

# CHomics Tutorial

Version 1.0, Feb, 2020

The screenshot shows the CHomics v1 web interface. At the top, there is a green header bar with the following items from left to right: a logo, "CHomics (v1)", "Toolbox", "My Analyses", "Admin", "Projects (1)", "Comparisons (12)", "Samples (30)", "Hello, Demo User", and "Sign Out". Below the header, there is a light blue banner with the text "My Experiments and Analyses" and a "Hide" link. Underneath this, there are three project cards:

- CHO Demo**: Last updated 2019-10-10. Contains 5 samples and 1 analysis. The analysis is labeled "CHO Demo RNA-Seq Analysis (✓ Finished)".
- RNA-Seq Data**: Last updated 2019-10-08. Contains 26 samples and 3 analyses. One analysis is labeled "Test1 (✓ Finished)".
- Test**: Last updated 2019-10-11. Contains 0 samples and 0 analyses.

Below the project cards, there are three more sections:

- My Private Projects**: Shows 1 private project and a "Show" link.
- All Comparisons**: Shows 0 comparisons and a "Search" link.
- List of Comparisons**: Shows 0 comparisons and a "Show" link.

At the bottom of the page, there is a footer bar with the text "Powered by CanvasXpress.js, D3.js, Plotly.js, Highcharts.js, R, Bioconductor, HOMER, WikiPathways, KEGG, Reactome and BioInfoRx Application Platform."

<http://chomics.org>

user:demo@bioinforx.com  
password:CHO\_demo

From the login page, you can use your email to register an account that is recommended, as you will be able to save results and upload your own data. Otherwise just use guest account to view public data.

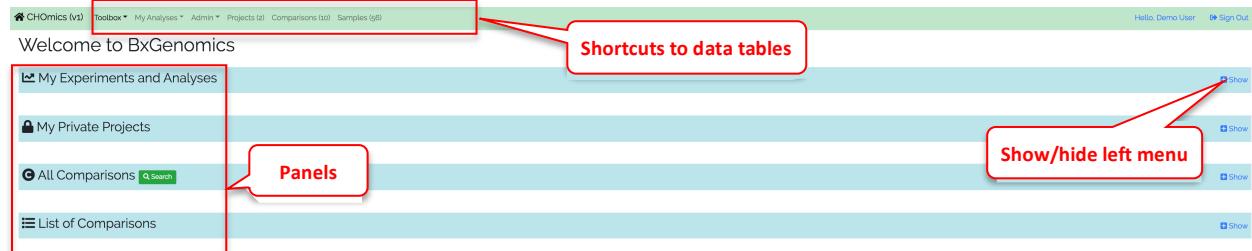
## Contents

1.	Overview of CHOmics .....	4
1.1	Menu Bar.....	4
1.2	Experiments and Analyses.....	4
1.3	Projects.....	5
1.4	Samples .....	5
1.5	Comparisons.....	6
1.6	Genes.....	8
2	Data Input.....	9
2.1	Upload fastq files to experiment.....	9
2.2	Upload data file to project .....	10
3	Data Analysis.....	12
3.1	RNAseq analysis pipeline.....	12
3.2	DE, GSEA and GO analysis .....	18
3.3	Saved Genes and Comparisons .....	21
3.4	Advanced Analysis.....	23
3.3.1	Correlation Tools.....	23
3.3.2	PCA Analysis .....	24
3.3.3	Meta-Analysis.....	26
4	Visualization.....	29
4.1	Visualize Gene Expression .....	29
4.1.1	View Gene Expression from multiple samples.....	29
4.1.2	View Gene Expression in Heatmap .....	32
4.1.3	Multi-omics Expression View .....	33
4.2	Visualize Comparison Data.....	34
4.2.1	Dashboard View of Comparison .....	34
4.2.2	Bubble Plot.....	36
4.2.3	Get significant genes from comparisons.....	41
4.2.4	Volcano Plot .....	44
4.3	Visualize functional pathway.....	46
4.3.1	Enrichment from Up and Down Regulated Genes .....	46
4.3.2	View Changed Genes from a Functional Term in Volcano Plot.....	47

4.3.3	View Enriched Pathways Directly from Comparison Details.....	49
4.3.4	Multi-layer visualization.....	50
4.3.5	Pathway Heatmap From Comparisons.....	54
5	Customized analysis pipeline .....	56
5.1	Use alternative tool or algorithm .....	56

# 1. Overview of CHomics

There are several panels stacked in the main interface. The recent experiment and projects are listed in the panel separately for quick access. You can also access them and other functions from the shortcuts at the top menu bar.



## 1.1 Menu Bar

In top menu bar, several shortcuts are listed for quick access of functions including: Toolbox, My Analysis, and Admin, Projects, Comparisons and Samples.

'Toolbox' contains a list of functional modules including: 'Import Project Data', 'Gene Expression Analysis', 'Comparison-based Analysis', 'Pathway Visualization' and 'Other tools'.

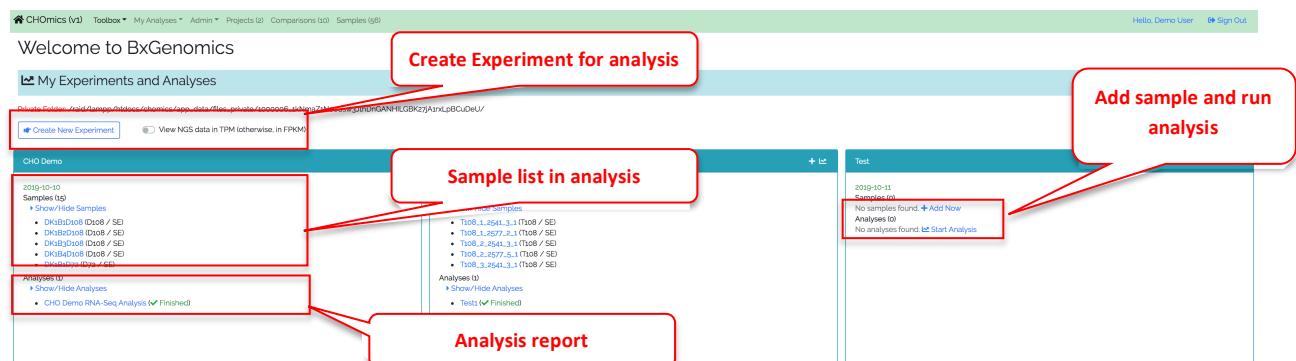
'My Analysis' provides quick access to the information of all 'Experiments', 'Samples' and 'Analysis'.

'Admin' allows the users to manage the data files from private folder, shared folder and overview the platforms applied to all data sets.

'Projects', 'Comparisons' and 'Samples' all provide searching function and access to specific project, comparison and sample respectively.

## 1.2 Experiments and Analyses

Experiment is designed for running the built-in RNA sequencing pipeline on the raw sequencing data. Once the experiment is created, users can upload raw fastq files and sample meta information, and then launch the built-in pipeline for analysis. After the analysis is completed, the analysis report is generated and the results can be exported as one 'Project' for visualization and cross-project comparison.



## 1.3 Projects

The project is used to perform data mining and data visualization. Users can either import analysis report from ‘Experiment’ or upload pre-processed data to create a project. In the project, users can easily explore different features of the data (e.g, Gene expression profiling, sample clustering, PCA, differential expression genes and pathways, etc), compare the analysis with the other projects or perform the meta-analysis by combining multiple projects.

The screenshot shows the CHOmics interface. At the top, there's a navigation bar with links like 'CHOmics (v1)', 'Toolbox', 'My Analyses', 'Admin', 'Projects (2)', 'Comparisons (16)', and 'Samples (56)'. On the right, it says 'Hello, Demo User' and 'Sign Out'. Below the navigation, there's a 'Welcome to BxGenomics' message. A red box highlights the 'Create Project' button. Another red box highlights the 'Sample and comparison list' section, which contains a sub-section for 'CH0 Demo - CH0 Demo RNA-Seq Analysis'. This section shows statistics: 'Created on 2019-10-23', 'There are 30 samples in this project.', and 'There are 5 comparisons in this project.' At the bottom of this section, there's a link to 'View Report'.

Each project mainly consists of samples including both meta information and omics profiling, and comparisons showing the statistical differences among samples.

## 1.4 Samples

A project may include many samples which can be searched by the ‘Sample’ in top menu bar. Each sample has its own properties including Species, CellType, DiseaseState,etc (details available by clicking the button on the left ends).

The screenshot shows the CHOmics interface focusing on samples. At the top, there's a 'Search' bar with dropdowns for 'Sample', 'Search options', 'Display setting', and 'Save as sample list'. Below the search bar, there are buttons for 'Advanced Search', 'Change Column Settings', 'Create Sample List', and 'Reset search conditions'. A red box highlights the 'Advanced Search' button. A note below the search bar says: 'Note: You can do quick search using the search box on the top, filter the table, or apply advanced search below.' Below the search area, there's a table with columns: ID, Name, Celltype, SampleSource, SamplingTime, and Treatment. A red box highlights the 'Name' column header. A red box also highlights the first row of the table, which contains sample details: ID DK1B1D108, Name DK1B1D108, Celltype Transcriptomics, SampleSource Transcriptomics, SamplingTime D108, and Treatment D108. The table also includes a 'Check/Uncheck All' checkbox at the top left and buttons for 'Column visibility', 'Copy', and 'CSV' at the top right. A red box highlights the 'Detailed view of sample' link at the top right of the table area.

ID	Name	Celltype	SampleSource	SamplingTime	Treatment
<input checked="" type="checkbox"/>	DK1B1D108		Transcriptomics		D108
<input checked="" type="checkbox"/>	DK1B2D108		Transcriptomics		D108
<input type="checkbox"/>	DK1B3D108		Transcriptomics		D108
<input type="checkbox"/>	DK1B4D108		Transcriptomics		D108
<input type="checkbox"/>	DK1B1D72		Transcriptomics		D72
<input type="checkbox"/>	DK1B2D72		Transcriptomics		D72
<input type="checkbox"/>	DK1B3D72		Transcriptomics		D72
<input type="checkbox"/>	DK1B1D84		Transcriptomics		D84
<input type="checkbox"/>	DK1B2D84		Transcriptomics		D84

To change columns displayed in the table, using the table settings (green button). Users can also select the samples to save them into the sample list. Samples from the list can be loaded to other analysis or visualization stools like heatmap.

Each sample has a gene expression profile. In CHOmics, there are multiple ways to analyze and visualize the samples including: correlation tool (noted by 'C'), gene expression plot( noted by 'E'), expression heatmap (noted by 'H'), and PCA analysis (noted by 'P').

Sample: DK1B1D108

» Search All Samples

Found in project: CHO Demo - CHO Demo RNA-Seq Analysis

Found in comparisons: D108vs.D72 D108vs.D84 D108vs.D96

**Tools for sample analysis**

**Gene Expression Correlation** **Gene Expression Plot** **Gene Expression Heatmap** **PCA Analysis**

Sample Details	
ID	1
Project Name	CHO Demo - CHO Demo RNA-Seq Analysis
Platform	NGS_Mouse
PlatformName	Generic Mouse NGS Platform
Species	Mouse
Description	CHO sample DK1-B1-D108. Time D108. Replicate B1
CellType	
DiseaseState	
Gender	
Organism	
SamplePathology	
Projects ID	1
Platforms ID	2
Platform Type	NGS
Samples ID	27
Name	DK1B1D108
SampleIndex	27
DiseaseCategory	
Ethnicity	
Infection	
Response	
SampleSource	Transcriptomics

## 1.5 Comparisons

Comparison is defined by the comparative analysis between two groups of samples including differential gene analysis and pathway enrichment analysis.

There are a lot of meta data available for each comparison. See the dashboard for an overview of key categories, and the detailed description of each comparison has the full information.

CHOMics (v1) Toolbox My Analyses Admin Projects (2) Comparisons (16) Samples (56) Hello, Demo User Sign Out

**Search Comparison** **Display Comparison** **Save to comparison list** **Save to sample list**

Note: You can do quick search via the search box on the top right, or use the search bar below, or apply advanced search.

Advanced Search  Change Column Settings  Create Comparison List  Create a Sample List  Reset search conditions

Check/Uncheck All

Column visibility Copy CSV Show **Detailed view of comparison** Search: [ ]

ID	Name	Case SampleIDs	Control SampleIDs
<input checked="" type="checkbox"/> B M H C V W R K 1	D84vs.D72	Show/Hide	Show/Hide
<input checked="" type="checkbox"/> B M H C V W R K 2	Dg6vs.D72	Show/Hide	Show/Hide
<input type="checkbox"/> B M H C V W R K 3	D108vs.D72	Show/Hide	Show/Hide
<input type="checkbox"/> B M H C V W R K 4	Dg6vs.D84	Show/Hide	Show/Hide
<input type="checkbox"/> B M H C V W R K 5	D108vs.D84	Show/Hide	Show/Hide
<input type="checkbox"/> B M H C V W R K 6	D108vs.D96	Show/Hide	Show/Hide

The selected comparisons can be saved to the comparison list (yellow button) for easy loading into the plotting tools.

Several options on each comparison for complicated visualization and analysis are also listed including: bubble plot of gene expressions(noted by 'B'), meta analysis (noted by 'M'), pathway heatmap plot(noted by 'H'), significant changes genes (noted by 'C'), volcano plot(noted by 'V'), Wikipathway mapping(noted by 'W'), and Rectome and KEGG pathway mapping (noted by 'R' and 'K' respectively).

CHOMics (v1) Toolbox My Analyses Admin Projects (2) Comparisons (16) Samples (56) Hello, Demo User Sign Out

Comparison: D84.vs.D72

[» Search All Comparisons](#) [» View Comparison Genes](#) [» Edit Comparison Details](#)

**Options for comparison analysis and visualization**

ID	Name	Project	Category	DiseaseState	Case Samples	Control Samples
1	D84vs.D72	CHO Demo - CHO Demo RNA-Seq Analysis		Unknown Disease	Show/Hide	Show/Hide

View Details

Up Regulated Genes [» Download SVG File](#)

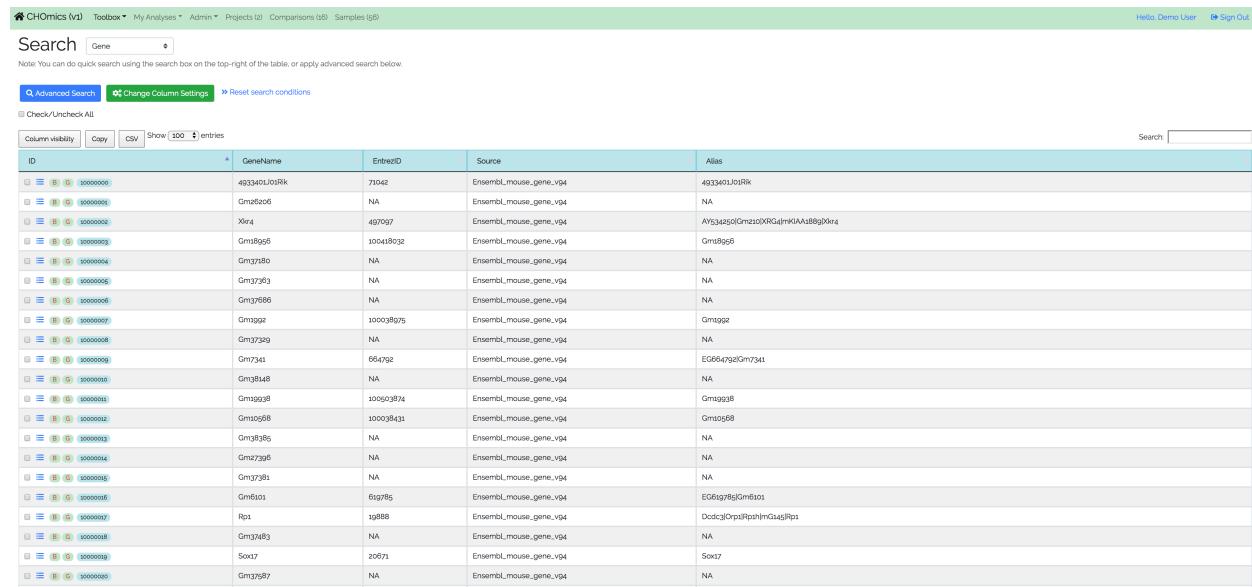
Biological Process

Number of Genes

Cellular Component  
Molecular Function  
KEGG  
Molecular Signature  
Interpro Protein Domain  
Wiki Pathway  
Reactome  
[» Enrichment Report](#)

## 1.6 Genes

The genome-wide gene expression values were detected in each sample using RNA-Seq or microarrays. All the human genes that have expression values are listed in gene table. The gene annotation from difference platforms were all mapped to NCBI gene ID (EntrezID) for consistence across platforms.



The screenshot shows a gene table in the CHOMics software. The table has columns for ID, GeneName, EntrezID, Source, and Alias. The table contains 20 rows of gene information. The first few rows include:

ID	GeneName	EntrezID	Source	Alias
Gm1860	4933401j0IRK	71042	Ensembl_mouse_gene_v94	4933401j0IRK
Gm2606		NA	Ensembl_mouse_gene_v94	NA
Xkr4		497097	Ensembl_mouse_gene_v94	AY534250 Gm210XKG mrkIAjd8g Xkr4
Gm18656		105418032	Ensembl_mouse_gene_v94	Gm18656
Gm2780		NA	Ensembl_mouse_gene_v94	NA
Gm17953		NA	Ensembl_mouse_gene_v94	NA
Gm17686		NA	Ensembl_mouse_gene_v94	NA
Gm1992		10003875	Ensembl_mouse_gene_v94	Gm1992
Gm17349		NA	Ensembl_mouse_gene_v94	NA
Gm7341		664792	Ensembl_mouse_gene_v94	EG664792 Gm7341
Gm18148		NA	Ensembl_mouse_gene_v94	NA
Gm19938		100503874	Ensembl_mouse_gene_v94	Gm19938
Gm10568		100038431	Ensembl_mouse_gene_v94	Gm10568
Gm17865		NA	Ensembl_mouse_gene_v94	NA
Gm27956		NA	Ensembl_mouse_gene_v94	NA
Gm17951		NA	Ensembl_mouse_gene_v94	NA
Gm6031		619785	Ensembl_mouse_gene_v94	EG619785 Gm6031
Rp1		19888	Ensembl_mouse_gene_v94	Dcds3 Rp1 Rp1 mG4d Rp1
Gm1783		NA	Ensembl_mouse_gene_v94	NA
Sox17		20571	Ensembl_mouse_gene_v94	Sox17
Gm17957		NA	Ensembl_mouse_gene_v94	NA

To find a gene, you can use gene symbol, gene description, gene alias, NCBI gene ID, Ensembl gene ID or Uniprot ID.

For some common genes, the symbols used in publications are often not the official symbol, and you can try search alias field. For example, TP53 is often referred to as P53 in publication. You need to search P53 in alias or tumor protein p53 in description to find it if you don't know its official symbol.

The NCBI Gene search <https://www.ncbi.nlm.nih.gov/gene> is a good source to get official gene symbols and IDs.

You can view full details of a gene by clicking the  button .

Gene: Gm18956

[» Search All Genes](#)

[G Gene Expression Plot](#) [B Gene Bubble Plot](#)

**View expression plot**

**View Bubble plot across comparisons**

Gene Details			
ID	10000003	Species	Mouse
GeneIndex	10000003	GeneName	Gm18956
EntrezID	100418032	Source	Ensembl_mouse_gene_v94
Description	predicted gene_18956	Alias	Gm18956
Ensembl	ENSMUSG00000102851	Unigene	NA
Uniprot	NA	TranscriptNumber	1
Strand	+	Chromosome	1
Start	3252757	End	3253236
ExonLength	480	GeneID	Gm18956
AccNum	NA	Biotype	processed_pseudogene

From gene details, you can access RNA-Seq data in a box plot, or view all comparisons including this gene in a bubble plot.

## 2 Data Input

### 2.1 Upload fastq files to experiment

After the experiment is created, users can upload fastq or fastq.gz files through remote URLs, server files or local files. The files are uploaded to the private folder named ‘Experiments’ automatically.

CHOMics (v1) Hello, Demo User Sign Out

Experiment: Test [Delete Experiment](#)

Valid data files in current experiment: 0

[Upload Files](#) [Manage files in Private Files](#)

No files added to this experiment yet.

