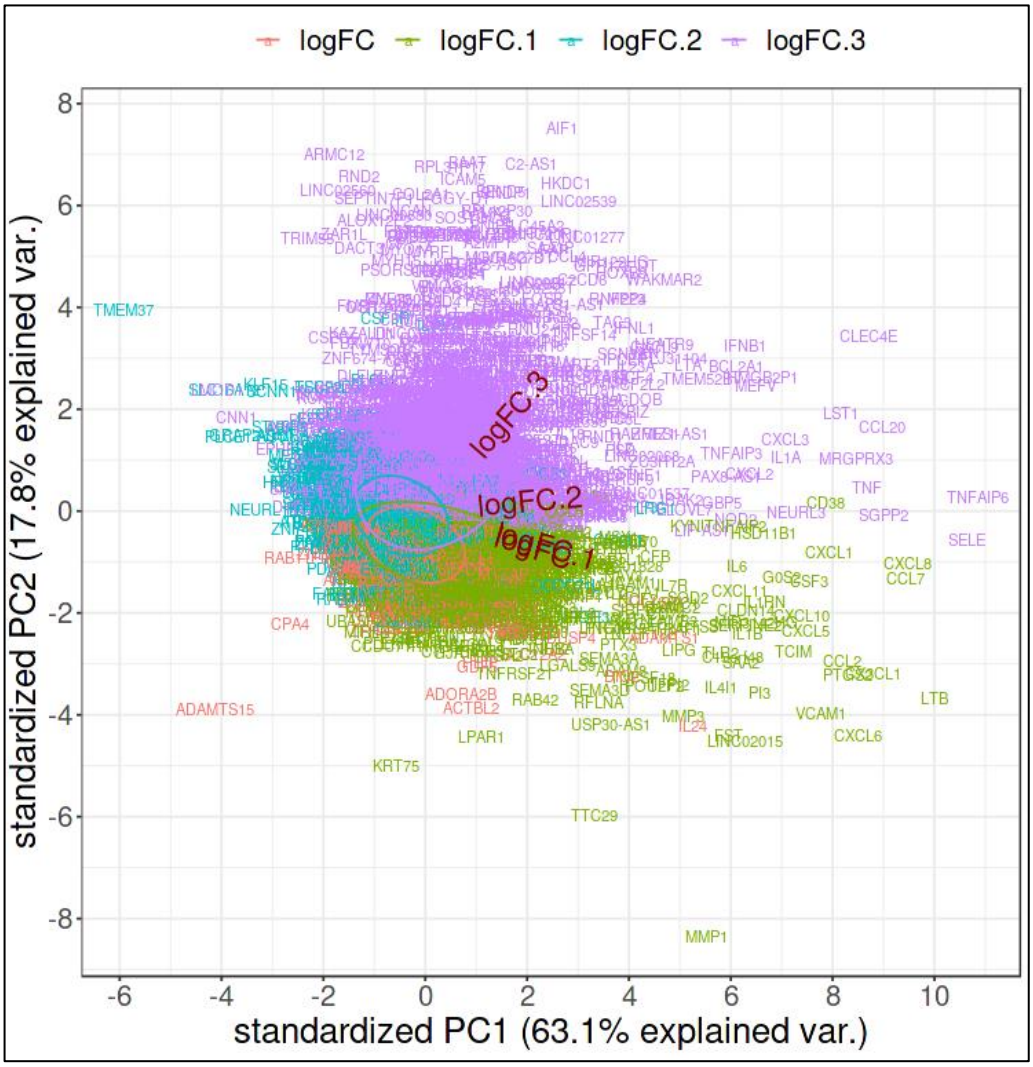
Figure 10. Web interface to perform a principal component analysis



A-interactive panel to allow a user to provide input. B-variable checkbox to generate a variable plot. C-drop down menu to compare the principal components generated in the biplot. D-is the output panel, presenting the variable plot. Eclipses in the output panel present the association between the variables (treatments) and contribution in the PCA. E-present the sectioned window to visualize PCA plots for transcriptomics and proteomics separately. F-input tabs to select image type, customize the dimension and resolution for PCA plot download.

In proteomics biplot PCA, we identified E1B999, P52176, Q2TBU0 and Q8SPP7 proteins in treatment 1(36h), 2(42h), 3(57h) and 4(81h) presented a positive correlation in both PCs. The complete result of the variable and biplot plot of proteomics is presented in Supplementary information 2 (**http://gofile.me/6DOhe/w5HfJ3qUa**).

Figure 11. Biplot of principal component analysis



This plot is showing the variable (logFC) and observation (genes) of transcriptomics data in a Biplot. The X and Y-axis present the principal component 1 and 2, respectively. Genes from treatment 1 (pink), treatment 2 (green), treatment 3 (aqua) and treatment 4 (purple) forming an overlap at a range of -2 to +2 in PC1 and PC2, showing the similarities. Gene CLEC4E of treatment 4 and CD38 of treatment 2 has the highest positive variability in PC1 and PC2. PC1 explain the 63.1% of data variability making it a good summary to measure, whereas, PC2 is highly influenced by the logFC.3 (treatment 4) and present 17.8% of data variability.

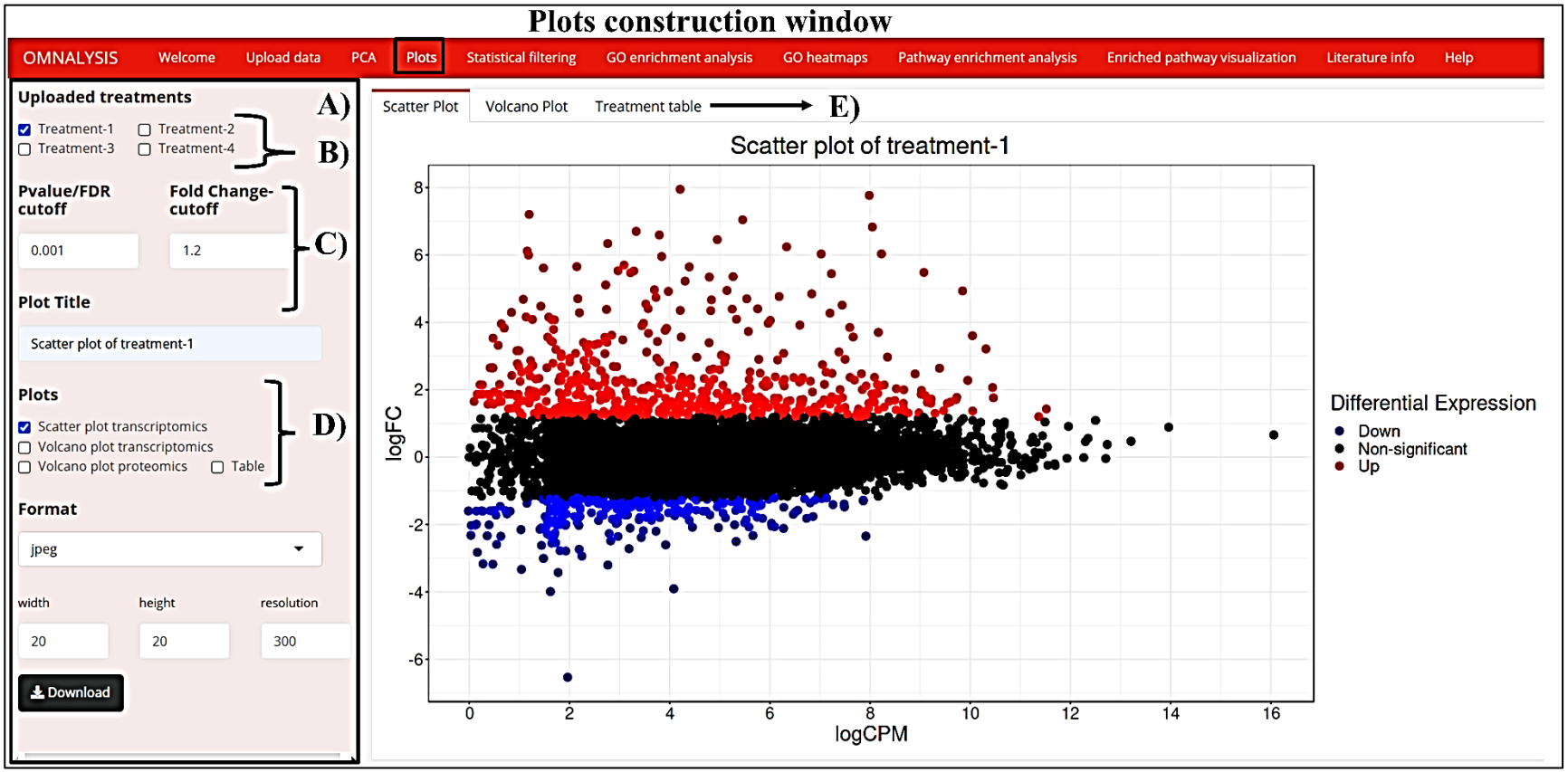
5.4 Plots

This is the fourth section in the OMnalysis tool. The interactive panel (Figure 12, A) consists of treatment selection (Figure 12, B), *P* value or FDR (Pvalue) and log fold change value cut off input tabs (Figure 12, C). To identify the differentially expressed up and down-regulated genes we applied 0.001 *P* value (Pvalue) and ±1.2 log fold change to 11,357 genes and plotted them using scatter plot transcriptomics or volcano plots transcriptomics checkboxes (Figure 12, D). To maximize the use of space, we divided the output panel into subsections to display plots and treatment tables separately (Figure 12, E). Using the table section, we checked the converted IDs of each uploaded treatment. Scatter and volcano plots of treatment 1 to 4 are presented in Supplementary information 3 (**http://gofile.me/6DOhe/d1VjMgSeX**).

Figure 13 compares the differential data of proteomics treatment 1 and 2 on the volcano plot, in which treatment 1 shows comparatively a lesser number of statistically significantly up-regulated genes, whereas, down-regulated genes were almost equal in both the treatments. Interestingly, the signaling molecule chemokine ligand (CXCL) and interleukin (IL) family was significantly up-regulated in both the treatments, showing the immune response induced by the *Borrelia burgdorferi* and *Neisseria meningitidis* pathogen.

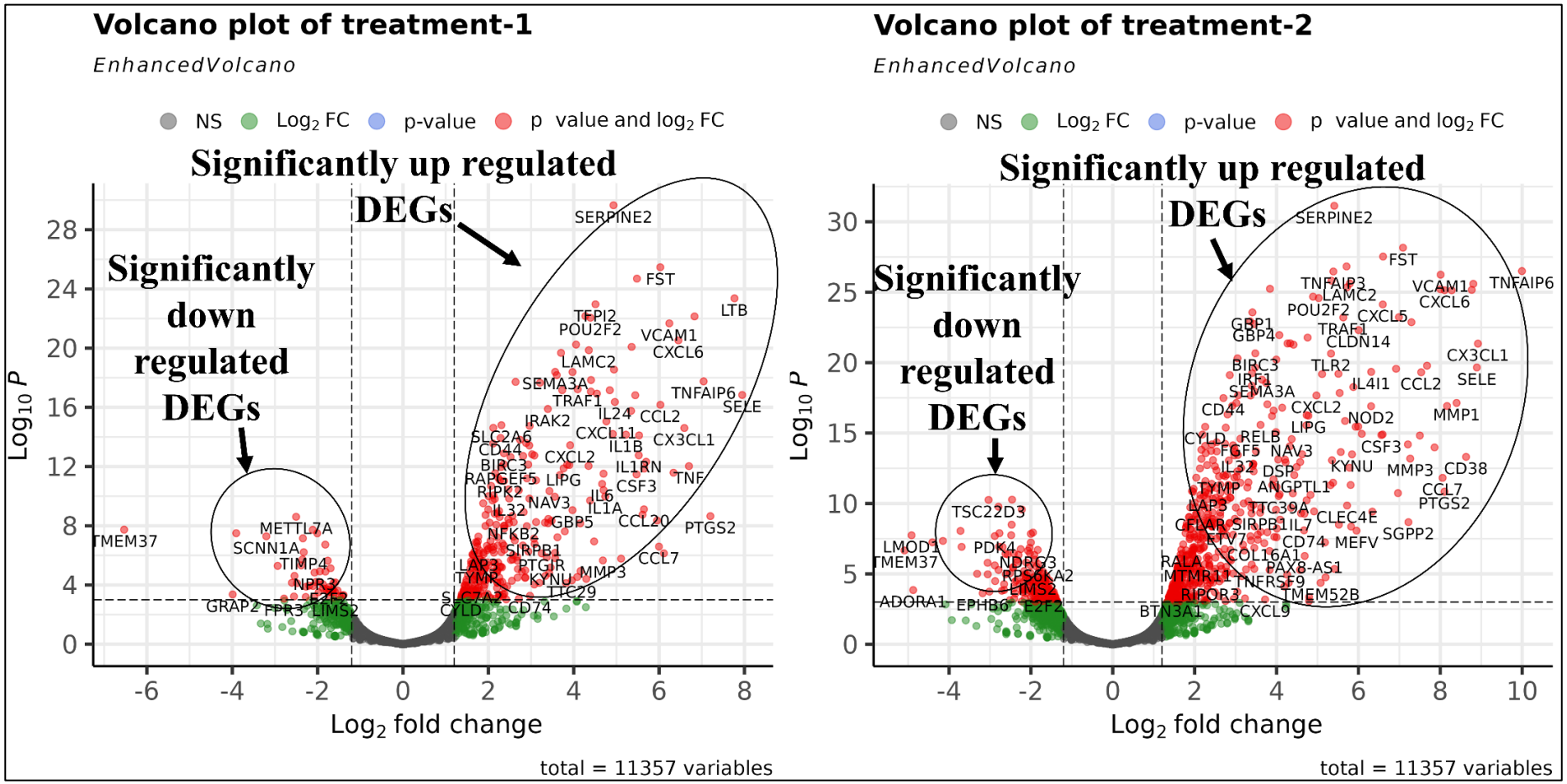
In the case of proteomics up and down-regulated proteins were visualized using the OMnalysis tool volcano plot proteomics checkbox. The volcano plot of 81-hours (treatment 4) and 57-hour treatment contain the highest number of observations after applying the threshold of log fold change ± 1.2 and Pvalue/FDR cutoff (FDR-adjusted P-value) 0.01. We identified Q8SPP7 protein as a highly up-regulated protein in all the proteomics treatments. In proteomics data logCPM column is absent, thus scatter plot visualization is beyond the capability of OMnalysis. The volcano plot of each treatment is presented in Supplementary information 3 (**http://gofile.me/6DOhe/d1VjMgSeX**).

Figure 12. Web interface for scatter and volcano plot



A-interactive panel that acts as an input to generate scatter and volcano plots for the uploaded treatments. B-each checkbox (1 to 4) is required to generate scatter and volcano plots for corresponding treatment. C-default Pvalue 0.001 and logFC ± 1.2 as a cutoff to show significant genes or proteins. D-multiple checkboxes to generate plots for transcriptomics and proteomics treatments. E- presents the sectioned output windows for Scatter Plot, Volcano Plot and Treatment table. Scatter Plot, showing red dots as up-regulated genes, blue dots as down-regulated genes and black dots as non-significant genes.

Figure 13. Volcano plot comparing treatment 1 and treatment 2



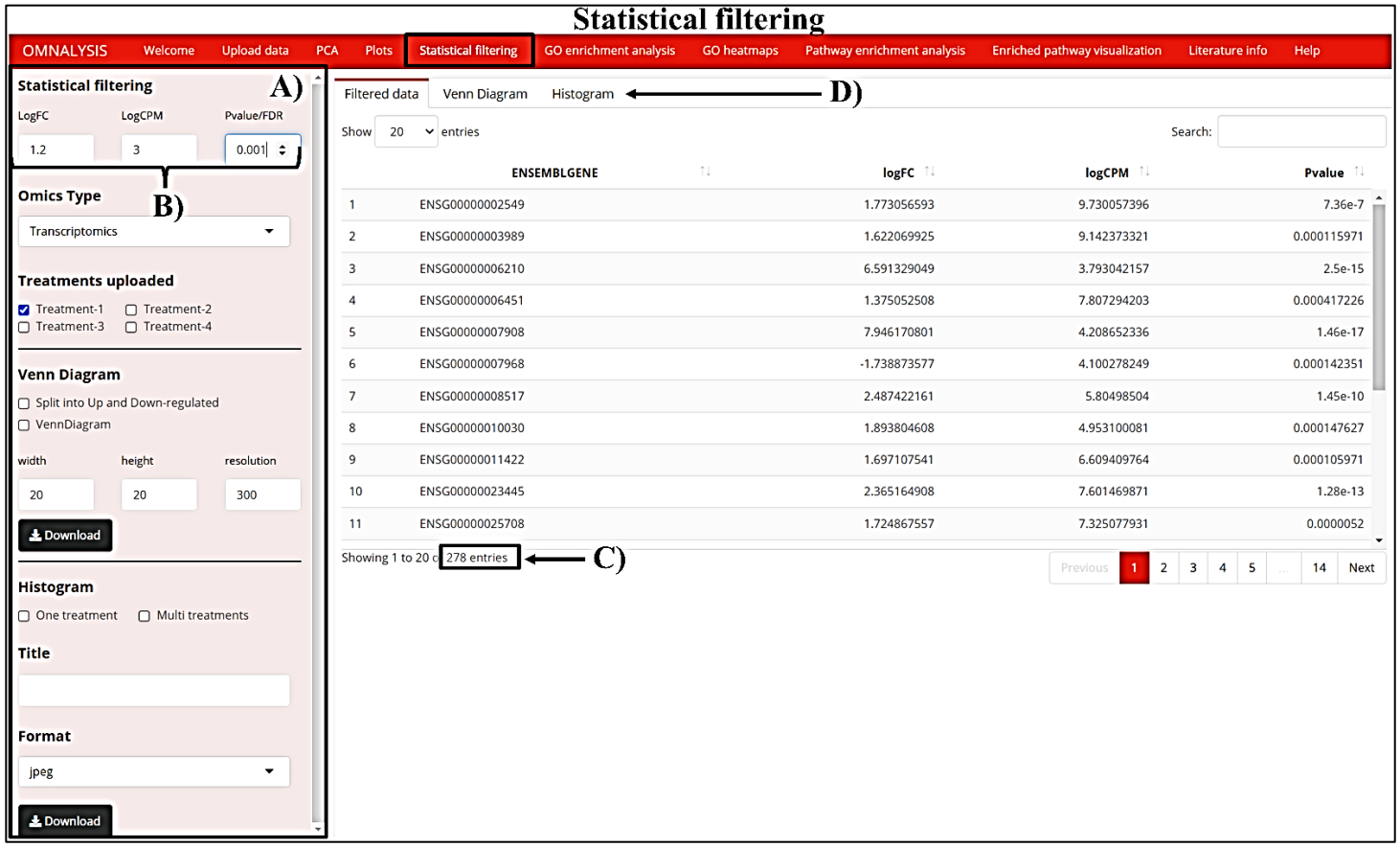
Volcano plots of transcriptomics treatment 1 and treatment 2 presenting the significantly up and down-regulated genes in red colour with their gene name. The grey and greens dots represent DEGs that are not significant at Pvalue > 0.001 (-log$_{10}$ 0.001 = 3) and logFC > 1.2 and < - 1.2.

5.5 Statistical filtering

This is the fifth section of the OMnalysis tool. The interactive panel (Figure 14, A) is populated with log count per million (logCPM), log fold change (logFC) and Pvalue/FDR numeric input boxes (Figure 14, B) to select differentially expressed genes (DEGs) or differentially expressed proteins (DEPs). We used logCPM greater than 3, logFC greater than +1.2 and less than -1.2 and Pvalue less than 0.001 thresholds to filter 11,357 genes. A total of 278 significantly DEGs were obtained in treatment 1 (Figure 14 C), 576 significantly DEGs in treatment 2, 69 significantly DEGs in treatment 3, whereas, 893 significantly DEGs were obtained in treatment 4. The complete set of transcriptomics statistically filtered results is presented in Supplementary information 4 (**http://gofile.me/6DOhe/miRakoBZ5**).

In the proteomics study, we applied logFC greater than +1.2 and less than -1.2, Pvalue (FDR-adjusted P-value) less than 0.01 threshold to the proteomics data. In total 74, 118, 214 and 264 statistically significantly DEPs were obtained in treatment 1, 2, 3 and 4, respectively. The complete set of proteomics statistically filtered results is presented in Supplementary information 4 (**http://gofile.me/6DOhe/miRakoBZ5**).

Figure 14. Web interface for filtering and diagram construction



A-presents the interactive panel used for filtering the DEGs and their visualization. B-to obtains statistically significant differentially expressed genes or proteins using logFC ± 1.2, logCPM >3 and Pvalue/FDR 0.001 values as a threshold. C-showing the number of statistically significantly DEGs obtained after applying the thresholds in treatment 1. D-display window is sectioned into three for separate visualization of Filtered data, Venn diagram, and Histogram.

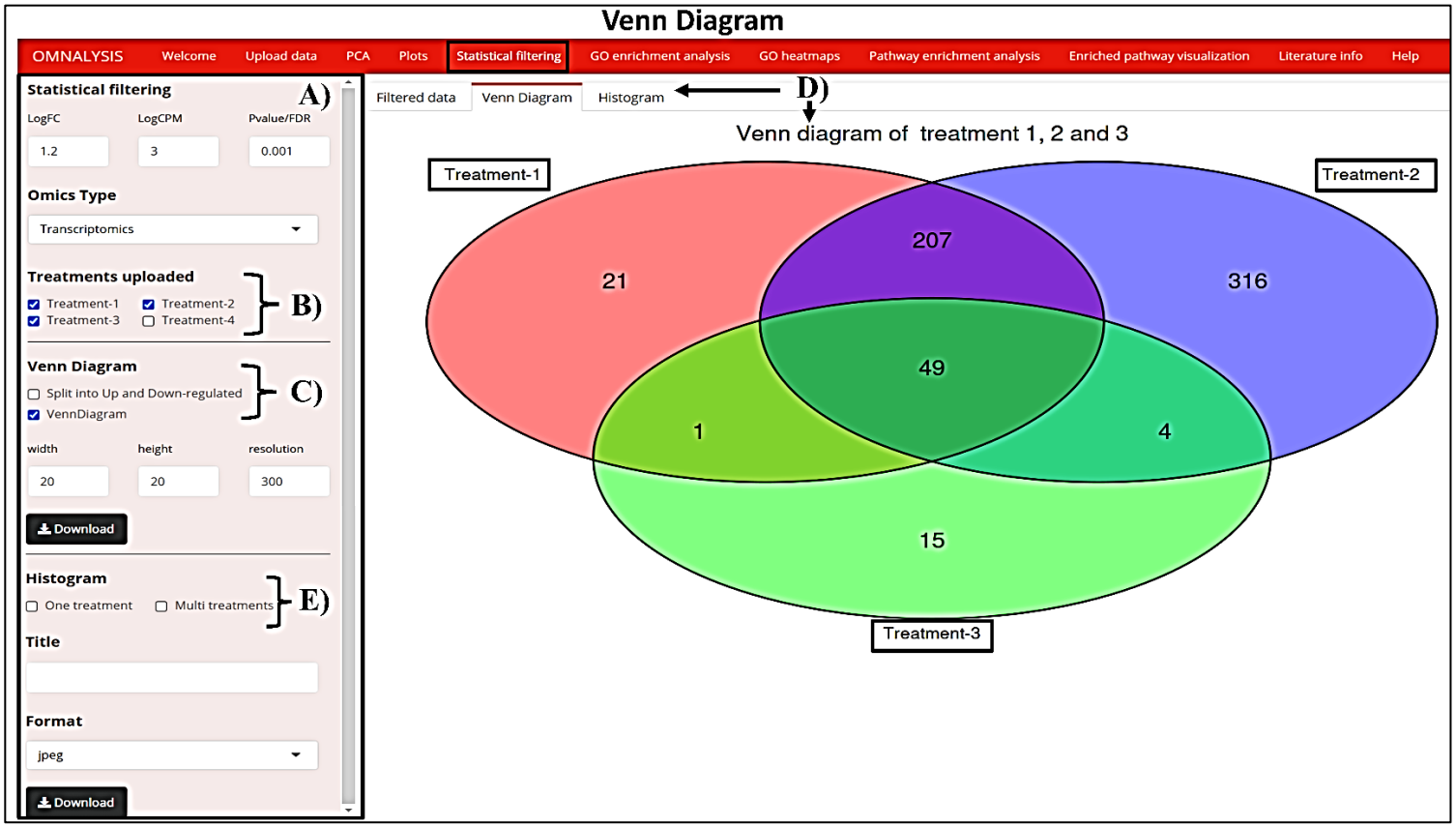
5.6 Venn diagram and histogram

This is the sixth section of the OMnalysis tool. The interactive panel (Figure 15, A) is populated with treatments (Figure 15, B), Venn diagram (Figure 15, C) and Histogram checkboxes. In the first three treatments, we obtained in total 923 significantly DEGs, among them, only 49 DEGs were common in all (Figure 15, D). Whereas, 256, 49, and 36, significantly DEGs were observed to be common in treatment (1 and 2), treatment (1, 2 and 3), and treatment (1, 2, 3 and 4), respectively. We used the default names of each circle (Treatment-1, 2 and 3) provided in the OMnalysis to differentiate the treatments in the Venn diagram (Figure 15, D).

