MICROBESONLINE

Virtual Institute for Microbial Stress and Survival

Site Guide & Tutorial

MICROBESONLINE

Site Guide & Tutorial

Table of Contents

Chapter 1

Tutorial Overview

This tutorial will give you a basic overview of the MicrobesOnline website and tools. Certain icons and text styles are used throughout the document to help better organize information and to assist you with navigating the site. Text indicated in **bold** will represent the names of hyperlinks or buttons that you should follow and text indicated in *italics* will be used to describe the name of a particular section of the website or a specific tool—for example you may see, "…to access *xyz* resource please click on the **Login** button." Text shown in a `fixed-width` font indicates text you should type, such as a URL.

Occasionally, icons will be used to indicate that there is more information about a specific site element or tutorial topic, such as a limitation, or more advanced usage of a particular tool. You should look for the corresponding caption box for an explanation. The following is a list of all icons used in this document:

 **Feature limitation or feature unsupported**

 **More information on a particular topic**

 **Important URL to bookmark or save**

 **Contact information**

 **Additional information tangentially related to a topic**

 **Alert or caution**

This tutorial is a work in progress and therefore is subject to change. Please make sure you have the latest version, which can be downloaded from the MicrobesOnline website, `http://microbesonline.org`. If you have any questions, please feel free to contact us via email to `gtlweb@vimss.lbl.gov`.

Site Overview

The MicrobesOnline site is rather large and complex. Most tools have multiple entry points and almost everything is interconnected. The following diagram shows only the core MicrobesOnline site features and how they are interconnected.

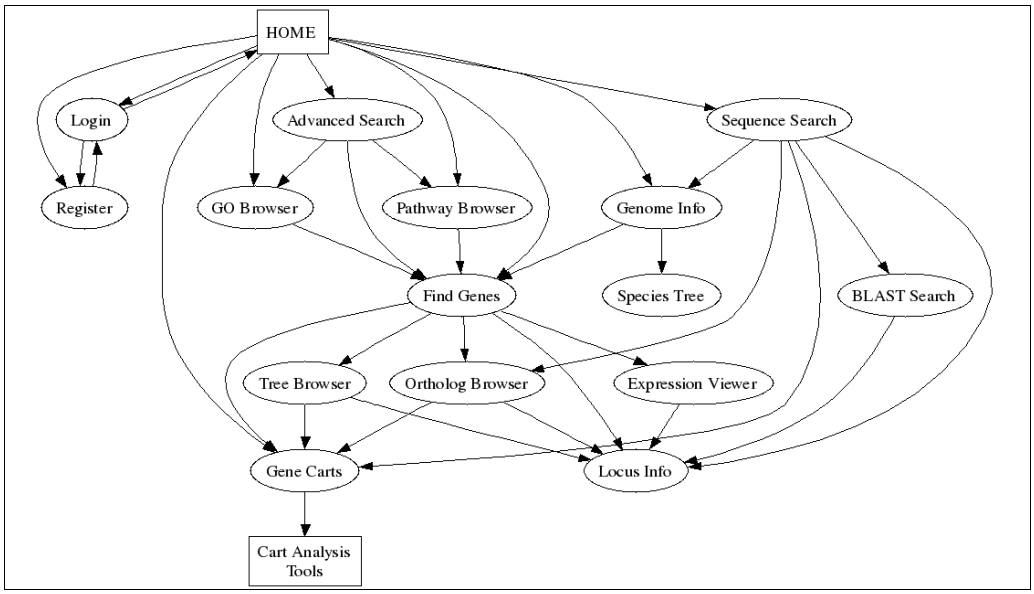


Figure 1. Overview of the MicrobesOnline site, showing only core components.

Some portions of the site and some functionality can only be accessed by registering for an account and logging in. You'll be able to create a *session gene cart* without registering but if you wish to save this cart or use any of the *cart analysis tools*, you will need to register for an account, or login if you've already registered. In addition, registering allows you to opt in to receive notifications from our staff concerning site downtime and updates to data and tools.

System Requirements

To access all of the features of the MicrobesOnline website, you must us a modern web browser that supports cookies and JavaScript. We recommend a minimum screen resolution of `1024x768` and one of the following browsers: Microsoft® Internet Explorer 6.0 or higher, or Mozilla Firefox 2.0 or higher. The formatting of the contents of our website is tested using Microsoft® Internet Explorer 7.0 and Mozilla Firefox 2.0. Use of other browsers may result in misalignment or unintended formatting of content.

Jalview is an external Java™-based multiple sequence alignment viewer that is used in MicrobesOnline. To use *Jalview* you must have a functional Java™ runtime environment (JRE). We recommend JRE 6 update 3, which is available for download at `http://www.java.com/en/download/manual.jsp`.

WARNING

Large images and tables generated by MicrobesOnline may take a long time to download and/or render on slower computers. On older computers, your browser may become unresponsive or crash. If this occurs, reduce your dataset size and try your query again.

Site Privacy Policy

The MicrobesOnline website uses cookies to maintain state about your session, such as whether you are logged in as a valid user. These cookies are restricted to the `microbesonline.org` domain. In order to better serve the needs of our users, we collect basic information sent to our web server by your web browser, such as your web browser type and version, your computer's operating system and version, and your screen resolution. This information is commonly referred to as the "user agent" and you should refer to your web browser's documentation for more information on what information is sent.

Certain portions of the MicrobesOnline site require registration. During the registration process we collect your name, an optional institution name, and your email address. This information is used internally only and is never released to outside third parties without your consent. Additionally, at your choosing, we may contact you at the supplied email address to notify you planned or unplanned site outages as well as data and site software updates. Your email address is never used for the purposes of distributing unsolicited commercial or bulk email.

If you have any questions or concerns, please contact us via email to `gtlweb@vimss.lbl.gov`.

Data Reliability Disclaimer

Much of the data presented on MicrobesOnline is determined using computational methods and therefore are subject to error. This section describes the general trustworthiness and caveats for various data found on the MicrobesOnline site.

DNA Sequences

DNA sequences for complete genomes are highly accurate. The local accuracy—the proportion of correctly identified bases—is usually >99.99%. Larger-scale assembly errors in which the genome is rearranged from its actual layout are possible, but for complete bacterial genomes, misassembly is very rare. Those assembly errors that do occur are mostly collapsed tandem repeats (e.g., the assembly shows one copy of the region rather than two). For incomplete genomes, however, missassembly may be much more common. You can see if a genome is complete or not from the *Genome Information* page.

Although the overall accuracy of genome sequences is very high, occasional base-calling errors or 1-nucleotide insertion/deletions (indels) do occur. In rare cases, this

will introduce a spurious frameshift into a protein-coding gene, so that it is not identified or is identified as a pseudogene. Sometimes the sequencing center will double-check these cases. Either the frameshift was spurious, and the gene will be corrected, or the region will be annotated as a *genuine frameshift* (that is, they double-checked and the frameshift truly is present).

Although most draft genomes are over 99% complete, it is difficult to know how complete the assembly is. Thus, no strong conclusions can be drawn from the absence of a gene in a draft genome. One test you can do yourself is to see how many tRNAs or ribosomal proteins are present.

Contamination is also an issue with draft genomes. Some draft "genomes" are even mixtures of two or more strains and when we identify these cases, we rename the genome to have an "spp." suffix.

Protein-coding Gene Predictions

Predicting protein-coding genes in bacterial genomes is over 95% accurate. However, the methods by which these genes are identified varies from genome to genome, and as many as a few percent of genes could be missing. If you suspect that a gene of interest to you might be present in some genomes but was not predicted, you can you use *Sequence Search* or *tblastn* to look for homology in the translated region of MicrobesOnline genomes.

A bigger problem with gene predictions is that many genuine bacterial genes do not have homologs in other organisms, or only have nearly-identical homologs in other strains. There is no way to know if any particular "ORFan" genes are genuine, especially if the gene is short.

Although gene predictors do a good job of determining whether or not a protein-coding gene is present in a given region of the genome, and in choosing the correct reading frame, the start codon predictions are probably much less accurate (perhaps 80-90% accurate).

Finally, there are a few ambiguous cases involving frameshifts. If a frameshift is present in what otherwise appears to be a protein-coding gene, the gene is usually classified as pseudogene and sometimes it is simply ignored. It is hard to know if the remaining fragment of the gene could be functional or if there is a programmed frameshift so that the protein is expressed despite the frameshift.

Non-coding RNA Gene Predictions

Almost all genomes include predictions for ribosomal RNAs and for tRNAs. Coverage of other non-coding RNAs is spotty.

Protein Homology

MicrobesOnline includes a variety of pre-computed gene families and gene homology information. These methods are quite trustworthy—genes that have the same domain assignment or are listed in the homologs page are virtually certain to be homologous.

The only exception is hits to the *superfamily* database—this database uses a weak statistical cutoff and thus the weaker hits (E>0.001) may not be reliable.

Although the gene homology is almost certainly correct, the functional implication of the homology is uncertain. General functions (e.g., transcription factor) are generally conserved, but the specific role (e.g., O_2 sensor) often is not.

Gene Names & Descriptions

The gene names and gene descriptions in MicrobesOnline are almost entirely derived from *RefSeq*, and most of those are in turn derived from the original genome project. For most genes, these names and descriptions are entirely based on homology, and they are often incorrect and/or misleading. Also, the quality of the annotations varies widely by organism—annotations in model organisms are usually better, however even annotations for genes in E. coli K12 can be misleading.

Here are some guidelines we recommend for checking if an annotation is reasonable.

Is it consistent with its domain structure & COG assignment? If you're not sure how to interpret the domain structure, find a characterized homolog, perhaps a distant one from *E. coli* or *B. subtilis*, and verify that the domain structures are the same.

Do close homologs in the *Tree Browser* have consistent annotations?

Are there characterized close homologs? Characterized genes are highlighted in the *Tree Browser* with green underlines and are highlighted on the homologs page with green **see papers** links. If yes, check the paper abstracts. If you don't find links to papers at MicrobesOnline, but the gene itself or its close homologs have specific gene names (e.g., rtsA), then information may be available but not linked to the gene sequences. We recommend that you search *PubMed* and the relevant model organism databases such as *EcoCyc* or *Subtilist*.

While using the *Tree Browser*, you may see conserved gene neighbors (e.g., across diverse bacteria, the same COG is adjacent to your gene of interest). Conserved gene neighbors, if present, provide strong evidence that the two genes have related functions. On the other hand, many genes do not have conserved neighbors, and in that case, no conclusions can be drawn.

EC Assignments & Metabolic Maps

EC assignments are obtained from *KEGG* and are almost entirely based on homology. Thus, many of them are incorrect. Non-specific EC assignments (e.g., 2.1.3.-) are particularly unreliable.

If you have an EC assignment of interest and want to check it, try the processes for checking a gene's annotation described above.

On the other hand, if you suspect an enzyme is present but no gene has been assigned its EC number, we suggest looking for known families that carry out the reaction. You can find such genes in *MetaCyc*. You can also look for gene-to-gene relationships with

other steps in the pathway. You can find gene neighbors in the *Tree Browser* or on the *regulon* page located in the *Locus Information* tool. In some organisms you can also look for co-expressed genes using the expression *profile search tool*.

Orthologs

MicrobesOnline identifies orthologous genes by analyzing phylogenetic trees and identifying subgroups that are mostly present only once per genome. Ideally, these "tree orthologs" are evolutionary orthologs—genes that diverged from each other when the two genomes diverged from each other—or xenologs – genes that were horizontally transferred between ancestors of these genomes, without any more recent duplications. These relationships usually, but not always, indicate conserved function. You can check the accuracy of the orthologs by using the *Tree Browser*. In particular, you can select genomes that you suspect might have an ortholog of your gene of interest and then click **update** to highlight genes from those genomes in the tree. The *Tree Browser* can also compare the gene's tree to the species tree, which can highlight horizontal transfer events.

Instead of "tree orthologs," a few features of the web site rely on MicrobesOnline Ortholog Groups (MOGs). These are similar to the tree orthologs and are faster to compute, but are less accurate. They are mostly used for automated analyses such as operon prediction.

Operon Predictions

The operon predictions in MicrobesOnline are around 85% correct and are based on comparative genomics analyses. They are also based on gene expression data if that is available. You can also use the expression data to see if the putative operon's expression pattern is consistent. Although gene expression data is often quite noisy, genes in the same operon usually have much more correlated expression than adjacent genes that are not in the same operon.

The operon predictions are designed to predict if two adjacent genes can sometimes be expressed from the same promoter. Bacterial operons often have internal promoters or partial terminators. In these cases, you will see some correlation in the genes' expression, but the expression patterns may be quite different.

Regulon Predictions

The regulon predictions are based on conserved gene neighbors. The predictions are not reliable but can be a useful pointer for functionally related genes or, by looking at the annotations of the linked-to genes, for the functional role of the gene itself.

Gene Expression Data (Microarrays)

Microarray data is of widely varying quality, depending on the platform, the number of replicates, and the skill of the experimenter. MicrobesOnline includes a z-score that gives a very rough estimate of the reliability of each measurement—values above 2 or below -2 indicate higher confidence. The z-score only takes into account the number of replicates, not the quality of the platform therefore if the probe on the chip is biased

and gives consistently incorrect results, MicrobesOnline will misleadingly show a strong z-value.

To help you assess the reliability of each microarray comparison, each comparison has a plots page. The plots include a volcano plot of z-value versus log-fold-change and an "agreement with operons" plot. The agreement plot is based on the assumption that genes in the same operon should have the same fold-change. A good experiment will have agreement above 50% for the most extreme groups of up- and down-changers.

An individual microarray measurement is more trustworthy if it is consistent with other genes in the operon, if any. Because each gene has its own probe(s), any systematic biases should affect the different genes in the operon differently. You can assess this using the operon heatmap view. If an experiment has multiple time points, then the fold-change of a gene should be similar for nearby time points.

Home Page

The MicrobesOnline home page can be divided into 5 major sections—the top navigation elements, the genome selector, the quick sequence search tool, the "about us" section, and the site highlights section. The first two sections, the top navigation elements and genome selector, are present on many of the MicrobesOnline pages.

Each section is described in greater detail on the following pages.

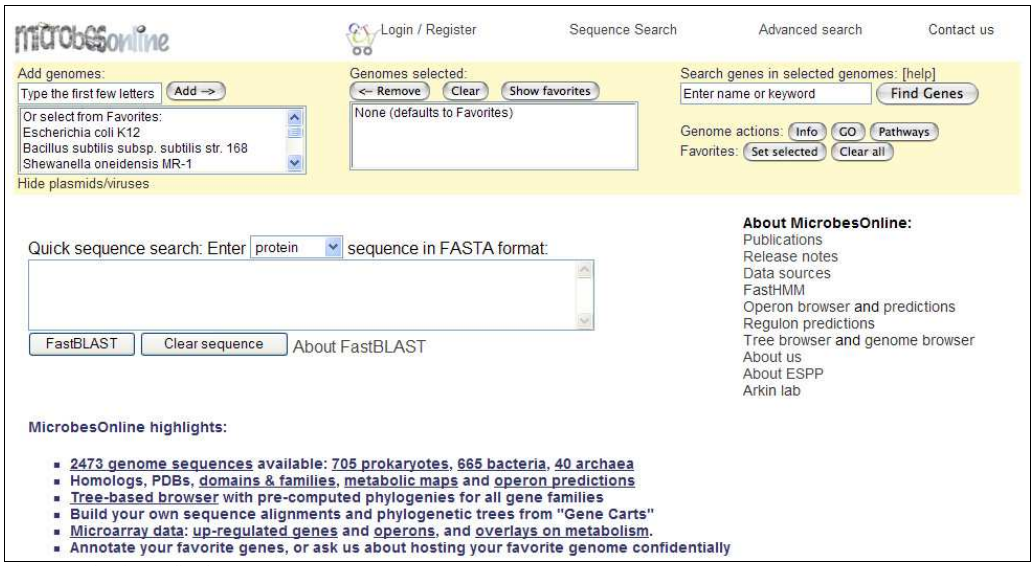


Figure 2. MicrobesOnline homepage

Top Navigation

The top navigation elements include a clickable MicrobesOnline logo and a series of four or five links, depending on whether you are currently logged in with your

MicrobesOnline account. If you are not logged into your account, you should see the following five links: **Login**, **Register**, **Sequence Search**, **Advanced search**, and **Contact Us**, as shown below.

Figure 3. Top navigation view when not logged in

If you are already logged into your MicrobesOnline account, you will see four links instead: **My Gene Carts**, **Sequence Search**, **Advanced search**, and **Contact Us**.

Figure 4. Top navigation view when logged in

QUICK TIP

Clicking the MicrobesOnline logo will always return you to the home page.

Login

The **Login** link in the top navigation allows you to sign in using your MicrobesOnline account. This is required to access certain portions of the MicrobesOnline site, including *gene cart analysis tools*.

Figure 5. Login screen

To login, you must supply the email address and password you used to register your account then click on the **Login** button. If you do not remember your password, click on the **forgotten** password link and enter your email address. We will reset your password and email it to you.

Finally, if you have not yet registered for an account you may do so by clicking on the **register here** link.

WARNING

The MicrobesOnline site requires that you enable cookies in your web browser. If you login but are unable to access login-only portions of the site, you may have cookies disabled. Please consult with your system administrator for assistance.

Register

If you wish to register for an account, please click on the **Register** link located in the top navigation of most pages. You will be asked to supply some basic information about yourself and to choose a password. You must supply a valid email address or MicrobesOnline will not be able to email you to reset your password if you've forgotten it. In addition, you may select your email notification preferences if you wish to be contacted by MicrobesOnline staff concerning site downtime, data updates, or changes to the site software.

> ⚠ **WARNING**
>
> User names and passwords are transmitted to the MicrobesOnline server without encryption. You should avoid using a password that is similar or identical to one you use on another site.

My Gene Carts

This link will open a pop-up window to the *Bioinformatics Workbench Cart Summary*, where you can see a list of all your saved carts and access *gene cart analysis tools*. You can find more information about gene carts in the *Gene Carts* section found later in this document.

Sequence Search

This link will redirect you to the *Sequence Search* tool. Please see this section for more information.

Advanced Search

This link will redirect you to the *Advanced Search* tool. Please see this section for more information.

Contact Us

This is an email link to `gtlweb@vimss.lbl.gov`. Feel free to contact us for any reason. Most inquiries are resolved within 24-48 hours however custom data requests or requests to host genomes may take longer. We appreciate your patience and thank you in advance for your feedback.

Genome Selector

The genome selector component is also present on many pages throughout MicrobesOnline. It allows you to quickly select a subset of genomes for obtaining *genome information*, limiting gene searches with the *find genes* tool, browsing *Gene Ontology*, and comparing the presence and absence of genes in the *Pathway Browser*. You can also use it to configure a set of favorite genomes, which become your default set each time you use MicrobesOnline.

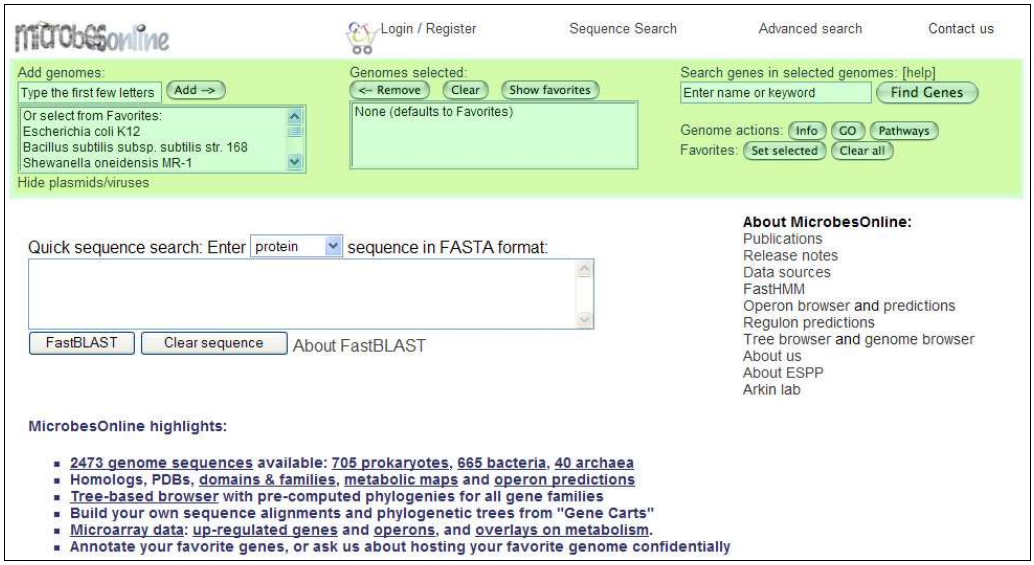


Figure 6. Home page showing the genome selector highlighted in green.

All of the actions in the genome selector apply only to the subset of genomes selected. If you are new to MicrobesOnline or if you do not have your own set of favorites configured, this subset defaults to the MicrobesOnline set of favorite genomes as indicated by the text "None (defaults to Favorites)", which include among others *E. coli K12, B. subtilis str 168,* and *S. oneidensis MR-1*.

Configuring Your Own Set of Favorite Genomes

To configure your own set of favorite genomes, select one or more genomes from the left pane and click on the **Add** button. Only the MicrobesOnline favorite genomes are displayed in the left pane by default. If the genome you wish to add is not in this list, enter the first few letters of its genus in the **Add genomes** text box and search through the resultant matches. You may refine your search by specifying the complete genus and part or all of the desired genome's species name. Beneath the left pane, you will find a toggle link that allows you to choose whether you would like virus and plasmid names to be searched as well. If the link reads **Show plasmids/viruses (slow)** then viruses and plasmids will **not** be searched. However, if the link reads **Hide plasmids/viruses**, then virus and plasmid names are searched. You should disable searching of plasmid and virus names to improve performance of the genome selector.

> **QUICK TIP**
>
> To select multiple genomes from the left pane to be added to your favorites you must hold down the `Ctrl` key on a PC or the `Shift` key on a Mac and use your mouse to left-click on the genomes you wish to select.

Once you are done selecting genomes you must save your list by clicking on the **Set selected** button, located on the right side of the genome selector display. Now you should see your favorite genomes show up in the right pane. You may modify this set at any time by using the **Remove** and/or **Clear** buttons in conjunction with the **Add** button.

> **NOT SUPPORTED**
>
> You may only select individual genomes to be added to your favorite genome set. Genomes groups can be differentiated from individual genomes because they contain a string showing how many genomes are contained within the group, for example `Escherichia (19 genomes)`.

To undo changes you've made to your set of favorite genomes, simply click on the **Show favorites** button above the right pane. To reset your favorite genomes to the default set, click the **Clear** button above the right pane then click on **Set selected**.

⚠ WARNING

Favorite genomes are currently saved to your browser using a cookie and are not saved to your MicrobesOnline profile. Therefore, the favorite genomes you set are only available in a subsequent session from the same browser on the same computer.

Finding Genes

There are a number of ways to find a gene in MicrobesOnline. The *Find Genes* section of the genome selector is one such way. It allows you to locate genes by using keywords, such as all or part of the gene's name, the gene's symbol, or other attributes as described below. A wildcard character of _ will match any single character and a wildcard character of % will match zero or more of any character. For example, the query `translation%factor` will match *translation factor*, *translation elongation factor*, *translation initiation factor*, etc.

The following table contains special keywords and a description of the type of search that will be performed for each special keyword.

Type	Format	Example	Description
EC numbers	`ECX.X.X.X`	`EC1.-.-.-`	Find all genes assigned the specified EC number
Gene Ontology term accession	`GO:XXXXXXX`	`GO:0007275`	Find all genes assigned the specified GO accession
VIMSS id	`VIMSS<id>`, or `<id>`	`VIMSS14482` or `14482`	Find the gene with the specified VIMSS id
Genbank GI number	`gi<id>`	`gi46578818`	Find the gene with corresponding Genbank GI number
SwissProt or UniProt id or accession		`BGAL_ECOLI` or `P00722`	Find the gene with corresponding SwissProt or UniProt id
InterPro domain accession	`IPRxxxxxx`	`IPR006067`	Find all genes with the specified domain conserved
COG cluster id	`COG<id>`	`COG2221`	Find all genes assigned to the specified COG cluster
Pfam accession	`PFxxxxx`	`PF01077`	Find all genes with hits to the specified Pfam
TIGRfam accession	`TIGRxxxxx`	`TIGR02064`	Find all genes with hits to the

			specified TIGRfam
Superfamily accession	SSFxxxxx[x]	SSF56014	Find all genes with hits to the specified Superfamily
Panther accession	PTHRxxxxx	PTHR15184	Find all genes with hits to the specified Panther
PIR accession	PIRSFxxxxxx	PIRSF010340	Find all genes with hits to the specified PIR
SMART accession	SMxxxxx	SM00382	Find all genes with hits to the specified SMART

Table 1. Keyword search special keywords

> ⚠ **WARNING**
>
> Keyword searches only apply to the selected genomes listed in the right pane of the genome selector. Overly broad searches may take a long to complete.

The results page shows each gene identified by the keyword search, one per row, including the VIMSS id of the gene (e.g., VIMSS422957), its symbol or name, its location on the genome including strand, the name of the genome to which the gene belongs, and optional annotation information.

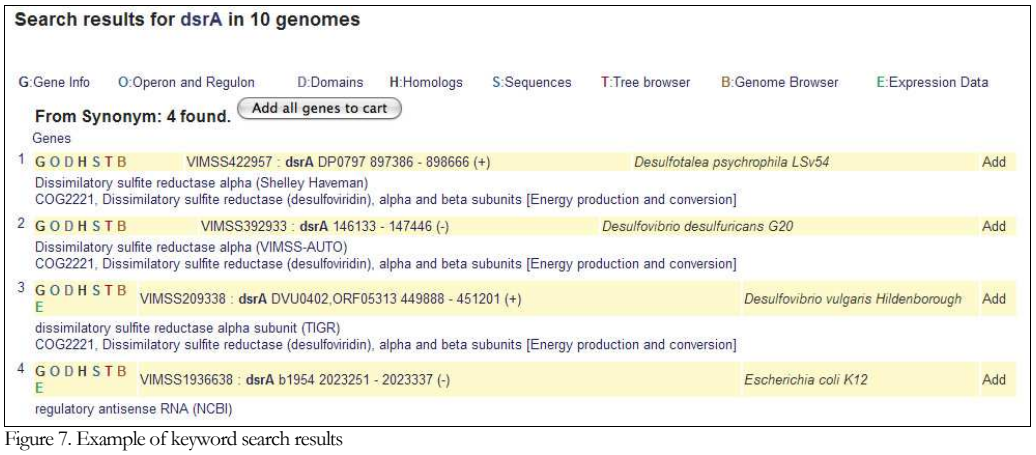


Figure 7. Example of keyword search results

You can add all identified genes to your *session gene cart* by click on the **Add all genes to cart** button, or individually add genes of interest using the **Add** link on the right, present for each gene displayed in the results. Please see the section on *Gene Carts* for more information.

Clicking on the genome name of a particular gene will load basic information about the genome, including the size and GC% content of each scaffold, and a distribution of COG assignments.

The links on the left side of each result allow you to retrieve more information about the gene. Each letter corresponds to a different type of information or additional tool that is available for each result. Expression data (E) are only available for certain genes in certain genomes, including *E. coli K12* and *B. subtilis*. For more information on the G O D H S links, see the section describing *Locus Information*. Additional information about the *Tree Browser* (T), *Ortholog Browser* (B), or *Expression Data Viewer* (E) is available in their respective sections of this tutorial.

Genome Actions: Info

Clicking on the **Info** button will retrieve summary information for the selected genomes or the MicrobesOnline set of favorite genomes if no genomes have been selected. For more information please see the *Genome Information* section.

Genome Actions: GO

Clicking on the **GO** button will launch the *GO Browser*, allowing you to browse the Gene Ontology hierarchy in the context of the selected genomes. For more information please see the *GO Browser* section.

Genome Actions: Pathways

Clicking on the **Pathways** button will launch the *Pathway Browser*, allowing you to browse the KEGG pathway hierarchy in the context of up to two selected genomes. For more information please see the *Pathway Browser* section.

Quick Sequence Search

This tool allows you to search for genes using sequence similarity using our proprietary FastBLAST heuristic, which is often orders of magnitude faster than regular BLAST searches. The *quick sequence search* tool is identical to the *Sequence Search* tool, with its interface included on the home page for more convenient access. For more information about the FastBLAST heuristic, click on the **About FastBLAST** link. Otherwise, please see the *Sequence Search* tool for more information.

