

img INTEGRATED MICROBIAL GENOMES

IMG Home Find Genomes Find Genes Find Functions Compare Genomes Analysis Carts MyIMG Using IMG

IMG Genomes

	finished/draft/perm draft	Total
Bacteria	907/837/4	1748
Archaea	68/9/0	77
Eukarya	19/57/0	76
Plasmids	1051/0/0	1051
Viruses	2605/1/0	2606
All Genomes	4650/904/4	5558

Genome by Metadata
IMG Statistics
Project Map

IMG 3.0: What's New

IMG 3.0 is the 18th release of the Integrated Microbial Genomes (IMG) genomic data management and analysis system. **IMG 3.0** was released on **December 21st, 2009**.

IMG 3.0 Content Genomes

The content of **IMG 3.0** has been updated with new microbial genomes available in **RefSeq version 37** (June 02, 2009). **IMG 3.0** contains a total of **5,558** genomes consisting of **1,748** bacterial, **77** archaeal, **76** eukaryotic genomes, **2,606** viruses (including bacterial phages), and **1,051** plasmids that did not come from a specific microbial genome sequencing project. Among these genomes, **4,650** are **finished** genomes, and **904** are **draft** genomes, and **4** are **permanent draft** (i.e. will never be finished) genomes.

The following **fungus genomes** have been included into **IMG 3.0**:

1. *Ajellomyces capsulatus* NAM1
2. *Aspergillus clavatus* NRRL 1
3. *Aspergillus flavus* NRRL3357
4. *Botryotinia fuckeliana* B05.10
5. *Candida dubliniensis* CD36
6. *Candida tropicalis* MYA-3404
7. *Chaetomium globosum* CBS 148.51
8. *Coccidioides immitis* RS
9. *Coprinopsis cinerea* okayama7#130
10. *Laccaria bicolor* S238N-H82
11. *Lachancea thermotolerans* CBS 6340

12. *Lodderomyces elongisporus* NRRL YB-4239
13. *Moniliophthora perniciosa* FA553
14. *Neosartorya fischeri* NRRL 181
15. *Penicillium chrysogenum* Wisconsin 54-1255
16. *Penicillium marneffeii* ATCC 18224
17. *Phaeosphaeria nodorum* SN15
18. *Pichia guilliermondii* ATCC 6260
19. *Pichia pastoris* GS115
20. *Podospira anserina* DSM 980
21. *Postia placenta* Mad-698-R
22. *Pyrenophora tritici-repentis* Pt-1C-BFP
23. *Schizosaccharomyces japonicus* yFS275
24. *Sclerotinia sclerotiorum* 1980 UF-70
25. *Talaromyces stipitatus* ATCC 10500
26. *Vanderwaltozyma polyspora* DSM 70294
27. *Zygosaccharomyces rouxii* CBS 732

Note that **42** microbial genomes from **IMG 2.9** were **replaced** in **IMG 3.0** because (1) a "Draft" genome has been replaced by its "Finished" version or (2) the composition of the genome has changed through the addition of new replicons (plasmids, chromosomes). For replaced genomes, the gene object identifiers (gene OIDs) for the protein-coding genes (CDS) were mapped to their new version in IMG. 3.0. See IMG [Data Evolution History](#) for details.

The **MetaCyc** (<http://www.metacyc.org/>) pathways in IMG 3.0 have been updated with MetaCyc version 13.5 released on Oct 7, 2009 (<http://biocyc.org/metacyc/release-notes.shtml>).

KEGG pathways, modules, compounds, reactions and KO terms have been updated based on KEGG release 52.0 (October 1, 2009: <http://www.genome.jp/kegg/docs/relnote.html>).

The **Pfam** collection of protein domain families has been updated based on Pfam version 24.0 (October 2009: <http://pfam.sanger.ac.uk/>). **Pfam clans** (<http://pfam.sanger.ac.uk/browse>) are provided as an additional classification of Pfam domain families.

Genes in IMG involved in regulatory interaction experiments controlling their expression are now linked to RegTransBase (<http://regtransbase.lbl.gov>) RegTransBase is a database of regulatory sequences and regulatory interactions on the transcriptional and posttranscriptional levels in prokaryotic genomes, that contains experimental data and predicted sites published in scientific journals.

