HSAPath: Activation Analysis for Human Cellular Pathways

Addendum to ESKAPE Act PLUS Manual

This application extends the functionality of ESKAPE Act PLUS to the human genome. Given data from high-throughput experiments from human subjects, HSAPath returns information on activated and repressed KEGG pathways and GO terms.

Most of manual developed for ESKAPE Act PLUS applies to HSAPath. The sections that are unique to the ESKAPE Act PLUS version of the application are highlighted in yellow. Section V, "Worked Example", outlines an analysis using differential gene expression data from pseudomonas samples. There is updated sample data for HSAPath, which can be downloaded in the application via the "HSA example data" button. After downloading the data you can follow the same procedure outlined in Section V, ignoring the steps regarding selection of bacterial strain.

The data file containing KEGG pathways, GO terms and their constituent genes was constructed using the EnrichmentBrowser (version 2.28.0) and the limma (version 3.54.1) R packages. Enrichment Browser was used to get GO terms and limma was used to get KEGG pathways.

ESKAPE Act PLUS

Activation Analysis for ESKAPE Pathogens and other Prokaryotes Labs Usually Study

User Manual

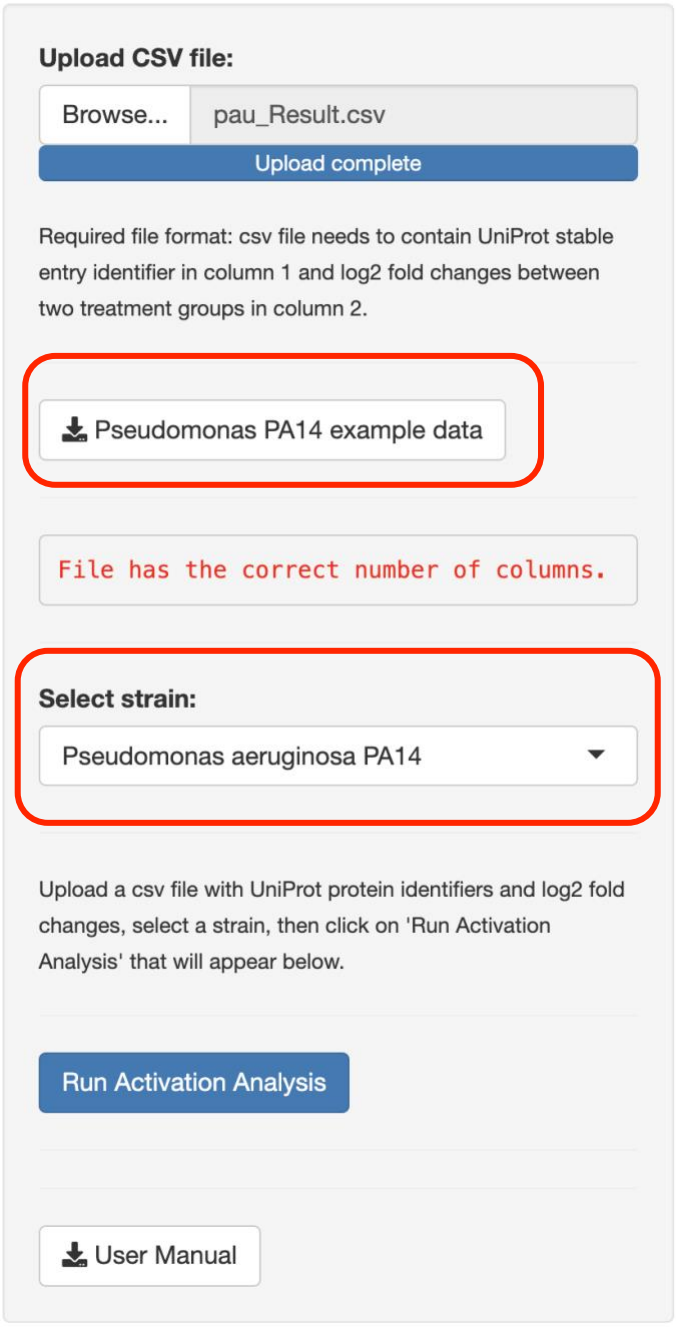
Second edition: September 2022

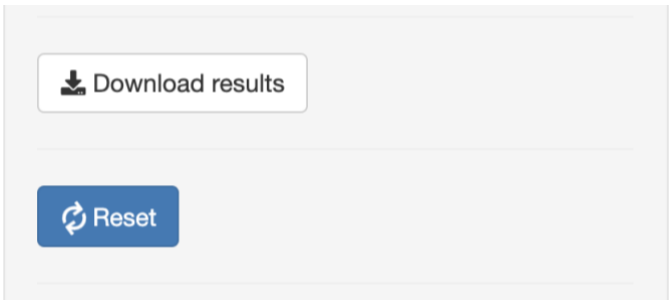
Thomas Hampton, PhD
Department of Microbiology and Immunology
The Geisel School of Medicine at Dartmouth
Hanover, NH 03755, USA
e-mail: Thomas.H.Hampton@Dartmouth.edu

If using this app advances your research, please cite our publication:

ESKAPE Act Plus: Pathway Activation Analysis for Bacterial Pathogens.
Koeppen K, Hampton TH, Neff SL, Stanton BA. mSystems. 2022 Oct 19:e0046822.
doi: 10.1128/msystems.00468-22. PMID: 36259735.

Contents





I. Quick Start Guide

1. Input file upload

Click the "Browse" button to upload a csv file with input data. The file should contain UniProt stable entry identifiers in column 1 and log2 fold changes in column 2.

Example data can be downloaded by clicking the "Pseudomonas PA14 example data" button.

2. Strain selection

From the dropdown menu, select the strain that corresponds to the UniProt entry IDs in the input file.

When using the example data, choose "Pseudomonas aeruginosa PA14".

3. Run activation analysis

Once an input file has been uploaded, a "Run Activation Analysis" button will appear. Click on it to start the Analysis.

This user manual can be downloaded with the "User Manual" button.

4. Obtain results

Results will appear in tabs in the main panel of the application (more details on page 12). Output files can be downloaded with the "Download results" button.

5. Reset application

The application can be reset to its original state with the "Reset" button.

