IMG 2.9 is the **17th** release of the Integrated Microbial Genomes (IMG) genomic data management and analysis system. **IMG 2.9** was released on **August 13th, 2009.**

IMG 2.9 Content

Genomes

The content of **IMG 2.9** has been updated with new microbial genomes available in **RefSeq version 35** (June 02, 2009). **IMG 2.9** contains a total of **5,115** genomes consisting of **1,535** bacterial, **71** archaeal, **49** eukaryotic genomes, **2,490** viruses (including bacterial phages), and **970** plasmids that did not come from a specific microbial genome sequencing project. Among these genomes, **4,368** are finished genomes, and **747** are draft genomes.

Note that **10** microbial genomes from **IMG 2.8** were **replaced** in **IMG 2.9** because (1) a "Draft" genome has been replaced by its "Finished" version or (2) the composition of the genome has changed through the addition of new replicons (plasmids, chromosomes).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. 2.9. See IMG Data Evolution History for details.

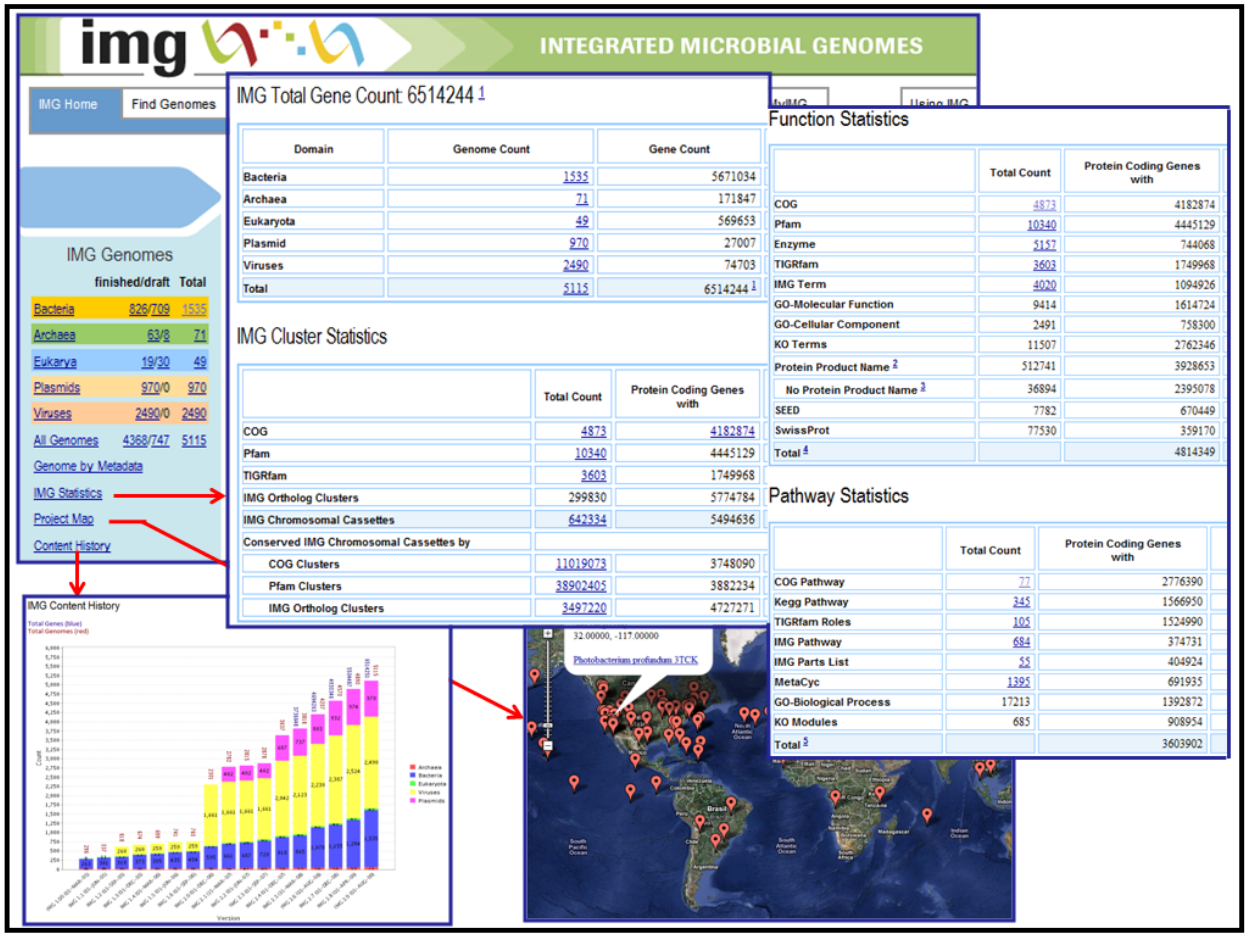
tRNA and **rRNA** (23S, 16S, 5S) genes missing from original RefSeq genome files are added using tRNAscan-SE v1.23 for tRNA genes and similarity comparisons to existing RNA genes. In IMG 2.9 **23** tRNA, **75** rRNA and **5** misc_RNA genes were added in **40** genomes.

IMG 2.9 also contains **protein expression** data from a recent *Arthrobacter chlorophenolicus* study[1].

[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

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 2.8**, **IMG 2.9** contains **6,514,244 genes**, an increase of **1,009,757** 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 2.9 User Interface

The User Interface (UI) has been extended in order to improve its overall functionality and usability.

The main UI changes include:

(a) **New features**

- (i) A new phylogenetic tree based **Genome Browser**.
- (ii) A gene cassette based **Gene Search**.
- (iii) Tools for exploring **Horizontally Transferred Genes**.
- (iv) A **dotplot** tool for comparing the genome sequences.
- (v) Tools for exploring **protein expression** data.

