RNASeq Studies IMG

Introduction

IMG contains **RNASeq** data from a variety of studies (Fig. 1)**.** This data can be explored by going to the OMICS -> RNASeq menu.

RNASeq Expression Studies

Studies | Studies by Ref. Data Set

Study Name (Proposal Name)	Data Set Type	Total Ref. Data Sets	Total RNA Seq Data Sets
Active sludge and wastewater microbial communities from Klosterneuburg, Austria	Metatranscriptome	25	25
Active sludge microbial communities of municipal wastewater-treating anaerobic digesters from various locations	Metatranscriptome	30	30
Anaerobic biogas reactor microbial communites from Washington, USA	Metatranscriptome	4	4
Anoxygenic and chlorotrophic microbial mat microbial communities from Yellowstone National Park, USA	Metatranscriptome	98	763
Aqueous microbial communities from the Delaware River and Bay under freshwater to marine salinity gradient to study organic matter cycling in a time-series	Metatranscriptome	59	59
Arabidopsis, maize, boechera and miscanthus rhizosphere microbial communities from different US locations	Metatranscriptome	32	216
Arctic peat soil microbial communities from the Barrow Environmental Observatory site, Barrow, Alaska, USA	Metatranscriptome	5	5
Avena fatua rhizosphere microbial communities from Hopland, California, USA	Metatranscriptome	59	850
Avena fatua rhizosphere microbial communities from Hopland, California, USA, for root-enhanced decomposition of organic matter studies	Metatranscriptome	61	929
Bacterial and archaeal communities from various locations to study Microbial Dark Matter (Phase II)	Metatranscriptome	1	1
Bacteroides fragilis 610 transcriptome (610_1)	Transcriptome	1	1
Bacteroides fragilis 610 transcriptome (610_2)	Transcriptome	1	1
Bacteroides fragilis 638R transcriptome (638R_1)	Transcriptome	1	1
Bacteroides fragilis 638R transcriptome (638R_2)	Transcriptome	1	1
Biogas fermentation microbial communities from biogas plants in Germany	Metatranscriptome	16	16
Blue grama grass rhizosphere eukaryotic and microbial communities from Sevilleta Long Term Ecological Research site, New Mexico, US	Metatranscriptome	1	82
Coastal salt marsh microbial communities from the Groves Creek Marsh, Skidaway Island, Georgia	Metatranscriptome	36	36
Compost Feedstock-Adapted Isolates	Metatranscriptome	1	17
Continental margin sediment microbial communities from various locations	Metatranscriptome	5	5
Coral microbial communities from various locations to study host-microbial communication	Metatranscriptome	24	24
Corn, switchgrass and miscanthus rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA	Metatranscriptome	8	8
Corynebacterium glutamicum codon depletion transcriptome	Transcriptome	1	6

Figure 1. List of all RNASeq Expression Studies

RNASeq data can be viewed either through the reference data sets (genomes) for a study or through the RNASeq data sets for a study. Each study can have one or many associated reference data sets (Fig. 2b). Each reference data set (or genome) for a given study has associated RNASeq data sets (Fig. 2a).

RNASeq Expression Studies

Ref. Data Set: Marine sediment microbial community from La Parguera, Puerto Rico - PR Tt Sediment 1 (PR Tt Sediment 1 - Bioluminescent Bay in La ParagueraPR, July 2012 Assem)

Select Samples | Single Sample Analysis | Pairwise Sample Analysis | Multiple Sample Analysis

hint: You may change the Normalization Method used to compute abundances of the genes in any given sample: Quantile Affine RPKM (Reads Per Kilobase of gene)

Select	GOLD ID	Data Set Name	Data Set ID	Study Name (Proposal Name)	Genes with Reads	Total Reads Count	Average Reads per Gene
	Gp0055727	Enrichment culture microbial communities from Fremont, California, USA - EucalyptusEnriched_D_E2P	530	Environmental microbial communities from Fremont, CA and La Paraguera, Puerto Rico	172	983	5.72
	Gp0055737	Enrichment culture microbial communities from Fremont, California, USA - PineEnriched_G_P2P	533	Environmental microbial communities from Fremont, CA and La Paraguera, Puerto Rico	168	1152	6.86
	Gp0055736	Enrichment culture microbial communities from Fremont, California, USA - PineEnriched_D_P2P	553	Environmental microbial communities from Fremont, CA and La Paraguera, Puerto Rico	11	21	1.91
	Gp0055738	Enrichment culture microbial communities from Fremont, California, USA - PineEnriched_G_P1P	554	Environmental microbial communities from Fremont, CA and La Paraguera, Puerto Rico	257	1795	6.98
	Gp0055744	Enrichment culture microbial communities from Fremont, California, USA - PineEnriched_D_P1P	606	Environmental microbial communities from Fremont, CA and La Paraguera, Puerto Rico	23	66	2.87

Select All | Clear All

Notes:

1 - Normalization Methods:
For additional information see: Quantile, Affine, RPKM, R doc

Figure 2a. RNASeq datasets associated with a genome for a selected study

Data Sets for RNASeq Study

Study Name (Proposal Name): Sequencing the transcriptome of cyanobacteria

Select Samples | Single Sample Analysis | Pairwise Sample Analysis | Multiple Sample Analysis

hint: You may change the Normalization Method used to compute abundances of the genes in any given sample: Quantile Affine RPKM (Reads Per Kilobase of gene)

Select All | Clear All

Filter column: Total Reads Count Filter text: Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector | Select Page | Deselect Page

Select	GOLD ID	Data Set Name	Data Set ID	Genes with Reads	Total Reads Count	Average Reads per Gene	Ref. Data Set
	Gp0060674	Synechococcus elongatus PCC 7942 transcriptome (Temperature 24 hours)	331	2713	14437225	5321.5	Synechococcus elongatus PCC 7942
	Gp0060671	Synechocystis sp. PCC 6803 transcriptome (light 1 hour)	344	3622	12454641	3438.61	Synechocystis sp. PCC 6803
	Gp0060684	Synechococcus elongatus PCC 7942 transcriptome (CO2 24 hours)	334	2713	12237501	4510.69	Synechococcus elongatus PCC 7942
	Gp0060662	Synechocystis sp. PCC 6803 transcriptome (pH 1 hour)	343	3622	11850679	3271.86	Synechocystis sp. PCC 6803
	Gp0060673	Synechococcus elongatus PCC 7942 transcriptome (Temperature 1 hour)	326	2713	11744286	4328.89	Synechococcus elongatus PCC 7942
	Gp0060667	Synechocystis sp. PCC 6803 transcriptome (Temperature 1 hour)	612	3622	11726524	3237.58	Synechocystis sp. PCC 6803
	Gp0060663	Synechocystis sp. PCC 6803 transcriptome (NaCl 24 hours)	338	3622	11475584	3168.3	Synechocystis sp. PCC 6803
	Gp0060665	Synechocystis sp. PCC 6803 transcriptome (CO2 24 hours)	346	3622	11294624	3118.34	Synechocystis sp. PCC 6803

Figure 2b. RNASeq datasets for a selected study, associated with multiple genomes

RNASeq expression data can also be explored starting with the **Organism Details** page for a genome (Fig. 3). The **Expression Studies** link, see bottom of Figure 3, leads to the list of studies for the genome of interest (Figure 4). The **genes with RNASeq data** link under Genome Statistics for the selected genome leads to a list of genes that have associated RNASeq data.

Synechococcus sp. PCC 7002

Add to Genome Cart | Browse Genome | BLAST Genome

About Genome

- Overview
- Statistics
- Expression Studies
- Methylomics Experiments
- Genes

Overview

Study Name (Proposal Name)	Synechococcus sp. PCC 7002
Organism Name	Synechococcus sp. PCC 7002
Taxon ID	641522654
NCBI Taxon ID	32049
GOLD ID in IMG Database	Study ID: Gs0033815 Project ID: Gp0002091
GOLD Analysis Project Id	Ga0031137
GOLD Analysis Project Type	Genome Analysis
Submission Type	Primary
JGI Analysis Project Type	Genome Analysis
External Links	NCBI/RefSeq:NC_010474; NCBI/RefSeq:NC_010475; NCBI/RefSeq:NC_010476; NCBI/RefSeq:NC_010477; NCBI/RefSeq:NC_010478; NCBI/RefSeq:NC_010479
Lineage	Bacteria; Cyanobacteria; unclassified; Synechococcales; Synechococcaceae; Synechococcus; Synechococcus sp. PCC 7002
Sequencing Status	Finished

Biosynthetic Clusters	13	-
Genes in Biosynthetic Clusters	78	2.41%
Fused Protein coding genes	210	6.49%
Protein coding genes coding signal peptides	126	3.89%
Protein coding genes coding transmembrane proteins	752	23.23%
Revised Genes	2	0.06%
Genes with RNASeq data	2823	
COG clusters	1329	69.98%
KOG clusters	449	23.64%
Pfam clusters	1862	73.89%
TIGRfam clusters	902	91.30%
Internal clusters	129	

Expression Studies

List of RNASeq Studies

RNASeq Differenital Expression Data

Figure 3. Genome page link to list of RNASeq Studies for this genome

RNASeq Expression Studies

Ref. Data Set: Synechococcus sp. PCC 7002

Select Samples | View in GBrowse | Single Sample Analysis | Pairwise Sample Analysis | Multiple Sample Analysis

hint: You may change the Normalization Method used to compute abundances of the genes in any given sample: Quantile Affine RPKM (Reads Per Kilobase of gene)

Select All | Clear All

Select	GOLD ID	Data Set Name	Data Set ID	Study Name (Proposal Name)	Genes with Reads	Total Reads Count	Average Reads per Gene
	Gp0060686	Synechococcus sp. PCC 7002 transcriptome (dark aerobic)	352	Transcription Profiling of the Model Cyanobacterium Synechococcus sp. Strain PCC 7002	2823	481909	170.71
	Gp0060687	Synechococcus sp. PCC 7002 transcriptome (dark anaerobic)	353	Transcription Profiling of the Model Cyanobacterium Synechococcus sp. Strain PCC 7002	2823	1033135	365.97
	Gp0060688	Synechococcus sp. PCC 7002 transcriptome (high light)	354	Transcription Profiling of the Model Cyanobacterium Synechococcus sp. Strain PCC 7002	2823	473131	167.6
	Gp0060689	Synechococcus sp. PCC 7002 transcriptome (standard1)	355	Transcription Profiling of the Model Cyanobacterium Synechococcus sp. Strain PCC 7002	2823	3005584	1064.68
	Gp0060690	Synechococcus sp. PCC 7002 transcriptome (standard2)	356	Transcription Profiling of the Model Cyanobacterium Synechococcus sp. Strain PCC 7002	2823	1470994	521.07

Figure 4. List of studies for a genome of interest

The **Gene Detail** page for a gene with associated RNASeq data has an **rnaseq** link (Fig. 5a) to expression studies for that gene (Fig. 5b), listing the experimental conditions and the reads count for each experiment.