Manager files in folder

Drag & Drop a File  
Or, select an option below:

[Remote URLs](#) [Server Files](#) [Local Files](#)

Experiment Details [Edit](#)  
Time Created: 2019-10-11  
Description: Test

Experiment Samples [Add Samples One by One or in Batch](#)  
No samples found.

Experiment Analyses

In the folder 'Experiments', there may be multiple subfolders corresponding to different experiments. Users can easily modify the folder or upload new files to the folder.

CHOMics (v1) Hello, Demo User Sign Out

Treeview All Files Experiments

Name	Size	Last Modified
e1	29.00 B	2019-10-08 11:33
e2	17.00 B	2019-10-10 09:15
e3	2.00 B	2019-11-20 13:28

External Link (Share among applications)  
[/Experiments/1](#)  
[/Experiments/2](#)  
[/Experiments/3](#)

Actions [New Folder](#) [New File](#) [Upload File](#) [Batch Upload](#)

Subfolders

Upload files to the folder

## 2.2 Upload data file to project

Besides raw RNA sequencing data (fastq files), CHOMics also allow the input of other types of data to start a project, including meta data (i.e.project and samples), expression data, and summary data. Those data should be uploaded in comma separated values (CSV) or tab separated values (TSV) with either

fixed or flexible format.

The screenshot shows the 'Import Project Data' section of the CHOmics(v1) interface. A red box highlights the 'Import Project Data' button. Another red box highlights the 'Import Data Files with...' dropdown, which has a sub-section titled 'Flexible Formats'. A third red box highlights the 'Import data file with fixed format' link. A fourth red box highlights the 'Import data file with flexible format' link. A note at the bottom left says 'Note: All files should be in CSV format.' A note at the bottom right says 'Note: All files should be in TSV format. The first row must contain column names, which have to be the exact names of database table fields. Each file must have at least two columns.'

Project file can be uploaded to create a new project. The project file contains some required information such as Name, Platform and other optional fields such as Disease, Description etc.

Sample file can be uploaded to register samples for a project. The sample file contains required information such as Name and Project\_Name and optional fields such as Description, Tissue, DiseaseState, SampleSource, Gender, etc.

Expression file can be uploaded with quantified expression measure at gene level. The expression could be transcriptomics, proteomics or other gene-level counts. The file is required to contain GeneName, SampleName, Value.

Comparison file and Comparison data file are used to upload summary results for statistical comparison test applied externally. Comparison file needs to contain the Project\_Name, Case\_SampleIDs, and Control\_SampleIDs while comparison data file contains statistical results such as GeneName, ComparisonName, Log2FoldChange, PValue, Adjusted PValue for each comparison.

The screenshot shows the 'Import Project Data' section of the CHOmics(v1) interface. A red box highlights the 'Choose File' buttons for 'Projects', 'Samples', 'Comparisons', 'Sample Expression Data', and 'Comparison Data'. A red box highlights the 'Example File' links for 'Example File', 'Example File 1', 'Example File 2', 'Example File 3', and 'Example File 4'. A red box highlights the 'Import different data files' and 'Example file format' buttons at the bottom. A note at the top says 'Detailed Explanation of File Formats' and 'Import Files with Flexible Formats'. A note at the bottom says 'Note: You can upload one or more files from one or multiple projects. You DON'T have to upload all five files.' A checkbox 'Update if sample expression data already imported' is shown. A note at the bottom right says 'Note: You can upload one or more files from one or multiple projects. You DON'T have to upload all five files.'

## 3 Data Analysis

### 3.1 RNAseq analysis pipeline

After fastq files are uploaded to the experiment by following the Section 2.1, users can start the analysis by applying the built-in pipeline mainly including: Raw Data QC (quality control), Alignment, Gene Counts and QC, and DEG, GSEA and GO analysis. After the analysis is completed, the results can be exported into a project for visualization.

The screenshot shows the CHOmics interface with the following details:

- Analysis Details:** Experiment: CHO Demo, Time Created: 2019-10-10 09:15:01, Name: CHO Demo RNA-Seq Analysis, Description: (Not set).
- Analysis Samples:** (Data Type SE), 15 samples are used. Buttons: Show All Samples, Select Samples and Files, Create Sample.
- Analysis Steps and Progress:** Buttons: Duplicate Analysis, Finalize Analysis. A red box highlights the "Analysis pipeline for RNASEq" section, which contains:
  - A tip: "Tip: Please select one or multiple analysis steps to get started."
  - Checkboxes for Step 1: Raw Data QC, Step 2: Alignment with Subread, Step 3: Gene Counts and QC, and Step 4: DEG, GSEA and GO Analysis.
  - Status: All steps are finished.
  - Step Files: Reports for D84 vs D72, Dg6 vs D72, D108 vs D72, Dg6 vs D84, D108 vs D84, D108 vs D96, and GSEA Analysis Report.
- Report files:** A red box highlights the "Report files" section, which lists the generated reports.

After completion of each step, a report is generated for summarizing the metrics in each step to quantify raw data QC, alignment with Subread method, and gene count distribution and sample/gene count QC, respectively.

In the report for raw data QC, all fastq files are verified in quality by software fastQC. Sequencing read information and quality control metrics are summarized for each individual fastq file.

The screenshot shows the BxGenomics - Raw Sequencing Data QC interface with the following details:

- Read information:** A red box highlights the "Read information" section, which includes:
  - The fastQC program is used to verify raw data quality of the Illumina reads.
  - The table below shows a summary of basic statistics. Click the file name to open the detailed report for each individual sample.
  - If the sequencing run is paired-end (PE), you may have two files per sample with R1 and R2 in the file names respectively.
- Sample fastq file:** A red box highlights the "Sample fastq file" section, which lists the following files:

Filename	Total Sequences	Sequences flagged as poor quality	Sequence length	%GC	#Total Deduplicated Percentage
NG-7391_T96_4_RNA20140328RA_llb4411B_2577_2_1.fastq.gz	32221046	0	51	52	47.2%
NG-7391_T108_1_RNA20140328RA_llb4411B_2541_3_1.fastq.gz	27636352	0	51	55	50.0%
NG-7391_T72_1_RNA20140328RA_llb4410B_2577_5_1.fastq.gz	47835535	0	51	53	42.7%
NG-7391_T108_4_RNA20140328RA_llb4412B_2577_2_1.fastq.gz	32151429	0	51	53	46.6%
NG-7391_T84_1_RNA20140328RA_llb4411B_2541_3_1.fastq.gz	25000801	0	51	53	53.3%
NG-7391_T84_4_RNA20140328RA_llb4411A_2577_2_1.fastq.gz	29244828	0	51	52	48.7%

The table below show pass/fail for several QC metrics. Click the file name to open individual reports. You can view fastQC documentation to get more information about the QC metrics.

Please note that for RNA-Seq data, it is normal to observe a few failed metrics, which usually will not affect subsequent data analysis. First, per base sequence content (and Kmer content) will often fail fastQC due to non-random base content at the first -12 bases. This is because the random primers used during reverse transcription step are actually not totally random in terms of base content. Second, the sequence duplication levels of RNA-Seq data are usually high because many transcripts are highly expressed.

### QC metrics

File Name	Basic Statistics	Per base sequence quality	Per tile sequence quality	Per sequence quality scores	Per base sequence content	Per sequence GC content	Per base N content	Sequence Length Distribution	Sequence Duplication Levels	Oversupersent sequences	Adapter Content
NG-7391_T96_4_RNA20140328RA_llb44118_2577_2_1.fastq.gz	PASS	PASS	WARN	PASS	FAIL	PASS	PASS	PASS	FAIL	WARN	PASS
NG-7391_T108_1_RNA20140328RA_llb44119_2541_3_1.fastq.gz	PASS	PASS	PASS	PASS	FAIL	WARN	PASS	PASS	WARN	FAIL	PASS
NG-7391_T72_1_RNA20140328RA_llb44108_2577_5_1.fastq.gz	PASS	PASS	WARN	PASS	FAIL	PASS	PASS	PASS	FAIL	PASS	PASS
NG-7391_T108_4_RNA20140328RA_llb44122_2577_2_1.fastq.gz	PASS	PASS	WARN	PASS	FAIL	PASS	PASS	PASS	FAIL	PASS	PASS
NG-7391_T84_1_RNA20140328RA_llb44111_2541_3_1.fastq.gz	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	PASS	WARN	WARN	PASS
NG-7391_T84_4_RNA20140328RA_llb44114_2577_2_1.fastq.gz	PASS	PASS	WARN	PASS	FAIL	PASS	PASS	PASS	FAIL	WARN	PASS

In the report for alignment, parameter setting and quality metrics (e.g, mapped, junctions,etc) for alignment are listed for each fastq file.

CHOMics (v1) Toolbox ▾ My Analyses ▾ Admin ▾ Projects (2) Comparisons (16) Samples (56) Hello, Demo User Sign Out

## BxGenomics - Sequence Alignment Logs

Subread: v1.5.0-p1 (<http://subread.sourceforge.net/>)

---

### Subjunc Settings

```
Function : Read alignment + Junction detection (RNA-Seq)
Threads : 6
Input file : /raid/lampp/htdocs/chomics/app_data/analysis/2_yr ...
Output file : /raid/lampp/htdocs/chomics/app_data/analysis/2_yr ...
Index name : /var/www/html/cho_genomics/app_data/files_core/PI ...
Phred offset : 33

Min votes : 1 / 14
Allowed mismatch : 3 bases
Max indels : 5
# of Best mapping : 1
Unique mapping : no
Hamming distance : no
Quality scores : no
```

**Summary:**

```
Processed : 27636352 reads
Mapped : 26781244 reads (96.9%)
Junctions : 111928
Indels : 46043

Running time : 12.6 minutes
```

In the report for Gene Counts and QC step, several metrics have been calculated and plotted for comprehensive evaluation of genes and samples, including: reads mapping to genes, distribution of detected genes, percentage of reads for highly expressed genes, normalization and boxplot of gene expression, sample grouping and clustering, sample correlation and outlier detection.

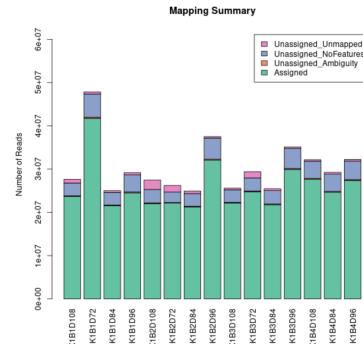
'1. Assign reads to genes' plots the mapping summary of reads to genes, showing the percentage of reads assigned to genes or unassigned due to unmapping, no features or ambiguous mapping.

BxGenomics - RNA-Seq QC Report

## 1. Assign Reads to Genes

The alignment bam files were compared against the gene annotation GFF file, and raw counts for each gene were generated using the [featureCounts](#) tool from subread. The graph below shows mapping and gene assignment summary. Click the graph to download the pdf version. You can also [download the csv file that contains the numbers](#).

- Download the raw gene counts in CSV format. This file lists the number of reads mapped to each gene.



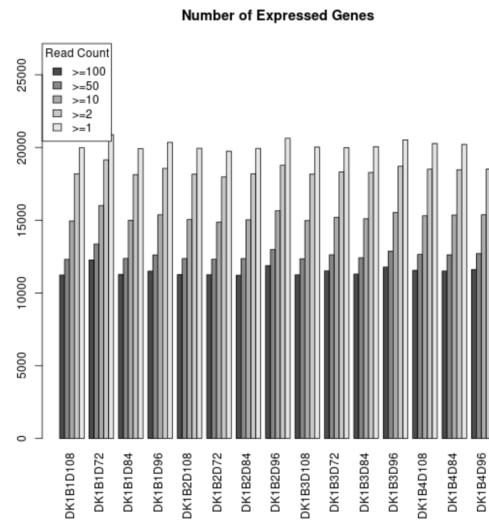
It is normal to observe some variation in number of reads across samples. However, samples with extremely low number of reads may not be suitable for downstream analysis, and we recommend checking the additional QC metrics below to identify potential outliers to exclude from downstream analysis.

'2. Number of Genes Detected' plots the number of expressed genes with read count from intervals of  $\geq 1$ ,  $\geq 2$ ,  $\geq 10$ ,  $\geq 50$  and  $\geq 100$ .

‘3. Percentage Reads from Most Highly Expressed Genes’ plots the percentage of reads mapped to the top expressed genes (up to 100 genes).

## 2. Number of Genes Detected

Next, we performed additional QC at gene level. We first looked at number of genes detected. We count the number of genes that have at least 1, 2, 10, 50 or 100 counts. In generally, number of genes with 2 or more counts can be used as a rough estimate of how many genes are expressed. Genes with only 1 read could be noise. In addition, the number of genes with 10 or more reads is a good indicator of how many genes have enough reads for downstream statistical analysis. Click the graph to download a pdf version, you can also download a csv file [containing the numbers](#).



We also try to detect outliers from this step. Any samples that show very small number of genes with 10 or more reads are potential outliers. The cutoff we used is 1/2 of the median across all samples.

## 3. Percentage Reads from Most Highly Expressed Genes

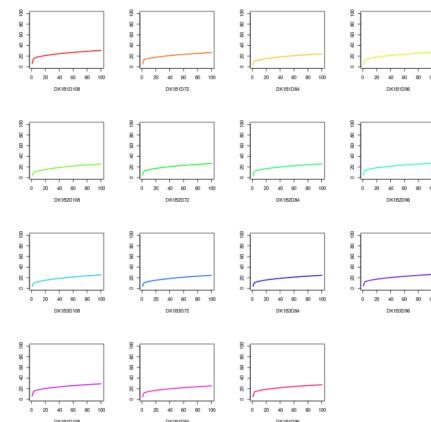
We also look at the percentage of reads belonging to the top genes. Basically we rank the genes by read counts, and compute the percentage of reads belonging to the top genes (up to top 100).

If majority of the reads come from top genes, then the sample probably has bottlenecking issues where a few genes were amplified many times by PCR during library preparation.

Most samples should have ~ 20% reads mapped to the top 100 genes.

If the top 100 genes account for more than 35% of all reads, we consider this sample as a potential outlier.

Click the graph to download a pdf version. You can also download a csv file [download the csv file that contains the numbers](#).



## ‘4. Normalization and Boxplot of Gene Expression’ evaluates gene expression after normalization by TMM method and then draws boxplot of normalized expression (logCPM: log of counts per million reads) after log2 transformation.

#### 4. Normalization and Boxplot of Gene Expression

The raw counts data were further processed by the following steps:

a) Remove genes that were not expressed. If a gene has counts per million (CPM) value  $>=1$  in at least two of the samples, we consider it expressed in the experiment and include it for downstream QC analysis. From 32871 total genes, 14212 genes are selected as expressed and used in downstream QC analysis.

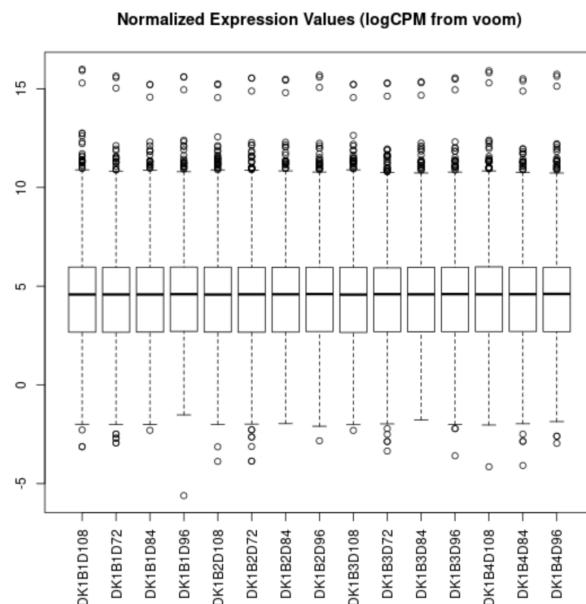
b) The [TMM normalization method](#) was used to scale samples to remove differences in the composition of the RNA population between samples. It is performed with the [edgeR package](#). The normalization factors for all samples are listed below. You can [download the csv file](#).

At this step, we also try to identify outliers that have extreme normalization factors ( $>1.5$  or  $<0.66$ ). Note sometimes samples with large biological differences can have extreme normalization factors.

Name	group	lib.size	norm.factors
DK1B1D108	1	23619476	0.927
DK1B1D72	1	41669151	1.010
DK1B1D84	1	21522805	1.028
DK1B1D96	1	24476888	0.995

c) The normalized gene counts were transformed to log2 scale using [voom method](#) from the [R Limma package](#). We created boxplot for each sample to summarize gene expression.

Since this is normalized data, most samples should look similar. Samples with high or low distribution may be outliers (or have large biological differences).

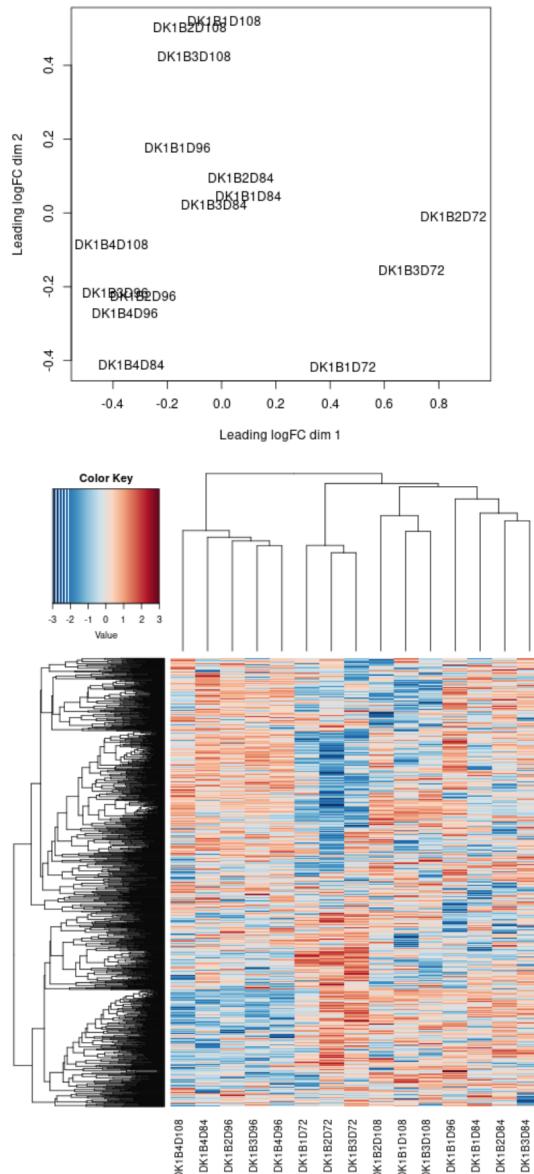


'5 Grouping and Clustering of Samples' plots the relationship among samples by multidimensional scale. Samples are clustered by hierarchical clustering method based on the expression of top genes with large variation( $SD/mean > 0.3$ ).

## 5. Grouping and Clustering of Samples

a) We first create multidimensional plot to view sample relationships. This is done using [R Limma package](#).

Here biological replicates should cluster together, and difference conditions ideally should separate from each other.



b) Very often hierarchical clustering can give better indication of the sample and gene relationships. We used [made4 package](#) from R to cluster samples and draw a heatmap.

We selected genes that have variable expression across samples to make the heatmap. These variable genes were chosen based on standard deviation (SD) of expression values larger than 30% of the mean expression values (Mean). If there are more than >5000 variable genes, we first remove genes with mean logCPM<1, then rank genes by SD/Mean to get the top 5000 genes.

The heatmap is created from 1124 variable genes.