(b) **Extended features**

- (i) The **Gene Search** of **Find Genes** has been extended with two genome filters and a new locus tag search capability.
- (ii) The **Ortholog Neighborhood Viewer** of **Gene Details** has been extended with the ability to reduce the number of displayed neighborhoods and to select genes on the viewer.
- (iii) The **Function Profile** tool available under **Analysis Carts /Functions** and **Compare Genomes** has been extended with a "slim view" option which allows to display only column identifiers instead of genome or function name abbreviations, and to remove the rows containing only zeros from the result of a function profile.

New Features

Genome Browser – Phylogenetic Tree View

A new version of the **Genome Browser** provides improved functionality is exploring and selecting genomes, as illustrated in Figure 1.

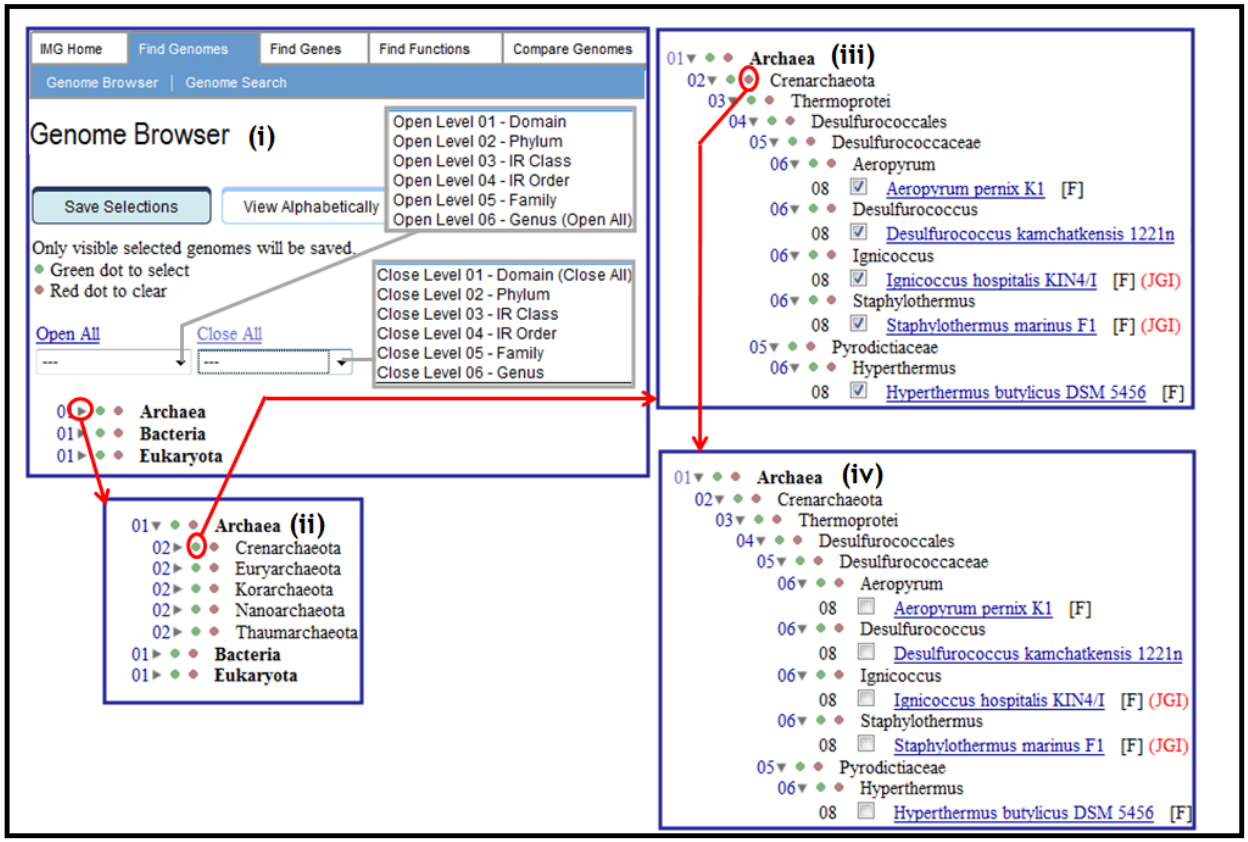


FIGURE 1. Genome Browser: Phylogenetic Tree View.

The phylogenetic tree based version of the **Genome Browser** initially displays the three genome domains, as illustrated in Figure 1(i). The genome display can be modified using the "Open All" and "Close All" options (see left side of Figure 1(i)), or using the open/close icons available at each level of the tree, as illustrated in Figure 1(ii).

Genomes can be **selected** either (a) individually using the select boxes associated with each strain, or (b) in groups using the green dot "select all" icons available at each level of the tree. For example, clicking the "select all" (green dot) icon associated with *Crenarchaeota*, as illustrated in Figure 1(ii), will both expand the subtree under this phylum down to individual genomes and select all these genomes, as illustrated in Figure 1(iii).

Genomes can be **unselected** (**cleared**) either (a) individually using the select boxes associated with each strain, or (b) in groups using the red dot "clear all" icons available at each level of the tree. For example, clicking the "clear all" (red dot) icon associated with *Crenarchaeota*, as illustrated in Figure 1(iii), will unselect (clear) all the genomes in this phylum, as illustrated in Figure 1(iv).

Gene Search – *Cassette Search*

Cassette Search is a new gene search tool provided under the **Find Genes** main menu option, as illustrated in Figure 2.