Using the split into up and down checkbox we obtained in total 754 (278+576), 347 (278+69), 1171 (278+893), 645 (576+69), 1469 (576+893) and 962 (69+893) significantly DEGs in (treatment 1, 2), (treatment 1, 3), (treatment 1, 4), (treatment 2, 3), (treatment 2, 4) and (treatment 3, 4), respectively. From them a total of 222, 50, 122, 53, 186 and 41 were up regulated and 34, 0, 10, 0, 23 and 2 were down regulated proteins in (treatment 1, 2), (treatment 1, 3), (treatment 1, 4), (treatment 2, 3), (treatment 2, 4) and (treatment 3, 4), (Figure 16). For better understanding, we used default names Up regulated 1 and Down regulated 1 for treatment 1, and Up regulated 2 and Down regulated 2 for treatment 2 (Figure 16). Detailed Venn diagrams of transcriptomics treatment 1, 2, 3 and 4 were presented in Supplementary information- 5 (**http://gofile.me/6DOhe/bpBl6fYxP**).

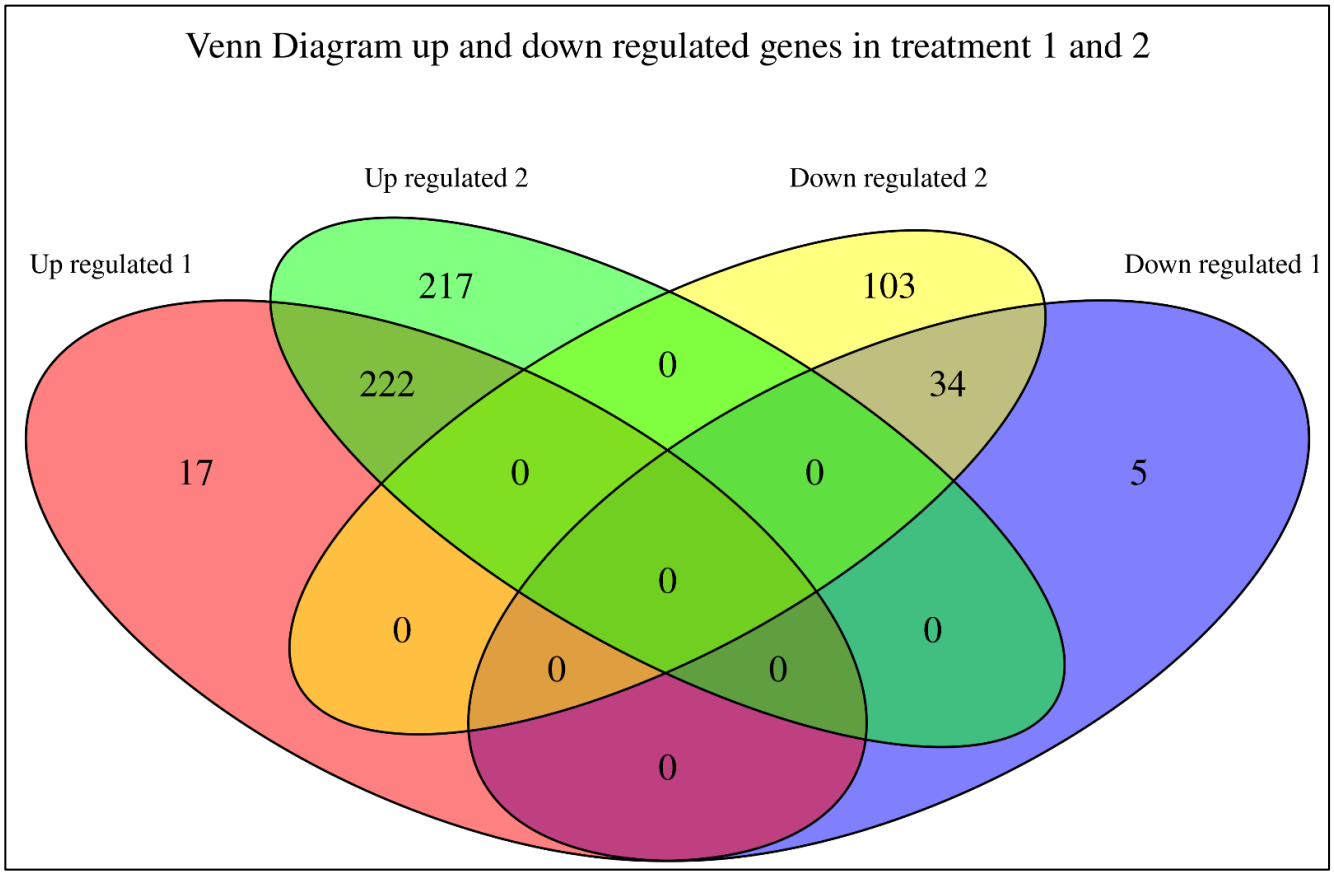
In the proteomics study, we found 68 in treatment 1 and 2, 49 in treatment 1, 2 and 3, 46 in treatment 1, 2, 3 and 4 significantly abundant common protein. Whereas, in split up and down regulation Venn diagram, we identified a total of 192 (118+74), 288 (214+74), 338 (264+74), 322 (118+214), 382 (118+264) and 478 (214+264) significantly abundance protein in (treatment 1, 2), (treatment 1, 3), (treatment 1, 4), (treatment 2, 3), (treatment 2, 4) and (treatment 3, 4), respectively. From them a total of 51, 49, 47, 82, 77 and 151 were up regulated and 17, 21, 18, 22, 19, and 33 were down regulated in (treatment 1, 2), (treatment 1, 3), (treatment 1, 4), (treatment 2, 3), (treatment 2, 4) and (treatment 3, 4). Detailed Venn diagrams of proteomics treatments 1, 2, 3 and 4 were presented in Supplementary information 5 (**http://gofile.me/6DOhe/bpBl6fYxP**).

Figure 15. Web interface for construction and visualization of Venn diagram and histogram



A-interactive panel for statistical filtering, Venn diagram and Histogram generation and download. B- treatment checkboxes to select treatments, in which, checkboxes treatment 1, 2 and 3 are selected to generate the Venn diagram. C-checkboxes to generate different types of Venn diagrams. D-presents the output with common DEGs present in treatments 1, 2 and 3 (49 genes were common) and sub sections to display filtered data, Venn diagram and Histogram. E-checkboxes to generate different types of histograms. Overlap of the colors shows the union between the significantly DEGs of each treatment.

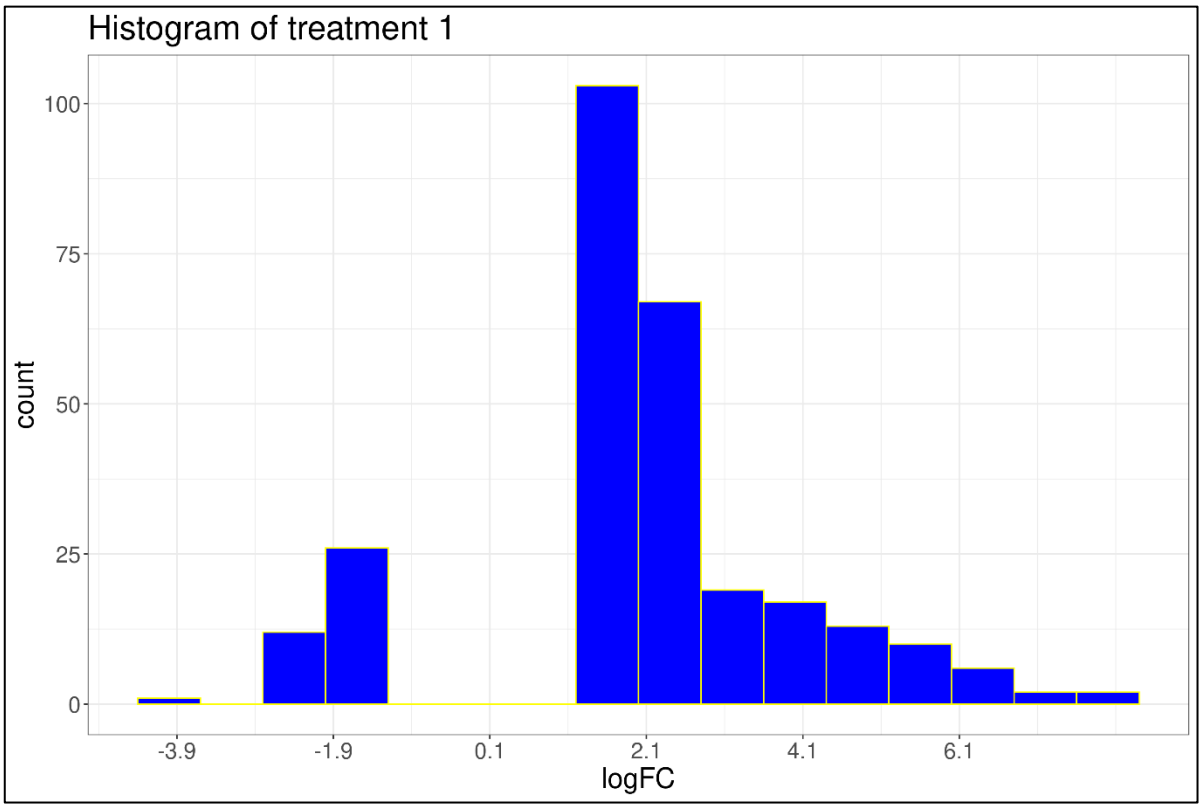
Figure 16. Graphical presentation of statistically differentially expressed genes (DEGs)



Graphical presentation of common up and down-regulated significantly DEGs in transcriptomics treatment 1 and treatment 2 ((222+17+34+5) + (217+222+103+34)) = 754). Each ellipse presented up and down-regulated DEGs in treatments 1 and 2. The number and overlap colours of the Venn diagram provides the common DEGs present in the treatments. In treatments 1 and 2, 222 and 34 DEGs were identified as common up-regulated and down-regulated.

To generate a histogram of the statistically significantly DEGs and proteins of each treatment, we used the histogram checkbox of the statistical filtering section (Figure 15, E). Figure 17 presents the distribution of up and down-regulated significantly DEGs between 1.5 to 8 and -1.4 to -3.9, respectively. Due to the page restriction, the histograms of transcriptomics and proteomics treatments are presented in Supplementary information 6 (**http://gofile.me/6DOhe/6YykI9FmR**).

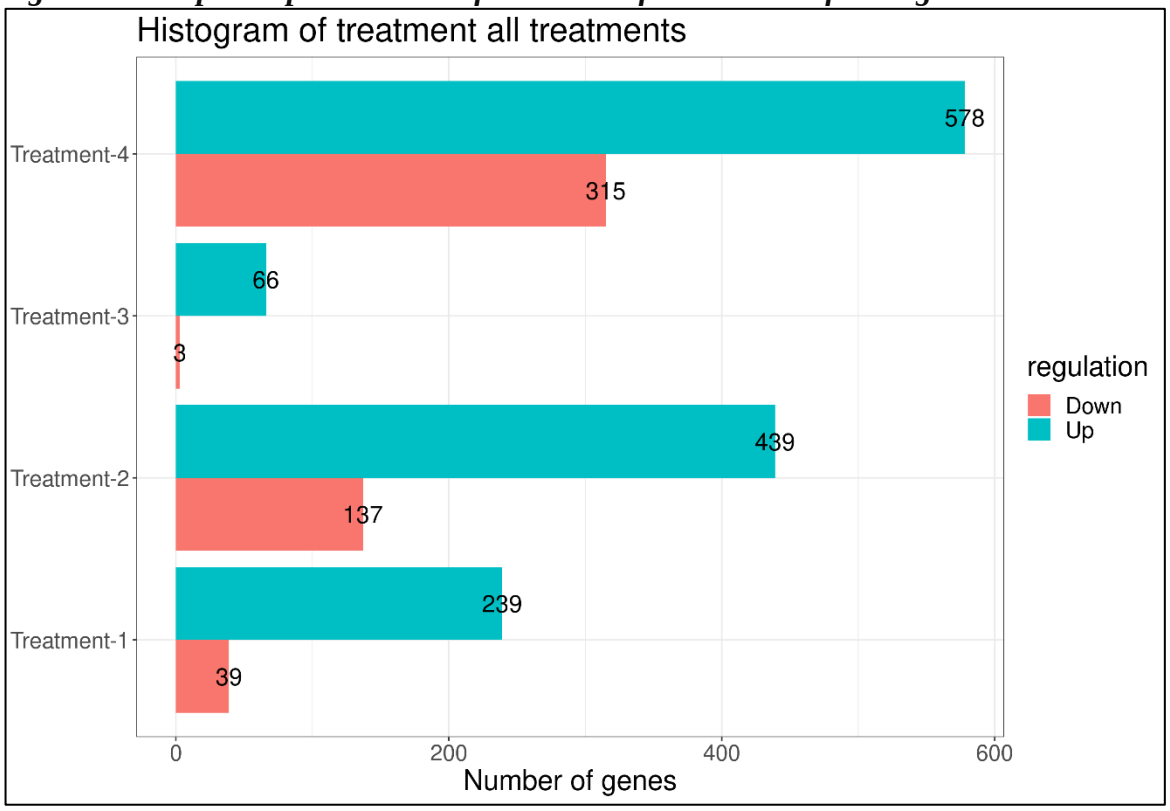
Figure 17. Graphical presentation of DEGs range



Histogram showing the frequency of the significantly DEGs lying on the log fold change in transcriptomics treatment 1. After statistical filtering, DEGs log fold change range reduces to -3.9 and 8 for down and up-regulated significantly DEGs, respectively. The highest number of up-regulated DEGs were observed at 1.6 logFC, whereas, -1.9 for down-regulated DEGs.

For the publication purpose, we used an all-treatment histogram checkbox to generate a single histogram that displays all the up and down-regulated significantly DEGs or proteins in treatments 1, 2, 3 and 4 (Figure 18). As shown in Figure 18, 278 (239 up and 39 down-regulated), 576 (439 up and 137 down-regulated), 69 (66 up and 3 down-regulated) and 893 (578 up and 315 down-regulated) significantly DEGs were obtained in transcriptomics treatment 1, 2, 3 and 4, respectively.

Figure 18. Graphical presentation of all DEGs after statistical filtering



The histogram presents the total number of up and down-regulated significantly DEGs obtained after applying statistical filtration in all the treatments. Red bars (down-regulated significantly DEGs) and blue bars (up-regulated significantly DEGs). The X-axis is a number of genes and the Y-axis is treatments. The number of significantly DEGs displayed on the histogram head.

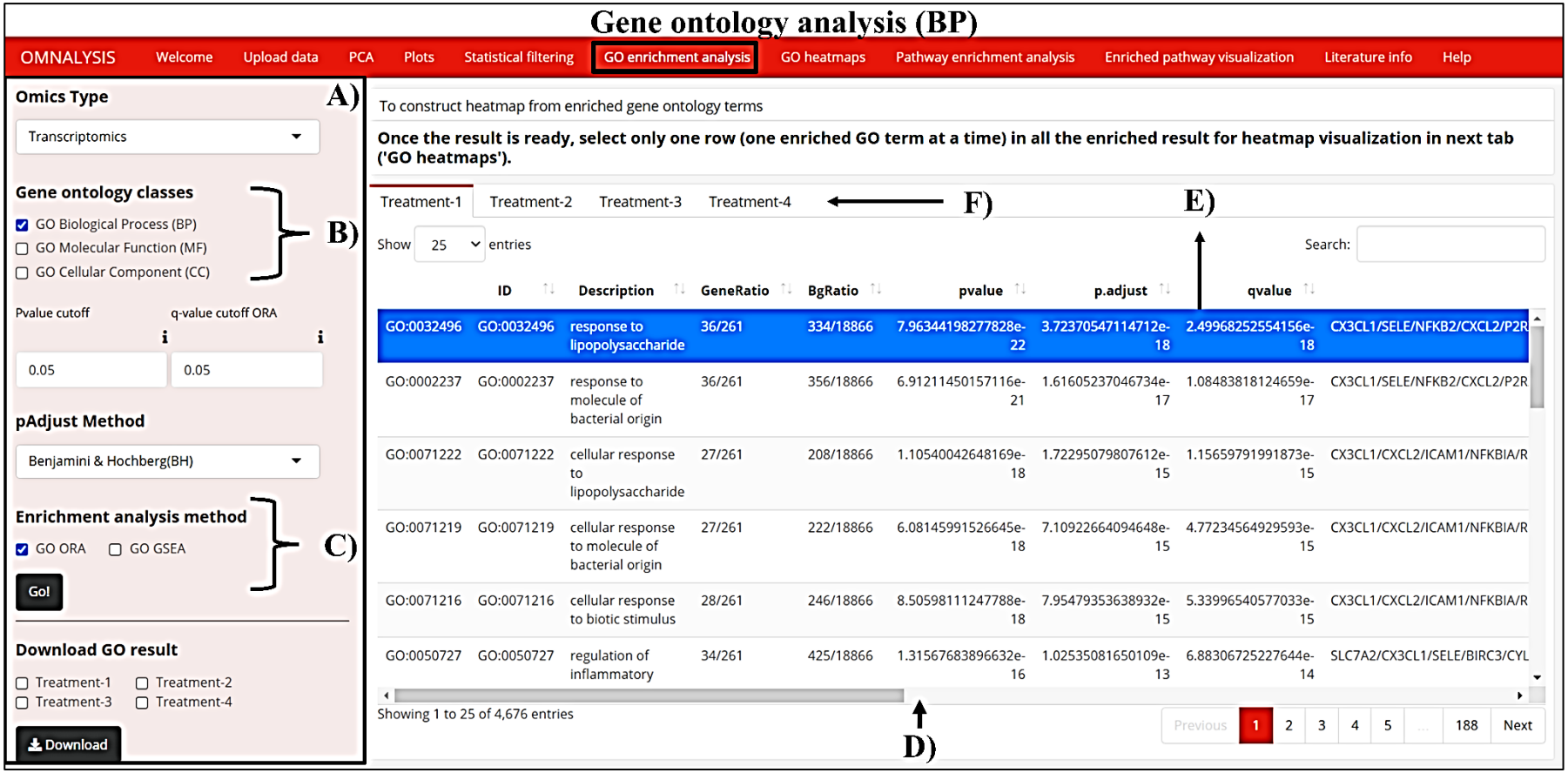
In the proteomics study, we observed the distribution range of significantly up-regulated proteins between 2.3 to 12, 2.3 to 13.3, 1.4 to 14.5 and 1.2 to 15 in treatment 1, 2 3 and 4, respectively. Whereas, significantly down-regulated proteins range between -1.7 to -7.7, -1.7 to -7.7, -2.6 to -12.6 and -1.2 to -12.8 in treatment 1, 2, 3 and 4, respectively. From, all treatment histograms we identified (52 up, 22 down), (92 up, 26 down), (164 up, 46 down) and (222 up, 38 down) regulated proteins in treatment 1, 2, 3 and 4, respectively. Due to the page restriction, the histograms of transcriptomics and proteomics treatments are presented in Supplementary information 6 (**http://gofile.me/6DOhe/6YykI9FmR**).

5.7 Gene ontology enrichment analysis

This is the sixth section of the OMnalysis web application. We used an interactive panel (Figure 19, A) to segregate the significantly differentially expressed genes into gene ontology classes (Biological processes (BP), Molecular function (MF) and cellular activity (CC)), (Figure 19, B). We subjected statistically significant DEGs to biological process (BP), using the ORA (Figure 19, C) and in total 4,676 terms were enriched in treatment 1 (Figure 19, D). Among them, response to lipopolysaccharides was significantly enriched at p-value cutoff 0.05 and q-value cutoff 0.05. The detailed result of the analysis in which Column 9 presents the significantly DEGs (36) mapped to the response to lipopolysaccharide and response to molecule of bacterial origin term in treatment 1 (Table 10, A). A comparison of the result revealed that due to the difference in BgRatio (Table 10 Column 4, 334/18866 and 356/18866), we obtained a response to lipopolysaccharides as the first hit; 334 and 356 present the size of the gene set annotated directly or indirectly to the term. To get a holistic picture of the biological activities, we used the other two classes of gene ontology (molecular function and cellular components) at Pvalue 0.05 cutoff and q-value cutoff 0.05. In the molecular function, we observed that 31 and 41 DEGs were overlapped to cytokine activity and receptor ligand activity. The analyzed complete result of transcriptomics treatment 1, 2, 3 and 4 using ORA and GO classes (BP, MF, and CC) is provided in Supplementary information 7 (**http://gofile.me/6DOhe/8QRQK46hJ**).

In proteomics using ORA and BP gene ontology class, we obtained sterol import and defense response as a top hit at Pvalue cutoff 0.05 and q-value cutoff 0.05 in treatment 1 (Table 10, B). The analyzed complete result of proteomics treatment 1, 2, 3 and 4 using ORA and GO classes (BP, MF, and CC) is provided in Supplementary information 7 (**http://gofile.me/6DOhe/8QRQK46hJ**).

Figure 19. The web interface of gene ontology analysis



A-interactive panel is a user interface for performing a functional interpretation of significantly DEGs or proteins. B-presents the checkboxes to select the gene ontology class. C- is to select the enrichment method (ORA or GSEA) and clicking on a Go button to execute the analysis. D-presents the result of treatment 1 in the transcriptomics study. E-indicates a biological process selected by the user for the next section of OMnalysis (i.e., GO heatmaps). F- sections to display the result of each treatment result (Treatment-1, 2, 3 and 4).