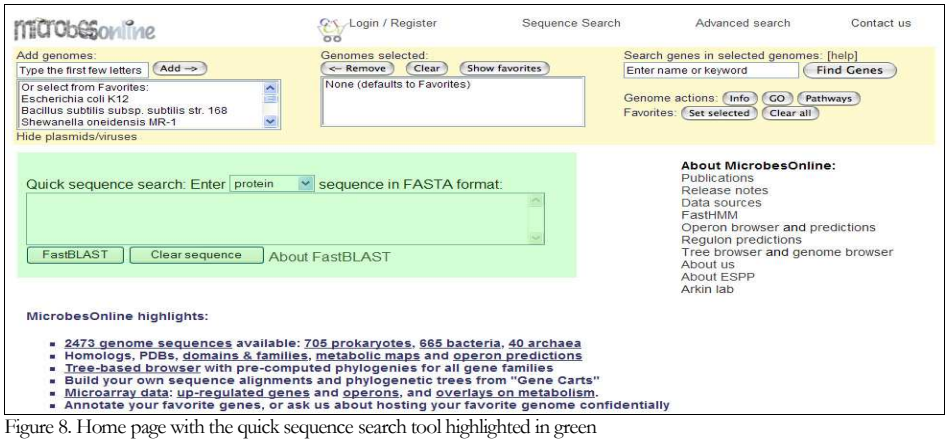


Figure 8. Home page with the quick sequence search tool highlighted in green

About MicrobesOnline

This section contains a list of links with information our users may find useful. Generally, this includes a list of publications, release notes, data sources, and information about analysis methods developed for MicrobesOnline.

MicrobesOnline Highlights

This section contains highlights from the most recent data and software release.

Sequence Search

The *Sequence Search* tool provides an interface for finding genes and other sequences using sequence similarity. The *Sequence Search* interface is present on the MicrobesOnline home page for convenience and is also located as a separate tool from the home page and other tools on the site. Both interfaces operate identically. *Figure 8* from the previous page highlights the *Sequence Search* interface found on the home page and *Figure 9*, below, shows the *Sequence Search* standalone tool.

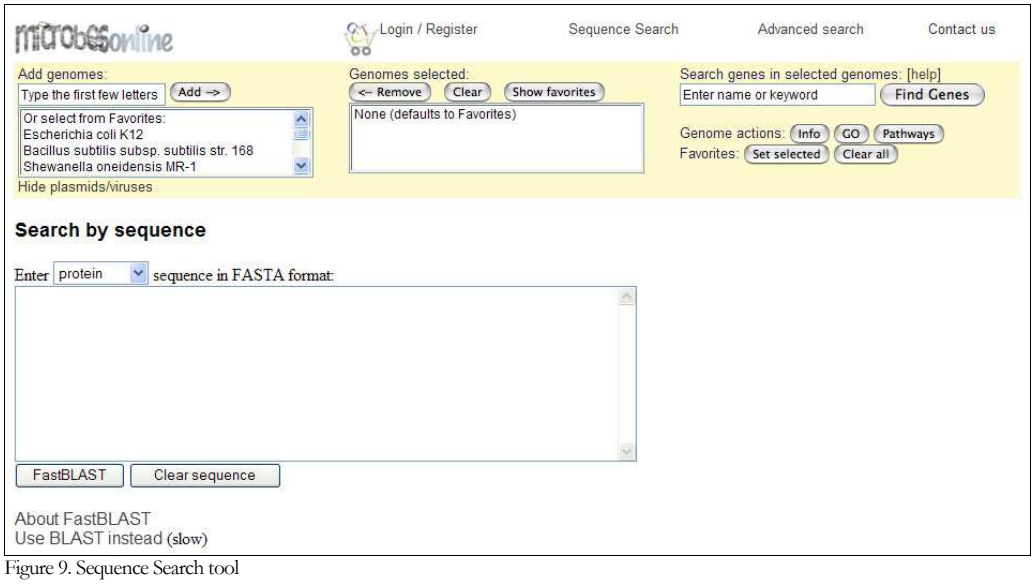


Figure 9. Sequence Search tool

A drop-down selection box allows you to select the query sequence type, either **protein** or **nucleotide**, and defaults to **protein**. Select the appropriate type then input your query sequence in FASTA format. A FASTA-formatted sequence contains a header line, called a "defline", and is followed by one or more lines of sequence as shown in the following example.

```
>mysequence_name
MAKHETPLLDQLESGPWPSFVTDIKRQAEKKPECWDILGILELSFKERITHWKHGGIVGV
FGYGGGIVGRYADVPERFPGVEHFHTIRVAQPSSKYYTSKNLRQLMNLWEKHGSGMTNFH
GSTGDVILLGTRTENLEPFFWDLTHEMGQDLGGSGSNLRTPACCLGTSRCEWACYDTQEA
CHSLTMHYQDEIHRPAFPYKFKFKFSGCPNDCVAAIARSDVAVIGTWKDDIRIDQAAVKG
YMANEFPANGGAFIGREWDAFDIQKEVIDLCPTNCMWMEDGELKIDDAECTRCMHCINVM
PRALRPGAQGGASICVGAKAPILDGAQFATLILPFIPVTKDNDFEELIEFIESIWDWWME
IGKNRERVGETMQRVGLPTFLRAVGVEALPQHVKYPRENPYVFWNEEEVEGGFERDVQEF
RARHAA
```

NOT SUPPORTED

You may only specify one query sequence as input to the *Sequence Search* tool at this time.

Once you've inputted your query sequence simply click on the **FastBLAST** button to begin the search. Please see the following section on how to interpret the results. Also, we have provided a link to *BLAST* available from the indicated *Sequence Search* page only, which will provide a more refined, though often slower search. For more information on the *BLAST Sequence Search* tool, please refer to its section below.

Protein Sequence Search Results

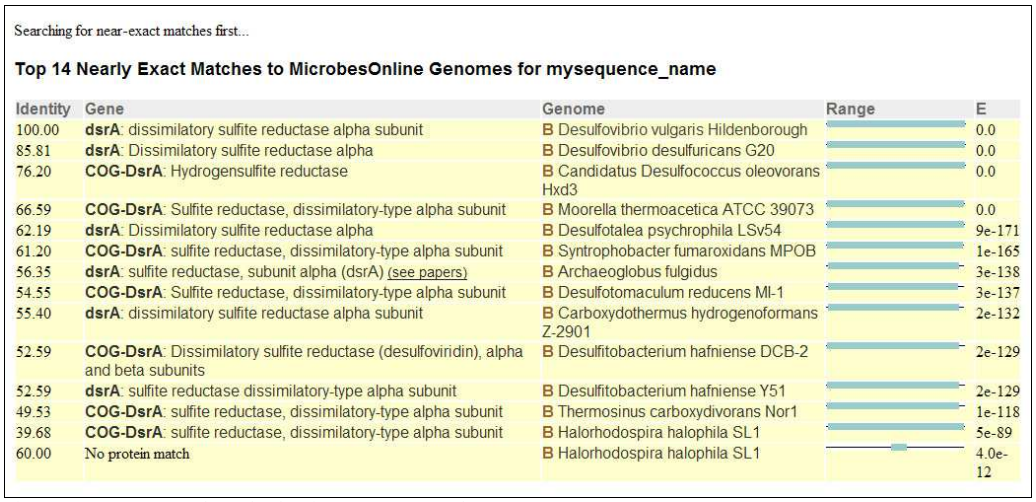
Sequence Search actually performs three types of searches in series. First, it looks for very highly conserved (near-exact) matches in all MicrobesOnline genomes using BLAT (📌 BLAST-Like Alignment Tool). Next, it will use our in-house 📌 FastHMM tool to identify conserved domains within the input query sequence. Finally, it will use our in-house 📌 FastBLAST tool to look for other potentially-distant hits.

REFERENCES

- BLAT (BLAST-Like Alignment Tool)
  `http://genome.ucsc.edu/FAQ/FAQblat.html`
- MicrobesOnline FastHMM
  `http://microbesonline.org/fasthmm`
- MicrobesOnline FastBLAST
  `http://microbesonline.org/about_fastblast.html`

Near-Exact Matches

You will see the text "*Searching for near-exact matches first…*" while the *Sequence Search* tool identifies near-exact matches. Once completed, the results will follow in tabular form, as shown in the figure below.



Searching for near-exact matches first...

Top 14 Nearly Exact Matches to MicrobesOnline Genomes for mysequence_name

Identity	Gene	Genome	Range	E
100.00	**dsrA**: dissimilatory sulfite reductase alpha subunit	B Desulfovibrio vulgaris Hildenborough		0.0
85.81	**dsrA**: Dissimilatory sulfite reductase alpha	B Desulfovibrio desulfuricans G20		0.0
76.20	**COG-DsrA**: Hydrogensulfite reductase	B Candidatus Desulfococcus oleovorans Hxd3		0.0
66.59	**COG-DsrA**: Sulfite reductase, dissimilatory-type alpha subunit	B Moorella thermoacetica ATCC 39073		0.0
62.19	**dsrA**: Dissimilatory sulfite reductase alpha	B Desulfotalea psychrophila LSv54		9e-171
61.20	**COG-DsrA**: sulfite reductase, dissimilatory-type alpha subunit	B Syntrophobacter fumaroxidans MPOB		1e-165
56.35	**dsrA**: sulfite reductase, subunit alpha (dsrA) (see papers)	B Archaeoglobus fulgidus		3e-138
54.55	**COG-DsrA**: Sulfite reductase, dissimilatory-type alpha subunit	B Desulfotomaculum reducens MI-1		3e-137
55.40	**dsrA**: dissimilatory sulfite reductase alpha subunit	B Carboxydothermus hydrogenoformans Z-2901		2e-132
52.59	**COG-DsrA**: Dissimilatory sulfite reductase (desulfoviridin), alpha and beta subunits	B Desulfitobacterium hafniense DCB-2		2e-129
52.59	**dsrA**: sulfite reductase dissimilatory-type alpha subunit	B Desulfitobacterium hafniense Y51		2e-129
49.53	**COG-DsrA**: sulfite reductase, dissimilatory-type alpha subunit	B Thermosinus carboxydivorans Nor1		1e-118
39.68	**COG-DsrA**: sulfite reductase, dissimilatory-type alpha subunit	B Halorhodospira halophila SL1		5e-89
60.00	No protein match	B Halorhodospira halophila SL1		4.0e-12

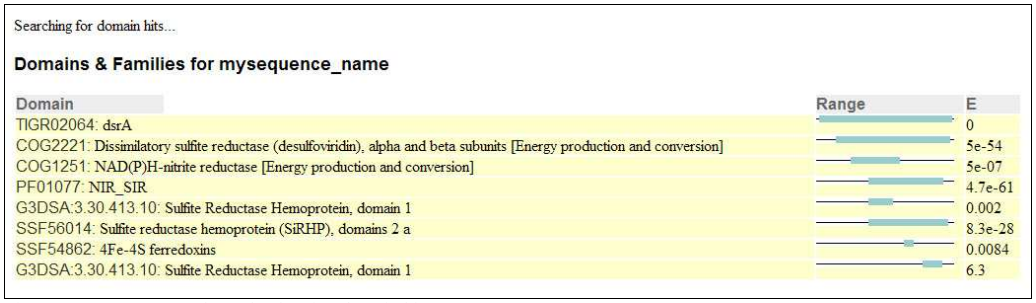
Figure 10. Sequence Search results showing near-exact matches.

The table shows the results of the BLAT search. The table headings, i.e. **Identity**, **Gene**, **Genome**, **Range**, and **E** are clickable to sort the data by the selected column. Sorting is initially in ascending order however clicking on the same column twice will change the sorting order to descending order. By default all matches to regions flanked by known genes are shown ordered by their %-identity in descending order followed by all other matches also sorted by %-identity in descending order. Other matches are typically labeled *"No protein match"* because they are hits to the six-frame translation of the genome where no gene has previously been identified.

For all hits to regions flanked by known genes, you can click on the gene symbol, name, or identifier that is displayed to retrieve more information for that gene in the *Locus Information* tool. The brown **B** link allows you to view the hit in the context of the gene neighborhood surrounding the candidate homolog. Clicking on the genome name will display more information on that genome in the *Genome Information* tool. The range shows a graphical summary of the hit span with respect to the input query sequence and the E value is a measure of confidence of the hit and represents the expectation that this hit would occur by chance. A lower E value means that a hit is potentially more significant.

Domain Hits

You will see the text *"Searching for domain hits…"* while FastHMM identifies conserved domain hits. The identified hits will be shown in tabular form, as shown in the following figure. The table headings can also be used to sort the results, as described previously.



Searching for domain hits...

Domains & Families for mysequence_name

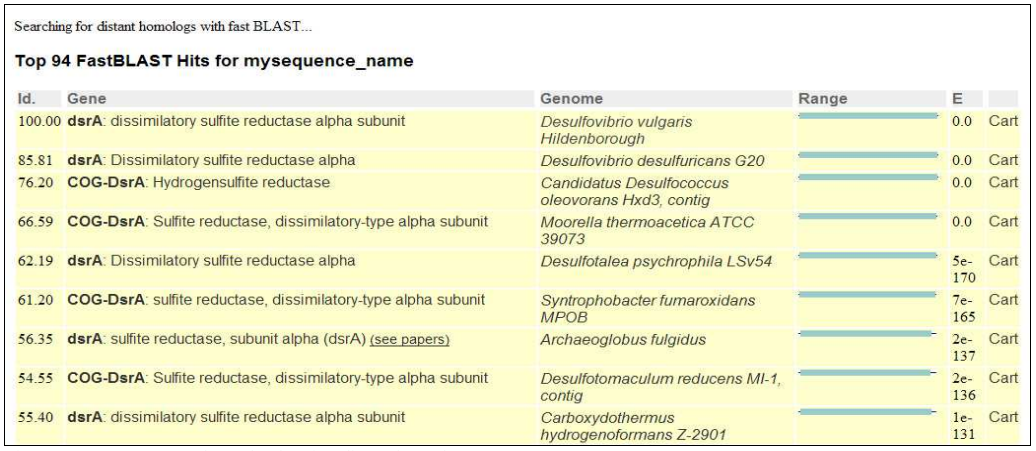
Domain	Range	E
TIGR02064: dsrA		0
COG2221: Dissimilatory sulfite reductase (desulfoviridin), alpha and beta subunits [Energy production and conversion]		5e-54
COG1251: NAD(P)H-nitrite reductase [Energy production and conversion]		5e-07
PF01077: NIR_SIR		4.7e-61
G3DSA:3.30.413.10: Sulfite Reductase Hemoprotein, domain 1		0.002
SSF56014: Sulfite reductase hemoprotein (SiRHP), domains 2 a		8.3e-28
SSF54862: 4Fe-4S ferredoxins		0.0084
G3DSA:3.30.413.10: Sulfite Reductase Hemoprotein, domain 1		6.3

Figure 11. Sequence Search results showing domain hits.

When available, the domain name will provide a hyperlink to an external data source for viewing more information about the domain. In the above example, clicking on the **TIGR02064** name will load the TIGR Comprehensive Microbial Resource (CMR) details of the TIGR02064 domain. The range and E value show the span of the hit with respect to the input query sequence and the confidence of the hit.

Distant Homologs

FastBLAST uses the domain hits generated from the previous search to significantly reduce the search space when identifying distant homologs. It is able to do this by masking out regions of the input query sequence covered by an identified domain hit. Domain hits are then used to discover similarity to other genes and finally any unmasked region of the input query sequence is used as a query in a global BLAST search of all genes. This search is generally much faster than a global BLAST search with minimal reduction of sensitivity, however if an expected hit isn't identified we recommend trying the slower global *BLAST Sequence Search*, described below.



Searching for distant homologs with fast BLAST...

Top 94 FastBLAST Hits for mysequence_name

Id.	Gene	Genome	Range	E	
100.00	**dsrA**: dissimilatory sulfite reductase alpha subunit	*Desulfovibrio vulgaris Hildenborough*		0.0	Cart
85.81	**dsrA**: Dissimilatory sulfite reductase alpha	*Desulfovibrio desulfuricans G20*		0.0	Cart
76.20	**COG-DsrA**: Hydrogensulfite reductase	*Candidatus Desulfococcus oleovorans Hxd3, contig*		0.0	Cart
66.59	**COG-DsrA**: Sulfite reductase, dissimilatory-type alpha subunit	*Moorella thermoacetica ATCC 39073*		0.0	Cart
62.19	**dsrA**: Dissimilatory sulfite reductase alpha	*Desulfotalea psychrophila LSv54*		5e-170	Cart
61.20	**COG-DsrA**: sulfite reductase, dissimilatory-type alpha subunit	*Syntrophobacter fumaroxidans MPOB*		7e-165	Cart
56.35	**dsrA**: sulfite reductase, subunit alpha (dsrA) (see papers)	*Archaeoglobus fulgidus*		2e-137	Cart
54.55	**COG-DsrA**: Sulfite reductase, dissimilatory-type alpha subunit	*Desulfotomaculum reducens MI-1, contig*		2e-136	Cart
55.40	**dsrA**: dissimilatory sulfite reductase alpha subunit	*Carboxydothermus hydrogenoformans Z-2901*		1e-131	Cart

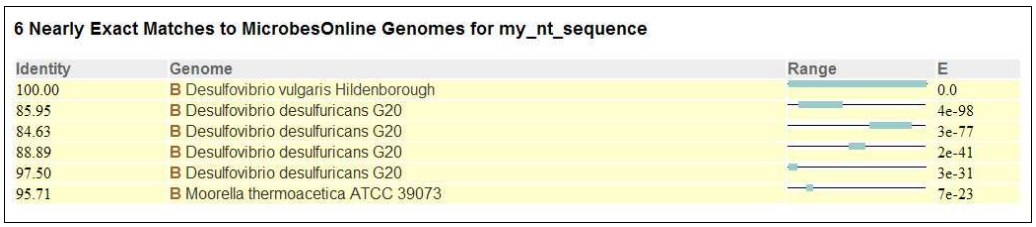
Figure 12. Sequence Search results showing distant homologs.

Identifying distant homologs using FastBLAST is often the most time-consuming of the three searches. You will see the message *"Searching for distant homologs with fast BLAST…"* while we perform this search and the results will be displayed in tabular form similar to that used in displaying near-exact matches however the hits displayed in this table are to actual genes as opposed to regions on genomes flanked by known genes therefore it is possible to add the representative gene to your *session gene cart* using the **Cart** link located to the right of each hit row.

Other than the addition of the **Cart** link, the remainder of the output remains consistent with the near-exact matches result. Clicking on a gene symbol or name will load more information about the gene from the *Locus Information* page and clicking on the genome name will load more information from the *Genome Information* page. The range is a graphical representation of the span of the hit with respect to the input query sequence and the E-value is the computed expectation that a particular hit would occur by chance. The E-values are based on a database size of 100 million letters rather than the actual size.

Nucleotide Sequence Search Results

When searching for sequences using a nucleotide input sequence, the *Sequence Search* tool only performs a near-exact genome search using MEGABLAST (`http://www.ncbi.nlm.nih.gov/blast/megablast.shtml`). The results are displayed in a table that is similar to the one used in the near-exact search for protein sequences, as shown below.



6 Nearly Exact Matches to MicrobesOnline Genomes for my_nt_sequence

Identity	Genome	Range	E
100.00	B Desulfovibrio vulgaris Hildenborough		0.0
85.95	B Desulfovibrio desulfuricans G20		4e-98
84.63	B Desulfovibrio desulfuricans G20		3e-77
88.89	B Desulfovibrio desulfuricans G20		2e-41
97.50	B Desulfovibrio desulfuricans G20		3e-31
95.71	B Moorella thermoacetica ATCC 39073		7e-23

Figure 13. Nucleotide Sequence Search results showing near-exact hits.

The results are initially sorted by %-identity in descending order. Like most results tables, you can change the sorting of the result rows by clicking on the table headings. Clicking the same heading multiple times will toggle between the default ascending sort and descending sort for a particular column.

Clicking the brown **B** link will redirect you to the *Ortholog Browser* where you can view the nucleotide hit in the context of candidate genes, if any. Clicking on the genome name will redirect you to the *Genome Information* page where you can find more information about the genome. The range and E value are identical to those used in the *Protein Sequence Search Results* section.

BLAST Sequence Search

We also offer a *BLAST Sequence Search* tool if you wish to perform your searches using BLAST only. While this is generally slower than the methods we employ in our *Sequence Search* tool, you may achieve slightly better sensitivity using BLAST. To access the *BLAST Sequence Search* tool, you must first access the *Sequence Search* tool. This is generally done by clicking on the **Sequence Search** link in the top navigation from most of the pages in MicrobesOnline. If you are unable to find this link, return to the MicrobesOnline home page first. On the *Sequence Search* tool interface, we provide a link to the *BLAST Sequence Search* tool by clicking on the **Use BLAST instead** link.

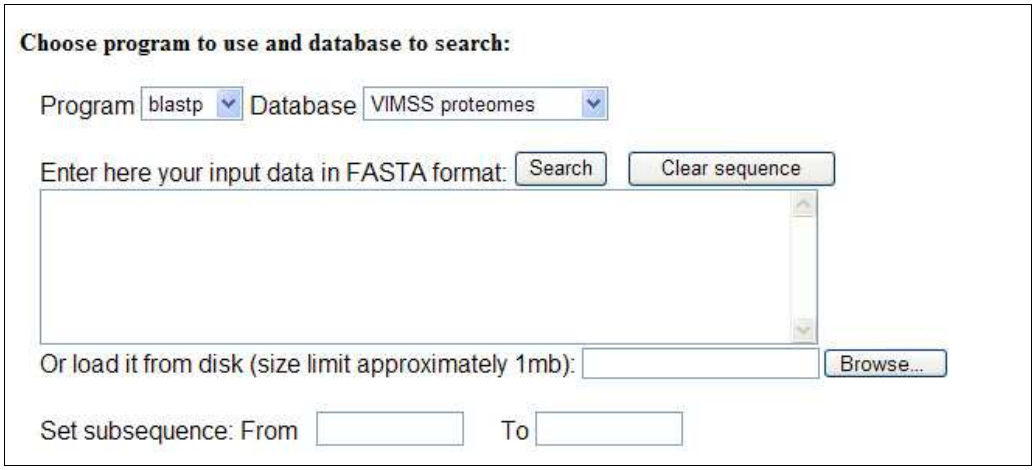


Figure 14. BLAST Sequence Search tool interface

The *BLAST Sequence Search* tool is generally meant for more advanced users as it provides less guidance and requires more input than our *Sequence Search* tool. You must be familiar with the different BLAST program types, such as **blastp** or **blastn**. Similarly, you must select a compatible database from the list of available databases in order to perform your search. The following table shows which of the VIMSS databases are compatible with the various BLAST program types.

Query Sequence	BLAST Program	VIMSS Databases
Nucleotide	blastn	VIMSS genomes, VIMSS transcriptomes
Protein	blastp	VIMSS proteomes
Nucleotide	blastx	VIMSS proteomes
Protein	tblastn	VIMSS genomes, VIMSS transcriptomes
Nucleotide	tblastx	VIMSS genomes, VIMSS transcriptomes

Table 2. BLAST Sequence Search valid input query sequence, BLAST program, and VIMSS database configurations.

At a minimum, you should select the BLAST **Program** and **Database**. To input your query sequences, simply paste them in FASTA format in the provided text box or select the file containing your sequences from your local computer by clicking the **Browse…** button. You may limit the search to a particular region of your input sequence by setting the **From** and **To** coordinates to the region you wish to search. Coordinates are 1-based, that is the first letter in your query sequence begins at position 1 and the range is inclusive. For example specifying **From 10** and **To 100** will result in a 91 residue subsequence beginning with the 10th residue of your input sequence.

> **WARNING**
>
> Specifying a subsequence range limits you to a single input query sequence. If you specify multiple sequences, only the first sequence will be used!

The **Clear sequence** button will clear the sequence entry text box and the **From** and **To** values if they're specified. To submit your BLAST query, simply click on the **Search** button. Depending on the number and size of your input query sequences your BLAST query may take as long as 5-10 minutes. *Your screen will not refresh until the analysis completes!*

The bottom half of the *BLAST Sequence Search* tool allows more advanced users to fine-tune how BLAST performs its search. We recommend using the default settings however you should try adjusting the settings if an expected hit is not found. Below we will explain how these options affect the overall search however you should refer to NCBI's BLAST help page for more information if you still have questions (➧ `http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml`).

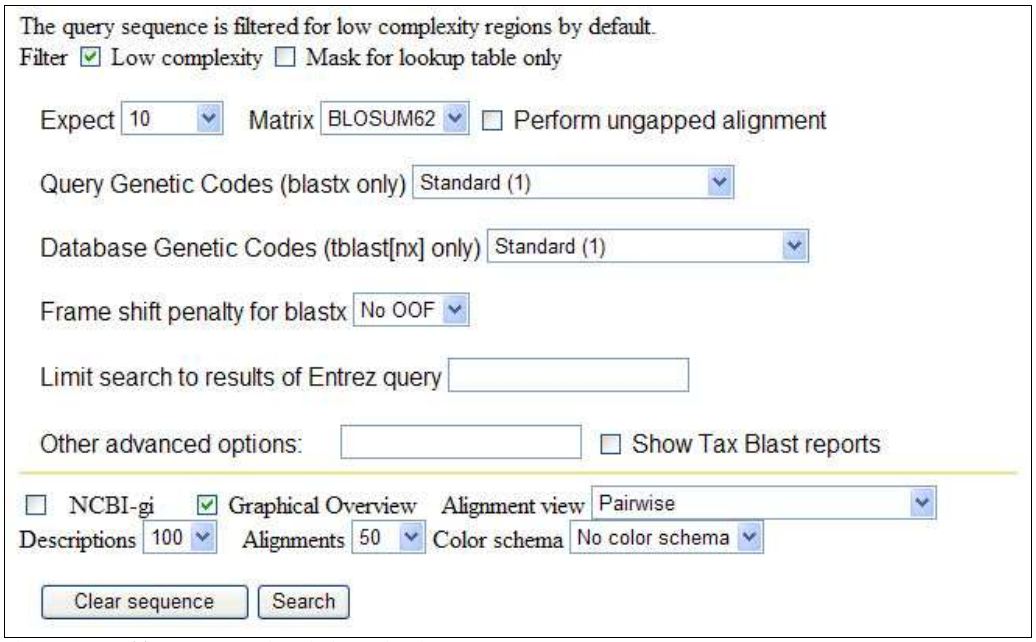


Figure 15. BLAST Sequence Search tool advanced options

Filter Low Complexity

Enabling this option will cause BLAST to mask out low complexity regions in your input query sequence. This will prevent BLAST from finding and extending seed alignments in or through low complexity regions. This behavior is slightly modified if you also enable the **Mask for lookup table only** option explained below. By default this option is enabled.

Mask For Lookup Table Only

By enabling this option in conjunction with the **Filter Low Complexity** option, you will allow BLAST to extend seed alignments through low complexity regions. However, BLAST still will not start an alignment in a low complexity region. By default this option is disabled.

Expect

Each hit produced by BLAST has an E value that represents the likelihood of the hit occurring by chance. Smaller E values tend to mean the hit is more significant. Setting the **Expect** threshold requires all hits have an E value less than or equal to the specified threshold. Using an initially small **Expect** threshold will result in a faster search time however you may inadvertently exclude hits. Conversely, if your BLAST query results in too many hits you should lower the **Expect** threshold to reduce some of the lower quality hits. By default this option is set to **10**.

Matrix

Changing the matrix affects the scoring that BLAST will use internally for matches and the penalties it uses for mismatches. This in turn could potentially alter the structure of returned hits. You shouldn't change this setting unless you are familiar with the different BLAST matrices and how they affect the overall search. By default this option is set to **BLOSUM62**.

Perform Ungapped Alignment

Once BLAST has identified candidate seed regions, it will extend these regions inserting gaps as needed to maximize the internal score of the hit. Enabling this option will disable insertion of gaps during the extension phase. By default this option is disabled.

Genetic Codes

Modifying these parameters alters BLAST translational queries and allows certain codons to be substituted during the translation phase. In general, these should be left as default unless you know that a particular input query and database selection are compatible with the selected genetic codes. For more information, please see the NCBI Genetic Codes website.