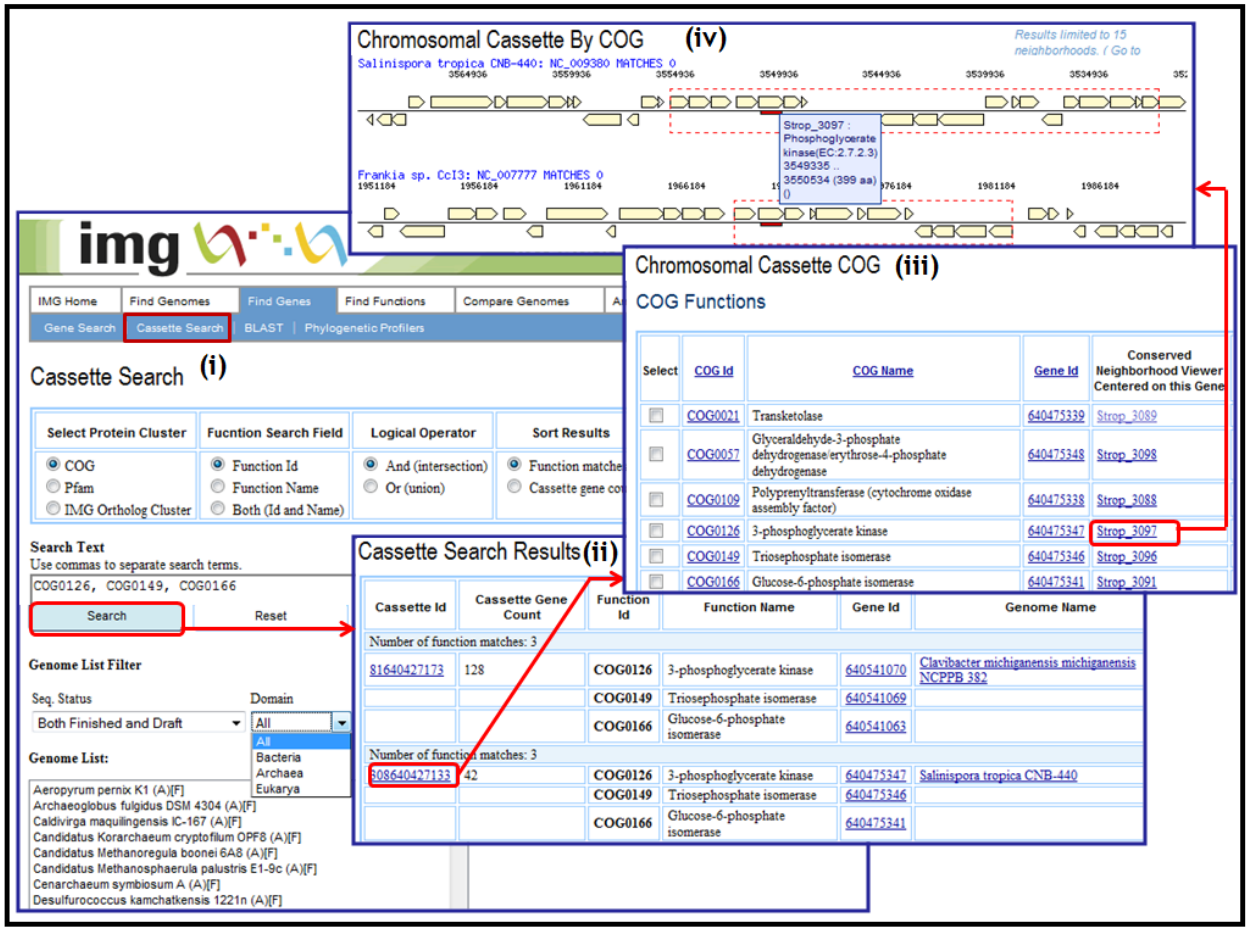


FIGURE 2. Gene Search- Cassette Search.

Cassette Search allows searching genes that are part of ***chromosomal cassettes***[2] involving specific protein clusters. First, select the protein cluster underlying the cassettes (COG, Pfam, or IMG Ortholog Cluster), the protein cluster identifier for the search (cluster/function identifier, cluster/function name, or both), the logical operator (either "and" or "or") used for the search expression and the order of presenting the search results. In the example shown in Figure 2(i), the search involves COG clusters, cluster identifiers are used for the search condition which involves "and": search genes that are part of cassettes involving COG0124 <u>and</u> COG0149 <u>and</u> COG 0166 (which are part of the Glycolysis pathway).

By default, the search is carried out across all the genomes in IMG. The search can be limited only to (i) finished or draft genomes using the **Seq Status** filter, (ii) bacterial, archaeal, or eukaryotic genomes using the **Domain** filter, or (iii) genomes selected using the **Genome List.**

[2] A chromosomal cassette is defined as a stretch of protein coding genes with intergenic distance smaller or equal to 300 base pairs.

The **Cassette Search Result** lists the genes that satisfy the search condition, together with the identifiers of the cassettes they are part of, their associated protein cluster identifiers and names, and their genomes, as illustrated in Figure 2(ii). The cassette identifiers provide links to the **Chromosomal Cassette** details page, as illustrated in Figure 2(iii). The **Chromosomal Cassette** viewer can be accessed from this page, as illustrated in Figure 2(iv).

Organism Details – Putative Horizontally Transferred Genes

Putative Horizontally Transferred Genes is available as part of a genome's **Organism Details** page, as illustrated in Figure 3. This tool allows exploring genes in a query genome that are likely horizontally transferred from genomes in phylogenetic groups that are different than the group the query genome belongs to.