Table 10. A detailed portion of gene ontology enrichment result

(A) Top 3 enriched biological process terms using over-represented analysis method for treatment 1 transcriptomics								
Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9
ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0032496	response to lipopolysaccharide	36/261	334/18866	7.96E-22	3.72E-18	2.50E-18	CX3CL1/SELE/NFKB2/CXCL2/P2RX7/ICAM1/NFKBIA/RIPK2/CSF3/CCL2/NFKB1/IL1A/TNFAIP3/CXCL6/IL1B/GCH1/IRAK2/IL6/TLR2/SMAD6/TIMP4/IL24/CLDN1/CXCL3/CXCL5/CXCL1/ZC3H12A/LGALS9/CXCL10/CXCL11/CXCL8/TRIB1/LTA/TNF/LYN/CCL5	36
GO:0002237	response to molecule of bacterial origin	36/261	356/18866	6.91E-21	1.62E-17	1.08E-17	CX3CL1/SELE/NFKB2/CXCL2/P2RX7/ICAM1/NFKBIA/RIPK2/CSF3/CCL2/NFKB1/IL1A/TNFAIP3/CXCL6/IL1B/GCH1/IRAK2/IL6/TLR2/SMAD6/TIMP4/IL24/CLDN1/CXCL3/CXCL5/CXCL1/ZC3H12A/LGALS9/CXCL10/CXCL11/CXCL8/TRIB1/LTA/TNF/LYN/CCL5	36
GO:0071222	cellular response to lipopolysaccharide	27/261	208/18866	1.11E-18	1.72E-15	1.16E-15	CX3CL1/CXCL2/ICAM1/NFKBIA/RIPK2/CSF3/CCL2/NFKB1/IL1A/TNFAIP3/CXCL6/IL1B/IRAK2/IL6/TLR2/IL24/CXCL3/CXCL5/CXCL1/ZC3H12A/CXCL10/CXCL11/CXCL8/TRIB1/TNF/LYN/CCL5	27
(B) Top 3 enriched biological process terms using over-represented analysis method for treatment 1 proteomics								
GO:0035376	sterol import	3/16	9/4535	2.99E-06	0.001242	0.000719	APOA1/APOC3/CD36	3
GO:0070508	cholesterol import	3/16	9/4535	2.99E-06	0.001242	0.000719	APOA1/APOC3/CD36	3
GO:0006952	defense response	8/16	338/4535	6.65E-06	0.001841	0.001066	APOA1/CATHL3/CATHL1/SERPINF2/CATHL4/VIM/CFB/PGLYRP1	8

Top three biological processes that were enriched in treatment 1 transcriptomics and treatment 1 of proteomics study using the over-represented method. The 1st and 2nd columns of the table are GO term ID and biological process name, respectively. Column 3rd is the Gene Ratio present size of unique DEGs or proteins overlaps against a unique set of genes present in the background distribution. Column 4th is Bg Ratio indicates the size of the annotated gene set against the all-unique collection of gene sets. The genes that were mapped to the go terms from the DEGs are shown in column 8. Column 9 indicate the number of genes mapped to a GO term.

Another method, gene set enrichment analysis (GSEA) was performed on the list of significantly DEGs sorted in descending order by log fold change (logFC) values. In the biological process, genes were overlapped in the response to bacterium term with positive normalized enrichment score (NES) in transcriptomics treatment 1 (Figure 20). It is noteworthy that, top two enriched terms provided evidence of hBMECs (human brain microvascular endothelial cells) responded to bacteria molecules when induced with *Borrelia burgdorferi*. Total 1,004 gene ontology terms were enriched in treatment 1 using GSEA and Benjamini & Hochberg (BH) pAdjustment method at Pvalue cutoff 0.5. The result of transcriptomics treatment 1, 2, 3 and 4 using GSEA and GO classes (BP, MF, and CC) is presented in Supplementary information 8 (**http://gofile.me/6DOhe/lHBHknqZk**).

In the case of proteomics, we performed GSEA using a list of treatment 1 proteins and obtained defense response to other organisms in BP, hydrolase activity in MF and extracellular matrix in CC as first hits with positive normalized enrichment score (NES). We also identified that a small number of significantly DEGs or proteins often fails to generate any result using the GSEA, thus, we used less stringent values (Pvalue cutoff range from 0.5 to 1). The result of proteomics treatment 1, 2, 3 and 4 using GSEA and GO classes (BP, MF, and CC) is presented in Supplementary information 8 (**http://gofile.me/6DOhe/lHBHknqZk**).

Figure 20. A part of the result using a GSEA.

Col. 1	Col. 2	Col. 3	Col. 4	Col. 4	Col. 5	Col. 5	Col. 6	Col. 7	Col. 8	Col. 9
ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalues	rank	leading_edge	core_enrichment
GO:0009617	response to bacterium	49	0.656379	2.631569	1.96E-10	4.15E-07	3.77E-07	68	tags=63%, list=24%, signal=58%	ENSG00000007908/ENSG00000169429/ENSG00000232810/ENSG00000006210/ENSG00000124875/ENSG00000108691/ENSG00000115009/ENSG00000163739/ENSG00000108342/ENSG00000169245/ENSG00000163735/ENSG00000125538/ENSG00000162892/ENSG00000166920/ENSG00000169248/ENSG00000115008/ENSG00000136244/ENSG00000137462/ENSG00000163734/ENSG00000154451/ENSG00000081041/ENSG00000166523/ENSG00000118503/ENSG00000226979/ENSG00000134070/ENSG00000168961/ENSG00000125730/ENSG00000090339/ENSG00000163874/ENSG00000173334/ENSG00000171236
GO:0071219	cellular response to molecule of bacterial origin	27	0.746545	2.655924	2.38E-10	4.15E-07	3.77E-07	67	tags=81%, list=24%, signal=68%	ENSG00000169429/ENSG00000232810/ENSG00000006210/ENSG00000124875/ENSG00000108691/ENSG00000163739/ENSG00000108342/ENSG00000169245/ENSG00000163735/ENSG00000125538/ENSG00000162892/ENSG00000169248/ENSG00000115008/ENSG00000136244/ENSG00000137462/ENSG00000163734/ENSG00000081041/ENSG00000118503/ENSG00000134070/ENSG00000090339/ENSG00000163874/ENSG00000173334

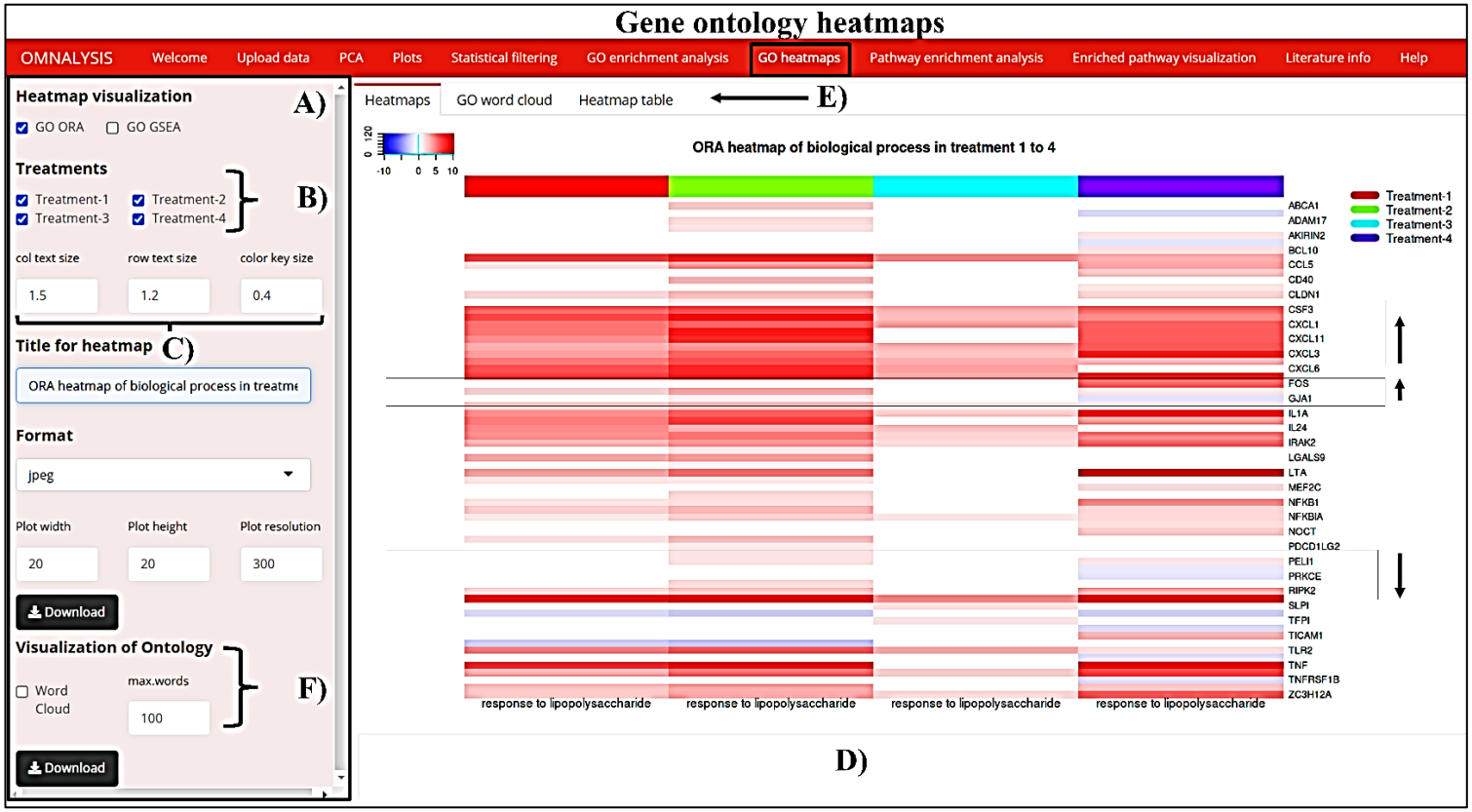
The portion of the enriched biological terms using gene set enrichment analysis of treatment 1 transcriptomics. The Col.1 and Col.2 of the table are GO term ID and biological process name, respectively. Col.3 presents the common genes in gene sets and expression data. Col.4 indicates the score of set size over-representation against the ranked set of the gene (sorted DEGs according to log fold change). Col.5 to Col.6 indicate the statistical values computed to find significant terms. Col.8 provides the percentage of contribution of genes to the enrichment score. Col. 9 lists the genes contributing to a biological process in Ensembl ID. Col.-column.

5.8 Gene ontology heatmaps

We used the seventh section of OMnaysis (GO heatmaps) to visualize and compare the DEGs or proteins that were overlapped to the specific biological processes using ORA or GSEA in interactive panel (Figure 21, A). We considered response to lipopolysaccharides term in transcriptomics treatment 1, 2, 3 and 4 (Figure 21, B) as significantly enriched using the ORA and selected it to compare the expression (log fold change) of overlapped DEGs on the heatmap. We resized the text in column and row and colour key size (Figure 21, C). From Figure 21, we observed that CXCL family genes were up-regulated in all four treatments. Also, the IL1 A gene is up-regulated with logFC values 5.7, 5.35, 2.33 and 9.8 in treatment 1, 2, 3 and 4, respectively, indicating the production of inflammation against the pathogen. However, DEGs such as SLPI and TNFRSF1B were downregulated or not evoked by the pathogens in all treatments (Figure 21, D). To visualize the result of each method, the display section is subdivided (Figure 21, E). The transcriptomics heatmaps of gene ontology terms enriched in classes (BP, MF, CC) using the ORA is presented in Supplementary information 9 (**http://gofile.me/6DOhe/8nZ0FFJs0**).

In proteomics, we considered defense response in biological process, identical protein binding in molecular function and extracellular space in cellular component for heatmaps generation using the ORA. The proteomics heatmaps of gene ontology terms enriched in classes (BP, MF, CC) using the ORA is presented in Supplementary information 9 (**http://gofile.me/6DOhe/8nZ0FFJs0**).

Figure 21. The web interface of enriched term visualization and comparison

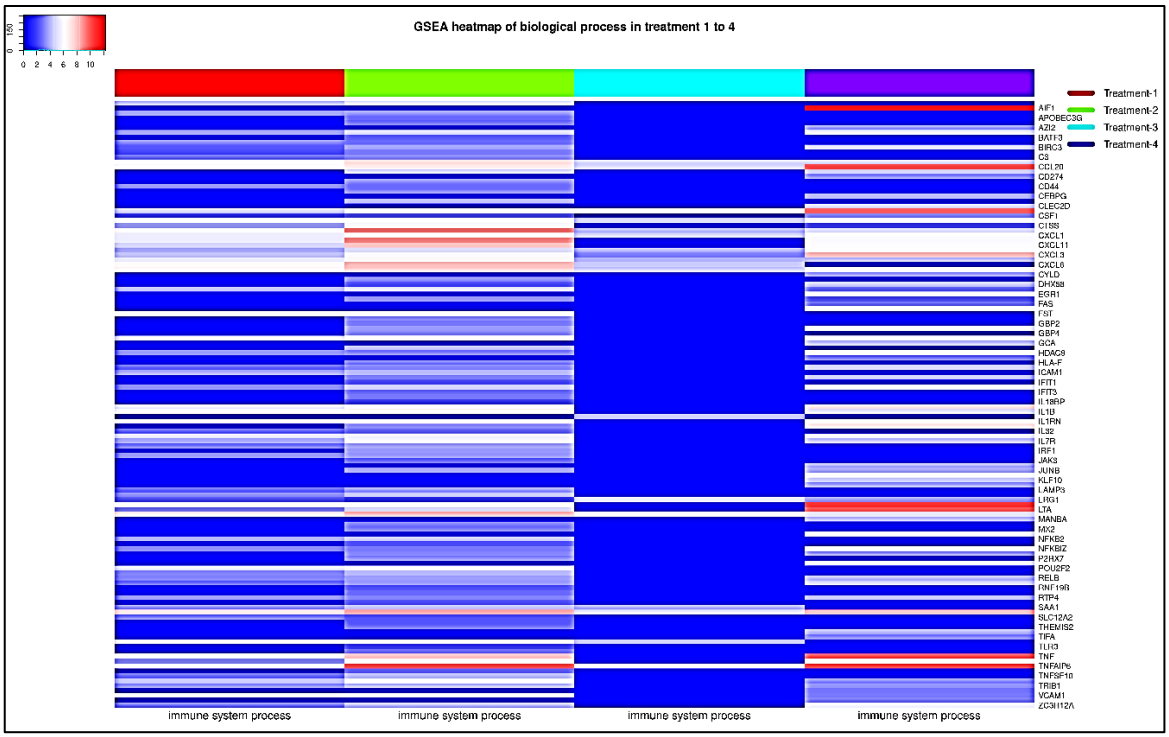


A-interactive panel provides multiple input options to generate, customize and download heatmap and word cloud. B-checkboxes to select the treatments, 1, 2, 3 and 4 were checked to generate the heatmap. C-provides numeric inputs to increase or decrease the font and color key size of the heatmap. D-presents generated heatmap with pseudo colors, red–up regulation, blue–down regulation and white - no change in the regulation of DEGs in response to the lipopolysaccharides biological process selected in the previous section of OMnalysis. E- shows the sectioned display window for Heatmaps, GO word cloud and Heatmap table. F-checkboxes to visualize GO terms in the word cloud and the number of GO terms. The range of the logFC is -10 to 10.

In transcriptomics using GSEA, we observed that the CXCL and Interleukin family genes were highly expressed in treatments 1, 2, 3 and 4, whereas, no down-regulated genes were identified in the immune system biological process. The transcriptomics heatmaps of gene ontology terms enriched in classes (BP, MF, CC) using the GSEA is presented in Supplementary information 10 (**http://gofile.me/6DOhe/qV0K8QqAo**).

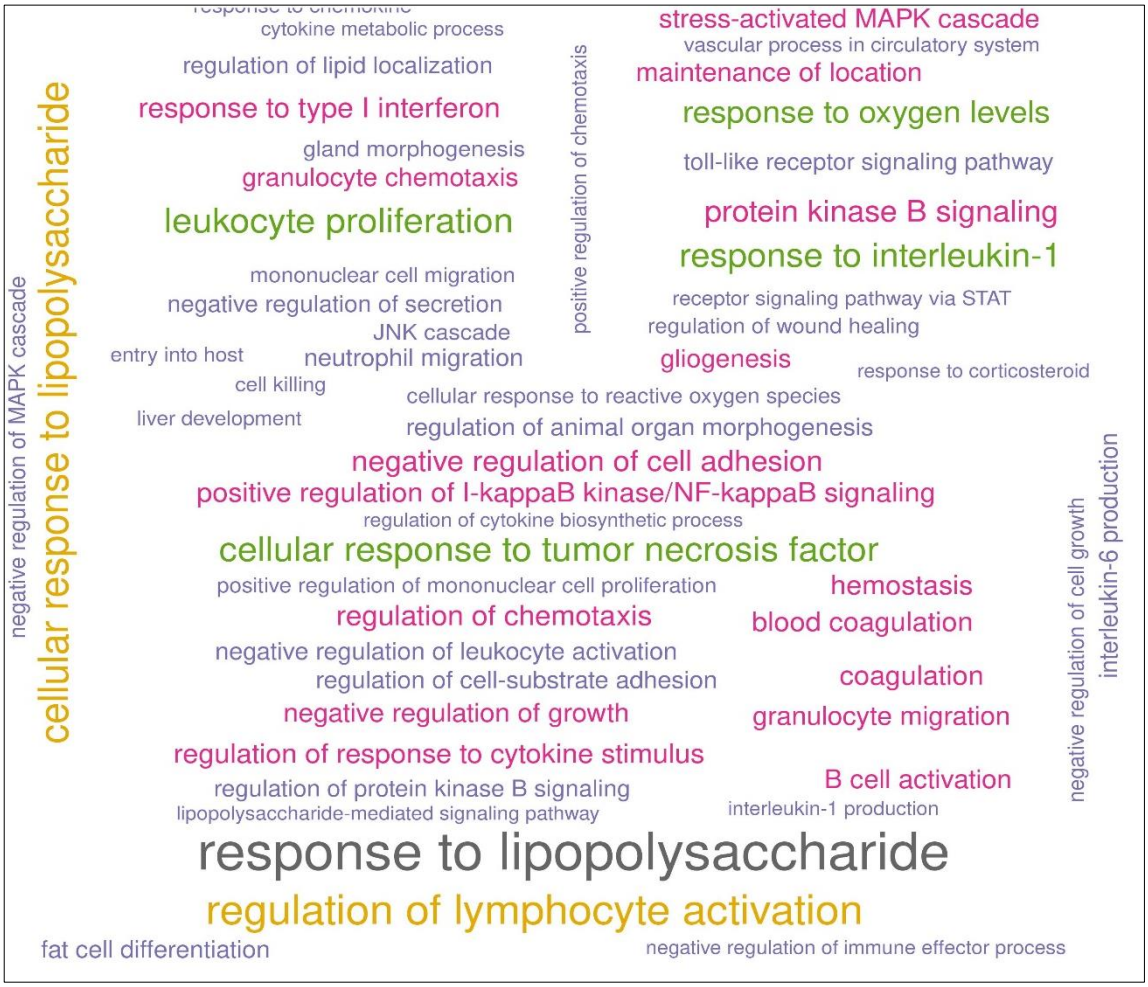
In proteomics, using GSEA we considered defense response to other organism in biological process, catalytic activity in molecular function and extracellular region in cellular component GO terms enriched in all treatments. In defense response to other organism biological process terms, we identified CATHL3, CATHL4, PGLYRP1 genes were highly expressed and no down-regulated genes were identified. The proteomics heatmaps of gene ontology terms enriched in classes (BP, MF, CC) using the GSEA is presented in Supplementary information 10 (**http://gofile.me/6DOhe/qV0K8QqAo**).

Figure 22. Heatmap comparing the biological processes



Heatmap showing the top biological process comparison between the uploaded treatments (1, 2, 3 and 4) using the GSEA. Legend red, green, aqua and blue represents treatments 1,2,3 and 4, respectively. Colour key scale to view a range of fold change (logFC values, i.e., 0 to 10), Showing the genes mapped to immune system process were highly expressed and no downregulation of genes was observed.

Figure 23. Word cloud of enriched biological terms



Word cloud showing the top hundred enriched biological process terms obtained in transcriptomics treatment 1 using the result of the ORA. Response to lipopolysaccharides and regulation of lymphocyte activation shows a larger font size concerning others.

Further, we used the word cloud option from the interactive panel to display a large number of enriched gene ontology terms (Figure 21, F). The size of the biological terms shown in Figure 23 was according to the frequency (higher the number of genes bigger the font of the text in the word cloud). The word cloud of gene ontology terms enriched in classes (BP, MF, CC) using the ORA and GSEA is presented in Supplementary information 9 (**http://gofile.me/6DOhe/8nZ0FFJs0**) and 10 (**http://gofile.me/6DOhe/qV0K8QqAo**), respectively.

5.9 Pathway enrichment analysis

The eighth section of OMnalysis is designed to perform a number of pathway enrichment analyses using the significantly differentially expressed genes (DEGs) or

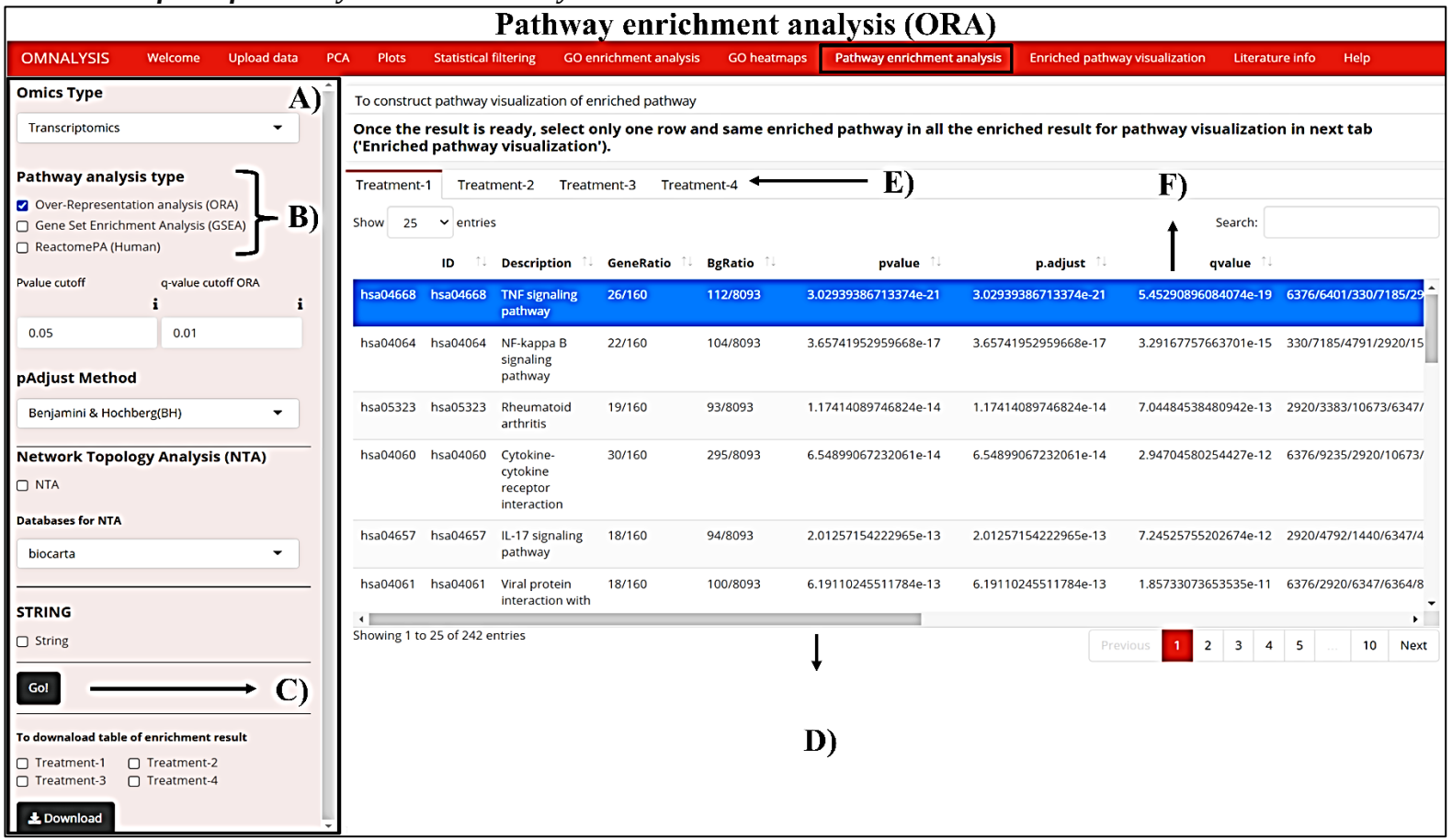
proteins. This section is divided into the interactive panel (Figure 24, A), used to unveil the mechanistic insight of the DEGs with the help of four types of pathway enrichment analysis methods (ORA, GSEA, and ReactomePA) (Figure 24, B) and action button Go (Figure 24, C).

