

- Find an article of interest, a dataset at the Gene Expression Omnibus (GEO), <http://www.ncbi.nlm.nih.gov/geo/>, or prepare your own data. In this example, we are using a paper (PMID: 6626502) that examines transcriptional responses in *Escherichia coli* MG1655 to a variety of stimuli. This strain has been designated as a reference strain by NCBI.

Open Access

Method

## A classification based framework for quantitative description of large-scale microarray data

Dipen P Sangurdekar<sup>\*†</sup>, Friedrich Srie<sup>nc\*</sup> and Arkady B Khodursky<sup>†\*</sup>

Addresses: <sup>\*</sup>Department of Chemical Engineering and Materials Science, University of Minnesota, Saint Paul, MN 55108, USA. <sup>†</sup>Biotechnology Institute, University of Minnesota, Saint Paul, MN 55108, USA. <sup>‡</sup>Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Saint Paul, MN 55108, USA.

Correspondence: Arkady B Khodursky. Email: khodu001@umn.edu

Published: 20 April 2006  
 Received: 11 November 2005  
 Genome Biology 2006, 7:R32 (doi:10.1186/gb-2006-7-4-r32)  
 Revised: 25 January 2006  
 The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2006/7/4/R32>  
 Accepted: 15 March 2006

© 2006 Sangurdekar et al.; licensee BioMed Central Ltd.  
 This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

---

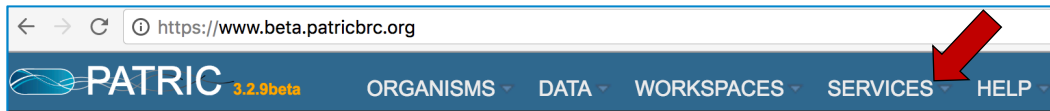
### Abstract

Genome-wide surveys of transcription depend on gene classifications for the purpose of data interpretation. We propose a new information-theoretical-based method to: assess significance of co-expression within any gene group; quantitatively describe condition-specific gene-class activity; and systematically evaluate conditions in terms of gene-class activity. We applied this technique to describe microarray data tracking *Escherichia coli* transcriptional responses to more than 30 chemical and physiological perturbations. We correlated the nature and breadth of the responses with the nature of perturbation, identified gene group proxies for the perturbation classes and quantitatively compared closely related physiological conditions.