Gene Detail

Gene Information
Find Candidate Product Name
Evidence For Function Predictions
Sequence Search
External Sequence Search
IMG Sequence Search
Homolog Display

Gene Information

Gene Information	
Gene ID	641610013
Gene Symbol	
Locus Tag	SYNPCC7002_A0002
IMG Product Name	periplasmic solute-binding protein of ABC transporter
Original Gene Product Name	periplasmic solute-binding protein of ABC transporter
IMG Product Source	
SwissProt Protein Product	
SEED	
IMG Term	
Description	similar to tll1975 of Thermosynechococcus elongatus
Genome	Synechococcus sp. PCC 7002
DNA Coordinates	1424..2353 (+)(930bp)
Scaffold Source	Synechococcus sp. PCC 7002: NC_010475 (3008047bp)

Protein Information	
Amino Acid Sequence Length	309aa
Fitness BLAST	Fitness BLAST searches for similar proteins that have mutant phenotypes. *Fitness BLAST is a 3rd party tool. Fitness BLAST Help* **View Fitness BLAST**
COG	COG0614 - ABC-type Fe3+-hydroxamate transport system, periplasmic component
KOG	
COG Function	Inorganic ion transport and metabolism
KOG Function	
EggNOG	
Families	- Coils:Coil - Gene3D:G3DSA:3.40.50.1980 - IPR002491 ABC transporter periplasmic binding domain (ProSiteProfiles PS50983) - SSF53807 "Helical backbone" metal receptor
Transmembrane Helices	No
Signal Peptide	No
Statistics	peptide
Expression	rnaseq
Pfam	pfam01497 - Peripla_BP_2
TIGRfam Role	

Figure 5a. Link to RNASeq studies from Gene Detail page

RNASeq Data for Gene

```
>Gene: 641610013 periplasmic solute-binding protein of ABC transporter
 [Synechococcus sp. PCC 7002: NC_010475]
```

Filter column: Data Set ID ▼ Filter text ▼: Apply

Export Page 1 of 1 << first < prev 1 next > last >> All ▼

Column Selector

Data Set ID ▲	Data Set Name	GOLD ID	Reads Count
352	Synechococcus sp. PCC 7002 transcriptome (dark aerobic)	Gp0060686	55
353	Synechococcus sp. PCC 7002 transcriptome (dark anaerobic)	Gp0060687	36
354	Synechococcus sp. PCC 7002 transcriptome (high light)	Gp0060688	16
355	Synechococcus sp. PCC 7002 transcriptome (standard1)	Gp0060689	221
356	Synechococcus sp. PCC 7002 transcriptome (standard2)	Gp0060690	114
357	Synechococcus sp. PCC 7002 transcriptome (standard3)	Gp0060691	127
358	Synechococcus sp. PCC 7002 transcriptome (OD 0.4)	Gp0060692	20
359	Synechococcus sp. PCC 7002 transcriptome (OD 1.0)	Gp0060693	46
360	Synechococcus sp. PCC 7002 transcriptome (OD 3.0)	Gp0060694	30
361	Synechococcus sp. PCC 7002 transcriptome (OD 5.0)	Gp0060695	40
362	Synechococcus sp. PCC 7002 transcriptome (low O2)	Gp0060696	20

Export Page 1 of 1 << first < prev 1 next > last >> All ▼

Figure 5b. Gene Details - RNASeq Expression Data for a gene

Sample Analysis

The samples for each RNASeq study (Fig. 2) can be analyzed in a number of ways (Fig. 6a). Samples can be selected to perform single, pairwise, or multiple sample analysis. Samples can also be analyzed in GBrowse to view differences in read coverage for individual genes under various conditions.

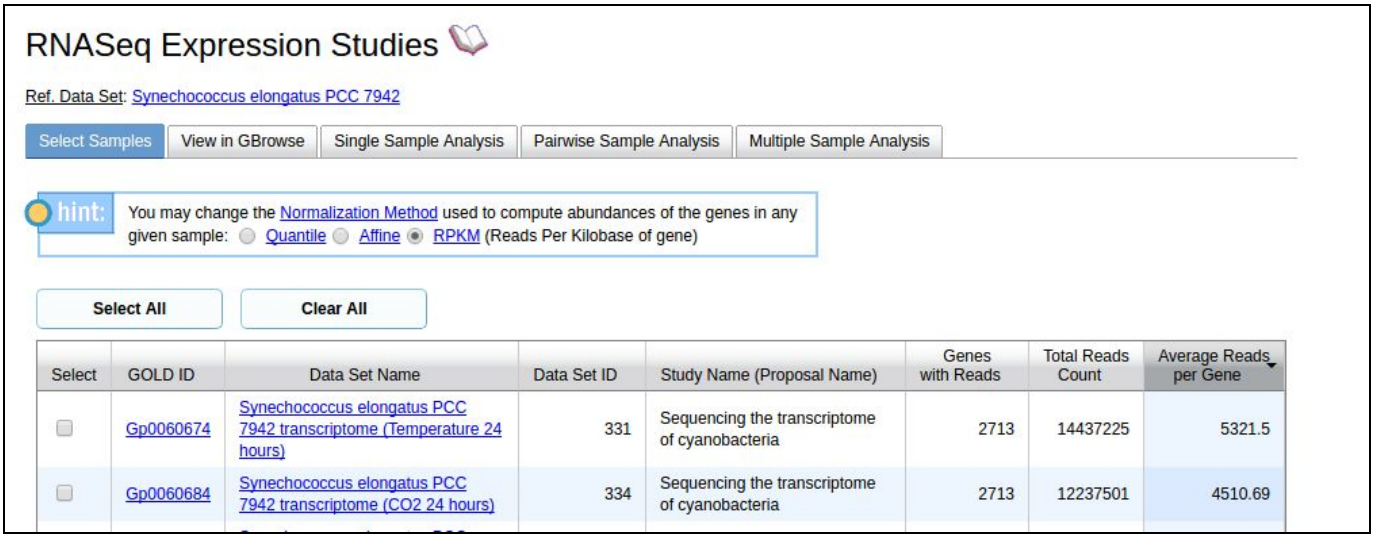


Figure 6a. RNASeq Study – Sample Analysis

Single Sample Analysis

A single sample can be analyzed to show gene coverage and to display information and links to cassettes, associated cog functions, kegg pathways and modules, and EC (Fig. 6c). This information is displayed when the user selects a sample and clicks on **Gene Expression Summary** (Fig. 6b).

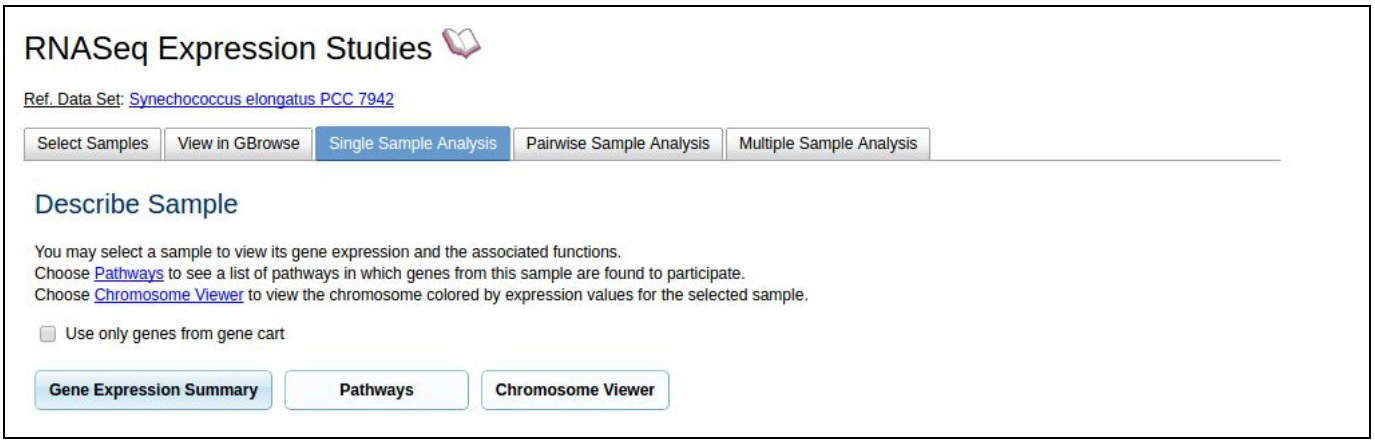


Figure 6b. Single Sample Analysis

RNASeq Expression Data for Selected Sample

Chromosome Viewer

Genome: Synechococcus sp. PCC 7002
Sample: Synechococcus sp. PCC 7002 transcriptome (dark aerobic) [Gp0060686]

Add Selections To Gene Cart | Select All | Clear All

Filter column: KEGG module Filter text: Apply

Export Page 1 of 29 << first < prev 1 2 3 4 5 6 7 8 9 10 next > last >> 100

Column Selector Select Page Deselect Page

Select	Gene ID	Locus Tag	Product Name	Synechococcus sp. PCC 7002 transcriptome (dark aerobic) [352]	Synechococcus sp. PCC 7002 transcriptome (dark aerobic) [352] Raw Count	Cassette ID	COG function	KEGG pathway	EC Number	KEGG module
	641612276	SYNPCC7002_A2266	D-alanyl-D-alanine dipeptidase superfamily	46.113	15	-	M - Cell wall/membrane/envelope biogenesis	Vancomycin resistance Two-component system	EC:3.4.13.22	Vancomycin resistance, D-Ala-D-Lac type (map01502, map02020)
	641612155	SYNPCC7002_A2145	D-Ala-D-Ala carboxypeptidase. Metallo peptidase. MEROPS family M15B	229.684	87	-		Vancomycin resistance Two-component system Peptidoglycan biosynthesis	EC:3.4.16.4	Vancomycin resistance, D-Ala-D-Lac type (map00550, map01502, map02020)
	641610620	SYNPCC7002_A0610	orotate phosphoribosyltransferase (EC 2.4.2.10)/orotidine-5'-phosphate decarboxylase (EC 4.1.1.23)	57.283	40	-	F - Nucleotide transport and metabolism	Pyrimidine metabolism Drug metabolism - other enzymes	EC:2.4.2.10	Uridine monophosphate biosynthesis, glutamine (+ PRPP) => UMP (map00240, map00983)

Figure 6c. Data for a Sample

The user can choose to view the chromosome for this organism colored by expression values for the selected sample by clicking on **Chromosome Viewer** (Fig. 6d).

Chromosome Viewer - Colored by EXPRESSION

Genome: Synechococcus sp. PCC 7002
Scaffold: Synechococcus sp. PCC 7002: NC_010475
(3008047bp, gc=0.50, coordinates **1-500000**)

Coloring is based on coverage of genes for sample: 352
Synechococcus sp. PCC 7002 transcriptome (dark aerobic)

Switch coloring to: Expression

hint: *Mouse over a gene to see details.*
red-high to green-low expression
*Query gene in red, RNAs in **black**, Pseudo genes in white*
My Gene in cyan or dashes
Query gene is marked by a red bar
Gene(s) with protein is marked by a purple bar
Gene(s) in Gene Cart is marked by a blue bar
CRISPR array

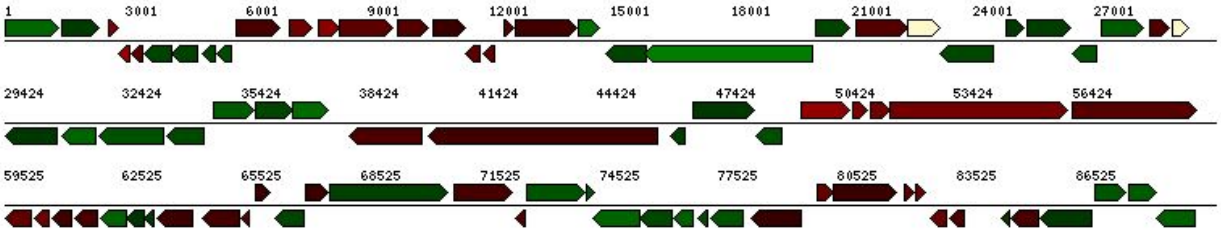


Figure 6d. Chromosome viewer for a Sample

Alternatively, the user can click on **Pathways** (Figure 6b) to view all pathways in which genes from this sample are found to participate (Figure 7).

