omics.data.edu.au

Pathway Tools “How-to” Guides

How to overlay transcriptome data for a single pathway using Pathway tools

This guide provides step-by-step instructions using Pathway Tools to overlay transcriptomic expression data on a **known pathway** of interest (example shown in Figure 1).

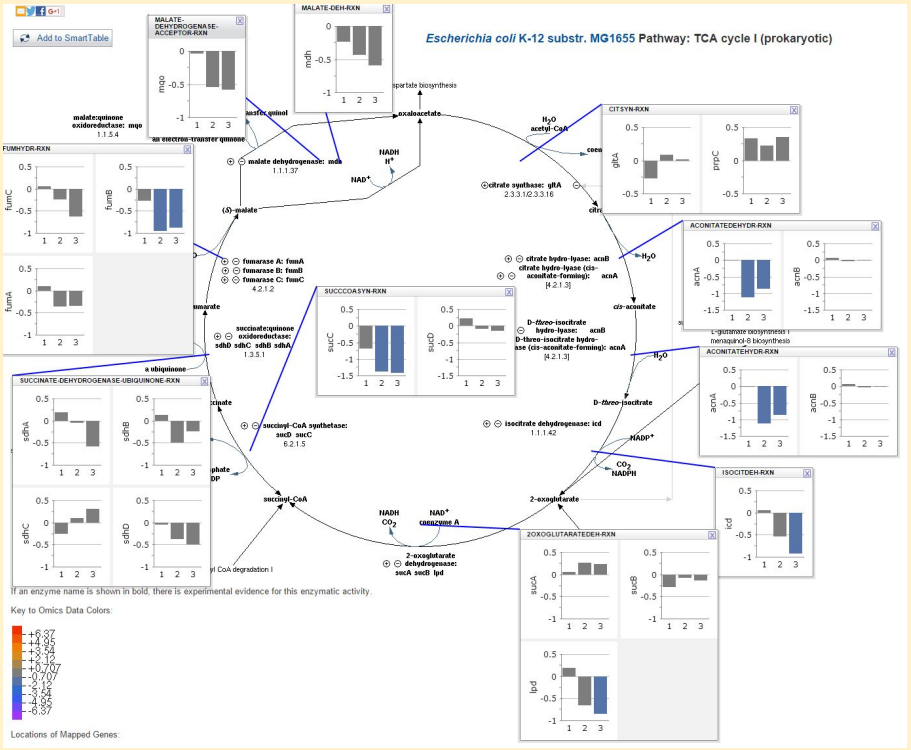


Figure 1 Overlay of experimental transcriptomic data on Metabolic Pathways. A close up of the TCA cycle E.coli K-12, ***with expression levels of constituent proteins shown in bar graphs across three experiments.***

This tutorial is for Pathway Tools version 19.5.

1 Introduction

Pathway Tools is a software suite from BioCyc database collection. Amongst its many functionalities (see http://bioinformatics.ai.sri.com/ptools/) , it supports the visual analysis of gene or protein expression and metabolomics datasets**,** such as overlaying omics data (eg transcriptomic expression values) onto diagrams of an organism's metabolic network, as shown in Figure 1.

By default, Pathway Tools contains four organism databases: (1) Homo sapiens, (2) Mus musculus, (3) ***E.coli K-12*** and (4) MetaCyc.

Pathway Tools also allow users to create their own organism-specific Pathway/Genome Database (PGDB) using the Pathologic tool. Creation of a customised database is covered in the online tutorial module "Pathway prediction and Annotations for new organisms".

2 Aim

This how-to guide will provide step-by-step instructions on how to use the overlay function of Pathway Tools using the web-browser.

This guide uses the metabolic pathways from the model organism ***Escherichia coli K-12*** and the data from the published study Varas ***et al***, 2017, ***Multi-level evaluation of Escherichia coli polyphosphate related mutants using global transcriptomic, proteomic and phenomic analyses***, Biochimica et Biophysica Acta 1861:871-883. In this study, the expression of all genes and proteins of ***E.coli K-12*** was measured using microarray and LC-ESI-MS/MS methods. The expressed were measured across three mutant strains (***Δppk1, Δppx, and ΔpolyP),*** which lack enzymes related to inorganic polyphosphate metabolism.