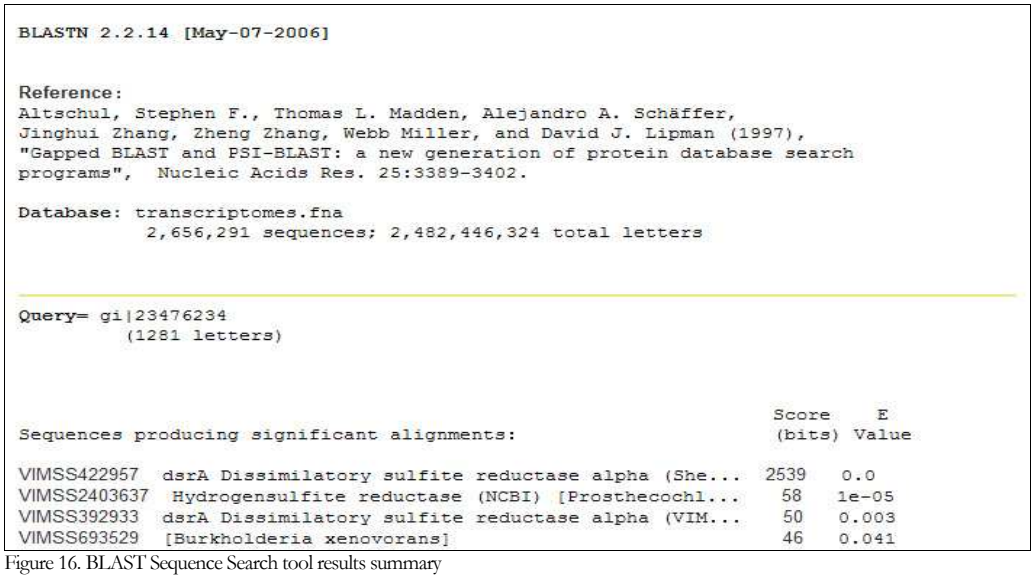
`http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c`

Frame Shift Penalty

This option determines whether frame shifts are allowed for translational BLAST programs and if so, the penalty that should be assumed for the frame shift. The value **No OOF** means *No out of frame* or that you do not wish to allow frame shifts.

BLAST Sequence Search Results

The results of the *BLAST Sequence Search* are returned in textual format with very little post-processing. The output is essentially raw BLAST output with a couple of exceptions. Each input query sequence will have its own set of results and each result set begins with a summary of all identified hits found within the criteria specified using the search tool.



```
BLASTN 2.2.14 [May-07-2006]


Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Database: transcriptomes.fna
           2,656,291 sequences; 2,482,446,324 total letters


Query= gi|23476234
         (1281 letters)



                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

VIMSS422957  dsrA Dissimilatory sulfite reductase alpha (She...  2539   0.0
VIMSS2403637  Hydrogensulfite reductase (NCBI) [Prosthecochl...    58   1e-05
VIMSS392933  dsrA Dissimilatory sulfite reductase alpha (VIM...    50   0.003
VIMSS693529  [Burkholderia xenovorans]                            46   0.041
```

Figure 16. BLAST Sequence Search tool results summary

Each hit is summarized on a single line including the database's sequence id, often a VIMSS id, a description of the sequence followed by the hit's bit score and E value. All VIMSS ids are clickable, for example **VIMSS422957** in the above figure, and clicking on a VIMSS id will redirect you to the *Locus Information* page.

Following the summary, each hit alignment is displayed in detail showing the region of the hit's span on the input query sequence and database sequence along with match/mismatch information. A summary of the bit score, E value, the number of identities (and %-identity), and the orientation of the hit are also provided. The following figure shows a sample of what each hit alignment will look like. Again the VIMSS id will be clickable.

NOT SUPPORTED

When BLASTing against the **VIMSS genomes** database, hyperlinks are not generated to the *Genome Information* page as you might anticipate.

```
>VIMSS422957 dsrA Dissimilatory sulfite reductase alpha (Shelley Haveman)
             [Desulfotalea psychrophila LSv54]
          Length = 1281

 Score = 2539 bits (1281), Expect = 0.0
 Identities = 1281/1281 (100%)
 Strand = Plus / Plus


Query: 1    atggcaaaacatgagacgcccttgttggatcagttggagagtggcccgtggccgagcttt 60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1    atggcaaaacatgagacgcccttgttggatcagttggagagtggcccgtggccgagcttt 60


Query: 61   gtaactgacattaaacgccaagctgagaagaagcccgagtgttgggatatcctcggtatc 120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 61   gtaactgacattaaacgccaagctgagaagaagcccgagtgttgggatatcctcggtatc 120


Query: 121  ttggaactgtctttcaaagagagaatcactcactggaagcacggcggaatcgttggtgtt 180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 121  ttggaactgtctttcaaagagagaatcactcactggaagcacggcggaatcgttggtgtt 180


Query: 181  tttggatacggtggtggtatcgttggtcgttatgcggatgtacctgagcgttttccaggt 240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 181  tttggatacggtggtggtatcgttggtcgttatgcggatgtacctgagcgttttccaggt 240
```

Figure 17. BLAST Sequence Search tool showing part of a single hit alignment.

Advanced Search

The *Advanced Search* tool allows you to locate terms from the various external databases used in annotating the genes in MicrobesOnline. Once a term has been located the *Advanced Search* tool provides a link to use the term in our *Find Genes* interface allowing you to locate all genes in the selected genomes that have been annotated to include said term.

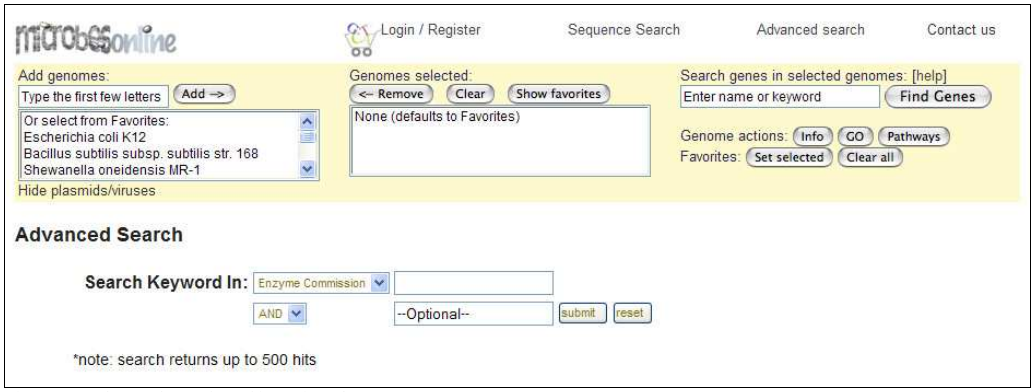


Figure 18. Advanced Search tool input interface

The input interface is simple. First, select from the list of available external data sources. At present, this includes **Enzyme Commission**, **KEGG Pathways**, **Gene Ontology**, and **COG**. Next specify one or more keywords in the text input field to the right of the data source drop-down selection box. You may enter additional keywords to filter your search as well as select how the two sets of keywords will be used to determine the final set of terms to display. By selecting **AND** only terms with both

sets of keywords will be returned. Selecting **OR** will return terms matching the first set of keywords or the second set of keywords, but not necessarily both. Finally, selecting **NOT** will return terms that match the first set of keywords but not the second set of keywords. The second set of keywords is ignored if the value is left blank or as the default value **--Optional--**. In this case, selecting the keyword set joining modifiers **AND**, **OR**, or **NOT** will have no effect on the returned results.

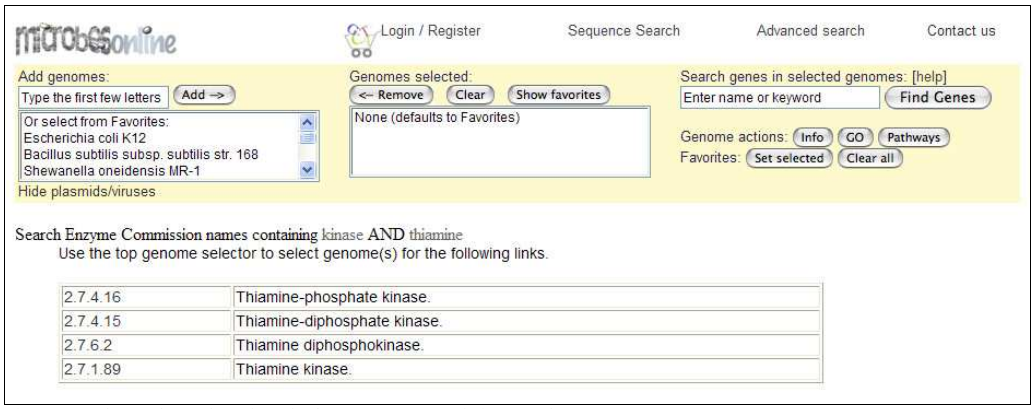


Figure 19. Advanced Search tool results for Enzyme Commission search

In the sample results shown above for the query **`kinase AND thiamine`**, four (4) results are returned. In general, the results will contain two columns—the first will provide a hyperlink to the *Find Genes* interface and show the term name and the second will show the term description. Clicking on a hyperlinked term will search for all genes that have been tagged with the selected term. Please see details on the *Find Genes* interface on page 8 of this guide for more information on how the tool works and how to interpret its results.

> ⚠ **WARNING**
>
> The order of keywords within a set matters. In the above example, searching for **`thiamine kinase`** would return **`EC2.7.1.89`** however searching for **`kinase thiamine`** would return no results.

> ⚠ **WARNING**
>
> Only term descriptions are searched. Searching for **`2.7.1.89`** would yield no results.

Genome Information

The *Genome Information* page has three basic modes—a summary display for showing information of two or more genomes, a more detailed single-genome view, and a gene

list view. Many tools that provide hyperlinks to the *Genome Information* page do so for a single genome however some, including the MicrobesOnline home page, provide a link to the summary view.

To access the *Genome Information* page from the MicrobesOnline home page, configure a set of genomes using the *Genome Selector* (page 6) then click on the **Info** button located on the right side of the *Genome Selector* display. Recall that if you do not configure a set of genomes, the default MicrobesOnline "favorites" will be used.

Summary View

The *summary view* provides a basic overview of the selected genomes. Often this is the set of genomes you've selected however other tools within MicrobesOnline may link to the *summary view* as well. Information on the selected genomes is presented in a table initially ordered by the genome name in ascending order. As with most tables on MicrobesOnline you can alter the sorting of the table by clicking on the column heading for the column you wish to use as the sorting criteria. Clicking a column heading multiple times will toggle between the default ascending sort and descending sort.

Information for 12 Genomes View Species Tree Show All Genomes, incl. Plasmids

Click on headings to sort the table. Please be patient if showing all genomes.

Genome	Phylum	Paper	Loaded	Complete	#Chr.	#Plasmids	#Genes
Bacillus subtilis subsp. subtilis str. 168	Firmicutes	yes	2007-05-08	yes	1	0	4225
Bdellovibrio bacteriovorus HD100	Proteobacteria	yes		yes	1	0	3623
Desulfotalea psychrophila LSv54	Proteobacteria	yes	2004-10-19	yes	1	2	3321
Desulfovibrio desulfuricans G20	Proteobacteria			yes	1	0	3291
Desulfovibrio vulgaris Hildenborough	Proteobacteria	yes	2005-02-08	yes	1	1	3633
Desulfuromonas acetoxidans DSM 684	Proteobacteria		2007-05-08	no			3318
Desulfuromonas spp.	Proteobacteria			no			5986
Escherichia coli K12	Proteobacteria	yes	2007-05-08	yes	1	0	4488
Escherichia coli W3110	Proteobacteria		2006-04-12	yes	1	0	4449
Geobacter metallireducens GS-15	Proteobacteria		2006-01-18	yes	1	1	3634
Geobacter sulfurreducens PCA	Proteobacteria	yes		yes	1	0	3470
Shewanella oneidensis MR-1	Proteobacteria	yes		yes	1	1	4637

Figure 20. Genome Information summary view of the MicrobesOnline "favorite" genomes

Clicking on the **View Species Tree** link display the MicrobesOnline *Species Tree* tool with the selected genomes highlighted. For more information on the *Species Tree* tool, please see the *Species Tree* section of this guide. Clicking on the **Show All Genomes, incl. Plasmids** link will refresh the *summary view* with all genomes present in MicrobesOnline including viruses and plasmids. As of the current release of MicrobesOnline this includes well over 2,400 genomes therefore loading this view and/or performing sorting operations may be time consuming.

Clicking on the individual genome names will load the *single-genome detail view* for that genome. Clicking on a phylum shown in the **Phylum** column will refresh the *summary view* to show all genomes from the selected phylum. Note that some phyla are extremely large therefore loading the view and/or performing sorting operations may be time consuming.

Finally, the **Selected Genomes** from the *Genome Selector* are automatically updated to show all of the genomes being displayed in the *summary view*. Therefore, using any of the other actions in the *Genome Selector* will apply to all genomes currently being displayed and not the last set of genomes you selected.

Single-Genome Detail View

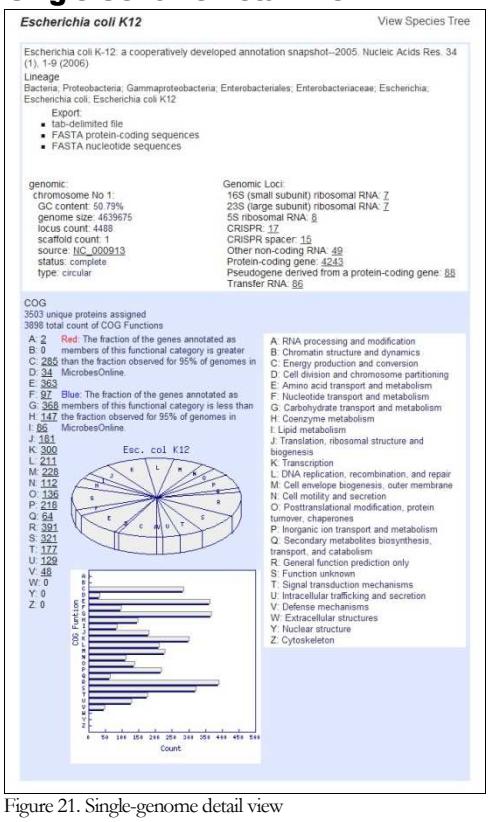


Figure 21. Single-genome detail view

The *single-genome detail view* provides a wealth of information about an individual genome including basic statistics about its sequence, paper references, loci statistics, and a distribution of genes by COG function.

Clicking on the **View Species Tree** link located in the upper-right corner of the *single-genome detail view* will redirect you to the *Species Tree* tool highlighting only the selected genome within the tree.

The first section of the display lists any papers in our database referencing the selected genome. In the example, a single paper reference *"Escherichia coli K-12: a cooperatively developed annotation snapshot"* is displayed. Clicking on a paper reference will redirect you to NCBI PubMed where you can read the paper abstract and access the full paper if you have the proper subscriptions.

The next section displays the full lineage of the genome from kingdom to sub-species or strain, if available. The individual lineage components are hyperlinked to the NCBI Taxonomy Browser.

The **Export** section provides links to download locus coordinates and other information in tab-delimited form, and both protein-coding genes and transcripts in FASTA format. This information is generally useful if you wish to perform analysis off-site.

The next section shows basic information about the genome's sequence(s) and loci, including the %-GC content, size, source, status and other information for each chromosome, and a distribution of loci according to type. The source sequence identifier is a hyperlink to view additional information about the sequence and will redirect you away from the MicrobesOnline site. The counts in the **Genome Loci** section each link to the *gene list view* showing the details of the genes representative in each count.

The last section, **COG**, shows a breakdown of the assignment of genes to COG functions. Each COG function category is assigned a letter from A-Z and the legend is displayed to the right along with the COG function description. For each COG function there is an associated count of the number of genes with that particular assignment, and both a pie and bar graph showing the same. If a particular COG function count is greater than zero, it will become a hyperlink to the *gene list view* showing the details of the genes representative in that count.

To return to the *summary view*, use your web browser's **Back** button.

Gene List View

The *gene list view* shows subsets of the genes in the selected genome that represent a particular count from which the view was hyperlinked. In the **Genome Loci** section of the *single-genome detail view*, genes counts are by type whereas in the **COG** section genes are shown by COG functional assignments. Clicking on the counts from either section will result in the *gene list view*.

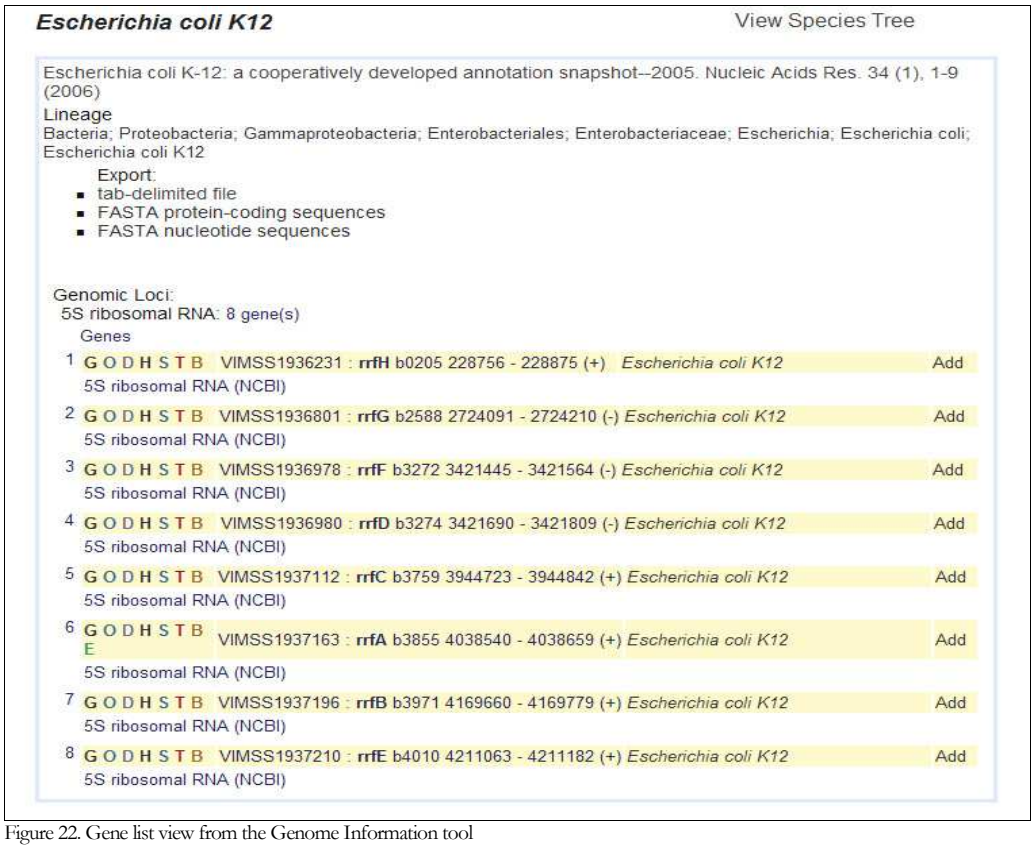


Figure 22. Gene list view from the Genome Information tool

The header of the *gene list view* is similar to the header of the *single-genome detail view*. Clicking on the **View Species Tree** link will redirect you to the *Species Tree* tool with only the selected genome highlighted within the tree. First, a list of paper references are displayed, which are each linked to NCBI PubMed. Next, the full lineage is

displayed. The **Export** section contains the same download links as in the *single-genome detail view*, providing a link to download genes in tab-delimited form or as proteins and transcripts in FASTA format.

The subsequent gene list is shown with a similar format to the *Find Genes* tool. Each gene is displayed as a single row. The first block of lettered links corresponds to additional tools within MicrobesOnline that can provide you with additional information about a gene. Specifically, the **G O D H S** links will allow you to access detailed information about the gene from the *Locus Information* tool. The **T** link will redirect you to the *Tree Browser* showing the gene tree that contains the selected gene. Note that not all genes belong to gene trees therefore occasionally you may see an error stating that the selected locus *"…is not in any pre-computed trees."* The **B** link will redirect you to the *Ortholog Browser* showing the gene in the context of its gene neighborhood and finally for genes where expression data is present, the **E** link will redirect you to the *Expression Data Viewer*.

Clicking on the genome name, although redundant in this case, will redirect you to the *single-genome detail view* of the *Genome Information* tool however using the **Back** button on your browser is generally faster and will accomplish this task as well. Lastly, you can add a gene to your *session gene cart* by clicking on its corresponding **Add** link, located right-most within each displayed result row. Finally, beneath each result row a description of the gene is displayed.

GO Browser

The *Gene Ontology (GO) Browser*, also known as *VertiGo*, provides an interface to traverse and browse the Gene Ontology hierarchy, which contains a hierarchical list of biologically relevant terms organized by biological process, cellular component, and/or molecular function. GO terms are assigned to genes through a mapping of conserved InterPro domains provided by the Gene Ontology Consortium. Identification of conserved InterPro domains is one part of our standard gene analysis pipeline. For more information about the Gene Ontology database, please visit the Gene Ontology Consortium's website.

`http://www.geneontology.org/GO.doc.shtml`

There are two views within the *GO Browser*—the *ontology browser view* allows you to navigate the Gene Ontology hierarchy and the *gene list view* allows you to view the list of genes that have been associated with a particular GO term and genome.

To update the list of genomes shown in the *GO Browser*, simply modify the list of genomes in the *Genome Selector* then click on the **Go** button located near the right side of the *Genome Selector* display.

Ontology Browser View

The *ontology browser view* is the standard view for the *GO Browser* and allows you to navigate the GO hierarchy in the context of representative gene counts from one or more genomes. Each selected genome is displayed as a column from left-to-right in the order in which the genomes are listed in the *Genome Selector*, along with navigation options and descriptions of the GO term hierarchy. The column headings show shortened forms of the selected genomes' names. For example the genome name *Escherichia coli K-12* is displayed as *Ecol K12* in the column heading. Each GO term is displayed on its own line and indentation is provided to show the hierarchy of the terms. To the left of each term, the corresponding counts of the number of genes associated with that term are displayed. Note that not all genes are assigned GO terms and some genes are assigned multiple differing GO terms, therefore the total gene counts shown to the left of the **all: all** GO term do not necessarily reflect the total gene count for each genome.

If the count for a particular genome and GO term is 500 genes or less, the count itself becomes a hyperlink that will redirect you to the *gene list view* for the selected genome and GO term otherwise the counts are not clickable.

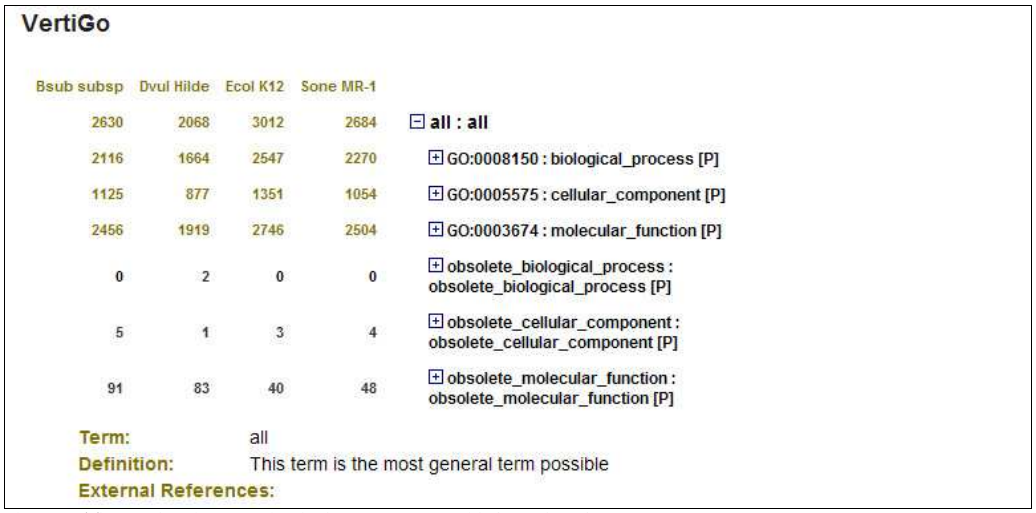


Figure 23. The ontology browser view in the GO Browser with four selected genomes

The **+** and **–** icons located to the left of each GO term are used to indicate whether there are additional terms for which the term is a parent. For example, the six terms shown in the above example are all *children* of the term **all: all**. To select a node to expand, click on the **+** or **–** icon located to the left of the term. If you wish to view the representative genes for a particular term and genome whose count is larger than 500 genes, you must expand the term as appropriate until its child terms yield smaller counts.

The selected term is always displayed beneath the count and term list, showing its name and definition. When available, external references are also provided.

Gene List View

The *gene list view* displays the list of genes for a selected genome and GO term with results shown using a display similar to the one used in the *Find Genes* tool. At the top of the *gene list view*, the selected GO term is shown highlighted in bold along with all parent terms and each corresponding count, for the selected genome. The selected term's details are displayed next, just as in the *ontology browser view*.

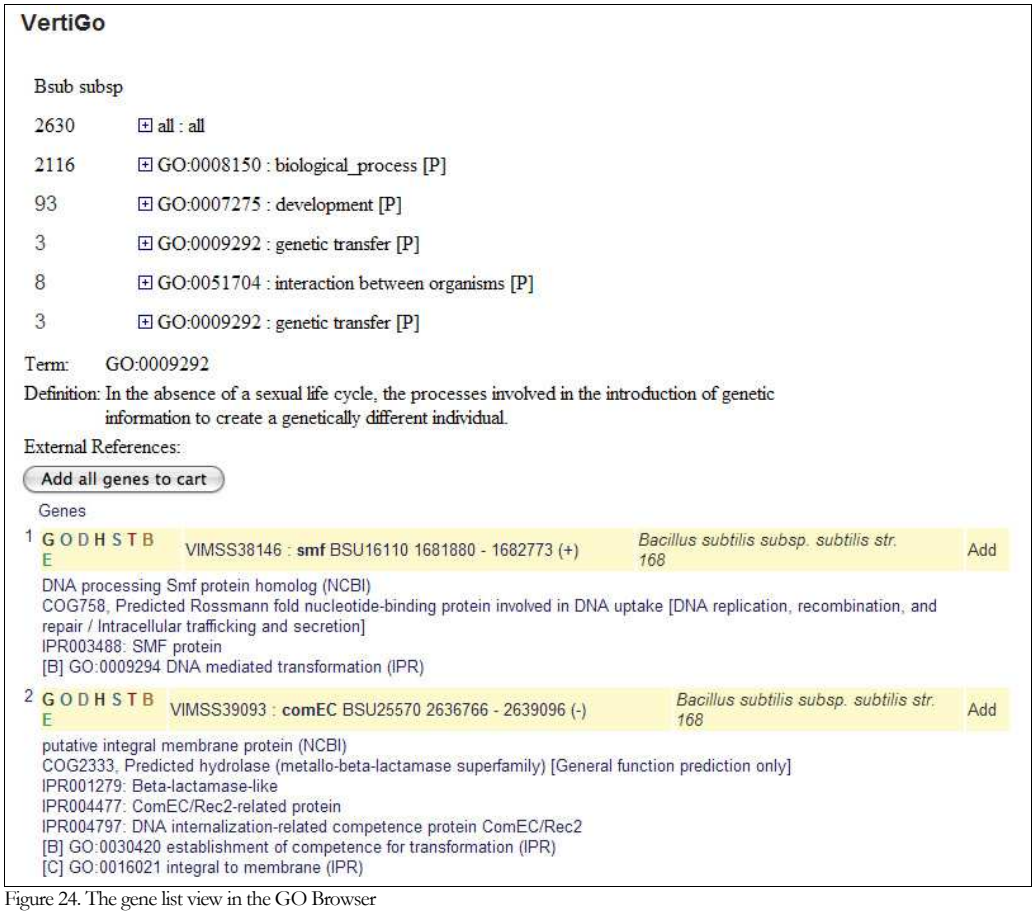


Figure 24. The gene list view in the GO Browser

The gene list is shown in row format with one gene per row. As with the *Find Genes* tool and the *gene list view* in the *Genome Information* tool, each row contains a series of lettered links. The **G O D H S** links redirect you to the *Locus Information* page. The **T** link will redirect you to the *Tree Browser* unless the selected gene isn't represented in a pre-computed gene tree, in which case an error is displayed. The **B** link will redirect you to the *Ortholog Browser* and finally, when expression data are available for the selected gene, the **E** link will redirect you to the *Expression Data Viewer*.