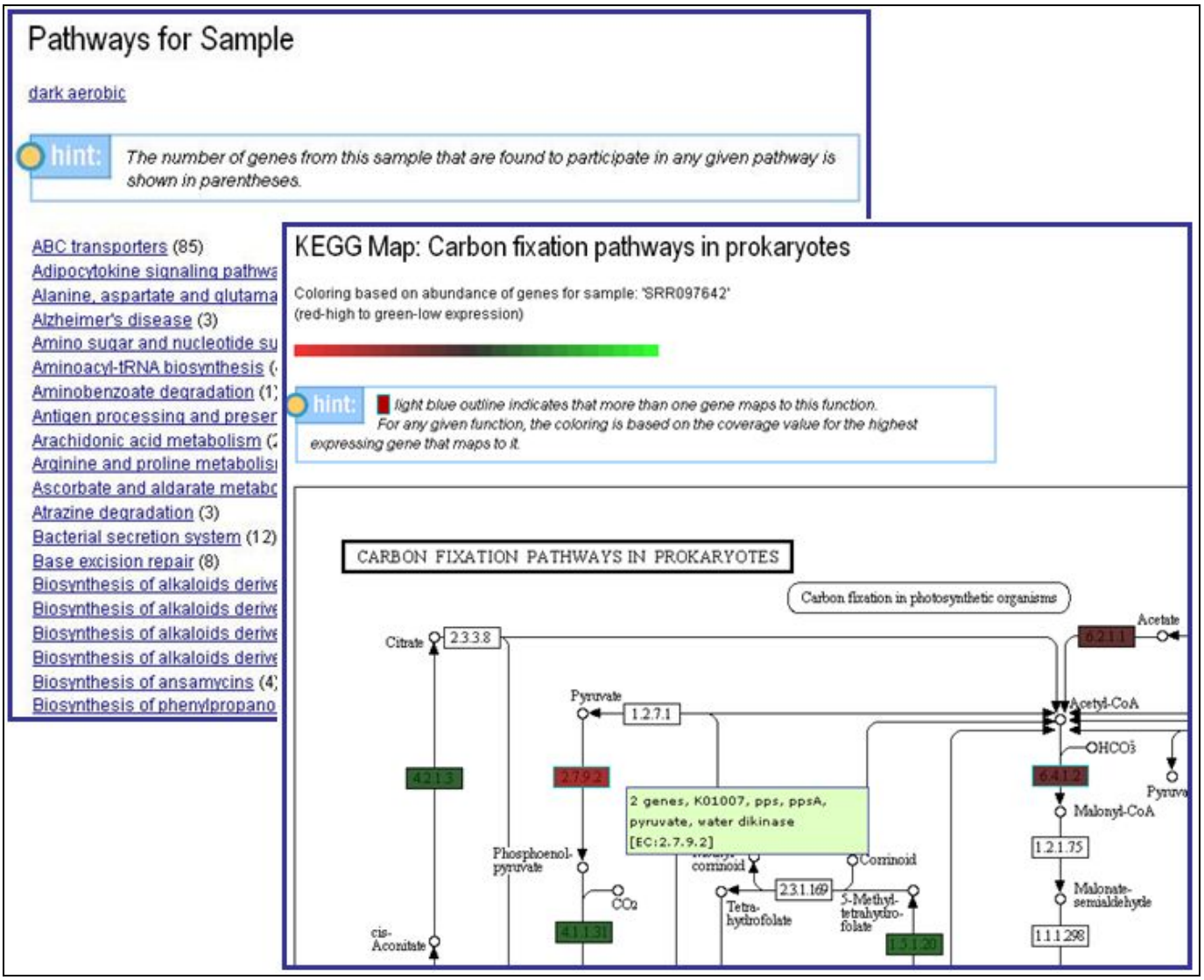


Figure 7. Pathways for a Sample

The kegg pathway link from the sample "expression data" page (Fig. 6c) displays the pathway for the selected sample. The pathway will be colored by gene expression (coverage) for that sample (Figure 7b).

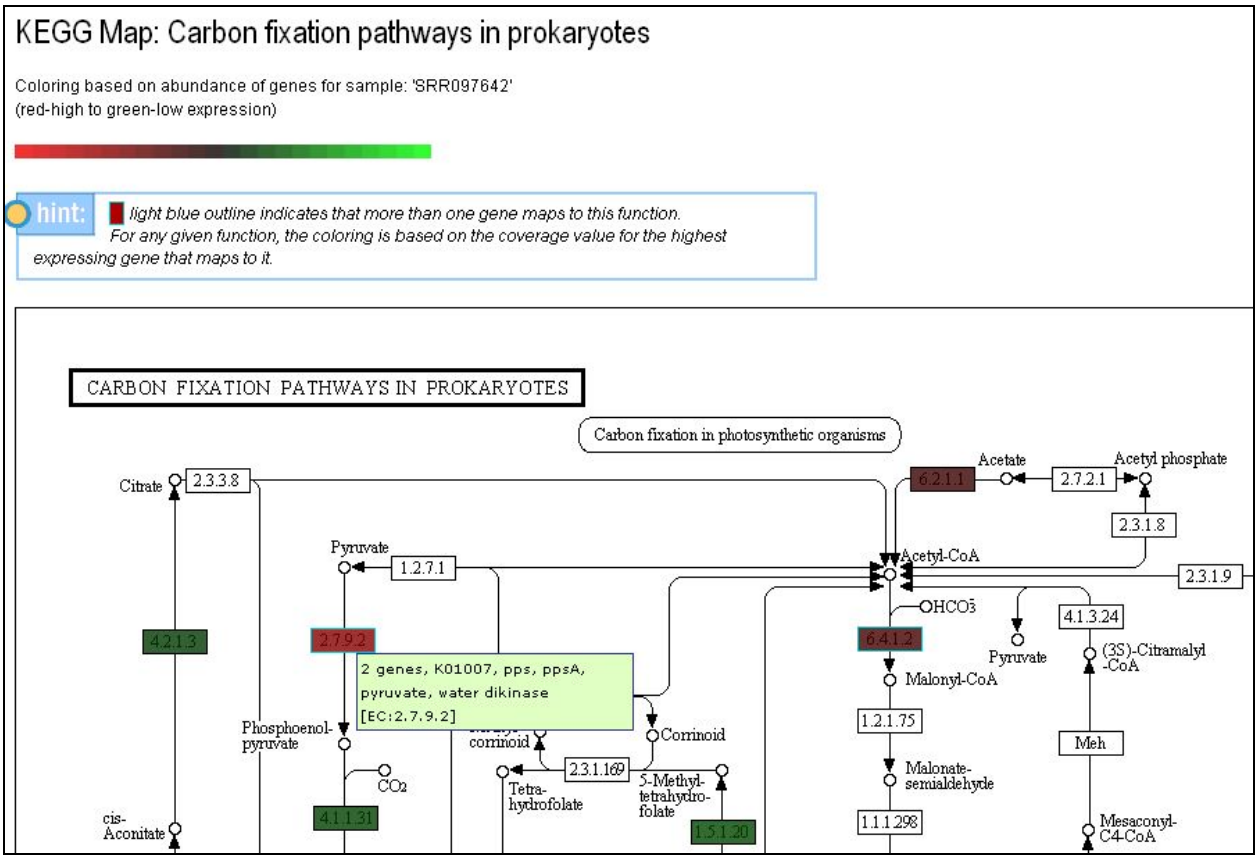


Figure 7b. Kegg Map colored by gene expression in single sample

Pairwise Sample Analysis

The user can select 2 samples of interest to see which genes are up-regulated and which are down-regulated between the two experimental conditions (Fig. 8).

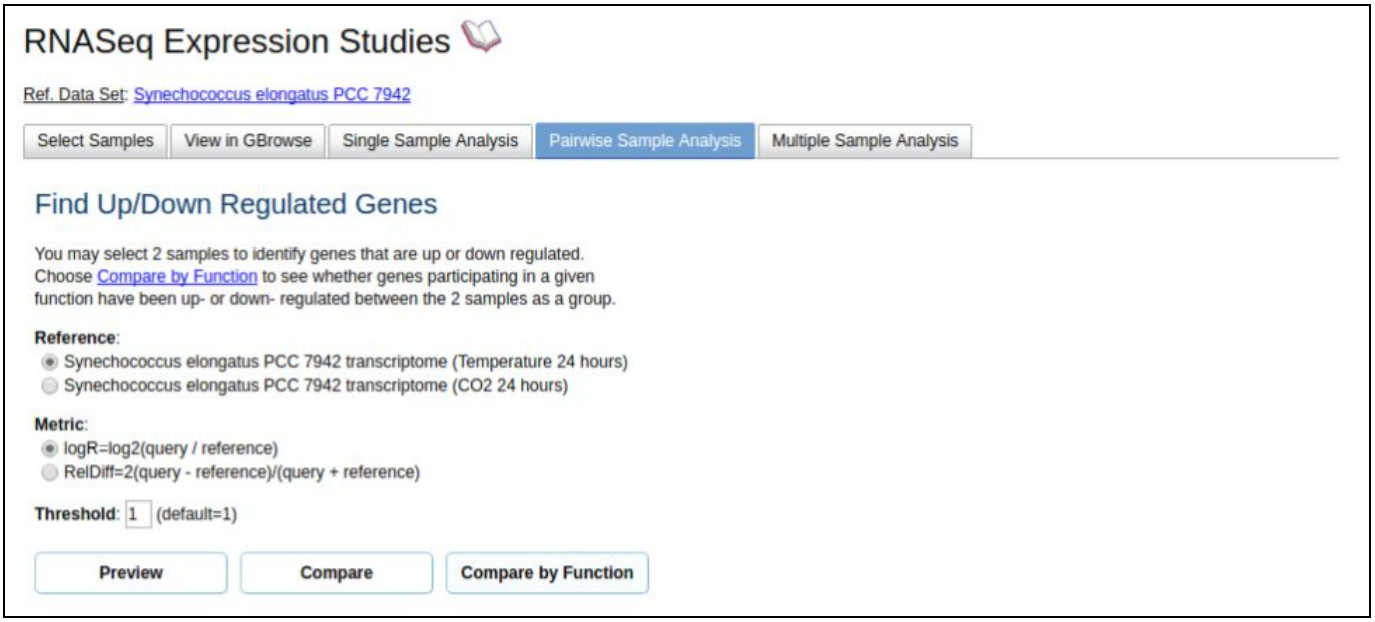


Figure 8. Pairwise Sample Analysis

The user can find all the genes whose expression levels differ by a specified threshold. The difference in expression is computed using either the logR=log2(query/reference) or the RelDiff=2(query-reference)/(query+reference) metric. The user can **preview** the comparison to

see the value spread in a histogram (Fig. 9). This helps in choosing which metric and what threshold value to use.

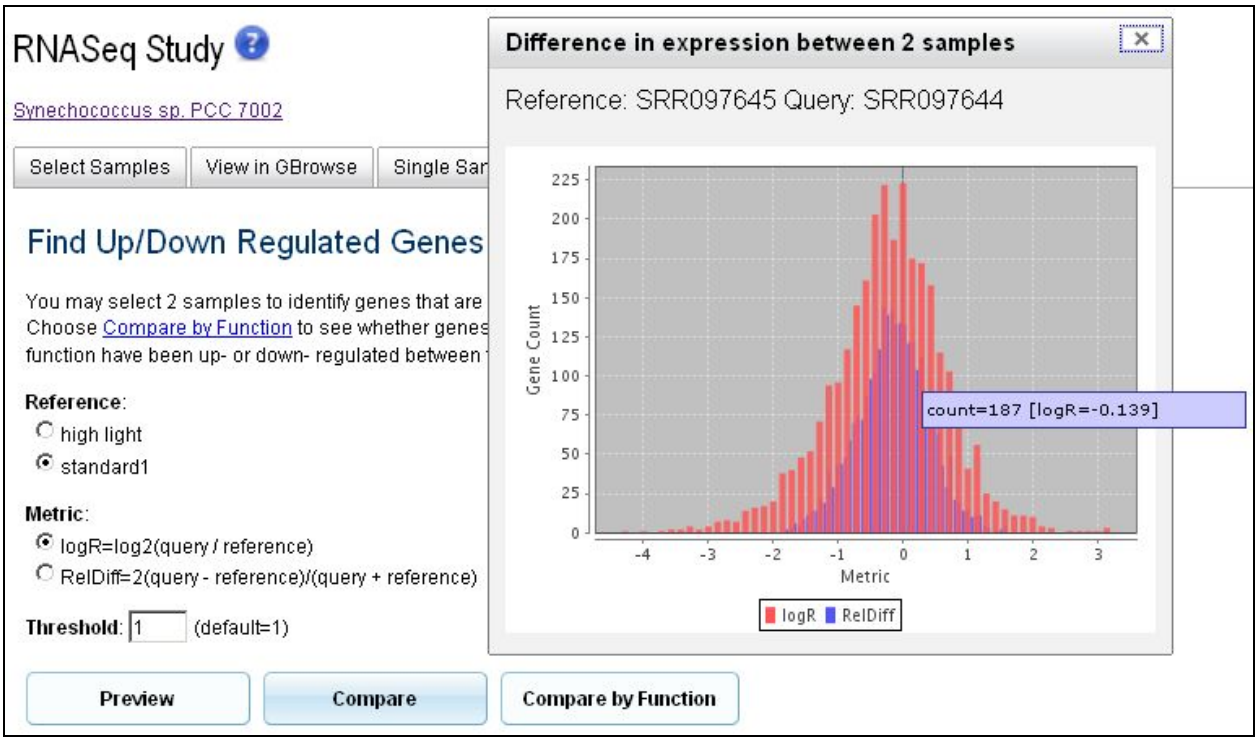
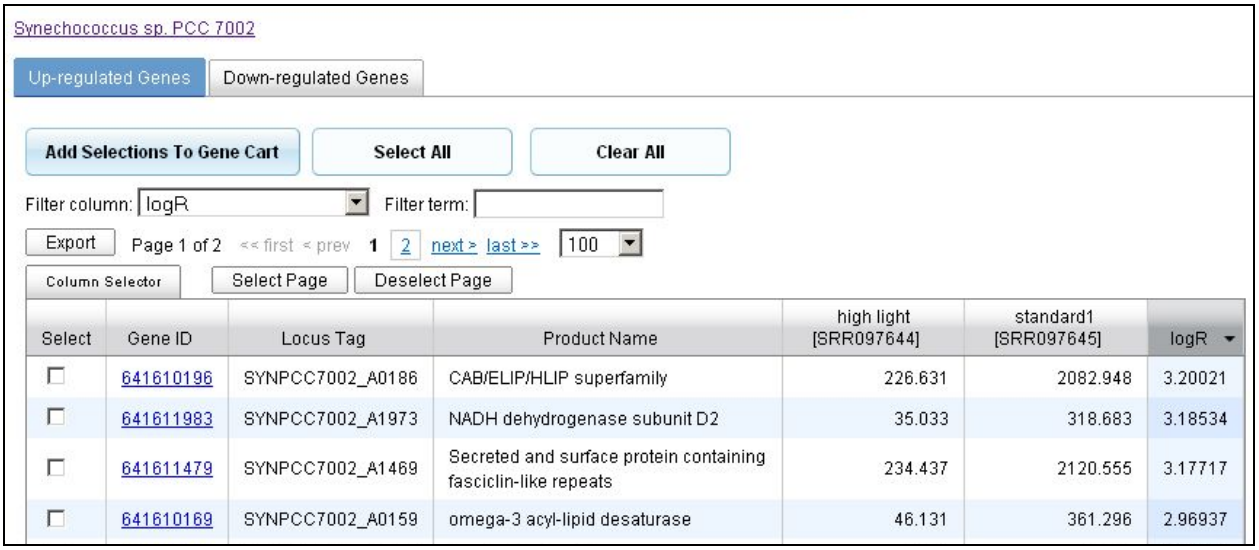


Figure 9. Pairwise Sample Analysis - Preview

When the user clicks **Compare**, the up- and down- regulated genes are presented in separate tabs (Fig. 10). The user can then select which genes to add to the Gene Cart.