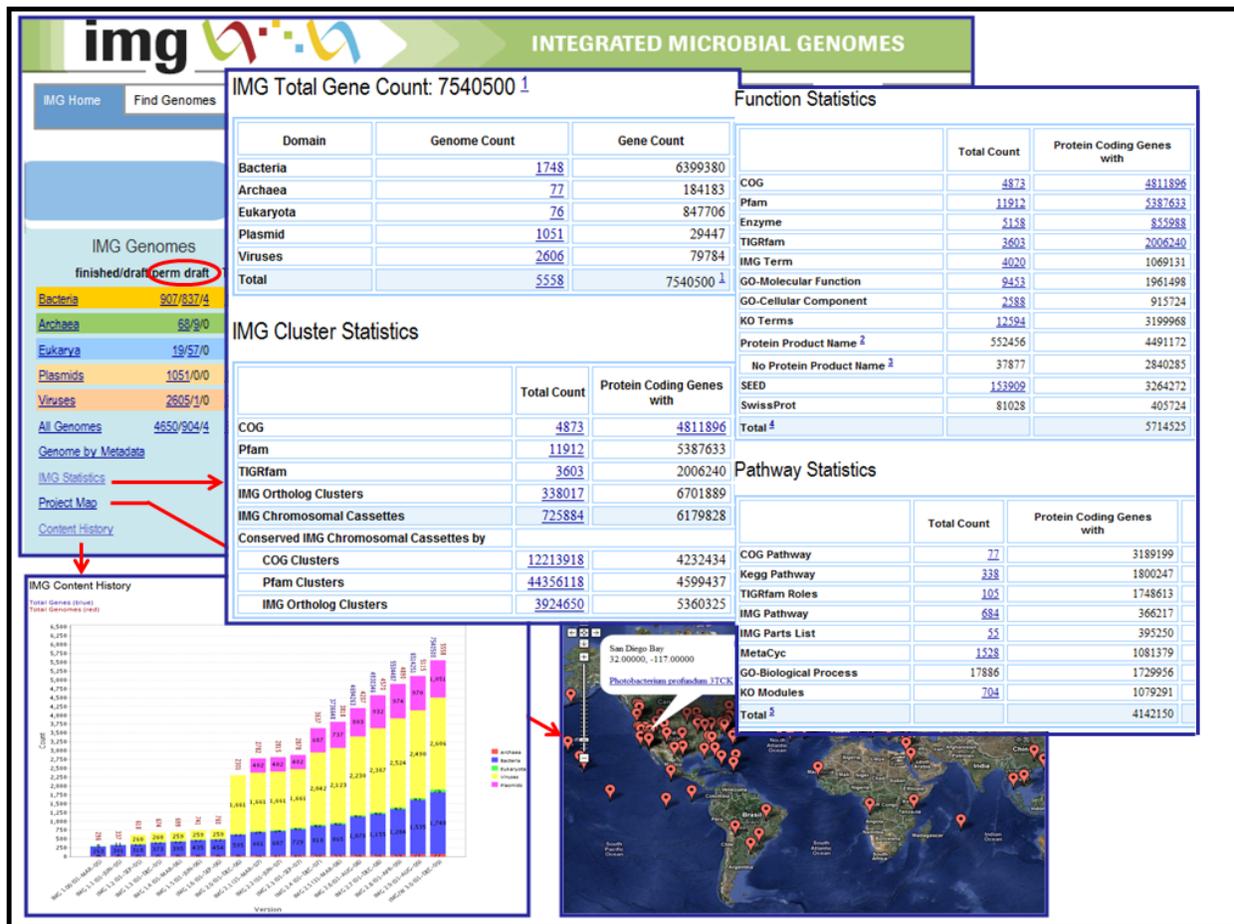
IMG 3.0 also contains **proteomic** data from recent *Arthrobacter chlorophenolicus* study¹, *Cryptobacterium curtum*, and *Brachybacterium faecium* studies.

¹ Unell, M., Abraham, P.E., Shah, M., Zhang, B. et al. (2009) Impact of Phenolic Substrate and Growth Temperature on the *Arthrobacter chlorophenolicus* Proteome. *J. Proteome Res.* **8** (4): 1953-1964.