Clicking on the genome name will redirect you to the *Genome Information single-genome detail* page. Clicking the individual **Add** links will add single genes from the list to your *session gene cart*. Clicking on the **Add all genes to cart** link will add all genes displayed on the page to your *session gene cart*.

Beneath each gene, details such as the gene description and a list of COG assignments, conserved InterPro domains and GO assignments are shown.

Tutorial: Finding Genes Associated with a GO Term

If you know the GO term id for the term whose associated genes you wish to look up, you can enter its GO term id into the *Find Genes* search box located toward the right side of the *Genome Selector* display. For example, if you wish to find all genes associated with **GO:0019684** you need only enter this term as a search query in the *Find Genes* search box.

In most cases, however, we imagine you haven't memorized all the GO term ids, but there may be a particular term whose associated genes you wish to find. You may not even know the GO term's name exactly, but that's okay. On the top navigation, select **Advanced search**. If you do not see this link, return to the MicrobesOnline home page by clicking the MicrobesOnline logo or entering `http://microbesonline.org` into your browser.

After the **Search Keyword In** text, click the drop-down select box and highlight **Gene Ontology**. In the text input field to the right, enter part of the GO term name. It's best just to use a single keyword from the term name that is most representative of the term. For example, let's say you wanted to look for the term *"photosynthesis, light reaction"* a good keyword to use would be **photosynthesis**. Click the **submit** button.

Look through the results for the desired term and click on its GO term id. This will automatically redirect you to the *GO Browser* where you can see the distribution of gene counts for the selected GO term and the selected genomes. To view the gene lists, click on the counts associated with the desired GO term.

This method is generally faster than browsing the GO hierarchy unless you know exactly where the desired term is located within the hierarchy!

Pathway Browser

The *Pathway Browser* allows you to navigate the Kyoto Encyclopedia of Genes and Genomes (KEGG; `http://www.kegg.com`) pathway maps in the context of predicted presence or absence of enzymes for up to two selected genomes. If more than two genomes are selected, you will receive an error. There are two views present in the *Pathway Browser*—the *map view* shows the selected KEGG pathway map and as well as the presence or absence of enzymes in the selected genomes, if any, and provides a means of navigating to other KEGG pathway maps. The *gene list view* is similar to other views of the same name in different tools—it provides a list of genes associated with a particular enzyme in the selected genomes.

Genes are assigned KEGG enzyme commission (EC) numbers through homology to the KEGG representative protein set. This analysis is performed as part of our standard gene analysis pipeline. Note that since EC assignments are based solely on homology that an assignment is not guaranteed to be correct. Likewise, a lack of an assignment does not indicate that a gene isn't associated with a particular enzyme.

Map View

The *map view* showing the top-level **Metabolic Pathways** map (not shown here) is the initial entry view of the *Pathway Browser*. Each *map view* shows the map title followed by other KEGG pathway maps associated with the pathway map you are viewing.

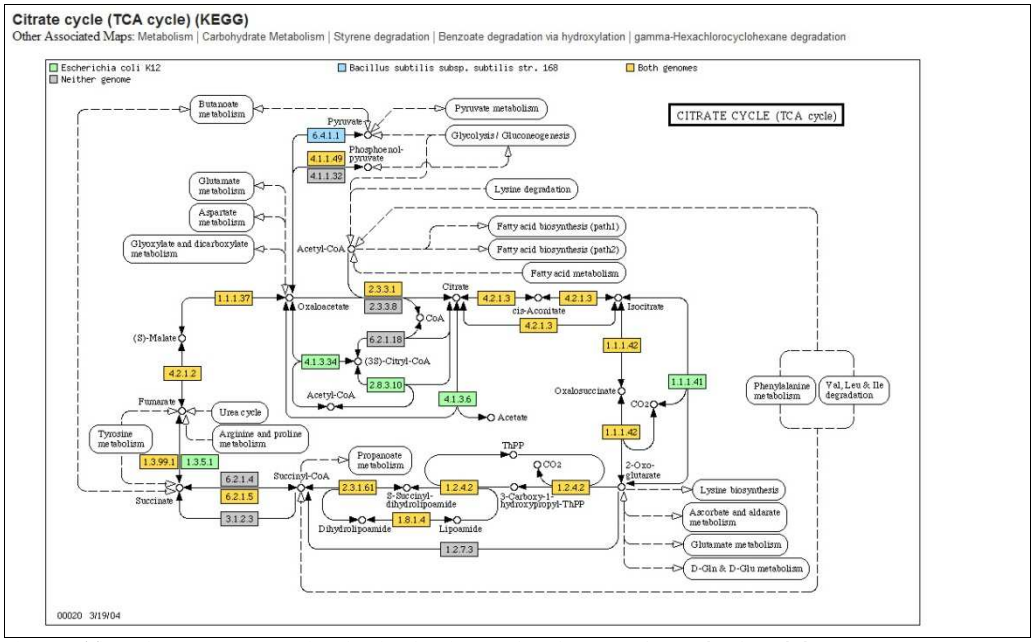


Figure 25. The Pathway Browser map view showing the Citrate cycle pathway map for *E. coli* and *B. subtilis*

Next, the pathway map itself is displayed. If you've selected one or more genomes, a legend is overlaid on the pathway map showing the colors used to indicate the presence or absence of enzymes, or in the case of two genomes whether at least one gene from both genomes are associated with a particular enzyme or if no genes from either genome are associated with a particular enzyme.

In the pathway map, enzymes are generally shown as rectangular boxes each of which contains the KEGG enzyme commission (EC) number assigned to the enzyme. Compounds are represented as small circular nodes and references to other pathways are generally shown in rectangle boxes with rounded corners.

Clicking on references to other pathways, or clicking on the pathway map names displayed as **Other Associated Maps** will refresh the *map view* with the newly selected pathway map. Clicking on compound nodes will currently redirect you to the KEGG

website showing details for the compound. In future releases of MicrobesOnline, we may color presence and absence of compounds based on metabolomics experiments.

If you've selected one or more genomes and a particular enzyme is represented by at least one gene from a genome, clicking its enzyme icon in the pathway map will redirect you to the *gene list view* showing all genes from the one or two selected genomes that have been associated with the selected enzyme. If you haven't selected any genomes or if an enzyme is not represented by any genes in the selected genomes, the enzyme icons will link to KEGG to provide you with more information about the enzyme.

Lastly, the date that appears in the lower-left corner of most KEGG pathway maps represents the date the map was last updated.

Gene List View

The *gene list view* shows all genes associated with the selected enzyme for the selected genomes. As with other *gene list views* in other tools, this view will provide basic information about the selected enzyme, including all KEGG pathway maps in which it is a member, followed by list of genes shown in row format. The gene list itself is identical to the gene list in the *gene list view* of the *GO Browser*. Please refer to this section on page 27 for help with interpreting the gene list.

Figure 26. The Pathway Browser gene list view

Tutorial: Identifying Metabolic Differences Between Two Genomes

The *Pathway Browser* can be used to help you identify metabolic differences between two genomes. To do this, use the *Genome Selector* to select the two genomes you're interested in comparing, then click on the **Pathways** button located toward the right side of the *Genome Selector* display. This will allow you to browse the KEGG metabolic pathways and to quickly identify differences the metabolic differences between the two genomes you've selected.

Instead of browsing all the KEGG metabolic pathways, you might instead be interested in specific pathways. To search for KEGG metabolic pathways by name, use the *Advanced Search* feature. To access this feature, click on the **Advanced search** link located in the top navigation of most pages on the MicrobesOnline site. If you do not see this link, return to the MicrobesOnline home page by clicking on the MicrobesOnline logo or entering `http://microbesonline.org` into your web browser, then click on the **Advanced search**.

Select **KEGG pathways** from the **Search Keyword In** drop-down selection box then, to the right, enter a keyword corresponding to the pathway you wish to browse. For example, if you wish to compare the *Purine metabolism* pathway of two genes enter the keyword **purine**, then click on the **submit** button.

The results will show all KEGG metabolic pathways with names containing the keyword you specified. Next, select the two genomes you wish to compare by using the *Genome Selector*. Finally, click on the KEGG metabolic pathway id, for example **`00230`** for *Purine metabolism*. This will take you to the desired metabolic pathway with the two desired genomes, without knowing exactly where the pathway is located within KEGG.

Tutorial Case Study: Free-living vs. Endosymbiont

Perhaps one of the most striking examples of the differences in metabolic capabilities between two organisms comes from comparing the metabolic pathway maps of a free-living organism to those of a highly reduced endosymbiont.

For example, when comparing the profiles of *E. coli K-12* to that of the insect endosymbiont *Buchnera aphidicola*, we see high levels of conserved metabolic capability for the synthesis of amino acids required by the insect host in the *Buchnera* genome (for example the ↳VAL, ILE, and LEU <`http://www.microbesonline.org/cgi-bin/kegg2.cgi?mapId=00290&taxId=83333&taxId=107806`>, and ↳LYS synthesis <`http://www.microbesonline.org/cgi-bin/kegg2.cgi?mapId=00300&taxId=83333&taxId=107806`> pathways), however the *Buchnera* genome is missing the required pathways for amino acids synthesized by its host (for example the ↳GLY and SER synthesis <`http://www.microbesonline.org/cgi-bin/kegg2.cgi?mapId=00260&taxId=83333&taxId=107806`> pathway).

Species Tree

The MicrobesOnline *Species Tree* viewer shows most MicrobesOnline bacterial and archaeal genomes according to their predicted evolutionary relationship. The tree is generated from a variety of support trees using *matrix representation of parsimony (MRP)* and the support trees are generated using concatenated alignments of highly conserved proteins primarily with the *maximum likelihood* method. A full description of the *Species Tree* method is located here:

➧ `http://www.microbesonline.org/treebrowseHelp.html#speciestree`

Links to the *Species Tree* are found in the *Genome Information* tool pages. In the *summary view*, clicking the **View Species Tree** link will redirect you to the *Species Tree* viewer showing all of the genomes displayed in the *summary view* highlighted in magenta. From the *single-genome detail view* and *gene list view*, clicking the **View Species Tree** link will redirect you to the *Species Tree* viewer with only the currently selected genome highlighted in magenta within the tree.

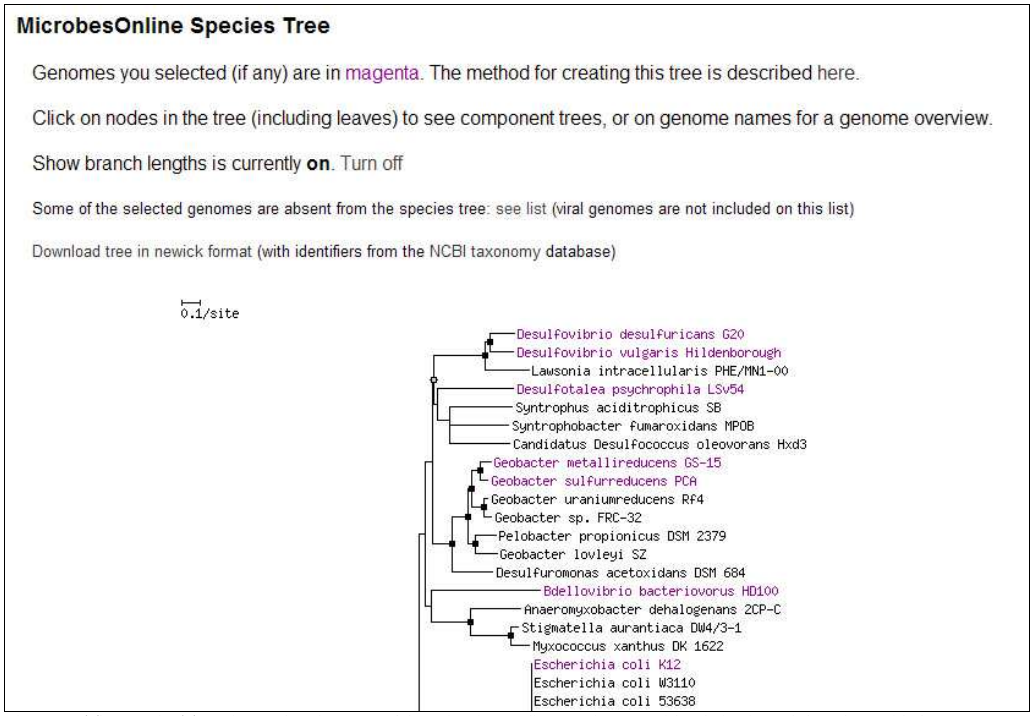


Figure 27. The Species Tree viewer showing part of the MicrobesOnline species tree with relative branch lengths enabled

Branch lengths in species trees are an approximate measure of distance between genomes. Our *Species Tree* viewer allows you to disable relative branch lengths by clicking on the **Turn off** link shown to the right of the current branch length rendering status. By default, rendering of relative branch lengths is enabled, as with the example shown previously. Part of the *Species Tree* viewer is shown below when relative branch lengths are disabled.

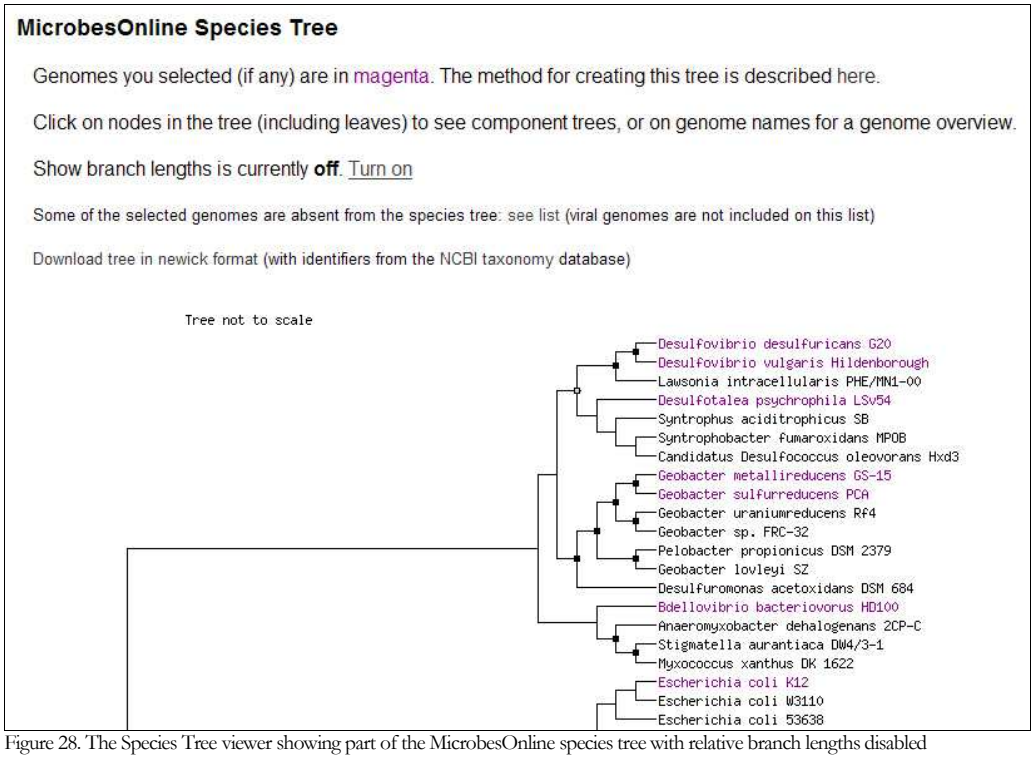


Figure 28. The Species Tree viewer showing part of the MicrobesOnline species tree with relative branch lengths disabled

The nodes within the species tree allow you to access the component trees that are built as an intermediate step of the final species tree and offer support for the arrangement of genomes and branch lengths chosen. One such component tree is shown below.

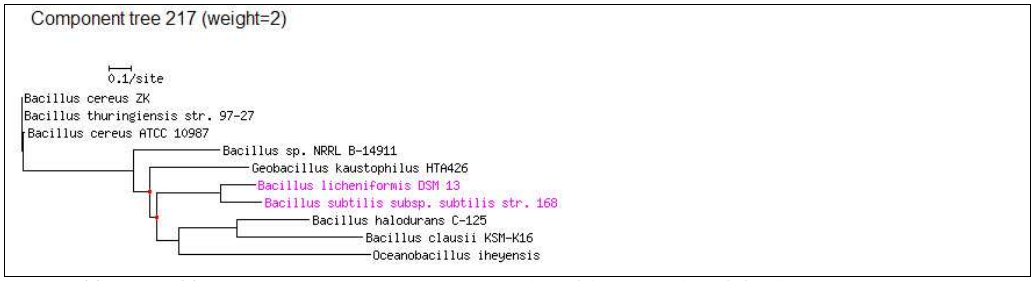


Figure 29. The Species Tree viewer showing one component tree in the *B. licheniformis* and *B. subtilis* split

If any of the selected genomes from the *Genome Information summary view* are not present in the final species tree, you will see a warning, such as the one shown in figure 28 that reads *"Some of the selected genomes are absent from the species tree."* To see the list of selected genomes that are absent from the species tree, click on the **see list** link.

Clicking on the genome names will redirect you to the *Genome Information* tool *single-genome detail view* for the selected genome.

Finally, if you wish to download out species tree for your own analysis or rendering, it is available for download in *Newick* format from our *Species Tree* viewer by clicking on

the **Download tree in newick format** link. Each node in the downloaded tree is labeled with the genome's NCBI taxonomy id rather than its full name.

Locus Information

The *Locus Information* tool is the primary method for retrieving all of the information associated with an individual gene. The information is neatly organized into five tabs labeled from left to right **Gene Info**, **Operon & Regulon**, **Domains & Families**, **Homologs**, and **Sequences**. A sixth tab **Add Annotation** is available for registered users wishing to add new annotation or to modify or delete existing annotation associated with a gene.

There are links to the *Locus Information* tool from just about every page in MicrobesOnline that allows the user to retrieve and display a list of genes. For example, the *Tree Browser*, *Ortholog Browser*, *Find Genes* interface, *Expression Data Viewer*, *GO Browser*, and *Pathway Browser* all provide links to the *Locus Information* tool.

Common Navigation

Each of the *Locus Information* tabs shares a common header that can be used to navigate between tabs and to access other linked tools or resources.

Figure 30. Locus Information common navigation header

As shown in figure 30, each tab is clearly labeled. The **<< Previous Gene** and **Next Gene >>** links can be used to access the previous and next gene in sequence by their location in the genome. It is important to note that there is no implied functional relationship between the gene you are viewing and the previous and next genes.

The drop-down selection box located below the **Gene Info** and **Operon & Regulon** tabs shows a list of other genomes with a putative ortholog to the gene you're viewing. To jump to the *Locus Information* page for an ortholog in another genome simply click the drop-down selection box and select the desired genome.

> ⚠ **WARNING**
>
> MicrobesOnline occasionally hosts private data for internal use or for collaborators. These genes may be present in the ortholog drop-down selection box however you must be given access in order to view data from private genomes. If you do not have access, you will see a corresponding error message.

The **Add to Cart** link will add the current gene to your *session gene cart*. If you're logged in and have previously analyzed this gene, a **Related Carts** link will be shown beneath the **Add to Cart** link that will allow you to view the carts that contain the specified locus and their associated jobs. The **Browse Genomes** section allows you to view the gene within the context of a gene neighborhood. Clicking on the **trees** link will redirect you to the *Tree Browser* and clicking on the **orthologs** link will redirect you to the *Ortholog Browser* with the selected tool centered on the current gene.

If gene expression data is present for the current gene, you will see a clickable link labeled **Gene expression** that will redirect you to the *Expression Data Viewer*.

Lastly, the **Printable Version** link provides a single view showing all of the information from the component tabs with basic formatting for a printer. This allows you to quickly print all of the most relevant information about a gene without cycling through each and every tab.

Gene Info

The *gene info* tab shows all of the basic annotation information for a gene where the annotation type is shown in the left column and the actual annotation data in the adjacent right column. In addition to basic sequence information about a gene, such as its unique **VIMSS Id**, **Organism**, P**osition**, length, and %-**GC content**, both automated and user-supplied annotations are also provided. The **Name**, **Synonym**, and **Description** fields provide descriptive information about the gene, including the synonym chosen as the gene's official name.

Some annotation, such as the **UniProt** reference, **PDB**, **COG**, and **EC Number** are determined through homology. In some cases, a significant hit was not detected and therefore some or all of these fields could be blank. **UniProt** references, if present, are sometimes associated with paper references. If present you will see a **see papers** link next to the **UniProt** annotation that will redirect you to NCBI PubMed showing the abstracts of those papers. The **PDB** references will contain one or more RCSB structure ids and links, along with a link **see PDBs** that will redirect you to a summary view of all associated structures. The COG assignment, if present, is listed on its own line with a link to the *Find Genes* interface by clicking on the assigned COG id and a link to the *Tree Browser* showing all genes assigned to the selected COG in a gene tree.

Each EC assignment is also listed on its own line with a link to the *Pathway Browser* tool's *gene list view* showing all genes from the selected genomes that have the selected EC assignment.

VIMSS ID	209338
Organism	*Desulfovibrio vulgaris Hildenborough* (Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough, complete genome)
Name	**dsrA**
Synonym	dsrA,DVU0402,ORF05313
External links	**Accession:** YP_009626 **GI:** 46578818 **RegTransBase:** 2407641
Type	Protein, 437 a.a.
GC content	61.26%
Position	449888 .. 451201 (+) on Scaffold ID: 1944
Description	dissimilatory sulfite reductase alpha subunit (TIGR)
UniProt	DSVA_DESVH (**100%** identical): Sulfite reductase, dissimilatory-type subunit alpha (EC 1.8.99.3) (Desulfoviridin subunit alpha) (Hydrogensulfite reductase alpha subunit) (**see papers**)
PDB	

Figure 31. The top portion of the *gene info* tab on the *Locus Information* page

TIGRFam and **InterPro** assignments represent conserved domains identified using FastHMM. **TIGRFam** hits, if present, do not provide hyperlinks however **InterPro** hits provide a link to the *Find Genes* interface by clicking on the **InterPro** hit id. **Gene Ontology** (GO) assignments are inferred from conserved **InterPro** domains and link to the *GO Browser* showing the selected term expanded in the *ontology browser view*.

RegTransBase data is still not available for many genes, however when present it provides a list of associated RegTransBase papers and experiments and also provides a link to **RegTransBase** to view more information about the listed papers and experiments.

The **External links** field shows external database cross-references when available. Generally, there will be links to NCBI by accession number and GI number as well as a link to **RegTransBase**.

Annotation History	**Annotator:** Shelley Haveman (University of Calgary) **Date:** 23-Apr-2004 09:33 **Name:** dsrA (Replace) **Annotator:** TIGR (TIGR) **Date:** 12-Apr-2004 17:46 **Name:** dsvA (Replace) **Description:** dissimilatory sulfite reductase alpha subunit (Replace) **Info:** Append ecNum 1.8.99.3 Gene model from GenBank

Figure 32. The annotation history portion of the *gene info* tab on the *Locus Information* page

The bottom of the *gene info* page shows the annotation history for the gene. Each new annotation, regardless of whether it is the addition of new information or the removal or modification of existing information, is logged and kept in the annotation history. Each edit contains the name of the **Annotator** including their registered organization name, the D**ate** and time of the edit, and a list of fields that were modified with the supplied value and the type of edit.

To see the versions of all external databases used in the analysis on the MicrobesOnline site, click on the **Version information for external databases** link that is shown at the bottom of the *gene info* tab.

Operon & Regulon

The *operon & regulon* tab displays information about the VIMSS predicted operon and regulon as well as a set of enriched GO terms based on the operon and regulon predictions. In some cases, the operon structure has been confirmed and the **Confirmed Operon** is displayed in addition to the VIMSS predicted operon.

Predicted/Confirmed Operons

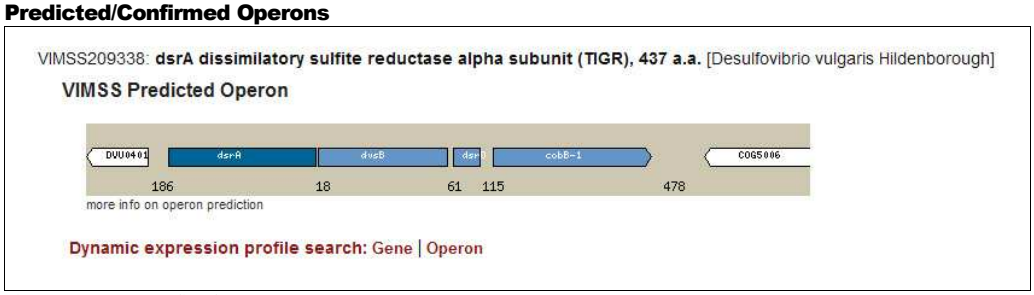


Figure 33. VIMSS predicted operon

The operon is shown in blue with the selected gene highlighted in dark blue. The gene name is displayed in the middle of each gene and the genes and intergenic regions are rendered relative to their actual lengths. The first genes upstream and downstream of the operon are also shown in white. All of the genes in the display can be clicked, redirecting you to the *Locus Information* page for that gene. If you would like more information on our operon prediction method, click on the **more info on operon prediction** link.

Finally, for *Desulfovibrio vulgaris Hildenborough*, *E. coli K-12*, and *S. oneidensis MR-1* if microarray experiment data is available, you can view dynamic expression profile searches based on the selected gene alone or based on the predicted or confirmed operon including the selected gene. If the gene does not belong to an operon of two or more genes, the **Operon** link will not be present. Clicking on the **Gene** or **Operon** links for genes in unsupported organisms will result in an error. You can find more information about expression profile searches in the *Expression Data View* section of this guide.

Predicted Regulons

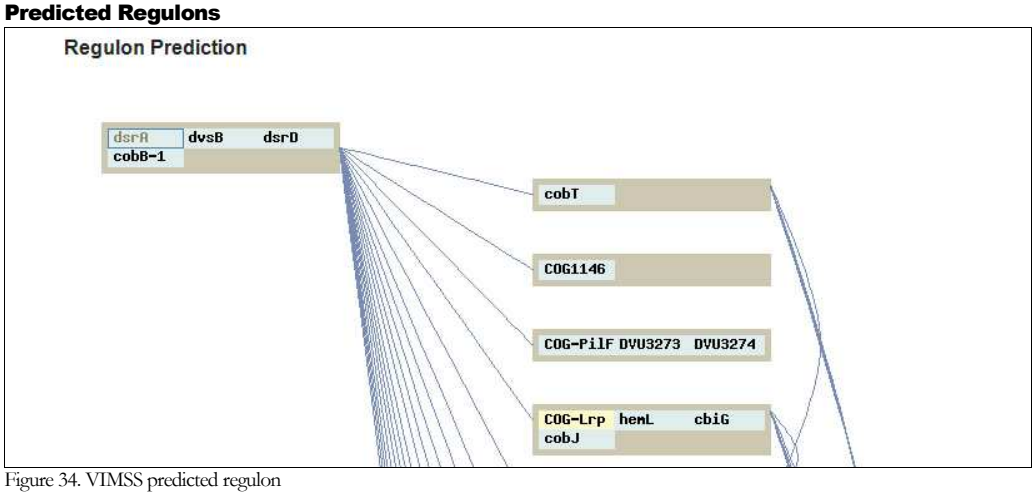


Figure 34. VIMSS predicted regulon

A regulon is a collection of genes believed to be under regulation by the same protein. The MicrobesOnline regulon predictions are quite complex, however the method is explained here:

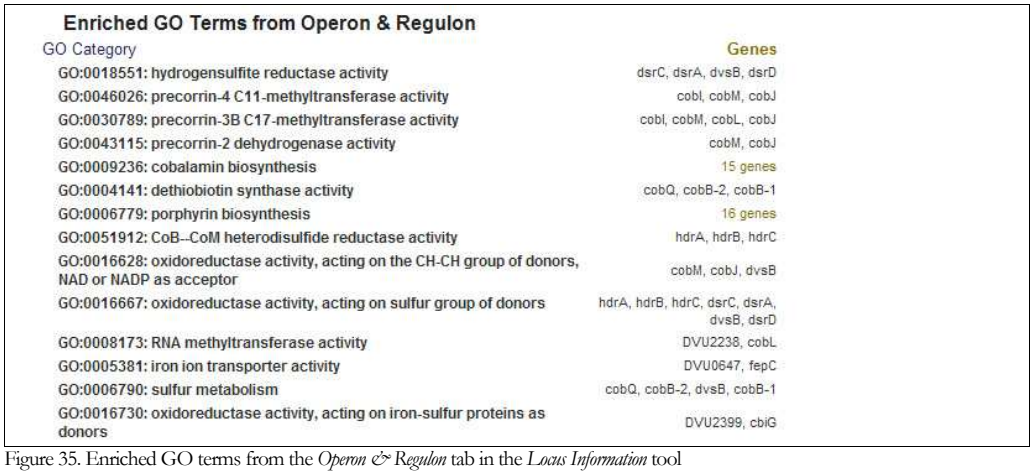
`http://www.microbesonline.org/about_regulon.html`

The regulon cluster of the selected gene is shown towards the upper-left of the graph, with the selected gene highlighted with a blue border. Genes in the regulon cluster are arranged by operons with each gene in the operon represented by a light blue box. Associations to other regulon clusters are formed using one or both of two methods, explained in the above reference. The first is a *gene neighbor method* which produces a score indicating how often orthologs of two genes within each regulon cluster are near each other in other bacteria. Links formed with this method are shown in blue. The second method allows linkages between regulon clusters if microarray data supports the linkage. These links are shown in red. If there is support from both the *gene neighbor method* and microarray data, the link is rendered in purple. Genes annotated as transcription factors, such as **COG-Lrp** in the above figure, are rendered in yellow.