Using ORA at Pvalue cutoff 0.05 and the FDR correction method as Benjamini and Hochberg (BH), we obtained the TNF signaling pathway as significantly enriched in the first transcriptomics treatment. A total of 242 pathways were enriched (Figure 24, D). The output panel is divided into subsections to visualize the analyzed results of treatments 1, 2, 3 and 4 (Figure 24, E). Table 12 provides the partial result of the pathways enriched using the ORA in transcriptomics treatment (1, 2, 3 and 4). The detailed result is provided in Supplementary information 11 (**http://gofile.me/6DOhe/UpAWrQ5nF**).

In Proteomics, we performed ORA at Pvalue cutoff 0.05 and observed Staphylococcus aureus infection and NOD-like receptor signaling pathway in top 5 enriched terms in proteomics treatment 1, 2, 3 and 4. The complete result of pathway enrichment in proteomics treatments using ORA is presented in Supplementary information 11 (**http://gofile.me/6DOhe/UpAWrQ5nF**).

To get more valuable insight we used an interactive panel (Figure 25, A), the second method GSEA (Figure 25, B) to perform the analysis. We used 0.05 as Pvalue cutoff and Benjamini and Hochberg (BH) FDR correction method, which resulted in a total of 45 significantly enriched pathways using the *KEGG* database (Figure 25 panel C).

Figure 24. Web-interface of Pathway enrichment analysis

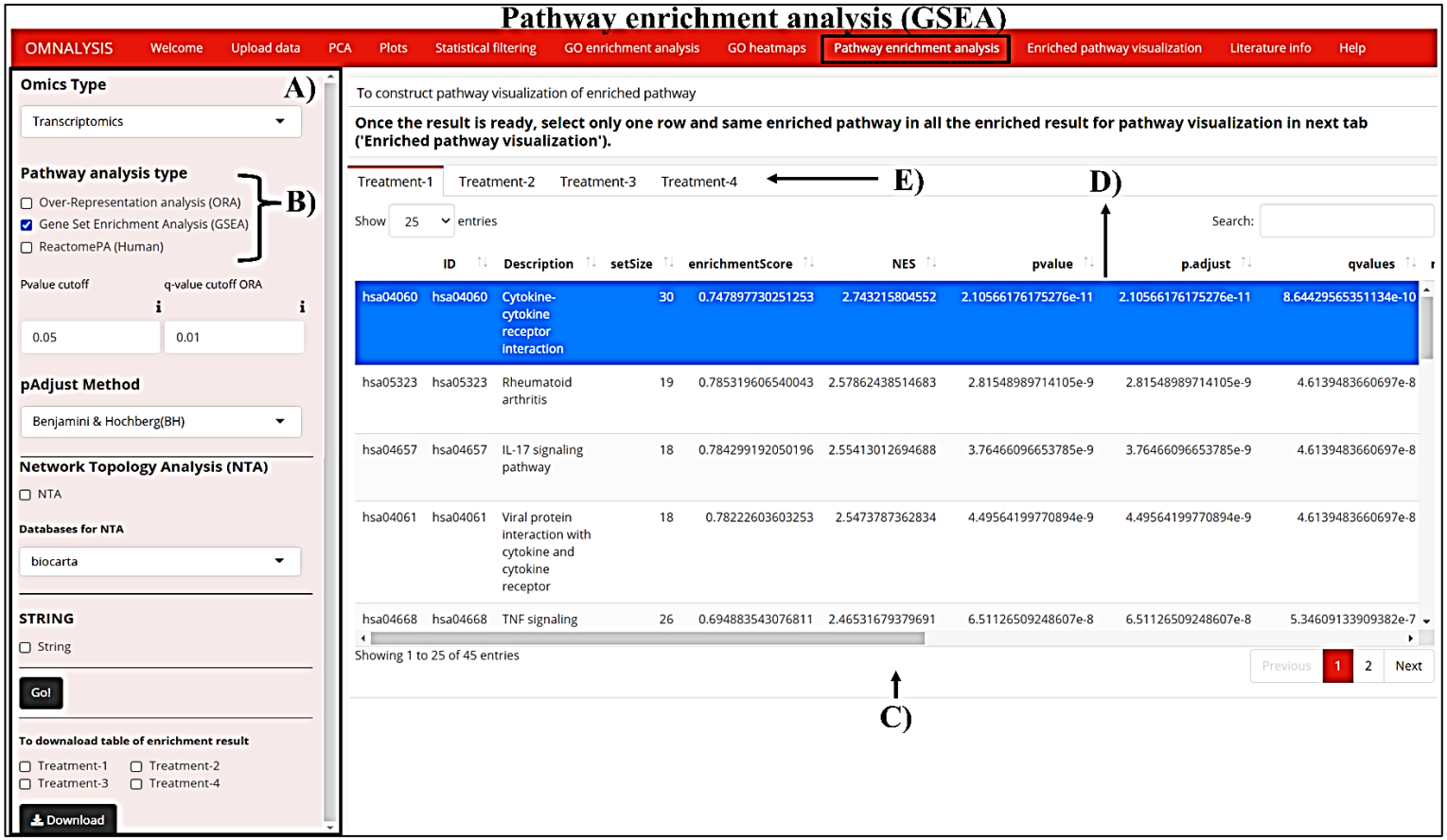


A-interactive panel for user input options. B-checkboxes to perform pathway enrichment analysis, in which ORA was checked. C- action button is provided to execute the analysis after the selection of omics type, enrichment method, cutoff value and pAdjust method. D-result of treatment 1 using ORA. E-provides sub sections to show the results of the analysis of each treatment (treatment 1, 2, 3 and 4). F-shows the selected enriched pathway to visualize the overlapped DEGs in the next section of OMnalysis (i.e., Enriched pathway visualization).

Table 11. Table comparing the enriched pathways in uploaded treatments.

Top 3 pathways enriched using ORA analysis for treatment 1		
Description	geneID (Entrez)	Count
TNF signaling pathway	6376/6401/330/7185/2920/3383/4792/182/6347/4790/6364/7128/6372/3659/3553/3569/7412/2921/6374/2919/3627/3726/1435/4049/7124/6352	26
NF-kappa B signaling pathway	330/7185/4791/2920/1540/3383/4792/10673/5971/23586/4790/7128/5328/3553/7412/2921/2919/3576/4049/4050/7124/4067	22
Rheumatoid arthritis	2920/3383/10673/6347/51561/3552/6364/6372/3553/3569/7097/2921/6374/2919/3576/1435/4050/7124/6352	19
Top 3 pathways enriched using ORA analysis for treatment 2		
TNF signaling pathway	8837/6376/6401/330/7185/602/2920/3383/4792/182/8717/6347/4790/6364/7128/6372/3659/3553/3569/7412/2921/6374/2919/197259/3627/3726/1435/4049/7124/6352	30
NF-kappa B signaling pathway	8837/330/7185/4791/2920/1540/3383/4616/4792/958/10673/8717/5971/23586/4790/7128/5328/3553/7412/2921/2919/3576/4615/4049/4050/7124/4067	27
Epstein-Barr virus infection	1870/864/960/4791/4938/3383/4616/4792/958/8717/5971/3718/23586/4790/4940/4939/6772/7128/3569/7097/6502/4794/10018/3455/6890/3627/6773/4615/9636/6891/3134/3105/7124/3106/4067/9641	36
Top 3 pathways enriched using ORA analysis for treatment 3		
TNF signaling pathway	6376/6401/602/2920/3383/4792/6347/6364/8809/7128/6372/3569/7412/2921/6374/2919/1435/9021	18
Rheumatoid arthritis	2920/3383/6387/6347/3552/6364/6372/3569/7097/2921/6374/2919/3576/1435/4050	15
Viral protein interaction with cytokine and cytokine receptor	6376/2920/6387/6347/6364/8809/6372/3569/11009/2921/6374/2919/3576/1435	14
Top 3 pathways enriched using ORA analysis for treatment 4		
TNF signaling pathway	8837/6376/6401/330/355/7133/5291/7185/7132/602/2920/3383/4792/208/6347/4790/329/6364/8809/7128/3659/3553/3569/7412/2921/6374/2919/197259/3627/2353/3726/3725/1435/9021/6300/4217/4049/7124/6352	39
NF-kappa B signaling pathway	8837/330/7185/7132/4791/2920/1540/3383/4616/4792/10673/5971/4790/329/1647/7128/3553/148022/7099/8915/7412/2921/2919/3576/4049/4050/7124	27
NOD-like receptor signaling pathway	330/2920/57506/4938/4792/8767/4793/6347/4790/24145/329/2635/2633/7128/3553/148022/3569/10010/7099/3708/115362/2634/115361/3428/2921/2919/3576/3725/29110/6300/7124/10628/6352	33

Figure 25. Web-interface of Pathway enrichment analysis using GSEA



A-is an interactive panel for providing input to the OMnalysis pathway enrichment section. B-is for enrichment analysis options with a Pvalue cutoff of 0.05, Gene Set Enrichment Analysis (GSEA) was checked. C-enrichment result of the treatment 1. D-presents the selected enriched pathway to visualize DEGs in the next section of OMnalysis (i.e., Enriched pathway visualization). E-provides subsections to displays the pathway enrichment result of treatments 1, 2, 3 and 4.

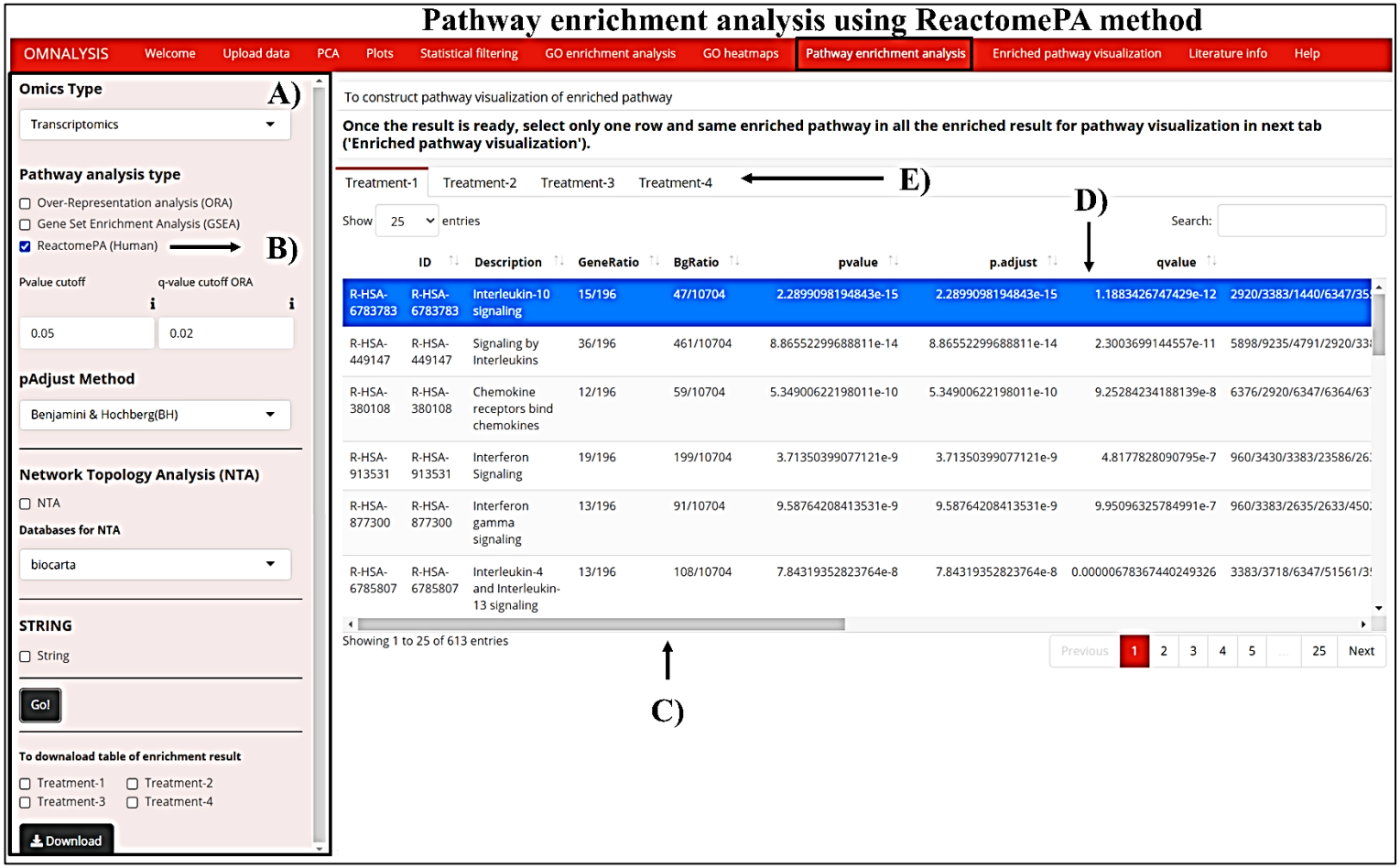
The most significant pathway identified as cytokine-cytokine receptor interaction (Figure 25, D) in treatment 1. The detailed result of treatments 1, 2, 3 and 4 (Figure 25, E) is presented in Supplementary information 12 (**http://gofile.me/6DOhe/G3T6Ei6nB**).

In proteomics treatments using GSEA, we identified the NOD-like receptor signaling pathway as the top enriched term with a positive normalized enrichment score (NES) in all four treatments. The complete result of proteomics GSEA pathway enrichment is presented in Supplementary information 12 (**http://gofile.me/6DOhe/G3T6Ei6nB**).

OMnalysis is also equipped with a peer-reviewed pathway analysis database Reactome. In the interactive panel (Figure 26, A), we selected the ReactomePA (Figure 26, B) checkbox at Pvalue cutoff 0.05 for treatment 1. In total 613 pathways were enriched using the DEGs of treatment 1 against the Reactome database (Figure 26, C). Among them, the top hit was interleukine-10 signaling in treatment 1 (Figure 26, D). The complete result of Reactome pathway analysis of treatment 1, 2, 3 and 4 (Figure 27, E) is provided in Supplementary information 13 (**http://gofile.me/6DOhe/xvCDxPW1O**).

Part of the result of the Reactome pathway analysis performed using treatments 1, 2, 3 and 4 is shown in Table 12. We observed that alpha-beta and interleukins pathways were highly enriched in treatments 1, 2, 3 and 4, providing evidence of a relationship between the cellular immune response against pathogens. We selected the enriched interleukine-10 signaling pathway to identify the expression of DEGs overlapped to it, in each treatment.

Figure 26. Web interface to perform Reactome pathway analysis



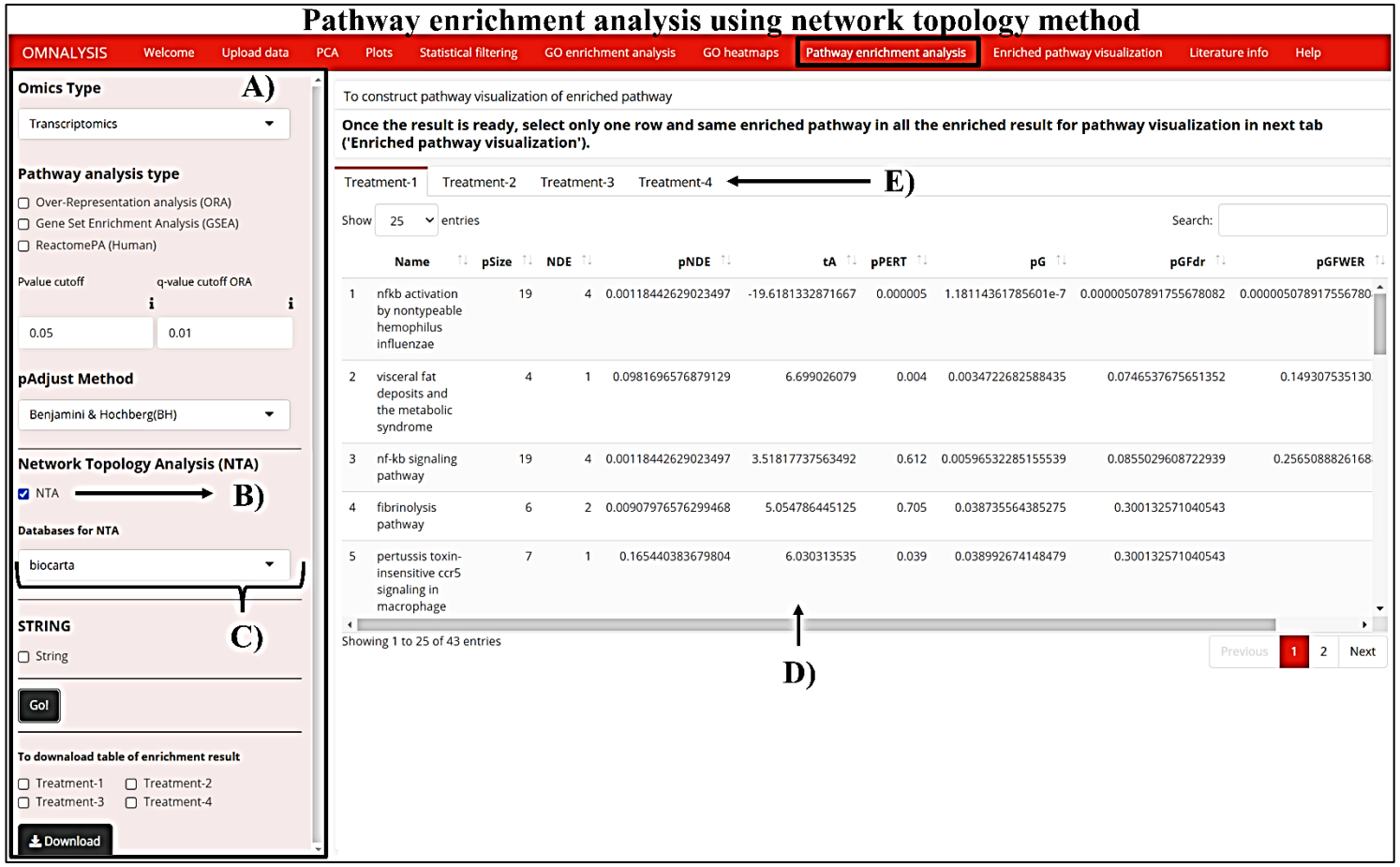
A-presents the interactive panel that was used to provide the input to perform pathway analysis using the Reactome database. B-ReactomePA checkbox selected among others. C-presents the enriched result of treatment 1 using ReactomePA. D-blue color shows the selected first hit of analysis to visualize it on the Reactome pathway. E- presents the output panel sectioned to visualize each treatment.

Table 12. Comparison of top three enriched pathways using Reactome database

Top three Reactome pathways enriched in transcriptomics treatment 1		
Description	geneID	Count
Interleukin-10 signaling	2920/3383/1440/6347/3552/6364/3553/3569/3557/2919/3627/3576/1435/7124/6352	15
Signaling by Interleukins	5898/9235/4791/2920/3383/4792/8767/3718/1440/6347/4790/51561/6648/3552/6364/1846/3553/3656/3569/3557/10068/7412/11009/2919/3575/3965/3667/3627/3576/3726/6288/1435/4128/7124/4067/6352	36
Chemokine receptors bind chemokines	6376/2920/6347/6364/6372/2921/6374/2919/3627/6373/3576/6352	12
Top three Reactome pathways enriched in transcriptomics treatment 2		
Interferon Signaling	960/3430/4938/3383/3717/23586/4940/4939/6772/2635/2633/3664/3437/3433/4502/3659/6737/10346/91543/8638/51191/5371/24138/115362/9246/4599/3455/2634/115361/7412/3660/7324/6773/3669/4600/11274/3434/9636/5696/3134/3105/3106	42
Interferon alpha/beta signaling	3430/4938/4940/4939/6772/3664/3437/3433/3659/91543/8638/24138/4599/3455/2634/3660/6773/3669/4600/11274/3434/9636/5696/3134/3105/3106	26
Signaling by Interleukins	5898/9235/5685/6196/4791/2920/3383/3717/3091/4792/8767/3718/1440/6347/4790/51561/6648/3977/604/3552/6364/6772/6382/1846/3553/3656/3601/3569/3557/10068/2247/1848/8878/7412/11009/2919/4088/3575/3965/3627/3576/6773/3726/4615/6288/1435/4128/57162/5696/1052/7124/5698/4067/6352	54
Top three Reactome pathways enriched in transcriptomics treatment 3		
Signaling by Interleukins	2920/3383/4792/1440/6347/6648/604/3552/6364/9173/8809/3656/3569/7412/11009/2919/3576/6288/1435/9021	20
Interleukin-10 signaling	2920/3383/1440/6347/3552/6364/3569/2919/3576/1435	10
Chemokine receptors bind chemokines	6376/2920/6387/6347/6364/6372/2921/6374/2919/3576	10
Top three Reactome pathways enriched in transcriptomics treatment 4		
Interleukin-10 signaling	7133/7132/2920/3383/1440/6347/3552/6364/3553/3569/3557/2919/3627/3576/1435/7124/6352	17
Signaling by Interleukins	7133/9846/5291/7132/25930/4791/2920/4208/3383/4792/8767/4793/1440/6347/4790/51561/6648/604/3552/6364/8809/1026/3553/3911/3656/3569/3557/5519/1848/5451/2308/3570/7412/2919/3575/3627/3576/2353/3726/5771/3725/29110/1435/9021/57162/5696/7124/6352	48
Interleukin-4 and Interleukin-13 signaling	7133/3383/6347/51561/604/3552/1026/3553/3911/3569/5451/2308/3570/7412/3576/2353/3726/9021/7124	19

To perform in-depth analysis, OMnalysis interactive panel (Figure 27, A) contain NTA (Figure 27, B) used four databases namely, biocarta (sets of protein participating in pathways), panther (Protein ANalysis THrough Evolutionary Relationships), NCI (Nature pathway Interaction Database) and pharmgkb (pharmacogenomics knowledge resources) to perform network topology analysis on transcriptomics or proteomics data (Figure 27, C). Figure 27, D shows the network topology analysis (NTA) using the biocarta database. In treatment 1 we identified that nfkb activation by nontypeable hemophilus influenzae pathway was one of the most significant pathways identified at FDR correction 0.05 cutoff, however, the perturbation of the pathway is inhibited in all the treatments (Table 13).