II. Introduction

1. ESKAPE pathogens

The group of bacteria collectively known as ESKAPE (***E****nterococcus faecium,* ***S****taphylococcus aureus,* ***K****lebsiella pneumoniae,* ***A****cinetobacter baumannii,* ***P****seudomonas aeruginosa and* ***E****nterobacter sp.)* are pathogens of concern due to their virulence and ability to develop antibiotic resistance. Collectively, they are the major cause of life-threatening hospital-acquired infections world-wide. The ESKAPE Act PLUS web application enables analysis of high-throughput experiments on ESKAPE pathogens and several other strains of bacteria that are of relevance to the biomedical research community. A complete list of supported strains is provided on page 10.

2. KEGG pathways and GO terms

ESKAPE Act PLUS uses biological pathway information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (M. Kanehisa and Goto 2000), obtained through the KEGGREST R package version 1.32.0 (Tenenbaum 2021), and gene ontology (GO) term annotations (Ashburner et al. 2000; Gene Ontology Consortium 2021) for biological processes, molecular functions and cellular components retrieved from the Universal Protein Resource (UniProt) (The UniProt Consortium 2021). KEGG pathway visualizations in the app are made possible via an API interface with KEGG Mapper's Color Tool (Minoru Kanehisa and Sato 2020; Minoru Kanehisa, Sato, and Kawashima 2022). In these visualizations, KEGG pathway images are overlayed with gene or protein level fold changes, allowing for easy visualization of pathway-level activation or repression. Genes or proteins with a positive fold change are highlighted in magenta, while genes or proteins with a negative fold change are depicted in green. Genes or proteins that were not detected in the experiment are shown in white. ESKAPE Act PLUS provides links to downloadable pathway level images for all significant KEGG pathways.

3. Activation analysis versus enrichment analysis

Many publicly available tools allow users to perform a gene set enrichment analysis for a biological pathway or GO term. Enrichment analysis typically requires a criterion such as a threshold p-value to identify a subset of differentially expressed genes or proteins. Differentially expressed genes are said to be "enriched" in genes performing specific biological functions if the subset of differentially expressed genes contains a higher proportion of genes performing these functions than one would expect by chance. Enrichment therefore suggests an association between experimental conditions and biological function. However, selecting different p-value cutoffs can

change enrichment results, and typical choices such as FDR < 0.05 may result in very small gene sets, limiting power. In addition, enrichment analysis does not consider the direction of the fold changes and thus does not predict the biological effect associated with the induction or repression of a given pathway or GO term. By contrast, ESKAPE Act PLUS uses all genes or proteins that were detected in a high-throughput experiment and their associated fold changes to predict overall activation or repression at the level of a biological pathway or GO term based on the fold changes of all genes associated with a given pathway or GO term.

4. Why use ESKAPE Act PLUS?

ESKAPE Act PLUS makes it easier for researchers working with prokaryotes to interpret high-throughput experiments such as RNA-seq and proteomics, leading to new insights and hypotheses. Specifically, ESKAPE Act PLUS provides biological pathway and GO term activation analysis, which were previously only available for eukaryotic systems and popular model organisms.

III. Input requirements

1. Input file format

Input files need to be files with a comma separated values format (.csv) that contain two columns without headers:

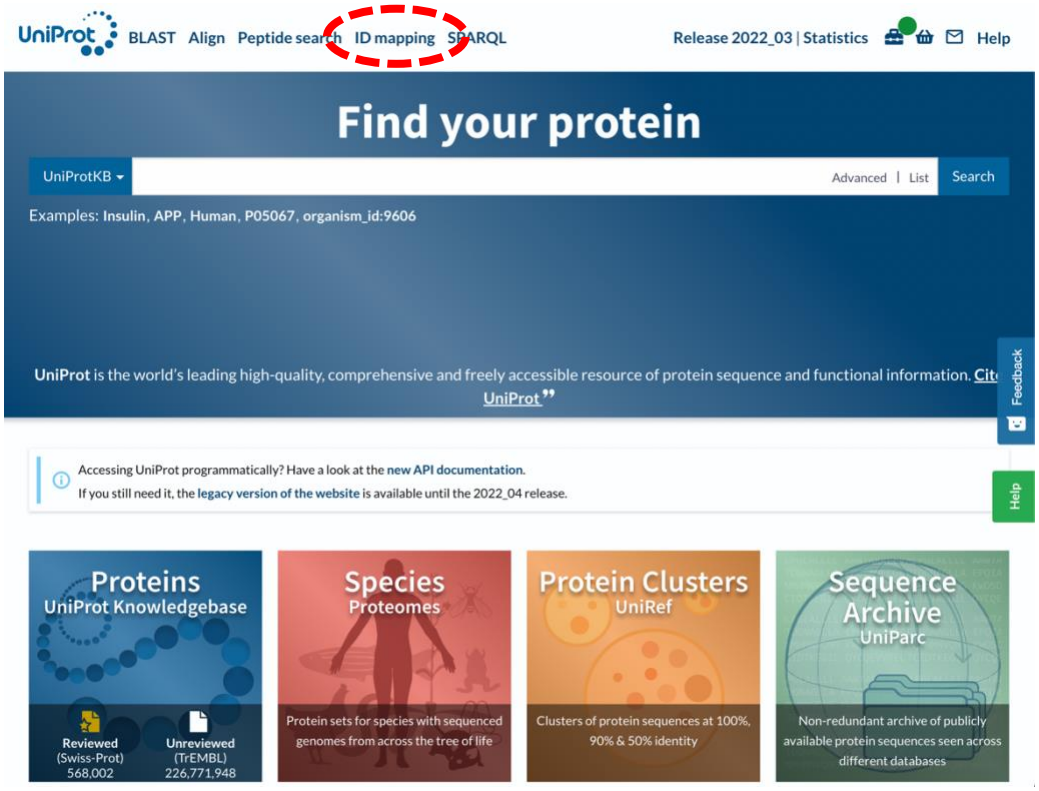
- **Column 1** contains UniProt identifiers (ID conversion is explained in section 2. below).
- **Column 2** contains log2 fold changes for the UniProt identifier specified in Column 1, for the comparison of interest. This column must contain numeric values only, no text, and may not contain any missing values.

Selecting "Download Pseudomonas PA14 example data" will download an example input data file for *Pseudomonas aeruginosa* strain PA14 ("pau_Result.csv") to illustrate the correct data format.