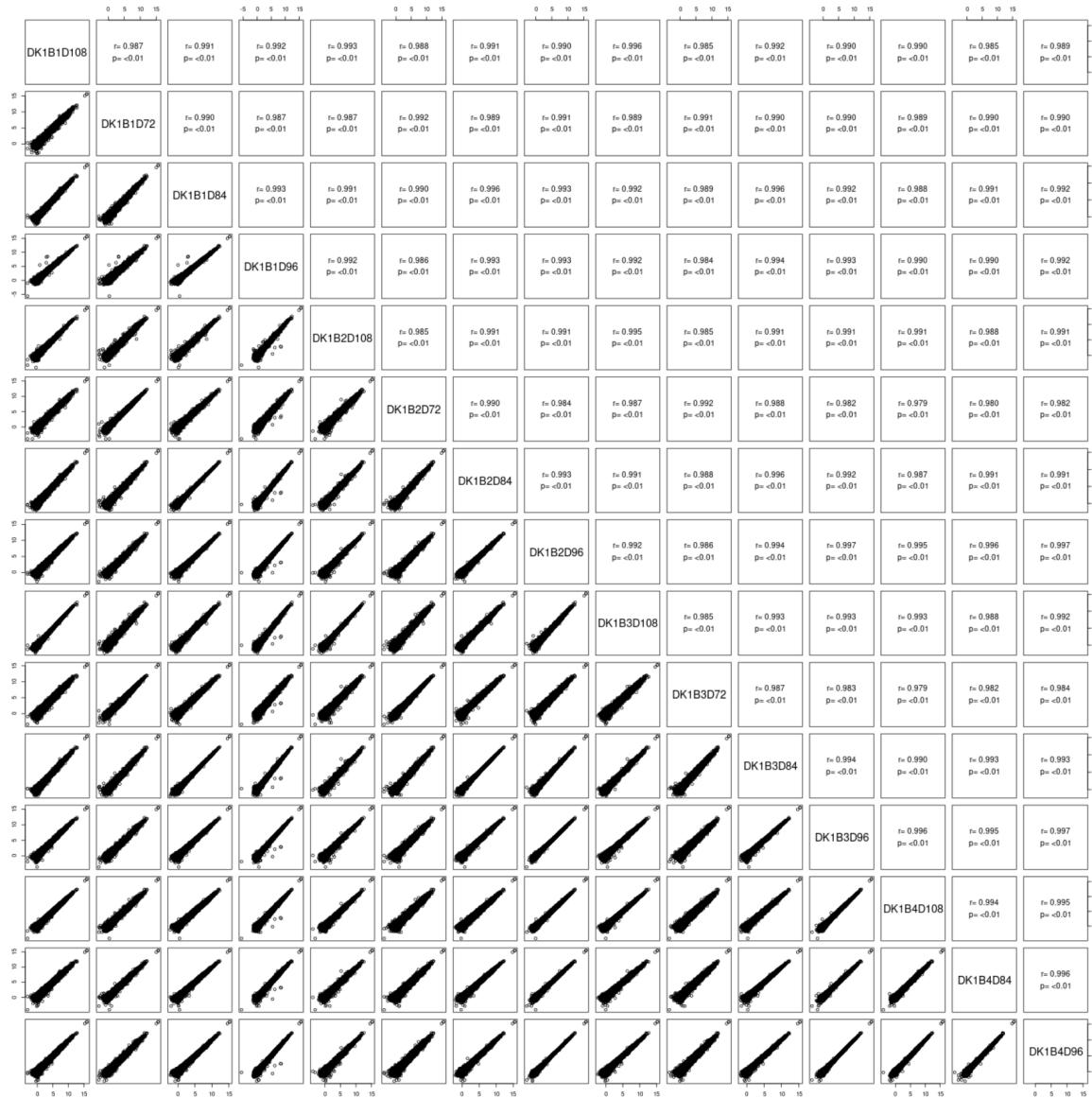
In the heatmap above, we selected genes that changed across samples (normally by SD/mean > 0.3), and plotted the relatively gene expression levels (blue is low, red is high). Gene names are not shown due to large number of genes used to create the heatmap. Both genes and samples are clustered in the heatmap. Normally biological replicates should cluster together, and ideally there should be up- or down- regulated genes between different conditions.

Heatmap can be used to detect overall patterns, as well as outlier samples.

'6 Sample correlation' creates scatter plots for the correlation between sample pairs. The idea is that biological replicates from the same group should look similar in the scatter plot, and should have high correlation values compared to the samples from other groups.

## 6. Sample correlation

We also created scatter plots for the correlation between sample pairs. If there are many samples, you may need to download the graph and view it at full size. Again, the idea here is that biological replicates should look similar in the scatter plot, and should have high correlation values.



## 3.2 DE, GSEA and GO analysis

After completing the first three steps for sample quality control and gene count readout, users can start statistical analysis as the last step of pipeline, including differential expression analysis (DEG), gene set enrichment analysis (GSEA) and gene ontology (GO) analysis.

DEG analysis is applied to compare gene expression between two groups, namely comparison. Users can design one or multiple comparisons for DEG analysis. In each comparison, differential expressed genes are identified by LIMMA model, followed by GSEA pathway analysis and GO enrichment analysis which explore the enrichment of DEGs in diverse pathways.

**Parameter for gene set analysis**

**Design comparison for DEG analysis**

The reports for DEG and pathway analysis are attached for each comparison after completion of analysis.

**Report for DEG, GSEA and GO analysis**

In the report for DEG analysis, the table summarizing DEGs with up- and down-regulation is listed along with a heatmap clustering the DEGs expression (up to top 1000 DEGs).

CHOMics (v1) Toolbox My Analytics Admin Projects (2) Comparisons (16) Samples (6) Hello, Demo User Sign Out

Comparisons D64vsD72 Dg6vsD72 D108vsD64 D108vsD84 D108vsD96

## Differentially Expressed Genes for D108.vs.D72

**1. Summary of Differentially Expressed Genes (DEGs)**

For comparison, the number of differentially expressed genes (DEGs) are shown below. Up-regulated genes are high in the first condition while down-regulated genes are high in the second condition.

Cutoff	Standard (two fold and FDR 0.05)
# of Up-regulated genes	171
# of Down-regulated genes	45
Total number of changed genes	216*
	233

\*These genes are reported in the table for differentially expressed genes (DEGs) and shown in the DEG heat maps.

**DEGs summary table**

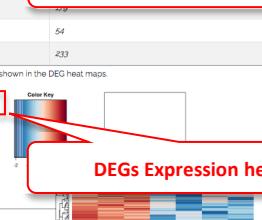
**Summary heat map showing all the differentially expressed genes (DEGs)**

Here log<sub>2</sub> gene expression levels were scaled and used to cluster genes and samples. Each row is a gene, each column is a sample. Gene names are not shown in this plot. Click image to download PDF version.

» View detailed heat map showing names for all DEGs (if there are more than 1000 DEGs, only the top 1000 genes are shown)

Sometimes, you may want to view gene-only clustering while leaving the sample order unchanged. You can view such heat maps here: [Summary](#) or [Detailed Plots](#)

**DEGs Expression heatmap**



The heatmap displays gene expression levels on a color scale from blue (low) to red (high). The y-axis lists genes, and the x-axis lists samples. A color key at the top indicates the scale from -3 to 3. The samples are grouped into four main clusters, corresponding to the four conditions listed on the left: D108 vs D72, D108 vs D64, D108 vs D96, and D108 vs D84. Within each cluster, genes are clustered by expression pattern.

In the report for GO enrichment analysis, barplots show the significance of enrichment of up- or down-regulated DEGs in different pathway databases, e.g., GO, KEGG, Wiki pathways, etc.

Gene Ontology Enrichment Analysis Results

### Introduction

This page displays the top 10 lists from functional enrichment of differential expressed genes.

Comparisons: D84vsD72, Dg6vsD72, Ds08vsD72, Dg6vsD64, Dr08vsDr64

**Pathway analysis for up-regulated DEGs**

Up Regulated Genes	» Download SVG File
Biological Process	
Cellular Component	
Molecular Function	
KEGG	
Molecular Signature	
Interpro Protein Domain	
Wiki Pathway	
Reactome	
» Enrichment Report	

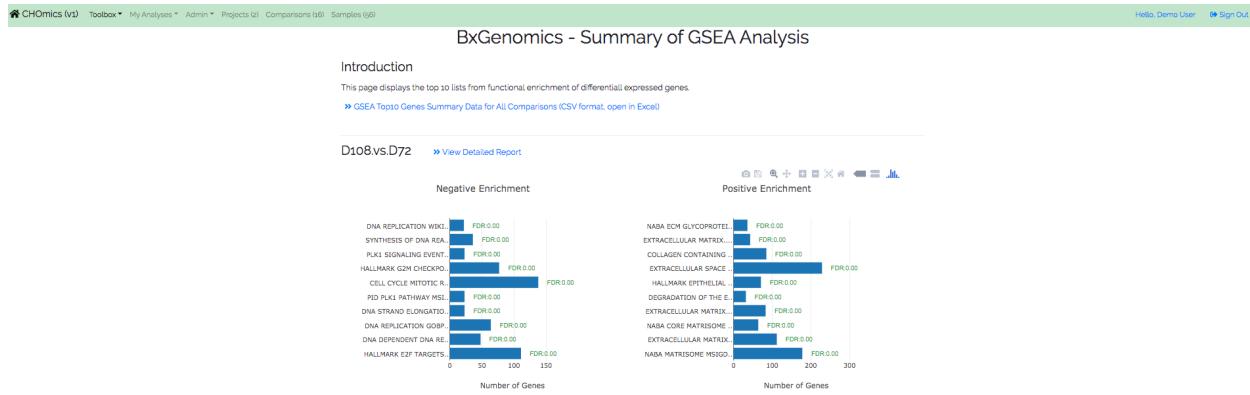
Pathway	log(p)
positive regulation of corticotropin secretion	-0.34
regulation of corticotropin secretion	-0.08
regulation of cAMP-mediated signaling	-0.68
negative regulation of immunoglobulin production	-0.89
positive regulation of cardiac muscle contraction	-0.90
regulation of interleukin-7 receptor proliferation	-0.21
positive regulation of striated muscle contraction	-0.75
regulation of interleukin-5 production	-0.37
protein refolding	-0.28
negative regulation of multi-cellular organismal process	7.19

**Pathway analysis for down-regulated DEGs**

Down Regulated Genes	» Download SVG File
Biological Process	
Cellular Component	
Molecular Function	
KEGG	
Molecular Signature	
Interpro Protein Domain	
Wiki Pathway	
Reactome	

Pathway	log(p)
response to hypoxia	-10.16
response to mercury ion	-9.94
response to decreased oxygen levels	-8.67
negative regulation of nucleic acid metabolism	-9.24
response to nitric oxide	-9.12
response to oxygen levels	-8.85
embryonic placenta development	-7.48
cellular response to dexamethasone stimulus	-7.98
response to oxygen-containing compound	-7.76
response to dexamethasone	-7.80

Similarly, in the report for GSEA analysis, enrichment results for up- and down- regulated DEGs in pathways from MigDB database are plotted with significance level (FDR), respectively.



### 3.3 Saved Genes and Comparisons

Customers can save selected genes or comparison for future use (e.g. multiple gene and multiple comparisons bubble plot). From gene search, check the genes you want to save, and click the yellow button "save selected genes"

ID	GeneName	EntrezID	Source	Alias
<input checked="" type="checkbox"/> 11000020	glutamate	57	NA	NA
<input checked="" type="checkbox"/> 11000021	glutamine	53	NA	NA
<input checked="" type="checkbox"/> 11000022			NA	NA
<input type="checkbox"/> 11000023			NA	NA
<input type="checkbox"/> 11000024	N-acetylglutamine	33943	NA	NA
<input type="checkbox"/> 11000025	glutamate_gamma-methylester	33487	NA	NA
<input type="checkbox"/> 11000026	pyroglutamine	46225	NA	NA
<input type="checkbox"/> 11000027	N-acetyl-l-aspartyl-glutamate_NAAG	35665	NA	NA

You can do the same with comparisons.

ID	Name	Case SampleIDs	Control SampleIDs
<input checked="" type="checkbox"/> D108.D72		Show/Hide	Show/Hide
<input checked="" type="checkbox"/> D108.D84		Show/Hide	Show/Hide
<input checked="" type="checkbox"/> D108.D96		Show/Hide	Show/Hide
<input type="checkbox"/> Protein_D108.D72		Show/Hide	Show/Hide
<input type="checkbox"/> Comparison_Protein_D108.D84		Show/Hide	Show/Hide
<input type="checkbox"/> Comparison_Protein_D108.D96		Show/Hide	Show/Hide

To view selected genes or comparison, click “My Results” link on the left menu and then “Gene Lists”.

CHOMics (v1) Toolbox My Analyses Admin Projects (12) Comparisons (12) Samples (30)

**Search**

Note: You can do quick search in the top-right of the table, or apply advanced search below.

**Advanced Search**

Check/Uncheck All

**ID**

- Comparison-based Analysis
  - Volcano Plot
  - Bubble Plot
  - Significantly Changed Genes
  - Pathway Heatmap
  - Export Comparison Data
- Pathway Visualization
  - KEGG Pathway View
  - Reactome Pathway View
  - WikiPathway View

Showing 1 to 6 of 6 entries

**Comparison-based Analysis**

- Volcano Plot
- Bubble Plot
- Significantly Changed Genes
- Pathway Heatmap
- Export Comparison Data

**Pathway Visualization**

- KEGG Pathway View
- Reactome Pathway View
- WikiPathway View

**Other Tools**

- My Saved Lists
- Functional Enrichment
- Overlap and Venn Diagrams
- Search Functional Gene Lists
- Compare Gene Lists
- Meta Analysis
- Manage Platforms

CHOMics (v1) Toolbox My Analyses Admin Projects (12) Comparisons (12) Samples (30) Hello, Demo User Sign Out

**My Saved Lists**

Category: (All)

Column visibility Copy CSV Show 100 entries

Name	Category	Count	Time
bubble_demo	Comparison	6	2020-02-25
bubble_demo	Gene	15	2020-02-25
DEGs_D108vsD72	Gene	169	2019-10-24
Genes_allcompare	Gene	19	2020-01-14
RNA and Proteomics	Sample	30	2019-11-03

Show 1 to 5 of 5 entries Previous 1 Next

One additional way to select and save genes or comparisons is from the ‘significantly changed genes’ in toolbox. Using the dynamic filters to choose the comparisons or genes you are interested in, and you can use the table at the bottom to save comparisons or genes.

CHOMics (v1) Toolbox My Analyses Admin Projects (12) Comparisons (12) Samples (30)

**Search**

Note: You can do quick search in the top-right of the table, or apply advanced search below.

**Advanced Search**

Check/Uncheck All

**ID**

- Comparison-based Analysis
  - Volcano Plot
  - Bubble Plot
  - Significantly Changed Genes
  - Pathway Heatmap
  - Export Comparison Data
- Pathway Visualization
  - KEGG Pathway View
  - Reactome Pathway View
  - WikiPathway View

Showing 1 to 6 of 6 entries

**Comparison-based Analysis**

- Volcano Plot
- Bubble Plot
- Significantly Changed Genes
- Pathway Heatmap
- Export Comparison Data

**Pathway Visualization**

- KEGG Pathway View
- Reactome Pathway View
- WikiPathway View

**Comparisons:**

Or, upload your comparison files:  No file chosen

**Display Options:**  Log2FC  PValue  FDR

**Fold Change Cutoff:**  Both Up- and Down-regulated

**Statistic Cutoff:**

**List Genes:**  Common Genes from All Comparisons  Genes from Any Comparisons

**4. Save comparison list**

**6. Save gene list**

**5. Select genes**

The screenshot shows the CHOMics interface with several tabs at the top: Save Comparison List, Bubble Plot, Pathway Heatmap, Meta Analysis, Export Comparison Data, WikiPathways, Reactome Pathways, KEGG Pathways, Save Gene List, Gene Expression Plot, Heatmap, Correlation Tool, PCA Analysis, and Export Expression Data. Below these are buttons for Column visibility and Copy. A search bar is at the top right. The main area displays a table of genes with various columns for Log2FC, PValue, and FDR. Red boxes highlight the 'Save comparison list' button, the 'Save gene list' button, and a row in the table labeled 'Adm'.

## 3.4 Advanced Analysis

Besides the above analyses, the CHOMics also provides several advanced tools.

### 3.3.1 Correlation Tools

Once the user has identified a gene of interest, the user can use correlation tools to find other genes that share similar (or opposite) profiles in terms of gene expression or fold change. First, enter the gene of interest, and samples to be used for correlation. In the example below, we entered a saved gene list, and 15 samples.

**1. Start here**

**2. Enter gene of interests**

**3. Enter or load saved samples**

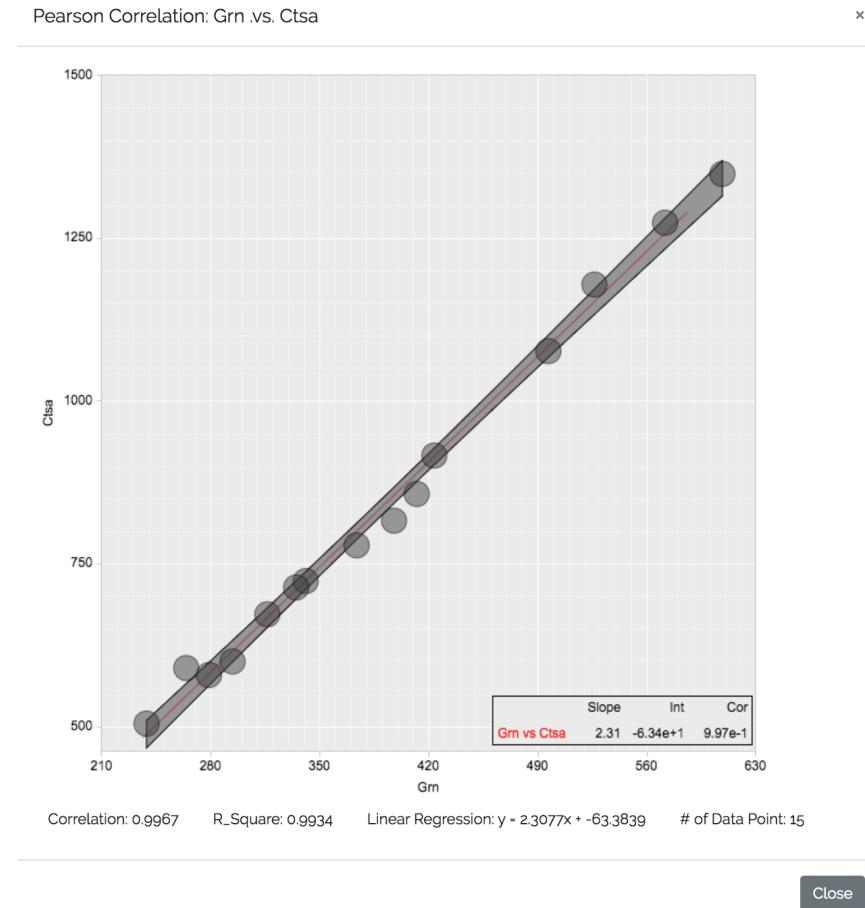
**4. Set options**

**5. Submit**

**The result shows a table of correlated genes, ranked by R<sup>2</sup>**

The screenshot shows the Correlation Tool interface. It includes sections for Genes (with a 'Load from saved list' button), Samples (with a 'Load from saved list' button), and Advanced Options (with 'Calculate the correlations against all available genes in database' and 'Calculate the correlations among the entered genes only' checkboxes). Below these are sections for Correlation Method (Person Correlation, Spearman Correlation, Enable Log Transform), and a 'Submit' button. A note says '1496 records found'. The results table at the bottom has columns for Source Gene, Matched Gene, Correlation Coef, R Square, # of Data Points, and Actions (with 'Plot' links).

Click the plot icon will show scatter plot of the target and the correlated gene (e.g, gene Grn vs. Ctsa).



### 3.3.2 PCA Analysis

You can select a set of samples and genes and use PCA plot to visualize the sample relationships on the target gene set.

CHOMics (v1)   Toolbox   My Analyses   Admin   Projects (2)   Comparisons (1)   Samples (56)   Hello, Demo User   Sign Out

**PCA Analysis**

1. Enter or load gene of interests

2. Enter or load saved samples

3. Select sample attributes for visualization

Genes:

Samples:

Note: You must enter one or more sample names.