Figure 27. Web-interface to perform network topology analysis (NTA)



A-interactive panel is populated with enrichment analysis methods. B-presents the checkbox required to perform NTA. C-provides the list of databases that supports NTA (biocarta, panther, NCI and pharmGkb). D-presents the terms obtained in treatment 1 against biocarta. E-presents the subsections to display the result of treatments 1, 2, 3 and 4.

Table 13. Part of Signaling activities identified using NTA method and biocarta database

Top three signaling activates activated or inhibited in transcriptomics treatment 1									
Col.1	Col.2	Col.3	Col.4	Col.5	Col.6	Col.7	Col.8	Col.9	Col.10
Name	pSize	NDE	pNDE	tA	pPERT	pG	pGFdr	pGFWER	Status
nfkb activation by nontypeable hemophilus influenzae	19	5	0.01115	-18.8263	5.00E-06	9.87E-07	9.47E-05	9.47E-05	Inhibited
mets affect on macrophage differentiation	13	6	0.000187	5.41478	0.252	0.000518	0.024856	0.049712	Activated
pertussis toxin-insensitive ccr5 signaling in macrophage	7	3	0.011636	6.880032	0.049	0.004829	0.101381	0.463578	Activated
Top three signaling activates activated or inhibited in transcriptomics treatment 2									
nfkb activation by nontypeable hemophilus influenzae	19	6	0.000317	-23.1764	0.001	5.05E-06	0.000445	0.000445	Inhibited
nf-kb signaling pathway	19	6	0.000317	7.780221	0.401	0.001266	0.055705	0.11141	Activated
visceral fat deposits and the metabolic syndrome	4	1	0.19504	8.165339	0.008	0.011644	0.280848	1	Activated
Top three signaling activates activated or inhibited in transcriptomics treatment 3									
hiv-1 nef: negative effector of fas and tnf	25	4	0.003664	-7.81771	0.22	0.006548	0.109595	0.196438	Inhibited
nfkb activation by nontypeable hemophilus influenzae	19	3	0.012345	-6.63445	0.074	0.007306	0.109595	0.219191	Inhibited
pertussis toxin-insensitive ccr5 signaling in macrophage	7	2	0.012951	3.246024	0.165	0.015276	0.152759	0.458276	Activated
Top three signaling activates activated or inhibited in transcriptomics treatment 4									
nfkb activation by nontypeable hemophilus influenzae	19	4	0.056201	-20.8951	5.00E-06	4.52E-06	0.000339	0.000339	Inhibited
visceral fat deposits and the metabolic syndrome	4	2	0.033028	10.30185	0.004	0.001312	0.049204	0.098408	Activated
mets affect on macrophage differentiation	13	5	0.002198	8.770737	0.116	0.002365	0.058281	0.177383	Activated

The top three signaling activates activated or inhibited using the set of DEGs. The first and the last column show the inhibition and activation of molecular activities in treatments.

Further, we performed NTA analysis using other databases and the significant part of the result is shown in Table 13, 14, 15, and 16 using biocarta, panther, NCI and pharmGkb databases, respectively. In Table 13, 14, 15 and 16, Col.1 and Col.2 shows the number of genes on the pathway and the number of significantly differentially expressed genes (DEGs) per pathway, respectively. Col.4 (pNDE) indicates the probability to observe at least significantly DEGs on the pathway using the hypergeometric analysis model. Moreover, Col.5 to Col.7 explain the observed total perturbation in the pathway (tA), probability of accumulation in the pathway by chance (nPERT) and global Pvalue (pG) generated by combining pNDE and pPERT. Columns eighth and ninth are false discovery rate and Bonferroni corrected global pG. Col.8 and Col.9 provides the FDR (pGFdr) and Bonferroni adjusted global Pvalue (pGFWER), respectively. Col.10 indicates the perturbation direction of the pathway, either activated or inhibited. By comparing Table 13 and 14 we found that the NDE value is less when the panther database was used, which may result in the inappropriate significant pathways in the treatments. Whereas, NDE value is higher when the NCI database was used, thus, IL-23 mediated signaling pathway was found to be most significant and activated in all treatments. We also identified that 11 significantly DEGs from the input list of treatment 1 found on the Cyclosporine Pathway. As a result, the tA value in Cyclosporine Pathway was increased and activated. The result of treatment (1, 2, 3 and 4) using NTA and databases (biocarta, panther, NCI and pharmGkb) provided in Supplementary information 14 (**http://gofile.me/6DOhe/GiQJv9VxP**).

Table 14. The first four signaling pathways were identified using the panther database

Top three signaling activates activated or inhibited in transcriptomics treatment 1									
Col.1	Col.2	Col.3	Col.4	Col.5	Col.6	Col.7	Col.8	Col.9	Col.10
Name	pSize	NDE	pNDE	tA	pPERT	pG	pGFdr	pGFWER	Status
Plasminogen activating cascade	4	3	0.001578	0	NA	0.001578	0.022096	0.022096	Inhibited
Gastrin_CCK2R_240212	44	7	0.043346	6.357428	0.13	0.034817	0.243721	0.487442	Activated
Notch signaling pathway	9	3	0.024941	1.267856	0.611	0.078998	0.368655	1	Activated
Top three signaling activates activated or inhibited in transcriptomics treatment 2									
Plasminogen activating cascade	4	2	0.015544	0	NA	0.015544	0.170983	0.170983	Inhibited
Gastrin_CCK2R_240212	44	3	0.412406	8.278777	0.05	0.100658	0.553617	1	Activated
Notch signaling pathway	9	1	0.38632	1.375271	0.321	0.382865	0.999406	1	Activated
Top three signaling activates activated or inhibited in transcriptomics treatment 3									
Transcription regulation by bZIP transcription factor	33	1	0.580975	-3.01564	0.055	0.141985	0.320414	0.85191	Inhibited
Gastrin_CCK2R_240212	44	1	0.68663	4.521147	0.054	0.15924	0.320414	0.955441	Activated
Toll receptor signaling pathway	12	2	0.037372	0	1	0.160207	0.320414	0.961243	Inhibited
Top three signaling activates activated or inhibited in transcriptomics treatment 4									
B cell activation	22	4	0.088274	-1.99277	0.138	0.065877	0.665592	0.724644	Inhibited
Gastrin_CCK2R_240212	44	5	0.259543	7.494566	0.118	0.137385	0.665592	1	Activated
Enkephalin release	17	1	0.750213	1.983221	0.078	0.224613	0.665592	1	Activated

Part of the result from network topology analysis shows the activation and inhibition of the identified pathways by DEGs.

Table 15. The result of NTA using nature pathway interaction databases (NCI)

Top three interaction activates activated or inhibited in transcriptomics treatment 1									
Col.1	Col.2	Col.3	Col.4	Col.5	Col.6	Col.7	Col.8	Col.9	Col.10
Name	pSize	NDE	pNDE	tA	pPERT	pG	pGFdr	pGFWER	Status
IL23-mediated signaling events	23	10	2.43E-06	2815.338	5.00E-06	3.18E-10	3.66E-08	3.66E-08	Activated
Calcineurin-regulated NFAT-dependent transcription in lymphocytes	25	10	6.05E-06	20.28925	5.00E-06	7.63E-10	4.39E-08	8.78E-08	Activated
Validated transcriptional targets of AP1 family members Fra1 and Fra2	29	10	2.82E-05	39.00016	5.00E-06	3.33E-09	1.28E-07	3.83E-07	Activated
Top three interaction activates activated or inhibited in transcriptomics treatment 2									
IL23-mediated signaling events	23	11	6.10E-09	2533.833	5.00E-06	9.80E-13	9.80E-11	9.80E-11	Activated
Validated transcriptional targets of AP1 family members Fra1 and Fra2	29	9	1.16E-05	28.25812	0.001	2.23E-07	1.11E-05	2.23E-05	Activated
AP-1 transcription factor network	52	14	3.06E-07	7.725886	0.192	1.04E-06	3.45E-05	0.000104	Activated
Top three interaction activates activated or inhibited in transcriptomics treatment 3									
IL23-mediated signaling events	23	7	1.27E-06	345.5437	0.238	4.84E-06	0.000276	0.000276	Activated
CD40/CD40L signaling	29	6	8.36E-05	-312.001	0.515	0.000476	0.006679	0.027114	Inhibited
Endogenous TLR signaling	17	5	5.46E-05	-6.07374	0.826	0.000497	0.006679	0.028307	Inhibited
Top three interaction activates activated or inhibited in transcriptomics treatment 4									
IL23-mediated signaling events	23	10	3.66E-06	2505.005	5.00E-06	4.71E-10	5.28E-08	5.28E-08	Activated
Validated transcriptional targets of AP1 family members Fra1 and Fra2	29	13	7.90E-08	18.57987	0.05	8.04E-08	4.50E-06	9.00E-06	Activated
ATF-2 transcription factor network	43	14	2.67E-06	96.34534	0.005	2.55E-07	8.64E-06	2.86E-05	Activated

Top three molecular interactions activated or inhibited in treatments.

Table 16. The result of NTA using pharmgkb database

Col.1	Col.2	Col.3	Col.4	Col.5	Col.6	Col.7	Col.8	Col.9	Col.10
Name	pSize	NDE	pNDE	tA	pPERT	pG	pGFdr	pGFWER	Status
Tacrolimus/Cyclosporine Pathway, Pharmacodynamics	31	11	8.13E-06	47.78496	0.003	4.52E-07	5.42E-06	5.42E-06	Activated
Peginterferon alpha-2a/Peginterferon alpha-2b Pathway (Hepatocyte), Pharmacodynamics	15	6	0.00048	-3.66935	0.424	0.001934	0.011602	0.023203	Inhibited
EGFR Inhibitor Pathway, Pharmacodynamics	44	8	0.015131	18.95479	0.112	0.012507	0.050028	0.150083	Activated
Vemurafenib Pathway, Pharmacodynamics	22	5	0.020963	14.45998	0.373	0.045751	0.137252	0.549008	Activated

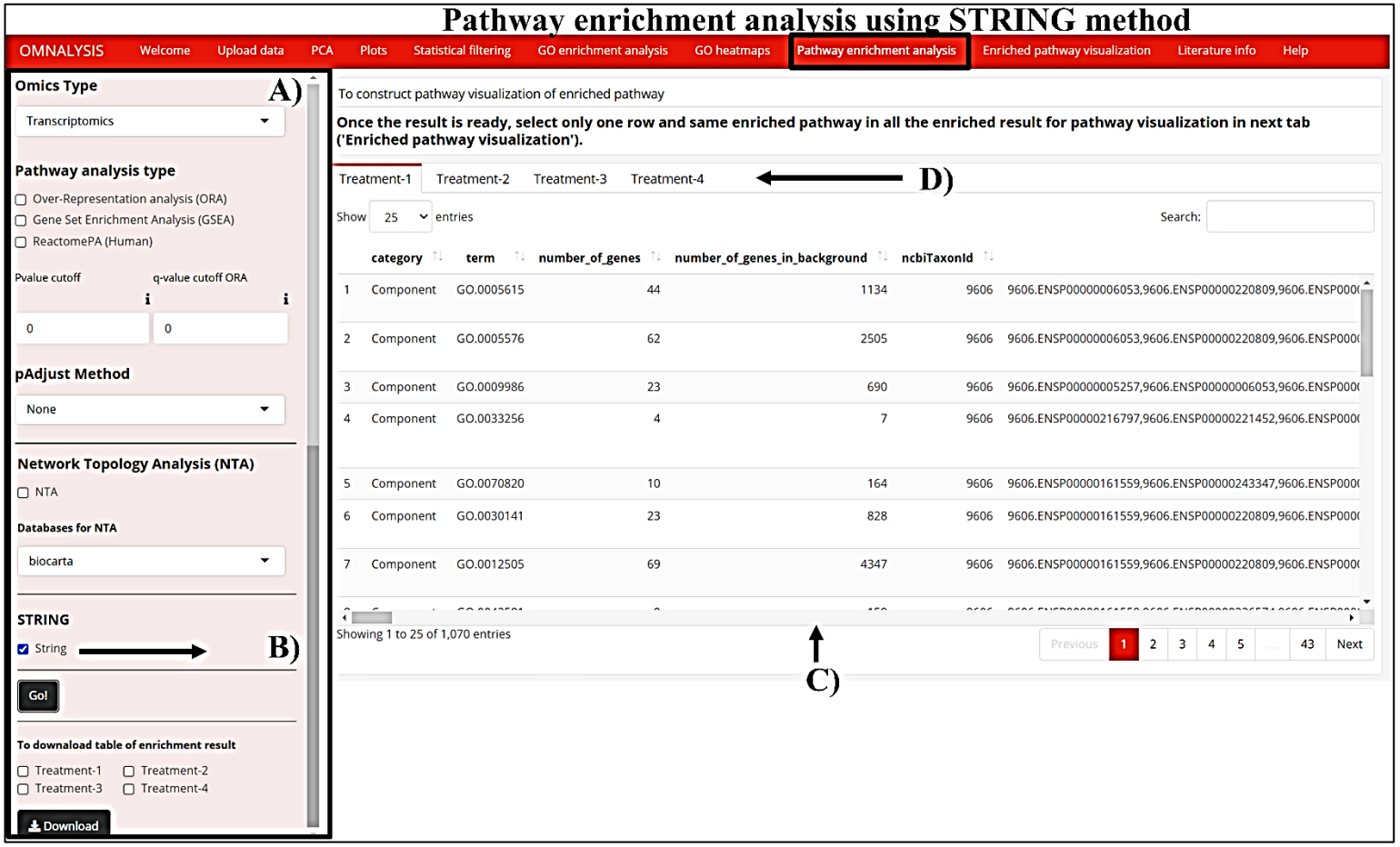
DEGs of treatment 1 mapped to the pathway related to pharmacodynamics using the pharmGkb database. In the first hit, 11 DEGs were overlapped.

OMnalysis is incorporated with the STRING to extend the functionality. In the interactive panel (Figure 28, A), we selected Omics type as transcriptomics and STRING checkbox (Figure 28, B) for treatment 1 to execute the analysis using the action button Go. At Pvalue cutoff 0.05, in total 1,070 terms (Figure 28, C) were enriched in treatment 1,969 terms in treatment 2,668 terms in treatment 3,599 terms in treatment 4 against STRINGdb (GO classes, KEGG, Pfam, SMART, Reactome, PubMed, InterPro, and UniProt Keywords). Among them, the top hit against each database is presented in Table 17.

In transcriptomics, we observed that process related to immune, function related to receptor signaling, pathways related to interleukin and TNF signaling, domain and motif related to interleukin and chemokines were highly enriched in treatments 1, 2, 3 and 4, providing evidence of a relationship between the cellular immune response against pathogens. The complete result of STRING analysis of transcriptomics treatment 1, 2, 3 and 4 (Figure 28, D) is provided in Supplementary information 16 (**http://gofile.me/6DOhe/9tbDuVumN**).

In proteomics, at Pvalue 0.05, we obtained a total of 286 terms enriched in treatment 1,405 in treatment 2,901 in treatment 3,1088 in treatment 4 against STRINGdb (GO classes, KEGG, Pfam, SMART, Reactome, PubMed, InterPro, and UniProt Keywords). The complete result of STRING analysis of proteomics treatments is provided in Supplementary information 16 (**http://gofile.me/6DOhe/9tbDuVumN**).

Figure 28. Web interface to perform STRING analysis



A-presents the interactive panel that was used to provide the input to perform protein-protein interaction using a string database. B- presents STRING checkbox was selected among others. C-presents the result of treatment 1. D-the output panel is sectioned into four (pathway result-1, 2, 3 and 4) display panels.

Table 17. Result table using STRING

Top enriched terms obtained against the various category of databases using transcriptomics treatment 1						
category	term	InputGenes	preferredNames	p_value	fdr	description
Component	GO.0005615	9606.ENSP00000006053,9606.ENSP00000220809,9606.ENSP00000225474,9606.ENSP00000225831,9606.ENSP00000226317,9606.ENSP00000228534,9606.ENSP00000242208,9606.ENSP00000243347,9606.ENSP00000245907,9606.ENSP00000252809,9606.ENSP00000259206,9606.ENSP00000260356,9606.ENSP00000261292	CX3CL1, PLAT,CSF3,CCL2,CXCL6,IL23A,INHBA,TNFAIP6,C3,GDF15,IL1RN,THBS1,LIPG	2.34E-14	8.17E-12	extracellular space
Function	GO.0005515	9606.ENSP00000005257,9606.ENSP00000006053,9606.ENSP00000161559,9606.ENSP00000216797,9606.ENSP00000220751,9606.ENSP00000220809,9606.ENSP00000221452,9606.ENSP00000222553,9606.ENSP00000225474,9606.ENSP00000225831,9606.ENSP00000226317,9606.ENSP00000226574,9606.ENSP00000228280,9606.ENSP00000228534	RALA,CX3CL1,CEACAM1,NFKBIA,RIPK2,PLAT,RELB,NAMPT,CSF3,CCL2,CXCL6,NFKB1	8.41E-17	5.08E-14	signaling receptor binding
InterPro	IPR037684	9606.ENSP00000359488,9606.ENSP00000359490,9606.ENSP00000359497,9606.ENSP00000359504,9606.ENSP00000359512	GBP5,GBP4,GBP2,GBP1,GBP3	7.36E-08	2.52E-05	Guanylate-binding protein, C-terminal
KEGG	hsa04668	9606.ENSP00000006053,9606.ENSP00000216797,9606.ENSP00000225831,9606.ENSP00000226574,9606.ENSP00000254958,9606.ENSP00000263341,9606.ENSP00000263464,9606.ENSP00000264832,9606.ENSP00000294728,9606.ENSP00000296026,9606.ENSP00000296027,9606.ENSP00000331736,9606.ENSP00000351671,9606.ENSP00000362994,9606.ENSP00000379110,9606.ENSP00000385675,9606.	CX3CL1,NFKBIA,CCL2,NFKB1,JAG1,IL1B,BIRC3,ICAM1,VCAM1,CXCL3,CXCL5,SELE,CCL20,TRAF1,CXCL1,IL6	5.62E-16	1.16E-13	TNF signaling pathway
Keyword	KW-0964	9606.ENSP00000006053,9606.ENSP00000161559,9606.ENSP00000220809,9606.ENSP00000222543,9606.ENSP00000222553,9606.ENSP00000225474,9606.ENSP00000225831,9606.ENSP00000226317,9606.ENSP00000228280,9606.ENSP00000228534,9606.ENSP00000241261,9606.ENSP00000242208,9606.ENSP00000245907,9606.	CX3CL1,CEACAM1,PLAT,TFPI2,NAMPT,CSF3,CCL2,CXCL6,KITLG,IL23A,TNFSF10,INHBA	1.02E-14	2.66E-12	Secreted
NetworkNeighborAL	CL.3630	9606.ENSP00000216797,9606.ENSP00000220751,9606.ENSP00000221452,9606.ENSP00000226574,9606.ENSP00000256458,9606.ENSP00000259206,9606.ENSP00000260010,9606.ENSP00000263339,9606.ENSP00000263341,9606.ENSP00000263464,9606.ENSP00000263642,9606.ENSP00000275015,9606.ENSP00000299663,9606.ENSP00000317891,9606.ENSP00000358983,9606.ENSP00000362994,9606.ENSP00000369213,9606.ENSP00000481570	NFKBIA,RIPK2,RELB,NFKB1,IRAK2,IL1RN,TLR2,IL1A,IL1B,BIRC3,IFIH1,NFKBIE,CLEC4E,TNIP1,NFKB2,TRAF1,DDX58,TNFAIP3	3.89E-17	2.55E-14	mixed, incl. I-kappaB kinase/NF-kappaB signaling, and interleukin-1-mediated signaling pathway

PMID	PMID.21655103	9606.ENSP00000216797,9606.ENSP00000220751,9606.ENSP00000225831,9606.ENSP00000226317,9606.ENSP00000226574,9606.ENSP00000245414,9606.ENSP00000256495,9606.ENSP00000258534,9606.ENSP00000259874,9606.ENSP00000260010,9606.ENSP00000263339,9606.ENSP00000263341,9606.ENSP00000263464,9606.ENSP00000263642,9606.ENSP00000264832,9606.ENSP00000266671	NFKBIA,RIPK2,CCL2,CXCL6,NFKB1,IRF1,BHLHE40,DRAM1,IER3,TLR2,IL1A,IL1B,BIRC3,IFIH1,ICAM1,PHLDA1	4.59E-44	2.92E-38	(2011) Inflammatory gene regulatory networks in amnion cells following cytokine stimulation: translational systems approach to modeling human parturition.
Pfam	PF00048	9606.ENSP00000006053,9606.ENSP00000225831,9606.ENSP00000226317,9606.ENSP00000296026,9606.ENSP00000296027,9606.ENSP00000351671,9606.ENSP00000379110,9606.ENSP00000427279	CX3CL1,CCL2,CXCL6,CXCL3,CXCL5,CCL20,CXCL1,CXCL2	4.72E-08	1.41E-05	Small cytokines (intecrine/chemokine), interleukin-8 like
Process	GO.0009605	9606.ENSP00000005257,9606.ENSP00000006053,9606.ENSP00000216797,9606.ENSP00000220751,9606.ENSP00000221452,9606.ENSP00000225474,9606.ENSP00000225831,9606.ENSP00000226317,9606.ENSP00000226574,9606.ENSP00000228534,9606.ENSP00000234111,9606.ENSP00000243457,9606.ENSP00000245185,9606.ENSP00000245414,9606.ENSP00000245907,9606.ENSP00000251642,9606.ENSP00000252809,9606.ENSP00000256458	RALA,CX3CL1,NFKBIA,RIPK2,RELB,CSF3,CCL2,CXCL6,NFKB1,IL23A,ODC1,KCNJ2,MT2A,IRF1,C3,DHX58,GDF15,IRAK2,BHLHE40,OASL,RTP4,TLR2,THBS1,LIPG,IL1B,IFIH1,HERC6,ICAM1	2.88E-31	1.19E-27	immune system process
RCTM	HSA-1280215	9606.ENSP00000005257,9606.ENSP00000220751,9606.ENSP00000221452,9606.ENSP00000225474,9606.ENSP00000225831,9606.ENSP00000228534,9606.ENSP00000245185,9606.ENSP00000256458,9606.ENSP00000259206,9606.ENSP00000263339,9606.ENSP00000263341,9606.ENSP00000263464,	RALA,RIPK2,RELB,CSF3,CCL2,IL23A,MT2A,IRAK2,IL1RN,IL1A,IL1B,BIRC3,ICAM1,VCAM1,CCL20,NFKB2,GBP5,IL24,CXCL1,IL18BP,IFI35,CXCL2,SOD2	3.61E-20	1.62E-18	Cytokine Signaling in Immune system
SMART	SM00199	9606.ENSP00000006053,9606.ENSP00000225831,9606.ENSP00000226317,9606.ENSP00000296026,9606.ENSP00000296027,9606.ENSP00000351671,9606.ENSP00000379110,9606.ENSP00000427279	CX3CL1,CCL2,CXCL6,CXCL3,CXCL5,CCL20,CXCL1,CXCL2	3.43E-08	4.59E-06	Intercrine alpha family (small cytokine C-X-C) (chemokine CXC).