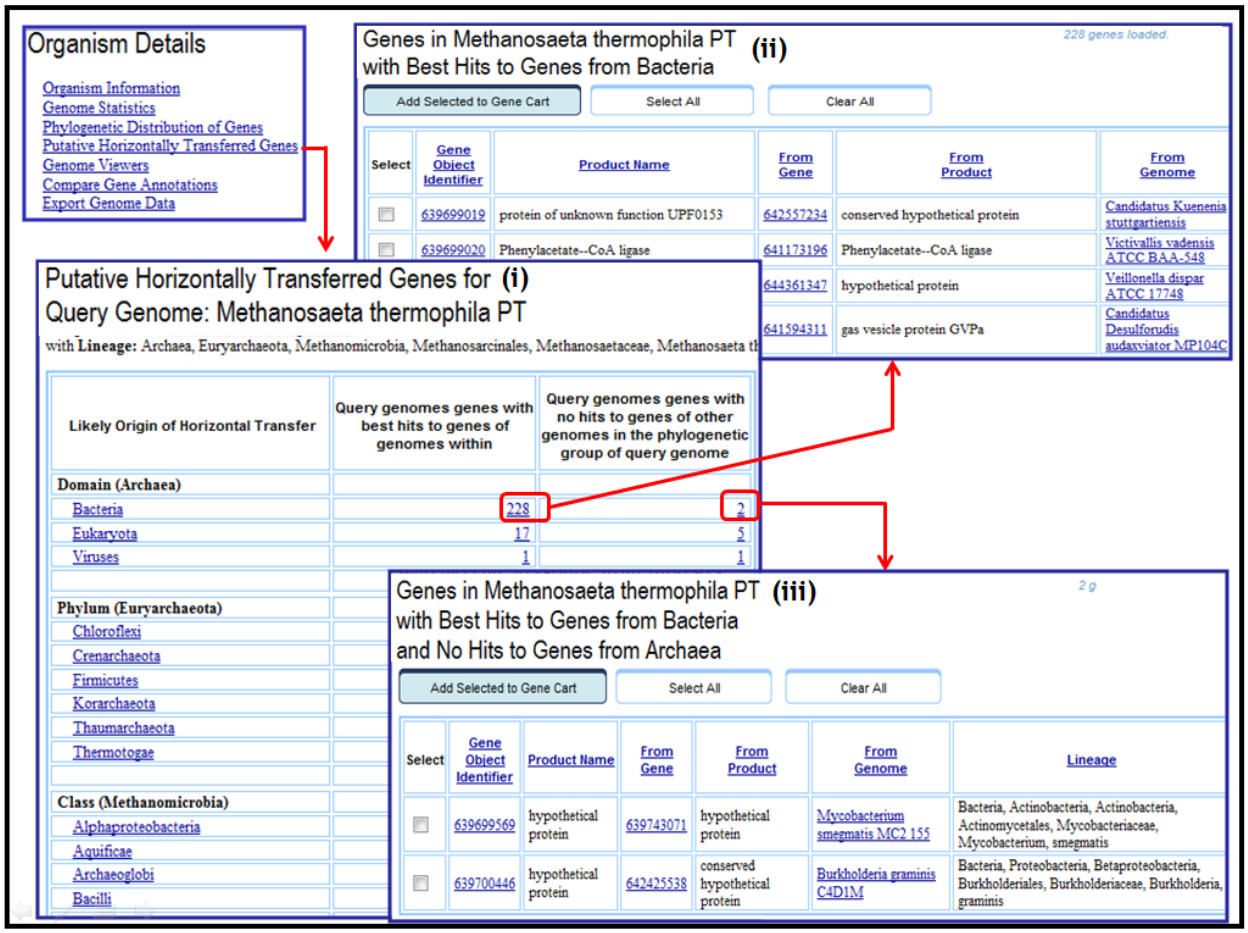


FIGURE 3. Organism Details – Putative Horizontally Transferred Genes.

Putative horizontally transferred genes are defined as genes that have best hits (best bitscores) to genes that don't belong to the phylogenetic group of the query genome. In this calculation we use not only the best hit (i.e. the hit with the best bitscore) but all the hits that have bitscore equal or greater than 95% of the best hit.

For a query genome ***QG***, such as *Methanosaeta thermophila PT*, two lists of genes are provided, as illustrated in Figure 3(i):

1. Genes **with best hits** (best bit score) to genes of genomes within a phylogenetic group (domain, phylum, class, etc.) that is different than the analogous group ***QG*** belongs to;
    - For example, as an archaeal genome, *Methanosaeta thermophila PT* has 228 genes with best hits to bacterial genomes, 17 genes with best hits to eukaryotic genomes, and 1 gene with best hits to viral genomes. These genes may be horizontally transferred genes from bacterial, eukaryotic, or viral genomes, respectively, as illustrated in Figure 3(ii).
2. Genes **with best hits** to genomes within a **phylogenetic group** (domain, phylum, class, etc.) that is **different** than the analogous group ***QG*** belongs to, **and no hits** to genes of genomes within the **same phylogenetic group** (domain, phylum, class, etc.) as the group ***QG*** belongs to;
    - For example, *Methanosaeta thermophila PT* has 2 genes with best hits to bacterial genomes and no hits to other archaeal genomes, as illustrated in Figure 3(iii), with a higher likelihood of being horizontally transferred from bacterial genomes.

Compare Genomes – *Dotplot*

Dotplot is a new tool for comparing the sequences of two genomes, as illustrated in Figure 4.