Select	Gene ID	Locus Tag	Product Name	high light [SRR097644]	standard1 [SRR097645]	logR
☐	641610196	SYNPCC7002_A0186	CAB/ELIP/HLIP superfamily	226.631	2082.948	3.20021
☐	641611983	SYNPCC7002_A1973	NADH dehydrogenase subunit D2	35.033	318.683	3.18534
☐	641611479	SYNPCC7002_A1469	Secreted and surface protein containing fasciclin-like repeats	234.437	2120.555	3.17717
☐	641610169	SYNPCC7002_A0159	omega-3 acyl-lipid desaturase	46.131	361.296	2.96937

Figure 10. Pairwise Sample Analysis - Compare

The user can click on **Compare by Function** to see whether genes participating in a given function are up- or down- regulated between the two samples as a group (Fig. 11).

Expression by Function for Selected Samples

Genome: Synechococcus elongatus PCC 7942

hint: *Click on a tab to select grouping of genes by kegg or by cog function. For each sample, the average of normalized expression values for each group of genes is displayed.*
Normalization: coverage
2 sample(s) selected

Reference sample: Synechococcus elongatus PCC 7942 transcriptome (Temperature 24 hours)
Query sample: Synechococcus elongatus PCC 7942 transcriptome (CO2 24 hours)

By KEGG | By COG

Filter column: KEGG Pathway Filter text: Apply

Export Page 1 of 2 << first < prev 1 2 next > last >> 100

Column Selector

KEGG Pathway	Gene Count	Gene Count Synechococcus elongatus PCC 7942 transcriptome (Temperature 24 hours) [331]	Average Expression Synechococcus elongatus PCC 7942 transcriptome (Temperature 24 hours) [331]	Gene Count Synechococcus elongatus PCC 7942 transcriptome (CO2 24 hours) [334]	Average Expression Synechococcus elongatus PCC 7942 transcriptome (CO2 24 hours) [334]	logR
2-Oxocarboxylic acid metabolism	21	21	292.372	21	238.601	-0.29321
ABC transporters	62	62	115.617	62	252.776	1.12850
Adipocytokine signaling pathway	1	1	116.508	1	211.875	0.86278
Alanine, aspartate and glutamate metabolism	17	17	372.283	17	323.137	-0.20425

Figure 11. Pairwise Sample Analysis – Compare by Function

The user can analyze the strength of the association between each pair of expression measurements by performing a Spearman's Rank Correlation (Figure 12a-c).

Spearman's Rank Correlation

You may analyze more closely the strength of the association between each pair of expression measurements by performing a Spearman's Rank Correlation. This involves 1) ranking the values under each condition and then 2) calculating the correlation on pairs of ranks. The resulting correlation values for each comparison are in the range of -1 to 1 (where values in between -1 and -0.5 indicate strong negative correlation, 0 means that the ranked value pairs are completely independent and there is no correlation, and values in between 0.5 and 1 indicate strong positive correlation).

Spearman

Figure 12a. Pairwise Sample Analysis – Spearman's Rank Correlation

Spearman's Rank Correlation

Samples compared:
Synechococcus elongatus PCC 7942 transcriptome (Temperature 24 hours)
Synechococcus elongatus PCC 7942 transcriptome (CO2 24 hours)

Spearman's Coefficient | Graph

Add Selections To Gene Cart | Select All | Clear All

Filter column: Spearman's Rho ▾ Filter text ▾: Apply

Export Page 1 of 107 << first < prev 1 2 3 4 5 6 7 8 9 10 next > last >> 25 ▾

Column Selector | Select Page | Deselect Page

Select	Gene ID	Locus Tag	Product Name	Spearman's Rho ▾
☐	637798390	Synpcc7942_B2658	hypothetical protein	1.000
☐	637798394	Synpcc7942_B2641	hypothetical protein	1.000
☐	637798454	Synpcc7942_0048	pilin polypeptide PilA-like	1.000
☐	637798519	Synpcc7942_0113	hypothetical protein	1.000
☐	637798529	Synpcc7942_0123	hypothetical protein	1.000
☐	637798649	Synpcc7942_0243	possible high light inducible polypeptide HliC	1.000

Figure 12b. Spearman's Rank Correlation Results

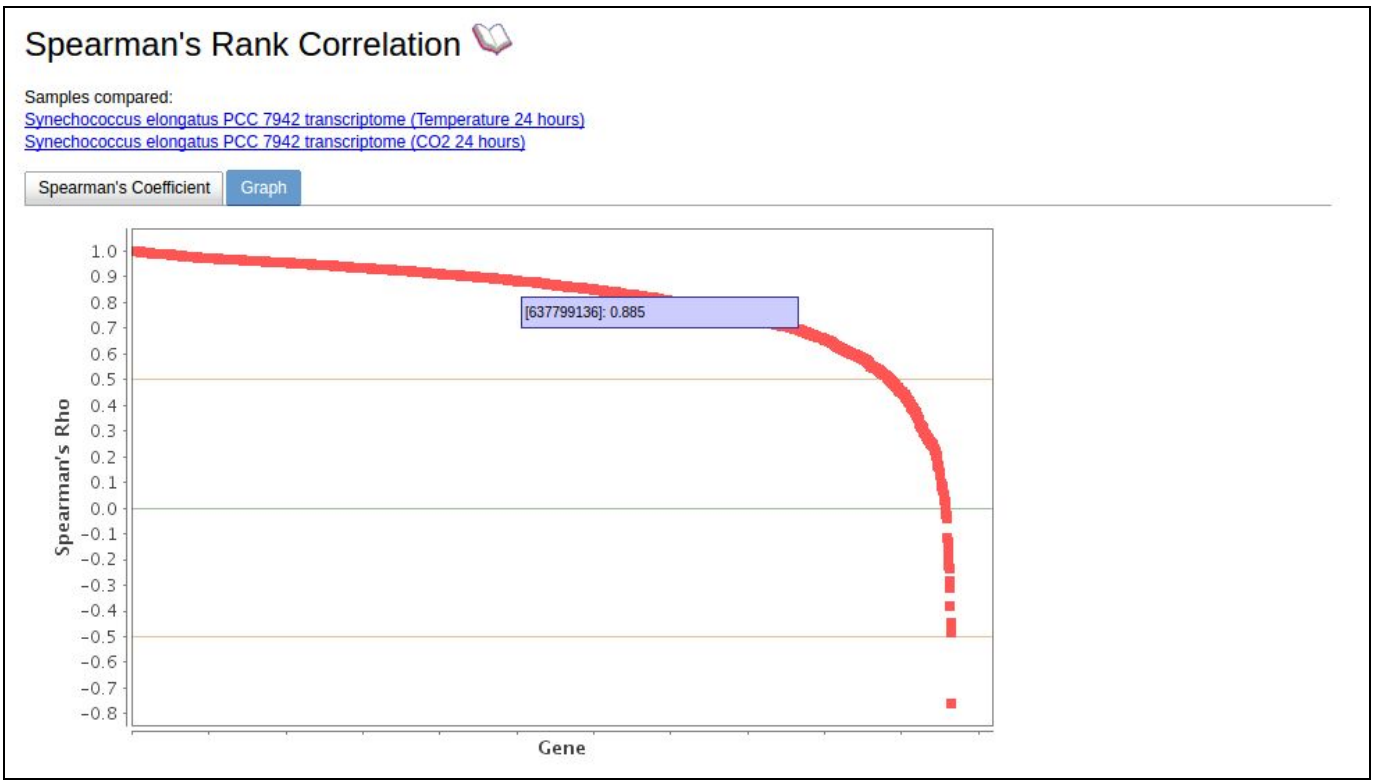


Figure 12c. Spearman's Rank Correlation - Graph

Two samples can also be analyzed using Linear Regression and R-squared to help determine if they are replicates (Fig. 13a-b).

Figure 13a. Pairwise Sample Analysis - Linear Regression

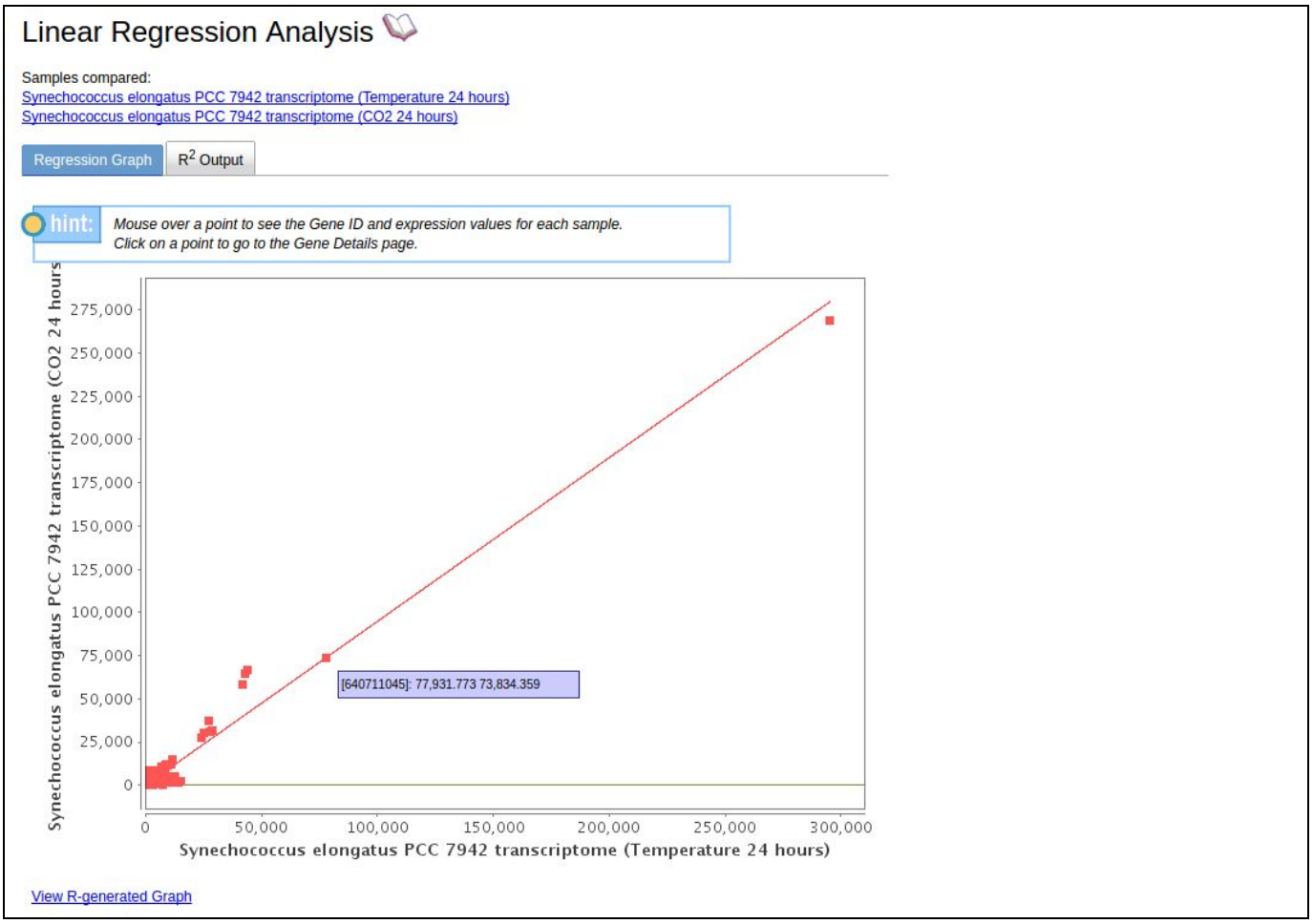


Figure 13b. Pairwise Sample Analysis - Linear Regression - Results

Please note: If the 2 samples selected for Pairwise Analysis are from different genomes (same study), then they can only be analyzed using **Compare by Function** (Fig. 13c).