Enriched GO Terms

This section shows all GO terms that are present in two or more genes shown in the predicted or confirmed operon, or in any regulon cluster. If a term is associated to two to 8 genes, the term is displayed along with a list of the associated genes. Each gene in this list is clickable and will redirect you to the *Locus Information* page for that gene. If a term is associated to more than 8 genes, the number of genes is displayed instead. In both cases, clicking on the GO term id and description will redirect you to the *GO Browser* with the selected term expanded.

The following figure shows an example of this section of the *operon & regulon* tab.



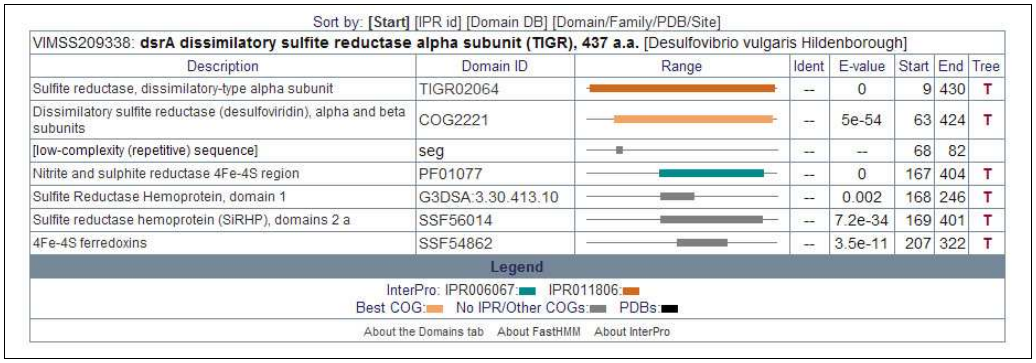
Enriched GO Terms from Operon & Regulon

GO Category	Genes
GO:0018551: hydrogensulfite reductase activity	dsrC, dsrA, dvsB, dsrD
GO:0046026: precorrin-4 C11-methyltransferase activity	cobI, cobM, cobJ
GO:0030789: precorrin-3B C17-methyltransferase activity	cobI, cobM, cobL, cobJ
GO:0043115: precorrin-2 dehydrogenase activity	cobM, cobJ
GO:0009236: cobalamin biosynthesis	15 genes
GO:0004141: dethiobiotin synthase activity	cobQ, cobB-2, cobB-1
GO:0006779: porphyrin biosynthesis	16 genes
GO:0051912: CoB--CoM heterodisulfide reductase activity	hdrA, hdrB, hdrC
GO:0016628: oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor	cobM, cobJ, dvsB
GO:0016667: oxidoreductase activity, acting on sulfur group of donors	hdrA, hdrB, hdrC, dsrC, dsrA, dvsB, dsrD
GO:0008173: RNA methyltransferase activity	DVU2238, cobL
GO:0005381: iron ion transporter activity	DVU0647, fepC
GO:0006790: sulfur metabolism	cobQ, cobB-2, dvsB, cobB-1
GO:0016730: oxidoreductase activity, acting on iron-sulfur proteins as donors	DVU2399, cbiG

Figure 35. Enriched GO terms from the *Operon & Regulon* tab in the *Locus Information* tool

Domains & Families

The *domains & families* tab shows hits from the selected gene to domains and gene families from various external databases. In our current analysis pipeline we search for conserved domains and gene families in TIGRFAM, Pfam, Superfamily, SMART, Panther, PIRSF, and Gene3D. In addition, we show hits to COG proteins and representative PDB proteins, as well as low-complexity or repetitive regions as determined by *seg*.



Sort by: **[Start]** [IPR id] [Domain DB] [Domain/Family/PDB/Site]

VIMSS209338: **dsrA dissimilatory sulfite reductase alpha subunit (TIGR), 437 a.a.** [Desulfovibrio vulgaris Hildenborough]

Description	Domain ID	Range	Ident	E-value	Start	End	Tree
Sulfite reductase, dissimilatory-type alpha subunit	TIGR02064		--	0	9	430	T
Dissimilatory sulfite reductase (desulfoviridin), alpha and beta subunits	COG2221		--	5e-54	63	424	T
[low-complexity (repetitive) sequence]	seg		--	--	68	82	
Nitrite and sulphite reductase 4Fe-4S region	PF01077		--	0	167	404	T
Sulfite Reductase Hemoprotein, domain 1	G3DSA:3.30.413.10		--	0.002	168	246	T
Sulfite reductase hemoprotein (SiRHP), domains 2 a	SSF56014		--	7.2e-34	169	401	T
4Fe-4S ferredoxins	SSF54862		--	3.5e-11	207	322	T

Legend

InterPro: IPR006067: IPR011806:
Best COG: No IPR/Other COGs: PDBs:

About the Domains tab About FastHMM About InterPro

Figure 36. *Domains & Families* tab in the *Locus Information* tool

For each identified hit, an id, description, range, and coordinates are shown. Mousing over the hit id will show the member database's name and clicking on the hit id will open a new browser window redirecting you to the website for the member database showing more detailed information about the target sequence, if applicable. For all domains aligned using an HMM, clicking on the graphical representation of the range or coverage of the hit will allow you to view the alignment in the *locus hmm alignment viewer*. For PDB hits, clicking on the range will allow you to view the alignment in the *locus pdb alignment viewer*. For all domain and gene family hits as well as hits to COG and PDB proteins, mousing over the displayed E value will show the score returned by alignment tool that identified the hit.

If the hit target sequence is present within the InterPro database, the corresponding InterPro description is shown and a link is provided to the InterPro website that will provide you with more information about the hit target. Mousing over the hit description will show the InterPro id, if available. The range shows the span of the hit relative to the selected gene's sequence and the hit span's bar is color-coded according to the legend displayed at the bottom of the results. In general, hits to the same InterPro id are colored the same. Unclassified hits are shown in grey.

MicrobesOnline provides pre-computed gene trees for selected families and when present the red **T** link will redirect you to that gene tree using the *Tree Browser*.

The results are shown ordered by their starting position relative to the selected gene. You can change the sort order by clicking on one of the **Sort by** links, which allows you to select the field you wish to use as your sorting criteria.

Locus HMM Alignment Viewer

The *locus hmm alignment viewer* allows you to view any HMM alignment used to identify a conserved domain or family in greater detail.

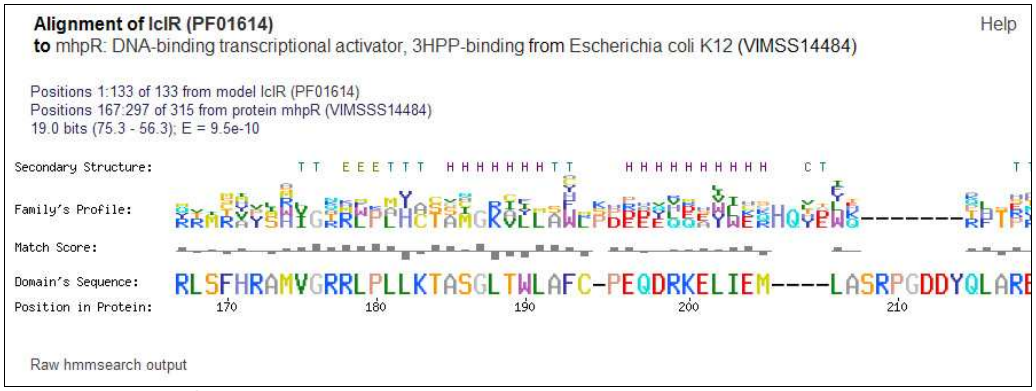


Figure 37. *Locus HMM alignment viewer*

The name of the model follows the text **Alignment of** and the name of the gene and genome appear below it, after the text **to**. Clicking on the gene name will return you to the *Domains & Families* tab of the *Locus Information* tool and clicking on the genome name will redirect you to the *Genome Information* tool for the indicated genome. A **Help** link allows you to read more detailed information about HMM alignments and the *locus hmm alignment viewer*.

Next, the location of the hit on the model and the gene as well as the bit score and E value of the alignment are displayed in summary form. Clicking on the name of the model will redirect you to the source database's detail view of the indicated model.

Next, you will find a graphical representation of the alignment, showing for each position in the hit, the family's profile, the domain sequence and position, and the match score. For some *Pfam* families, there is structural or active site information. If available, this information is displayed above the alignment labeled as the **Secondary Structure** or **Active Site**, respectively.

Finally, you can view the raw results of the HMM alignment program (*hmmsearch*) by clicking on the **Raw hmmsearch output** link.

Locus PDB Alignment Viewer

The *locus pdb alignment viewer* allows you to view any alignment to a PDB reference gene in greater detail.

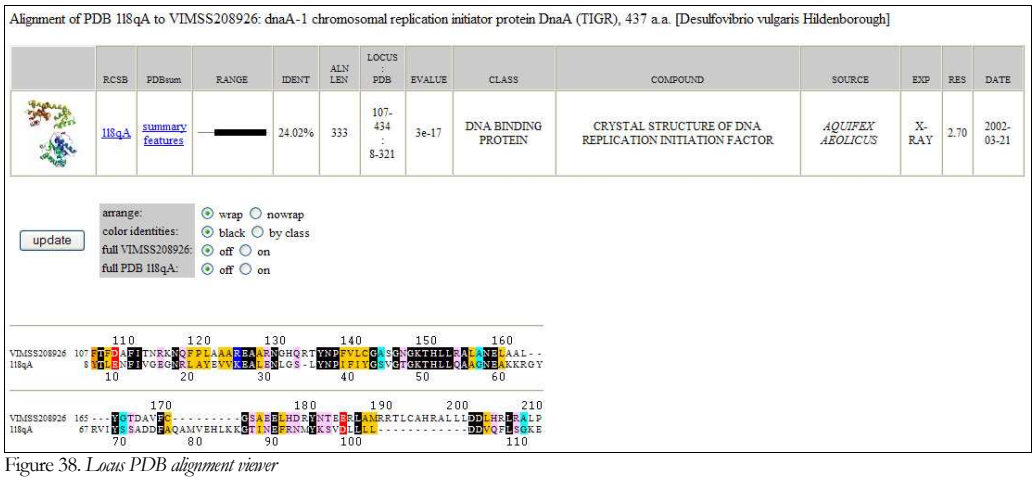


Figure 38. *Locus PDB alignment viewer*

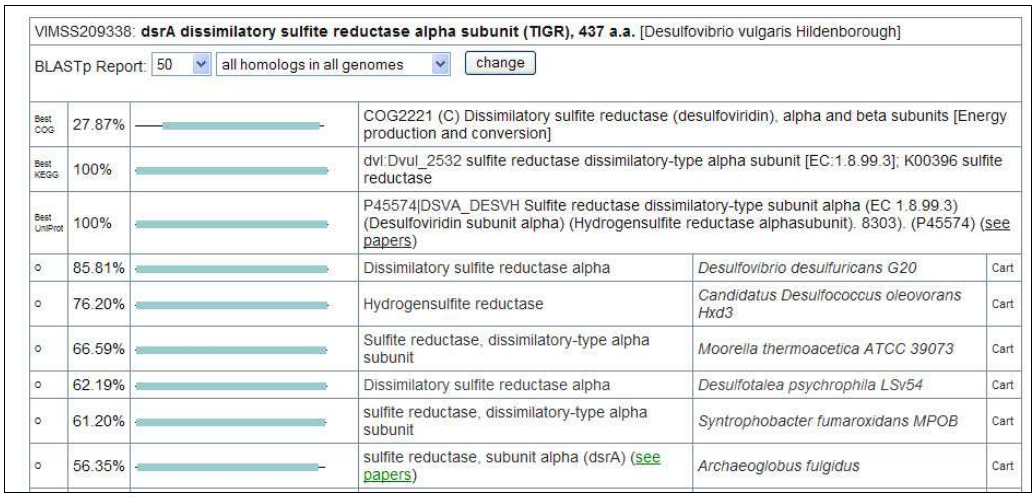
The name of the PDB reference gene and MicrobesOnline gene, length, and associated genome are displayed at the top of the viewer. The table that follows contains details of the PDB reference gene and alignment to the MicrobesOnline gene, including external hyperlinks to the PDB and PDBsum for additional information. The **RANGE** shows a graphical representation of the coverage of the hit with respect to the reference gene and the coordinates of the spans of the hit on the MicrobesOnline gene and PDB reference gene are shown in the **LOCUS..PDB** column. **IDENT** shows the %-identity of the hit and the length of the alignment can be found in the **ALN LEN** column. The **EVALUE** column shows the E-value of the alignment. The rest of the columns contain information from the reference database.

The next section contains formatting options for the alignment view, which follows. The **arrange** setting allows you to choose whether long alignments should be **wrap**ped to multiple lines, the default, or if they should be shown on a single line requiring horizontal scrolling. The **color identities** setting controls how matches should be colored in the alignment view. By default, identities within the alignment are shown in white with a black background with all other similar but non-identical matches colored according to their class. Using the **by class** option will color all identical and non-identical matches according to class. Setting the **full** setting to **on** will display the indicated sequence in the alignment in full. You must click on the **update** button to refresh the alignment view.

Homologs

The *homologs* tab shows the relationship of the selected gene to other genes through sequence similarity. Every gene in MicrobesOnline is compared to all other genes in

MicrobesOnline, as well as to the KEGG proteins, SwissProt/UniProt, COG, and NCBI viral sequences. Sequence similarity is determined using FastBLAST, which is designed to find the closest homologs quickly with little reduction in sensitivity.



VIMSS209338: **dsrA dissimilatory sulfite reductase alpha subunit (TIGR), 437 a.a.** [Desulfovibrio vulgaris Hildenborough]

BLASTp Report: 50 | all homologs in all genomes | change

Best COG	27.87%	COG2221 (C) Dissimilatory sulfite reductase (desulfoviridin), alpha and beta subunits [Energy production and conversion]		
Best KEGG	100%	dvl:Dvul_2532 sulfite reductase dissimilatory-type alpha subunit [EC:1.8.99.3]; K00396 sulfite reductase		
Best UniProt	100%	P45574|DSVA_DESVH Sulfite reductase dissimilatory-type subunit alpha (EC 1.8.99.3) (Desulfoviridin subunit alpha) (Hydrogensulfite reductase alphasubunit). 8303). (P45574) (see papers)		
o	85.81%	Dissimilatory sulfite reductase alpha	*Desulfovibrio desulfuricans G20*	Cart
o	76.20%	Hydrogensulfite reductase	*Candidatus Desulfococcus oleovorans Hxd3*	Cart
o	66.59%	Sulfite reductase, dissimilatory-type alpha subunit	*Moorella thermoacetica ATCC 39073*	Cart
o	62.19%	Dissimilatory sulfite reductase alpha	*Desulfotalea psychrophila LSv54*	Cart
o	61.20%	sulfite reductase, dissimilatory-type alpha subunit	*Syntrophobacter fumaroxidans MPOB*	Cart
o	56.35%	sulfite reductase, subunit alpha (dsrA) (see papers)	*Archaeoglobus fulgidus*	Cart

Figure 39. *Homologs* tab in the *Locus Information* tool

The report shows the selected gene on top, including its VIMSS id, gene name, description, length, and source genome. The initial report shows the best COG, KEGG, and UniProt hits, if any, along with up to 50 of the closest homologs in all MicrobesOnline genomes.

The best COG, KEGG, and UniProt hits, if any, are always displayed at the top of the report and indicated with a corresponding notation in the left-most column. These hits include the %-identity, a graphical summary showing the span of the hit with respect to the selected gene and a description of the COG, KEGG, and/or UniProt sequences including their external database id. KEGG protein hits provide a hyperlink from the protein id to the KEGG website that allows you to view more information about the specific KEGG protein. UniProt hits also provide a hyperlink from the UniProt id to the UniProt website, and in addition may include a link to NCBI PubMed to view related publications if referenced publications exist. This link is available by clicking on the **see papers** hyperlink. Homologs with one or more associated PDB reference genes will be shown with a **see PDBs** link. Clicking on this link will show a summary list of all associated PDB reference genes.

The remainder of the report shows up to 50 (by default) of the closest homologs to the selected gene in all other MicrobesOnline genomes, in descending order by %-identity. The left-most column is used to indicate the predicted relationship between the selected gene and the hit gene, where **o** designates a putative ortholog and **p** designates a paralog, a homolog in the same genome as the selected gene. If the value is empty the relationship is unknown but doesn't necessarily preclude the hit gene from being an ortholog.

Mousing over the %-identity value for each hit will show the hit's bit score and E value. Each hit also includes a graphical summary showing the span of the hit in relation to the selected gene. Mousing over the graphical summary will display the span of the hit in the selected gene and the hit gene and clicking on the graphical summary will open a new window allowing you to view the selected alignment at the sequence level, as shown in the following figure.

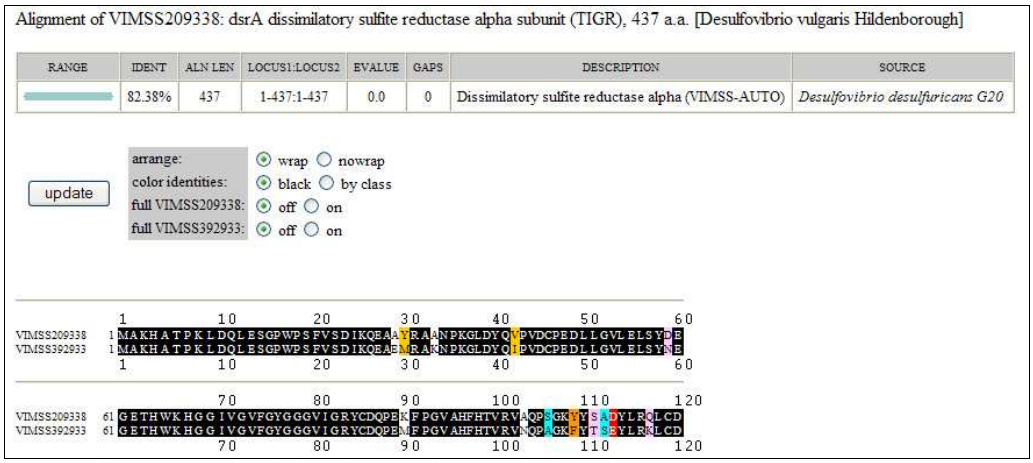


Figure 40. Alignment viewer from the homologs tab in the *Locus Information* tool

To the right of graphical summary, each hit also includes the hit gene's description, which is also a hyperlink to view that gene in the *Locus Information* tool. Additionally, if there are associated paper references, a link is provided to the right of the hit gene's description, **see papers**, which will redirect you to NCBI PubMed. To the right of the hit gene's description, the hit gene's source genome is displayed and linked to the *Genome Information* tool. Finally, there is a **Cart** link that will add the selected hit gene to your *session gene cart*.

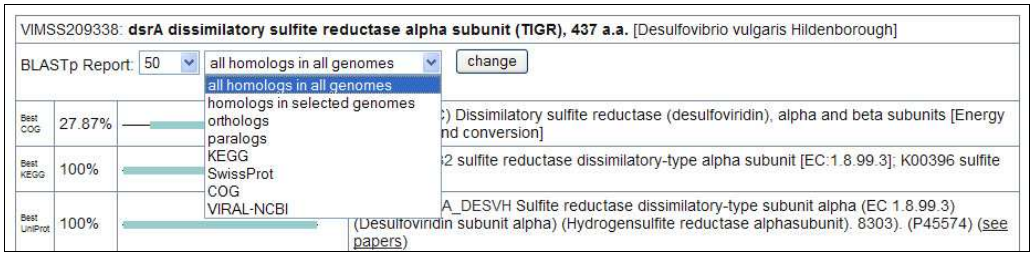


Figure 41. Homolog display options

In addition to being able to change the maximum number of displayed homologs in the report, you can also select which subset of homologs to display by selecting the appropriate option from the drop-down selection box located immediately to the left of the **change** button. By default **all homologs in all genomes** is selected.

To restrict the report to homologs within the genomes selected in the *Genome Selector*, change the drop-down option to **homologs in selected genomes**, then click **change**. Selecting **orthologs** will restrict the report to predicted orthologs of the selected gene and similarly, selecting **paralogs** will restrict the report to paralogs of the selected gene.

If you wish to view only hits to the KEGG proteins, select the **KEGG** option from the drop-down selection box. Similarly, you can strict the report to only genes in SwissProt/UniProt by selecting the **SwissProt** option. Selecting the **COG** option will restrict the report to only COG proteins. Finally, if you wish to view homology only to NCBI viral sequences, select the **VIRAL-NCBI** option.

After changing the homology type selection or the maximum number of hits to report, you must click the **change** button to apply the changes to the report.

Sequences

The *sequences* tab contains various sequences related to the selected gene. As with the other tabs, the selected gene is prominently displayed at the top of the tab including its VIMSS id, name, description, length, and source genome.

Four sequences are available for each gene—its **Protein** sequence and **Nucleotide** sequence, as well as up to 250 nucleotides **Upstream** and **Downstream** of the the gene transcript. Sequences are displayed in FASTA format wrapped to 60 letters per line.

Add Annotation

The *add annotation* tab is only accessible to registered users. For more information on registering, please see the *Register* section on page 9. This tab allows you to change certain annotation values for the selected gene, including its **Name**, **Description**, **EC number** assignments, and **Gene Ontology** assignments. In addition, it allows you to record a **Comment**, which will be logged with your annotation changes.

In order to assist you with making changes to the annotation, most of the basic information displayed on the *gene info* tab is also displayed as read-only fields on the *add annotation* table. This includes the selected gene's VIMSS id, organism, synonyms, position, COG annotation, if any, InterPro annotation, if any, as well as the current annotation history.

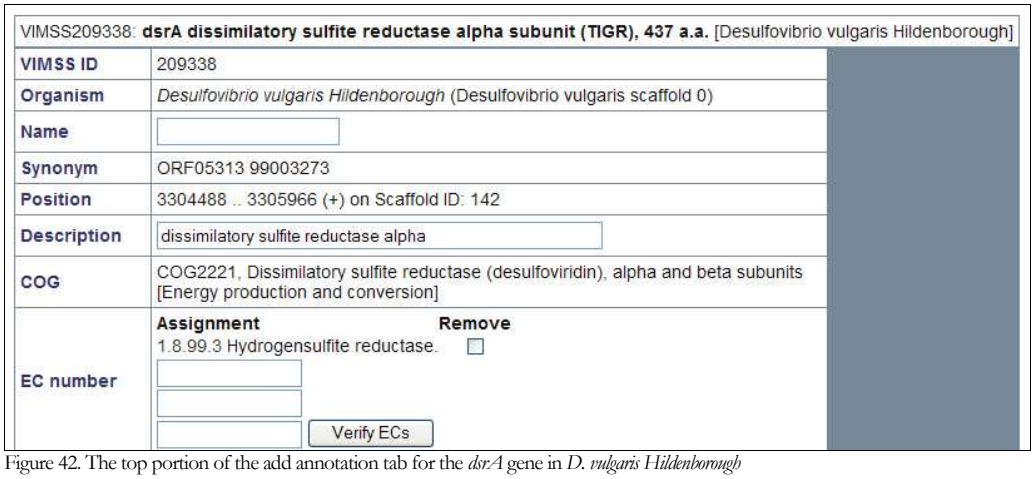


Figure 42. The top portion of the add annotation tab for the *dsrA* gene in *D. vulgaris Hildenborough*

Tree Browser

The *Tree Browser* is a tool allows you to view gene trees of a selected gene of interest within the context of its gene neighborhood or species tree, depending on the selected view.

The *Tree Browser* has two views—the *gene context view* is similar to the view described above. It allows you to view the selected gene of interest along with its closest homologs in the context of each genes' neighborhood. The *species tree view* allows you to compare the gene tree of the selected gene of interest with the species tree. Most entry points into the *Tree Browser* will link to the *gene context view*.

Gene Tree Options

Regardless of the selected view, there are common tree clustering and rendering options which affect the gene tree that is displayed. These options are always located above the rendered output.

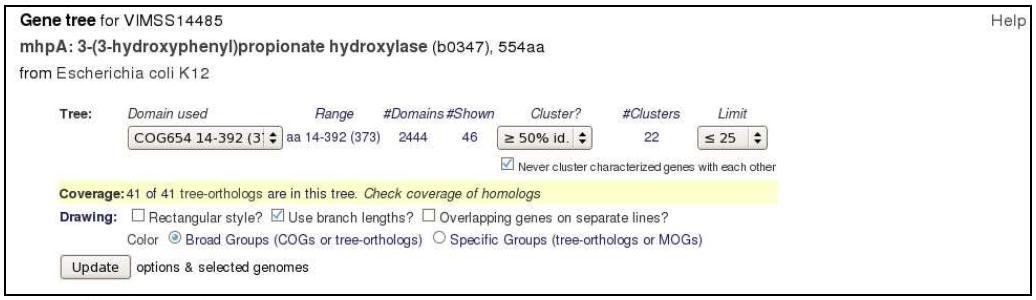


Figure 43. *Tree Browser* options

First, the *Tree Browser* will identify the gene of interest, showing its VIMSS id, gene name, gene description, length, and source genome. A **Help** link located in the upper-right corner of the display can provide online assistance for the *Tree Browser*. If you are logged in, a **Related Carts** link will be displayed to the left of the **Help** link, which will allow you to view all carts containing the gene whose tree you are viewing, along with all associated jobs.

Tree Options

Gene trees are computed using conserved domains or families from the COG, Pfam, TIGRfam, PIRSF, SMART, SuperFam, and Gene3D families, as well as by aligning genes that do not belong to these families with FastBLAST. By default, the *Tree Browser* will select a tree that has many aligned positions and contains as many tree-orthologs as possible. However you may choose the tree with the **Domain used** drop-down select box.

The range of the selected domain or gene family hit is displayed as a starting and ending position relative to the selected gene of interest and also includes the length of

the hit in parentheses. The **#Domains** and **#Shown** counts indicate how many total domains are included in this gene tree versus the total number of domains shown given the gene tree options specified.

The **Cluster?** Option allows you to set the %-identity threshold required to cluster similar genes together, or to disable clustering. By default, genes that are **≥ 50% id** are clustered together in the gene tree. The **Never cluster characterized genes with each other** checkbox controls whether characterized genes can be clustered with other characterized genes. By default, a characterized gene can be clustered with other uncharacterized genes but never with another characterized gene.

The **#Clusters** count shows the total number of gene clusters rendered in the gene tree. To change the limit on the maximum number of gene clusters, change the value in the **Limit** drop-down select box to reflect the new desired value.

Coverage

The *Tree Browser* lets you verify whether or not all of the close homologs to the selected gene of interest are represented in the gene tree. By default, it reports whether all of the tree-based orthologs are in this tree. If you request, it will also check whether all of the top FastBLAST homologs are in this tree. To use this feature, simply click on the **Check coverage of homologs** link. This option remains active even in subsequent gene tree updates unless disabled.

Drawing

These options control the actual rendering of the gene tree and in the case of the *gene context view*, the method used to color orthologs. By default, the gene tree is rendered with straight lines connecting the various internal nodes however you can change this option by checking the **Rectangular style** checkbox. Similarly, the default gene tree is rendered with branch lengths, that is rendered branch lengths are proportional to computed branch lengths. To disable this feature, simply uncheck the **Use branch lengths** checkbox.