5.10 Enriched pathway visualization

This is the ninth section of the OMnalysis tool. In the interactive panel (Figure 29, A), we selected pathway ORA checkbox (Figure 29, B) and colors from the drop-down menu according to the log fold change value (logFC) of significantly DEGs identified in each treatment (Figure 29, C) to visualize them on the pathway. The TNF signaling pathway is shown with the log fold change value (logFC) of significantly DEGs present in all four treatments (Figure 29, D). The enlarged view shows that the CXCL family is highly expressed in all the treatments, whereas the JAG1 gene shows no expression in all treatments. The display panel is subdivided into four sections (Figure 29, E) each section displays the output of the method selected in the interactive panel.

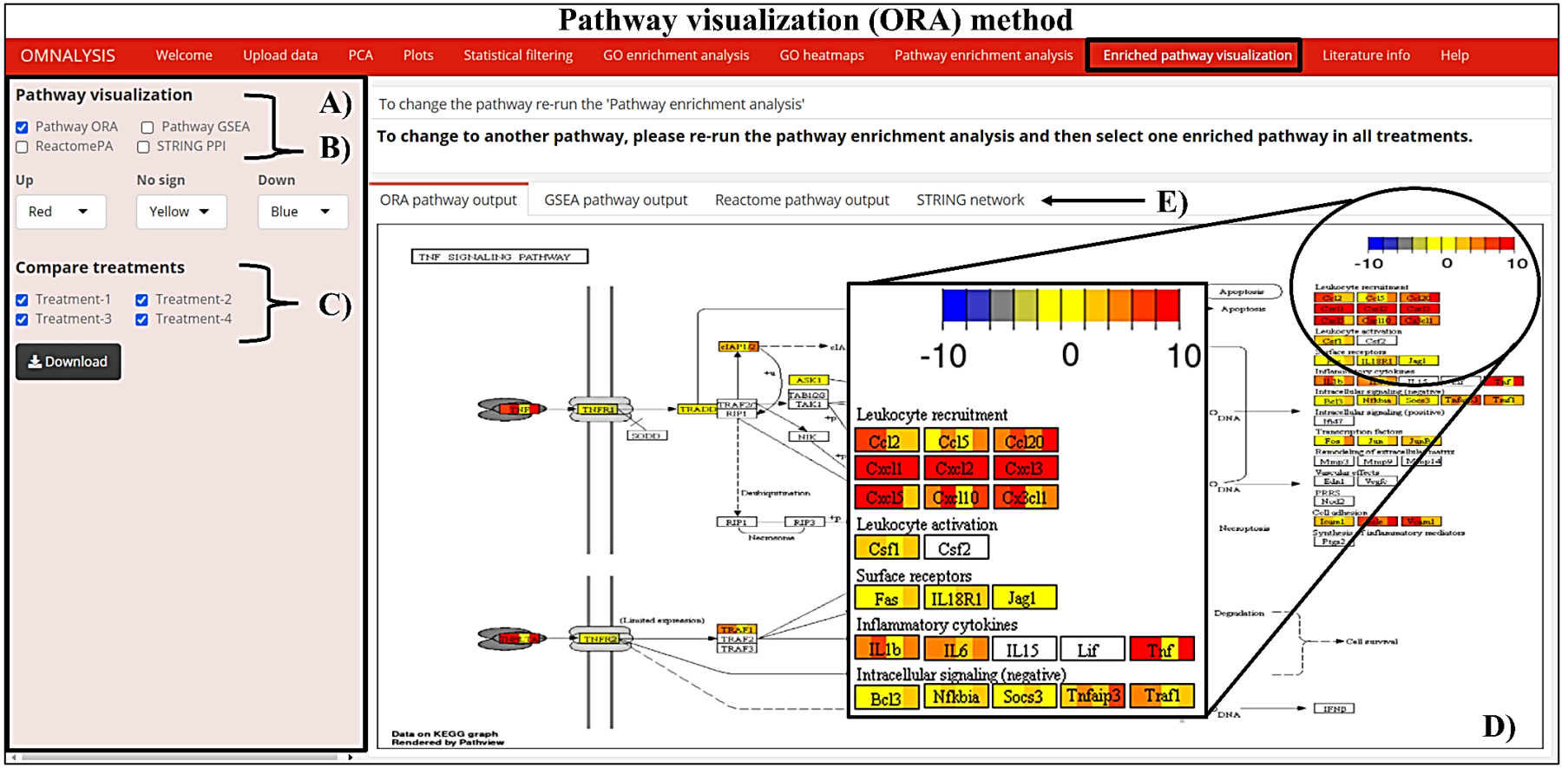
Using an interactive panel (Figure 30, A), we selected pathway GSEA (Figure 30, B) with four treatments (Figure 30, C) to compare the expression (logFC) of DEGs on the cytokine-cytokine receptor-mediated pathway (Figure 30, D). We identified inflammation inducing and antiviral response genes were highly expressed in treatments, however, in treatment 3 some genes were not significantly expressed. Pathways enriched using ORA and GSEA are presented in Supplementary information 15 (**http://gofile.me/6DOhe/gKmvF4xoZ**).

Using an interactive panel (Figure 31, A), we selected the Reactome pathway checkbox (Figure 31, B) with four treatments (Figure 31, C) to compare the expression (logFC) of DEGs on the interleukine-10 signaling pathway (Figure 30, D). We identified that majority of the genes in the interferon alpha-beta signaling pathway are marked as red or yellow indicating high or no expression in treatment (1, 2, 3 and 4). Pathway with the reactomePA method is presented in Supplementary information 16 (**http://gofile.me/6DOhe/Qm9eB8YJl**).

For the STRING protein-protein interaction (PPI) visualization, we used an interactive panel (Figure 32, A). In which the four analysis visualization checkboxes (Figure 32, B) were provided to select the type of output in the output panel. We selected treatment checkbox 1 (Figure 32, C) to identify the expression of DEGs or DEPs on the generated network. From the protein-protein interaction network the result of the transcriptomics treatment 1 (Figure 32, D), we identified CXCL and IL6, IL1A, IL1B were highly up-regulated. The interaction evidence provided in the network is co-expression and text

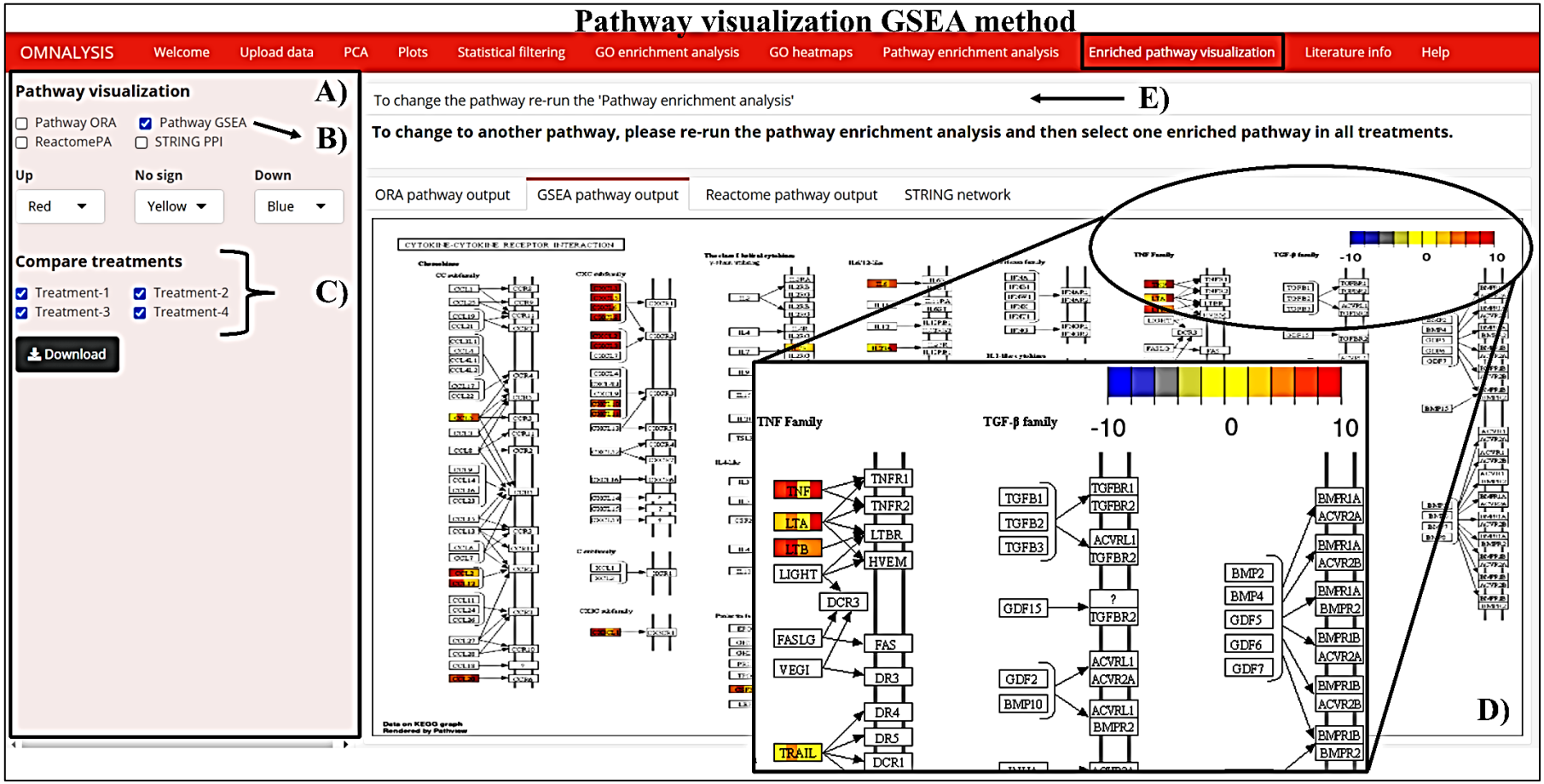
mining. PPI network of transcriptomics and proteomics treatment 1,2,3 and 4 is presented in Supplementary information 16 (**http://gofile.me/6DOhe/Qm9eB8YJl**).

Figure 29. Web-interface for ORA enriched pathway visualization



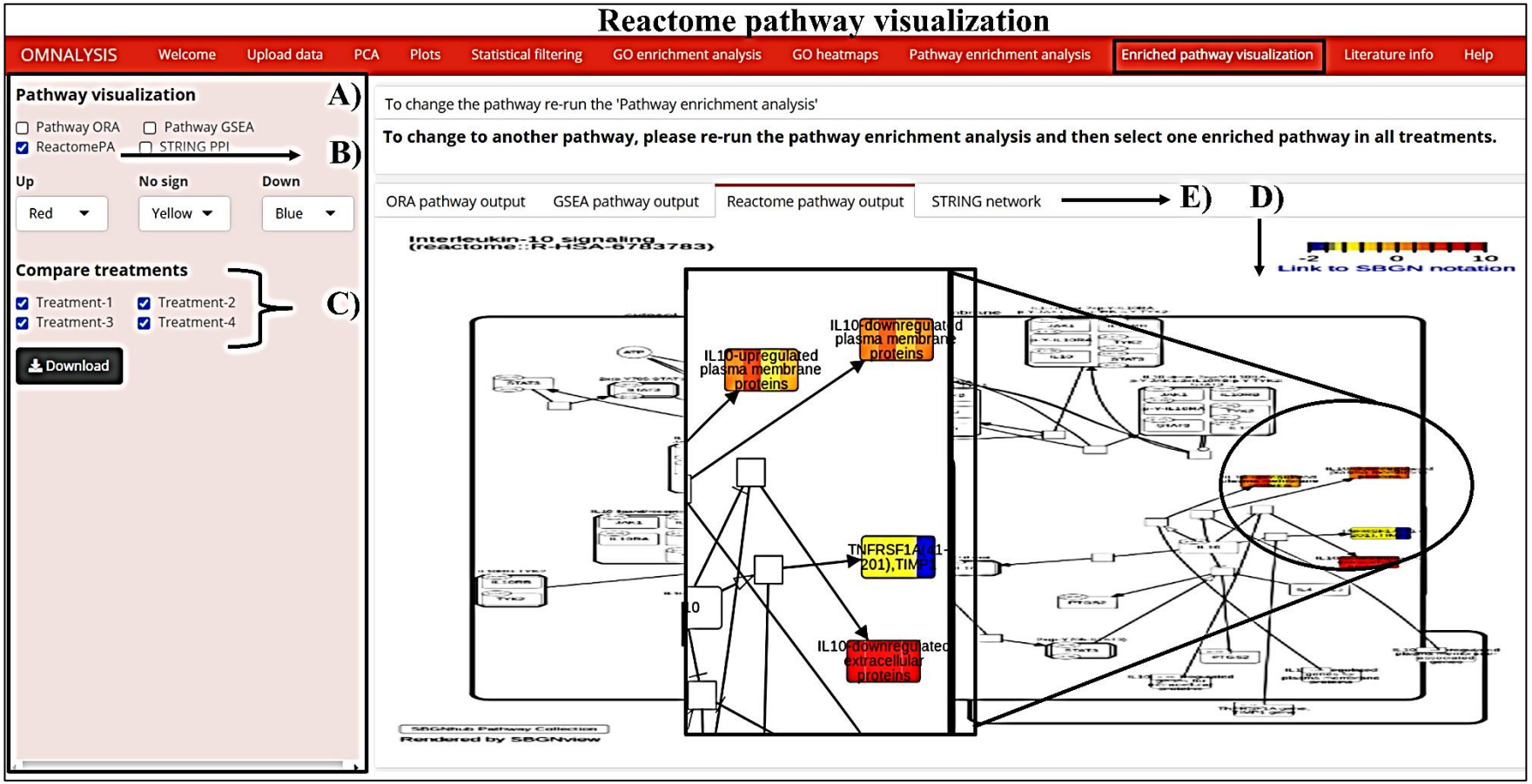
A-interactive panel for users to provide input for pathway visualization. B-presents the checkboxes available for pathway visualization, in which ORA was selected. C-checkboxes are available for comparison of expression profile among the treatments, in which treatment 1, 2, 3 and 4 is checked. D-presents the display panel to visualize the enriched pathway and zoomed view of DEGs red-up regulated, blue-down regulated and yellow-no significant expression in treatment 1, 2, 3 and 4.

Figure 30. Web interface for GSEA enriched pathway visualization



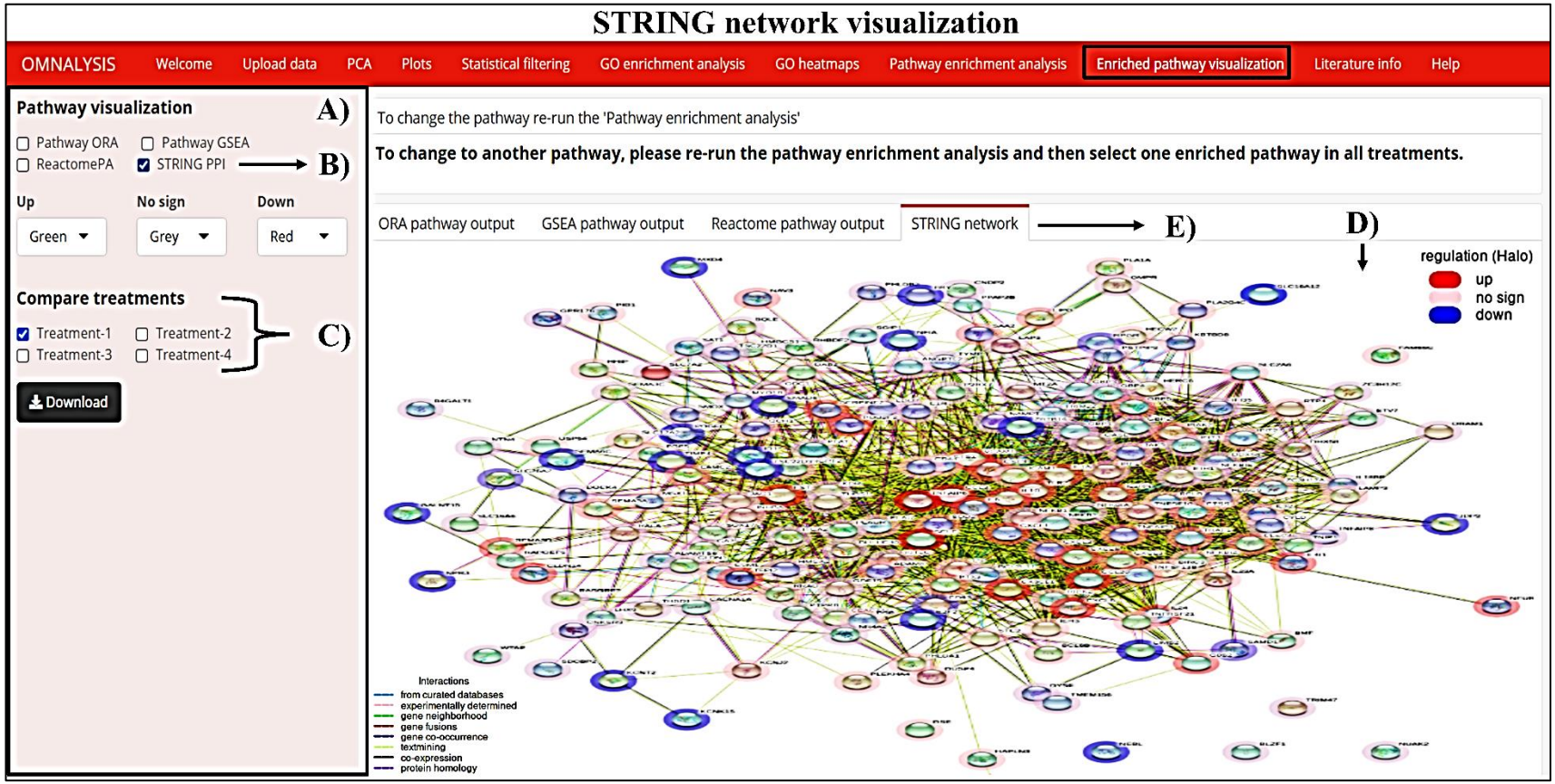
A-in interactive panel for users to provide inputs. B-checkbox Pathway GSEA was selected to visualize the GSEA enriched pathway. C-provides checkboxes for each treatment to compare the expression profile, in which treatment 1, 2, 3 and 4. D-window for pathway visualization. The zoomed view presents red-up regulation, blue-down regulation and yellow-no significant expression.

Figure 31. Web interface for Reactome pathway visualization



A-interactive panel for users to provide inputs for pathway visualization. B-ReactomePA checkbox selected for the visualization of pathway enriched using Reactome database. C-provides treatment checkboxes, in which treatments 1, 2, 3 and 4 are selected. D-panel for visualization of the enriched pathway. Zoomed pathway section showing gene box red-up regulated, blue-down regulated and yellow-no significant expression. E-the output panel is sectioned into 4 display panels for each enrichment method.

Figure 32. Web interface for STRING analysis visualization



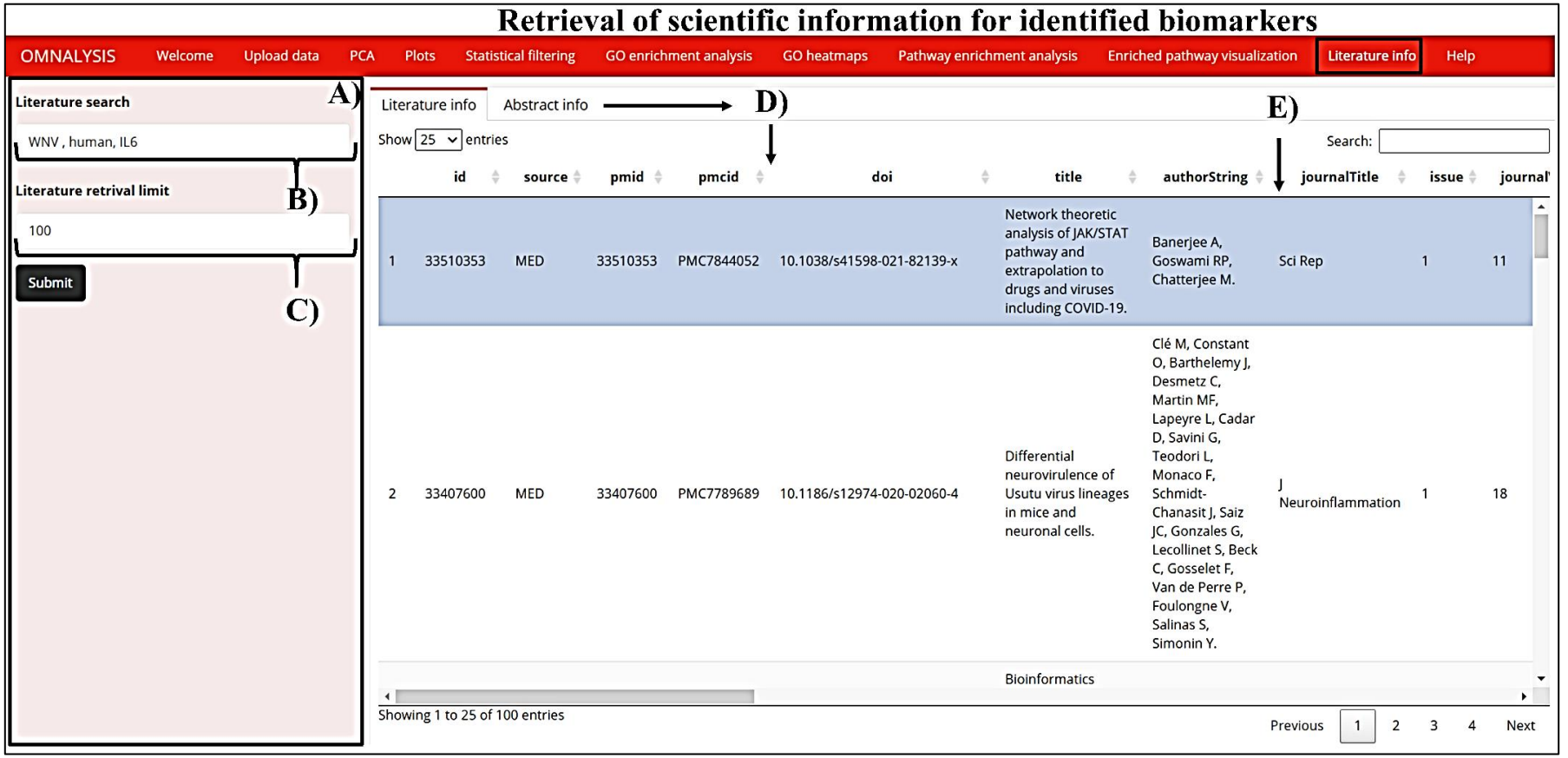
A-interactive panel for users to provide inputs for network visualization. B-STRING PPI checkbox selected for the visualization of the protein-protein interaction network. C-provides treatment checkboxes to display the PPI network of the selected treatment. D-panel display result of interaction in the interactive panel. Halo on the nodes displays expression (logFC), red-up regulated, blue-down regulated and pink-no significant regulation. The interactions legend shows the source of interaction between edges and nodes. E-sectioned output panel to display each analysis result.

5.11 Literature information

This is the tenth section of the OMnalysis tool and is divided into interactive and output panels. We used literature search in an interactive panel (Figure 33, A) and input keywords such as gene, species and pathogens (Figure 33, B) to fetch the classified scientific information from the literature repositories. By setting a limit of retrieval to 100 (Figure 33, C) and checking the submit button, we obtained literature hits (Figure 33, D) and then we selected one row (Figure 33, E) to pull the abstract information. In the interactive section, the maximum limit of retrieval of literature at once is 5000 using the *Europe PMC* R interface (LEVCHENKO et al., 2018).