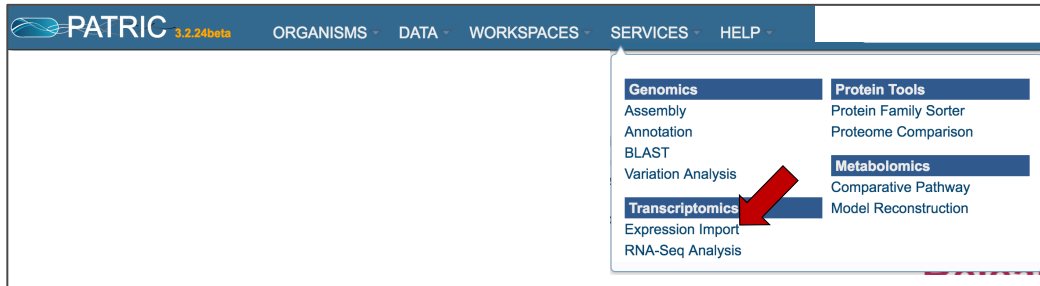
- The Sangurdekar paper has a link to all the log expression data for 3046 genes across 246 different conditions.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	UID	NAME	5min	15min	30min	60min	120min	180min	240min	300min	360min	420min	480min	540min	600min	660min	720min	780min	840min	900min	960min	1020min
2	B00001	thrL	0.784	0.522	0.538	0.357	-0.273	0.088	-0.438	-0.505	-0.396	-0.222	-0.129	-0.199	-0.091	0.391	0.093	0.841	0.594	0.316	-0.247	0.2
3	B00002	thrA	0.215	-0.037	0.005	0.044	0.818	-2.409	-1.653	-1.929	-1.1	-0.681	-0.095	-0.331	-1.172	0.058	-0.007	-0.167	-0.023	0.01	-0.47	-0.4
4	B00003	thrB	0.193	0.155	0.229	0.159	0.728	-0.545	-1.156	-0.921	-1.129	-0.662	-0.346	-0.454	-1.017	0.135	0.093	0.19	0.354	0.448	-0.043	-0.4
5	B00004	thrC	0.42	-0.071	0.127	0.108	0	-1.793	-1.587	-1.582	-1.268	-0.888	-0.392	-0.48	-0.882	0.221	0.15	0.219	0.21	0.454	-0.294	-0.1
6	B00006	yaaA	-0.008	0.299	0.172	0.085	0.626	0.832	0.273	0.327	-0.073	-0.015	0.062	0.11	0.007	0.493	0.137	0.306	0.136	0.159	0.384	0.1
7	B00008	talB	2.032	1.656	1.854	1.726	-0.535	0.813	0.678	0.648	0.232	0.291	0.39	0.418	0.553	0.477	0.448	-0.104	-0.877	-0.807	0.799	0.4
8	B00009	mog	0.571	0.421	0.155	0.199	-0.369	-0.333	-0.192	-0.033	0.128	0.082	0.056	-0.015	0.236	-0.058	-0.054	-0.208	-0.181	-0.126	0.051	0.1
9	B00010	yaaH	-1.796	-1.308	-1.206	-0.684	-0.185	0.21	0.405	0.28	0.093	0.132	0.041	-0.166	-0.023	0.302	0.234	0.297	0.342	0.161	0.442	0.1
10	B00012	hcgA	0.598	-0.273	-0.088	-0.135	-0.107	0.665	0.301	0.682	-0.176	0.081	-0.125	0.351	0.177	0.202	0.298	0.74	0.733	0.479	0.1	
11	B00013	yaaI	-0.803	0.406	-0.782	-0.791	0.361	-0.391	-0.225	-0.063	-0.072	-0.051	0.051	-0.041	-0.005	0.312	0.194	0.344	0.275	0.145	0.185	0.1
12	B00014	dnaK	1.757	1.358	1.946	1.73	2.936	-1.895	-1.737	-1.899	-0.122	0.662	0.73	0.73	0.477	1.023	1.6	1.181	0.245	-0.516	1.217	1.1
13	B00015	dnaJ	1.248	1.322	1.239	1.068	2.342	-0.069	-0.99	-0.525	-0.175	0.34	0.297	0.219	0.081	0.567	1.3	1.214	0.204	0.095	0.968	1.1
14	B00016	yB1_1	0.158	-0.95	-0.104	0	1.662	-0.572	-0.252	-0.641	-0.134	-0.113	-0.083	-0.138	-0.221	-0.902	-0.132	-0.452	-0.622	-0.596	0.102	0.4
15	B00017	yB2_1	0.479	-0.255	-0.304	0.051	0.919	-0.113	0.007	-0.049	-0.318	-0.19	-0.115	-0.33	-0.241	-0.646	-0.2	-0.413	-0.585	-0.428	-0.321	-0.1
16	B00018	gcf	0.823	-0.111	-0.059	0.009	0.112	-0.589	-0.295	0	0.064	-0.033	0.474	-0.154	0.054	-0.269	-0.114	-0.535	-0.153	-0.435	-0.004	0.1
17	B00019	nhaA	-0.154	-0.343	-0.612	-0.432	1.406	0.442	0.445	0.828	0.346	0.136	0.135	0.309	0.171	0.991	0.88	0.864	0.746	0.749	1.049	0.4
18	B00020	nhaB	-0.33	-0.558	-0.761	-0.563	0.327	0.325	0.144	0.604	-0.249	-0.417	-0.359	-0.469	-0.459	0.446	0.291	-0.244	0.535	0.098	0.373	0.1