To make the browser within the *gene context view* more concise, all genes within the gene context of a single genome are rendered on a single line. This can make it difficult to see highly overlapping genes even though they should be rare. To render overlapping genes on different lines, simply check the **Overlapping genes on separate lines** checkbox.

Finally, you can change the default ortholog coloring method, which relies on the COGs or orthologs. This is fast, but distantly related genes may belong to the same COG; also, genes that are not in the top track will probably not be colored unless they belong to COGs. You can get more accurate results by using the narrow coloring (using orthologs for all genes).

Updating

After changing any of the tree browser options, you must click the **Update** button to refresh the display to include the new options.

Gene Context View

The *gene context view* shows a gene tree along with a genome browser. The selected gene is at the top. The tree shows that gene along with its close neighbors, which could be from the same genome, or even from the same gene if there has been a domain duplication. For each of these homologs, the track in the Gene Context section is centered on that homolog, and the homolog's name is shown in green text.

One feature that makes the *Tree Browser* particularly useful is the ability to cluster very similar genes into subtrees, thus reducing the clutter of the overall display. This allows you to quickly identify genes or clusters of genes that are likely to share the same function as the selected gene of interest.

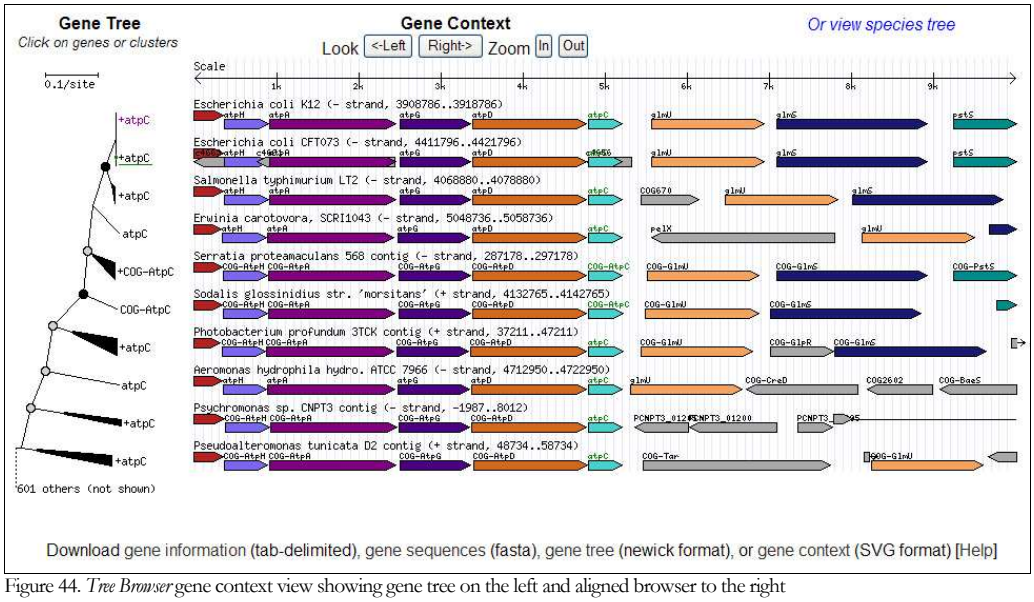


Figure 44. *Tree Browser* gene context view showing gene tree on the left and aligned browser to the right

The *gene context view* browser provides basic navigation through the four buttons located under the **Gene Context** title, labeled **<-Left**, **Right->**, **In**, and **Out**. Clicking on the **<-Left** or **Right->** buttons will shift the current view by 33% of the viewable range in the direction selected. Clicking on the **In** or **Out** buttons will change the viewable range by 1.5x in the direction selected.

To switch to the *species tree view*, click on the **Or view species tree** link located in the upper-right corner of the *gene context view*.

Mousing over a gene feature or gene cluster will show a representative name and description of the gene feature or cluster. For gene features in the browser, the gene's VIMSS id and source are also displayed. Mousing over internal nodes in the gene tree will show the bootstrap value of the selected node.

Clicking on gene features or gene clusters is controlled by a context-sensitive menu. For genes in the browser portion of the *gene context view*, you can choose to access the

Locus Information tool for the selected gene in the current window using the **View gene(s)** menu option or in a new window using the **as popup** menu option. You can redraw the *Tree Browser* using a new gene of interest by selecting the **Recenter** menu option and finally you can add the selected gene to your *session gene cart* by selecting the **Add to cart** menu option.

Clicking on single-gene features in the gene tree will yield the same context-sensitive menu as with the browser portion, as described in the previous paragraph. Clicking on a gene cluster yields a menu containing all of the same options, with the addition of an **Expand** option, which will expand the cluster. Once a cluster is expanded and the view has refreshed, the cluster's root will show a red **–** sign which you can click to collapse the cluster to its original state. Genes or gene clusters that are not shown in the current display are indicated by the dashed line at the bottom of the gene tree showing the total number of excluded genes. If the **Use branch lengths** option is enabled, a scale is shown at the top of the gene tree to give the scale of the rendered branches.

At the very bottom of the *gene context view*, four links allow you to download datasets related to the *gene context view*. The **gene information** link will allow you to download a tab-delimited file containing all of the representative genes in the current view, including the coordinates of the domain or gene family hit, the name of the gene, description, and source genome with its taxonomy id.

The **gene sequences** link will allow you to download all of the representative genes in the current view in FASTA format.

You can also download the current gene tree in Newick format by clicking on the **gene tree** link. Finally, you can download the entire *gene context view* in scalable vector graphics (SVG) format for inclusion in a paper or other publication.

Species Tree View

The *species tree view* allows you to compare the gene tree side by side with the species tree. The generated species tree will be shown with the genome of the selected gene of interest at the top aligned with the gene of interest in the gene tree.

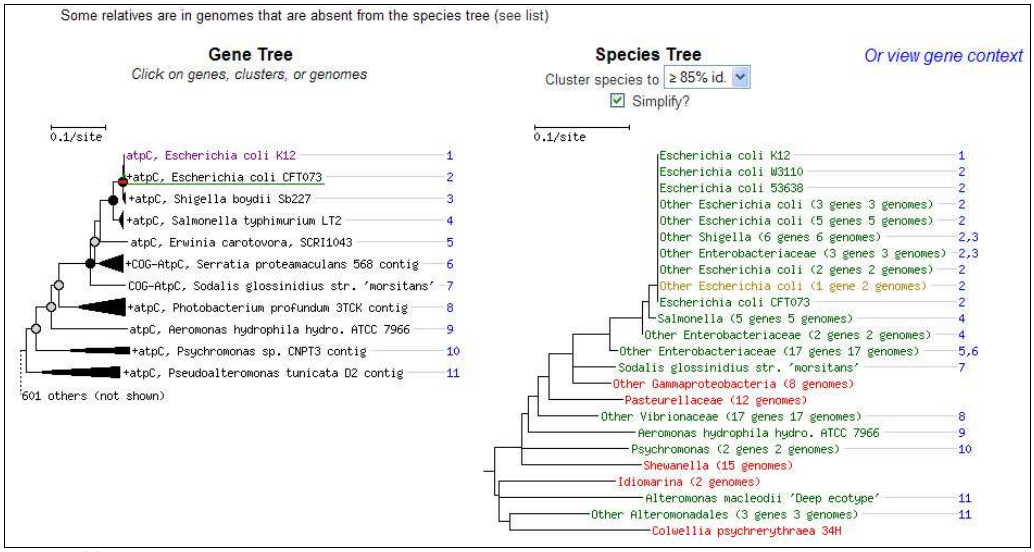


Figure 45. The *species tree view* in the *Tree Browser*

Genomes that contain one or more genes from the shown gene tree will be shown in green in the species tree. Related genomes that do not contain any of the genes in the shown gene tree will be shown as red in the species tree. A genome may be shown in red even if it contains genes that are in the gene tree, because they are too distant to be in the shown portion of the tree.

By default, closely related genomes are condensed into groups to make the species tree more compact and comprehensible. Genome groups are indicated in the species tree with the number of genes and genomes that comprise the genome group. Mousing over a genome group's name will show the group's member genomes. Clicking on a genome group's name will redirect you to the *Genome Information summary view* for all the genomes represented by the selected group. A genome group will be shown in yellow in the species tree if at least one but not all of the member genomes contain one or more genes in the shown portion of the gene tree.

To help show you whether or not a gene tree is similar to its corresponding genome species tree, all of the genes or gene clusters in the shown gene tree are labeled with blue numbers. A number is shown to the right of a genome or genome group if that genome or genome group contains the corresponding gene from the gene tree.

Clicking on the labels in the gene tree will show the context-sensitive menu described in the *gene context view*, above.

Tutorial: Examining the Evolutionary History of a Gene

The tree-browser shows you the evolutionary history of a gene and what genes are nearby. This can help you spot questionable annotations. It can also highlight genes that are transferred together between genomes, which usually reflects a functional relationship.

For example, see the screenshot below of the tree-browser for purR from *E. coli* K12 (VIMSS15779). We have highlighted key features with red circles and boxes. In the gene tree, the genes from K12 are in **magenta** and you can see that K12 contains both purR and a paralog rbsR. Both purR and rbsR have been characterized, and the orthologs from the O157:H7 EDL933 strains are linked to papers (see green underlines). You would click on those genes to see the papers and confirm that purR regulates purine synthesis and that rbsR regulates the uptake and degradation of ribose.

You can also see that rbsR is near rbsA, rsbC, rbsD, and rbsK, while purR is near ydhP or COG-AraJ. This suggests that both rbsR and purR's function has been conserved, and you can split the tree into the purR genes (in *E. coli* out to *Haemophilus*) and the rbsR genes (in *E. coli* out to *Aeromonas*). Although a function for these genes can be confidently assigned, many of them just have the generic COG-PurR label. You can also see that more distant relatives, at the bottom of the tree, are also near ribose genes. If you hover on the genome name for track at the bottom, it shows you the lineage of that organism is in the Betaproteobacteria. (*E. coli* is in the Gammaproteobacteria, a distant division.) This suggests that rbsR might have been transferred between these organisms.

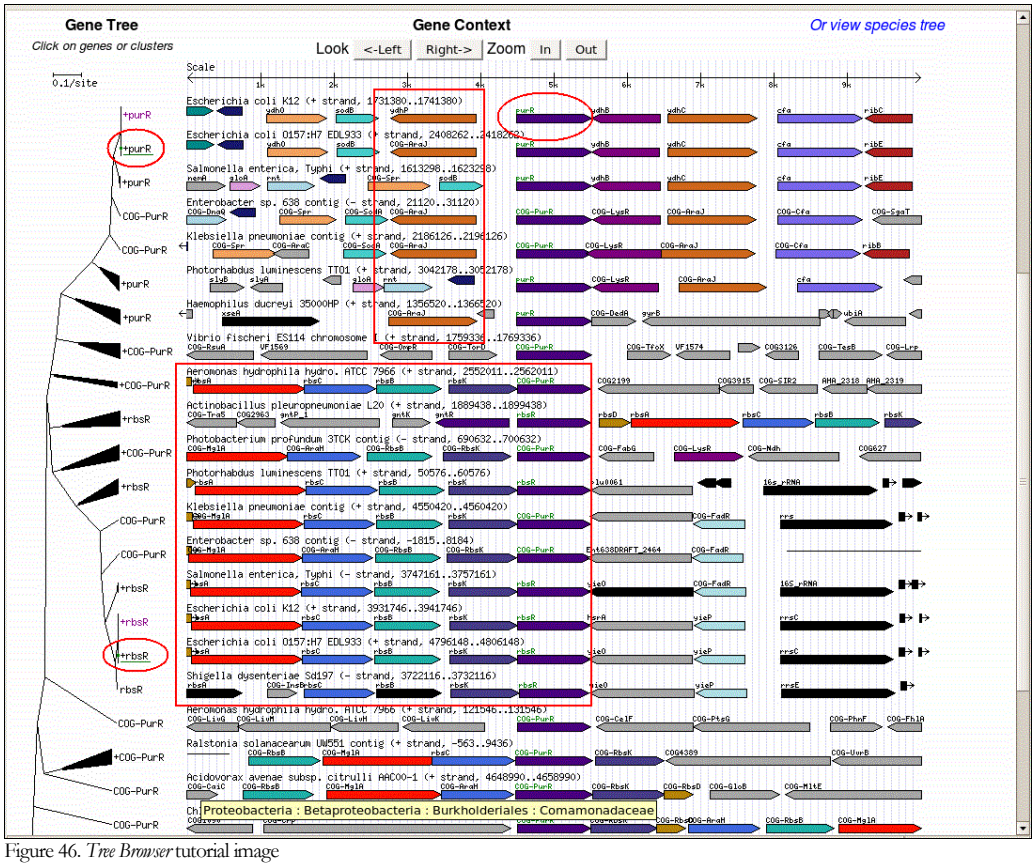


Figure 46. *Tree Browser* tutorial image

If you use the tree browser to compare the gene tree to the species tree, you can see that rbsR and purR resulted from a gene duplication within the gamma-Proteobacteria, and that their common ancestor was acquired by horizontal gene transfer (HGT). In the figure, we've highlighted the genes that were annotated as purR and rbsR above. You can see that most relatives of *E. coli* contain both genes. A likely point for the duplication is highlighted. Because these genes are absent from more distant relatives (e.g., from most Shewanella and other Alteromonadales), it appears that the common ancestor of rbsR and purR was acquired by HGT. Furthermore, in the gene context view, you can see that these HGT relatives (e.g. in Ralstonia) also have ribose genes nearby. This shows that rbsR and the ribose genes have been transferred together, and strongly suggests that these distant homologs have the function of rbsR, and probably not that of purR.

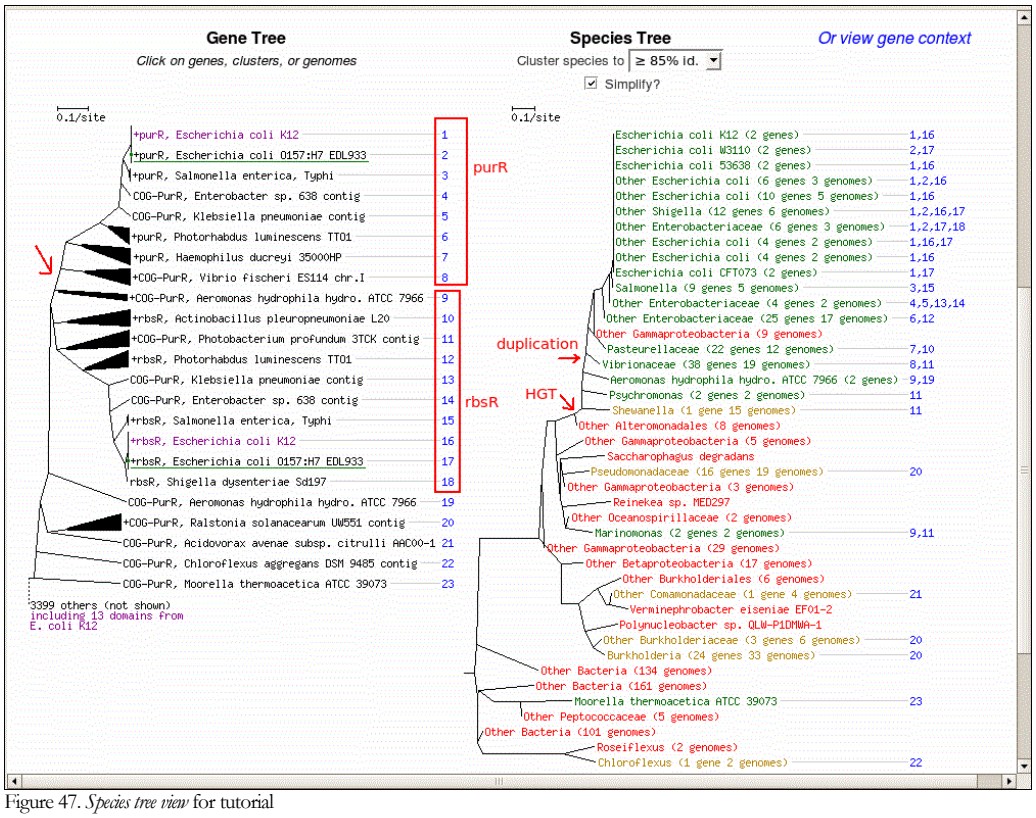


Figure 47. *Species tree view* for tutorial

Finally, once you use the tree-browser to get an initial impression, MicrobesOnline includes several tools that you should use to verify that you are correct:

1. Verify that homologs are present in the tree. It is possible that the gene tree is misleading because the close homologs of the gene of interest are missing from the tree. Use this feature to check if that is the case. If there is a problem with this tree, try selecting a different one, or use the homologs page and the add cart feature to build your own.

1. Build your own tree. MicrobesOnline has a huge number of large trees, and so we cannot use the most accurate methods to build these trees. You can build a more accurate tree by adding all relevant homologs to a gene cart. (Click on the gene tree or the gene context view to do this.) We also recommend using a more stringent clustering threshold in the tree browser (at least 70%). Then, save the cart, build a multiple sequence alignment, and build a tree. You can view your new tree in a tree-browser-like view as well (click on "View this tree with gene neighborhood context" at the bottom).

Ortholog Browser

The *Ortholog Browser* allows you to view a gene of interest within the context of its gene neighborhood aligned with orthologs of the gene of interest in other selected genomes. The genes in the browser display are colored such that genes predicted to be orthologous or those within the same COG cluster are assigned the same color. This allows you to quickly see whether a gene and surrounding genes are conserved and whether their ordering is also conserved in other genomes. Genes that do not have an ortholog shown are displayed in grey.

Browser Display

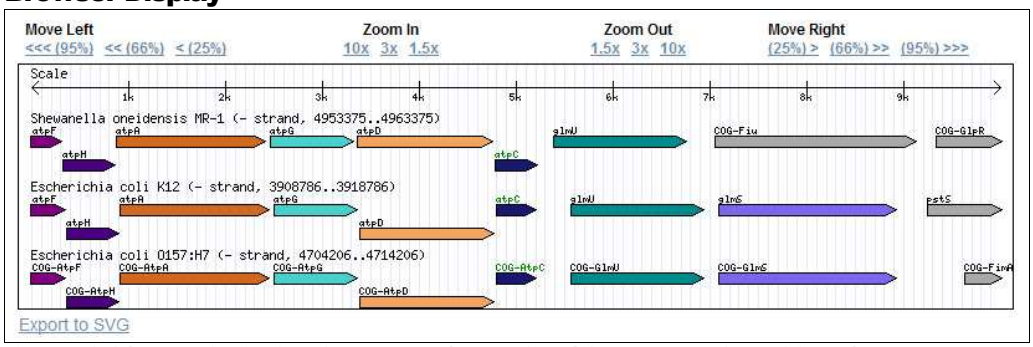


Figure 48. *Ortholog Browser* showing *S. oneidensis MR-1*, *E. coli K-12*, and *E. coli O-157:H7* centered around the *atpC* gene

The browser display shows one *track* per selected genome and the ordering of genome tracks is determined by the order in which the genomes are selected in the browser options. The view is always centered on the selected gene of interest, which in the case of the example is the *atpC* gene in *E. coli K-12* and its orthologs in the selected genomes. The current region of the genome being displayed is shown to the right of the genome name in the track label, including the strand.

The selected gene of interest and its orthologs in other genomes are always aligned with their names highlighted in green. The surrounding genes are then rendered according to their relative spacing on each genome and labeled with their gene name. In some cases, there is no assigned gene name so the COG cluster name is used instead. This is indicated by the **COG-** prefix in the gene name.

Clicking on a gene feature will bring up a menu. You can then navigate to the *Locus Information* tool for the selected gene, use the selected gene as the new gene of interest

(**Recenter on feature**) or to add the selected gene to your *session gene cart* (**Save as feature of interest**). Mousing over a gene feature in the browser will display an extended description of the gene along with its VIMSS id and source.

The browser's navigation toolbar is located above the browser display. You can shift the view without affecting the range being displayed or you can zoom in or out, changing the viewable range. Each navigation option is labeled with the relative amount of the change and each amount is relative to the current viewable range. For example, the **<<< (95%)** link will shift the view to the left of the current view by 95% of the current viewable range. If the viewable range is 10,000 bp, the view will be shifted 9,500 bp to the left in each genome, although the view is still anchored in each genome by the selected gene of interest and its orthologs. Clicking on the **1.5x** link under **Zoom In** will reduce the viewable range from 10,000 bp to approximately 6,667 bp, for example. When zooming in or out, the view always remains centered on the selected gene of interest and its orthologs.

Finally, if you wish to export the current browser view as a Scalable Vector Graphics (SVG) file for inclusion in a paper or other publication, you can click on the **Export to SVG** link located near the bottom-left of the browser display.

Browser Options

To change which genomes are displayed in the browser, select the desired genomes from the **Select Organism(s) to Display** selection box. The genomes that are currently being displayed are always shown at the top with the remainder of genomes shown below, sorted alphabetically. Only individual genomes may be selected. To select multiple genomes, hold the `Ctrl` key on a PC or the `Shift` key on a Mac and use your mouse to select the desired genomes. You must click on the **Update Display** button to refresh the browser display. The genome of the selected gene of interest is always displayed, even if it is not explicitly added.

> ⚠ **WARNING**
>
> The colors assigned to orthologs or COG cluster members are not necessarily consistent from view to view because the colors are assigned dynamically based on the current view.

Coloring

By default, genes are colored by relatively broad groups (mostly by COG). You can also color more narrow groups (by tree orthologs for genes in the anchor track and by MicrobesOnline Ortholog Groups for other genes). This is can be useful if you are viewing many genomes and aren't sure if two genes that belong to the same COG are actually closely related.

Expression Data Viewer

The *Expression Data Viewer* tool is the general microarray expression experiment viewer and profile search tool. At present, MicrobesOnline contains microarray experiment data for several genomes including *E. coli K-12*, *D. vulgaris Hildenborough*, and *S. oneidensis MR-1*. Due to the limited amount of microarray experiment data for some genomes, profile searches are only available for experiments in *E. coli K-12*, *D. vulgaris Hildenborough*, and *S. oneidensis MR-1*.

The *Expression Data Viewer* has several components. The first is an *experiment browser*, which will allow you to search for specific experiments in selected genomes and/or involving the selected experimental conditions. You can also search for experiments genes using their VIMSS id or by our internal microarray experiment ids. The second component is an *expression experiment viewer* that provides a detailed report for each microarray experiment, including a list of up-regulated and down-regulated genes, plots, and experimental conditions. For gene and operon-centric views, the *Expression Data Viewer* includes a *gene expression viewer*. Finally, for some datasets, you can perform gene expression profile searches using the *profile search tool*.

Each of these components is described in greater detail in the subsequent sections.

> ⚠ **WARNING**
>
> Some microarray expression heatmap images are extremely large and may take a long time to download and render in your browser. Your browser may appear to be locked up while rendering large images.

Experiment Browser

The *experiment browser* allows you to search through the MicrobesOnline microarray data compendium for specific experiments by selecting the desired organism and experimental condition, or by specifying a list of MicrobesOnline gene ids (VIMSS ids) and microarray experiment ids. The former search is useful in identifying the available experiments for the selected organism or organisms while the latter search is useful in identifying experiments involving specific genes of interest. The *experiment browser* is available by clicking on the **Microarray** link from the MicrobesOnline home page.

Browsing By Organism or Experimental Condition

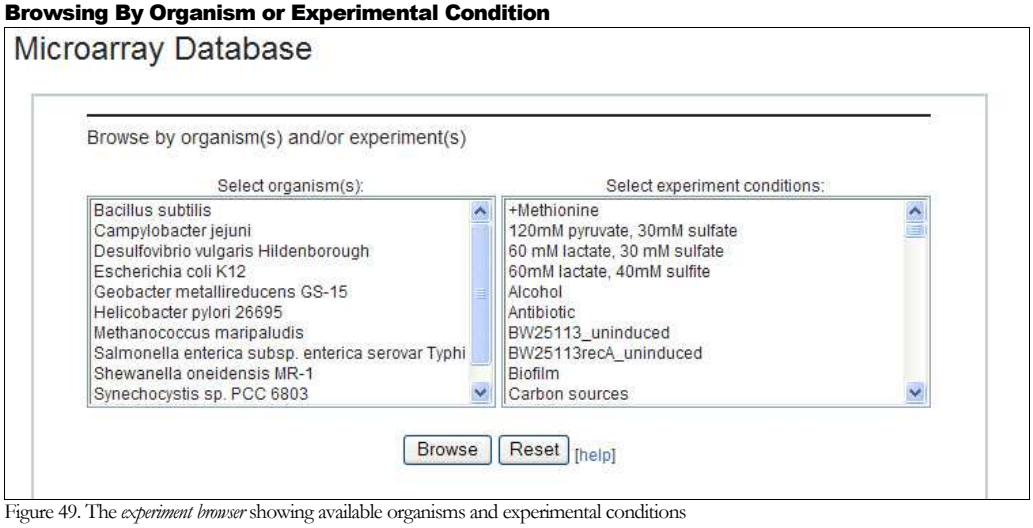


Figure 49. The *experiment browser* showing available organisms and experimental conditions

To search for microarray experiment data by organism and/or experimental condition, simply select the desired organism or organisms from the **Select organism(s)** selection box and the desired experimental condition or conditions from the **Select experiment conditions** selection box. To view the list of matching experiments, click on the **Browse** button or clear the selections in both boxes click on the **Reset** button. To browse all available experiments, simply click the **Browse** button without selecting an organism or experimental condition.

Specifying both an organism or organisms and an experimental condition or conditions requires both sets of criteria to be true in the listed experiments. If the result set is empty you will be returned to the *experiment browser* with an error message stating *"No public data found in VIMSS Gene Expression Database."*

If at least one experiment matches the specified criteria, the experimental result set matching the specified criteria will be displayed. One experiment is displayed per row and includes the experiment's internal id, organism, stress, basic experiment protocol information, referenced publications, and the available data series. Each experiment has at least one report set in its data series, however most have two or more.

A report set represents the measurement at a specific time or level of stress in the experiment with multiple report sets representing the change in expression over time and/or level of stress. Each report set is represented by its specific treatment (time point and stress level) and the control used as the basis for that report set.

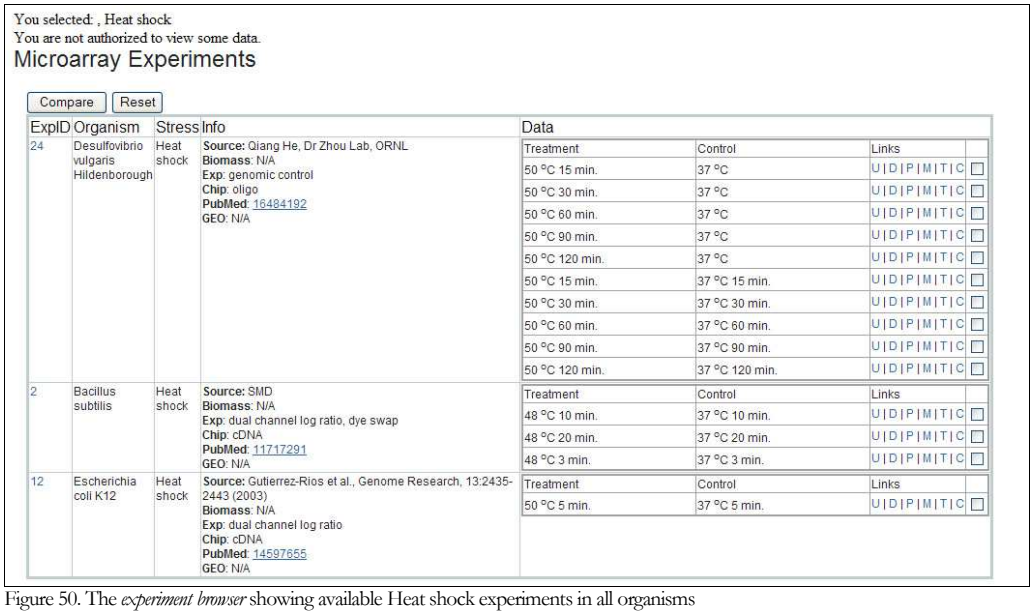


Figure 50. The *experiment browser* showing available Heat shock experiments in all organisms

If the result set includes one or more experiments matching the selected criteria to which you do not have access, you will see a warning displayed near the top of the *experiment browser* that says *"You are not authorized to view some data."*

Each experiment and data series within an experiment contains a number of hyperlinks. Clicking on the experiment id (**ExpID**) of an experiment will show only the data available for that experiment. If an experiment has one or more referenced publications, each is listed in the **Info** column under the **PubMed** heading using the publications PubMed id. Each PubMed id is a hyperlink to NCBI PubMed to view the abstract of the referenced publication.