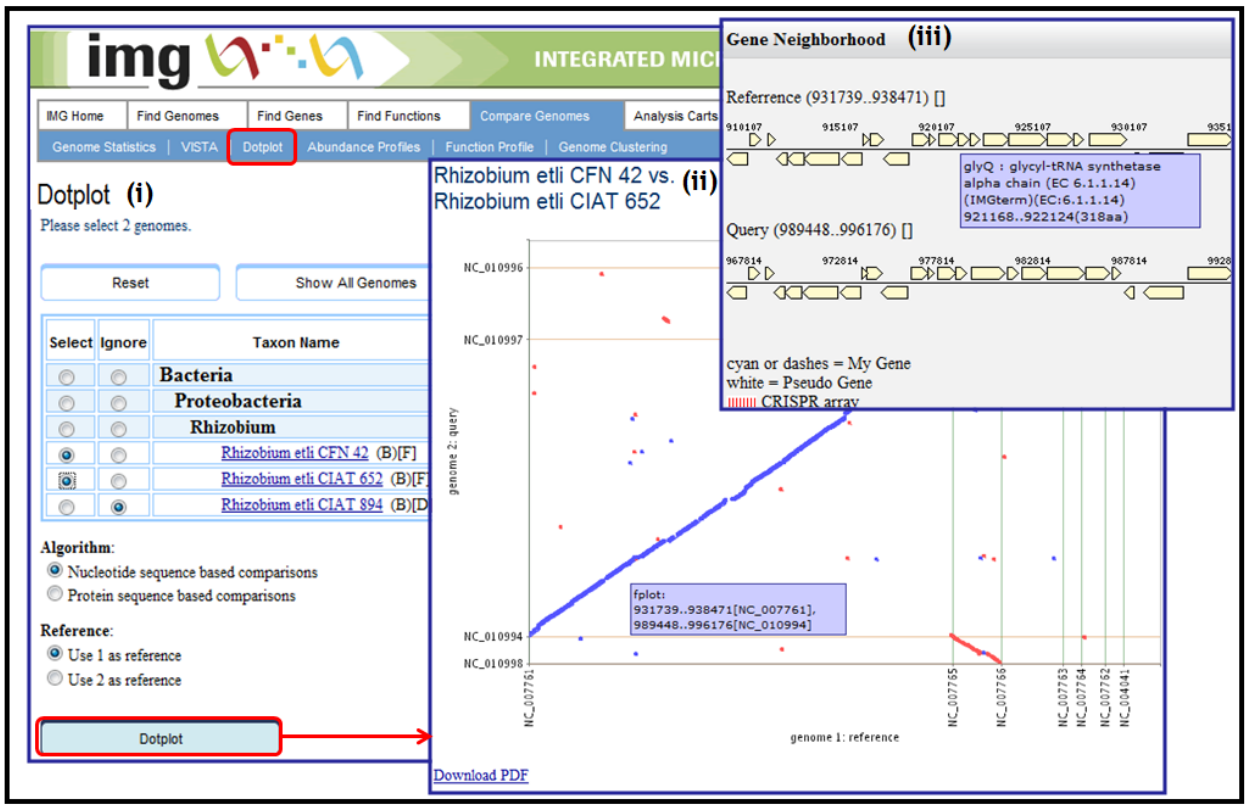


FIGURE 4. Compare Genomes- Dotplot.

Dotplot employs the program ***Mummer*** to generate dotplot diagrams between two genomes, whereby **nucleotide sequences** are used for genomes with fairly similar sequences and

protein sequences are used for genomes with less similar nucleotide sequences. **Dotplot** can be applied on two genomes that are selected from the list of all available or preselected genomes, as illustrated in Figure 4(i). Select whether to use **nucleotide based comparison** for genomes with similar nucleotide sequences or **protein based comparison** for genomes with dissimilar nucleotide sequences, and select the genome that will be used as reference.

The dotplot, illustrated in Figure 4(ii), consists of **blue** points for the regions of similarity found on parallel strands (fplot) and **red** points for the regions of similarity found on antiparallel strands (rplot). The tooltip for each point shows which plot it is and the coordinates and scaffolds of the alignment. Clicking on a point brings up the chromosomal neighborhood of the alignment of the two genomes, as illustrated in Figure 4(iii).

Organism Details – *Protein Expression Studies*

IMG 2.9 contains **protein expression** data from a recent *Arthrobacter chlorophenolicus* study**.** These data can be explored using the tools illustrated in Figures 5 and 6.