IMG Statistics

A new **sequencing** status, **permanent draft**, designating genomes that are not targeted for finishing, is available for genomes in IMG 3.0. Four genomes in IMG 3.0 have been specified as permanent draft, as illustrated in the figure below.

Various statistics are provided via the **IMG Statistics** link on the home page of IMG, as shown below, including **IMG Total Gene Count** which consists of counting all the genes (protein coding genes, RNA genes) in IMG, except obsolete genes. Compared to **IMG 2.9**, **IMG 3.0** contains **7,540,500 genes**, an increase of **1,026,256 genes**.



The **Content History** link on the home page of IMG leads to a bar chart, as illustrated above, representing the growth of IMG in terms of number of genomes and genes since the release of its first version in March 2005. Note that the number of genes is recorded only for the last five versions of the system.

The **Project Map** link on the home page of IMG leads to a Google Map, as illustrated above, displaying the location of isolation sites for genomes that are associated with longitude/latitude coordinates in GOLD (<http://www.genomesonline.org/>).

IMG 3.0 User Interface

The User Interface (UI) has been extended with two new tools: **Scaffold Cart** for analyzing individual scaffolds and/or groups of scaffolds, and **Artemis ACT** synteny viewer for pair-wise genome DNA sequence comparison. The UI has been also extended to display additional details **protein expression** data details.

New Features

Organism Details – Scaffold Cart

A new **Scaffold Cart** provides support for analyzing individual scaffolds (rather than all the scaffolds/contigs of a genome) or groups of scaffolds, as illustrated in Figure 1.

Organism Information (i)

Organism Name	Rhizobium leguminosarum bv. trifolii WSM1325
Taxon Object ID	644736401

Genome Statistics

	Number
DNA, total number of bases	7418122
DNA coding number of bases	6484855
DNA G+C number of bases	4507991
DNA scaffolds	5
CRISPR Count	0
Plasmid Count	5

Chromosome Viewer (ii)

Scaffolds and contigs for Rhizobium leguminosarum bv. trifolii WSM1325

User Selectable Coordinates

Select	Scaffold	Length (bp)	GC	Type	Topology	No. Genes
<input type="checkbox"/>	Rhizobium leguminosarum bv. trifolii WSM1325 plasmid pR132501: NC_012848	828924	0.60	plasmid	circular	769
<input type="checkbox"/>	Rhizobium leguminosarum bv. trifolii WSM1325 plasmid pR132502: NC_012858	660973	0.61	plasmid	circular	661
<input checked="" type="checkbox"/>	Rhizobium leguminosarum bv. trifolii WSM1325 plasmid pR132503: NC_012853	516088	0.59	plasmid	circular	560
<input type="checkbox"/>	Rhizobium leguminosarum bv. trifolii WSM1325 plasmid pR132504: NC_012852	350312	0.61	plasmid	circular	315
<input type="checkbox"/>	Rhizobium leguminosarum bv. trifolii WSM1325 plasmid pR132505: NC_012854	294782	0.60	plasmid	circular	287
<input checked="" type="checkbox"/>	Rhizobium leguminosarum bv. trifolii WSM1325: NC_012850	4767043	0.61	chromosome	circular	4700

Add to Scaffold Cart Select All Clear All

FIGURE 1. Scaffold Cart: Selection.

The **Scaffold Cart** tools provide support for analyzing genomes at the level of individual scaffolds and contigs, such as individual chromosomes and plasmids. For a genome of interest, such as *Rhizobium leguminosarum bv. trifolii WSM1325*, the selection of scaffolds and contigs starts by following the **DNA scaffolds** link in the **Genome Statistics** section of the **Organism Details** page, as illustrated in Figure 1(i). The scaffolds and contigs that are listed on the **Chromosome Viewer** page can be then selected and added to the **Scaffold Cart** as illustrated in Figure 1(ii), where the chromosome and one of the plasmids of *Rhizobium leguminosarum bv. trifolii WSM1325* are selected for further analysis.

Scaffold Cart provides several tools for analyzing and handling scaffolds, including tools for including the genes of one or several scaffolds into the **Gene Cart**, associating a name with selected scaffolds for further analysis, computing a function profile across selected scaffolds in

the cart, and for examining the phylogenetic distribution of genes for one or several scaffolds in the cart.