Sample Attributes: (3 selected)

Age    CellType    Collection    DiseaseCategory    DiseaseState    Ethnicity    Flag Remark    Flag To Remove    Gender    Infection    Organism    RIN Number    RNASeq Assignment Rate    RNASeq Mapping Rate    RNASeq Total Read Count  
 Response    SamplePathology    SampleSource    SampleType    SamplingTime    Symptom    Tissue    TissueCategory    Transfection    Treatment    Uberon ID    Uberon Term    Check All    Check None

The system will use FactorMineR package to run PCA analysis and display the results. Several PCA metrics are plotted for interactive visualization:

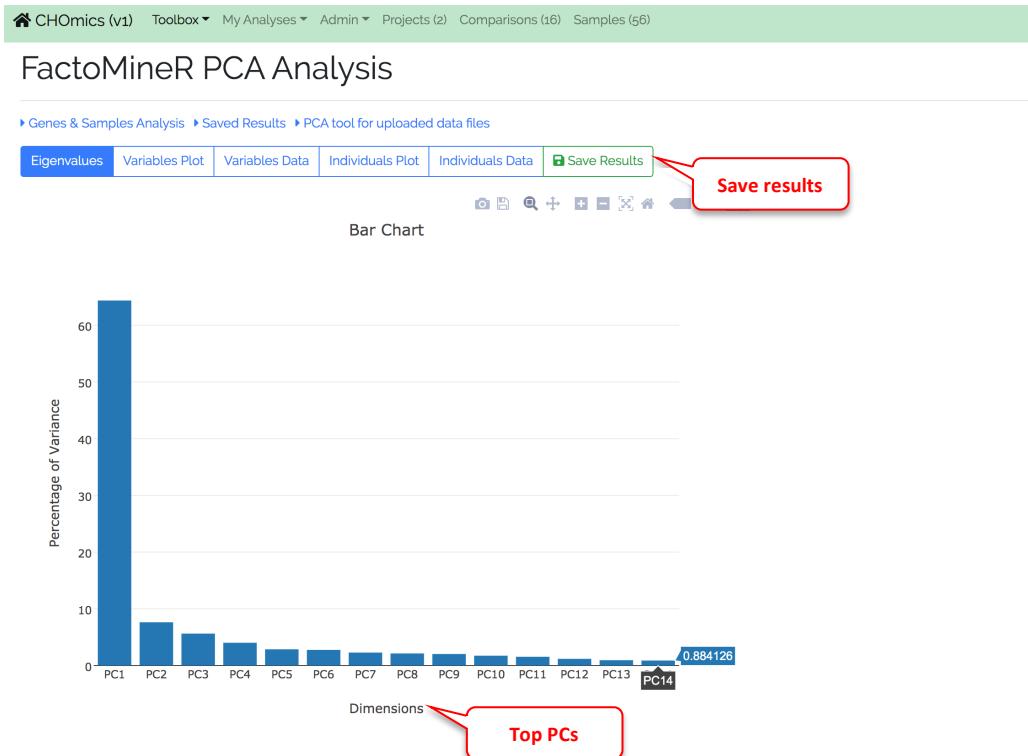
Eigenvalues plot the percentage of variance explained by top PCs.

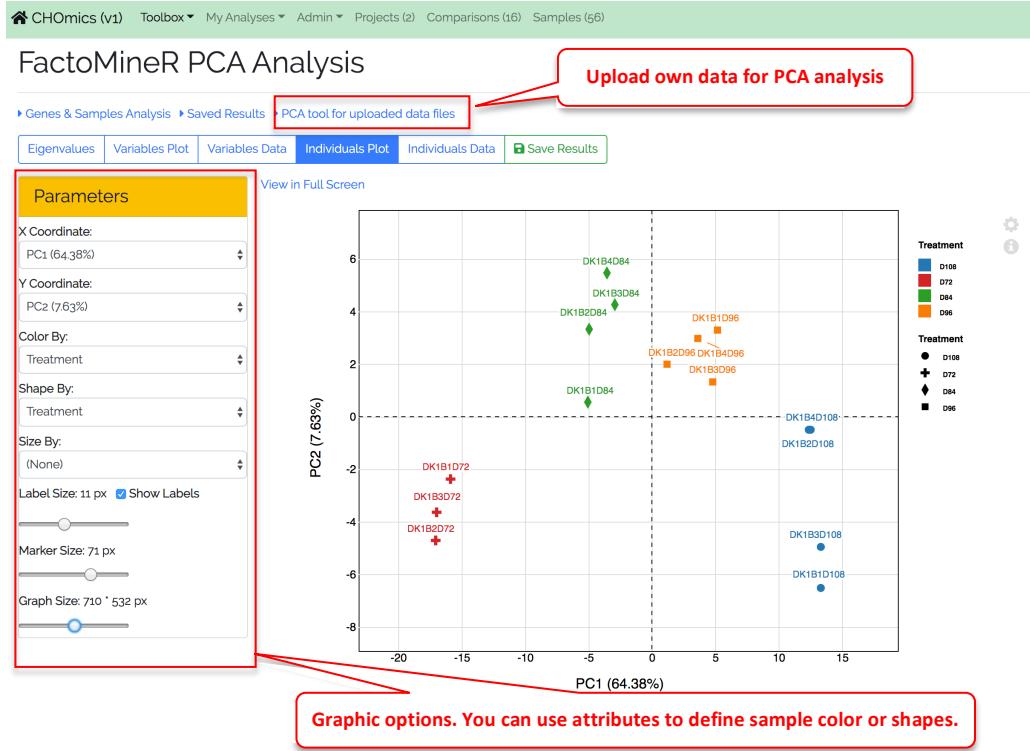
Variables Plot shows the weights of top contributing genes in each PCs.

Variable Data summarizes the weights of each gene in each individual PC.

Individuals Plot shows the relationship of samples on the spanned space by different PCs.

Individual Data summarizes the score vector of each sample in each individual PC.





The PCA results can be saved. Users can load it in the future.

#### My PCA Saved Results

My PCA Saved Results		
Show 100 entries		
Title	Description	Actions
Test_PCA		<a href="#">Delete</a>

Showing 1 to 1 of 1 entries

Users can upload their own data matrix or pre-calculated data for PCA analysis and visualization.

#### PCA Scatter Plot Tool

► Genes & Samples Analysis ► Saved Results ► PCA tool for uploaded data files

Upload Files: Data file is required

Data matrix for PCA

Data File:	<input type="file"/> Choose File	No file chosen	<a href="#">Example Data File</a>
Attributes File:	<input type="file"/> Choose File	No file chosen	<a href="#">Example Attributes File</a>
Variance File:	<input type="file"/> Choose File	No file chosen	<a href="#">Example Variance File</a>
Format:	<input checked="" type="radio"/> csv	<input type="radio"/> txt / tsv	

### 3.3.3 Meta-Analysis

Meta-Analysis can be used to identify genes that are changed consistently across multiple projects. It is listed as one functional module in toolbox panel. In the example below, we are looking for the most significant DEGs in three comparisons.

The screenshot shows the CHOMics (v1) interface for Meta Analysis. It includes sections for 'Saved Meta Analysis Results' (with a red box around 'Save meta-analysis results'), 'Enter or load comparison list' (with a red box around 'Enter or load comparison list'), 'Enter or load gene list' (with a red box around 'Enter or load gene list'), 'Attributes to show in results table' (with a red box around 'Attributes to show in results table'), and 'Optional parameter setting for cutoff on meta-analysis statistical results' (with a red box around 'Optional parameter setting for cutoff on meta-analysis statistical results'). A 'Submit' button is at the bottom.

The meta-analysis pipeline will compute three types of results:

- 1) Maximum p-value (maxP). This method targets on DEGs have small *p*-values in "all" comparisons. We recommend using maxP if you are looking for DEGs that are common among several studies.
- 2) Fisher's p-value. The Fisher's method sums up the log-transformed p-values obtained from individual studies. This p-value combination method is useful if you want to identify DEGs in any of the comparisons.
- 3) We also applied simple counting method to report the frequency a gene is classified as up or down-regulated DEG from all the comparisons. The default DEG cutoff is two-fold change and FDR<0.05. but user can change the cutoff.

In most cases, combining maxP (smaller values are more significant) and the counting method (e.g. up-regulated in 50% of studies) will give the most biological relevant results for consistently regulated genes across comparisons.

The screenshot shows the CHOMics Meta Analysis Result interface. It includes sections for 'Optional tools for visualizing and saving results' (with a red box around 'Optional tools for visualizing and saving results'), 'Filtering' (with a red box around 'Filtering'), 'Columns in results table' (with a red box around 'Columns in results table'), and a 'Select genes for visualization and other analysis' button (with a red box around it). Below is a table with columns: GeneName, EntrezID, Up.Per, RankProd, RP\_logFC, RP\_Pval, RP\_FDR, Combined\_Pval\_Fisher, Combined\_Pval\_maxP, Combined\_FDR\_Fisher, and Combined\_FDR\_maxP. A red box highlights the first few rows of the table.

GeneName	EntrezID	Up.Per	RankProd	RP_logFC	RP_Pval	RP_FDR	Combined_Pval_Fisher	Combined_Pval_maxP	Combined_FDR_Fisher	Combined_FDR_maxP
Mb21d1	214763	66.6667	8.7760	0.8800	0.0033	0.2758	0.0000	0.0000	0.0000	0.0000
Snrnd14e	100302594	33.3333	79260	1.0500	0.0023	0.3927	0.0000	0.0000	0.0000	0.0000
Hspa1a	193740	33.3333	157900	0.8400	0.0199	0.6728	0.0001	0.0001	0.0002	0.0004
LOC100689269	100689269	33.3333	149500	0.7800	0.0170	0.7182	0.0000	0.0000	0.0000	0.0000
LOC100689270	100689270	33.3333	139000	0.7400	0.0138	0.7744	0.0000	0.0000	0.0000	0.0000
Calcr	12311	33.3333	36.0400	0.7500	0.1531	1.2320	0.0000	0.0000	0.0000	0.0000
LOC113832837	113832837	33.3333	33.8000	0.3300	0.1341	1.2590	0.0000	0.0000	0.0001	0.0000

In the above example, we used a relatively loose filtering criterion (`N.data.points>1`, and `up-regulation in percentage>30%` of studies, and `Combined_Pval_MaxP <=0.0001`) because only small number of genes pass the stringent default criterion.

Display Options

---

N.data.points >=

Percentage Up-Regulated >=

Percentage Down-Regulated >=

Combined\_Pval\_MaxP Cutoff <=

RP\_Pval <=

Pval Fisher Cutoff <=

Number of records to show:  100  1000  3000 (limit)

---

Update Settings Cancel

The data table shows the genes that pass the filters. We can sort the table by maxP value. A different filter can be applied to get down-regulated genes.

The results can be saved for future access. There are also links to several other tools. The download meta data link will save a CSV file that contain results from all genes.

Next, we will choose all the genes that pass filter by checking the box for all listed genes, and use bubble plot to visualize the results.

#### Bubble Plot Multiple

[» Single Gene Plot](#)

Genes: [» Load from saved lists](#) [Q Load functional gene sets](#) [X Clear](#)

Calcr  
Hsp90aa1  
LOC100689269  
LOC100689270  
LOC113832837  
Mb21d1

**Selected gene list**

Comparisons: [» Load from saved lists](#) [Q Search and Select](#) [Q Select a Project](#) [X Clear](#)

D108 vs D96  
D84 vs D72  
D96 vs D84

**Comparison list for meta-analysis**

Note: You must enter one or more gene names.

Note: You must enter one or more comparison names.

[Toggle Advanced Settings](#)

Chart Height Scale Factor:

Chart Left Margin Scale Factor:

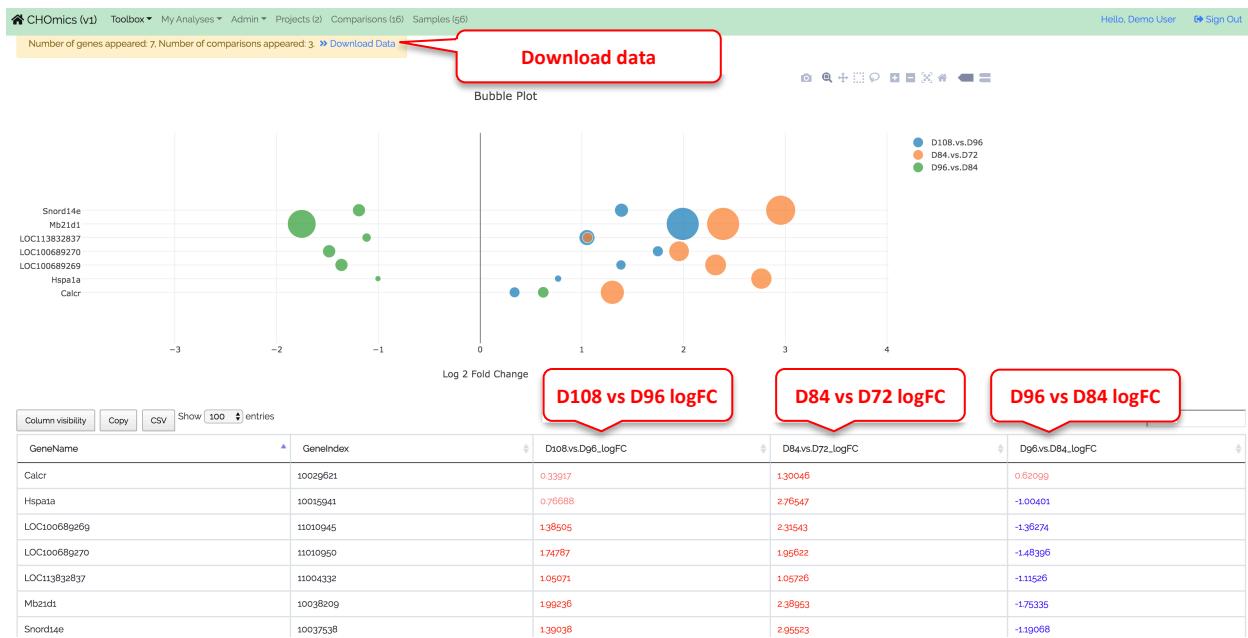
Show Columns in Table:

Log2FC  P-Value  FDR

**Unclick p-value and FDR to show only logFC**

**Submit**

The resulting bubble plot will show all three comparisons for each gene.



The data table below the bubble plot can also be used for filtering. Remember in the advanced settings, we choose to display logFC only, this makes it easier to look for genes that are reverted in different time points. The logFC values are color coded (red, increase, blue, decrease), therefore we can see that most of genes show upregulation in D84, and then downregulation in D96 and then upregulation in D108.

You can also redo the plot, check all columns to include p-value and FDR in the table, and export the results to excel file.

The workflow above uses up-regulated genes as example. You can get down-regulated genes from the filter step in meta-analysis result page.

## 4 Visualization

### 4.1 Visualize Gene Expression

CHOMics provides tool to easily visualize gene expression level across multiple genes, samples and omics. For each gene, you can view its expression levels across multiple samples.

#### 4.1.1 View Gene Expression from multiple samples

Choose the Gene Expression tool from Toolbox -> Gene Expression Plot from top menu, and enter the official symbol of genes or load gene list from saved lists. Alternatively, in the gene details page, click View Gene Expression link.

**Gene Expression Tool**

**1. Enter or load gene symbol or list**

**2. Enter or load sample list**

**3. Select data platform and type**

**4. (Optional) Choose attributes, apply data filters**

As an optional step, you can choose what sample attributes to pass to the plot, and use data filter to choose only a subset of data points.

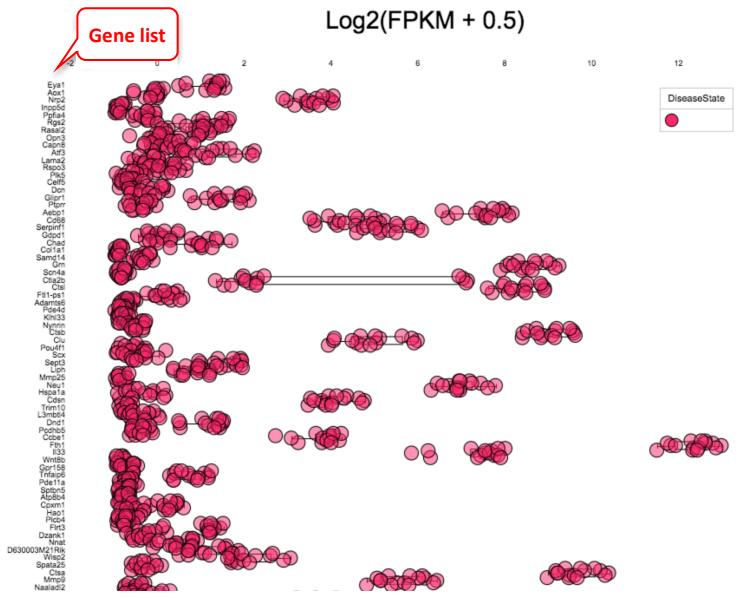
The Data Filter can be very useful if there are too many data points, and you want to focus on a few diseases or tissue types.

The screenshots below show default boxplot showing all samples by different time points (i.e, treatment).

## Summary of Data

- 169 genes found.
- 15 samples found: DK1B1D108, DK1B2D108, DK1B3D108, DK1B4D108, DK1B1D72, DK1B2D72, DK1B3D72, DK1B1D84, DK1B2D84, DK1B3D84, DK1B4D84, DK1B1D96, DK1B2D96, DK1B3D96, DK1B4D96

Download: [Raw Data File](#)



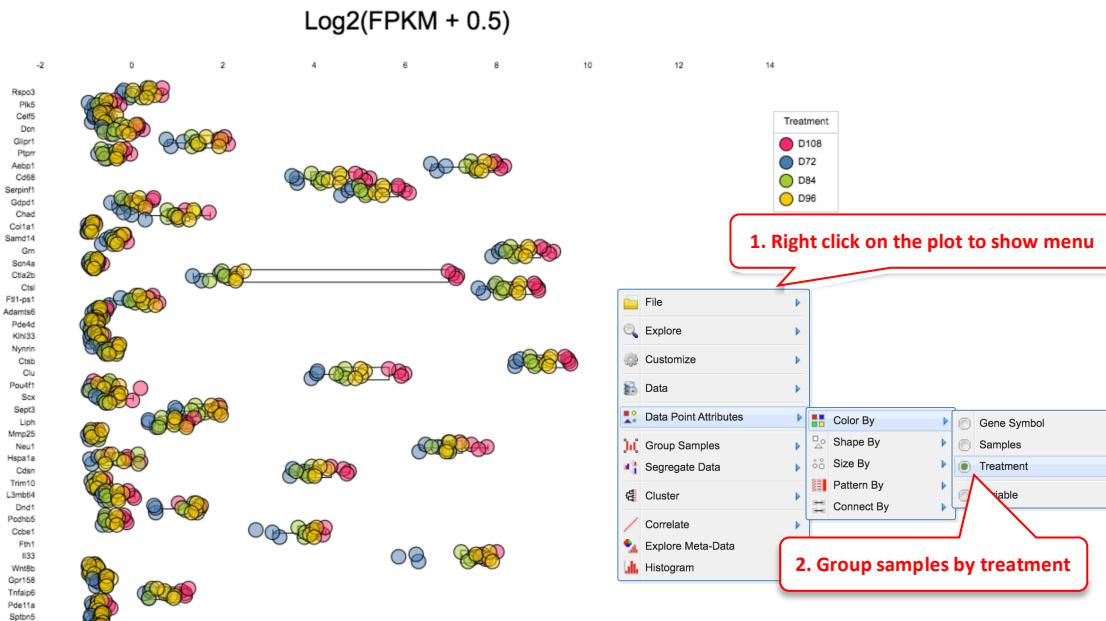
## Customize Gene Expression Plot

The boxplot is created using CanvasXpress (<https://canvasxpress.org>) plug-in, and sample grouping and coloring can be customized by the user. In the example below, we show how data points are colored.

## Summary of Data

- 169 genes found.
- 15 samples found: DK1B1D108, DK1B2D108, DK1B3D108, DK1B4D108, DK1B1D72, DK1B2D72, DK1B3D72, DK1B1D84, DK1B2D84, DK1B3D84, DK1B4D84, DK1B1D96, DK1B2D96, DK1B3D96, DK1B4D96

Download: [Raw Data File](#)



#### 4.1.2 View Gene Expression in Heatmap

Heatmap can be useful to visualize gene profiles from multiple samples. It can also provide information about how genes and samples cluster.

The screenshot shows the CHOMics interface for creating a heatmap. A red box labeled "1. Start here" points to the top navigation bar. Another red box labeled "2. Enter or Load Saved Genes" points to the "Genes" input field containing gene names like Admrt6, Admrt7, Adh1, Aebp1, Af2, and Ampd3. A third red box labeled "3. Enter or Load Saved Samples" points to the "Samples" input field containing sample names like DK1B3D84, DK1B4D84, DK1B1D96, DK1B2D96, DK1B3D96, and DK1B4D96. A fourth red box labeled "4. (Optional) Choose attributes overlaying on heatmap" points to the "Sample Attributes" section, which includes checkboxes for various sample characteristics. A fifth red box labeled "5. (Optional) Change data transformation, clustering options" points to the "Advanced Options" button. A sixth red box labeled "6. Plot" points to the "Plot" button at the bottom left.