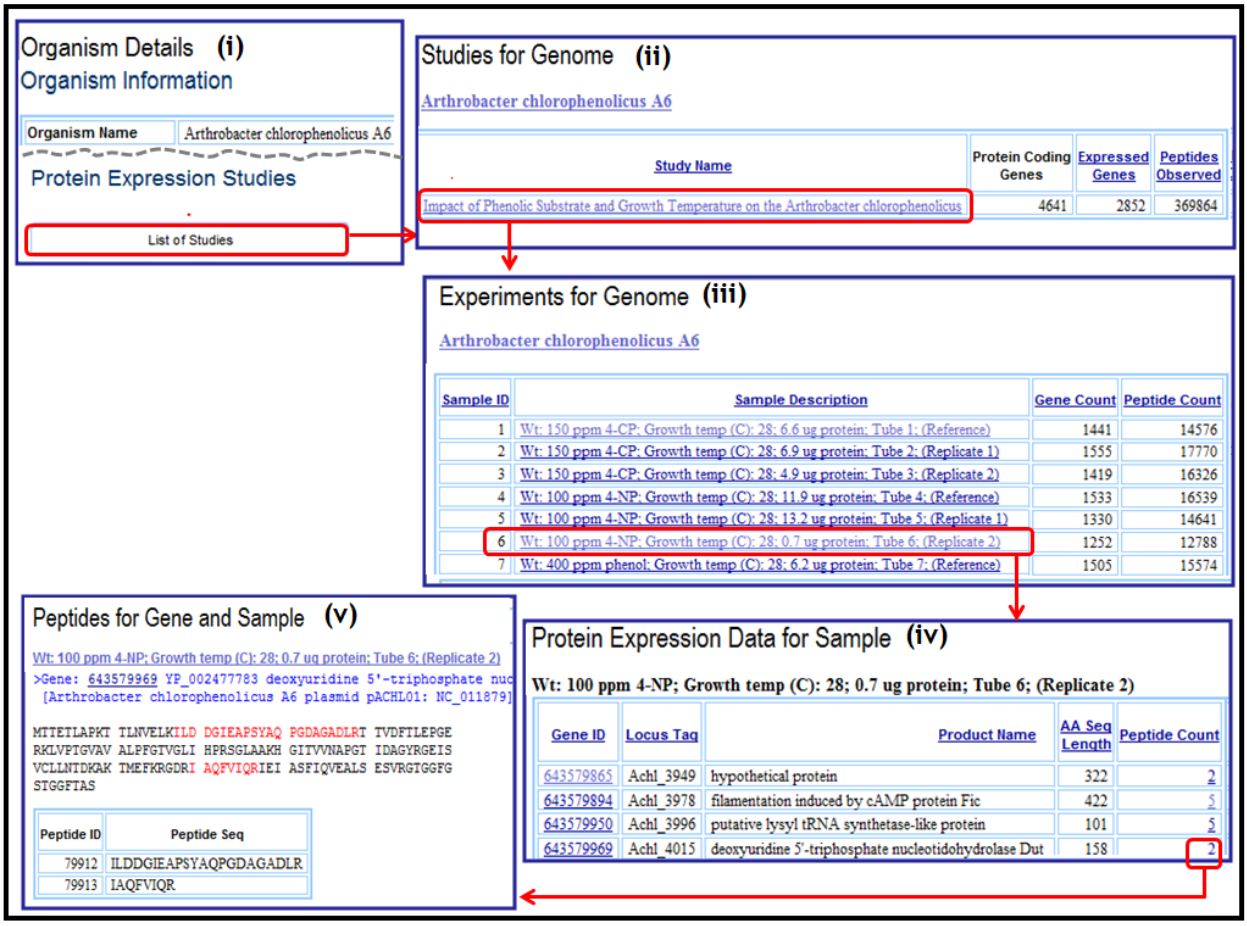


FIGURE 5. Organism Details- Protein Expression Studies.

The **Organism Details** for a genome associated with protein expression studies provides two starting points for exploring protein expression data. First, a **Protein Expression Studies** link on this page, 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). 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 leads to the peptide details page, illustrated in Figure 5(v), where the peptide sequences are displayed aligned on the gene's protein sequence.

The **Organism Details** provides another starting point for exploring protein expression data through the number of expressed genes in the Genome Statistics section, as illustrated in Figure 6(i). This number leads to the list of expressed genes, with each gene linked to its **Gene Detail** page, as illustrated in Figure 6(ii).

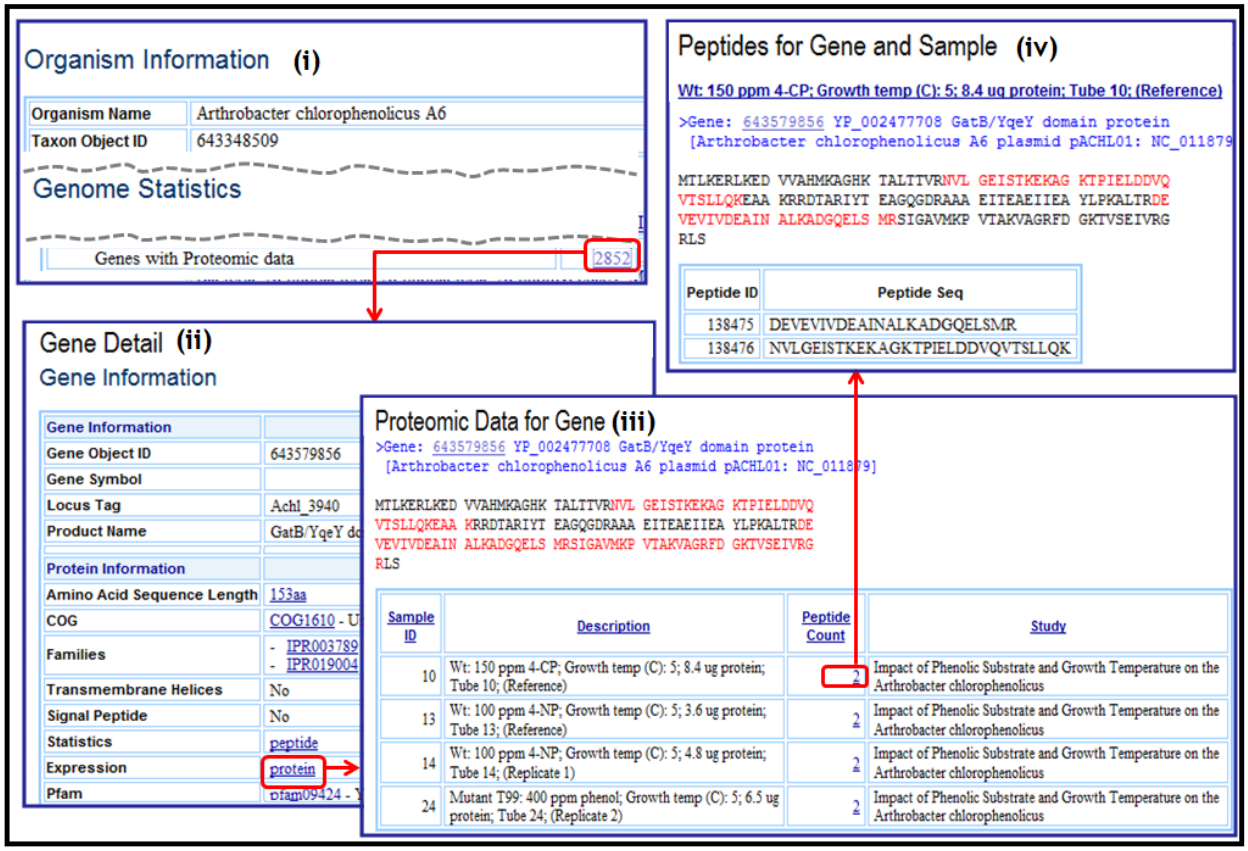


FIGURE 6. Gene Details- Protein Expression Data.