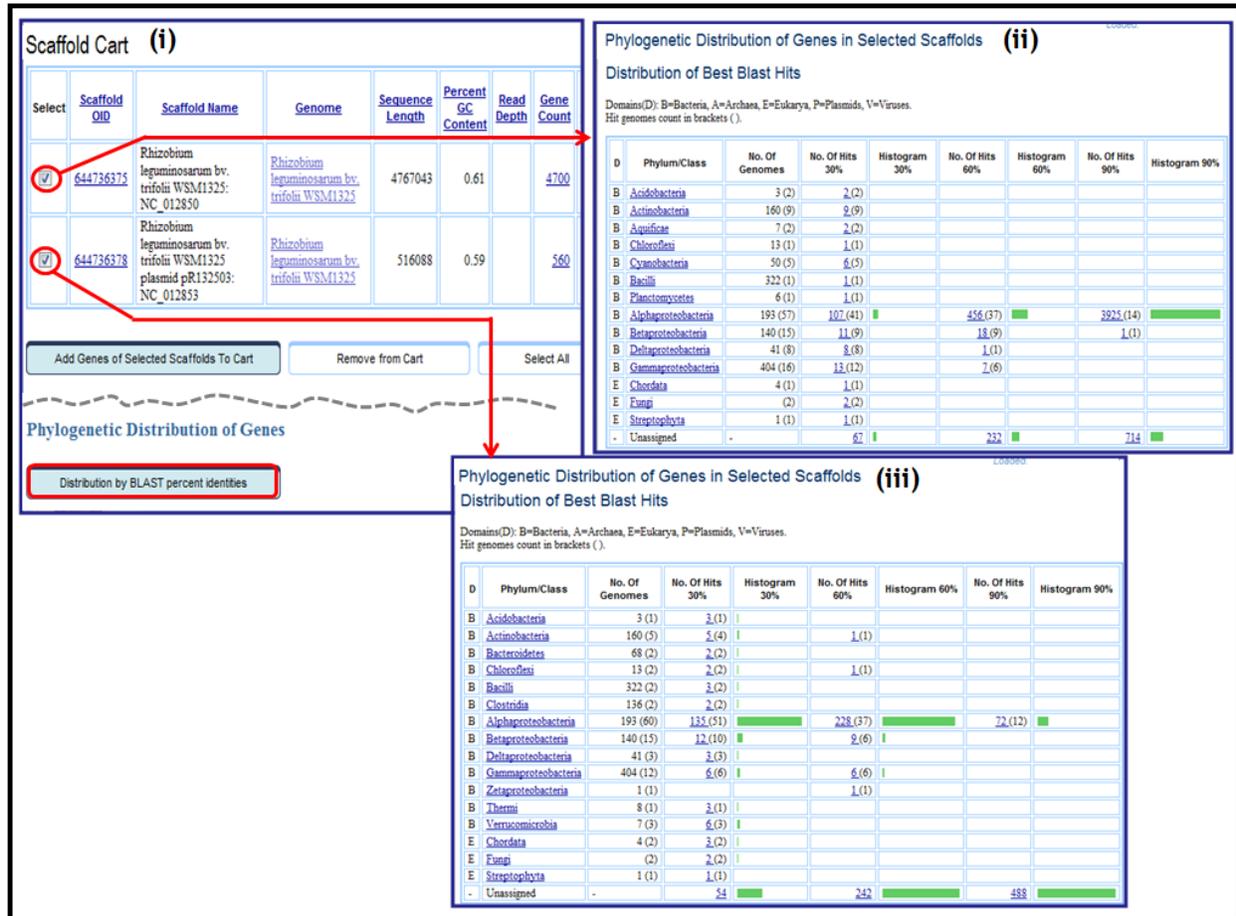


FIGURE 2. Scaffold Cart: Comparing Phylogenetic Distribution of Genes.