Select Samples | Single Sample Analysis | Pairwise Sample Analysis | Multiple Sample Analysis

Find Up/Down Regulated Genes

You may select 2 samples to identify genes that are up or down regulated.
Choose Compare by Function to see whether genes participating in a given function have been up- or down- regulated between the 2 samples as a group.

Reference:
- (●) Synechococcus elongatus PCC 7942 transcriptome (pH 1 hour)
- ( ) Synechocystis sp. PCC 6803 transcriptome (S6-2 test)

Metric:
- (●) logR=log2(query / reference)
- ( ) RelDiff=2(query - reference)/(query + reference)

Threshold: 1 (default=1)

Please note: If the 2 samples selected are from different genomes, then they cannot be compared using gene expression analysis

Preview | Compare | **Compare by Function**

Spearman's Rank Correlation

You may analyze more closely the strength of the association between each pair of expression measurements by performing a Spearman's Rank Correlation. This involves 1) ranking the values under each condition and then 2) calculating the correlation on pairs of ranks. The resulting correlation values for each comparison are in the range of -1 to 1 (where values in between -1 and -0.5 indicate strong negative correlation, 0 means that the ranked value pairs are completely independent and there is no correlation, and values in between 0.5 and 1 indicate strong positive correlation).

Spearman

Linear Regression

You may select 2 samples to compare using linear regression analysis and R-squared.
This analysis can help determine whether the selected samples are replicates.

Regress

Figure 13c. Pairwise Analysis - comparing samples from 2 different genomes

Multiple Sample Analysis

Multiple samples can be analyzed to show gene coverage and to display information and links to cassettes, associated cog functions, kegg pathways and modules, and EC (Figure 14). This information is displayed when the user selects samples and clicks on **Gene Expression Summary**. The results are similar to those of Single Sample Analysis (Fig. 6c).

RNASeq Expression Studies

Ref. Data Set: Synechococcus elongatus PCC 7942

Select Samples | View in GBrowse | Single Sample Analysis | Pairwise Sample Analysis | Multiple Sample Analysis

Describe Samples

You may select samples to compare and describe. If you choose to first cluster the samples to have the results colored by the cluster grouping, go to Map Clusters to Pathways below. Choose Expression by Function to see, for each selected sample, the average expression of genes participating in a given function as a group.

☐ Use only genes from gene cart

Gene Expression Summary | Expression by Function

Cluster Samples

You may select samples and cluster them based on the abundance of the expressed genes. Proximity of grouping indicates the relative degree of similarity of samples to each other.

Clustering Method:
- ◉ Pairwise complete-linkage (default)
- ○ Pairwise single-linkage
- ○ Pairwise average-linkage

Distance Measure:
- ◉ Pearson correlation (default)
- ○ Spearman's rank correlation
- ○ Euclidean distance
- ○ City-block distance (Manhattan)

Minimum number of samples in which a gene should appear in order to be included: 3 (default=3, total=15)

Cut-off threshold: 0.8 (default=0.8)

☐ Use only genes from gene cart

Cluster | Map Clusters to Pathways

Figure 14. Multiple Sample Analysis

The samples can be clustered before displaying gene expression summary. This way the results will be colored by the cluster groupings (see below under “Sample Clustering”).

The kegg pathway link from the “expression data” page, displays the pathway for multiple samples that the user selected. The pathway will be colored by percentage of samples in which a given gene is expressed (Fig. 15).

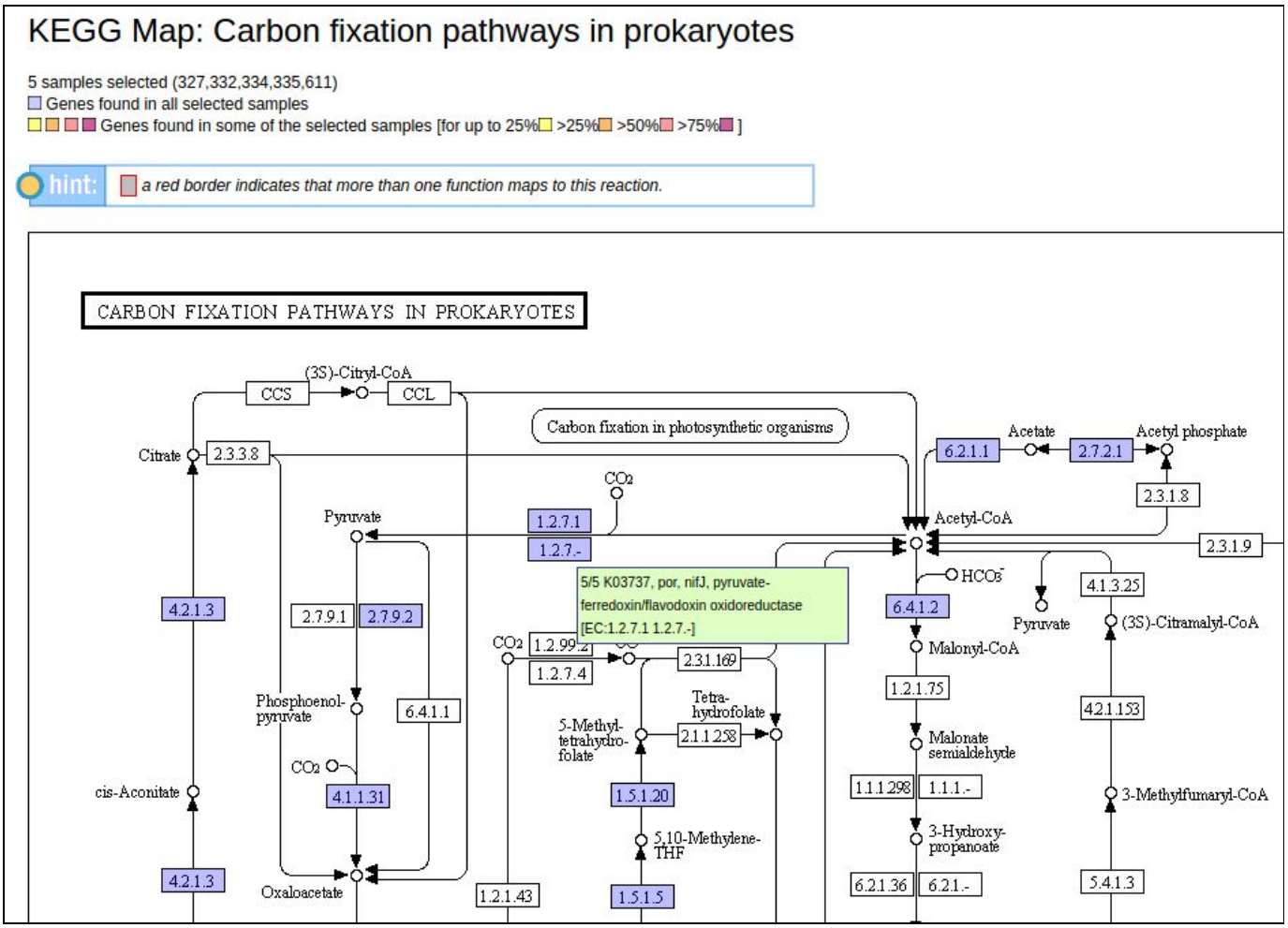
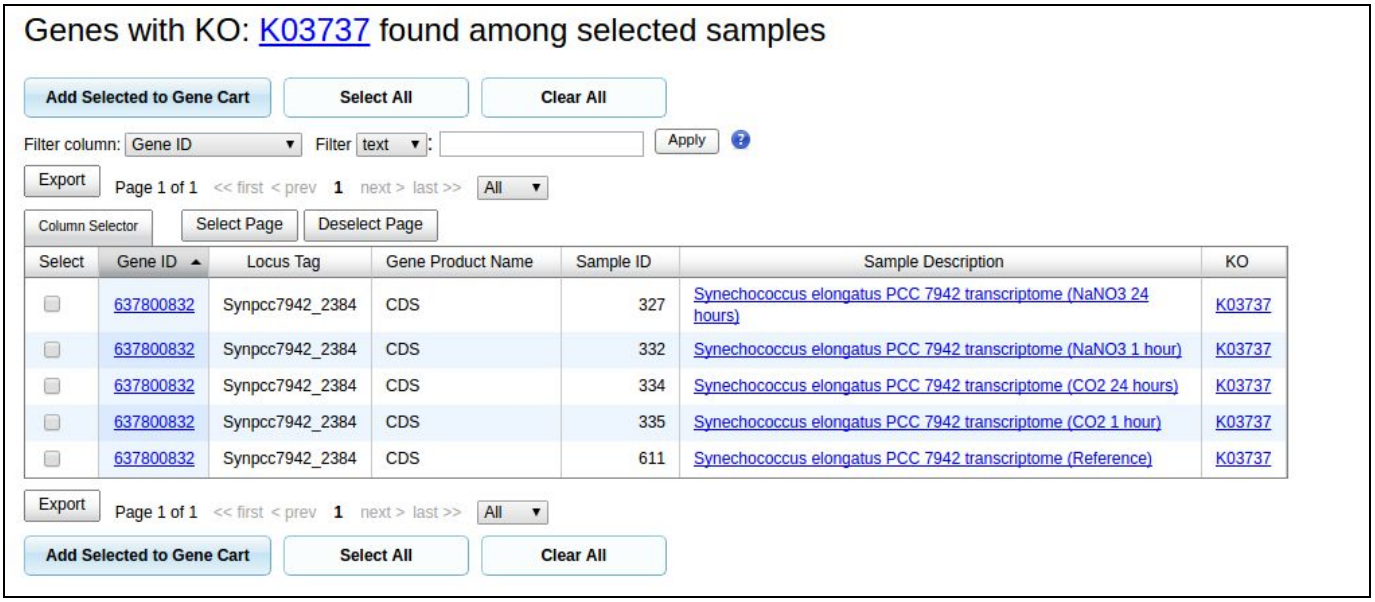


Figure 15. Pathway for Samples (unclustered)

Each region of interest on the kegg map links to a list of genes for the selected samples that fall in that region (Fig. 16).



Genes with KO: K03737 found among selected samples

Add Selected to Gene Cart | Select All | Clear All

Filter column: Gene ID Filter text : Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector | Select Page | Deselect Page

Select	Gene ID	Locus Tag	Gene Product Name	Sample ID	Sample Description	KO
	637800832	Synpcc7942_2384	CDS	327	Synechococcus elongatus PCC 7942 transcriptome (NaNO3 24 hours)	K03737
	637800832	Synpcc7942_2384	CDS	332	Synechococcus elongatus PCC 7942 transcriptome (NaNO3 1 hour)	K03737
	637800832	Synpcc7942_2384	CDS	334	Synechococcus elongatus PCC 7942 transcriptome (CO2 24 hours)	K03737
	637800832	Synpcc7942_2384	CDS	335	Synechococcus elongatus PCC 7942 transcriptome (CO2 1 hour)	K03737
	637800832	Synpcc7942_2384	CDS	611	Synechococcus elongatus PCC 7942 transcriptome (Reference)	K03737

Export Page 1 of 1 << first < prev 1 next > last >> All

Add Selected to Gene Cart | Select All | Clear All

Figure 16. List of genes for selected samples that fall on a given region of interest

On the “Multiple Sample Analysis” page, the user can also click on **Expression by Function** (Fig. 14) to see, for each selected sample, the average expression of genes participating in a given function as a group (Fig. 17).