The **Protein Information** section of the **Gene Detail** for an expressed gene provides a link to a **Proteomic Data** page, as illustrated in Figure 6(iii), which displays the list of experiments/ samples involving the expressed gene. Each sample is associated with a description consisting of the experimental conditions and the number of peptides observed for the expressed gene as part of this experiment. The number of peptides leads to a page displaying peptide details page, illustrated in Figure 6(iv), where the peptide sequences are displayed aligned on the gene's protein sequence.

Extended Features

Find Genes –*Gene Search*

The **Gene Search** of **Find Genes** has been extended with two genome filters and a new locus tag search capability, as illustrated in Figure 7.

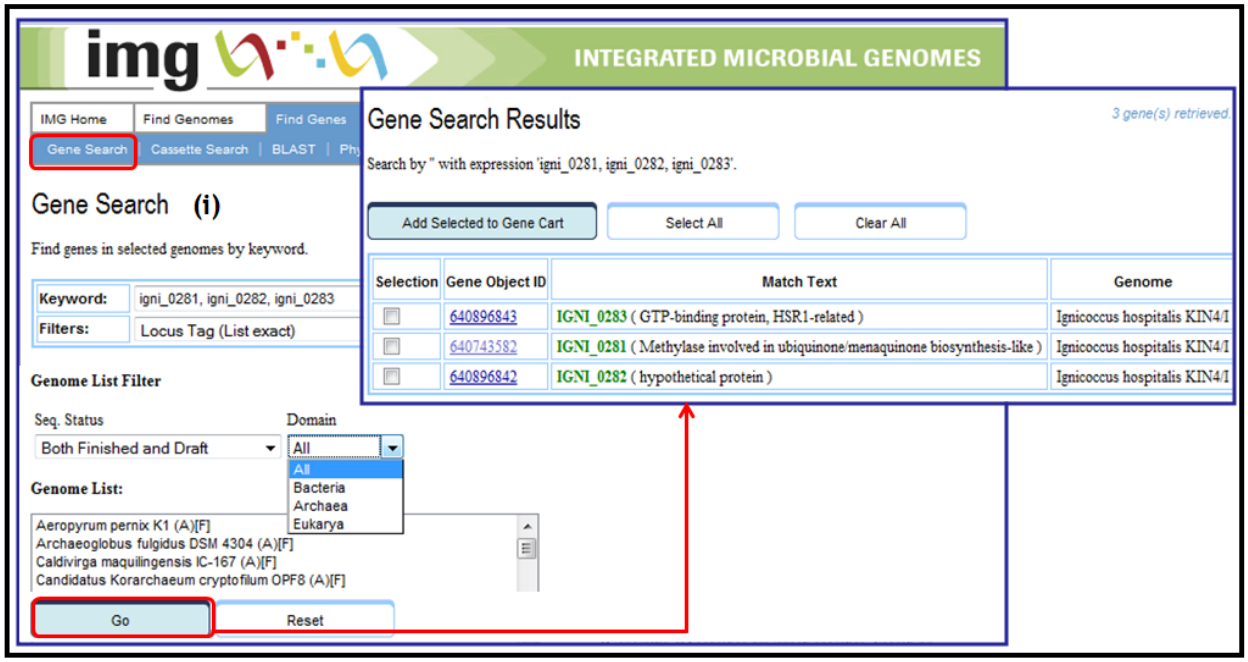


FIGURE 7. Find Genes- Gene Search.

Two filters allow limiting the gene search only to (i) finished or draft genomes using the **Seq Status** filter, or (ii) bacterial, archaeal, or eukaryotic genomes using the **Domain** filter.

A new "Locus Tag (List exact)" filter allows searching genes using a list of locus tags separated by commas, as illustrated in Figure 7(i). The **Gene Search Result** lists the genes that are associated with one of the locus tags in this list, as illustrated in Figure 7(ii).