19	B00021	nhaB_1	-0.02	0.006	-0.104	-0.102	-0.786	0.939	1.053	1.181	-0.118	-0.221	-0.267	-0.401	-0.287	-1.391	-1.003	-1.03	-0.655	-0.397	-0.276	-0.1
20	B00022	nha_1	0.11	-0.219	-0.127	-0.425	-0.447	0.564	0.413	0.693	-0.069	-0.22	-0.392	-0.409	-0.187	-0.843	-0.91	-0.94	-0.469	-0.216	-0.5	-0.1
21	B00023	rpsT	1.555	1.707	1.195	0.878	0.55	0.212	0.277	0.203	0.222	0.213	0.275	0.287	0.377	0.478	0.221	0.568	0.445	0.416	-0.229	-0.1
22	B00025	rifB	0.612	0.337	0.552	0.447	0.718	0	-0.124	-0.23	0.236	0.261	0.372	0.449	0.262	-0.35	-0.234	-0.715	-0.459	-0.37	-0.196	0
23	B00026	lcs	0.485	0.398	0.182	0.022	1.34	-0.242	-0.424	-0.507	-0.219	0.33	0.469	0.595	0.451	-0.412	-1.225	-1.222	-1.837	-1.384	-0.42	-0.1
24	B00027	tpaA	-0.013	0.069	-0.394	-0.468	0.151	-0.552	-0.251	-0.596	-0.778	-0.114	0.104	0.082	0.056	0.051	-0.256	-0.899	-0.709	-0.313	-0.156	-0.4
25	B00030	yaaF	0.088	0.072	0.09	0.062	-0.635	-0.592	0	0	-0.065	-0.055	0.002	-0.009	0.032	0.32	0.185	0.352	0.305	0.035	-0.057	0.1
26	B00031	dapB	2.1	2.056	1.217	1.748	-0.224	-0.138	-0.436	-0.631	-0.089	-0.235	-0.225	-0.272	-0.197	0.034	0.073	-0.579	-0.227	0.118	-0.491	-0.4
27	B00032	carA	2.309	2.381	2.833	2.773	-0.235	-0.504	-0.904	-1.189	0.8	0.645	-0.278	-0.229	-0.006	-1.896	-1.868	-2.028	-2.092	-2.531	-2.003	-0.4
28	B00033	carB	2.218	1.536	1.49	1.74	-0.795	-0.673	-0.666	-1.143	0.129	0.324	-0.312	0.006	-0.091	-1.436	-1.3	-1.604	-1.942	-1.996	-0.591	0.1
29	B00034	cafF	1.423	0.883	0.63	0.795	-0.003	-0.118	-0.212	-0.264	-0.072	0.083	0.114	0.11	0.234	-0.417	-0.215	-0.514	-0.277	-0.136	-0.757	-0.1
30	B00035	calE	1.39	0.704	0.508	0.525	0.008	-0.053	-0.07	0.108	0.034	0.076	0.031	0.311	0.317	-0.063	0.23	0.003	0.248	0.295	-0.179	0.1
31	B00036	calD	0.951	0.381	0.096	0.087	0.473	-0.48	-0.399	-0.409	0.157	0.033	0.136	0.175	0.21	-0.44	0.077	0.325	0.549	0.578	-0.172	0.1
32	B00037	calC	-0.751	-0.528	-0.894	-0.825	-0.083	-0.516	-0.389	-0.513	0.062	-0.092	-0.122	-0.217	0.001	0.559	0.266	0.563	0.596	0.719	0.26	0.1
33	B00038	calB	-0.111	-0.43	-0.899	-0.771	-0.383	-0.562	-1.137	-0.24	0.166	0.008	0.025	0.077	0.371	0.117	0.161	0.218	0.321	0.331	0.263	0.1
34	B00041	flaA	0.263	0.083	0.299	0.147	-0.078	-0.828	-0.599	-1.002	0.16	0.058	0.152	0.138	0.137	0.578	0.295	0.667	0.587	0.535	0.207	0.1
35	B00047	kefC	-1.222	-1.308	-1.549	-1.2	0.071	0.19	0.177	0.05	-0.126	-0.078	0.022	0.007	-0.148	0.532	0.174	0.499	0.473	0.516	0.131	-0.4
36	B00048	foiA	-0.436	-0.404	-0.561	-0.582	0.497	0.618	0.142	0.384	0.138	0.043	0.133	0.416	0.109	-0.358	0.048	-0.101	-0.064	-0.133	-0.469	0.1
37	B00049	foiB	-0.413	-0.261	-0.363	-0.083	0.136	0.614	-0.136	-0.363	-0.198	-0.043	-0.133	0.063	-0.093	-0.096	-0.070	-0.333	-0.364	-0.133	-0.143	0.1