2. UniProt ID conversion

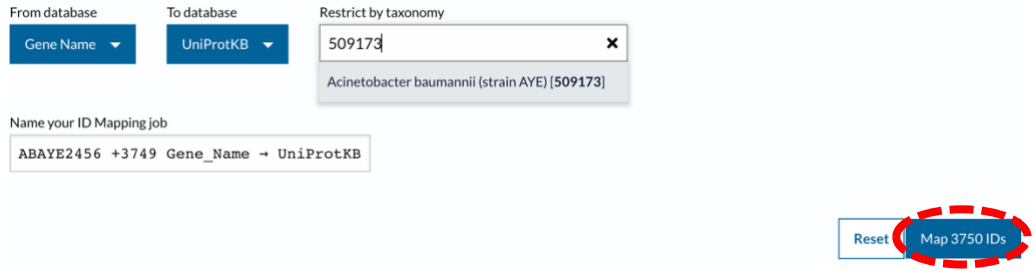
Step 1 – go to the UniProt homepage

Step 2 – select "ID mapping":

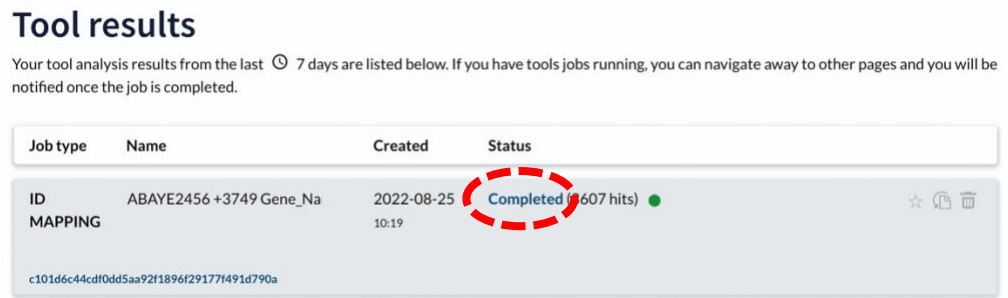


Step 3 – paste your gene identifiers into the provided window or upload a file containing your gene identifiers:

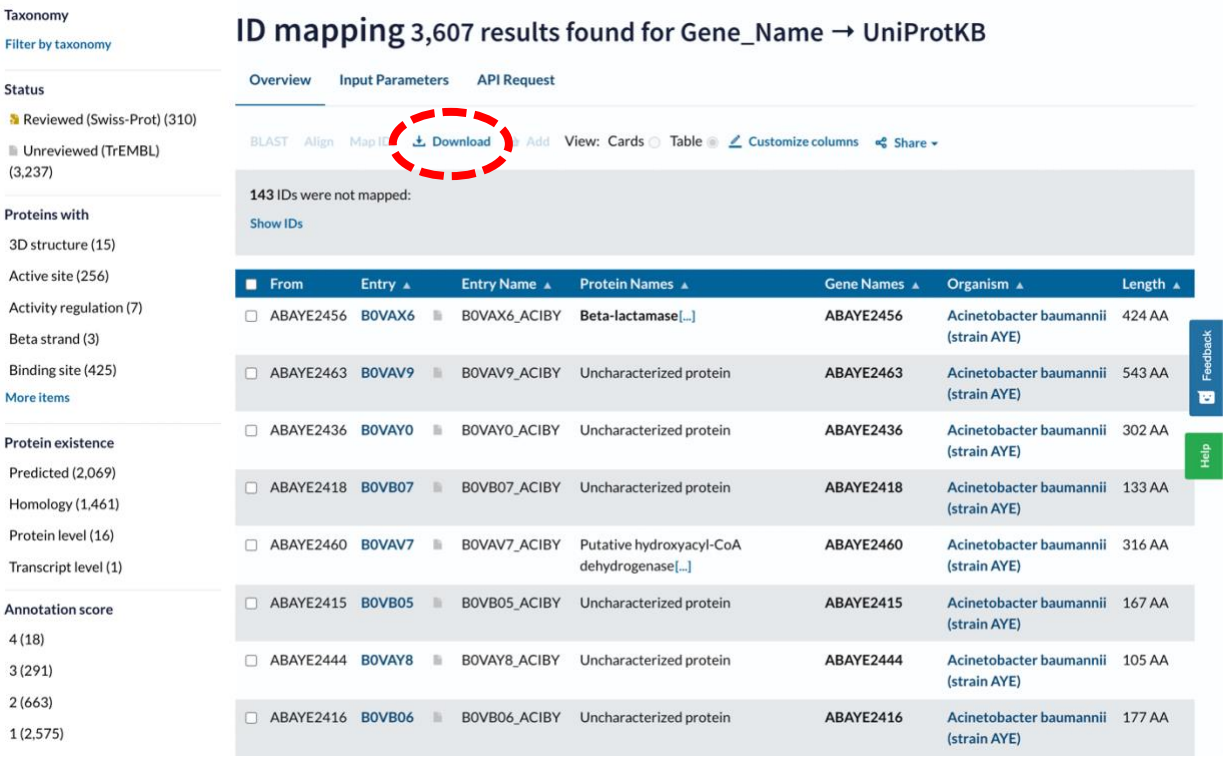
Step 4 – Select from "Gene name" (or other identifier, if applicable) to "UniProtKB", enter the Taxon ID for your strain of interest, select the organism once the popup window appears and click the "Map IDs" button. A table with the Taxon IDs for all supported strains can be found on page 10. The example below is for *Acinetobacter baumannii* stain AYE (Taxon ID 509173):