In the **Data** column, all available report sets within the data series are shown for each experiment, one per row, including the specific **Treatment** and **Control** descriptions. Each report set also includes a set of lettered links that allow you to view more details about the specific report set in the *expression experiment viewer*. This information is organized in tabs much like the *Locus Information* tool and each lettered link corresponds to a specific tab in the display. The **U** link shows the list of up-regulated genes in the report set while the **D** link shows the list of down-regulated genes. Both of these lists allow you to view the gene sets within the context of their corresponding operons as well as in their corresponding KEGG metabolic pathways. In the future, you'll also be able to view the gene sets within the context of their Gene Ontology assignments.

The **P** link will show all related plots for the selected report set and the **M** link will take you to the integrated KEGG metabolic pathway browser. This pathway browser is identical to the MicrobesOnline *Pathway Browser* except that gene expression levels are displayed in the context of metabolic maps instead of presence and absence of genes. The **T** link will display the report set's TIGR roles, if any, and the **C** link will display the report set's COG roles.

You can also mouseover the lettered links to obtain a brief description of the information that is shown, or click any of the lettered links and navigate to the desired data by using the *expression experiment viewer*'s internal tabs.

The checkbox located next to each report set in an experiment's data series allows you to compare the selected experiments for correlation. To do this, select the experiments you wish to compare then click on the **Compare** button located above the list of experiments. You must select at least two experiments and selecting three or more will result in a pair-wise comparison of all selected experiments.

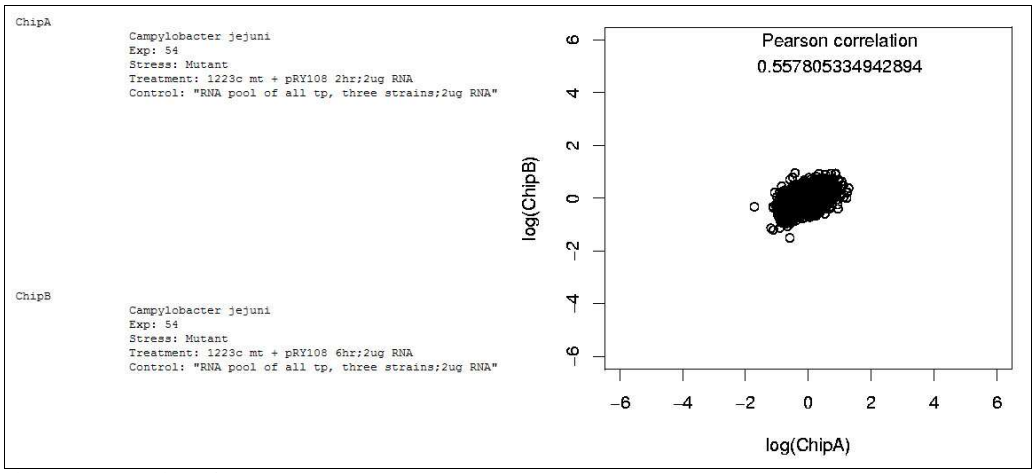


Figure 51. A comparison of two report sets

You can clear all selected experiments by clicking on the **Reset** button also located above the list of experiments, to the right of the **Compare** button.

Browsing By VIMSS Id or Experiment Id

Another method to locate experiments of interest is to search for the inclusion of specific genes of interest. The bottom half of the *experiment browser* allows you to accomplish this. You can also limit the search to specific experiments by specifying a list of experiment ids. Specifying one or more experiment ids without specifying at least one gene of interest will show all matching experiments using the same display as with browsing experiment data by organism or experimental condition, as described in the previous section. Specifying one or more genes of interest will return all experiments containing at least one gene from the specified list, while specifying one or more genes of interest with one or more experiment ids will return a list of all experiments containing the gene of interest in the set of specified experiments.

The displayed result set when specifying a set of genes of interest is slightly different than simply browsing experiments as in the previous section. Each experiment containing one or more of the specified genes of interest is displayed in the result set.

ExpID	Organism	Stress	Info
12	Escherichia coli K12	Heat shock	**Source:** Gutierrez-Rios et al., Genome Research, 13:2435-2443 (2003) **Biomass:** N/A **Exp:** dual channel log ratio **Chip:** cDNA **PubMed:** 14597655 **GEO:** N/A

Gene Name	Links	Description	50 °C 5 min. vs. 37°C 5 min.
purR	I | GP | O | OP	DNA-binding transcriptional repressor, hypoxanthine-binding	logR=-0.17 z=-1.00 n=NA

40	Escherichia coli K12	pH	**Source:** Blattner Lab **Biomass:** N/A **Exp:** Affymetrix **Chip:** Affymetrix **PubMed:** N/A **GEO:** N/A

Gene Name	Links	Description	pH2 10 min. vs. wild type 10 min.
purR	I | GP | O | OP	DNA-binding transcriptional repressor, hypoxanthine-binding	logR=-0.21 z=-1.00 n=NA

Figure 52. Browsing all experiments containing the *pur*R gene in *E. coli K-12*

The first four columns of the display are identical to the display used when browsing experiments. The first column contains the experiment's internal id (**ExpID**), the second column contains the organism name, the third contains the stress description and the fourth contains basic experiment protocol information and includes a list of referenced publications, if any.

Instead of displaying the data series for each experiment however, each gene from the specified genes of interest that are contained within an experiment is displayed in a sub-table with one gene per row. Each row in this sub-table shows the gene's name, which is a hyperlink to the *Locus Information* tool, a set of lettered links, the gene's description and the gene's expression level in each report set in the data series. Each report set is displayed as an additional column in the sub-table. The example above shows two experiments each of which only has one report set. Clicking on a report set headings in the sub-table will redirect you to the *expression experiment viewer* for the selected report set.

The **I** link also links to the *Locus Information* tool for the indicated gene. Clicking on the **GP** link will redirect you to the *profile search tool* for the indicated gene while clicking on the **OP** link will redirect you to the *profile search tool* for the operon containing the indicated gene. The **O** link will show the expression levels for all genes in the operon containing the indicated gene.

Expression Experiment Viewer

Each experiment often contains multiple report set, where each report set represents the expression level at a given time point and/or level of stress. Multiple report sets record and allow you to view the change in expression over time or over varying levels of stress. The *expression experiment viewer* allows you to view the details on individual report sets.

The information is organized into seven different tabs, excluding a **Home** tab which will simply redirect you to the *experiment browser*. Each tab is described in more detail in the following sections.

Figure 53. Header of the *expression experiment viewer* for an *E. coli K-12* heat shock experiment report set

Every tab in the *expression experiment viewer* contains the same basic header and navigation options. This header shows the organism name, **Stress** condition, including the treatment description and control description, the internal experiment id (**VIMSS Experiment ID**), and publication reference (**Source**). The **Experiment condition** drop-down selection box shows a list of all available report sets for the selected experiment with the current selected report set as the default selected option. You can quickly navigate to other report sets by selecting the desired report set stress condition and clicking on the **submit** button. The **Color by minimum absolute Z-score** drop-down selection box allows you to specify the Z-score significance cut-off when viewing gene expression in the KEGG metabolic pathways tab. To update this option, select a new cut-off and click on **submit**.

The description of each tab is contained within the tab itself and is the hyperlink for selecting a particular tab to view. By default, unless you selected to view a specific tab, the **Up-regulated** gene list tab will be displayed. Each tab is described in more detail below.

Up-regulated Tab

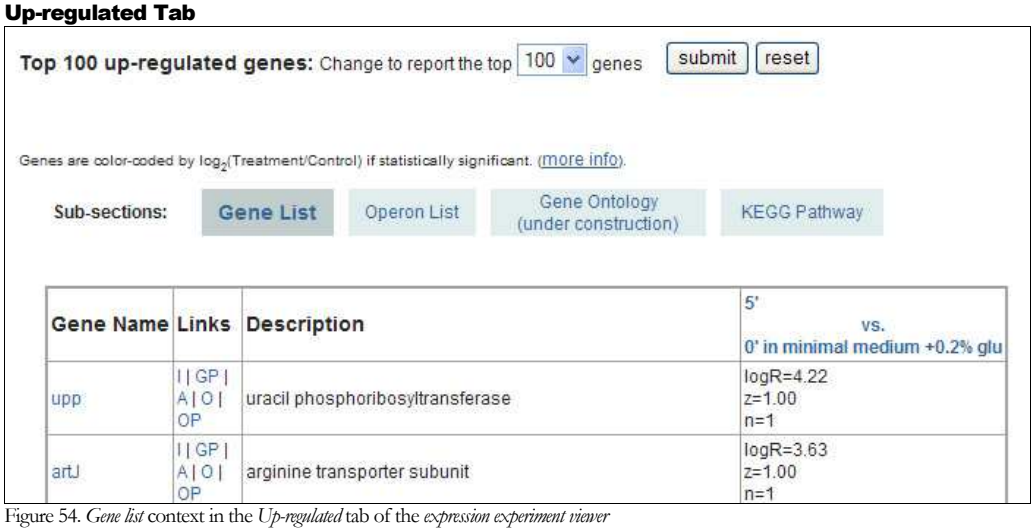


Figure 54. *Gene list* context in the *Up-regulated* tab of the *expression experiment viewer*

The *up-regulated tab* contains a list of genes that were up-regulated under the stress conditions of the selected report set. This list can be viewer in a variety of ways, which

are available from the links in the **Sub-sections** menu. The *gene list* simply shows a list of up to 100 (by default) of the top up-regulated genes displayed in descending order by $\log_2$-ratio (**logR**) value. You can change the maximum number of top up-regulated genes to include by selecting from the available values in the **Change to report the top** drop-down selection box then clicking on the **submit** button.

Each gene in the *gene list* view includes its name, a set of lettered links, its description, and its expression level and significance as measured by its $\log_2$-ratio (**logR**) value and Z-score (**z**). The gene name and **I** link will redirect you to the *Locus Information* tool for the indicated gene. The **GP** and **OP** links will redirect you to the *profile search tool* in the *Expression Data Viewer* for the gene's expression profile and gene's operon's expression profile, respectively. The **A** link will redirect you to the *experiment browser* showing all experiments containing the indicated gene and the **O** link will show the expression levels for each gene in the indicated gene's operon across all report sets. You can mouseover the lettered links to view a brief description of where you will be redirected if you click the corresponding link.

You can view the top up-regulated genes within the context of their operons by clicking on the **Operon List** link under **Sub-sections**. Each displayed operon contains one or more of the top up-regulated genes and includes a **details** link that will show you the expression levels for all genes in the selected operon for the selected experiment across all report sets. The **website** link will redirect you to the *Locus Information* tool's *operon & regulon* tab. The operon structure is shown including the names of all component genes. Mousing over the gene names will show the gene description and individual expression levels for the selected report set.

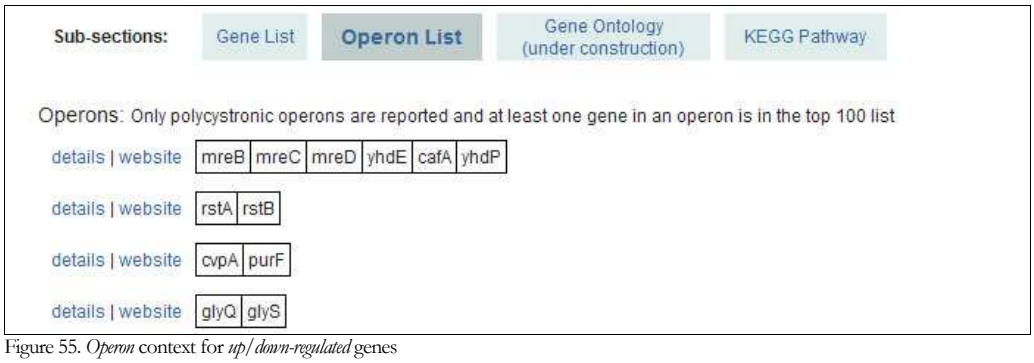


Figure 55. *Operon* context for *up/down-regulated* genes

The *gene ontology* view is currently under construction and no data is displayed.

The *KEGG pathway* view will show the top up-regulated genes according to their presence on the KEGG metabolic pathway maps. The pathway maps are shown in descending order by top up-regulated gene count labeled by their KEGG metabolic pathway map id and description followed by a gene count.

Figure 56. *KEGG pathway* context for *up/down-regulated* genes

Clicking on a KEGG metabolic pathway map id will redirect you to the *expression experiment viewer*'s internal KEGG metabolic pathway browser, which will show you the up-regulated genes in the selected metabolic pathway. Clicking on the count that follows the metabolic pathway description will show you the expression levels of all top up-regulated genes in the indicated KEGG metabolic pathway for the selected experiment across all report sets.

Down-regulated Tab

The *down-regulated tab* in the *expression experiment viewer* is identical to the *up-regulated tab* except that it shows the top down-regulated genes instead of the top up-regulated genes. Headers, navigation, and displayed results are identical. Please refer to the *up-regulated tab* section above for more information.

Plots Tab

The *plots tab* shows four plots to help you assess the quality of the experiment report set data. The first is a histogram of $\log_2$-ratio values by the number of genes. The second plot shows the cumulative probability of $\log_2$-ratio values. The third, a volcano plot, shows $\log_2$-ratio values compared to their corresponding absolute Z-score. The fourth plot shows operon-based estimates of local accuracy by showing the Z-scores compared to the percentage of true changers.

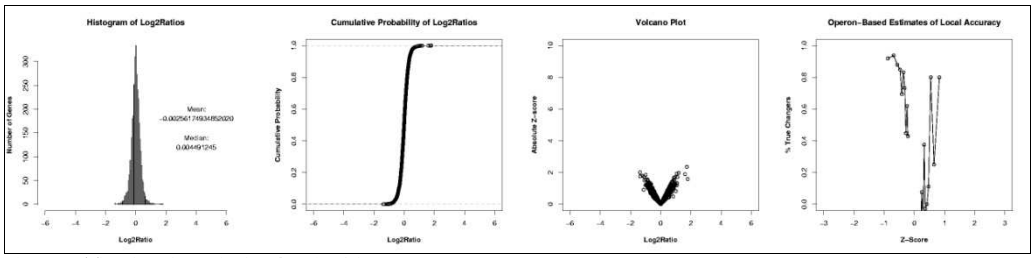


Figure 57. The plots shown in the *plots* tab of the *expression experiment viewer*

KEGG Maps Tab

The *KEGG maps tab* contains a KEGG metabolic pathway browser similar to the MicrobesOnline *Pathway Browser*. It allows you to browse KEGG metabolic pathways showing enzymes as up-regulated or down-regulated based on associated gene expression in the selected experiment. The interface looks identical to the *Pathway Browser* except two color ranges are used to indicate up-regulated (red) and down-regulated (blue) enzymes. Enzymes shown in grey were not significantly up-regulated

or down-regulated in the current comparison and enzymes shown in white are not represented by genes in the current genome. In cases where an enzyme is represented by two or more genes you may see different color bands. In these cases, the box representing the enzyme is divided into equal portions according to the number of genes and each gene is colored according to its regulation in the experiment.

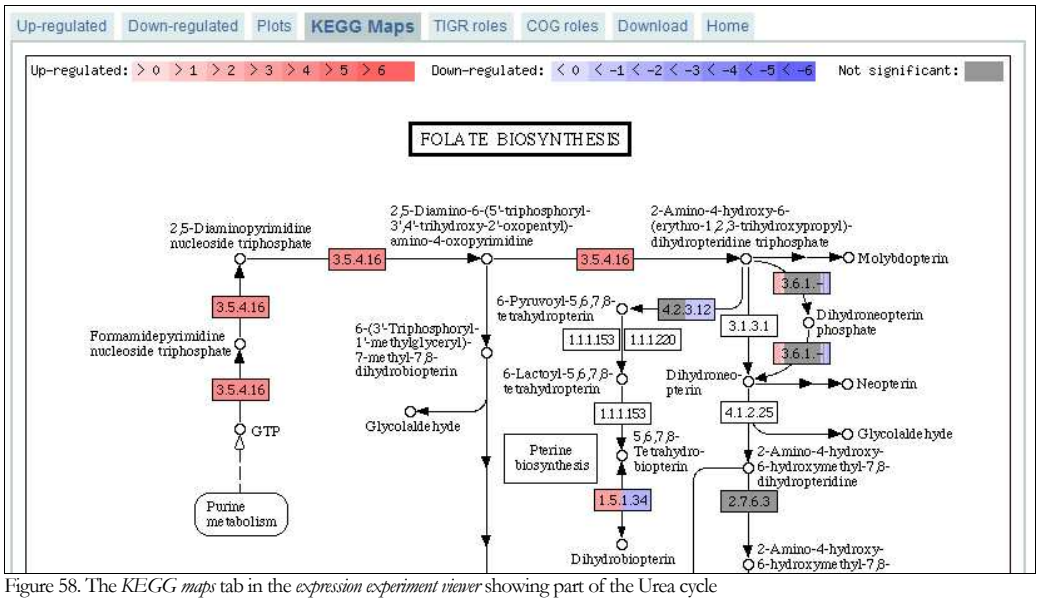


Figure 58. The *KEGG maps* tab in the *expression experiment viewer* showing part of the Urea cycle

Links to related pathways are shown in rectangles with rounded corners and compounds are shown as circular nodes. Clicking on compounds or non-represented enzymes will redirect you to the KEGG website to view additional information about the selected compound or enzyme. Clicking on pathway references will redraw the *KEGG maps tab* showing the selected pathway. Clicking on a represented enzyme will show the enzyme's associated gene's expression level in the current experiment across all of its comparisons.

TIGR Roles Tab

The *TIGR roles tab* shows the distribution of genes in the selected experiment report set according to their assigned TIGR role and the number of genes assigned to each role that were up-regulated or down-regulated. Some genes are not assigned TIGR roles and therefore they will not be represented in the data shown on the *TIGR roles tab* even if they are present within the selected experiment report set.

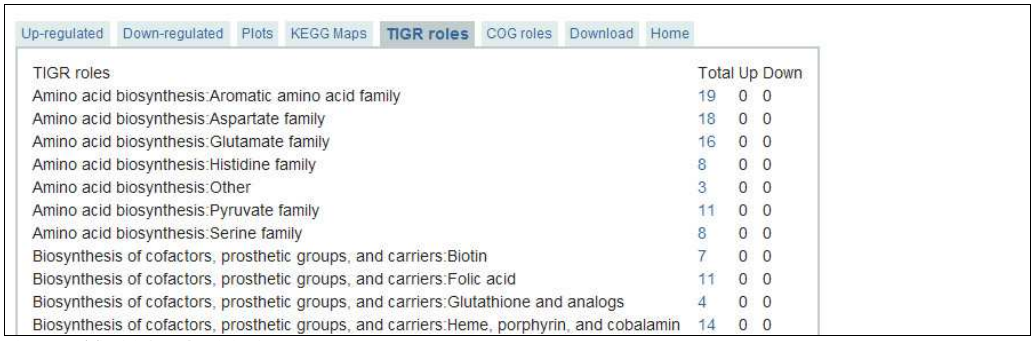


Figure 59. The *TIGR roles* tab in the *expression experiment viewer*

Each of the displayed counts for each TIGR role, if non-zero, is a hyperlink that will show the expression levels of the total, up-regulated, or down-regulated genes assigned to the indicated TIGR role for the selected experiment across all report sets, respectively.

COG Roles Tab

The *COG roles tab* shows the distribution of genes in the selected experiment report set according to their assigned COG functional category and the number of genes in each category that were up-regulated or down-regulated. Some genes are not assigned a COG functional category and therefore they will not be represented in the data shown on the *COG roles tab* even if they are present within the selected experiment report set.

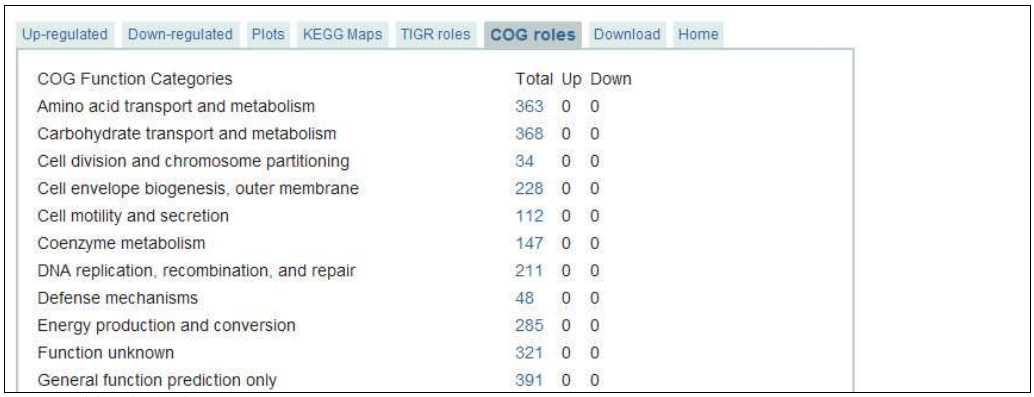


Figure 60. The *COG roles* tab in the *expression experiment viewer*

Each of the displayed counts for each COG functional category, if non-zero, is a hyperlink that will show the expression levels for the member genes, in the current experiment across all report sets.

Download Tab

The *download tab* allows you to download the raw report data in a tab-delimited format that is compatible with Microsoft® Excel. Not all report sets are available for download and therefore may result in a *"Page not found"* error when clicking on the *download tab*. Please contact us at `gtlweb@vimss.lbl.gov` to request a specific dataset. If the raw report set data is available, you will be prompted by your web browser to save the Excel file.

Gene Expression Viewer

The *gene expression viewer* shows a graphical heatmap of the expression levels of the selected gene of interest, all other genes in the operon containing the selected gene of interest, and the first upstream and downstream gene of the operon in all experiments and their full data series (all report sets) which contain the gene of interest. Since this tool is gene-centric, it is found as a link **Gene expression** from the *Locus Information* tool. This link may not be present for some genes, indicating the lack of any experiments containing that gene.

The header portion of the *gene expression viewer* is constant across all genes with the exception of the positive and negative correlation gene expression profile search links.

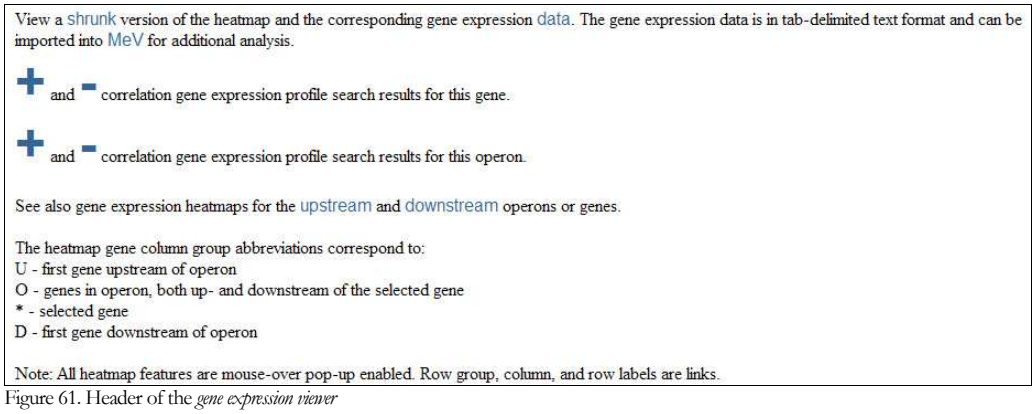


Figure 61. Header of the *gene expression viewer*

Some datasets are extremely large resulting in large heatmap images, which may take an extremely long time to load and which may be slow to access on slower computers. We recommend use of the **shrunk** link to view the same gene expression heatmap with much smaller data cells. You may also download the raw gene expression data in tab-delimited format by clicking on the **data** link in the first sentence.

The large **+** and **−** icons are displayed if gene expression profile search results are available for the gene of interest or its operon. Some genes may have gene expression data but have insufficient data in order to generate a statistically significant gene expression profile search. These genes cannot be used with the *profile search tool* described in the next section. The **upstream** and **downstream** links allow you to view the gene expression heatmaps of the upstream and downstream genes of the operon containing the selected gene of interest.

Next, a legend explains the symbols used for abbreviation purposes in the gene expression heatmap. Genes are displayed as columns in the heatmap and each gene belongs to a group based on whether it is the selected gene of interest (*****), part of the selected gene of interest's operon (**O**), or the upstream (**U**) or downstream (**D**) gene of the selected gene of interest's operon. Genes are identified by their corresponding gene name or symbol. Clicking on a gene's name will redirect you to the *Locus Information* tool for the selected gene and mousing over the gene name will provide you

with additional information about the selected gene, including its description and COG assignment, if any.

The experimental data series is arranged with report sets as rows arranged into experiment groups. Each report set contains the specific stress conditions including the stress level and time point, while the experiment groups contain the high-level stress description. In the **shrunk** heatmap view, report set and gene labels are replaced by horizontal and vertical bars, respectively. Clicking on a report set label will redirect you to the *expression experiment viewer* for the selected report set and clicking on the experiment group label will redirect you to the *experiment browser* showing all experiments with the same stress condition in all organisms.

Each cell in the gene expression heatmap represents the expression level for a specific gene in a specific experiment report set. Cells are color coded according to the scale presented to the right of the gene expression heatmap, an indication of the relative amount of up- or down-regulation based on its $\log_2$-ratio value. If a particular $\log_2$-ratio value has a significant Z-score, its colored cell is given a border according to the Z-score scale shown under the $\log_2$-ratio value scale. Some cells may not contain any data and they are rendered according to the **No Data** cell in the legend.

Mousing over a cell in the gene expression heatmap displays the cell's $\log_2$-ratio value as well as summaries of its gene and report set labels. This is of more importance in the **shrunk** heatmap, where gene and report set labels are not rendered.

None of the cells in the gene expression heatmap are clickable.

You may save the gene expression heatmap to your computer by right-clicking on the heatmap on a PC or left-clicking and holding the left mouse button on a Mac, then selecting the **Save Picture** or **Save Image** option in the pop-up window.

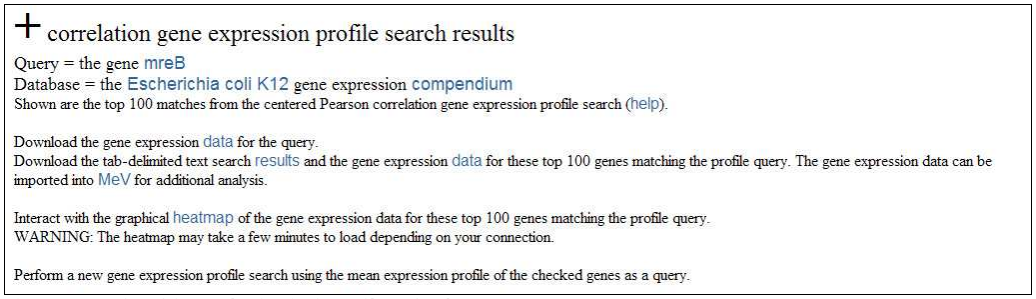
Profile Search Tool

A gene or operon expression profile search allows you to identify other genes that match the expression profile of the gene or operon you've selected. This is useful in identifying other genes that could potentially be regulated using the same mechanism as the selected gene or operon. You can select from positive (**+**) correlation searches or negative (**–**) correlation searches.

The *profile search tool* has two operating modes. The first, and default entry view, is the profile search *gene list view* showing all genes with expression levels statistically similar to the selected gene or operon. The *profile heatmap view* is a graphical heatmap representation of the selected gene or operon's expression level and the expression levels of the genes identified to be similar to the input profile.

Gene List View

The *gene list view* presents the results of the expression profile search as a list of genes whose expression is similar to the input profile. Similarity is determined by using a linear (Pearson) correlation.


+ correlation gene expression profile search results

Query = the gene mreB
Database = the Escherichia coli K12 gene expression compendium
Shown are the top 100 matches from the centered Pearson correlation gene expression profile search (help).

Download the gene expression data for the query.
Download the tab-delimited text search results and the gene expression data for these top 100 genes matching the profile query. The gene expression data can be imported into MeV for additional analysis.