Consider an example in which the phylogenetic distributions of best BLASTp hits are compared for the genes on the chromosome and one of the plasmids of *Rhizobium leguminosarum* bv. *trifolii* WSM1325, as illustrated in Figure 2(i). The phylogenetic distribution of the best BLASTp hits for the genes of the chromosome, shown in Figure 2(ii), indicates that the vast majority of the genes have their best BLASTp hits within the class of *Alphaproteobacteria* with greater than 90% identity. Furthermore, the majority of the hits to *Alphaproteobacteria* turn out to be to the members of *Rhizobiaceae* family and the genus *Rhizobium*, again with greater than 90% sequence identity (not shown). Conversely, only a small fraction of the plasmid genes have hits with greater than 90% identity to organisms from *Alphaproteobacteria*, as shown in Figure 2(iii), and the distribution of the hits between the members of *Rhizobiaceae* family is quite different from that of the chromosomal genes. These differences indicate that while the majority of chromosomal genes are vertically inherited, the plasmid carries a number of genes that may have been horizontally transferred from other lineages and may reflect a broader host specificity of this plasmid.

Synteny Viewers – Artemis ACT

ACT (Artemis Comparison Tool) is a viewer based on Artemis that allows examining pair-wise genome DNA sequence comparisons, as illustrated in Figures 3 and 4. **ACT** is grouped together with **VISTA** and **Dotplot** viewers under the new **Synteny Viewers** submenu of **Comparing Genomes**, as illustrated in Figure 3(i).

ACT Genome Selection (ii)

Select at least 2 genomes to compare against.
Max. 5 selections.

Genome List Filter

Seq. Status: Finished
Domain: Bacteria

Genome List:

- Psychromonas ingrahamii 37 (B)[F]
- Ralstonia eutropha H16 (B)[F]
- Ralstonia eutropha JMP134 (B)[F]
- Ralstonia metallidurans CH34 (B)[F]
- Ralstonia pickettii 12D (B)[F]
- Ralstonia pickettii 12J (B)[F]
- Ralstonia solanacearum GM1000 (B)[F]
- Renibacterium salmoninarum ATCC 33209 (B)[F]
- Rhizobium etli CFN 42 (B)[F]
- Rhizobium etli CIAT 652 (B)[F]
- Rhizobium leuconacearum bv. trifolii WSM1325 (B)[F]

Pairwise Selection (iii)

Order	Genome Id	Genome Name
1	637000229	Ralstonia eutropha JMP134
2	637000230	Ralstonia metallidurans CH34

Pairwise Selection (iv)

Order	Genome Id	Genome Name
1	637000230	Ralstonia metallidurans CH34
2	637000229	Ralstonia eutropha JMP134

Contig Reorder - Artemis - ACT (v)

Ralstonia metallidurans CH34

Ignore	Order	Ext Accession	Scaffold Name	Seq. Length
<input checked="" type="checkbox"/>	1	NC_007971	Ralstonia metallidurans CH34 plasmid 1: NC_007971	233720
<input checked="" type="checkbox"/>	2	NC_007972	Ralstonia metallidurans CH34 plasmid 2: NC_007972	171459
<input type="checkbox"/>	3	NC_007973	Ralstonia metallidurans CH34 chromosome 1: NC_007973	3928089
<input type="checkbox"/>	4	NC_007974	Ralstonia metallidurans CH34 chromosome 2: NC_007974	2580084

Ralstonia eutropha JMP134

Ignore	Order	Ext Accession	Scaffold Name	Seq. Length
<input checked="" type="checkbox"/>	1	NC_007336	Ralstonia eutropha JMP134 megaplasmid: NC_007336	634917
<input checked="" type="checkbox"/>	2	NC_007337	Ralstonia eutropha JMP134 plasmid 1: NC_007337	87688
<input type="checkbox"/>	3	NC_007347	Ralstonia eutropha JMP134 chromosome 1: NC_007347	3806533
<input type="checkbox"/>	4	NC_007348	Ralstonia eutropha JMP134 chromosome 2: NC_007348	2726152

FIGURE 3. Artemis ACT: Genome and Contig Selection and Ordering.