3 Data

The transcriptomic data files are sourced from the Genome Expression Omnibus (GEO) database on NCBI. The following samples from series GSE29954 were downloaded, these contained the gene expression levels relative to wild-type ***E.coli*** K-12 for:

- ***Δppk1 (***GSM741272)
- ***Δppx*** (GSM741273)
- ***ΔpolyP*** (GSM741274)

and the metadata file GSE29954_raw_data_IDs_ALL_columns.txt.gz with information about the gene IDs and symbols used in the study. The gene symbols are used in the pathway overlay feature to map the genes (and proteins) in the expression data files to the metabolic pathway.

Pathway Tools simply takes in any data with 2 or more columns, where the first column contains a gene symbol and subsequent columns are the expression level values for different experimental conditions. As such, while the example data used in this guide was derived from microarray experiments, the following step-by-step guide to overlay the gene expression onto the metabolic pathways can also be applied to NGS-derived data.

We have already performed the data transformation required to match the format required by pathway tools. The input file for this guide is **"all_mutants_transcriptomic_modified.txt"**. Box

3.1, below, describes the steps used for transforming and concatenating the GSM datasets into one file.

3.1 Data transformation

As noted above, Pathway Tools takes in data with 2 or more columns, where the first column contains a gene symbol and subsequent columns are the expression values. Here we describe how the data transformation was done for this guide.

Each of the raw files (GSM741272, GSM741273 and GSM741274) have two columns: ID_REF and log2 value. The ID_REF values are project specific gene identifiers assigned by Varas ***et al*** (e.g. 4.3.1.17) and need to be converted to a gene symbol since gene symbols are used within Pathway Tools to denote genes/proteins. In this case, the gene annotations were supplied in the accompanying GSE29954_raw_data_IDs_ALL_columns.txt file, which was used to map the ID_REF values to their corresponding gene symbol.

Following this conversion, the three modified 'raw' files (each with two columns: Gene Symbol and gene expression level, were then combined into one master file **"all_mutants_transcriptomic_modified.txt"** with the following four columns:

1. gene symbol
2. log2(wt_ Δ***ppk1***) – relative gene expression level of wild type vs ppk1 mutant
3. log2(wt_ Δ***ppx***) - relative gene expression level of wild type vs ppx mutant
4. log2(wt_ Δ***polyP***) – relative gene expression level of wild type vs (ppk1 + ppx) mutant

This file "all_mutants_transcriptomic_modified.txt" is used as input for this guide.