You can enter genes and samples in the box, or load pre-saved genes and samples quickly from your collection. Be default, we will log2 transform the gene expression data, perform scaling of the data across samples for each gene, and limit the scaled value to -3 to 3 before displaying the data in heatmap. This works well in most situations. However, advanced users can change the options. For example, if you want to keep the order of samples as you entered, just uncheck “Cluster Samples”.

#### Advanced Options

##### Data Options

Enable Log2 Transform

Value To Be Added For Log Transformation:

Enable Z-Score Transform

Enable Upper Limit

Enable Lower Limit

Cluster Genes

Cluster Samples

##### Display Options

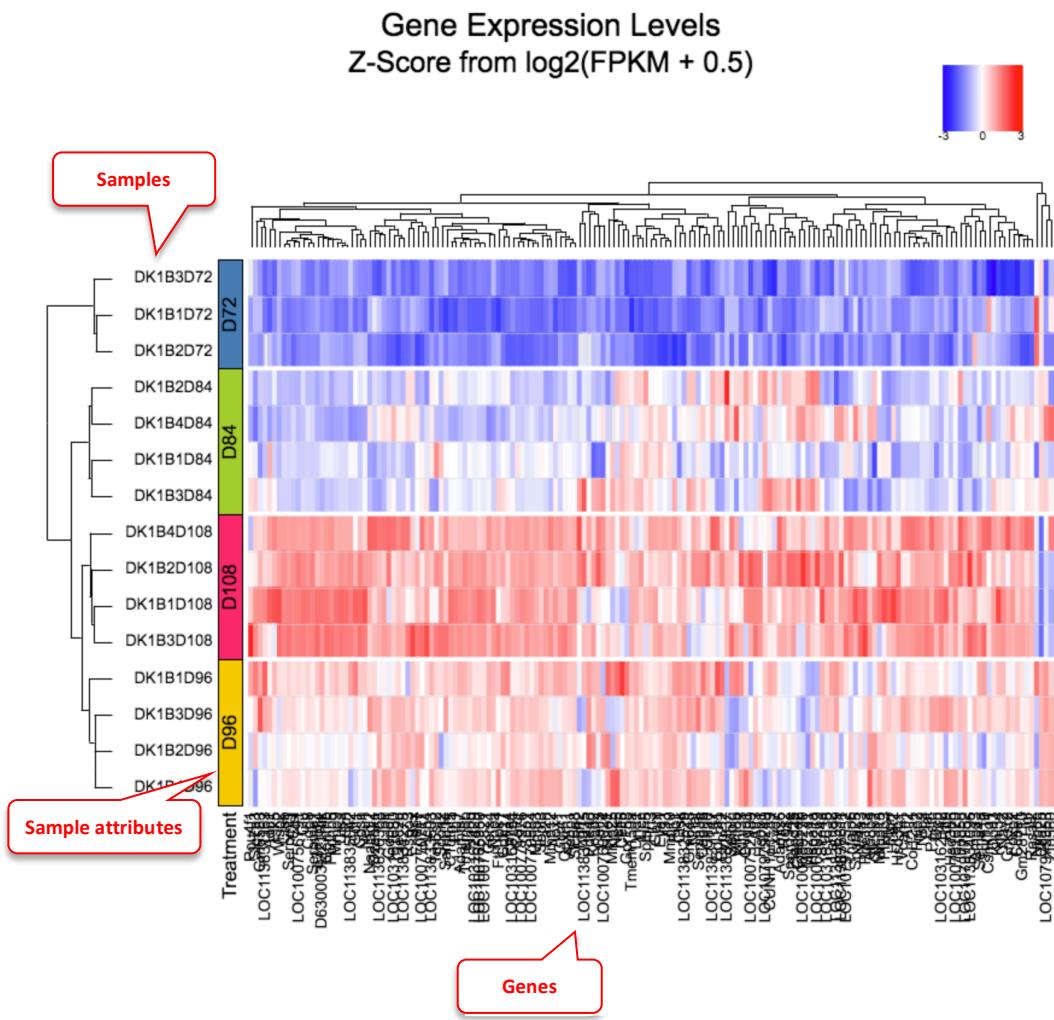
Overlay Samples

Display Gene Names

Display Sample IDs

The heatmap is rendered by CanvassXpress. You can change the plot size if needed.

Download: [Raw Data File](#) [Heatmap Data File](#) [Bookmark URL](#)



In the example heatmap, we entered a few significantly differential expressed genes between time D72 vs time D108. From heatmap clustering, we can see that the samples are clearly clustered by time points with increase of expression on most of genes along with time.

#### 4.1.3 Multi-omics Expression View

Besides the plotting of transcriptomics data, CHOmics also enables the visualization of other types of omics data such as proteomics, and the comparison across omics.

Here is an example of comparing gene expression (transcriptomics) and protein expression (proteomics) of gene CTSA at different time points, using the ‘Gene Expression Plot’ tool. By right clicking the plotting area, users can group the samples by different treatment time points while segregating the data by omics type(i.e, Samplesource).

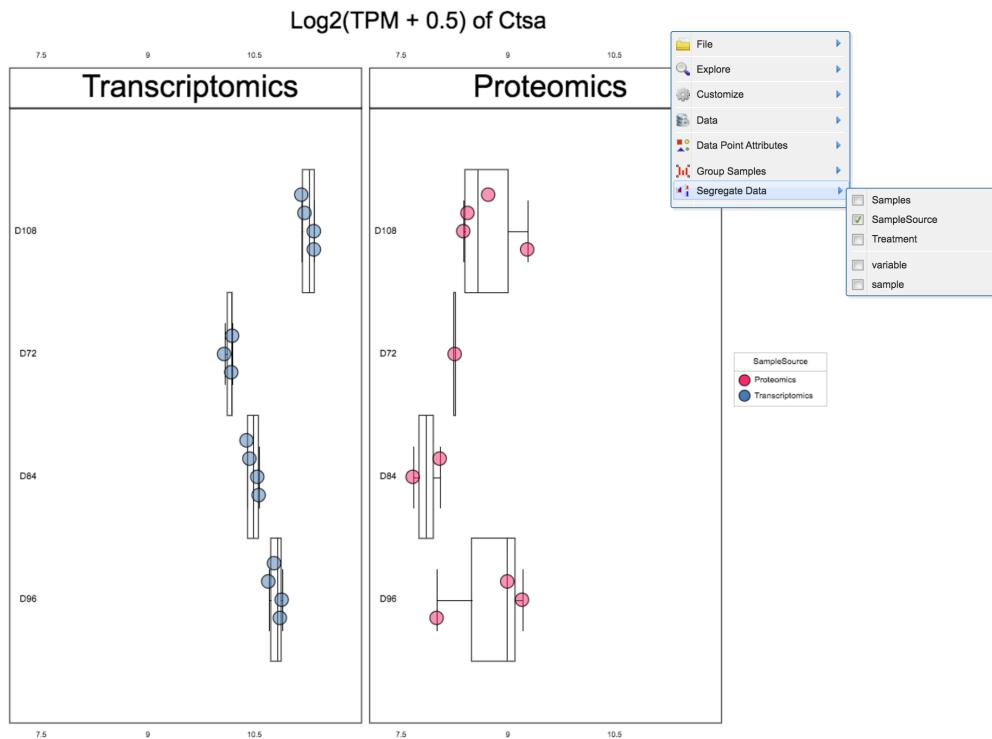
CHOMics (v1) Toolbox ▾ My Analyses ▾ Admin ▾ Projects (2) Comparisons (16) Samples (56) Hello, Demo User Sign Out

[Plot](#) [Reset All](#)

### Summary of Data

- 1 gene found: Ctsa
- 25 samples found: DK1B1D108, DK1B2D108, DK1B3D108, DK1B4D108, DK1B1D72, DK1B2D72, DK1B3D72, DK1B4D72, DK1B1D84, DK1B2D84, DK1B3D84, DK1B4D84, DK1B1D96, DK1B2D96, DK1B3D96, DK1B4D96, P\_DK1-B1-D108, P\_DK1-B2-D108, P\_DK1-B3-D108, P\_DK1-B4-D108, P\_DK1-B3-D72, P\_DK1-B4-D72, P\_DK1-B3-D84, P\_DK1-B4-D84, P\_DK1-B2-D96, P\_DK1-B3-D96, P\_DK1-B4-D96

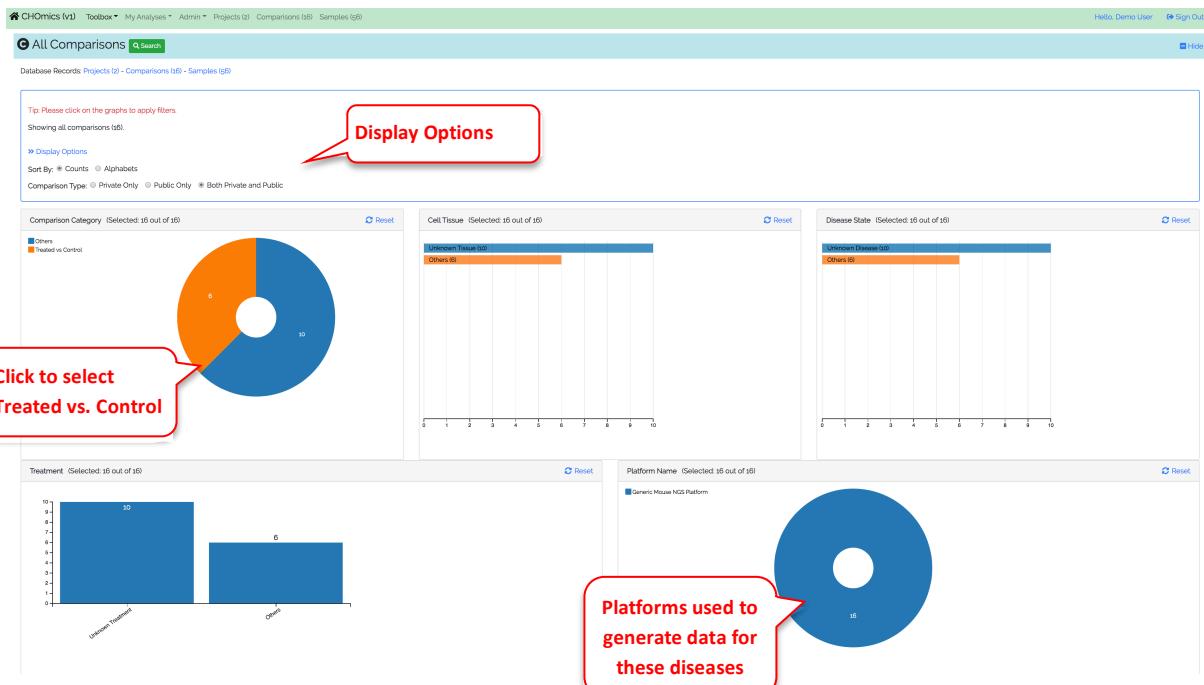
Download: [Raw Data File](#)



## 4.2 Visualize Comparison Data

### 4.2.1 Dashboard View of Comparison

The dashboard shows a summary of all the comparisons.



The above dashboard shows the comparisons from different Categories, Cell Type, Disease State, Treatment, Platform, etc. Below the dashboard, there is also a table listing all the comparisons.

In addition, users can set Dashboard Preference to change how the comparison summary is displayed.

## Dashboard Options

### Comparison Category:

Show Top 15 Only

### Cell Tissue:

Hide "Unknown Tissue"  Hide "Others" Type  Show Top 15 Only

### Disease State:

Hide "Unknown Disease"  Hide "normal control"  Hide "Others" Type  Show Top 15 Only

### Treatment:

Hide "Unknown Treatment"  Hide "Others" Type  Show Top 15 Only

### Platform Name:

Hide "Generic" Types  Show Top 15 Only

**Save** **Close**

#### 4.2.2 Bubble Plot

Bubble plot is another useful demonstration of gene or gene set in comparisons. For each gene, you can view all the available comparisons in a bubble chart.

CHomics (v1) Toolbox ▾ My Analyses ▾ Admin ▾ Projects (1) Comparisons (12) Samples (30)

**Start here**

## Bubble Plot

» Multiple genes .vs. multiple comparisons

Gene Name: **Tgm2** Gene Symbol

Please enter the gene name, e.g., BRWD1-IT2

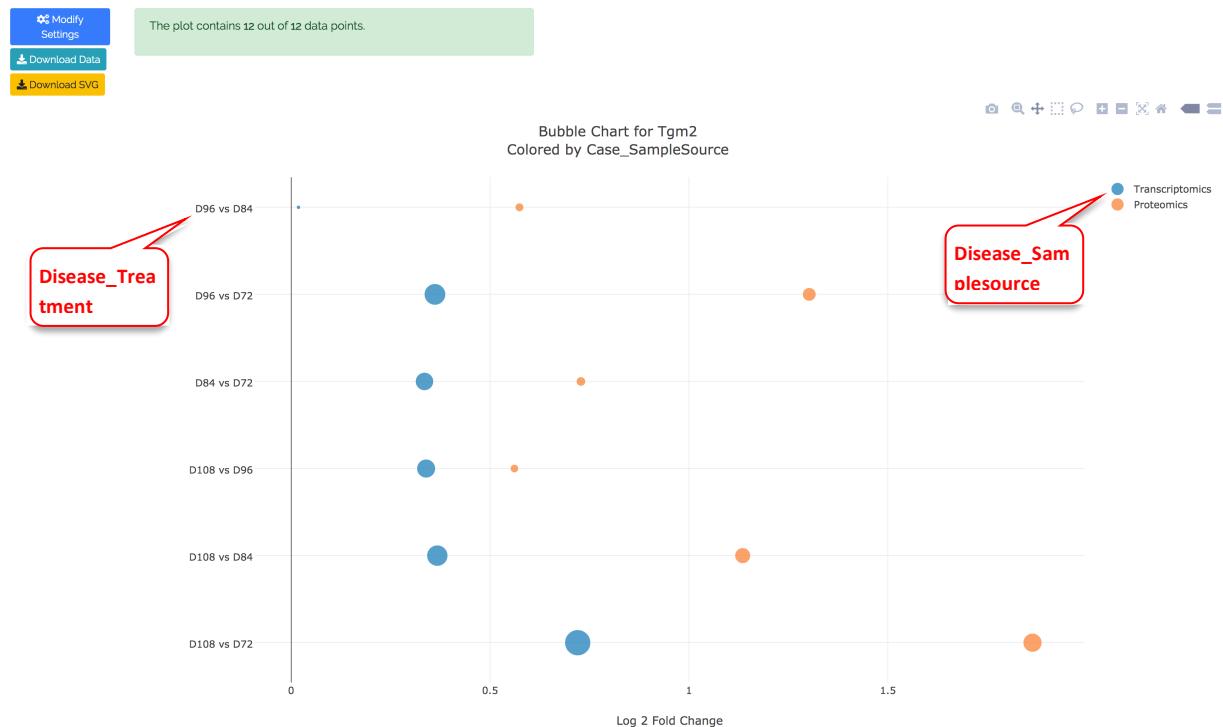
Y-axis Field: Case\_Treatment

Coloring Field: **Case\_SampleSource**

Comparison Type: All Comparisons

**> Next Step**

The default settings work for most users. After clicking the Next Step button, you will see a plot like:

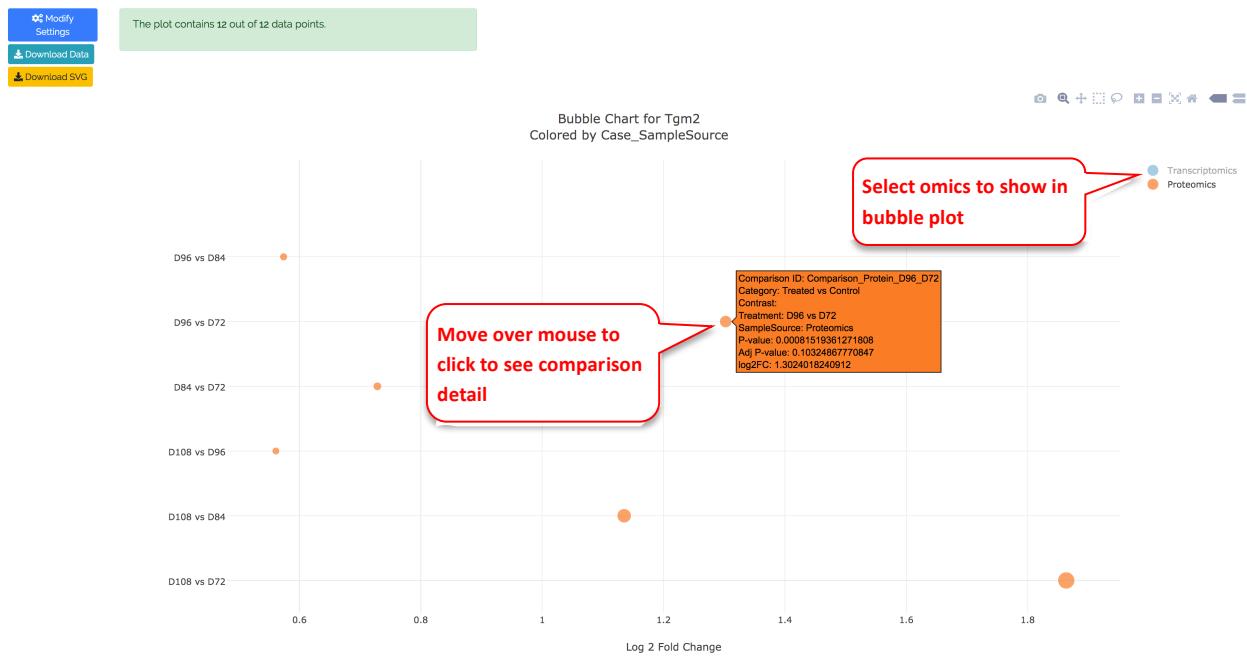


In the bubble plot, the X-axis shows log2 Fold Change of the comparison, the Y-axis shows 'Case\_treatment'. Each dot represents the comparison result of this gene from one comparison. The color of the dot represent 'Case\_Samplesource' (i.e, here we set as omics type), and the size of the dot represent significance (-log10(FDR), larger is more significant).

The user can click and unclick the color legend at right to select or deselect omics types. When mouse over a dot, more details are shown. And the user can also click the dot to link to other graphs.

The tool bars at top right corner allows the user to zoom and pan the graph.

The screenshot below shows the same bubble chart after selecting one omics type (i.e,transcriptomics), and zoom into a portion of the chart.



## Data Filter and Advanced Settings in Bubble Plot

In addition, advanced users can change settings by click "Modify Settings Button". For example, the user may want to show a selected list of diseases. After clicking Customize in Case\_Treatment, user can select which treatments to display in the pop-up window.