Expression by Function for Selected Samples

Genome: Synechococcus elongatus PCC 7942

hint: *Click on a tab to select grouping of genes by kegg or by cog function. For each sample, the average of normalized expression values for each group of genes is displayed.*
Normalization: coverage
5 sample(s) selected

By KEGG | By COG

Filter column: KEGG Pathway Filter text: Apply

Export Page 1 of 2 << first < prev 1 2 next > last >> 100

Column Selector

KEGG Pathway	Gene Count	Gene Count Synechococcus elongatus PCC 7942 transcriptome (NaNO3 24 hours) [327]	Average Expression Synechococcus elongatus PCC 7942 transcriptome (NaNO3 24 hours) [327]	Gene Count Synechococcus elongatus PCC 7942 transcriptome (NaNO3 1 hour) [332]	Average Expression Synechococcus elongatus PCC 7942 transcriptome (NaNO3 1 hour) [332]	Gene Count Synechococcus elongatus PCC 7942 transcriptome (CO2 24 hours) [334]	Average Expression Synechococcus elongatus PCC 7942 transcriptome (CO2 24 hours) [334]	Gene Count Synechococcus elongatus PCC 7942 transcriptome (CO2 1 hour) [335]	Average Expression Synechococcus elongatus PCC 7942 transcriptome (CO2 1 hour) [335]	Gene Count Synechococcus elongatus PCC 7942 transcriptome (Reference) [611]	Average Expression Synechococcus elongatus PCC 7942 transcriptome (Reference) [611]
2-Oxocarboxylic acid metabolism	21	21	198.403	21	260.226	21	238.601	21	289.412	21	261.360
ABC transporters	62	62	254.888	62	751.176	62	252.776	62	345.642	62	165.054
Adipocytokine signaling pathway	1	1	115.861	1	137.556	1	211.875	1	116.034	1	97.200
Alanine, aspartate and glutamate metabolism	17	17	173.258	17	179.914	17	323.137	17	236.934	17	185.411

Figure 17. Multiple Sample Analysis – Expression by Function

Sample Clustering

For each study, the experiments (samples) can be selected for **clustering** based on coverage values for the genes expressed in each sample (Fig. 18). The user can specify the clustering method and the distance measure to use, as well as the minimum number of samples in which a gene should appear in order to be included in the calculation.

Cluster Samples

You may select samples and cluster them based on the abundance of the expressed genes.
Proximity of grouping indicates the relative degree of similarity of samples to each other.

Clustering Method:
- (●) Pairwise complete-linkage (default)
- ( ) Pairwise single-linkage
- ( ) Pairwise average-linkage

Distance Measure:
- (●) Pearson correlation (default)
- ( ) Spearman's rank correlation
- ( ) Euclidean distance
- ( ) City-block distance (Manhattan)

Minimum number of samples in which a gene should appear in order to be included: 4 (default=3, total=15)

Cut-off threshold: 0.8 (default=0.8)

☐ Use only genes from gene cart

Cluster | Map Clusters to Pathways

Figure 18. Clustering of RNASeq samples

Both samples and genes are clustered. What will be displayed in IMG are a hierarchical cluster tree of samples and a normalized heat map of coverage values for each gene for each sample

(Fig. 19). On the heat map, the samples link to sample pages and the genes link to gene pages. Clicking on a sample in the cluster tree **sorts** the heat map on descending coverage values for that sample.

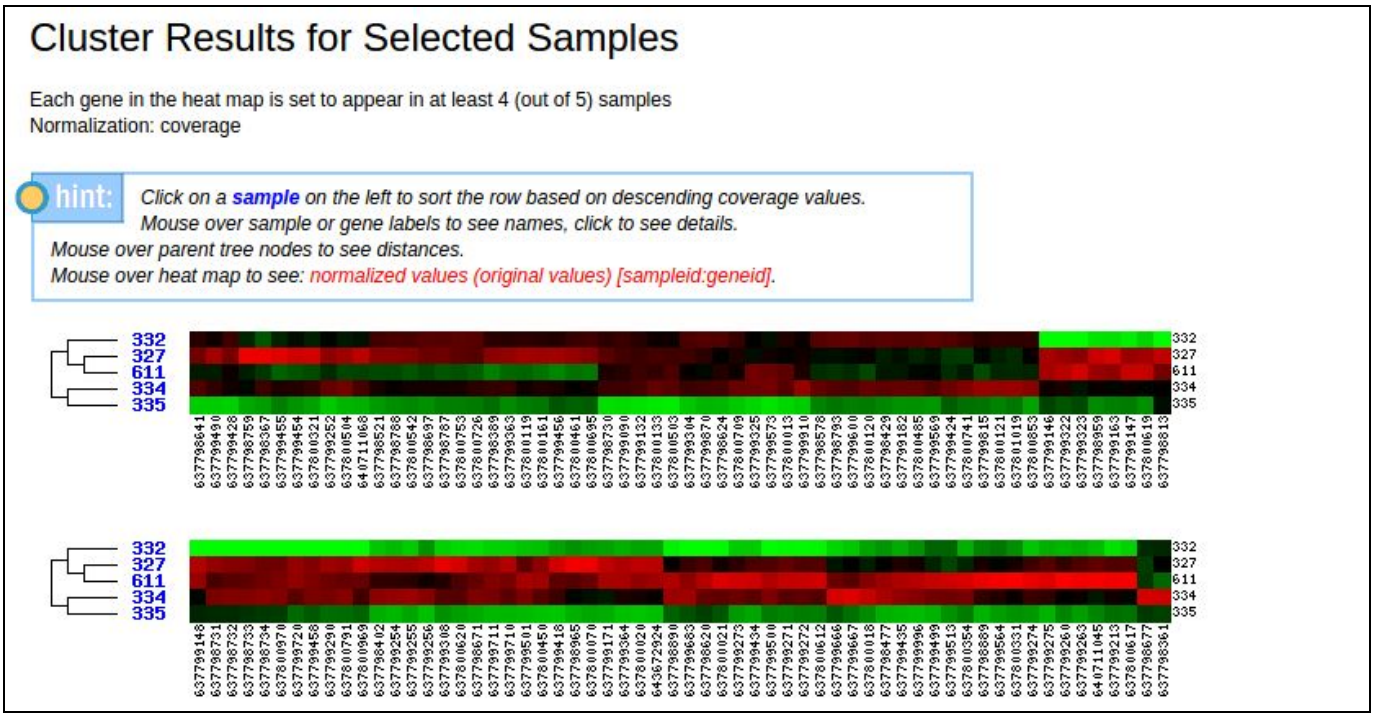


Figure 19. Clustering results

Multiple samples can be clustered to create groupings of genes. Expression data for the samples can then be sorted by these cluster groups (Fig. 20). This is done when the user clicks on **Map Clusters to Pathways** under the "Multiple Sample Analysis" tab (Fig. 18).

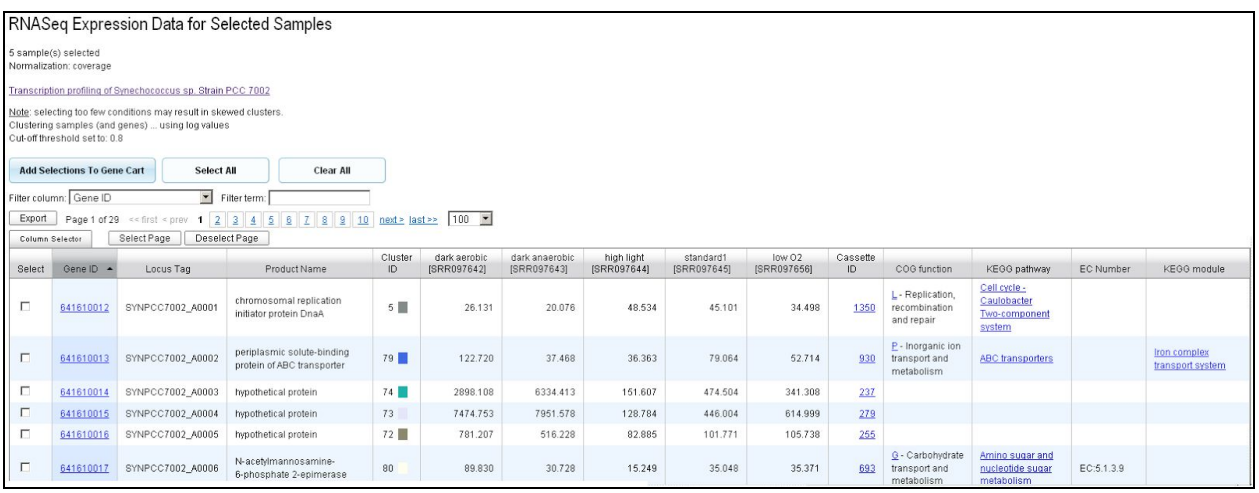


Figure 20. Map clusters to pathways

The kegg pathway link from the sample "expression data" page, displays the pathway colored by the cluster groupings (Figure 21).

KEGG Map: Carbon fixation pathways in prokaryotes

hint: *Samples have been clustered into cluster groups. Functions with genes belonging to the same cluster are colored by the color of that cluster group (see Cluster ID).*
indicates that genes mapping to this function belong to different clusters.

Hide Cluster Colors

Selecting a gene, finds it on the map and marks it with ▽
Gene expression data coloring is based on red (high) to green (low) expression.

Filter column: dark anaerobic [SRR097643] Filter term:

Export Page 1 of 2 << first < prev 1 2 next > last >> 10

Column Selector

Select	Gene ID	Cluster ID	Locus Tag	Product Name	KO	dark aerobic [SRR097642]	dark anaerobic [SRR097643]	high light [SRR097644]
☑	641610260	74	SYNPCC7002_A0250	phosphoenolpyruvate synthase	K01007	0.167	1.099	-0.564
☑	641612103	57	SYNPCC7002_A2093	malate dehydrogenase, NAD-dependent	K00024	0.231	1.073	-0.360
☑	641610899	57	SYNPCC7002_A0889	5,10-methylenetetrahydrofolate dehydrogenase; methenyl tetrahydrofolate cyclohydrolase	K01491	0.070	1.071	-0.020
☐	641611104	73	SYNPCC7002_A1094	succinate dehydrogenase iron-sulfur protein subunit	K00240	0.051	1.046	-0.758
☐	641610901	74	SYNPCC7002_A0891	succinate-CoA ligase beta chain	K01903	0.240	0.944	-0.364
☐	641612579	73	SYNPCC7002_A2569	succinate dehydrogenase flavoprotein subunit	K00239,K00244	0.403	0.867	-0.662
☐	641611693	89	SYNPCC7002_A1683	aconitate hydratase 2	K01682	0.108	0.786	0.253
☐	641610900	90	SYNPCC7002_A0890	succinyl-CoA synthetase, alpha chain	K01902	0.439	0.610	0.194
☐	641611848	71	SYNPCC7002_A1838	acetate--CoA ligase	K01895	0.541	0.576	-0.535
☐	641610848	5	SYNPCC7002_A0838	isocitrate dehydrogenase, NADP-dependent	K00031	-0.447	-0.316	0.484

Export Page 1 of 2 << first < prev 1 2 next > last >> 10

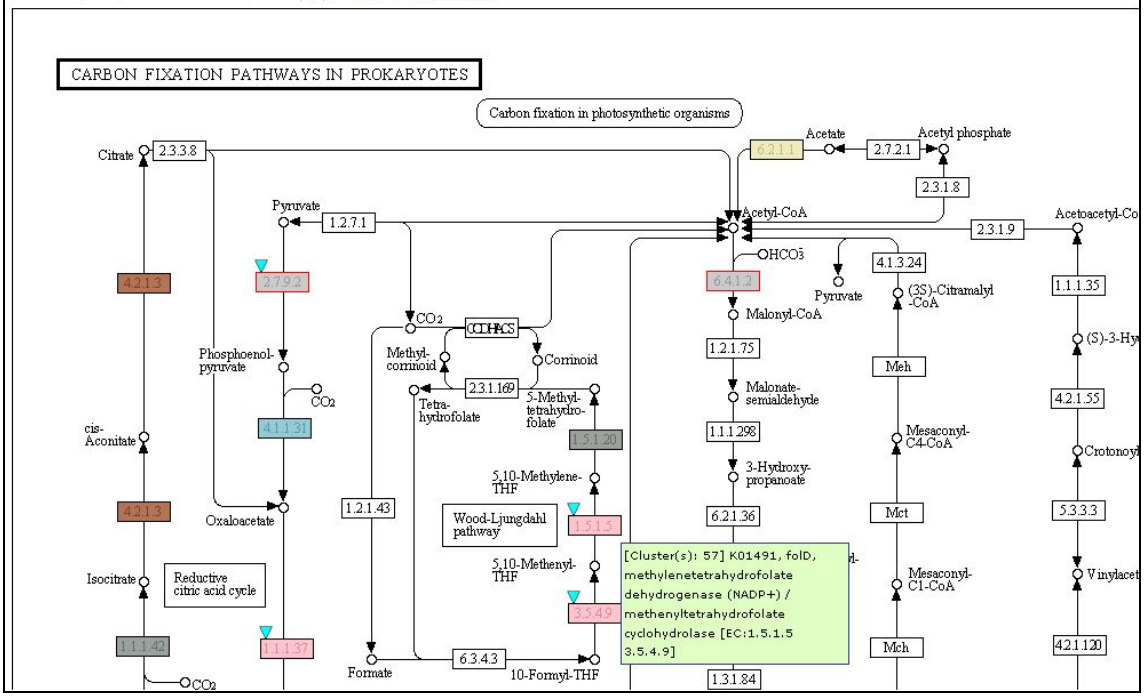


Figure 21. Multiple Sample Analysis – Map Clusters to Pathway

View in GBrowse

Please Note: GBrowse functionality is currently unavailable as it is being transitioned to JBrowse.