3. To upload this data, open the Services tab in PATRIC and click on Expression Import



4. 2. Click Experiment Import.



5. This will take you to the page where you can upload your own experimental data, or other data that you have found.

Differential Expression Import  
Transform differential expression data for viewing on PATRIC

Experiment Data ⓘ	Experiment Information ⓘ
EXPERIMENT DATA FILE <input type="text"/>	EXPERIMENT TITLE <input type="text" value="Title"/>
EXPERIMENT TYPE <input type="text" value="Transcriptomics"/>	EXPERIMENT DESCRIPTION <input type="text" value="Description"/>
TARGET GENOME <input type="text" value="Optional"/>	ORGANISM NAME <input type="text" value="e.g. Bacillus Cereus"/>
	PUBMED ID <input type="text" value="Optional"/>
	OUTPUT FOLDER <input type="text" value="Experiments"/>
Optional Metadata ⓘ	
METADATA FILE <input type="text"/>	
<input type="button" value="Reset"/> <input type="button" value="Submit"/>	

6. The data should be loaded into a particular format into PATRIC, an example of which can be seen below (A). The first column has the gene locus tag (RefSeq, GenBank, or PATRIC), and the second and later columns will have the log ratios for each gene in each condition. The column headers in the first row can have the name of that condition. You can see in B below that the second column has the gene name. That must be deleted before it can be uploaded.

## 1. Gene Matrix:

**A**

Gene ID	Comparison1	Comparison2
b0002	0.767	-1.316
b0003	0.815	-1.841
b0004	0.856	-1.643


	A	B	C	D	E	F	G	H
1	UID	NAME	5min .02M S	15min .02M	300min .02M	600min .02M	C410+0Ecb0	10' trp star
2	B0001	thrL	0.784	0.522	0.528	0.357	-0.273	0.088
3	B0002	thrA	0.215	-0.037	0.005	0.044	0.818	-2.409
4	B0003	thrB	0.193	0.155	0.229	0.159	0.728	-0.545
5	B0004	thrC	0.42	-0.071	0.127	0.108	0	-1.793
6	B0006	yaaA	-0.008	0.299	0.172	0.085	0.626	0.832
7	B0008	talB	2.032	1.656	1.854	1.726	-0.535	0.813
8	B0009	mog	0.571	0.421	0.155	0.199	-0.369	-0.333
9	B0010	yaaH	-1.796	-1.308	-1.206	-0.684	-0.185	0.21
10	B0012	htgA	0.598	-0.273	-0.088	-0.135	-0.107	0.665
11	B0013	yaaI	-0.803	-0.406	-0.782	-0.791	0.361	-0.391
12	B0014	dnaK	1.757	1.358	1.946	1.73	2.936	-1.895
13	B0015	dnaJ	1.248	1.322	1.239	1.068	2.342	-0.069
14	B0016	yi81_1	0.158	-0.95	-0.104	0	1.662	-0.572
15	B0017	yi82_1	-0.479	-0.255	-0.304	0.051	0.919	-0.113
16	B0018	gef	0.823	-0.111	-0.059	0.009	0.112	-0.589
17	B0019	nhaA	-0.154	-0.343	-0.612	-0.432	1.406	0.442
18	B0020	nhaR	-0.33	-0.558	-0.761	-0.563	0.327	0.325
19	B0021	insB_1	-0.02	0.006	-0.104	-0.102	-0.786	0.939

**B**

- To upload the data sheet, which can either be txt or excel, click on the folder in the “Experiment Data” box that follows the text box under the words “Experiment Data File”.