Gene Details –*Ortholog Neighborhood Viewer*

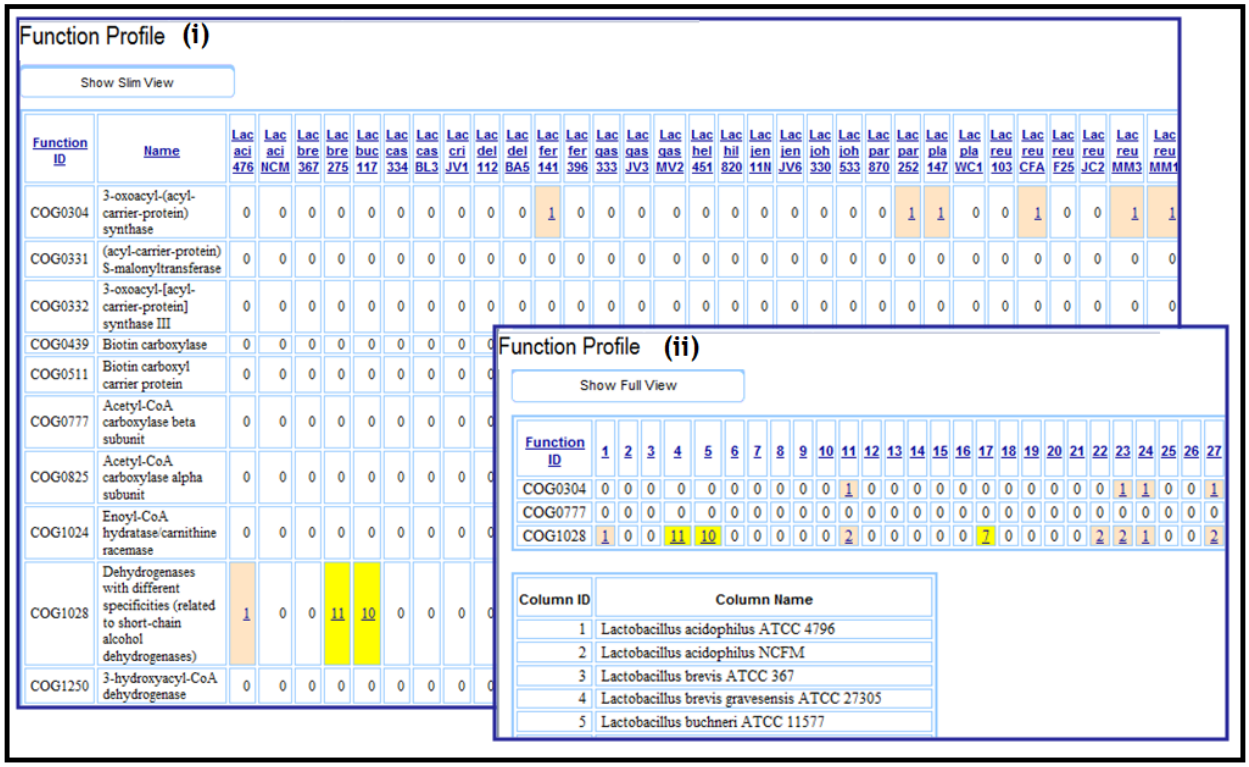
The **Ortholog Neighborhood Viewer** of **Gene Details** has been extended with the ability to reduce the number of displayed neighborhoods and to select genes on the viewer, as illustrated in Figure 8.

FIGURE 8. Gene Details- Ortholog Neighborhood Viewer.

The **Ortholog Neighborhood Viewer** displays the chromosome neighborhoods for the gene of interest and its orthologs in the genomes that have been selected using the **Genome Browser.** The **Ortholog Neighborhood Viewer** has been extended with the ability to reduce the number of displayed neighborhoods as illustrated in Figure 8(i), where the first two neighborhoods are selected and then redisplayed as illustrated in Figure 8(ii). The viewer has been also extended with the ability of selecting genes on the viewer, as illustrated in Figure 8(ii), where a gene is selected and then displayed as illustrated in Figure 8(iii). Genes selected on the viewer can then be added to the **Gene Cart**.

Function Profile –*Slim View*

The **Function Profile** tool available under **Analysis Carts /Functions** and **Compare Genomes** has been extended with a "slim view" option, as illustrated in Figure 9(i). This option allows to (i) display only column identifiers instead of genome or function name abbreviations, and (ii) remove the ***rows*** containing only zeros from the result of a function profile, as illustrated in Figure 9(ii).



Function ID	Name	Lac aci 476	Lac aci NCM	Lac bre 367	Lac bre 275	Lac buc 117	Lac cas 334	Lac cas BL3	Lac cri JV1	Lac del 112	Lac del BA5	Lac fer 141	Lac fer 396	Lac gas 333	Lac gas JV3	Lac gas MV2	Lac hel 451	Lac hil 820	Lac jen 11N	Lac jen JV6	Lac joh 330	Lac joh 533	Lac par 870	Lac par 252	Lac pla 147	Lac pla WC1	Lac reu 103	Lac reu CFA	Lac reu F25	Lac reu JC2	Lac reu MM3	Lac reu MM1
COG0304	3-oxoacyl-(acyl-carrier-protein) synthase	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	1	1
COG0331	(acyl-carrier-protein) S-malonyltransferase	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
COG0332	3-oxoacyl-[acyl-carrier-protein] synthase III	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
COG0439	Biotin carboxylase	0	0	0	0	0	0	0	0	0																						
COG0511	Biotin carboxyl carrier protein	0	0	0	0	0	0	0	0	0																						
COG0777	Acetyl-CoA carboxylase beta subunit	0	0	0	0	0	0	0	0	0																						
COG0825	Acetyl-CoA carboxylase alpha subunit	0	0	0	0	0	0	0	0	0																						
COG1024	Enoyl-CoA hydratase/carnithine racemase	0	0	0	0	0	0	0	0	0																						
COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)	1	0	0	11	10	0	0	0	0																						
COG1250	3-hydroxyacyl-CoA dehydrogenase	0	0	0	0	0	0	0	0	0																						

Function ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
COG0304	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1
COG0777	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
COG1028	1	0	0	11	10	0	0	0	0	0	2	0	0	0	0	0	7	0	0	0	0	2	2	1	0	0	2

Column ID	Column Name
1	Lactobacillus acidophilus ATCC 4796
2	Lactobacillus acidophilus NCFM
3	Lactobacillus brevis ATCC 367
4	Lactobacillus brevis gravesensis ATCC 27305
5	Lactobacillus buchneri ATCC 11577

FIGURE 9. Function Profile- Slim View.

Column names can be viewed by mousing over the column identifiers and are also provided in a list at the bottom of the function profile result page, as illustrated in Figure 9(ii).