The samples for an RNASeq study can also be analyzed in GBrowse to view differences in read coverage for individual genes under various conditions (Fig. 22).

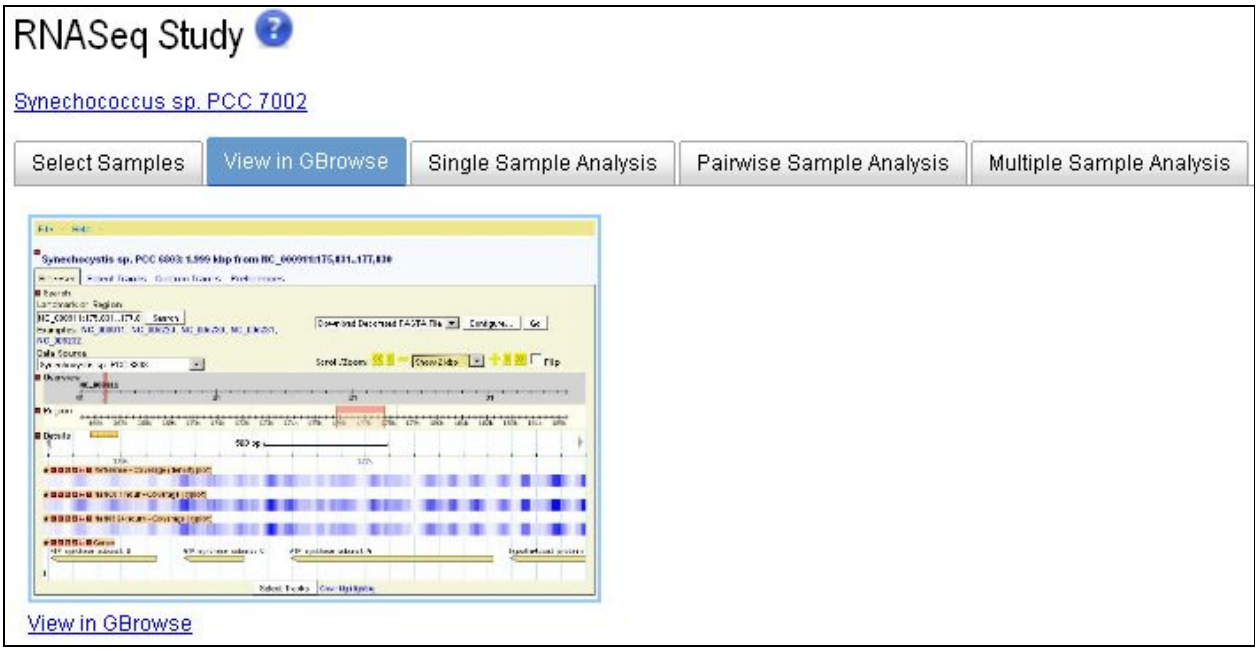


Figure 22. View in GBrowse

Clicking on **View in GBrowse** link or on the image of GBrowse will open the data for this study in GBrowse (Fig. 23).

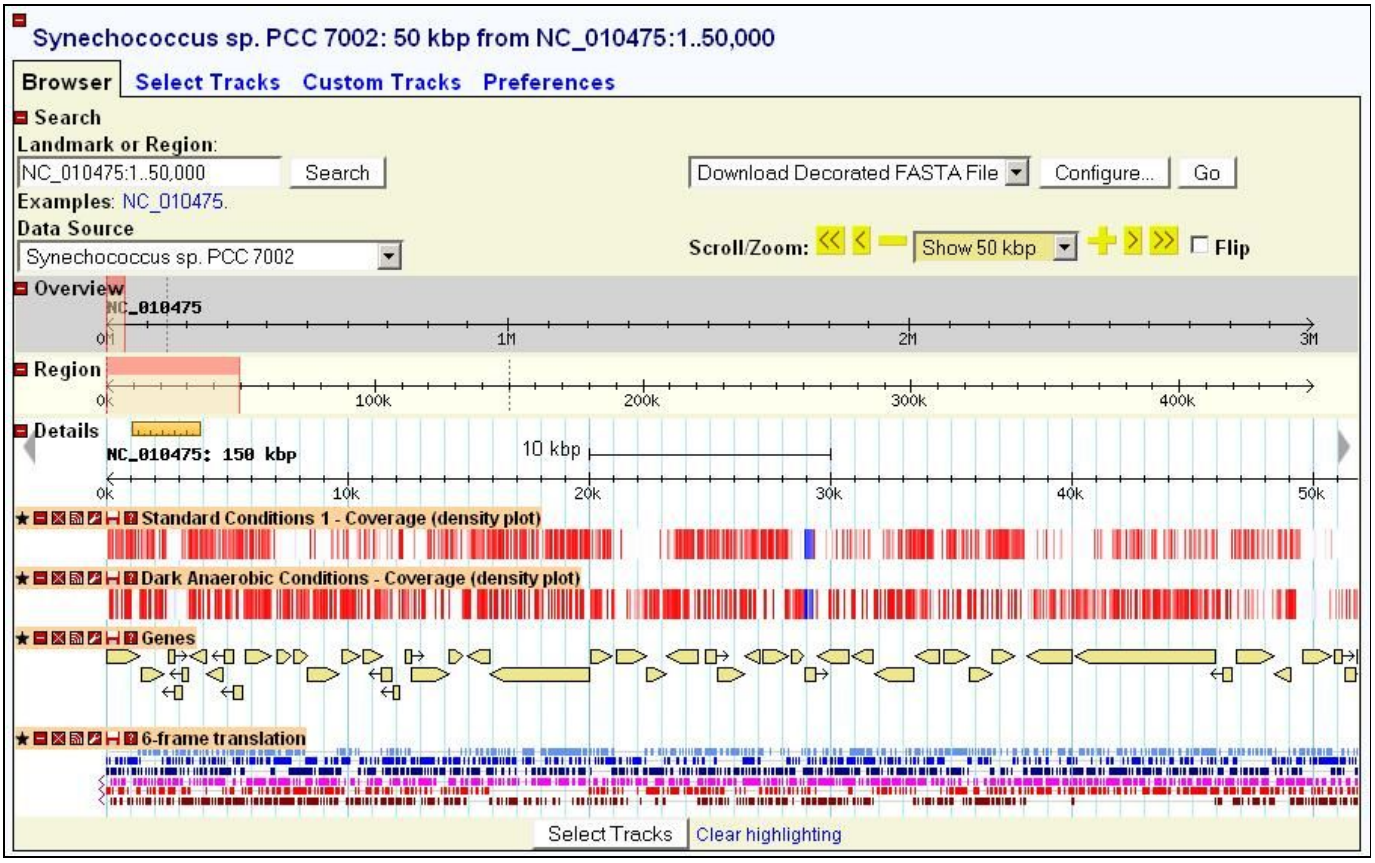


Figure 23. Study loaded into GBrowse

After the data is loaded, the user needs to click on “Select Tracks” to select which data to display (Fig. 24).

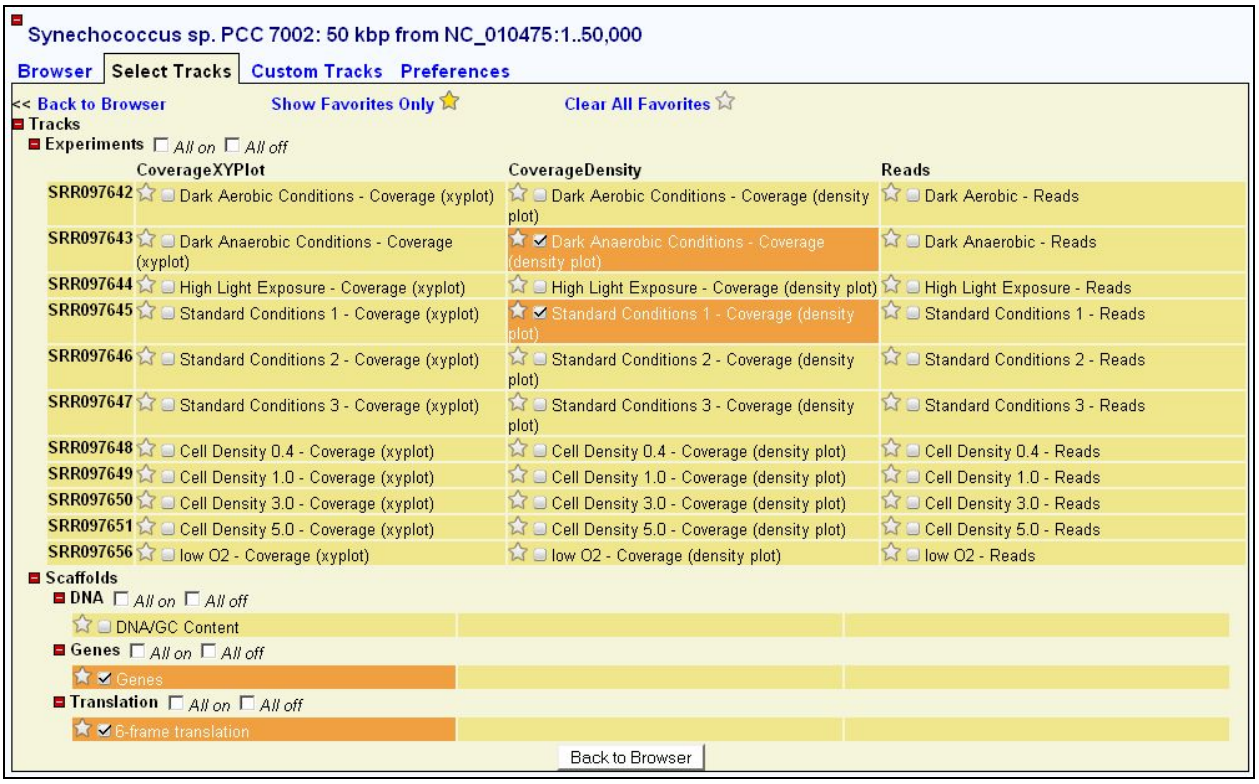


Figure 24. GBrowse – Select Tracks

The user can also zoom closer on the chromosome and see the actual DNA and nucleotide sequences for the genes (Fig. 25).