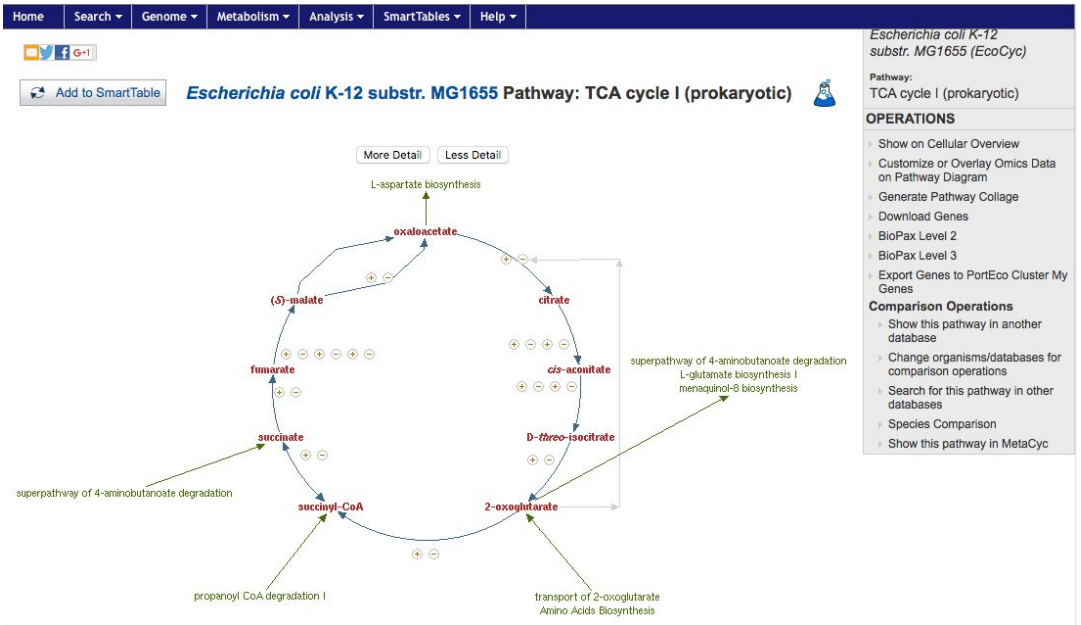
4. Step-by-step guide

Varas et al (2017) noted that expression of genes associated with the TCA cycle are significantly overrepresented in all three mutants, therefore we focus on the TCA pathway as an example to demonstrate the overlay feature.

Step	Instruction	Comments
1	Go to http://abrpi.genome.edu.au:9655/	Note: this is a private hosted instance of Pathway Tools that is part of the larger OMICs platform (http://omics.data.edu.au/use/).
2	Select **"Escherichia coli K-12 substr. MG1655 (EcoCyc)"** as the organism database by