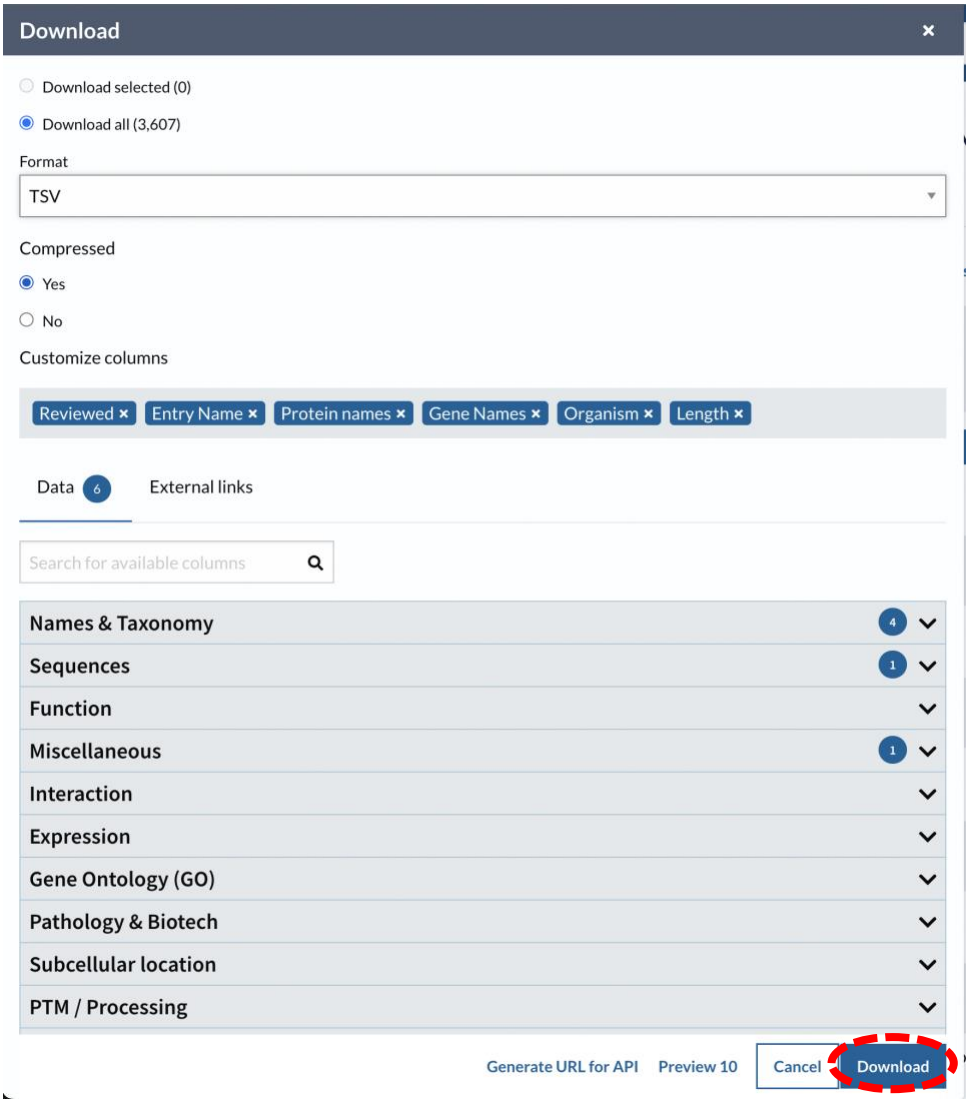
Step 5 – On the following page, select the most recent ID mapping job:



Step 6 – Verify that a sufficient number of identifiers were successfully mapped. Small numbers of mapped identifiers (e.g., fewer than 200) may reduce the number of significant KEGG pathways and GO terms that can be identified. Bear this in mind, then click the "Download" button:



Step 7 – select "TSV" from the dropdown menu and click "Download":



Step 8 – Create the input data file for ESKAPE Act PLUS by merging the Tab-separated file downloaded from UniProt (containing the original gene identifiers and the UniProt entry IDs required by ESKAPE Act PLUS) with the table containing gene identifiers and fold changes. Instructions on how to merge data using different platforms can be found in the links below:

Merging tables in Excel

Merging tables using R

Merging tables using Python

IV. Overview of capabilities

1. Supported strains

ESKAPE Act PLUS currently supports 23 strains from 13 species, including the six ESKAPE pathogens as well as several other species that are of relevance to the biomedical research community. Listed below are the KEGG, UniProt and Taxon identifiers for each strain.

Species	Strain	KEGG ID	UniProt ID	Taxon ID
Acinetobacter baumannii	AYE	aby	ACIBY	509173
Acinetobacter baumannii	MDR-ZJ06	abz	ACIBM	497978
Bacteroides fragilis	NCTC 9343	bfs	BACFN	272559
Bacteroides ovatus	3725 D1 iv	boa	BACOV	28116
Bacteroides thetaiotaomicron	VPI-5482	bth	BACTN	226186
Clostridioides difficile	630	cdf	CLOD6	272563
Clostridioides difficile	R20291	cdl	CLODR	645463
Enterobacter cloacae	ATCC 13047	enc	ENTCC	716541
Enterococcus faecium	DO	efu	ENTFD	333849
Escherichia coli	K-12 MG1655	eco	ECOLI	511145
Escherichia coli	K-12 W3110	ecj	ECOLI	316407
Escherichia coli	O157:H7 EDL933	ece	ECO57	155864
Escherichia coli	O157:H7 Sakai	ecs	ECO57	386585
Escherichia coli	BL21(DE3)	ebd	ECOBD	469008
Klebsiella pneumoniae	MGH 78578	kpn	KLEP7	272620
Prevotella melaninogenica	ATCC 25845	pmz	PREMB	553174
Pseudomonas aeruginosa	UCBPP-PA14	pau	PSEAB	208963
Pseudomonas aeruginosa	PAO1	pae	PSEAE	208964
Staphylococcus aureus	COL	sac	STAAC	93062
Staphylococcus aureus	USA300 FPR3757	saa	STAA3	367830
Staphylococcus aureus	NCTC8325	sao	STAA8	93061
Staphylococcus aureus	Newman	sae	STAAE	426430
Streptococcus sanguinis	SK36	ssa	STRSV	388919

2. Number of KEGG pathways and GO terms for each strain

ESKAPE Act PLUS will calculate activation for KEGG pathways and GO terms that have 4 or more associated genes in the input data provided. For example, if the input data contain only three genes on a specific path, no activation score will be calculated. The total number of possible KEGG pathways and GO terms available for each strain are listed below.