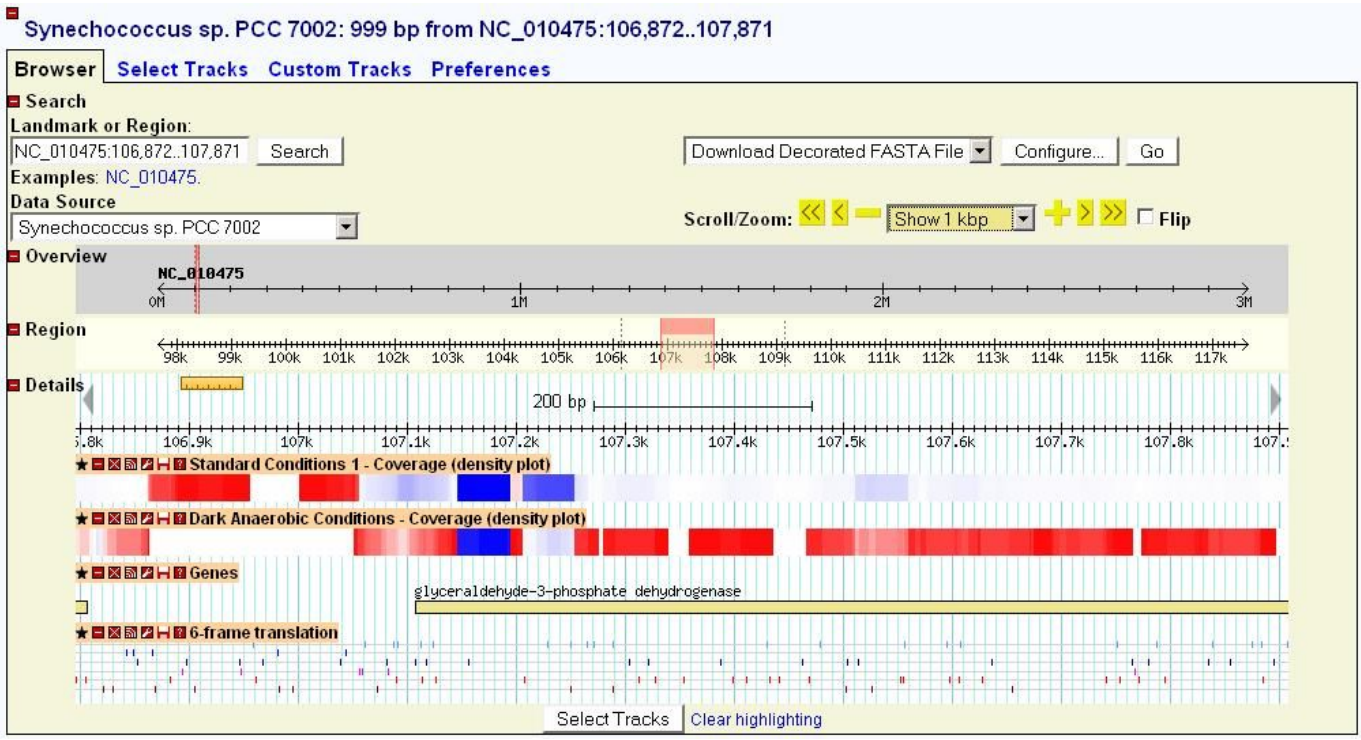


Figure 25. GBrowse – Zoom

Chromosome Viewer – Color by "Expression"

There are 2 ways to color the chromosome by expression.

From the sample page - from each experimental sample page, the user can select to go to the Chromosome Viewer page (Fig. 26).

RNASeq Expression Data for Selected Sample

dark aerobic

Chromosome Viewer

Transcription profiling of Synechococcus sp. Strain PCC 7002

Add Selections To Gene Cart | Select All | Clear All

Filter column: Gene ID Filter term:

Export Page 1 of 29 << first < prev 1 2 3 4 5 6 7 8 9 10 next > last >> 100

Column Selector | Select Page | Deselect Page

Select	Gene ID	Locus Tag	Product Name	dark aerobic [SRR097642]	Cassette ID	COG function	KEGG pathway	EC Number	KEGG module
☐	641610012	SYNPCC7002_A0001	chromosomal replication initiator protein DnaA	26.131	1350	L - Replication, recombination and repair	Cell cycle - Caulobacter Two-component system		
☐	641610013	SYNPCC7002_A0002	periplasmic solute-binding protein of ABC transporter	122.720	930	P - Inorganic ion transport and metabolism	ABC transporters		Iron complex transport system
☐	641610014	SYNPCC7002_A0003	hypothetical protein	2898.108	237				
☐	641610015	SYNPCC7002_A0004	hypothetical protein	7474.753	279				
☐	641610016	SYNPCC7002_A0005	hypothetical protein	781.207	255				
☐	641610017	SYNPCC7002_A0006	N-acetylmannosamine-6-phosphate 2-epimerase	89.830	693	G - Carbohydrate transport and metabolism	Amino sugar and nucleotide sugar metabolism	EC:5.1.3.9	

Figure 26. Chromosome Viewer link from Sample page

From the genome page - if a genome has samples associated with it, the "Scaffolds and Contigs" page allows the user to display the list of all available samples (by clicking on the "Select RNASeq Samples" button – Fig. 27) and to choose one sample to color the chromosome by the expression of genes for that sample.

Chromosome Viewer

Scaffolds and contigs for Synechococcus sp. PCC 7002

Select RNASeq Samples

User Selectable Coordinates

Add to Scaffold Cart | Select All | Clear All

Filter column: Scaffold Filter term:

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector | Select Page | Deselect Page

Select	Scaffold	Length (bp)	GC	Type	Topology	No. Genes	Coordinate Range
☐	Synechococcus sp. PCC 7002 plasmid pAQ1: NC_010476	4809	0.49	plasmid	circular	3	1..4809
☐	Synechococcus sp. PCC 7002 plasmid pAQ3: NC_010477	16103	0.46	plasmid	circular	17	1..16103
☐	Synechococcus sp. PCC 7002 plasmid pAQ4: NC_010478	31972	0.44	plasmid	circular	30	1..31972
☐	Synechococcus sp. PCC 7002 plasmid pAQ5: NC_010479	38515	0.43	plasmid	circular	39	1..38515

Figure 27. User can choose samples to use for coloring the chromosome viewer

The Chromosome Viewer that opens up displays coloring by “expression”, that is, coverage values for each gene in the given sample (Fig. 28).

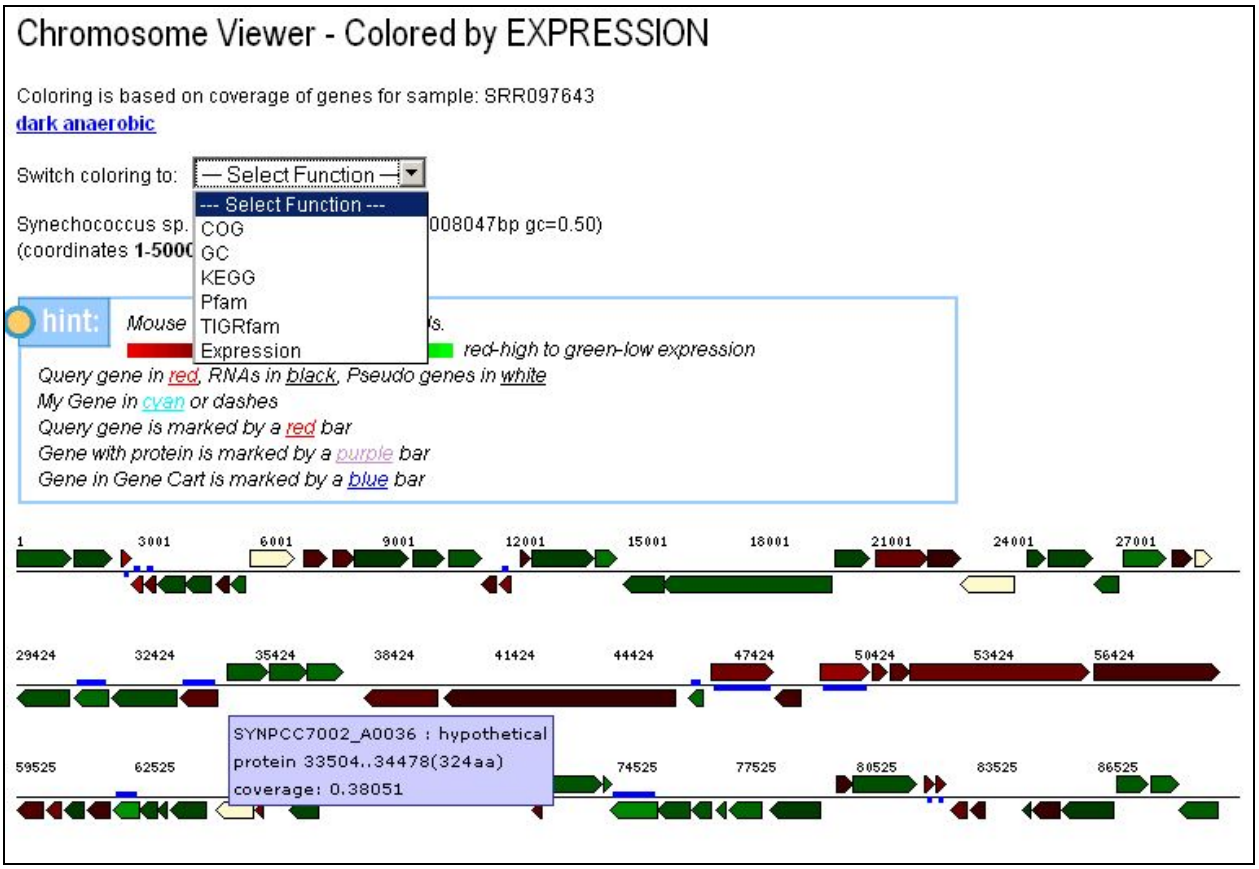


Figure 28. Chromosome Viewer colored by gene expression in a given sample

Analyzing genes of interest

The user can choose to limit the analysis to only the genes that were added to the Gene Cart (Fig. 29 below). For example, after a pairwise analysis the user can add the top 100 genes, showing the greatest logR value, to the Gene Cart. The user can go back and select all samples. Then, returning to the “Cluster Samples” page under “Multiple Sample Analysis” tab (Fig. 14), the user can check the “Use only genes from gene cart” and then click **Cluster** (Fig. 18). Alternatively, the user can use “Gene Expression Summary” to see the Gene Cart genes in the context of the KEGG maps, chromosomal neighborhood, and cassettes that they belong to.

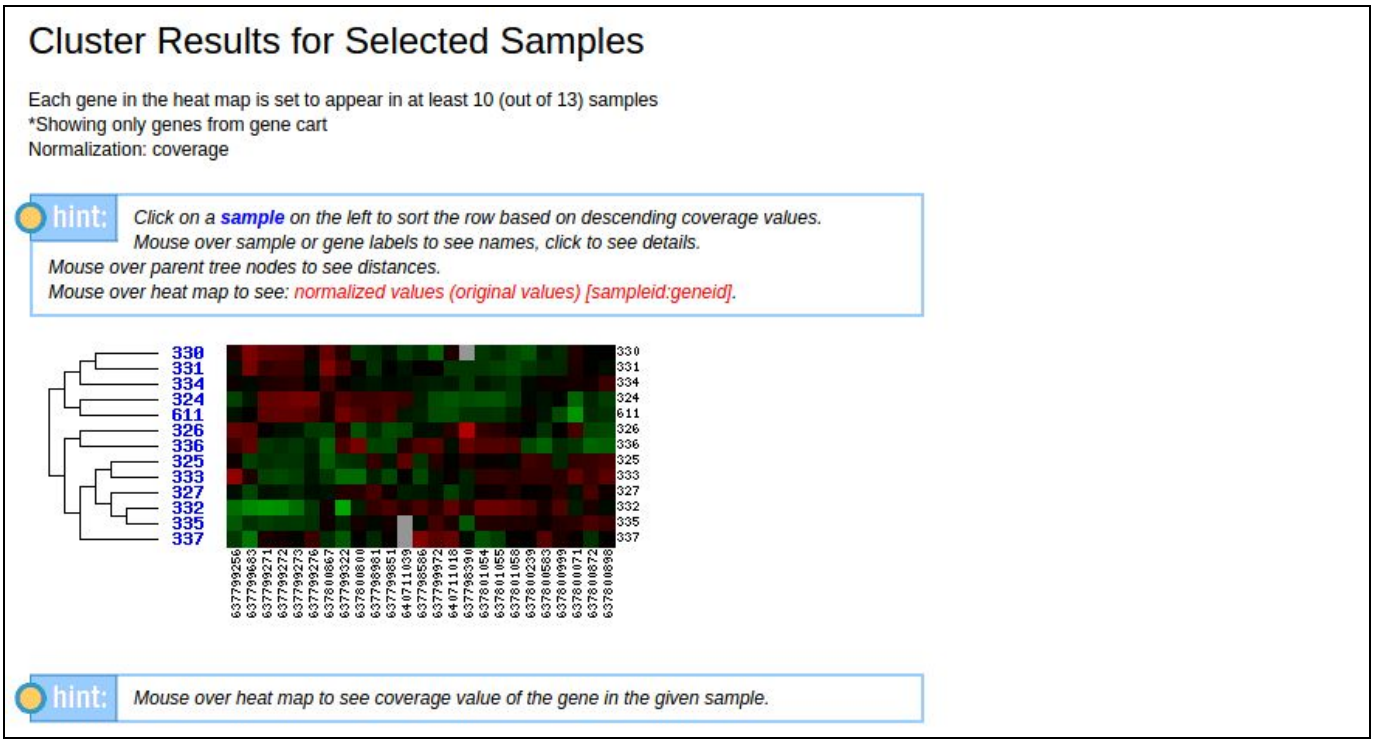


Figure 29. Clustering results – using only genes from Gene Cart