Differential Expression Import  
Transform differential expression data for viewing on F


**Experiment Data** ⓘ

EXPERIMENT DATA FILE  
 

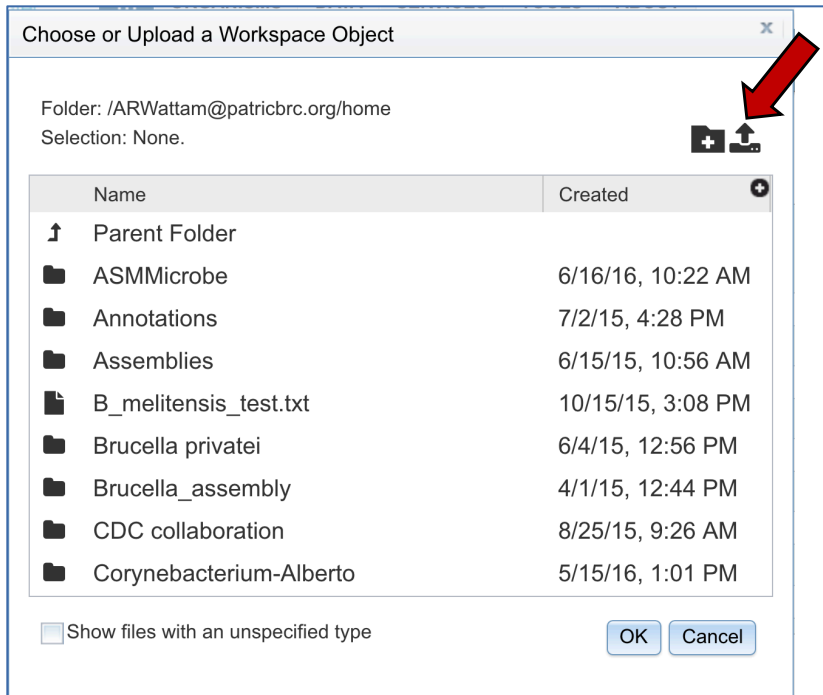
EXPERIMENT TYPE  
Transcriptomics

TARGET GENOME  
Optional

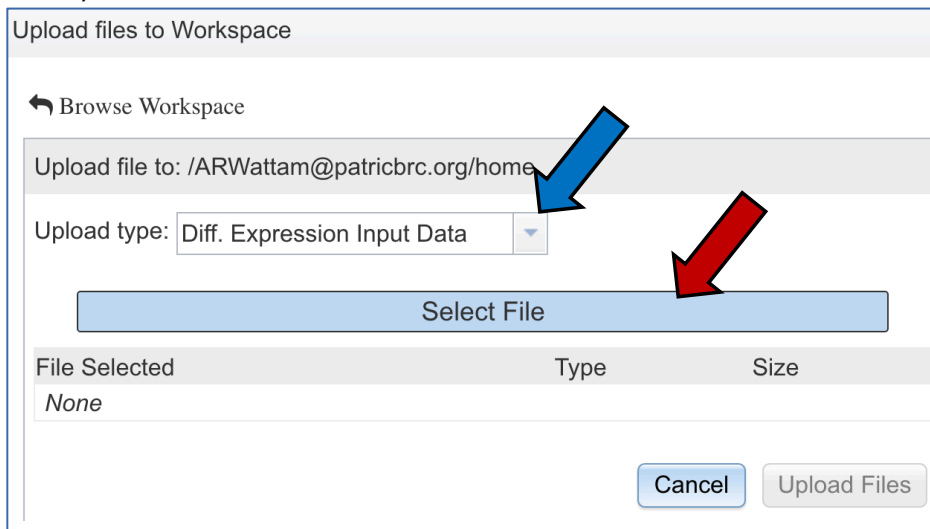
**Optional Metadata** ⓘ

METADATA FILE  
 

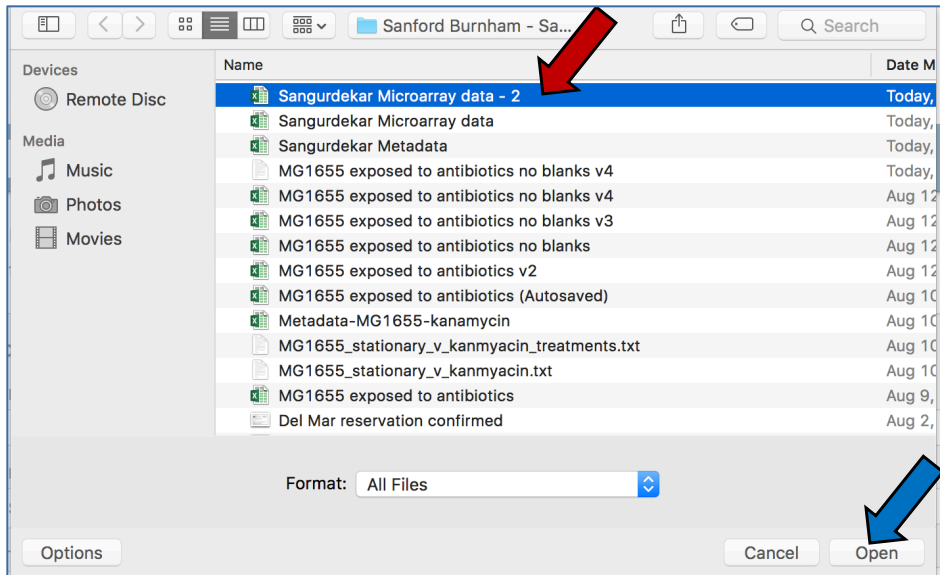
- This will open a window that shows you data you have in uploaded or collected in PATRIC. If you had previously uploaded this data, you could find it here. To upload new data, click on the upload icon.



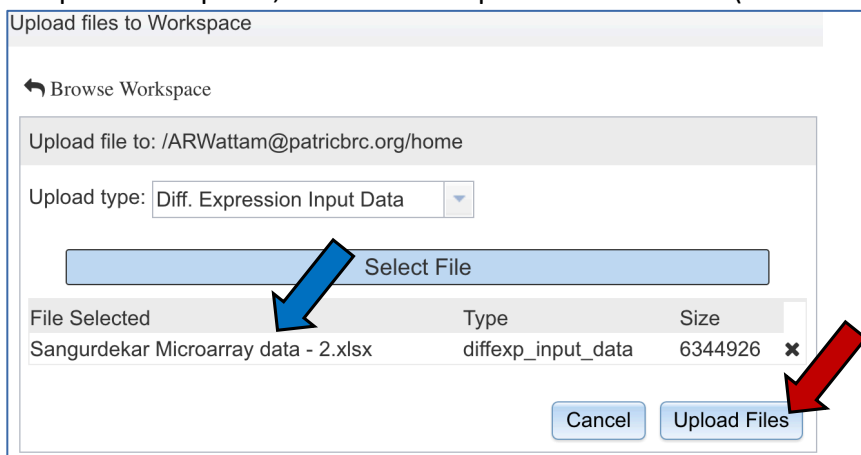
9. The pop-up widow will flip, and you will be able to select a file. Make sure you have the right type of data selected (Blue arrow). Then click on the blue “Select File” bar (Red arrow).



10. Select the file you want to upload (Red arrow) and then open it (Blue arrow).



11. The name of the file you selected will now appear on the page (Blue arrow). To complete the upload, click on the “Upload Files” button (Red arrow).



12. The next task is to select the target genome. It has to be the same as the genome in the experimental data you uploaded. In this instance, the reference genome was Escherichia coli MG1655. If you start typing MG1655 in the text box under “Target Genome” (Red arrow), PATRIC will display the best hits. There are quite a few there...how do you know that the one shown as selected (Blue arrow) is the right one.