Marker Area  P-Value  Adjusted P-Value

Y-axis Setting  Case\_Treatment  
 Show Top 10  Show Top 20  Show All  Customize

Coloring Setting  Case\_SampleSource  
 Show Top 10  Show Top 20  Show All  Customize

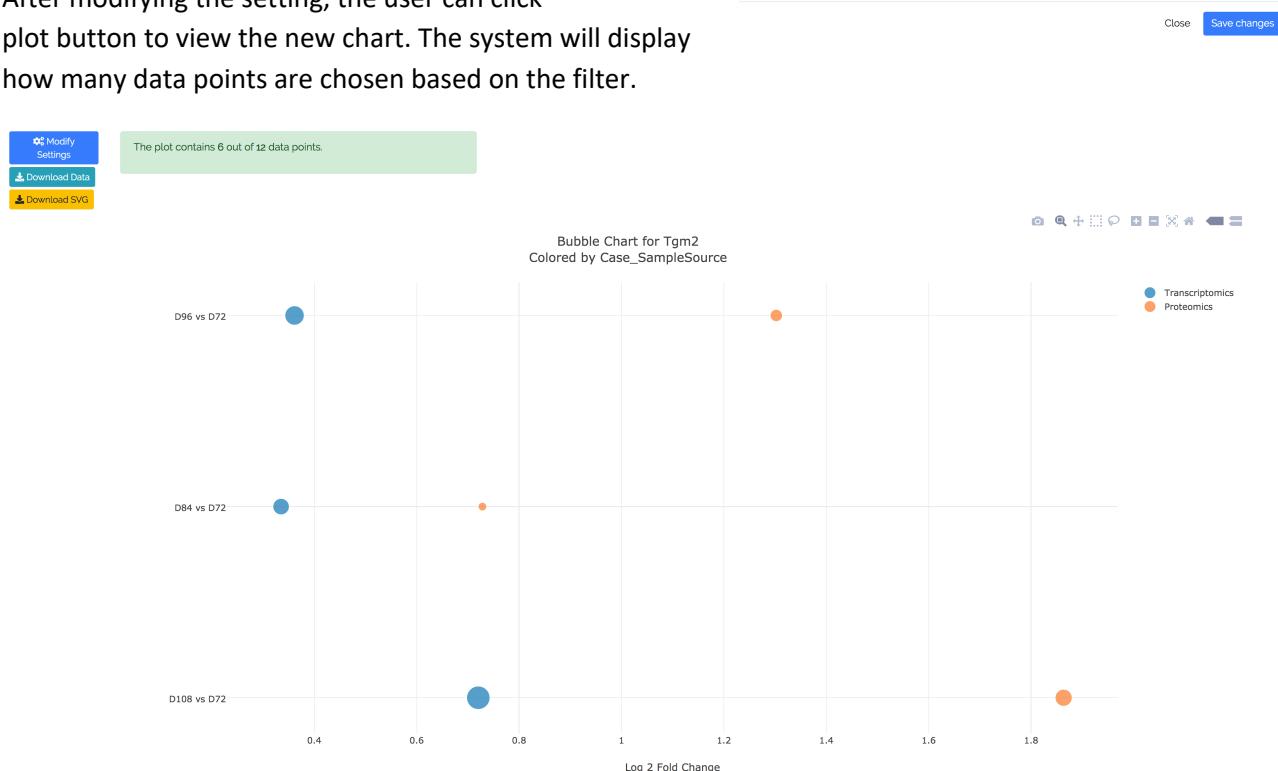
**Plot**

The plot contains 12 out of 12 data points.

Name	Occurrence
D84 vs D72	2
D96 vs D72	2
D108 vs D72	2
D96 vs D84	2
D108 vs D84	2
D108 vs Dg6	2

**Modify Settings** **Download Data** **Download SVG**

After modifying the setting, the user can click plot button to view the new chart. The system will display how many data points are chosen based on the filter.



## Bubble Plot of Multiple Genes and Multiple Comparisons

It can be useful to look at a set of genes (e.g. all differentially expressed genes, or genes from a certain pathways) in a set of related comparisons (e.g. all from the same disease).

To view this type of bubble plot, select the link for Multiple Genes vs. multiple comparisons.

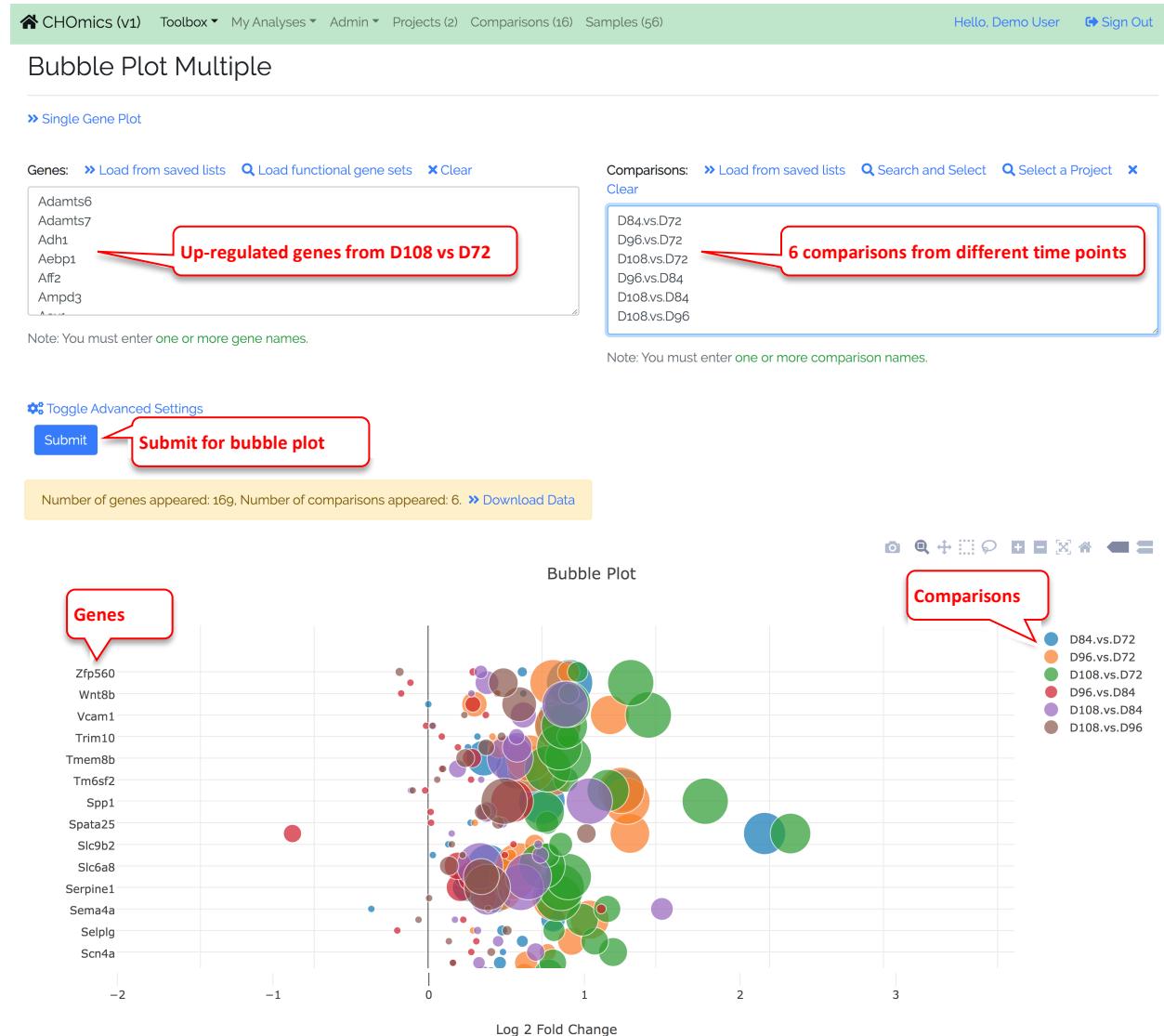
The screenshot shows the CHOMics interface with a green header bar containing links for Home, Toolbox, My Analyses, Admin, Projects (2), Comparisons (16), and Samples (56). A red box highlights the 'Start here' button in the top right corner of the header. Below the header, the title 'Bubble Plot' is displayed. A blue arrow points to the '» Multiple genes .vs. multiple comparisons' link. Another red box highlights the 'Bubble plot of gene set' link. On the right side, there are input fields: 'Gene Name:' with 'jak3' typed in, 'Y-axis Field:' with 'Case\_DiseaseState', 'Coloring Field:' with 'Case\_SampleSource', and 'Comparison Type:' with 'All Comparisons'. A green 'Next Step' button is located below these fields.

In the Genes and Comparisons Bubble plot window, you can now enter the symbols of the genes, and the comparison names. However, it is much easier to use the saved genes and saved comparisons features, or other tools from the system to quickly get a get set. Please see below for details.

The screenshot shows the 'Bubble Plot Multiple' configuration page. At the top, there are two sections: 'Genes:' and 'Comparisons:'. Both sections have 'Load from saved lists' buttons (highlighted with red boxes) and 'Search and Select' buttons. Below each section is a note: 'Note: You must enter one or more gene names.' and 'Note: You must enter one or more comparison names.' respectively. At the bottom left is a 'Toggle Advanced Settings' link and a 'submit' button. A red arrow points to the 'Load saved gene list' button in the 'Genes:' section, and another red arrow points to the 'Load saved comparisons' button in the 'Comparisons:' section.

In the example below, we use dashboard to select 6 comparisons that are for different time points in CHO cell lines. We save the comparisons and load in the bubble plot tool. For gene list, we get the up-regulated genes from comparison D72 vs D108, and paste into the gene names fields.

In the bubble plot, the gene symbols are listed in Y-axis. The X-axis represents logFC, and color of the bubble represents comparison; the size of the bubble represents the significance.

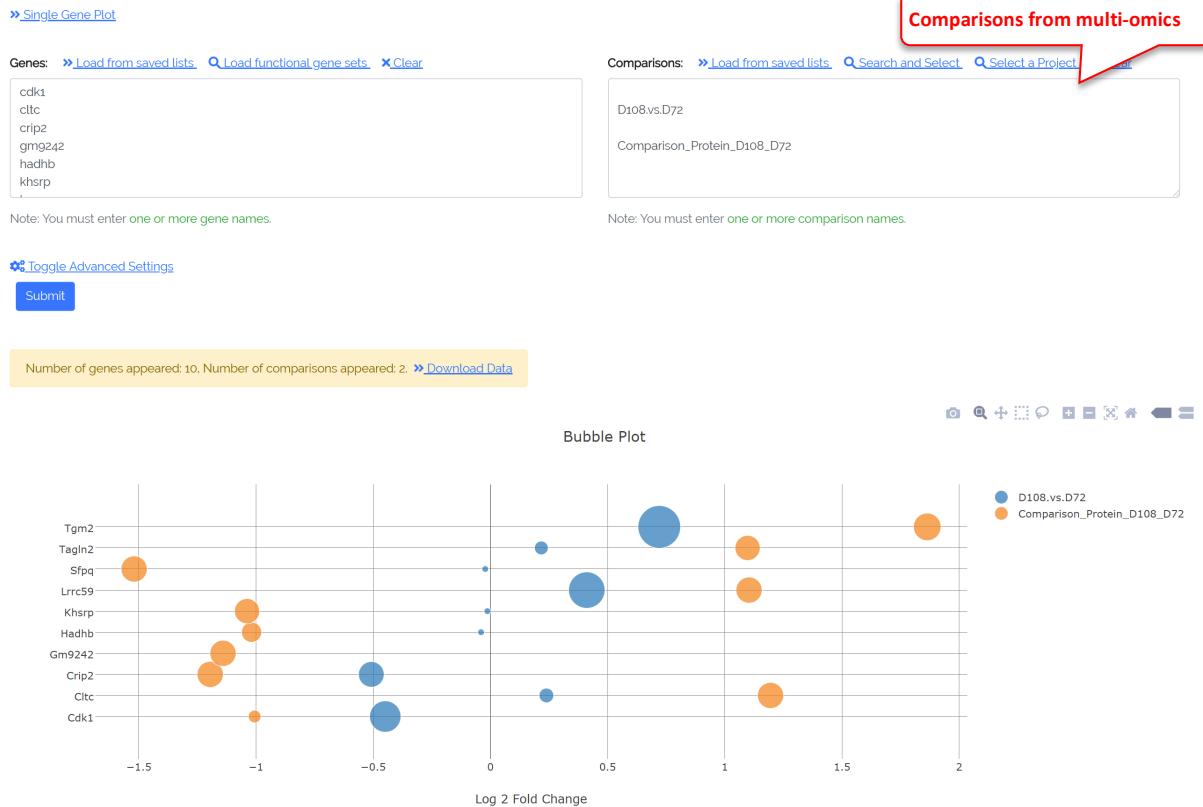


In the legend, the color keys for comparisons are shown. You can click the color key in the legend to hide/show comparisons. The size of the color dot in the legend correlates to the largest bubble for that comparison, which is the most significant gene with the smallest FDR.

### Bubble Plot of Multiple Omics data

Similar to the bubble plot of multiple genes across multiple comparisons, users can further compare the genes on the comparisons from different omics data.

## Bubble Plot Multiple



### 4.2.3 Get significant genes from comparisons

Another way to get a gene set to visualize in the genes/comparisons bubble plot is to filter for significantly changed genes. To do this, first select a few comparisons from the dash board, and click the "View Significantly Changed Genes" button.

CHOrnics (v1) Toolbox [My Analyses](#) [Admin](#) [Projects](#) [Comparisons](#) (12) [Samples](#) (30) Hello, Demo User [Sign Out](#)

[My Private Projects](#) [Show](#)

[All Comparisons](#) [Search](#) **Significant changed genes** [Show](#)

[List of Comparisons](#) [Hide](#)

[Save Comparison List](#) [Bubble Plot](#) [Significantly Changed Genes](#) [Pathway Heatmap](#) [Meta Analysis](#) [Export Comparison Data](#) [WikiPathways](#) [Reactome Pathways](#) [KEGG Pathways](#)

[Gene Expression Plot](#) [Heatmap](#) [Correlation Tool](#) [PCA Analysis](#) [Export Expression Data](#)

Column Visibility [Copy](#) [CSV](#) Show 10 entries

Name	ComparisonCategory	Case_Tissue	Case_DiseaseState	Case_Treatment	PlatformName
Dg6vs.D84	Others	Unknown Tissue	Unknown Disease	Dg6 vs D84	Generic Mouse NGS Platform
Dg6vs.D72	Others	Unknown Tissue	Unknown Disease	Dg6 vs D72	Generic Mouse NGS Platform
D84vs.D72	Others	Others	Others	D84 vs D72	Generic Mouse NGS Platform
D108vs.Dg6	Others	Unknown Tissue	Unknown Disease	D108 vs Dg6	Generic Mouse NGS Platform
D108vs.D84	Others	Unknown Tissue	Unknown Disease	D108 vs D84	Generic Mouse NGS Platform
<input checked="" type="checkbox"/> D108vs.D72	Others	Unknown Tissue	Unknown Disease	D108 vs D72	Generic Mouse NGS Platform
Comparison_Protein_Dg6.D84	Treated vs Control	Others	Others	Dg6 vs D84	Generic Mouse NGS Platform
Comparison_Protein_Dg6.D72	Treated vs Control	Others	Others	Dg6 vs D72	Generic Mouse NGS Platform
Comparison_Protein_D84.D72	Treated vs Control	Others	Others	D84 vs D72	Generic Mouse NGS Platform
Comparison_Protein_D108.Dg6	Treated vs Control	Others	Others	D108 vs Dg6	Generic Mouse NGS Platform

Showing 1 to 10 of 12 entries

Previous 1 2 Next

Dashboard filter.

In table, select comparisons and view significantly changed genes.

In the significantly Changed Genes window, the comparisons from the previous page are already loaded. You can add or remove comparisons if needed.

Now select direction (up-, down-, or both), and use the logFC cutoff and FDR value to get a list of genes. Depending on the comparisons, sometimes you may need to adjust the logFC and FDR values to get a good list of genes. In general, for bubble plot, using <100 genes will make the graph easier to read.

Once you are happy with the gene list, you can save it. You can also export the list for later use.

CHOMics (v1) Toolbox My Analyses Admin Projects (1) Comparisons (12) Samples (30)

## Significantly Changed Genes

Comparisons: [Load from saved lists](#) [Search and Select](#) [Select a Project](#) [Clear](#)

List of selected comparisons

D108.vs.D72  
D108.vs.D84  
D108.vs.Dg6  
D84.vs.D72  
Dg6.vs.D72  
Dg6.vs.D84

Or, upload your comparison files: [Choose File](#) No file chosen [Demo Data](#)

Display Options:  Log2FC  PValue  FDR

Fold Change Cutoff:  Both Up- and Down-regulated

Statistic Cutoff:    Increase or decrease threshold to select genes

List Genes:  Common Genes from All Comparisons  Genes from Any Comparisons

[Submit](#) [Start Over](#)

[Save Comparison List](#) [Bubble Plot](#) [Pathway Heatmap](#) [Meta Analysis](#) [Export Comparison Data](#) [WikiPathways](#) [Reactome Pathways](#) [KEGG Pathways](#)

[Save Gene List](#) [Gene Expression Plot](#) [Heatmap](#) [Correlation Tool](#) [PCA Analysis](#) [Export Expression Data](#)

Column visibility Copy CSV Show 10 entries

Showing 1 to 10 of 15 entries

	GeneName	Description	D108.vs.D72 - Log2FC	D108.vs.D72 - PValue	D108.vs.D72 - FDR	D108.vs.D84 - Log2FC	D108.vs.D84 - PValue	D108.vs.D84 - FDR	D108.vs.Dg6 - Log2FC
	Cd68	CD68 antigen	15075	0.0000	0.0000	0.9910	0.0000	0.0001	0.5591
	Clu	clusterin	18516	0.0000	0.0000	12528	0.0000	0.0000	0.8301
	Ctsb	cathepsin B	11143	0.0000	0.0000	0.7742	0.0000	0.0000	0.4140

## View Significantly Changed Genes in Bubble Plot

Back to the bubble plot, you can load the saved comparisons and saved genes and view the plot.

The screenshot shows a table of significantly changed genes. At the top, there are two buttons: 'Save to comparison list' (highlighted with a red box and arrow) and 'Save to gene list' (also highlighted with a red box and arrow). Below these buttons is a navigation bar with various links like 'Save Comparison List', 'Bubble Plot', etc. Underneath the navigation bar is a search/filter section with 'Column visibility', 'Copy', 'CSV', and 'Show 100 entries'. The main table has columns for GeneName, Description, Log2FC, D108 vs. D72 - PValue, D108 vs. D72 - FDR, D108 vs. D84 - Log2FC, D108 vs. D84 - PValue, D108 vs. D84 - FDR, D108 vs. D96 - Log2FC, D108 vs. D96 - PValue, D108 vs. D96 - FDR, D84 vs. D72 - Log2FC, and D84 vs. D72 - PValue. The table lists 15 genes, with most showing down-regulation (negative Log2FC values).