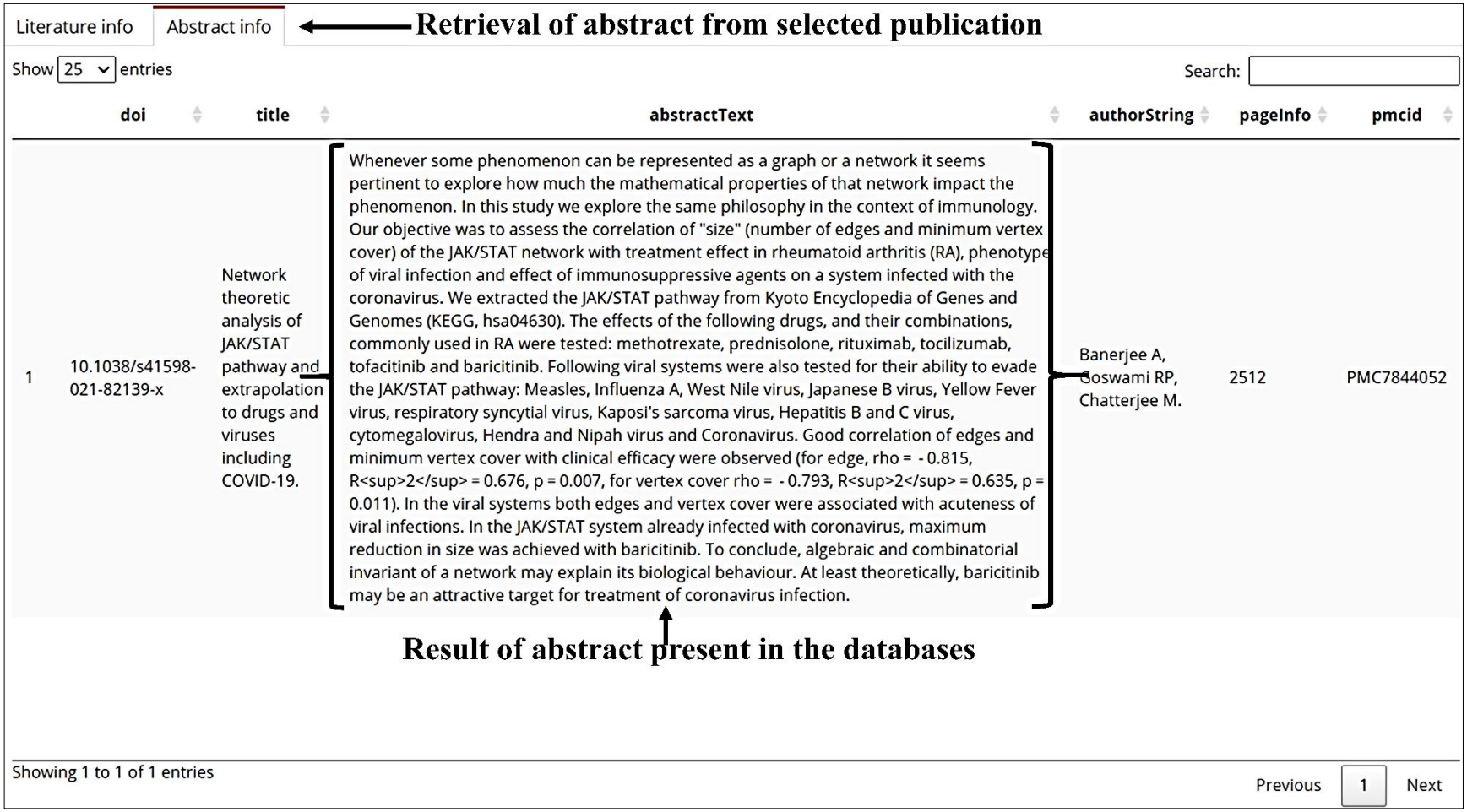
From the search, we obtained details that include source, pmcid, doi, title, authors detail, Journal, issue, publication year, citation count and others. Figure 34 displays the abstract information (doi, title, abstract and pmcid) of the literature search selected in the literature info.

Figure 33. Web interface for fetching scientific literature



A-interactive panel for the user input. B-text input tab to provide keywords gene, species and disease information to retrieve the scientific literature from Europe PMC. C- numeric input tab to limit the retrieval of scientific literature. D-is the output panel for results and further divided into Abstract info. (E) blue colour shows the most related search selected for abstract information.

Figure 34. The display panel of the literature info tab

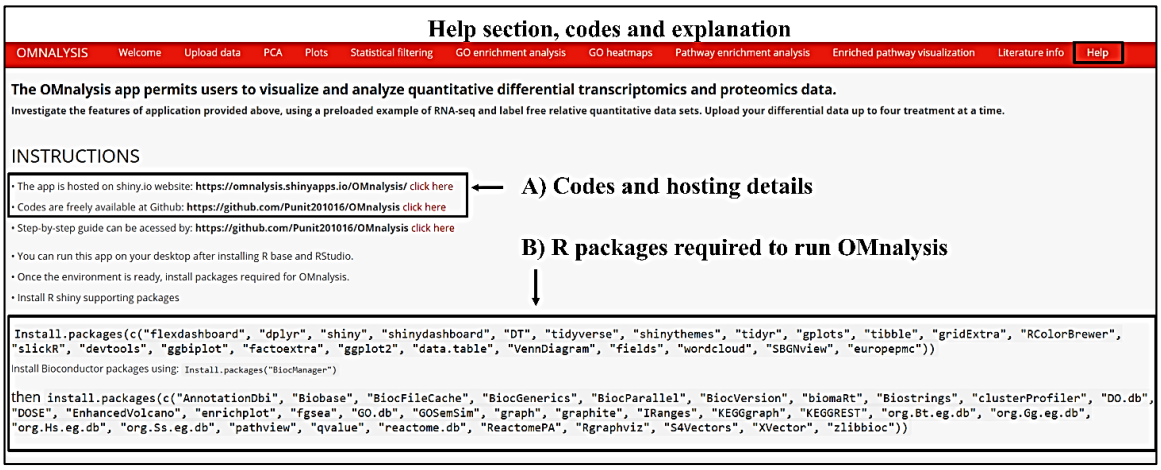


Result of abstract fetched from Europe PMC using europepmc R package. Details of doi, publication title, abstract, authors and pmcid are shown in the tabular form. Curley bracket shows the abstract information.

5.12 Help

This is the eleventh section of the software OMnalysis tool. We provided detailed information indicating the availability of codes by providing the link to *GitHub* repositories (Figure 35, A). The step-by-step guide of each section is provided to access using the link. The R and Bioconductor packages that were used in the OMnalysis development (Figure 35, B).

Figure 35. Web-interface of help tab



(A) provides information about the developed codes and the hosting address of the OMnalysis tool. (B) provides R packages and dependencies used to develop the OMnalysis tool. (C) is the explanation of tabs, checkbox, requirements of the OMnalysis and detailed user manual.

6 Discussion

Taking the advantage of Shiny, Bioconductor, flexdashboard, and markdown we were able to integrate and develop a user-friendly web tool OMnalysis. Using this tool, researchers can walk through various levels of exploration of quantitative data, which includes publication-ready plots, functional interpretation, pathway analysis, and scientific literature. Also, by leveraging the benefits of OMnalysis, the user will be able to analyze four differential expression data simultaneously, derived from quantitative transcriptomics or proteomics experiments. Till to date, few tools (*iDEP*, *DEBrowser*, *IRIS-EDA*, etc.) are developed, which accept count data to analyze differential expression, however, this approach is complicated for the biologist in terms of selecting the normalization methods. The available normalization methods are CPM (count per million), TPM (transcripts per kilobase million), FPKM/RPKM (fragment/reads per kilobase of transcript per million mapped reads), *DESeq2* - the median of rations (LOVE et al., 2014), and *EdgeR* - trimmed mean of M values (ROBINSON et al., 2010). Hence, biologists often use data matrices that contain a list of genes or proteins with expression values and statistical components for downstream differential expression analysis. Although the later option inherits some limitations e.g., inability to perform differential expression analysis using count or spectral data and lack of metadata table, it benefits the larger research community by minimizing the time to obtain the result and to understand the normalization methods. Some tools like *iDEP* (DIJK et al., 2018), *ShinyNGS* (MANNING, 2016) and *DEBrowser* (KUCUKURAL et al., 2019) require additional metadata table to provide the information related to samples and study design. To this background, the OMnalysis is built to provide the researcher with a user-friendly web application, with no metadata dependency, and with streamlined analysis tabs. It covers peer-reviewed, curated, and updated databases, and it enables advanced visualization in form of plots, mapping of the expression data (logFC) on pathways and networks using pseudo colors. A single enriched pathway decorated with the colored map of expression value provides the user with a holistic view of biological activities in different treatments.

We compared the OMnalysis with existing freeware used for DEGs analysis and exploration (e.g., *iDEP*, *IRIS-EDA*, *START* app, *DEBrowser*, etc.), in terms of input data requirements, types of visualization, ease of use, and database used for GO and Pathway enrichment analysis. The details of this comparison are provided in (Table 19). In contrast to OMnalysis, *IRIS-EDA* and *START* app does not support gene ID conversion. The *iDEP*

web application requires a manual update of the database to support gene ID conversion. *DEBrowser* although performs batch effect correction and DEGs analysis using count data, requires an R environment to generate a web interface, which could be a bottleneck in analysis for most biologists as they hold minimum programming knowledge. In contrast, OMnalysis is an online application and doesn't depend on the R environment.

Table 18. Comparison of a bioinformatics platform for downstream analysis.

	iDEP	***IRIS-EDA***	***START App***	***ShinyNGS***	***DEBrowser***	***OMnalysis***
Input data	Count data, meta-data	Count data	Count or expression data	Count or expression data	Count data	Expression data (list of DEGs or Proteins)
PCA	✓	✓	✓	✓	✓	✓
Volcano/Scatter plot	✓	✓	✓	✗	✓	✓
GO analysis	✓	✗	✗	✓	✓	✓
Pathway analysis	✓	✗	✗	✗	✓	✓
Gene ID conversion	✓	✗	✗	✗	✗	✓
Pathway databases	KEGG, STRING API	✗	✗	KEGG	KEGG	KEGG, Reactome, NCI, Panther, biocarta, PharmGKB, STRING
Literature retrieval	✗	✗	✗	✗	✗	✓
Application type	Online	Online	Online	Require R for online	Require R for online	Online
Stand-alone R code	✓	✓	✓	✓	✓	✓
Pathway visualization and STRING network	✓	✗	✗	✗	✓	KEGG, Reactome, PPI network

Note: ✗ and ✓ indicates non-available and available function in the tool, respectively.

When it comes to the representation of data in the form of scatter and volcano plots to demonstrate the level of expression against the level of significance (*P* value) or the number of transcripts (logCPM), the OMnalysis provides an option with customizable resolution and dimensions of the images and enables various image format to download (png, jpeg, tiff, and pdf). The *iDEP* generates scatter and volcano plots only in esp format, whereas the *IRIS-*

EDA and *START* applications produce plots only in png formats. For statistical filtering of non-significant genes, the *IRIS-EDA* uses adjusted *P* value and fold change, whereas in OMnalysis can filter out the non-significant genes based on fold change, *P* values, or adjusted *P* value and log counts per million in case of transcriptomic data.

In the *IRIS-EDA* tool, the enrichment analysis and functional interpretation are extended by providing the weblinks of third-party web servers (*DAVID*, *UCSC Genome Browser*, etc.). Whereas, in OMnalysis the functional interpretation and enrichment analysis is integrated and supported by ORA and its extension GSEA. ORA uses the hypergeometric test (FALCON et al., 2008) and GSEA uses the Kolmogorov-Smirnov statistics (SMIRNOV, 1948) to perform GO enrichment analysis. ORA and GSEA perform multiple hypothesis testing using the gene set against the GO dataset, however, in each run, it may add some false-positive results. To control false positives, OMnalysis supports adjustment of the *P* value using multiple hypothesis correction methods (e.g., Holm, Hochberg, Hommel, Bonferroni, Benjamini and Hochberg, BY and FDR) and a *P* value (Pvalue cutoff) to gain more reliable information. OMnalysis also provides options to segregate the DEGs or DEPs into various GO classes (biological processes, molecular functions, cellular component (ASHBURNER et al., 2000)), which is not available in the *IRIS-EDA* tool. Note that, when interpreting the result from GO and pathway enrichment analysis, a user must be cautious and try a combination of methods and databases to obtain the comprehensive result. Thus, OMnalysis integrated approach provides the user to perform the enrichment analysis and functional prediction in one web application.

Heatmaps are one of the best representation tools to compare the level of expression among various treatments. Users often use third-party *heatmapper* (BABICKI et al., 2016) web application, which requires an input (in manually arranged tabular form) populated with columns such as unique gene or protein IDs, GO IDs, and expression values for each treatment. To this end, we integrated the Heatmap function in the OMnalysis that automatically arrange the table and generates the heatmaps. It also identifies duplicate gene or protein names and filters out those redundancies. Furthermore, in comparison to *IRIS-EDA* and *START* app, OMnalysis provides customizable (key size and font size) and downloadable publication-ready heatmaps.

Finally, using the *R Shiny* platform and *Bioconductor* packages, we were able to integrate several functionalities into OMnalysis. The streamlined functionalities include uploading of expression data, PCA to identify correlation and variability among treatments,

plots to visualize differential expressions, statistical filtering to segregate the candidate according to the statistical significance, GO enrichment analysis, heatmaps to compare expression among treatments, pathway enrichment analysis, and pathway visualization capabilities. All together the OMnalysis provides the user with a comprehensive explanation of the transcriptomics and proteomics data. To our knowledge, no integrated web tool provides visualization of pathways based on *KEGG* and *Reactome*, and visualization of PPI network using *STRING* in one place. OMnalysis with higher flexibility, easy-to-use interface, multiple visualizations, and extensive coverages of curated databases outperforms many of the currently available web applications available to explore and analyze the quantitative transcriptomics and proteomics data.

7 Conclusion

OMnalysis has integrated an array of scattered packages and curated databases to provide a user-friendly data analysis tool. The overall functionality was tested on the four real datasets of transcriptomics and proteomics. Using these datasets, we were able to perform series of downstream analysis, starting from PCA and visualization of differentially expressed candidates in single or multiple treatments in the form of scatter or volcano plots, Venn diagrams and histograms. Further, this tool was able to segregate gene sets based on any of the three gene ontology classes (biological processes, molecular functions, cellular component) with seven possible multiple hypothesis correction methods and two types of enrichment analysis (ORA and GSEA). This tool provided a different view on transcriptomics and proteomics data using three enrichment methods (ORA, GSEA and *ReactomePA*) and network topology analysis using four different databases (PANTHER, biocarta, NCI and PharmGKB). Additionally, STRING gave an overall picture of enrichment and interaction among molecules. Comparing with the other tools, OMnalysis provides more customizable and functional options. We envisage developing an advanced version of OMnalysis, which will include more animal species, omics types, additional pathway networks (e.g., Wiki pathways, Pathbank, etc.), and characterization of functional units of discovered biomarkers (genes, proteins, and metabolites). Currently, we have added an option to download the set of codes, so that bioinformaticians can extend the functionality of the OMnalysis tool. With the existing capabilities, we are confident that OMnalysis will be a useful web application for researchers, with no or less bioinformatics experience, who want to analyze quantitative transcriptomic and proteomic data into actionable biological insight.

8 References

1. AFGAN, E., BAKER, D., VAN DEN BEEK, M., et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res*. 2016, W1, 44, W3-W10.
2. ALEXA, A., RAHNENFUHRER, J. topGO: enrichment analysis for gene ontology. R package version 2.24.0. Retrieved from **http://bioconductor.org/packages/topGO/**. 2016.
3. ALEXEYENKO, A., LEE, W., PERNEMALM, M., et al. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*. 2012, 13, 226.
4. ALLAIRE, J., XIE, Y., MCPHERSON, J., et al. Rmarkdown: Dynamic Documents for R. R package version 2.8. Retrieved from **https://github.com/rstudio/rmarkdown**. 2020.
5. ALONGE, M., SOYK, S., RAMAKRISHNAN, S., et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol*. 2019, 1, 20, 224.
6. AMUNUGAMA, R., JONES, R., FORD, M., et al. Bottom-Up Mass Spectrometry–Based Proteomics as an Investigative Analytical Tool for Discovery and Quantification of Proteins in Biological Samples. *Advances in Wound Care*. 2013, 9, 2, 549-557.
7. ANDERS, S., HUBER, W. Differential expression analysis for sequence count data. *Genome Biology*. 2010, 10, 11.
8. ANDERS, S., PYL, P.T., HUBER, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015, 2, 31, 166-169.
9. ANDREWS, S. FastQC: a quality control tool for high throughput sequence data. *Available at **http://www.bioinformatics.babraham.ac.uk/projects/fastqc/***. 2010.
10. ANTHONY, K., SKINNER, M.A., BUCHOFF, J.R., et al. The NCI-Nature Pathway Interaction Database: A comprehensive resource for cell signaling information. *Cancer Research*. 2011, 71.
11. ARIYARATNE, P.N., SUNG, W.-K. PE-Assembler: de novo assembler using short paired-end reads. *Bioinformatics*. 2010, 2, 27, 167-174.
12. ASHBURNER, M., BALL, C.A., BLAKE, J.A., et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000, 1, 25, 25-29.

13. BABICKI, S., ARNDT, D., MARCU, A., et al. Heatmapper: web-enabled heat mapping for all. *Nucleic Acids Res*. 2016, W1, 44, W147-153.
14. BAIROCH, A., APWEILER, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*. 2000, 1, 28, 45-48.
15. BATEMAN, A., COIN, L., DURBIN, R., et al. The Pfam protein families database. *Nucleic Acids Research*. 2004, 32, D138-D141.
16. BEAVIS, R.C. Using the global proteome machine for protein identification. *Methods Mol Biol*. 2006, 328, 217-228.
17. BENGTSSON, H., BENGTSSON, M.H., LAZYLOAD, T. 'R. utils'- Varios Programming Utilities. R package version version 2.10.1. Retrieved from **https://CRAN.R-project.org/package=R.utils**. 2020.
18. BLIGHE, K., RANA, S., LEWIS, M. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version 1. 8. 0. Retrieved from **https://github.com/kevinblighe/EnhancedVolcano**. 2020.
19. BOLGER, A.M., LOHSE, M., USADEL, B. Trimmomatic: A flexible read trimming tool for Illumina NGS data. *Bioinformatics*. 2014, 15, 30, 2114-2120.
20. CARLSON, M. org. Hs. eg. db: Genome Wide Annotation for Human. R package version 3. 2. 3. Retrieved from **https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html**. 2019.
21. CHAN, C.-H., CHAN, G.C., LEEPER, T.J., et al. rio: A Swiss-army knife for data file I/O. R package version 0. 5. Retrieved from **https://CRAN.R-project.org/package=rio**. 2018.
22. CHANDRAMOULI, K., QIAN, P.-Y. Proteomics: Challenges, Techniques and Possibilities to Overcome Biological Sample Complexity. *Human Genomics and Proteomics*. 2009, 2009, 1-22.
23. CHANG, W., CHENG, J., ALLAIRE, J., et al. Shiny: web application framework for R. R package version 1. 6. 0. Retrieved from **https://CRAN.R-project.org/package=shiny**. 2017.
24. CHANG, W., PARK, T., DZIEDZIC, L., et al. shinythemes: Themes for Shiny. R package version 1. 2. 0. Retrieved from **https://CRAN.R-project.org/package=shinythemes**. 2018.

25. CHANG, W., RIBEIRO, BARBARA, B. shinydashboard: Create Dashboards with 'Shiny'. R package version 0. 7. 1. Retrieved from **https://CRAN.R-project.org/package=shinydashboard**. 2019.
26. CHAO, H.P., CHEN, Y., TAKATA, Y., et al. Systematic evaluation of RNA-Seq preparation protocol performance. *BMC Genomics*. 2019, 1, 20, 571.
27. CHEN, H., BOUTROS, P.C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*. 2011, 12, 35.
28. CINGOLANI, P., PLATTS, A., WANG, L.L., et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012, 2, 6, 80-92.
29. COLAERT, N., GEVAERT, K., MARTENS, L. RIBAR and xRIBAR: Methods for Reproducible Relative MS/MS-based Label-Free Protein Quantification. *Journal of Proteome Research*. 2011, 7, 10, 3183-3189.
30. CONESA, A., GOTZ, S., GARCIA-GOMEZ, J.M., et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005, 18, 21, 3674-3676.
31. CONESA, A., MADRIGAL, P., TARAZONA, S., et al. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 2016, 17, 13.
32. CORLEY, S.M., MACKENZIE, K.L., BEVERDAM, A., et al. Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols. *BMC Genomics*. 2017, 1, 18, 399.
33. COSTA-SILVA, J., DOMINGUES, D., LOPES, F.M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLOS ONE*. 2017, 12.
34. COTTRELL, J.S. Protein identification using MS/MS data. *Journal of Proteomics*. 2011, 10, 74, 1842-1851.
35. CROFT, D., O'KELLY, G., WU, G., et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*. 2011, Database issue, 39, D691-697.
36. DARD-DASCOT, C., NAQUIN, D., D'AUBENTON-CARAFA, Y., et al. Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *BMC Genomics*. 2018, 1, 19, 118.

37. DEL FABBRO, C., SCALABRIN, S., MORGANTE, M., et al. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLOS ONE*. 2013, 12, 8, e85024.
38. DENNIS, G., SHERMAN, B.T., HOSACK, D.A., et al. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology*. 2003, 9, 4, 1-11.
39. DEPRISTO, M.A., BANKS, E., POPLIN, R., et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011, 5, 43, 491-498.
40. DEUTSCH, E.W. The PeptideAtlas Project. *Methods Mol Biol*. 2010, 604, 285-296.
41. DICKER, L., LIN, X., IVANOV, A.R. Increased Power for the Analysis of Label-free LC-MS/MS Proteomics Data by Combining Spectral Counts and Peptide Peak Attributes. *Molecular & Cellular Proteomics*. 2010, 12, 9, 2704-2718.
42. DIJK, E.L.V., JASZCZYSZYN, Y., NAQUIN, D., et al. The Third Revolution in Sequencing Technology. *Trends in Genetics*. 2018, 9, 34, 666-681.
43. DILLIOTT, A.A., FARHAN, S.M.K., GHANI, M., et al. Targeted next-generation sequencing and bioinformatics pipeline to evaluate genetic determinants of constitutional disease. *Journal of Visualized Experiments*. 2018, 134, 2018, 1-10.
44. DIZ, A.P., CARVAJAL-RODRIGUEZ, A., SKIBINSKI, D.O. Multiple hypothesis testing in proteomics: a strategy for experimental work. *Mol Cell Proteomics*. 2011, 3, 10, M110 004374.
45. DOBIN, A., DAVIS, C.A., SCHLESINGER, F., et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013, 1, 29, 15-21.
46. DONG, X., VEGESNA, K., BROUWER, C., et al. SBGNview: Data Analysis, Integration and Visualization on All Pathways. R package version 1. 4. 1. Retrieved from **https://github.com/datapplab/SBGNview**. 2021.
47. DURINCK, S., SPELLMAN, P.T., BIRNEY, E., et al. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*. 2009, 8, 4, 1184-1191.
48. EGELHOFER, V., HOEHENWARTER, W., LYON, D., et al. Using ProtMAX to create high-mass-accuracy precursor alignments from label-free quantitative mass spectrometry data generated in shotgun proteomics experiments. *Nature Protocols*. 2013, 8, 595-595.