Species	Strain	KEGG paths	GO terms
Acinetobacter baumannii	AYE	92	1622
Acinetobacter baumannii	MDR-ZJ06	93	1516
Bacteroides fragilis	NCTC 9343	85	1398
Bacteroides ovatus	3725 D1 iv	80	1285
Bacteroides thetaiotaomicron	VPI-5482	83	1482
Clostridioides difficile	630	83	1339
Clostridioides difficile	R20291	84	1240
Enterobacter cloacae	ATCC 13047	94	2210
Enterococcus faecium	DO	93	2310
Escherichia coli	K-12 MG1655	94	3936
Escherichia coli	K-12 W3110	94	3939
Escherichia coli	O157:H7 EDL933	94	2267
Escherichia coli	O157:H7 Sakai	70	1237
Escherichia coli	BL21(DE3)	92	2105
Klebsiella pneumoniae	MGH 78578	99	2164
Prevotella melaninogenica	ATCC 25845	99	2929
Pseudomonas aeruginosa	UCBPP-PA14	101	1921
Pseudomonas aeruginosa	PAO1	71	985
Staphylococcus aureus	COL	88	1378
Staphylococcus aureus	USA300 FPR3757	86	1357
Staphylococcus aureus	NCTC8325	85	1334
Staphylococcus aureus	Newman	82	1776
Streptococcus sanguinis	SK36	70	1191

3. Statistics

ESKAPE Act PLUS uses a binomial test to assess whether a given KEGG pathway or GO term contains significantly more induced or repressed genes or proteins than would be expected by chance. KEGG pathway and GO term activation analysis were performed as previously described (Hampton et al. 2018; Goodale et al. 2019; Koeppen et al. 2021). Under the null hypothesis, about 50% of the genes or proteins associated with any KEGG pathway or GO term would be expected to respond to treatment with a FC > 0, while the other 50% would respond with a FC < 0. Statistically significant divergence from this 50:50 percent split is assessed using binomial tests based on the fold changes of all genes or proteins associated with a KEGG pathway or GO term. The universe for the binomial test is all genes or proteins from the input file (rather than all genes or proteins known for a given species). Using identifiers found in the source file as the reference avoids the possibility of introducing biases related to tissue type and sample preparation.

Pathway enrichment p-values from the binomial test are FDR-corrected to account for the total number of KEGG pathways or GO terms that were tested. Only KEGG pathways or GO terms with an FDR-corrected p-value < 0.05 are considered significantly induced or repressed and displayed in figures and tables of the in-app results output. The downloadable results .csv files also contain the uncorrected p-values for each KEGG pathway and GO term to allow users to identify borderline cases with uncorrected p-values < 0.05 that just barely missed the significance cutoff of FDR < 0.05.

4. App output

ESKAPE Act PLUS generates two kinds of output – in-app output and downloadable output. Following an activation analysis, the in-app output appears in tabs in the main panel of the application, while downloadable output becomes available via the "Download results" button that appears in the left control panel of the app.

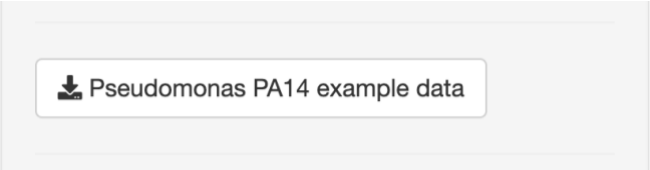
In-app output:

- Number of user-provided UniProt Entry identifiers that mapped to the internal KEGG and GO database as well as the number of significant KEGG pathways and GO terms.
- Boxplots for all significant (FDR < 0.05) KEGG pathways and GO terms.
- Tables with significant (FDR < 0.05) KEGG pathways and GO terms, containing estimates (< 0.5 = repressed, > 0.5 = activated), median fold changes, p-values from the binomial test and FDR-corrected p-values.
- Pathway level visualizations: the KEGG table contains clickable links to KEGG Mapper pathway level visualizations for each significant KEGG pathway. See page 4 for more details on pathway level visualizations.

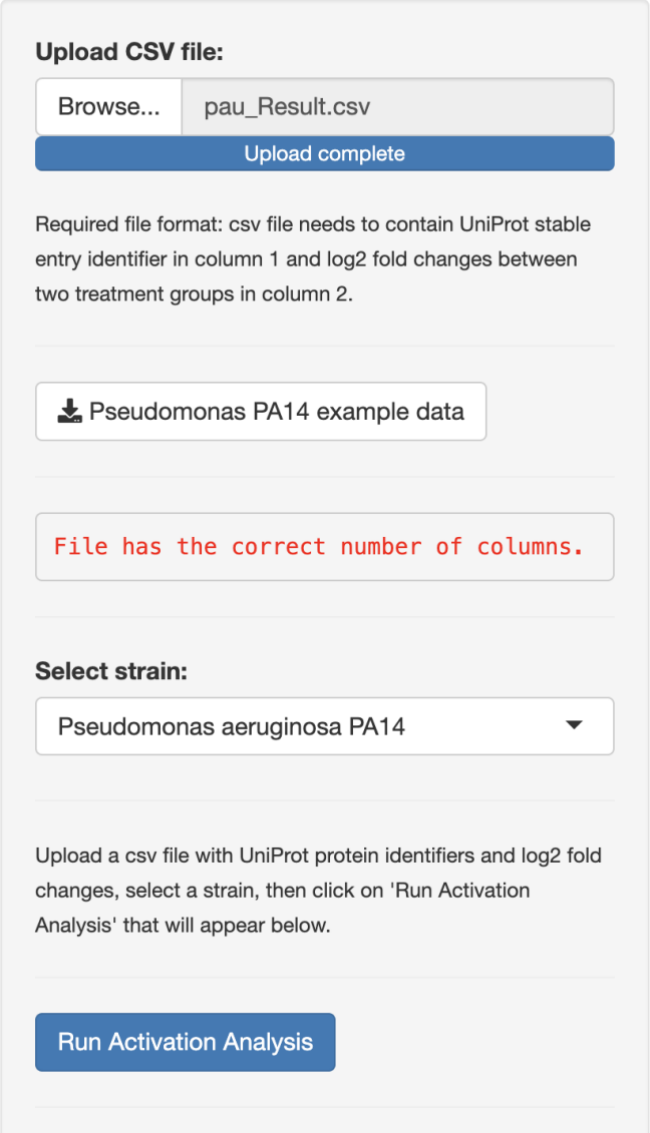
Downloadable output:

- Boxplots for all significant (FDR < 0.05) KEGG pathways and GO terms.
- Tables with estimates (< 0.5 = repressed, > 0.5 = activated), median fold changes, p-values, and FDR-corrected p-values for all KEGG pathways and GO terms that were tested.
- html file ("Pathways.html") with links to pathway level visualizations for all significant KEGG pathways generated by KEGG Mapper.