clicking the **"change organism database"** under the search column at the top right corner of the webpage.	The E.coli K12 MG1655 database contains the gene identifier and metabolic pathway information for E.coli Strain K12. This has been prepared previously by EcoCyc Staff.
3	Click on **"Pathways"** (in table #3)	Navigation to and selection of the (prokaryotic) TCA cycle pathway.
4	Click on **"Generation of Precursor Metabolites and Energy (42 instances)"**	

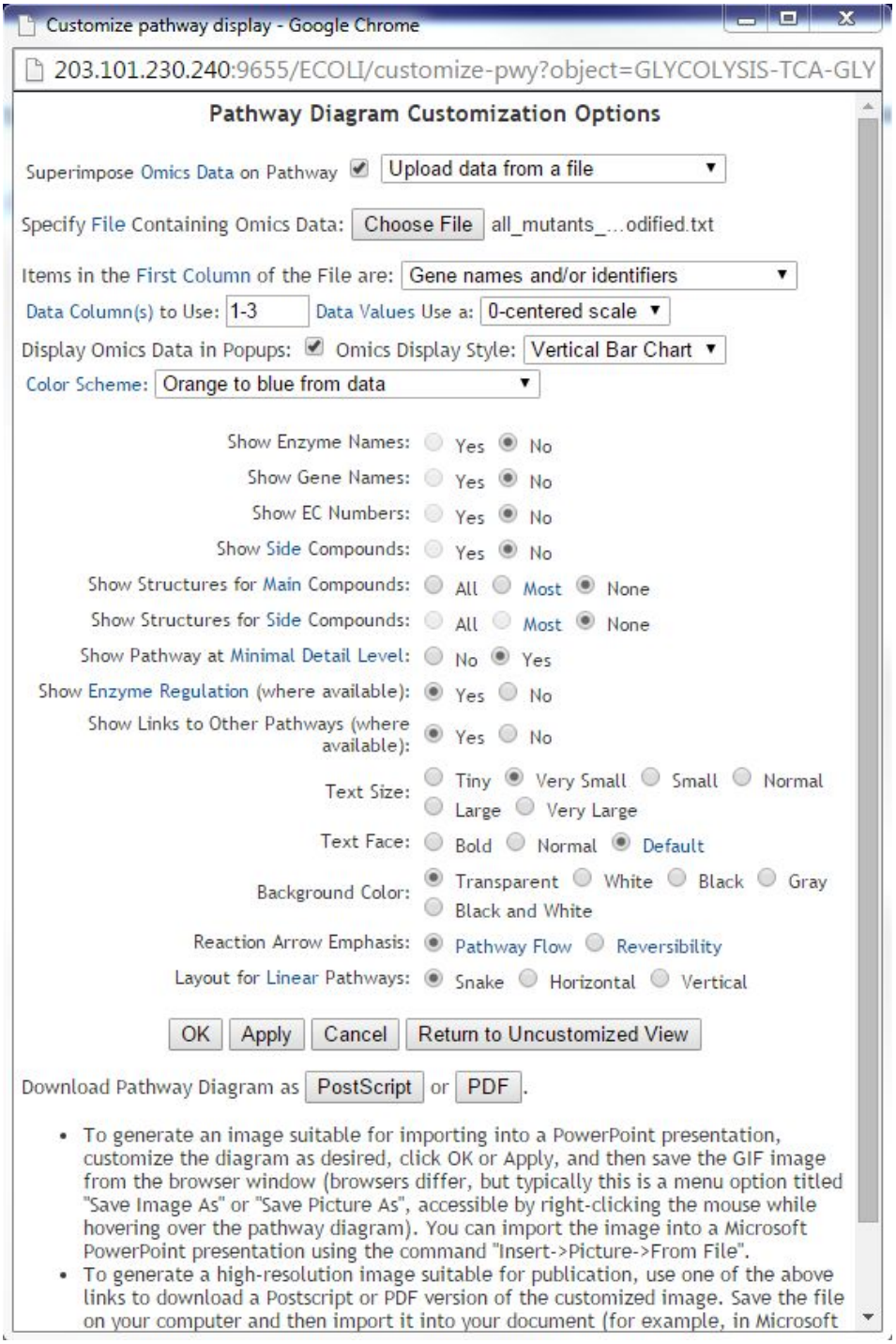
5	Click on **"TCA cycle(2)"**	
6	Click on **"TCA cycle I (prokaryotic)"**	