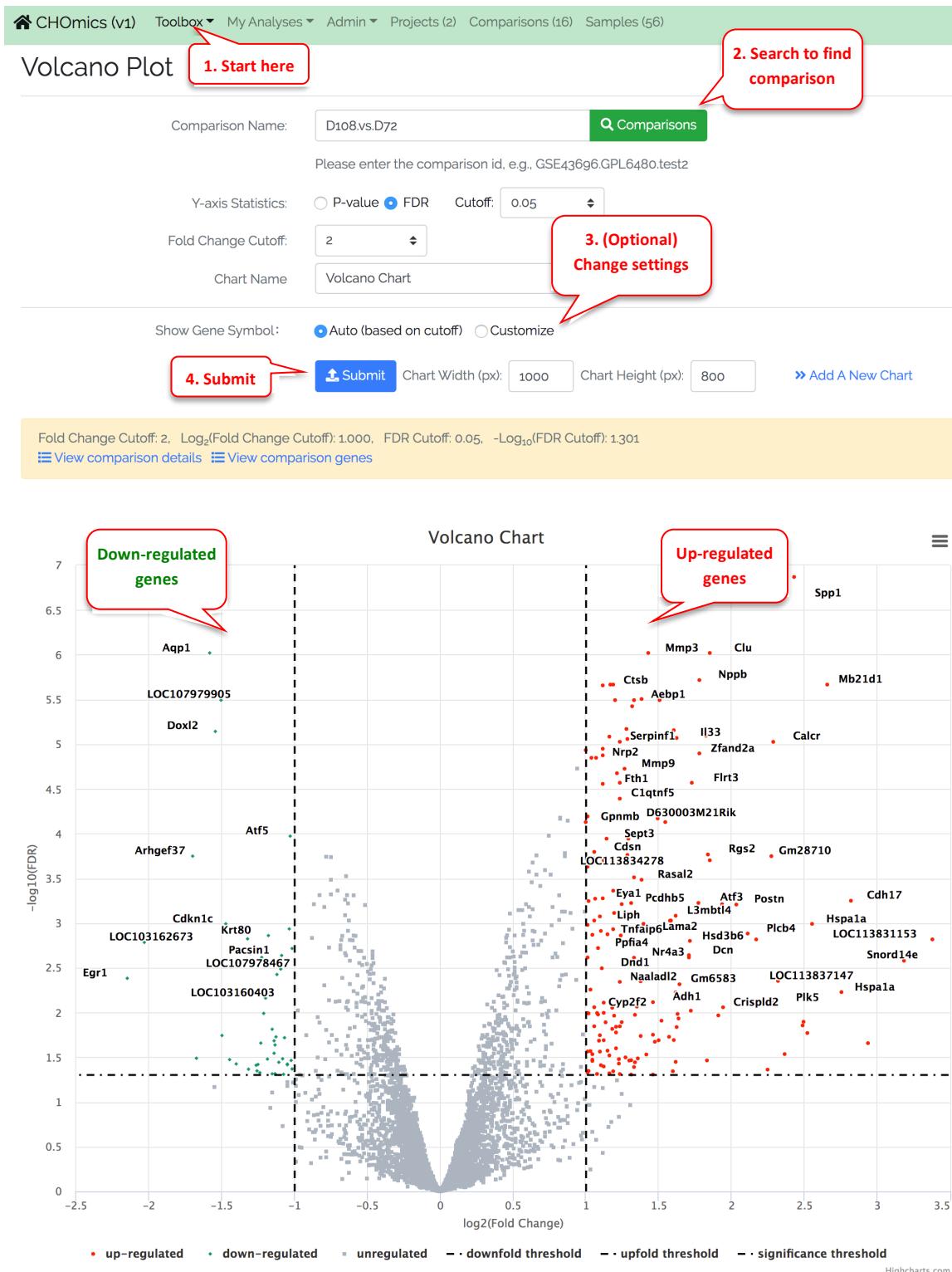
GeneName	Description	Log2FC	D108 vs. D72 - PValue	D108 vs. D72 - FDR	D108 vs. D84 - Log2FC	D108 vs. D84 - PValue	D108 vs. D84 - FDR	D108 vs. D96 - Log2FC	D108 vs. D96 - PValue	D108 vs. D96 - FDR	D84 vs. D72 - Log2FC	D84 vs. D72 - PValue
Cd68	CD68 antigen	15075	0.0000	0.0000	0.9910	0.0000	0.0001	0.5591	0.0000	0.0061	0.4969	0.0002
Clu	clusterin	18516	0.0000	0.0000	12528	0.0000	0.0000	0.8301	0.0000	0.0003	0.5806	0.0000
Ctsb	cathepsin B	11143	0.0000	0.0000	0.7742	0.0000	0.0000	0.4140	0.0000	0.0006	0.3224	0.0005
Ctsl	cathepsin L	11178	0.0000	0.0000	0.7777	0.0000	0.0000	0.3962	0.0000	0.0008	0.3222	0.0039
Doxi2	diamine oxidase-like protein 2	-15420	0.0000	0.0000	-0.9831	0.0000	0.0002	-0.4795	0.0002	0.0186	-0.5774	0.0000
Grn	granulin	10717	0.0000	0.0000	0.7742	0.0000	0.0001	0.4449	0.0000	0.0049	0.2799	0.0036
LOC100771976		11626	0.0000	0.0000	0.6455	0.0000	0.0002	0.2467	0.0009	0.0364	0.4993	0.0001
LOC103162429		12001	0.0000	0.0000	0.5994	0.0000	0.0002	0.2880	0.0003	0.0204	0.5830	0.0000
Mmp19	matrix metallopeptidase 19	0.9750	0.0000	0.0000	0.5372	0.0000	0.0001	0.1829	0.0014	0.0440	0.4200	0.0000
Mmp3	matrix metallopeptidase 3	14305	0.0000	0.0000	0.7182	0.0000	0.0001	0.2204	0.0013	0.0416	0.6943	0.0000
Nppb	natriuretic peptide type B	17812	0.0000	0.0000	0.9140	0.0000	0.0002	0.4367	0.0001	0.0100	0.8486	0.0000
Plat	plasminogen activator, tissue	10157	0.0000	0.0001	0.6146	0.0000	0.0005	0.2990	0.0005	0.0255	0.3822	0.0015

In the example below, it can be seen that most significant genes come from down-regulated direction from the first four comparisons.

The screenshot shows the 'Bubble Plot Multiple' page. On the left, under 'Genes', there is a list of genes: Cd68, Clu, Ctsb, Ctsl, Doxi2, Grn, Mmp19, LOC103162429, LOC100771976, Nppb, Mmp3, Mmp19, Plat, Nppb, Mmp3, Mmp19, and Grn. A red box highlights the 'Load saved gene list' button. On the right, under 'Comparisons', there is a list of comparisons: D108 vs. D72, D108 vs. D64, D108 vs. D96, D84 vs. D72, D96 vs. D72, and D96 vs. D84. A red box highlights the 'Load saved comparison list' button. Below these lists, there is a note: 'Note: You must enter one or more gene names.' and 'Note: You must enter one or more comparison names.' At the bottom, there is a 'Submit' button and a message: 'Number of genes appeared: 15, Number of comparisons appeared: 6. >> Download Data'. The main area shows a bubble plot with 'Log 2 Fold Change' on the x-axis (ranging from -1.5 to 2.5) and 'PValue' on the y-axis. The bubbles are colored according to the comparison they belong to, with sizes indicating their significance. A legend on the right side maps colors to comparisons: blue for D108 vs. D72, orange for D108 vs. D84, green for D108 vs. D96, red for D84 vs. D72, purple for D96 vs. D72, and brown for D96 vs. D84.

#### 4.2.4 Volcano Plot

Volcano plot is useful to view a top level summary of how many genes are significantly up- or down-regulated in a comparison.



You can use mouse to drag over an area to zoom in.

Mouse over a point will show the gene details. Click the data point will show you links to other graphs.

## View Multiple Volcano Plots Together

Users can also show multiple comparisons side-by-side. If needed, the user can also highlight the same group of genes across the volcano plots.

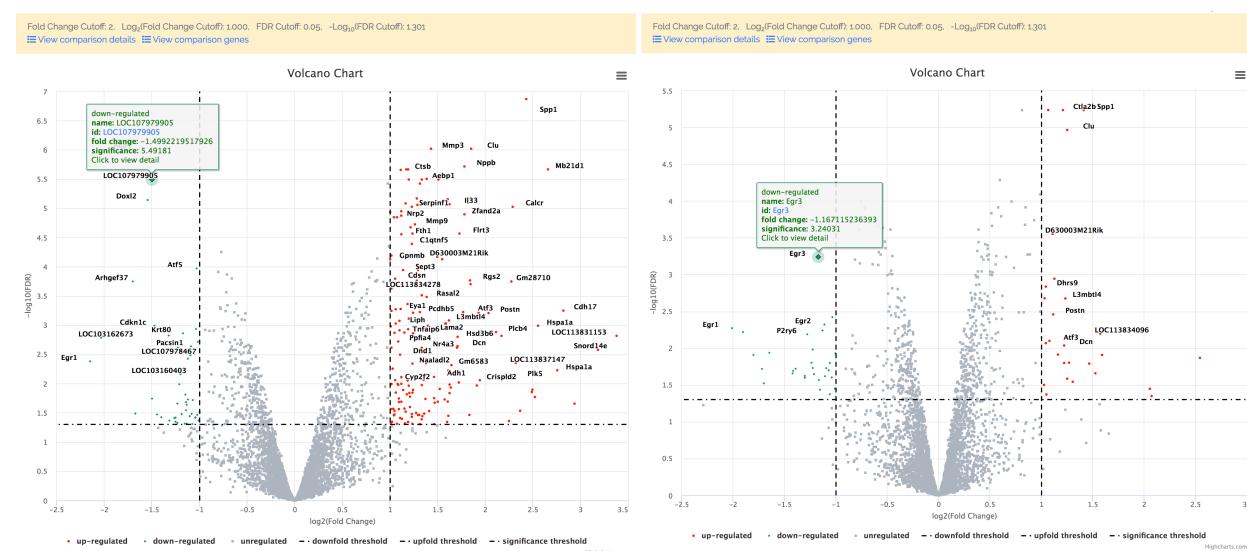
Screenshot of the CHOMics interface showing two volcano plots side-by-side. The top navigation bar includes: Home, CHOMics (v1), Toolbox, My Analyses, Admin, Projects (2), Comparisons (16), Samples (56).

The first comparison (top plot): Comparison Name: D108.vs.D72, Y-axis Statistics: FDR Cutoff: 0.05, Fold Change Cutoff: 2, Chart Name: Volcano Chart. A red callout box labeled "The first comparison" points to the comparison name.

The second comparison (middle plot): Comparison ID: D108.vs.D84, Y-axis Statistics: FDR, Chart Name: Volcano Chart. A red callout box labeled "The second comparison" points to the comparison name.

Customization options (bottom): Show Gene Symbol: Auto (based on cutoff) or Customize, Submit button, Chart Width (px): 1000, Chart Height (px): 800, Add A New Chart button. A red callout box labeled "Use 'Customize' to highlight genes entered in the text box" points to the "Customize" option. Another red callout box labeled "Add more comparisons if needed" points to the "Add A New Chart" button.

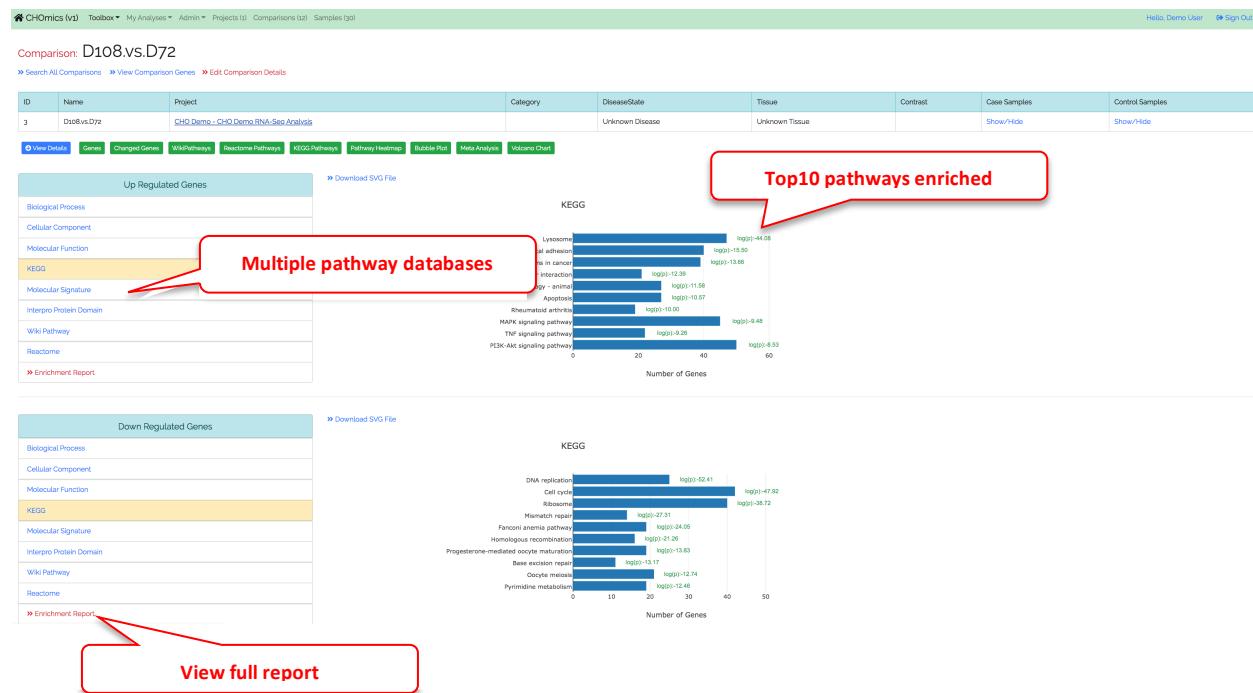
The resulting volcano plots are shown as below. Selected genes are shown as orange dots.



## 4.3 Visualize functional pathway

### 4.3.1 Enrichment from Up and Down Regulated Genes

When you view details of a comparison, the functional enrichment results are shown. Briefly, for each comparison, we generated the up- and down- regulate gene lists, and use these lists to compare with all genes in the genome to identify functions that are significantly enriched.



In the example above, this comparison is between D108 vs D72, and the top up-regulated biological processes are response to virus, immune effector process.

Click the left menu will switch the bar charts for different categories (Gene Ontology, KEGG, Molecular signature, Protein domain etc).

The bar charts here show the top 10 categories. To view complete results, click the Enrichment Report.

Gene Ontology Enrichment Results

Text file version of complete results (e.g. open with Excel)

[Download enrichment results file](#)

Enriched Categories

Functional Terms

Top10 pathways enriched

GO Tree	TermID	Term	Enrichment	logP	Genes in Term	Target Genes in Term	Fraction of Targets in Term	Total Target Genes	Total Genes	Actions
MSigDB	LEE_BMP2_TARGETS_DN	LEE_BMP2_TARGETS_DN	3.9176322471508e-31	-70.0146504296904	807	153	0.13871260199458	1103	15720	<a href="#">Show Genes</a>
Gene Ontology	GO:0044424	intracellular part	3.36395799347635e-25	-56.3517766476754	13240	933	0.76980198019802	1212	20830	<a href="#">Show Genes</a>
Gene Ontology	GO:0005622	intracellular	6.67887260173594e-25	-55.66667781237414	13317	936	0.772277227722772	1212	20830	<a href="#">Show Genes</a>
Gene Ontology	GO:0005488	binding	9.43954474027105e-25	-55.3197196725229	12660	892	0.758503401360544	1176	20347	<a href="#">Show Genes</a>
Gene Ontology	GO:0005730	nucleolus	5.912051254420940e-24	-53.4860493786918	813	125	0.103136313531353	1212	20830	<a href="#">Show Genes</a>
Gene Ontology	GO:004464	cell part	5.2835457378631e-23	-51.2952383306056	15460	1037	0.855610561056106	1212	20830	<a href="#">Show Genes</a>
Gene Ontology	GO:0005623	cell	6.33897499094402e-23	-51.1127400568371	15465	1037	0.855610561056106	1212	20830	<a href="#">Show Genes</a>
MSigDB	GSE6674_ANTI_IGM_VS_CPG_STIM_BCELL_DN	GSE6674_ANTI_IGM_VS_CPG_STIM_BCELL_DN	153486643903669e-21	-47.9258448890668	186	56	0.0507706255666364	1103	15720	<a href="#">Show Genes</a>
MSigDB	KRIGE_RESPONSE_TO_TOSEDOSTAT_24HR_DN	KRIGE_RESPONSE_TO_TOSEDOSTAT_24HR_DN	161445682388097e-21	-47.8752883847299	865	140	0.126926563916591	1103	15720	<a href="#">Show Genes</a>
Gene Ontology	GO:0043229	intracellular organelle	5.91195034341862e-21	-46.577311685508	11590	829	0.683993399339934	1212	20830	<a href="#">Show Genes</a>

Showing 1 to 10 of 1001 entries

Previous 1 2 3 4 5 - 101 Next

In the enrichment report, the full list of functional terms are shown by order of p-value.

#### 4.3.2 View Changed Genes from a Functional Term in Volcano Plot

From the bar chart, click a functional term, and you have the option to view these genes in a volcano plot.



Once you click the link in the popup window, volcano plot will be generated for the comparison with the changed genes from the selected term highlighted.

## Volcano Plot

Comparison of interest

[Back to Comparison Details](#)

Comparison Name: D10B.vs.D72

[Comparisons](#)

Please enter the comparison id, e.g., GSE43696.GPL6480.test2

Y-axis Statistics:

P-value  FDR

Cutoff:

0.05

Filtering genes from selected pathway

Fold Change Cutoff:

2

Chart Name: Volcano Chart

Show Gene Symbol:

Auto (based on cutoff)  Customize

Genes: [» Load from saved lists](#) [Load functional gene sets](#) [Clear](#)

Cd68  
Cln5  
Ctns  
Ctsa  
Ctsb  
Ctsd  
Ctsl  
Galc  
Gba  
Grb  
Arsa  
Hexb  
Psap  
Cd63  
Dnase2a  
Gnptab  
Naglu  
Hexb  
Lamp1  
Atp6vob  
Ctns  
Nagpa  
Mifsd8  
Ctsk  
Ctsw

Gene list to be shown

[Submit](#)

Chart Width (px): 1000

Chart Height (px): 800

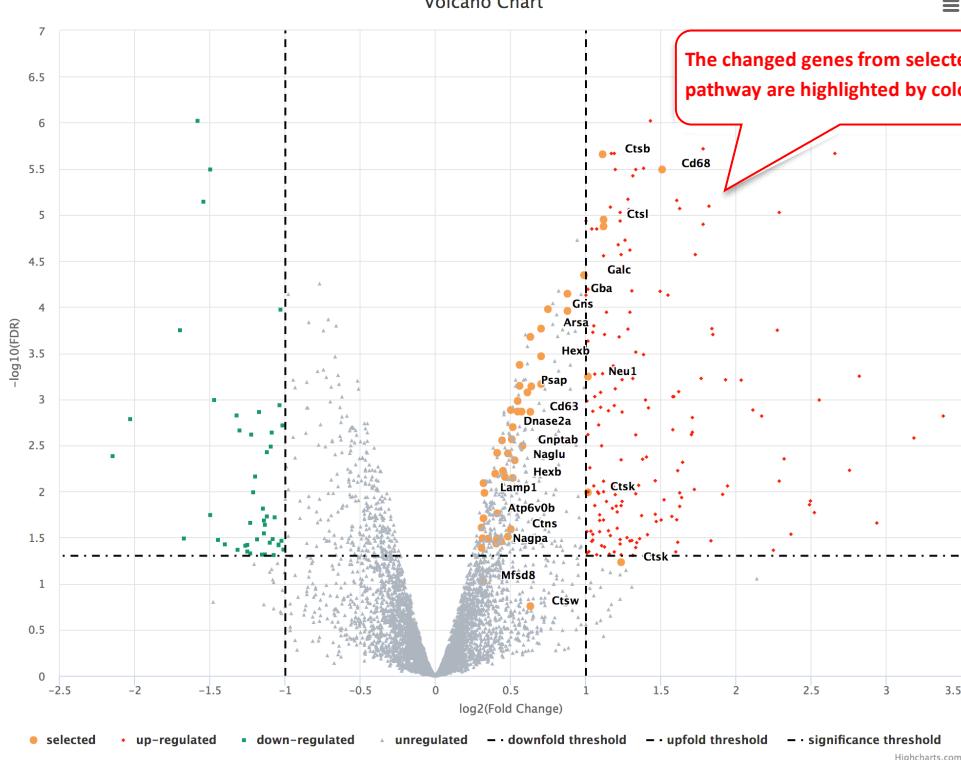
[» Add A New Chart](#)

Fold Change Cutoff: 2, Log<sub>2</sub>(Fold Change Cutoff): 1.000, FDR Cutoff: 0.05, -Log<sub>10</sub>(FDR Cutoff): 1.301

[View comparison details](#) [View comparison genes](#)