V. Worked example



1. Download *Pseudomonas aeruginosa* strain PA14 example data file **"pau_Result.csv"**

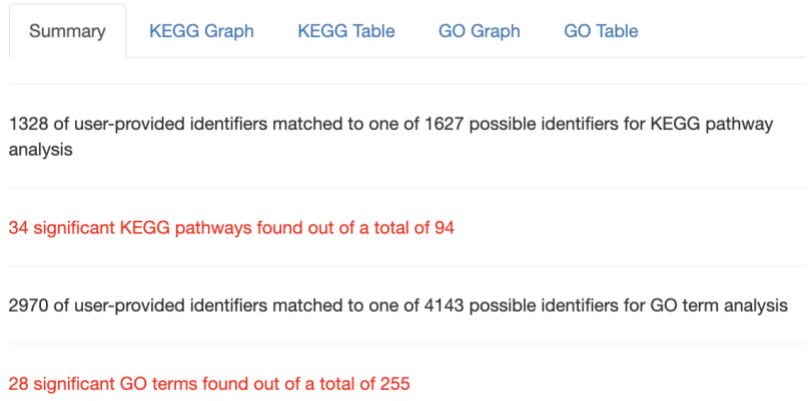


2. Upload example data file "pau_Result.csv" using the **"Browse…"** button.

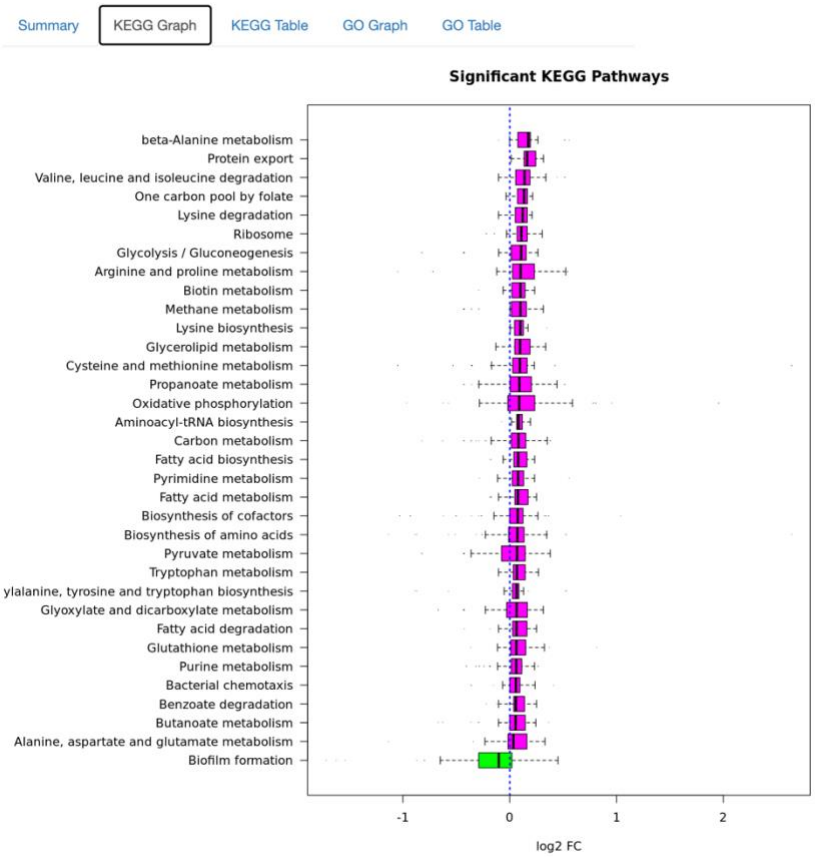
3. Select strain **"Pseudomonas aeruginosa PA14"**

4. Click **"Run Activation Analysis"**

5. The **"Summary"** tab shows how many user-provided identifiers mapped to the internal database for KEGG pathway and GO term activation analysis and how many KEGG pathways and GO terms were significantly induced or repressed:



6. The **"KEGG Graph"** tab contains a boxplot of log2 fold changes for each gene or protein on significant KEGG pathways. Magenta indicates pathway induction and green indicates repression:



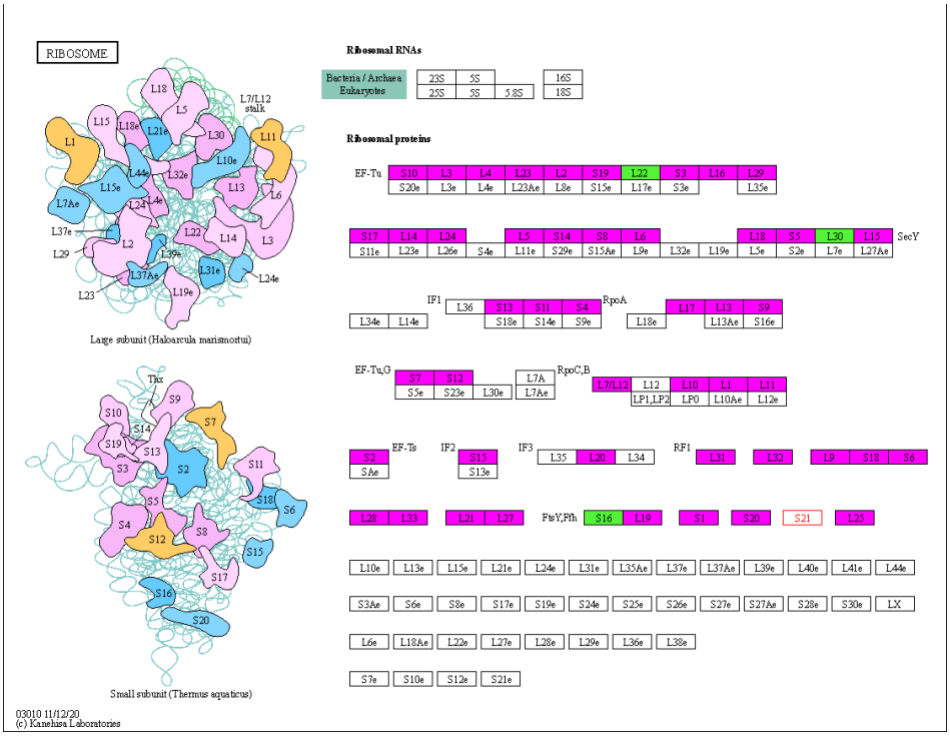
7. The **"KEGG Table"** tab provides a searchable and sortable table of significant KEGG pathways with pathway activation estimates (< 0.5 = repressed, > 0.5 = activated), median fold changes, p-values from the binomial test and FDR-corrected p-values:

Summary | KEGG Graph | KEGG Table | GO Graph | GO Table

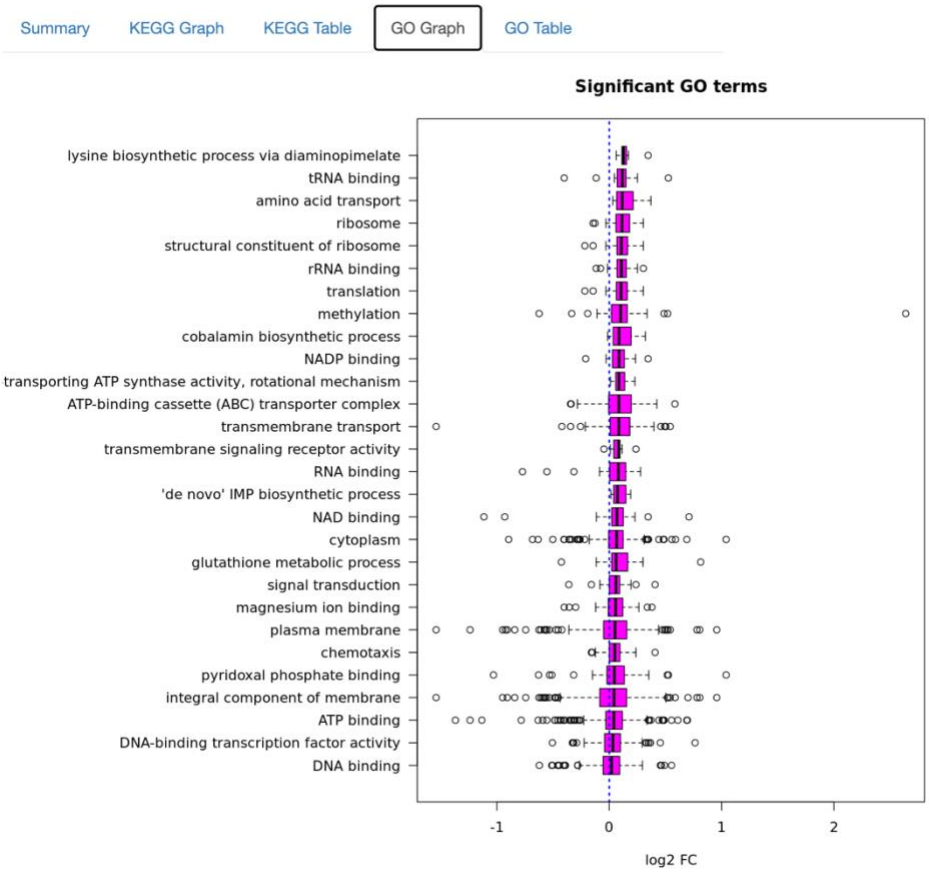
Show 50 entries Search:

Path	Estimate	Median_fold_change	P_value	FDR
Biofilm formation	0.33	-0.1	0.00219490890778059	0.00949158066743457
Pyruvate metabolism	0.69	0.07	0.00379370563537163	0.0132077159157383
Glyoxylate and dicarboxylate metabolism	0.7	0.06	0.00456153274640824	0.0147856578676681
Alanine, aspartate and glutamate metabolism	0.71	0.03	0.0166738478001207	0.0460982850944515

The Path column contains clickable links to pathway images generated by KEGG mapper, such as this example image:



8. The **"GO Graph"** tab contains a boxplot of log2 fold changes for each gene or protein associated with a significant GO term. Magenta indicates induction and green indicates repression:

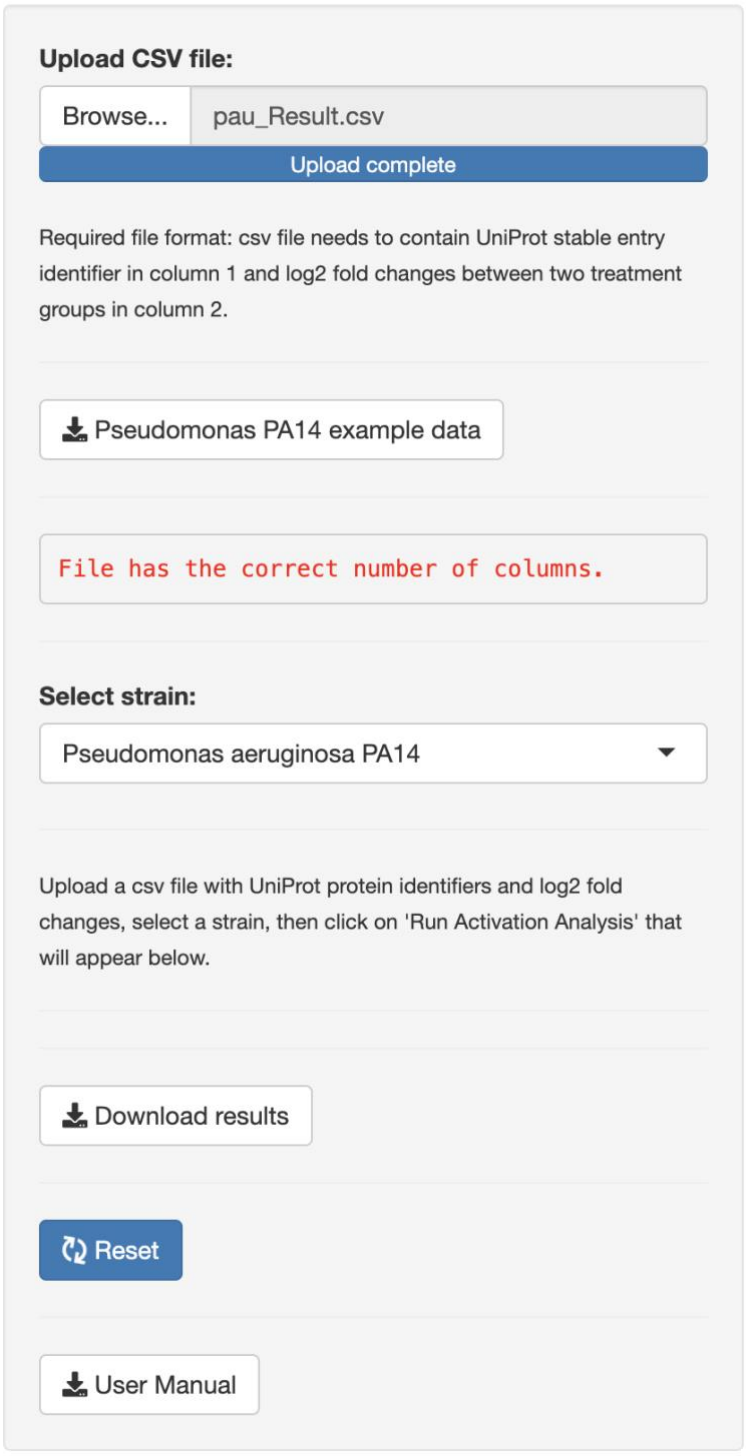


9. The **"GO Table"** tab provides a searchable and sortable table of significant GO terms with pathway activation estimates (< 0.5 = repressed, > 0.5 = activated), median fold changes, p-values from the binomial test and FDR-corrected p-values:

Summary | KEGG Graph | KEGG Table | GO Graph | GO Table

Show 50 entries Search:

GO_term	Estimate	Median_fold_change	P_value	FDR
DNA binding	0.61	0.02	0.000407810912196372	0.00649948641312968
integral component of membrane	0.62	0.04	0.00000341509554582281	0.00012407052026402
DNA-binding transcription factor activity	0.64	0.03	0.000142346183659903	0.00302485640277294



10. All results generated by the app (boxplots of significant KEGG pathways and GO terms as well as csv files with results for all KEGG pathways and GO terms and an html file with links to KEGG Mapper pathway level visualizations) can be downloaded with the **"Download results"** button.

11. The **"Reset"** button returns the app to its initial state before running a new analysis.

VI. Troubleshooting

Error message	Probable cause	Solution
Error – file has fewer than 2 columns!	The input file that was uploaded contains fewer than 2 columns of data	Upload an input file that contains UniProt entry IDs in column 1 and log2 fold changes in column 2
Error – file has more than 2 columns!	The input file that was uploaded contains more than 2 columns of data	Upload an input file that only contains 2 columns - UniProt entry IDs in column 1 and log2 fold changes in column 2
ID error - double check selected strain	The user-provided identifiers do not match the UniProt identifiers of the selected strain. Either the selected strain does not match the UniProt IDs of the input file or the input file does not contain the correct UniProt Entry IDs.	Double-check that the correct strain is selected that corresponds to the UniProt Entry IDs provided in the input file. If the input file does not contain the correct UniProt IDs, please refer to the section on ID conversion (page 6) to fix this problem.
Data error - The second column of the uploaded file needs to contain numeric values only, no missing values.	Verify that the second column of the input file contains a numeric fold change for each observation and does not contain any missing values.	Upload a corrected input file that contains a numeric value for each observation in the second column (no empty fields or rows).
Data error - The second column of the uploaded file needs to contain numeric values only, no text.	The second column of the input file contains text (rather than numeric values) for at least one of the rows. The input file contains a header with column names.	Upload a corrected input file without column name headers that contains numeric values only in the second column.

VII. References

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium." *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Gene Ontology Consortium. 2021. "The Gene Ontology Resource: Enriching a GOld Mine." *Nucleic Acids Research* 49 (D1): D325–34. https://doi.org/10.1093/nar/gkaa1113.

Goodale, Britton C., Thomas H. Hampton, Emily N. Ford, Craig E. Jackson, Joseph R. Shaw, Bruce A. Stanton, and Benjamin L. King. 2019. "Profiling MicroRNA Expression in Atlantic Killifish (Fundulus Heteroclitus) Gill and Responses to Arsenic and Hyperosmotic Stress." *Aquatic Toxicology (Amsterdam, Netherlands)* 206 (January): 142–53. https://doi.org/10.1016/j.aquatox.2018.11.009.

Hampton, Thomas H., Craig Jackson, Dawoon Jung, Celia Y. Chen, Stephen P. Glaholt, Bruce A. Stanton, John K. Colbourne, and Joseph R. Shaw. 2018. "Arsenic Reduces Gene Expression Response to Changing Salinity in Killifish." *Environmental Science & Technology* 52 (15): 8811–21. https://doi.org/10.1021/acs.est.8b01550.

Kanehisa, M., and S. Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28 (1): 27–30. https://doi.org/10.1093/nar/28.1.27.

Kanehisa, Minoru, and Yoko Sato. 2020. "KEGG Mapper for Inferring Cellular Functions from Protein Sequences." *Protein Science: A Publication of the Protein Society* 29 (1): 28–35. https://doi.org/10.1002/pro.3711.

Kanehisa, Minoru, Yoko Sato, and Masayuki Kawashima. 2022. "KEGG Mapping Tools for Uncovering Hidden Features in Biological Data." *Protein Science: A Publication of the Protein Society* 31 (1): 47–53. https://doi.org/10.1002/pro.4172.

Koeppen, Katja, Amanda Nymon, Roxanna Barnaby, Laura Bashor, Zhongyou Li, Thomas H. Hampton, Amanda E. Liefeld, et al. 2021. "Let-7b-5p in Vesicles Secreted by Human Airway Cells Reduces Biofilm Formation and Increases Antibiotic Sensitivity of P. Aeruginosa." *Proceedings of the National Academy of Sciences of the United States of America* 118 (28): e2105370118. https://doi.org/10.1073/pnas.2105370118.

Tenenbaum, Dan. 2021. *KEGGREST: Client-Side REST Access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)* (version 1.32.0).

The UniProt Consortium. 2021. "UniProt: The Universal Protein Knowledgebase in 2021." *Nucleic Acids Research* 49 (D1): D480–89. https://doi.org/10.1093/nar/gkaa1100.