The TCA cycle will be loaded:



7	In the Operations Panel, click on **"Customize or Overlay Omics Data on Pathway Diagram"**	This option is found in the grey menu on the right hand side of the window.

A new window will appear:



The following are **mandatory** settings:

8	**Superimpose Omics Data on Pathway:** check this box	
9	**Specify File Containing Omics Data:** select the file containing the gene symbols and expression data	In this example, use the **all_mutants_transcriptomic_modified.txt** file. Navigate to a local copy and upload it.

10	**Items in the First Column of the File are:** Gene names and/or Identifiers	As noted in section 4.1, the first column (column "0") of this file contain gene symbols whereas columns 1-3 contain gene expression levels of the 3 mutants relative to wildtype E.coli K12.
11	**Data Column(s) to Use:** 1-3	
12	**Data Values Use a:** 0-centered scale	Implies that the numerical data of your file can contain positive and negative values. The value 0 is considered to be the centre of the numerical values provided in your data file.
13	**Display Omics Data in Popups:** checked	
14	**Omics Display style:** Vertical Bar Chart	
15	Leave all other fields as their default settings.	
16	Click on **Apply**	

When the file has been uploaded, you should see something similar this Figure 2:

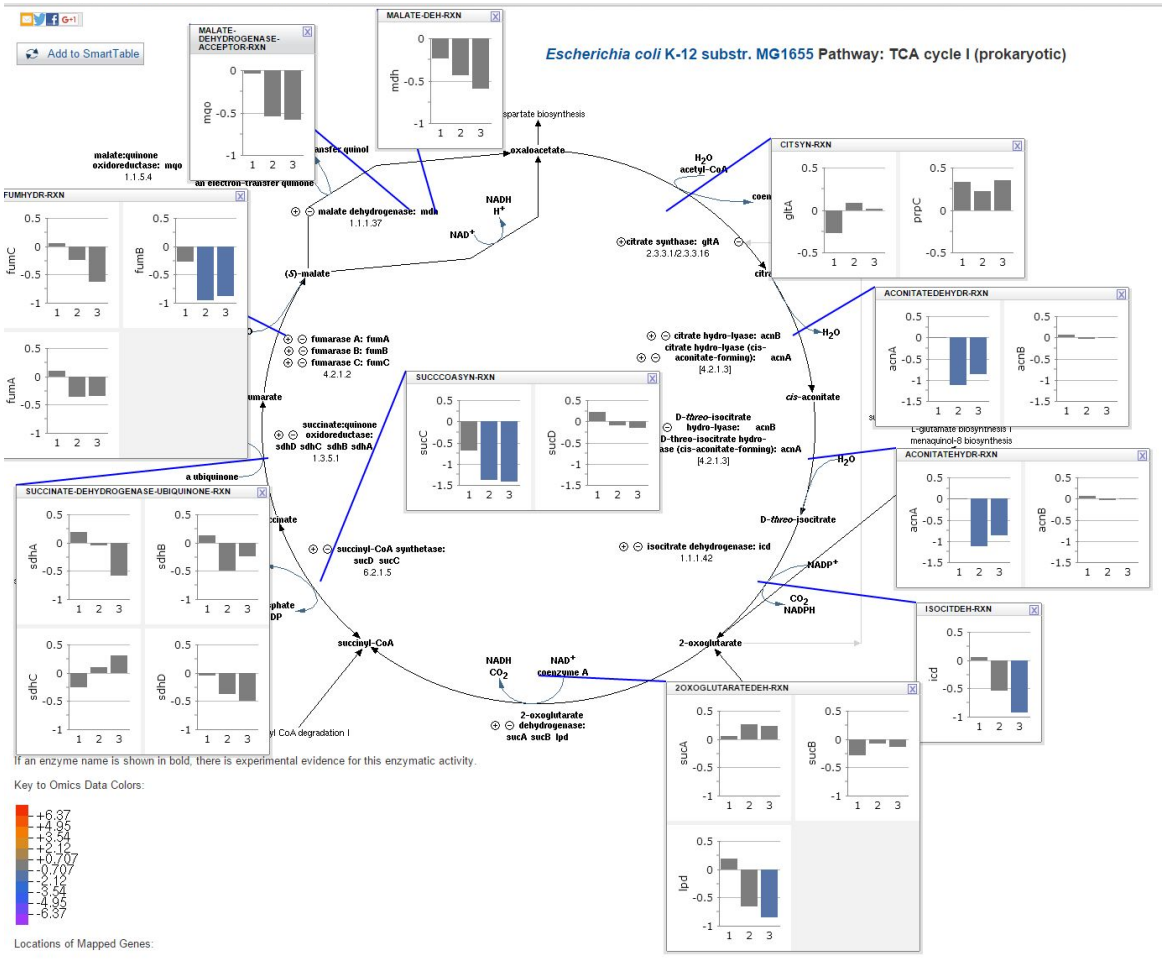
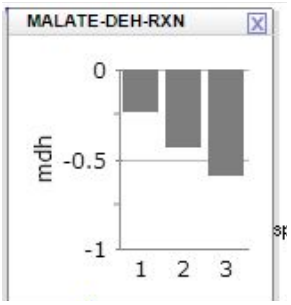


Figure 2 TCA cycle I (prokaryotic) with expression barplot.

Figure 2 shows the overlay of the transcriptomic data (all_mutants_transcriptomic_modified.txt) on the TCA cycle I pathway. Each bar plot shows the expression value on the log2 scale (as was provided by the input dataset) for the corresponding gene(s). Each vertical bar is the expression of the gene from the different experimental condition. For example, for the “mdh” gene:

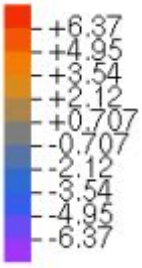


- Column one: wild type vs ppk1 mutant
- Column two: wild type vs ppx mutant
- Column three: wild type vs ppk1 + ppx mutant

The number of bars that are visible depends on the number of columns in the data file being uploaded.

You can change which columns in the data to show by modifying the **“Data Column(s) to Use”** field in step 11.

The colour of the bars represents the expression range as shown by the colour legend at the bottom. Grey bars indicate values in the range between +0.707 and -0.707 and blue bars indicate values in the range between -0.707 and -2.12.



If you need further assistance please contact us at omicsdataservices@lists.unimelb.edu.au