49. ENGSTROM, P.G., STEIJGER, T., SIPOS, B., et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods*. 2013, 12, 10, 1185-1191.
50. FALCON, S., GENTLEMAN, R., 2008, Hypergeometric testing used for gene set enrichment analysis, In: Bioconductor case studies. Springer, pp. 207-220.
51. FELLOWS, I. Package 'wordcloud'. R package version 2. 6. Retrieved from **https://CRAN.R-project.org/package=wordcloud**. 2018.
52. FONSECA, N.A., RUNG, J., BRAZMA, A., et al. Tools for mapping high-throughput sequencing data. *Bioinformatics*. 2012, 28, 3169-3177.
53. FRESE, K.S., KATUS, H.A., MEDER, B. Next-Generation Sequencing: From Understanding Biology to Personalized Medicine. *Biology (Basel)*. 2013, 378-398.
54. FULLER, C.W., MIDDENDORF, L.R., BENNER, S.A., et al. The challenges of sequencing by synthesis. *Nature Biotechnology*. 2009, 27, 1013-1013.
55. GAULTON, A., HERSEY, A., NOWOTKA, M., et al. The ChEMBL database in 2017. *Nucleic Acids Res*. 2017, D1, 45, D945-D954.
56. GENTLEMAN, R.C., CAREY, V.J., BATES, D.M., et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004, 10, 5, R80.
57. GENTZEL, M., KOCHER, T., PONNUSAMY, S., et al. Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics*. 2003, 8, 3, 1597-1610.
58. GRABHERR, M.G., HAAS, B.J., YASSOUR, M., et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*. 2011, 7, 29, 644.
59. GULCICEK, E.E., COLANGELO, C.M., MCMURRAY, W., et al. Proteomics and the analysis of proteomic data: an overview of current protein-profiling technologies. *Current protocols in bioinformatics*. 2005.
60. H. R. FULLER, G.E.M. Quantitative Proteomics Using iTRAQ Labeling and Mass Spectrometry. *Integrative Proteomics*. 2012.
61. HARRIS, M.A., CLARK, J., IRELAND, A., et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 2004, Database issue, 32, D258-261.
62. HASIN, Y., SELDIN, M., LUSIS, A. Multi-omics approaches to disease. *Genome Biology*. 2017, 1-15.

63. HERNANDEZ-DE-DIEGO, R., TARAZONA, S., MARTINEZ-MIRA, C., et al. PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res*. 2018, W1, 46, W503-W509.
64. HERNANDEZ, D., FRANÇOIS, P., FARINELLI, L., et al. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research*. 2008, 5, 18, 802-809.
65. HOWE, E.A., SINHA, R., SCHLAUCH, D., et al. RNA-Seq analysis in MeV. *Bioinformatics*. 2011, 22, 27, 3209-3210.
66. HRDLICKOVA, R., TOLOUE, M., TIAN, B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. 2017, 1, 8.
67. HUANG DA, W., SHERMAN, B.T., LEMPICKI, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009, 1, 37, 1-13.
68. IANNONE, R., ALLAIRE, J., BORGES, B. flexdashboard: R markdown format for flexible dashboards. R package version 0. 5. 2. Retrieved from **https://CRAN.R-project.org/package=flexdashboard**. 2018.
69. JAZAYERI, S.M., MELGAREJO MUÑOZ, L.M., ROMERO, H.M. RNA-SEQ: A GLANCE AT TECHNOLOGIES AND METHODOLOGIES. *Acta Biológica Colombiana*. 2015, 20, 23-35.
70. JECK, W.R., REINHARDT, J.A., BALTRUS, D.A., et al. Extending assembly of short DNA sequences to handle error. *Bioinformatics*. 2007, 21, 23, 2942-2944.
71. JONES, P., BINNS, D., CHANG, H.Y., et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014, 9, 30, 1236-1240.
72. KANEHISA, M., GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000, 1, 28, 27-30.
73. KASSAMBARA, A., MUNDT, F. Package 'factoextra'. Extract and visualize the results of multivariate data analyses. R package version 1. 0. 7. Retrieved from **https://CRAN.R-project.org/package=factoextra**. 2017.
74. KENT, W.J., SUGNET, C.W., FUREY, T.S., et al. The human genome browser at UCSC. *Genome Research*. 2002, 6, 12, 996-1006.
75. KERRIEN, S., ALAM-FARUQUE, Y., ARANDA, B., et al. IntAct-open source resource for molecular interaction data. *Nucleic Acids Res*. 2007, 35, D561.

76. KESHAVA PRASAD, T.S., GOEL, R., KANDASAMY, K., et al. Human Protein Reference Database--2009 update. *Nucleic Acids Res*. 2009, Database issue, 37, D767-772.
77. KIM, D., LANGMEAD, B., SALZBERG, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*. 2015, 4, 12, 357-360.
78. KIM, D., PERTEA, G., TRAPNELL, C., et al. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*. 2013, 4, 14.
79. KLEIN, T.E., ALTMAN, R.B. PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base. *Pharmacogenomics Journal*. 2004, 1, 4, 1-1.
80. KOBOLDT, D.C., CHEN, K., WYLIE, T., et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*. 2009, 17, 25, 2283-2285.
81. KOBOLDT, D.C., STEINBERG, K.M., LARSON, D.E., et al. The next-generation sequencing revolution and its impact on genomics. *Cell*. 2013, 1, 155, 27-38.
82. KOCKMANN, T., TRACHSEL, C., PANSE, C., et al. Targeted proteomics coming of age – SRM , PRM and DIA performance evaluated from a core facility perspective. *Proteomics*. 2016, 2183-2192.
83. KOHLER, U., LUNIAK, M.J. Data inspection using biplots. *The Stata Journal*. 2005, 2, 5, 208-223.
84. KOLKER, E., HIGDON, R., HOGAN, J.M. Protein identification and expression analysis using mass spectrometry. *Trends Microbiol*. 2006, 14, 229-235.
85. KOTOKA, E., ORR, M. Modifying SAMseq to account for asymmetry in the distribution of effect sizes when identifying differentially expressed genes. *Stat Appl Genet Mol Biol*. 2017, 5-6, 16, 291-312.
86. KRZYWINSKI, M., SCHEIN, J., BIROL, I., et al. Circos: an information aesthetic for comparative genomics. *Genome Research*. 2009, 9, 19, 1639-1645.
87. KUCUKURAL, A., YUKSELEN, O., OZATA, D.M., et al. DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics*. 2019, 1, 20, 6.
88. KUKURBA, K.R., MONTGOMERY, S.B. RNA sequencing and analysis. *Cold Spring Harbor Protocols*. 2015, 11, 2015, 951-969.

89. KUMAR, R., ICHIHASHI, Y., KIMURA, S., et al. A high-throughput method for Illumina RNA-Seq library preparation. *Front. Plant Sci.* 2012, August, 3, 1-10.
90. LAGARRIGUE, S., MARTIN, L., HORMOZDIARI, F., et al. Analysis of Allele-Specific Expression in Mouse Liver by RNA-Seq: A Comparison With eQTL Identified Using Genetic Linkage. *Genetics*. 2013, 3, 195, 1157.
91. LANGMEAD, B., SALZBERG, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012, 4, 9, 357-359.
92. LAW, C.W., CHEN, Y., SHI, W., et al. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*. 2014, 2, 15.
93. LEE, J.H., ANG, J.K., XIAO, X. Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. *RNA (A publication of RNA society)*. 2013, 19, 725-732.
94. LEEK, J.T., TAUB, M.A., RASGON, J.L. A statistical approach to selecting and confirming validation targets in -omics experiments. *BMC Bioinformatics*. 2012, 13, 150.
95. LEVCHENKO, M., GOU, Y., GRAEF, F., et al. Europe PMC in 2017. *Nucleic Acids Res*. 2018, D1, 46, D1254-D1260.
96. LI, H., HANDSAKER, B., WYSOKER, A., et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009, 16, 25, 2078-2079.
97. LI, W. Volcano plots in analyzing differential expressions with mRNA microarrays. *Journal of Bioinformatics and Computational Biology*. 2012, 1757-6334 10.
98. LIAO, Y., SMYTH, G.K., SHI, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2013, 7, 30, 923-930.
99. LICATA, L., BRIGANTI, L., PELUSO, D., et al. MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Research*. 2012, D1, 40, 857-861.
100. LIGTENBERG, W. A set of annotation maps for reactome. R package version 1. 74. 0. Retrieved from **https://bioconductor.org/packages/reactome.db/**. 2019.
101. LOHMANN, K., KLEIN, C. Next generation sequencing and the future of genetic diagnosis. *Neurotherapeutics*. 2014, 4, 11, 699-707.
102. LORAINE, A.E., BLAKLEY, I.C., JAGADEESAN, S., et al. Analysis and visualization of RNA-Seq expression data using RStudio, Bioconductor, and Integrated Genome Browser. *Methods Mol Biol*. 2015, 1284, 481-501.

103. LOVE, M.I., HUBER, W., ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014, 12, 15, 550.
104. LUO, R., LIU, B., XIE, Y., et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 2012, 1, 1, 2047-2217X-2041-2018.
105. LUO, W., BROUWER, C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*. 2013, 14, 29, 1830-1831.
106. MA, S., DAI, Y. Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics*. 2011, 6, 12, 714-722.
107. MACARTHUR, J., BOWLER, E., CEREZO, M., et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*. 2017, D1, 45, D896-D901.
108. MAGRANE, M., UNIPROT, C. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)*. 2011, 009.
109. MANNING, J. ShinyNGS: Shiny apps for NGS data. Retrieved from **https://github.com/pinin4fjords/shinyngs**. 2016.
110. MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 2011, 1, 17, 3.
111. MAXAM, A.M., GILBERT, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*. 1977, 2, 74, 560-564.
112. MCDERMAID, A., MONIER, B., ZHAO, J., et al. Interpretation of differential gene expression results of RNA-seq data: review and integration. *Brief Bioinform*. 2019, 6, 20, 2044-2054.
113. MCLAREN, W., GIL, L., HUNT, S.E., et al. The ensembl variant effect predictor. *Genome Biology*. 2016, 1, 17, 1-14.
114. MCPHERSON, J., ALLAIRE, J. rsconnect: Deployment Interface for R Markdown Documents and Shiny. R package version 0.8.18. Retrieved from **https://CRAN.R-project.org/package=rsconnect**. 2021.
115. MICHELE ARAÚJO PEREIRA, F.S.V.M.M.C.M.F., PATRÍCIA GONÇALVES PEREIRA, C., 2017, Application of Next-Generation Sequencing in the Era of Precision Medicine, In: Applications of RNA-Seq and Omics Strategies - From Microorganisms to Human Health. Intech open.

116. MILANEZ-ALMEIDA, P., MARTINS, A.J., GERMAIN, R.N., et al. Cancer prognosis with shallow tumor RNA sequencing. *Nat Med*. 2020, 2, 26, 188-192.
117. MORTENSEN, P., GOUW, J.W., OLSEN, J.V., et al. MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J Proteome Res*. 2010, 1, 9, 393-403.
118. MUDALIAR, M., TASSI, R., THOMAS, F.C., et al. Mastitomics, the integrated omics of bovine milk in an experimental model of Streptococcus uberis mastitis: 2. Label-free relative quantitative proteomics. *Mol Biosyst*. 2016, 9, 12, 2748-2761.
119. MUELLER, L.N., RINNER, O., SCHMIDT, A., et al. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics*. 2007, 19, 7, 3470-3480.
120. MURRELL, P., 2018, R graphics. R package version version 3.6.2. Retrieved from **https://cran.r-project.org/src/contrib/Archive/grid/**.
121. NAHNSEN, S., BIELOW, C., REINERT, K., et al. Tools for Label-free Peptide Quantification *Molecular & Cellular Proteomics*. 2013, 3, 12, 549-556.
122. OUGHTRED, R., RUST, J., CHANG, C., et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci*. 2021, 1, 30, 187-200.
123. PAGÈS, H., CARLSON, M., FALCON, S., et al. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. R package version 1.52.0. Retrieved from **https://bioconductor.org/packages/AnnotationDbi**. 2020.
124. PATEL, R.K., JAIN, M. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLOS ONE*. 2012, 2, 7, e30619.
125. PATRO, R., MOUNT, S.M., KINGSFORD, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*. 2014, 5, 32, 462-464.
126. PEPKE, S., WOLD, B., MORTAZAVI, A. Computation for chip-seq and rna-seq studies. *Nature Methods*. 2009, 11S, 6, S22.
127. PERKINS, D.N., PAPPIN, D.J.C., CREASY, D.M., et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS*. 1999, 18, 20, 3551-3567.

128. PERTEA, M., PERTEA, G.M., ANTONESCU, C.M., et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*. 2015, 3, 33, 290-295.
129. PEVZNER, P.A., TANG, H., WATERMAN, M.S. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*. 2001, 17, 98, 9748-9753.
130. PICO, A.R., KELDER, T., VAN IERSEL, M.P., et al. WikiPathways: pathway editing for the people. *PLoS Biol*. 2008, 7, 6, e184.
131. PITT, J.J. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *The Clinical biochemist. Reviews*. 2009, 1, 30, 19-34.
132. RAPLEE, I.D., EVSIKOV, A.V., MARIN DE EVSIKOVA, C. Aligning the Aligners: Comparison of RNA Sequencing Data Alignment and Gene Expression Quantification Tools for Clinical Breast Cancer Research. *J Pers Med*. 2019, 2, 9.
133. RAUDVERE, U., KOLBERG, L., KUZMIN, I., et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. 2019, W1, 47, W191-W198.
134. RITCHIE, M.E., PHIPSON, B., WU, D., et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015, 7, 43, e47.
135. ROBERTSON, G., SCHEIN, J., CHIU, R., et al. De novo assembly and analysis of RNA-seq data. *Nature Methods*. 2010, 11, 7, 909-912.
136. ROBINSON, M.D., MCCARTHY, D.J., SMYTH, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010, 1, 26, 139-140.
137. RONSEIN, G.E., PAMIR, N., VON HALLER, P.D., et al. Parallel reaction monitoring (PRM) and selected reaction monitoring (SRM) exhibit comparable linearity, dynamic range and precision for targeted quantitative HDL proteomics. *Journal of Proteomics*. 2015, 113, 388-399.
138. ROST, H.L., SACHSENBERG, T., AICHE, S., et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods*. 2016, 9, 13, 741-748.
139. ROUILLARD, A.D., GUNDERSEN, G.W., FERNANDEZ, N.F., et al. The harmonizome: a collection of processed datasets gathered to serve and mine

knowledge about genes and proteins. *Database-the Journal of Biological Databases and Curation*. 2016.

140. RSTUDIO TEAM. RStudio: integrated development for R. RStudio version 1.4.1103. Retrieved from **http://www.rstudio.com/**. 2015.
141. SALES, G., CALURA, E., CAVALIERI, D., et al. g raphite-a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*. 2012, 1, 13, 1-12.
142. SANGER, F., AIR, G.M., BARRELL, B.G., et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*. 1977, 5596, 265, 687-695.
143. SANTE, T., VERGULT, S., VOLDERS, P.-J., et al. ViVar: a comprehensive platform for the analysis and visualization of structural genomic variation. *PLOS ONE*. 2014, 12, 9, 113800.
144. SCHUBERT, O.T., ROST, H.L., COLLINS, B.C., et al. Quantitative proteomics: challenges and opportunities in basic and applied research. *Nat Protoc*. 2017, 7, 12, 1289-1294.
145. SCHULZ, M.H., ZERBINO, D.R., VINGRON, M., et al. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*. 2012, 8, 28, 1086-1092.
146. SHANNON, P., MARKIEL, A., OZIER, O., et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*. 2003, 11, 13, 2498-2504.
147. SIMPSON, J.T., WONG, K., JACKMAN, S.D., et al. ABySS: a parallel assembler for short read sequence data. *Genome Research*. 2009, 6, 19, 1117-1123.
148. SIMS, D., SUDBERY, I., ILOTT, N.E., et al. Sequencing depth and coverage : key considerations in genomic analyses. *Nature Publishing Group*. 2014, 2, 15, 121-132.
149. SMIRNOV, N. Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics*. 1948, 2, 19, 279-281.
150. SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V.K., et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005, 43, 102, 15545-15550.
151. SYED, F., GRUNENWALD, H., CARUCCIO, N. Next-generation sequencing library preparation : simultaneous fragmentation and tagging using in vitro transposition. *Nature Methods*. 2009, 11, 6.

152. SZKLARCZYK, D., GABLE, A.L., LYON, D., et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019, D1, 47, D607-D613.
153. SZKLARCZYK, D., MORRIS, J.H., COOK, H., et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*. 2017, D1, 45, D362-D368.
154. TARAZONA, S., FURIO-TARI, P., TURRA, D., et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res*. 2015, 21, 43, e140.
155. TARAZONA, S., GARCÍA, F., ALBERTO FERRER, J., et al. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet.journal*. 2012, 17B, 18.
156. TARCA, A.L., DRAGHICI, S., KHATRI, P., et al. A novel signaling pathway impact analysis. *Bioinformatics*. 2009, 1, 25, 75-82.
157. TEAM, R.C. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from **http://www.R-project.org/**. 2013.
158. THANKASWAMY-KOSALAI, S., SEN, P., NOOKAEW, I. Genomics Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics*. 2017, 3-4, 109, 186-191.
159. THOMAS, P.D., CAMPBELL, M.J., KEJARIWAL, A., et al. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research*. 2003, 9, 13, 2129-2141.
160. THORVALDSDÓTTIR, H., ROBINSON, J.T., MESIROV, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*. 2013, 2, 14, 178-192.
161. TIPNEY, H., HUNTER, L. An introduction to effective use of enrichment analysis software. *Human Genomics*. 2010, 3, 4, 202.
162. TRAPNELL, C., ROBERTS, A., GOFF, L., et al. RNA-seq TopHat and Cufflinks. *Nature Protocols*. 2012, 3, 7, 562-578.
163. TYAGI, P., BHIDE, M. History of DNA Sequencing *Folia Veterinaria*. 2020, 2, 64, 66-73.

164. TYANOVA, S., TEMU, T., COX, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*. 2016a, 12, 11, 2301-2319.
165. TYANOVA, S., TEMU, T., SINITCYN, P., et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods*. 2016b, 9, 13, 731-740.
166. VAN BREUKELEN, B., VAN DEN TOORN, H.W., DRUGAN, M.M., et al. StatQuant: a post-quantification analysis toolbox for improving quantitative mass spectrometry. *Bioinformatics*. 2009, 11, 25, 1472-1473.
167. VENTER, J.C., ADAMS, M.D., MYERS, E.W., et al. The sequence of the human genome. *Science*. 2001, 5507, 291, 1304-1351.
168. VERHAAK, R.G., SANDERS, M.A., BIJL, M.A., et al. HeatMapper: powerful combined visualization of gene expression profile correlations, genotypes, phenotypes and sample characteristics. *BMC Bioinformatics*. 2006, 7, 337.
169. VERHOECKX, K.C., BIJLSMA, S., DE GROENE, E.M., et al. A combination of proteomics, principal component analysis and transcriptomics is a powerful tool for the identification of biomarkers for macrophage maturation in the U937 cell line. *Proteomics*. 2004, 4, 4, 1014-1028.
170. VU, V.Q. Ggbiplot: A ggplot2 based biplot. R package version 0.55. Retrieved from **https://github.com/vqv/ggbiplot**. 2016.
171. WAGNER, G.P., KIN, K., LYNCH, V. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences*. 2012, 4, 131, 281-285.
172. WANG, K., LI, M., HAKONARSON, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. 2010, 16, 38, e164.
173. WANG, L., FENG, Z., WANG, X., et al. DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*. 2009a, 1, 26, 136-138.
174. WANG, L., WANG, S., LI, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*. 2012, 16, 28, 2184-2185.
175. WANG, W., QIN, Z., FENG, Z., et al. Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene*. 2013, 1, 518, 164-170.

176. WANG, Z., GERSTEIN, M., SNYDER, M. Expressed sequence tags (ests). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. 2009b, 1, 10, 57-63.
177. WARNES, M.G.R., BOLKER, B., BONEBAKKER, L., et al. Package 'gplots': Various R programming tools for plotting data. Package version 3. 1. 1. Retrieved from **https://CRAN.R-project.org/package=gplots**. 2016.
178. WARREN, R.L., SUTTON, G.G., JONES, S.J., et al. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*. 2007, 4, 23, 500-501.
179. WATSON, J.D., CRICK, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*. 1953, 4356, 171, 737-738.
180. WICKHAM, H., CHANG, W., WICKHAM, M.H. Package 'ggplot2': Create Elegant Data Visualisations Using the Grammar of Graphics. Package version 3. 3. 3. Retrieval from **https://CRAN.R-project.org/package=ggplot2**. 2016.
181. WICKHAM, H., FRANCOIS, R., HENRY, L., et al. dplyr: A grammar of data manipulation. R package version 1. 0. 5. Retrieval from **https://CRAN.R-project.org/package=dplyr**. 2015.
182. WILLIAMS, C.R., BACCARELLA, A., PARRISH, J.Z., et al. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics*. 2016, 1, 17, 103-103.
183. XIE, Y. knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29. Retrieval from **https://CRAN.R-project.org/package=knitr**. 2019.
184. XIE, Y., CHENG, J., TAN, X. DT: A Wrapper of the JavaScript Library "DataTables". R package version 0. 18. Retrieved from **https://CRAN.R-project.org/package=DT**. 2018.
185. YAN, J., RISACHER, S.L., SHEN, L., et al. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief Bioinform*. 2018, 6, 19, 1370-1381.
186. YANG, I.S., KIM, S. Analysis of Whole Transcriptome Sequencing Data : Workflow and Software. *Genomics Inform*. 2015, 119-125.
187. YANG, J., ZHANG, S., ZHANG, J., et al. Identification of key genes and pathways using bioinformatics analysis in septic shock children. *Infection and drug resistance*. 2018, 11, 1163-1174.

188. YATES, A.D., ACHUTHAN, P., AKANNI, W., et al. Ensembl 2020. *Nucleic Acids Res*. 2020, D1, 48, D682-D688.
189. YU, G., HE, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*. 2016, 2, 12, 477-479.
190. YU, G., WANG, L.-G., HAN, Y., et al. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*. 2012, 5, 16, 284-287.
191. ZALLEN, D.T. Despite Franklin's work, Wilkins earned his Nobel. *Nature*. 2003, 6953, 425, 15-15.
192. ZERBINO, D.R. Using the Velvet de novo assembler for short-read sequencing technologies. *Current protocols in bioinformatics*. 2010, 1, 31, 1-11.15. 12.
193. ZERBINO, D.R., BIRNEY, E. Velvet : Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*. 2008, 821-829.
194. ZHANG, Z.H., JHAVERI, D.J., MARSHALL, V.M., et al. A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data. 2014, 8, 9.
195. ZHAO, S., GUO, Y., SHENG, Q., et al. Heatmap3: an improved heatmap package with more powerful and convenient features. *BMC Bioinformatics*. 2014, 10, 15, P16.
196. ZHOU, Q., SU, X., WANG, A., et al. QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. *PLOS ONE*. 2013, 4, 8.