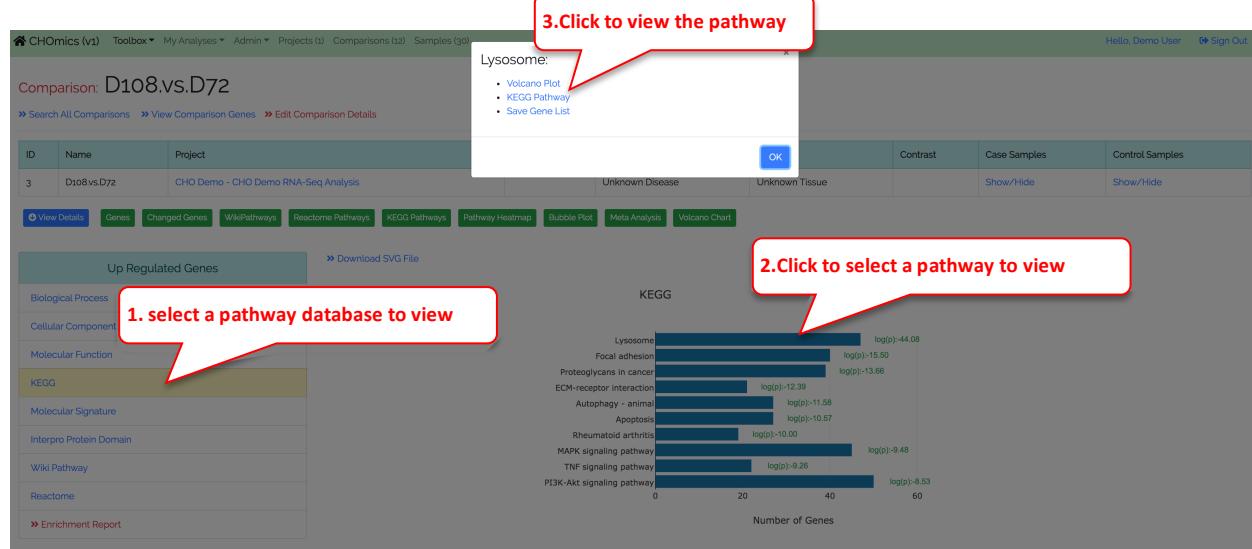
### Volcano Chart

The changed genes from selected pathway are highlighted by colors

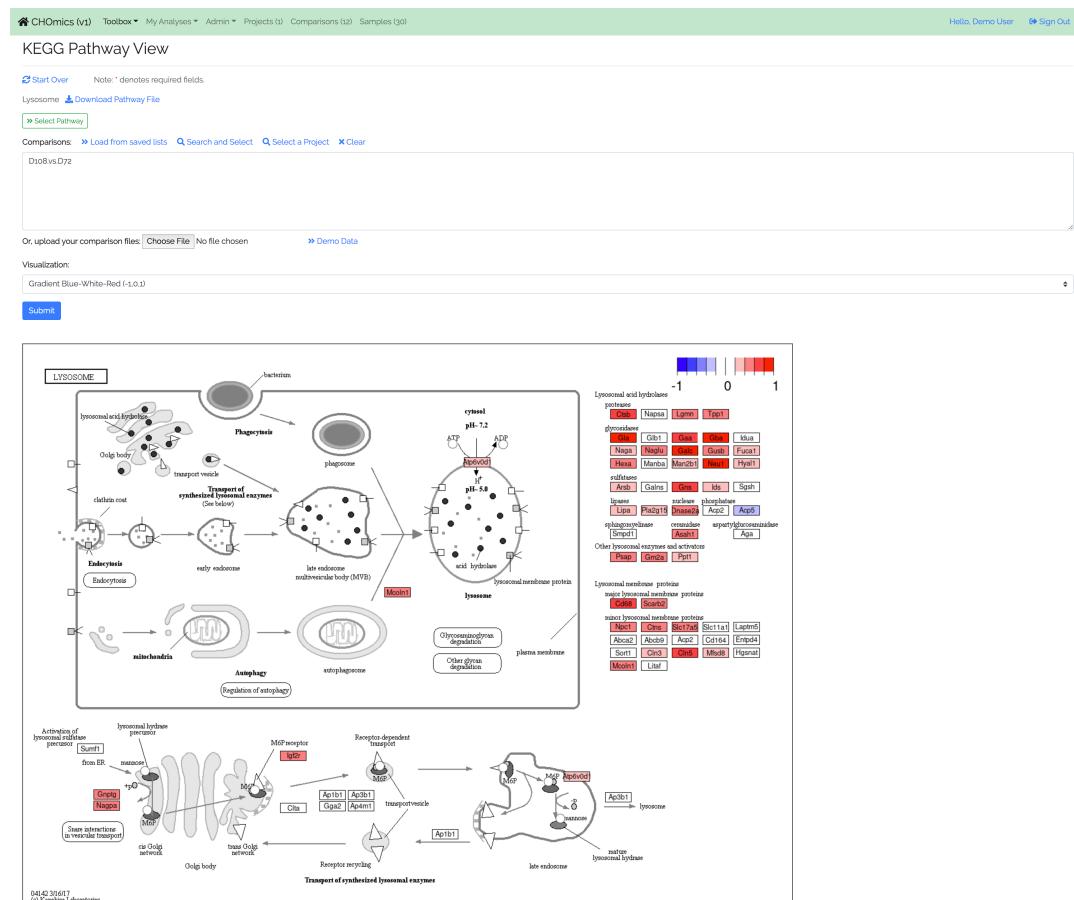


#### 4.3.3 View Enriched Pathways Directly from Comparison Details

From the bar chart, if you are viewing KEGG or wikipathway database, clicking the pathway name and you have the option to view pathway plot.



This will automatically open the pathway visualization page, and preload the pathway and comparison. Click submit to view the pathway.

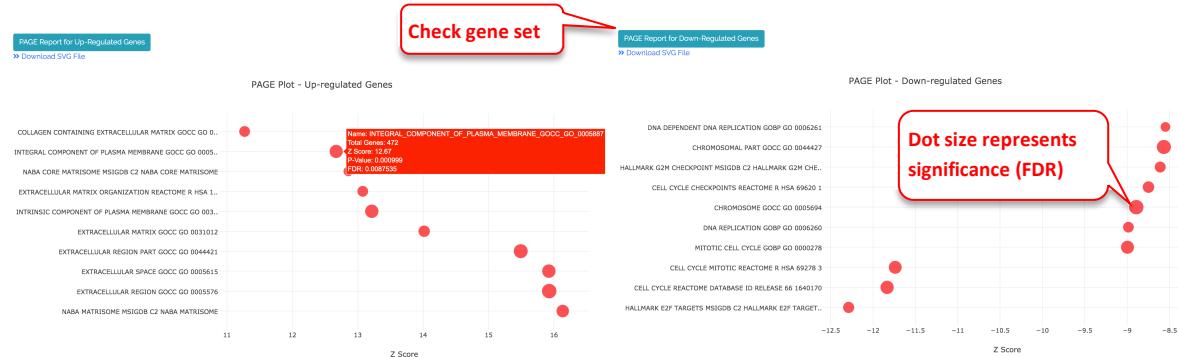


## Gene Set Enrichment from Ranked Genes

For each comparison, we produce a rank file for all genes using logFC. We use PAGE (Parametric Analysis of Gene Set Enrichment) to identify significant biological changes. PAGE can be more sensitive for comparisons where the logFC is relatively small, but most genes in a functional set show the same direction of change.

The predefined gene sets were from MSigDB.

For each comparison, the top up-regulated and down-regulated gene sets are plotted.



To view the full list of gene sets, you can click the report for genes as shown in following figure.

### UP-Regulated PAGE Report [Back to Comparison Details](#)

Gene Set Name	# Genes	Z Score	P Value	FDR
A_TETRASACCHARIDE_LINKER_SEQUENCE_IS_REQUIRED_FOR_GAG_SYNTHESIS.REACTOME_R_HSA_1974475_1	15	3.5026	0.000999	0.0087535
ACTIN_BASED_CELL_PROJECTION_GOCC_GO_0098858	123	3.8617	0.000999	0.0087535
ACTIN_BINDING_GOMF_GO_0003779	256	4.1771	0.000999	0.0087535
ACTIN_FILAMENT_BASED_PROCESS_GOBP_GO_0030029	340	3.499	0.000999	0.0087535
ACTION_POTENTIAL_GOBP_GO_0001508	36	5.7785	0.000999	0.0087535
ACTIVATION_OF_IMMUNE_RESPONSE_GOBP_GO_0002253	127	3.6063	0.000999	0.0087535
ACTIVATION_OF_MATRIX_METALLOPROTEINASES.REACTOME_DATABASE_ID_RELEASE_66_1592389	15	8.8731	0.000999	0.0087535
ACTIVE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY_GOMF_GO_0022B04	174	5.2265	0.000999	0.0087535
ADAPTIVE_IMMUNE_RESPONSE_BASED_ON_SOMATIC_RECOMBINATION_OF_IMMUNE_RECEPTEORS_BUILT_FROM_IMMUNOGLOBULIN_SUPERFAMILY_DOMAINS_GOBP_GO_0002460	83	3.3924	0.000999	0.0087535
ADAPTIVE_IMMUNE_RESPONSE_GOBP_GO_0002260	124	3.7329	0.000999	0.0087535

### 4.3.4 Multi-layer visualization

If you are interested in a particular pathway, sometimes it is useful to map the RNA-Seq or microarray data to the pathway for visualization.

**1. Start here**

The screenshot shows the CHOmics v1 interface. At the top, there's a navigation bar with links like 'Toolbox', 'My Analyses', 'Admin', 'Projects (1)', 'Comparisons (12)', and 'Samples (30)'. Below this, there are two main sections: 'My Experiments' and 'CHO Demo'. 'My Experiments' contains a 'Private Folder' path and a 'Create New Experiment' button. 'CHO Demo' shows a project from '2019-10-10' with 15 samples and 1 analysis. On the right, there are sections for 'Gene Expression Analysis' (with options like 'Gene Expression Plot', 'Heatmap', 'Correlation Tool', 'PCA Analysis', 'Export Expression Data') and 'RNA-Seq Data' (with a list of samples and analyses). A red box labeled '2. Select pathway for visualization' points to the 'Pathway Visualization' section under 'Other Tools'.

**KEGG Pathway View**

**Start Over** Note: \* denotes required fields.

Glycolysis / Gluconeogenesis  **3. Choose pathway**

**4. Choose comparison**

Comparisons: [Load from saved lists](#) [Search and Select](#) [Select a Project](#) [Clear](#)

D108.vs.D72

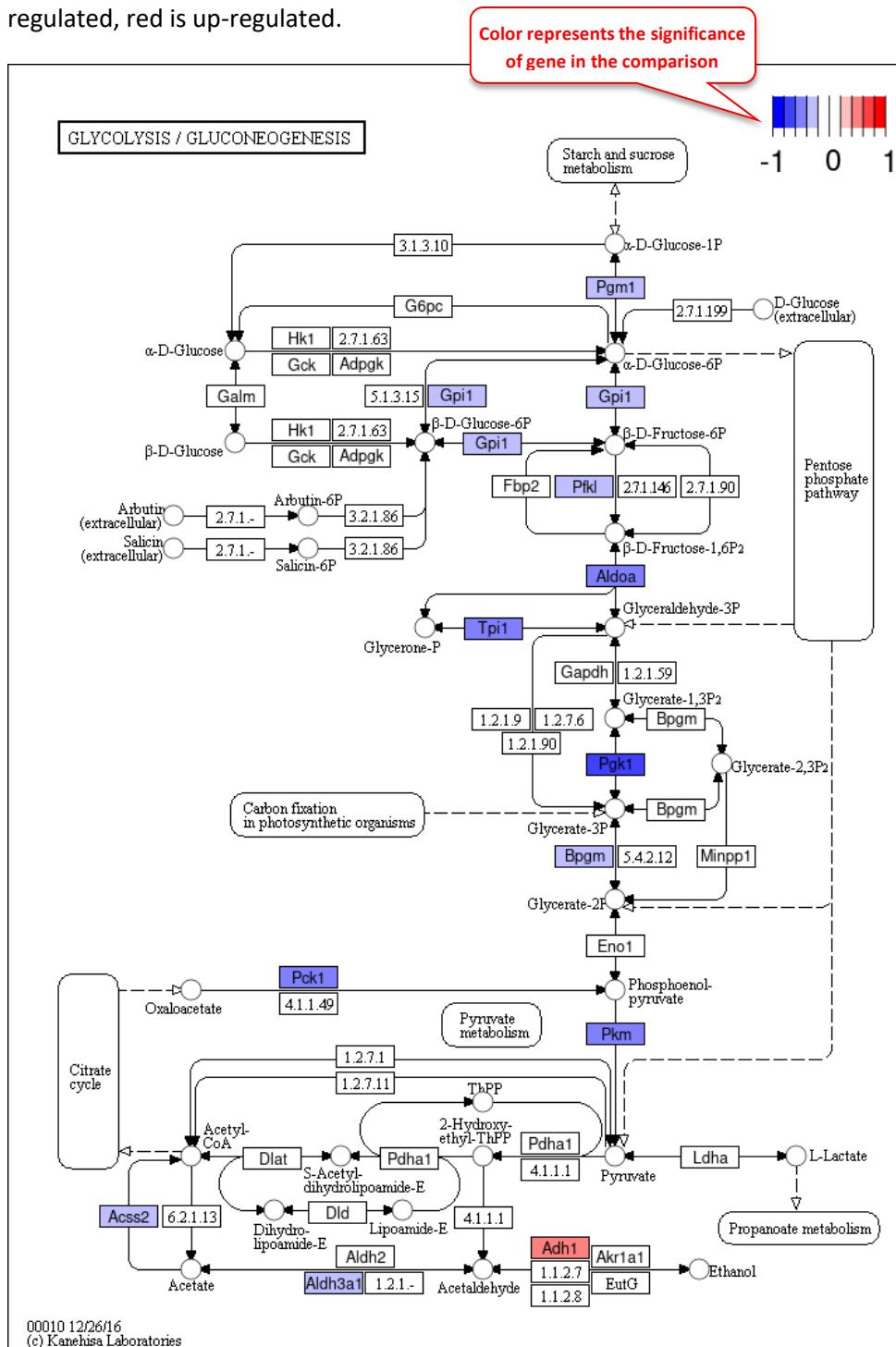
Or, upload your comparison files:  No file chosen [Demo Data](#)

Visualization: Gradient Blue-White-Red (-1,0,1) **5. (Optional) change visualization settings**

**Submit** **6. Submit to view the pathway**

This screenshot shows the 'KEGG Pathway View' page. It has a 'Start Over' button and a note about required fields. Below is a search field for 'Glycolysis / Gluconeogenesis' with a red box labeled '3. Choose pathway'. There's also a 'Comparison' dropdown with a red box labeled '4. Choose comparison'. Below these are buttons for loading saved lists, searching, selecting a project, and clearing. A 'D108.vs.D72' comparison is listed. There's a file upload section with a 'Choose File' button and a 'No file chosen' message, along with a 'Demo Data' link. Under 'Visualization', there's a color gradient selection ('Gradient Blue-White-Red (-1,0,1)') with a red box labeled '5. (Optional) change visualization settings'. At the bottom is a 'Submit' button with a red box labeled '6. Submit to view the pathway'.

In the pathway plot, typically we use red-blue color scale to show the log2 Fold Change. Blue is down-regulated, red is up-regulated.

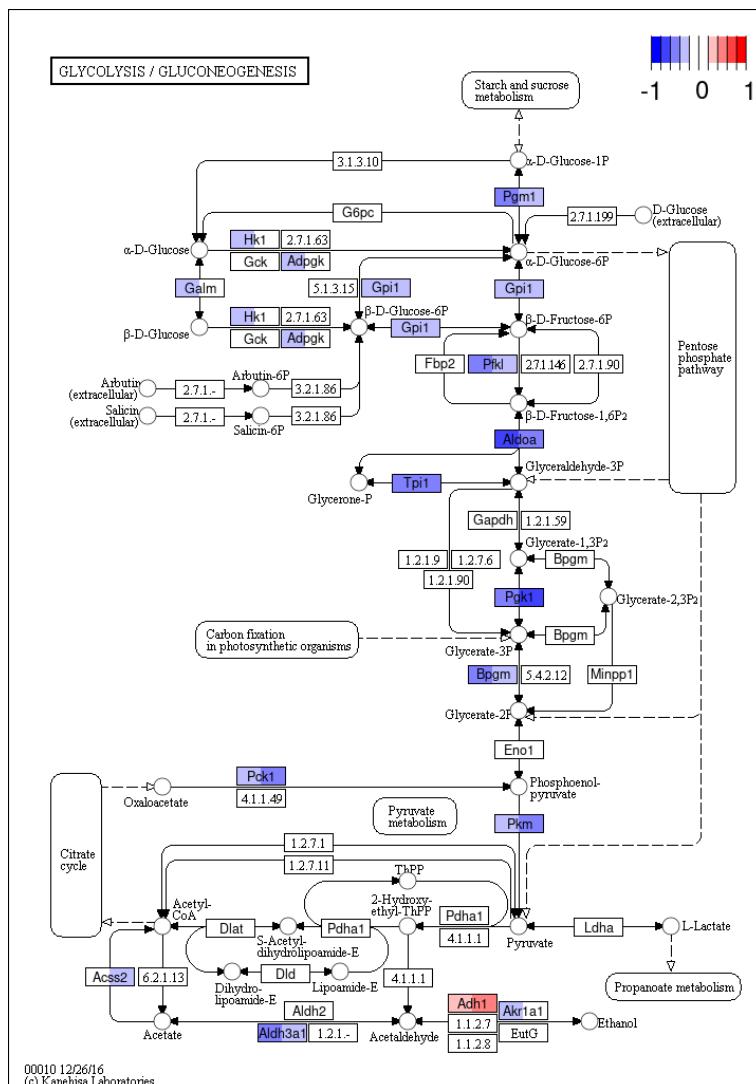


Pathway Plot from Several Comparisons

The user can add multiple comparisons from the pathway plot tool by clicking Add Comparison link. Besides showing log2 Fold Change, the user can also show statistical significance by clicking Enable Second Visualization Columns.

The screenshot shows the KEGG Pathway View interface. At the top, there is a navigation bar with links: CHOmics (v1), Toolbox, My Analyses, Admin, Projects (1), Comparisons (12), and Samples (30). Below the navigation bar, the title "KEGG Pathway View" is displayed. Underneath the title, there is a "Start Over" link and a note: "Note: \* denotes required fields." A link to "Glycolysis / Gluconeogenesis" is shown, along with a "Download Pathway File" button. A green button labeled "» Select Pathway" is visible. Below these, there is a section titled "Comparisons:" with links: "» Load from saved lists", "Search and Select", "Select a Project", and "Clear". Two comparison entries are listed: "D84.vs.D72" and "D108.vs.D72". A red box and arrow highlight the "Choose multiple comparison" link next to the second entry. Further down, there is a section for uploading comparison files with a "Choose File" button and a message "No file chosen". A "» Demo Data" link is also present. The "Visualization:" section includes a color gradient selection: "Gradient Blue-White-Red (-1,0,1)". At the bottom, a blue "Submit" button is located.

The pathway plot will now have multiple color bars corresponding to the different comparisons.



### 4.3.5 Pathway Heatmap From Comparisons

Users can display the enriched pathways from several related comparisons, and visualize the top enriched pathways across comparisons. Users can mix public data and inhouse comparisons.

The screenshot shows the CHOmics Pathway Heatmap interface with the following features highlighted:

- Select pathway database:** A dropdown menu set to KEGG.
- Choose multiple comparisons:** A red box highlights the "Comparisons" dropdown menu.
- The pathways are ranked by the most significant value:** A red box highlights the "Show Type" dropdown menu, which includes options for Top 10, Top 20, Top 50, and Top 100.
- Refresh Pathway & GeneSets:** A green button at the bottom right.

The heatmap shows pathways in rows, comparisons in columns. The statistical significance is color-coded (log P-value, or Z-score). Pathways are sorted by the negative logP values from the highest to the lowest.



From the pathway heatmap, users can click any data point to view details.

CHOMics (v1) Toolbox ▾ My Analyses ▾ Admin ▾ Projects (1) Comparisons (12) Samples (30)

Pathway Names and GeneSets:

Up-Regulated Genesets

- Transcriptional misregulation in cancer
- Human papillomavirus infection
- Fluid shear stress and atherosclerosis
- Pi3K-Akt signaling pathway
- TNF signaling pathway
- MAPK signaling pathway
- Galactose metabolism
- Rheumatoid arthritis
- Protein processing in endoplasmic reticulum

[Draw Heatmap](#)

Up-Regulated Genesets vs. Comparisons, log<sub>10</sub>(p-value) [Save SVG](#)

Data Actions

- [View Comparison Detail](#)
- [View Data in Volcano Plot](#)

Close

Fructose and mannose metabolism  
Progesterone-mediated oocyte maturation  
Huntington's disease  
Base excision repair  
Alzheimer's disease

x: D108 vs. D72  
y: Lysosome  
z: 44.082867304813  
logP: -44.0828673048135  
number of genes: 47

## 5 Customized analysis pipeline

### 5.1 Use alternative tool or algorithm

The analysis pipeline is modular, each step can be modified by users to use an alternative method if desired. The users should be familiar with the Linux bash to run the analysis steps and be familiar with php programming to make modification to the source code.

The full analysis pipeline has four steps, and each step is listed in a bash file in the analysis folder in the system.

- step\_0.sh FASTQC of raw data
- step\_1.sh Alignment to genome
- step\_2.sh Gene count
- step\_3.sh DEG detection and functional enrichment

These bash files are created by PHP programs chomics/app/bxgenomics/bxgenomics\_exe\_analysis.php, when users launch analysis pipeline online in a web browser via chomics/app/bxgenomics/analysis.php. For example, the current pipeline uses subread to perform alignment. If users want to modify the pipeline to change it to use the STAR program for alignment, they need the following steps:

- 1) Install STAR program on the server, prepare STAR index for the CHO genome.
- 2) Check the commands in step\_1.sh, and change the commands as needed. In this case, the subread command (subjunc step) needs to be replaced by the equivalent STAR command. Since STAR can sort the bam files, the samtools sort step can be omitted. Finally, the STAR output file is named as SampleIDAligned.sortedByCoord.out.bam, an extra step is needed to rename it to SampleID.sorted.bam, so step2.sh can output gene count files with the correct sample names.
- 3) Edit PHP program chomics/app/bxgenomics/bxgenomics\_exe\_analysis.php, find the part that generates step\_1.sh (The section is marked as “Step 1. Alignment with Subread”), and then make changes accordingly.
- 4) Test the updated system to make sure it works as expected.