### Experiment Data

EXPERIMENT DATA FILE

Sangurdekar Microarray data - 2.xlsx

EXPERIMENT TYPE

Transcriptomics

TARGET GENOME

Optional

### Experiment Data

EXPERIMENT DATA FILE

Sangurdekar Microarray data - 2.xlsx

EXPERIMENT TYPE

Transcriptomics

TARGET GENOME

MG1655

Escherichia coli strain K-12 substrain MG1655 [511145.209]

Escherichia coli str. K-12 substr. MG1655 JW5437-1 [511145.209]

Escherichia coli str. K-12 substr. MG1655 K-12 [511145.208]

Escherichia coli str. K-12 substr. MG1655star [879462.4]

Escherichia coli str. K-12 substr. MG1655 K-12 [511145.180]

Escherichia coli str. K-12 substr. MG1655 K-12 [511145.181]

Escherichia coli str. K-12 substr. MG1655 K-12 [511145.182]

Escherichia coli str. K-12 substr. MG1655 K-12 [511145.183]

Escherichia coli str. K-12 substr. MG1655 K-12 [511145.184]

Escherichia coli str. K-12 substr. MG1655 K12 [511145.12]

### Experiment

EXPERIMENT TITLE

Title

EXPERIMENT DESCRIPTION

Description

ORGANISM NAME

Organism name

13. The number that follows the name is the genome ID. If you click the Organism tab at the top of any PATRIC page, and then click on *Escherichia*, you will go to the landing page for that genus. Clicking on genome list will show all the genomes in *Escherichia*. MG1655 is a reference genome. You can find the genome ID by clicking on “Reference” in the filter (Red arrow). This will only show the reference genomes in the table

Taxon View

Bacteria » Proteobacteria » Gammaproteobacteria » Enterobacteriales » Enterobacteriaceae » **Escherichia** ( 4837 Genomes )

Overview

Phylogeny

Taxonomy

Genomes

Sequences

Features

Specialty Genes

Protein Families

Pathways

Transcriptomics

DOWNLOAD

KEYWORDS

REFERENCE

REFERENCE\_GENOME x

HIDE

APPLY

Collection Year	Disease	Genome Status	Isolation Country	Public	Reference Genome
1922 (1)	Urinary tract infection (2)	Complete (6)	Japan (1)	true (6)	Reference (6)
1997 (1)	Hemolytic uremic syndrome (HUS) (1)				Representative (7)
	Hemorrhagic colitis (1)				

Genome Name	Genome ID	Genome Status	Sequence	PATRIC CDS	Isolation Country	Host Name	Disease	Collection Year	Completion Date
Escherichia coli IA139	585057.6	Complete	1	5094			Urinary tract infection		12/17/08
Escherichia coli O104:H4 str. 201	1133852.3	Complete	4	5479			Hemolytic uremic syndrome		9/26/12
Escherichia coli O83:H1 str. NRG	685038.3	Complete	2	4686					11/29/10
Escherichia coli UMN026	585056.7	Complete	3	5198			Urinary tract infection		12/17/08
Escherichia coli O157:H7 str. Sak	386585.9	Complete	3	5602	Japan		Hemorrhagic colitis	1997	3/28/00
Escherichia coli str. K-12 substr. MG1655	511145.12	Complete	1	4549				1922	9/4/97

14. Click on the MG1655 genome (Red Arrow). This will load the vertical green bar with all the downstream processes that you can do with this genome. Click on the Genome icon (Blue arrow).

Overview Phylogeny Taxonomy **Genomes** Sequences Features Specialty Genes Protein Families Pathways Transcriptomics

DOWNLOAD KEYWORDS REFERENCE REFERENCE\_GENOME x HIDE APPLY


Collection Year	Disease	Genome Status	Isolation Country	Public	Reference Genome
1922 (1)	Urinary tract infection (2)	Complete (6)	Japan (1)	true (6)	Reference (6)
1997 (1)	Hemolytic uremic syndrome (HUS) (1)				Representative (7)
	Hemorrhagic colitis (1)				