The first step of using ACT involves selecting **up to 5** genomes of interest, as illustrated in Figure 3(ii), where *Ralstonia eutropha JMP134* and *Ralstonia metallidurans* are selected from the list of genomes. Two filters (Sequencing Status, Domain) can be used for reducing the list of all genomes available in IMG. The selected genomes are compared in pairs, where the pairs are determined by the order in which the genomes were selected. For example, if genomes *G1*, *G2*, *G3*, and *G4* were selected, the *G1* will be compared to *G2*, *G2* will be compared to *G3*, and *G3* will be compared to *G4*. The pairwise comparisons can be adjusted using the Pairwise Selection page, as illustrated in Figure 3(iii), where by typing “2” in the “Order” column for *Ralstonia eutropha JMP134*, the order of the two selected genomes is switched, as illustrated in Figure 3(iv).

Next, all the scaffolds and contigs of the selected genomes are displayed in the **Contig Reorder** page, as illustrated in Figure 3(v). This page can be used to remove the contigs that should not be included in the comparison, as well as for reordering the contigs that are included

in the comparison. In the example shown in Figure 3(v), the plasmids for both *Ralstonia eutropha JMP134* and *Ralstonia metallidurans* are marked as “ignored”, while the order in which the chromosomes will be displayed with **ACT** is left unchanged. The **Contig Reorder** page can be refreshed using the “Update” button, in order to remove the contigs marked as “ignored”, as illustrated in Figure 4(i).

Once the list of contigs involved in the comparison and their order has been finalized, several files are prepared for ACT, as illustrated in Figure 4(ii), including the FASTA and Genbank files for each genome and the result of Mega BLAST.

Figure 4 illustrates the Artemis ACT workflow. Panel (i) shows the 'Contig Reorder - Artemis - ACT' interface for *Ralstonia metallidurans* CH34 and *Ralstonia eutropha* JMP134. It features two tables for selecting contigs to ignore or reorder. The 'Update' button is highlighted with a red box, and a red arrow points to the 'Next' button. Panel (ii) shows the 'Artemis - ACT' interface, which lists 'ACT Data Files' (FASTA and Genbank files for both genomes and a MEGABLAST output file) and a 'Run ACT' button. A red arrow points from the 'Run ACT' button to panel (iii). Panel (iii) displays the MEGABLAST comparison results, showing two sequence tracks with matching regions highlighted in red and blue bands. The top track is labeled 'Emet_0001', 'Emet_0002', and 'Emet_0003'. The bottom track is labeled 'Reut_A0001', 'Reut_A0002', and 'Reut_A0003'. The x-axis for both tracks ranges from 0 to 6400.

FIGURE 4. Artemis ACT: Computing MEGABLAST and displaying comparisons.

ACT allows comparing genome sequences and associated annotations², as illustrated in Figure 4 (iii). ACT allows examining regions of similarity, insertions and rearrangements. For a pair of sequences, one is designated the query sequence and the other the subject sequence. The sequences are joined by colored bands that represent the matching regions: red and blue bands represent the forward and reverse matches, respectively. Double clicking on one of these bands will centre the associated matching regions in each sequence, and the sequences can be

² Carver, T.J., Rutherford, K.M., Berriman, M. et al. (2005) ACT: the Artemis Comparison Tool. *Bioinformatics* **21**(16): 3422-3423.

scrolled individually or together. A highlighted match region turns yellow and information about that BLAST hit is shown, namely the score and percentage identity. ACT re-uses the Artemis display the sequence and features³. ACT is written in Java and can run on any Java enabled platform, including UNIX, Macintosh and Windows.

Extended Features

Organism Details – Protein Expression Studies

IMG 3.0 contains protein expression data from three protein expression studies. The tools for displaying these data have been extended, as illustrated in Figure 5.