→

Genome Name	Genome ID	Genome Status	Sequenced	PATRIC CDS	Isolation Country	Host Name	Disease	Collection Year	Completion Date
<input type="checkbox"/> Escherichia coli IAI39	585057.6	Complete	1	5094			Urinary tract infection		12/17/08
<input type="checkbox"/> Escherichia coli O104:H4 str. 2011	1133852.3	Complete	4	5479			Hemolytic uremic syndrome		9/26/12
<input type="checkbox"/> Escherichia coli O83:H1 str. NRG	685038.3	Complete	2	4686					11/29/10
<input type="checkbox"/> Escherichia coli UMN026	585056.7	Complete	3	5198			Urinary tract infection		12/17/08
<input type="checkbox"/> Escherichia coli O157:H7 str. Sak	386585.9	Complete	3	5602	Japan		Hemorrhagic colitis	1997	3/28/00
<input checked="" type="checkbox"/> Escherichia coli str. K-12 substr. MG1655	511145.12	Complete	1	4549				1922	9/4/97

1 - 6 of 6 results

15. This will take you to the landing page for that genome, where you can see the specific genome ID (red box).

 **Escherichia coli str. K-12 substr. MG1655**

Length: 4639675bp, Chromosomes: 1, Plasmids: 0, Contigs: 0

**Organism Info**

Genome ID	511145.12
Genome Name	Escherichia coli str. K-12 substr. MG1655
NCBI Taxon ID	511145
Genome Status	Complete
Strain	K12
MLST	MLST.Escherichia_coli_1.10
Reference	Reference
Genome	

16. Next you will need to fill in all the metadata in the “Experiment Information” box (Red arrow). The PubMed ID is optional. When the information is filled out, click the submit button at the bottom of the page (Blue arrow).

**Differential Expression Import**  
Transform differential expression data for viewing on PATRIC

**Experiment Data**

EXPERIMENT DATA FILE  
Sangurdekar Microarray data - 2.xlsx

EXPERIMENT TYPE  
Transcriptomics

TARGET GENOME  
Escherichia coli str. K-12 substr. MG

**Optional Metadata**

METADATA FILE

**Experiment Information**

EXPERIMENT TITLE  
Sangurdekar MG1655 data

EXPERIMENT DESCRIPTION  
MG1655 stationary v stress conditions

ORGANISM NAME  
Escherichia coli str. K-12 substr. MG1655

PUBMED ID  
6626502

OUTPUT FOLDER  
Experiments

Reset Submit

17. A message appears underneath indicating that the data has been submitted (Red arrow).

Sangurdekar Microarray data - 2.xlsx

EXPERIMENT TYPE  
Transcriptomics

TARGET GENOME  
Escherichia coli str. K-12 substr. MG

**Optional Metadata**

METADATA FILE

Title

EXPERIMENT DESCRIPTION  
MG1655 stationary v stress conditions

ORGANISM NAME  
Escherichia coli str. K-12 substr. MG1655

PUBMED ID  
6626502

OUTPUT FOLDER  
Experiments

Reset Submit

Conversion should be finished shortly. Check workspace for results.

18. Clicking on the Jobs button at the bottom of the page will show you the status of the job you submitted.

<div> <div>Uploads 0/0</div> <div>Jobs 138/0/24</div> </div>						
Status	Submit	App	Output Name	Start	Completed	
completed	8/18/16, 9:23 AM	DifferentialExpression	Sangurdekar MG1655 data	8/18/16, 9:23 AM	8/18/16, 9:24 AM	
completed	8/17/16, 6:19 PM	DifferentialExpression	MG 1655 exposed to kanamycin	8/17/16, 6:19 PM	8/17/16, 6:19 PM	
completed	8/16/16, 11:05 AM	Proteome Comparison	B microti v GenBank microti	8/16/16, 11:05 AM	8/16/16, 11:06 AM	
completed	8/15/16, 3:57 PM	DifferentialExpression	MG1655 exposed to antibiotics 5	8/15/16, 3:57 PM	8/15/16, 3:57 PM	
completed	8/15/16, 3:08 PM	DifferentialExpression	MG1655 exposed to kanamycin 3	8/15/16, 3:08 PM	8/15/16, 3:09 PM	
completed	8/12/16, 7:48 AM	DifferentialExpression	MG1655 exposed to kanamycin v2	8/12/16, 7:48 AM	8/12/16, 7:49 AM	
completed	8/12/16, 7:40 AM	DifferentialExpression	MG1655 exposed to	8/12/16, 7:40 AM	8/12/16, 7:41 AM	