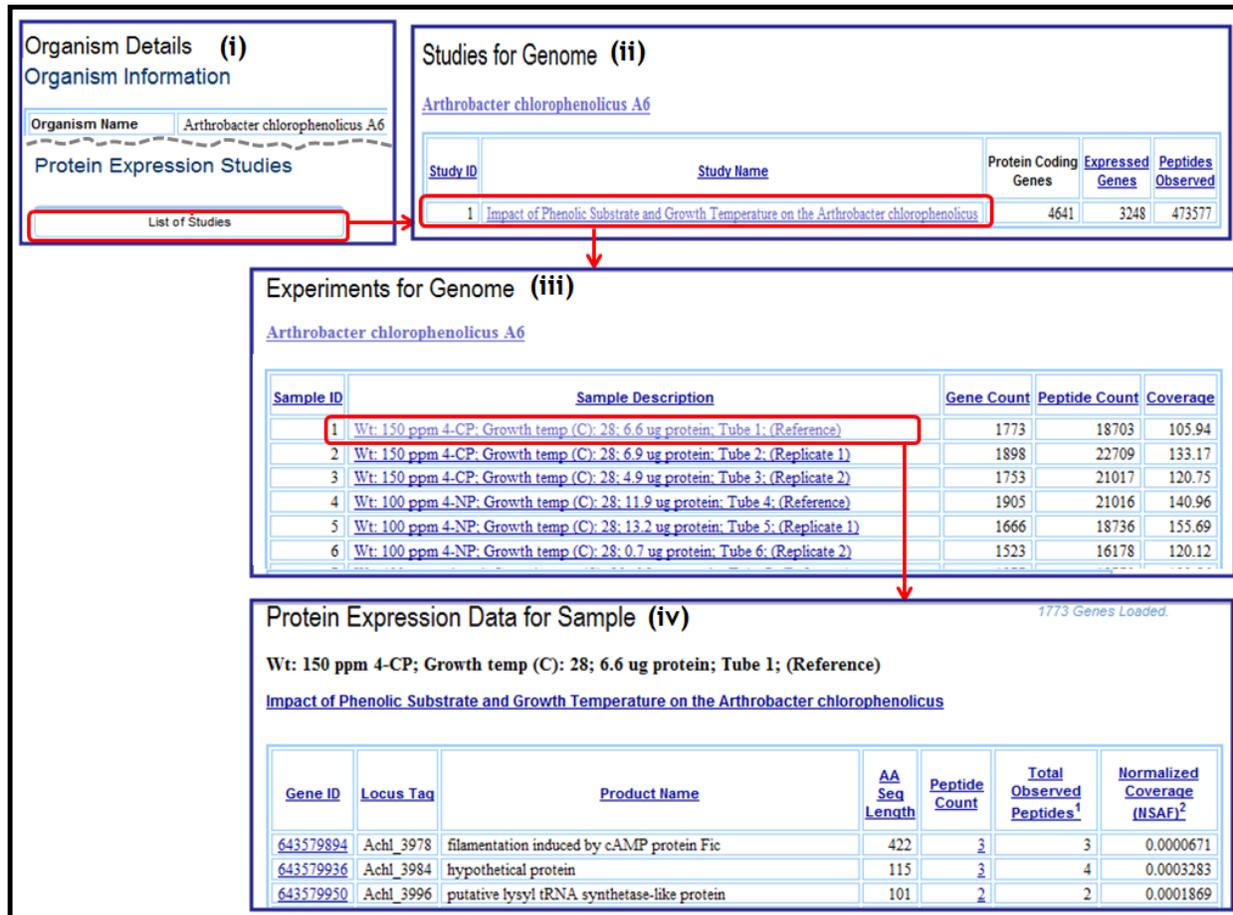


FIGURE 5. Organism Details- Protein Expression Studies.

The **Protein Expression Studies** link on the **Organism Details** for a genome associated with protein expression studies, illustrated in Figure 5(i), leads to the list of studies for the genome of interest, as illustrated in Figure 5(ii). Each protein expression study is displayed together with the number of expressed genes and observed peptides. The **Study Name** provides a link to the list of experiments/samples for the study, as illustrated in Figure 5(iii).

³ Berriman, M. and Rutherford, K. (2003) Viewing and annotating sequence data with Artemis. *Brief. Bioinformatics*, 4, 124–132.

The description for each sample consists of the experimental conditions and provides a link to the protein expression data for the sample organized per expressed gene, as illustrated in Figure 5(iv). For each expressed gene, the number of observed peptides and **normalized coverage** (NSAF) are provided. **NSAF** is the coverage for an expressed gene in an experiment divided by the total coverage of all the genes in that experiment, whereby the **coverage** for a gene is defined as the sum of the sizes of all observed peptides for the gene divided by the size of the gene⁴.

⁴ Florens, L., Carozza, M.J., Swanson, S.K. et al. (2006) Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors, *Methods* **40**(4): 303-311.