Interact with the graphical heatmap of the gene expression data for these top 100 genes matching the profile query.
WARNING: The heatmap may take a few minutes to load depending on your connection.

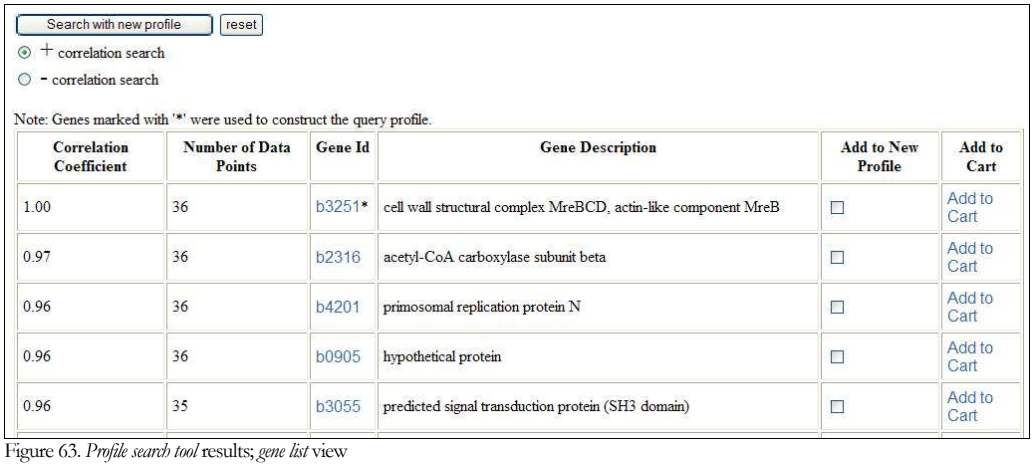
Perform a new gene expression profile search using the mean expression profile of the checked genes as a query.

Figure 62. Header of the *gene list* view in the *profile search tool*

The selected gene or operon is shown as the **Query**. Each displayed gene is a hyperlink to the *Locus Information* tool for the indicated gene. The **Database** indicates which organism's genes will be searched against the input profile and the **compendium** link will redirect you to the *experiment browser* showing all experiments for the indicated organism. Clicking on the organism name will redirect you to the *Genome Information* tool.

You can download the gene expression data for the input gene or operon by clicking on the **data** link in the *"Download the gene expression data for the query"* statement. You can download the list of genes from the profile search by clicking on the **results** link as well as download the gene expression for all of the genes returned by the profile search by clicking on the **data** link.

Finally, you can switch to the *profile heatmap view* by clicking on the **heatmap** link.


Search with new profile | reset

+ correlation search

- correlation search

Note: Genes marked with '*' were used to construct the query profile.

Correlation Coefficient	Number of Data Points	Gene Id	Gene Description	Add to New Profile	Add to Cart
1.00	36	b3251*	cell wall structural complex MreBCD, actin-like component MreB	☐	Add to Cart
0.97	36	b2316	acetyl-CoA carboxylase subunit beta	☐	Add to Cart
0.96	36	b4201	primosomal replication protein N	☐	Add to Cart
0.96	36	b0905	hypothetical protein	☐	Add to Cart
0.96	35	b3055	predicted signal transduction protein (SH3 domain)	☐	Add to Cart

Figure 63. *Profile search tool* results; *gene list* view

The genes from the target organism whose expression levels are well correlated with the input gene or operon expression levels are displayed in tabular form in descending order by Pearson correlation coefficient. Genes used to construct the input profile are marked by an asterisk in the **Gene Id** column. The **Number of Data Points**

corresponds to the number of expression comparisons considered in the profile search. The gene id is a hyperlink that will redirect you to the *Locus Information* tool for the indicated gene and the **Add to Cart** hyperlink will add the indicated gene to your *session gene cart*.

The profile search tool is particularly powerful as it allows you to select genes from the results of a previous profile search to use in building a new profile. This new profile can in turn be used to search for genes with well-correlated expression levels. This also allows you to rebuild operon profile searches excluding genes that may not actually be part of the operon.

To use this feature, select the genes you wish to use in a new profile search by checking its corresponding box under the **Add to New Profile** column. Then select whether you want a positive or negative correlation profile search by selecting the appropriate option. Finally, click on the **Search with new profile** button to generate a new profile search.

To clear all of the selected genes, click on the **reset** button.

Profile Heatmap View

Figure 64. *Profile heatmap* view

The *profile heatmap view* displays the input expression profile on top with experiment report sets oriented as columns and grouped into experiment column groups according to stress.

Below the input expression profile are the expression levels of all the genes identified in the profile search as being well-correlated with the input expression profile. The $\log_2$-ratio value scale is displayed to the right of the expression profile result gene list's heatmap.

Experiment report set and gene labels are condensed just as in the **shrunk** view of the *gene expression viewer*.

Clicking on the experiment group labels, such as **Heat shock** in the example to the left will redirect you to the *experiment browser* showing all experiments with the selected condition, across all organisms.

Clicking on a condensed experiment report set label will redirect you to the *expression experiment viewer* for the selected experiment report set and clicking on a condensed gene label will redirect you to the *Locus Information* tool for the selected gene. You can view condensed label names and additional information by mousing over the condensed label.

Mousing over the cells in the expression profile results heatmap will show the individual expression levels for the selected experiment report set and gene. As with the **shrunk** view in the *gene expression viewer*, the experiment report set and gene labels are also displayed in the mouseover information.

None of the heatmap cells are clickable.

You may save the gene or operon expression profile heatmap and/or the profile search heatmap to your computer by right-clicking on the desired heatmap on a PC or left-clicking and holding the left mouse button on a Mac, then selecting the **Save Picture** or **Save Image** option in the pop-up window.

Tutorials

This section contains tutorials for the most common usage scenarios of gene expression data on the MicrobesOnline site.

Accessing Gene Expression Data for a Gene of Interest

To view the gene expression data for a specific gene, type the gene name or gene id as a search query in the *Find Genes* tool search box. In the list of results, each gene will have a series of lettered links of which the **E** link (if present) leads to a heatmap view of the expression data for that gene in the context of its operon and first genes upstream and downstream of the operon.

Accessing Gene Expression Data for an Operon of Interest

To view the gene expression data for a specific operon, type a gene name or gene id of a gene in the operon as a search query in the *Find Genes* tool search box. Click the **O** link for the gene of interest which will lead to the 'Operon & Regulon' tab of the *Locus Information* page for that gene. On the 'Operon & Regulon' tab, click on the 'Gene expression' link just below the row tabs. This will bring up a heatmap view of the expression data for that gene in the context of its operon and first genes upstream and downstream of the operon.

Performing a Gene Expression Profile Correlation Search for a Gene of Interest

To perform a gene expression correlation profile search for a specific gene, type the gene name or gene id as a search query in the *Find Genes* tool search box. In the list of results each gene will have a series of lettered links and the **E** link (if expression data is available) leads to a heatmap view of the expression data for that gene in the context of its operon and first genes upstream and downstream of the operon. At the top of the heatmap view page there are **+** and **–** links to perform positive and negative correlation gene expression profile searches.

Performing a Gene Expression Profile Correlation Search for an Operon of Interest

To perform a gene expression correlation profile search based on the mean expression of all genes in a specific operon, type a gene name or gene id of a gene in the operon as a search query in the *Find Genes* tool search box. Click the **O** link for the gene of interest to arrive at the 'Operon & Regulon' tab for that gene. On the 'Operon & Regulon' tab of the *Locus Information* page, click on the 'Operon' *Dynamic expression profile search* link to perform a positive correlation gene expression profile search.

Viewing Gene-Gene Expression Correlations for an Operon

To view gene-gene expression correlations for a specific operon, type a gene name or gene id of a gene in the operon as a search query in the *Find Genes* tool search box. Click the **O** link for the gene of interest to arrive at the 'Operon & Regulon' tab for that gene. On the 'Operon & Regulon' tab of the *Locus Information* page, click on the 'Genes in the operon with downstream and upstream genes' *Gene expression correlations* link to view a heatmap of gene-gene expression correlations for the genes in this operon as well as the first genes upstream and downstream of the operon.

Accessing Gene Expression Data for a Set of Genes in a Gene Cart

To view gene expression data for a set of genes in the cart, navigate to the 'My gene carts' page and click on the named link of a specific cart. Then scroll down to the bottom of the cart contents page and click on 'View heatmap of gene expression data for these genes' which brings up a heatmap view of the expression data for these genes.

Viewing Gene-Gene Expression Correlations for a Set of Genes in a Gene Cart

To view gene-gene expression correlations for a set of genes in the cart, navigate to the 'My gene carts' page and click on the named link of a specific gene cart. Then scroll down to the bottom of the cart contents page and click on 'View heatmap of gene expression correlations for these genes' which brings up a gene-gene expression correlation matrix.

Performing a Gene Expression Correlation Profile Search for a Set of Genes in a Gene Cart

To perform a gene expression correlation profile search based on the mean expression of all genes in this cart navigate to the 'My gene carts' page and click on the named link of a specific gene cart. Then scroll down to the bottom of the carts content page and click on 'Perform a gene expression profile search with the mean expression profile of these genes' to perform a positive correlation gene expression profile search.

Gene Carts

Gene carts in MicrobesOnline are a convenient method for saving sets of genes and for analyzing them. There are two types of gene carts. The first is a *session gene cart*, one that is associated with your current browser session. *Session gene carts* are not persistent and are deleted as soon as you close your browser window, however they may be permanently saved if you are a registered user of MicrobesOnline. *Saved gene carts* are gene carts associated with your registered user profile and they are accessible any time you access MicrobesOnline by logging in with your account.

Session Gene Cart

The **Cart** or **Add to Cart** links present throughout MicrobesOnline will add the indicated gene to your *session gene cart*. Upon adding a gene to your *session gene cart*, a **Genes of Interest** window will pop-up showing the newly added gene and any other genes you've previously added in the same session. This window can be left open or closed at your discretion and closing the window does not remove or clear your *session gene cart*.

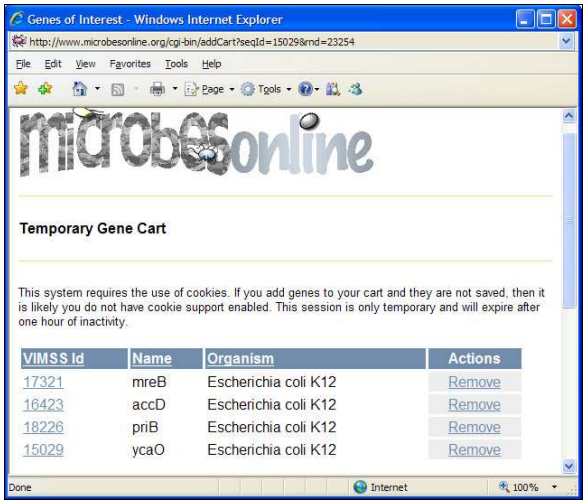


Figure 65. *Session gene cart* pop-up window showing four genes from *E. coli K-12*

There are three sections to the *session gene cart*. The first shows the cart's contents in tabular format. You may click the headers of the table to sort the results on the selected column, however only ascending sort is supported. Each gene's **VIMSS Id**, **Name**, and source **Organism** are displayed, along with a link to **Remove** the corresponding gene. Clicking on the **VIMSS Id** of a gene will redirect you to the *Locus Information* tool for the selected gene.

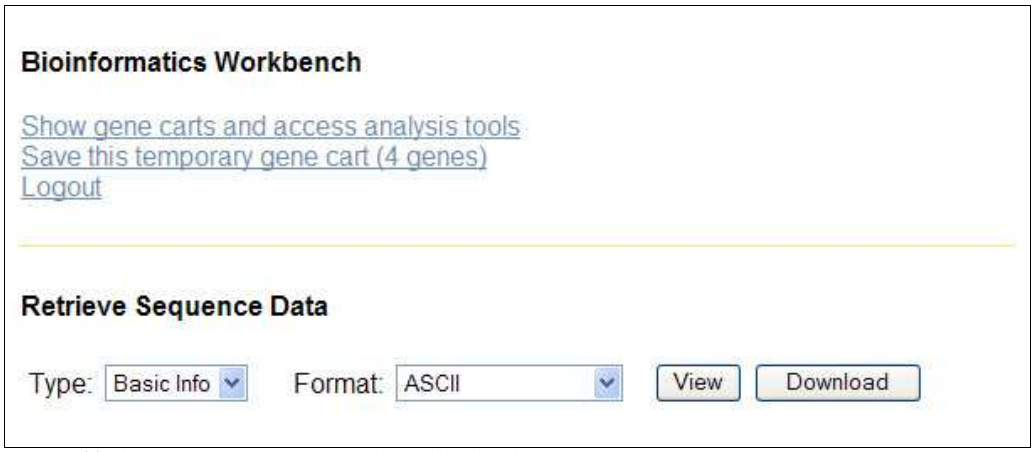


Figure 66. The bottom of the *session gene cart* contains additional tools

The other two sections of the *session gene cart* are the *bioinformatics workbench* and *sequence retrieval* sections. The *bioinformatics workbench* options are only available for registered MicrobesOnline users and allow you to save your *session gene cart* into a permanent gene cart and to access a variety of *gene cart analysis tools*.

The *sequence retrieval* section allows you to quickly retrieve gene information and/or sequences by selecting the appropriate option from the **Type** drop-down selection box. The **Format** specifies whether you want the data sent as plain-text (**ASCII**) or in a compressed format (Unix ***compress***, ***gzip***, or z**ip** are supported). The **View** button will pop-up a new browser window and display the selected information in plain-text

format. Clicking on **Download** or selecting any of the compressed data formats will result in a file being downloaded to your computer.

To save a *session gene cart*, simply click on the **Save this temporary gene cart** link. The *session gene cart* will be saved using the name of the first gene added to the cart and you will be redirected to the *gene cart viewer*, which presents a similar view of the cart contents as the *session gene cart*, but allows you to access the analysis tools. For more information, please see the *gene cart viewer* section.

Gene Cart Summary

The *gene cart summary* shows a list of all saved gene carts and your *session gene cart*, if you've created one, along with basic actions that allow you to copy, edit, or delete the cart, and to share it with other users. You can access the *gene cart summary* by clicking on the *My Gene Carts* link from the MicrobesOnline home page's *top navigation* menu, or by clicking the *Show gene carts and access analysis tools* link from the *session gene cart*'s *Bioinformatics Workbench* section.

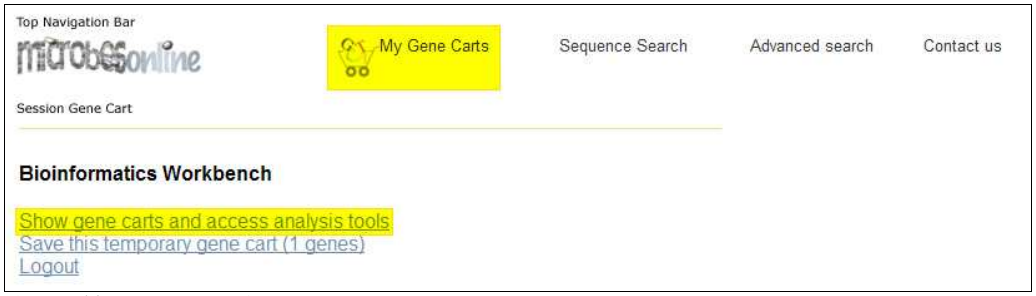


Figure 67. Two ways to access the *gene cart summary*

The top half of the *gene cart summary*, labeled *Bioinformatics Workbench*, provides a series of links, most of which allow you to view job results and associated data generated from your gene carts. The **View all** links show a summary of all available job results according to the job result type specified in the link. The job results summaries show basic information about each job, including the name of the job and associated gene cart, the date and time the job ran, associated parameters you selected when running the job and a summary of the results. The views are similar though each job type may have additional information not present for other job types therefore each summary view is described in greater detail in their respective subsequent sections.

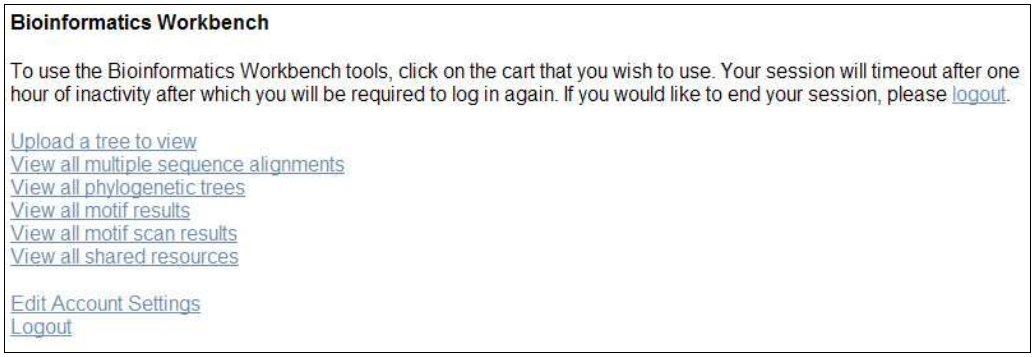


Figure 68. The *bioinformatics workbench* section of the *gene cart summary*

Shared resources refer to any gene cart or associated job result that you are sharing to other users or that other users are sharing with you. Clicking on the **View all shared resources** link provides an overview of which users have access to your data and the level of access, as well as other resources on MicrobesOnline to which you have access. To learn more about sharing resources with other users, see the *resource access control* section.

Clicking on the **Upload a tree to view** will allow you to access the *upload tree interface*, which allows you to upload pre-computed trees in a variety of formats for visualization and further analysis using MicrobesOnline. Please see the *upload tree interface* section for more information.

The **Edit Account Settings** link allows you to edit basic properties of your account, including your name, password, and email preferences. Please see the *account settings* section for more information.

Clicking on the **Logout** link will remove your login credentials thus preventing others from accessing your data.

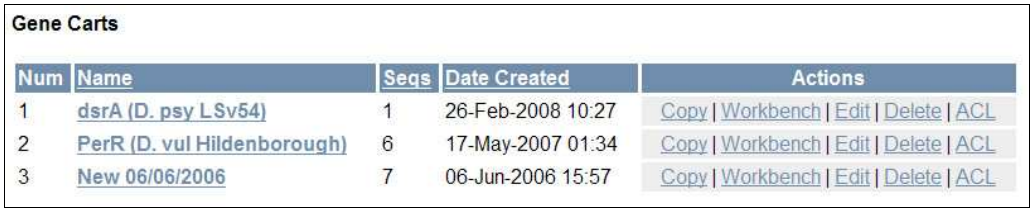


Figure 69. The *gene carts* section of the *gene cart summary*

The *Gene Carts* section of the *gene cart summary* shows a list of available gene carts including your *session gene cart* if you've created one. The list of gene carts can be sorted in ascending order according to the selected column by clicking on its column heading. For saved gene carts, clicking on the gene cart name will open the *gene cart viewer* for the selected gene cart, and you may use the links in the *Actions* column to perform basic operations on gene carts, such as copying, editing, removing, and configuring access control. Clicking on the **Copy** link will copy the gene cart's contents to your *session gene cart* and return you to the *gene cart summary*. Editing a gene cart allows you to add or remove genes and change the name and

access control for your cart. Deleting a gene cart will remove the cart from your user profile along with all associated job result data. To perform analysis using a gene cart, select the **Workbench** link corresponding to the gene cart you wish to use. This will open the *gene cart viewer*. To view or modify access control lists for a cart, click on the **ACL** link. For more information on the *gene cart viewer* or *resource access control*, see the appropriate subsequent sections.

Your *session gene cart* is labeled as *Temporary* when viewed in the *gene cart summary* and not all operations are available. You may view, edit, or delete your *session gene cart* by using the appropriate link. Clicking on the **View/Edit** link will take you to the *session gene cart* page and clicking the **Delete** link will empty the contents of your *session gene cart* without asking for confirmation. To access the **Workbench** you must save your *session gene cart* first by clicking on the **Save** link. You will be returned to the *gene cart summary* after your *session gene cart* has been saved and it will now be displayed in the cart list with an automatically assigned name of the first gene added to the cart.

Gene Cart Viewer

The *gene cart viewer* displays details about a gene cart including its list of sequences, and provides links to retrieve associated sequences, access *workbench tools*, view associated jobs, and modify its access control for sharing.

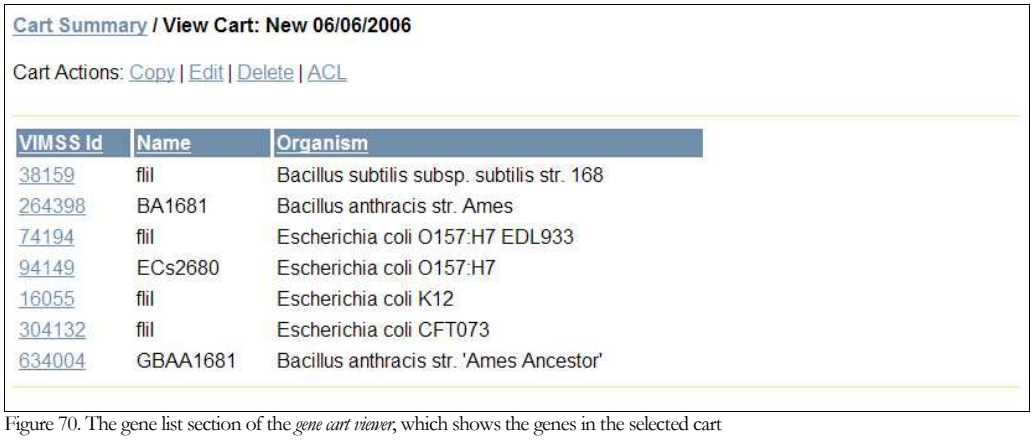


Figure 70. The gene list section of the *gene cart viewer*, which shows the genes in the selected cart

The first section of the *gene cart viewer* shows a list of the genes in the selected gene cart, one gene per line including the gene's VIMSS id, name or gene symbol, and associated organism. By default, the genes are listed in the order in which they were added to the gene cart however you may change the ordering by clicking on any of the column headings. Only sorting in ascending order is supported at this time. The *Cart Actions* links function identically to their counterparts on the *gene cart summary*.

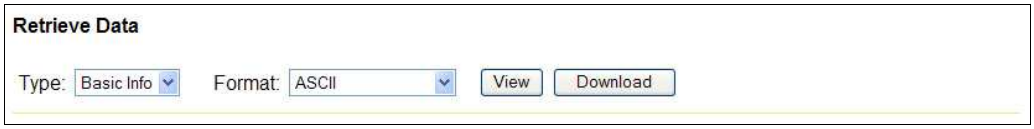


Figure 71. The *retrieve data* section of the *gene cart viewer* allows you export information and sequences of the genes in the selected cart

The *retrieve data* section of the *gene cart viewer* functions identically to the *retrieve data* section of the *session gene cart*. This tool can be used to retrieve information about the genes in your cart, including their transcript and protein sequences, and basic information about the gene, such as its beginning and ending coordinates, source scaffold, and organism. You may select the data to view or download, and for downloads you may optionally select a compression method. For large files, it is recommended that you compress the data and download the data to your local computer prior to viewing it. For more information on this section, please see the *session gene cart* section.

The *bioinformatics workbench* section provides access to different analysis tools and other data associated with the genes in a cart, such as associated microarray heatmap data and job results.

Figure 72. The *bioinformatics workbench* section of the *gene cart viewer* provides access to various analyses tools

Clicking on the **View Cart in Genome Browser** link opens the *cart browser*, a genome browser with one track for each gene in the selected cart. Please see the *gene cart browser* section for more information on using this tool.

To build a multiple sequence alignment or a gene tree, click on the **Create a Multiple Sequence Alignment** link. Gene trees can only be built from multiple sequence alignments or uploaded using the **Upload a tree to view** link. For more information on building multiple sequence alignments, or building or uploading gene trees, please see the *multiple sequence alignment tool*, *tree building tool*, and *upload tree interface* sections respectively.

Clicking on the **View heatmap of expression data for these genes** link will open the *microarray heatmap viewer* showing all available experiments and conditions for the genes in the selected cart. For more information on using this tool, please see the *microarray heatmap viewer* section.

Clicking on any of the **View Saved …** links will show you all temporary and saved job results for the desired job type. For more information on using these tools, please see the *workbench tools* section.

Finally, the **View/Modify Cart Permissions** opens the *resource access control* tool, which allows you to share the selected cart with other members on MicrobesOnline. Please see the *resource access control* section for more information.

Gene Cart Browser

The *gene cart browser* is much like the *Ortholog Browser* and will display each gene in the selected cart and its neighbors aligned, instead of an anchor gene and its orthologs in other organisms. This tool can be used for a variety of purposes including to visually confirm whether there are genes in the respective neighborhoods of the genes in your cart that are also reported as orthologs.

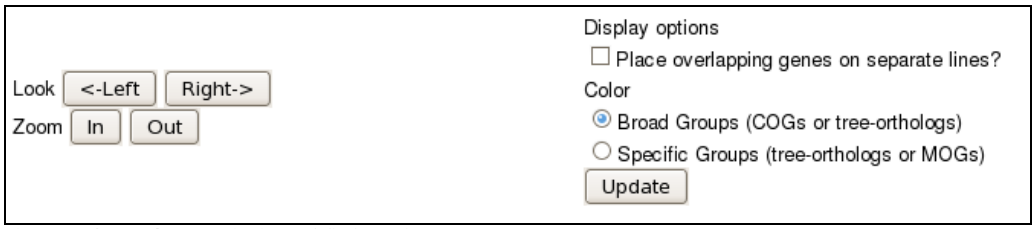


Figure 73. *Gene cart browser* navigation and display options

To navigate the *gene cart browser* display, use the buttons labeled **Left**, **Right** , **In**, and **Out**. Clicking on the **Left** or **Right** buttons will shift the current display left or right, respectively, by 33% of the current viewable range, which by default is 10,000 bp. Clicking on the **In** or **Out** buttons will change the zoom of the current display by 1.5x in or 1.5x out, respectively, changing the effective viewable range. The *Display options* allow you to change basic display properties. Checking the box labeled **Place overlapping genes on separate lines?** will cause overlapping genes in a browser track to be rendered on different lines. You can also change the coloring scheme.

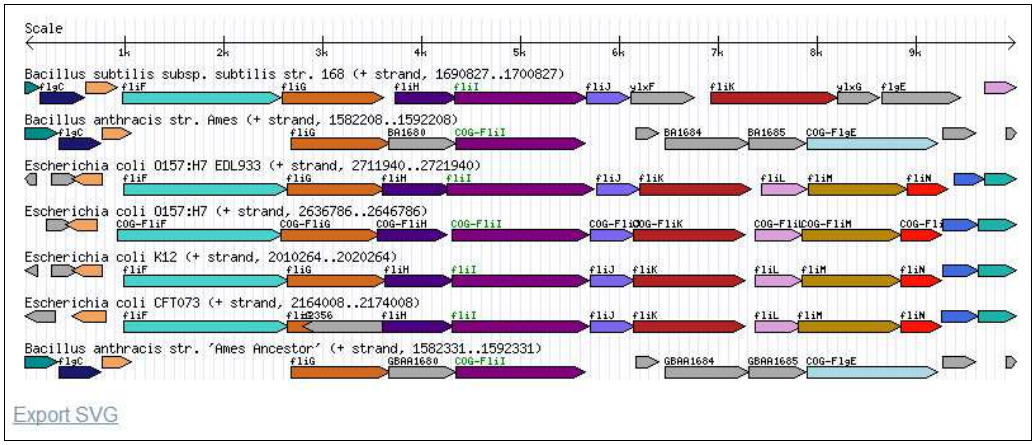


Figure 74. The browser display shows genes in your cart centered and highlighted in green along with their gene neighborhoods

The browser display shows one track per gene in the selected cart. Each track is labeled above with the name of the organism in which the gene exists, along with the current viewable range and strand. Clicking on the organism name label of a track will open the *Genome Information* tool for the selected organism. You can mouse over a gene to get more information and click on a gene to view it or to add it to a cart.

Below the browser display, you can click the **Export SVG** link to build a vector-based graphic of the current view. These are infinitely re-scalable without loss of resolution and are ideal for use in publications as opposed to the lower resolution web graphic displayed in your browser.

Workbench Tools

MicrobesOnline provides you with numerous tools with which you can perform further analysis of the sequences in your gene carts. The following diagram shows the types of analysis that are available including the specific analysis tools we offer for each type.

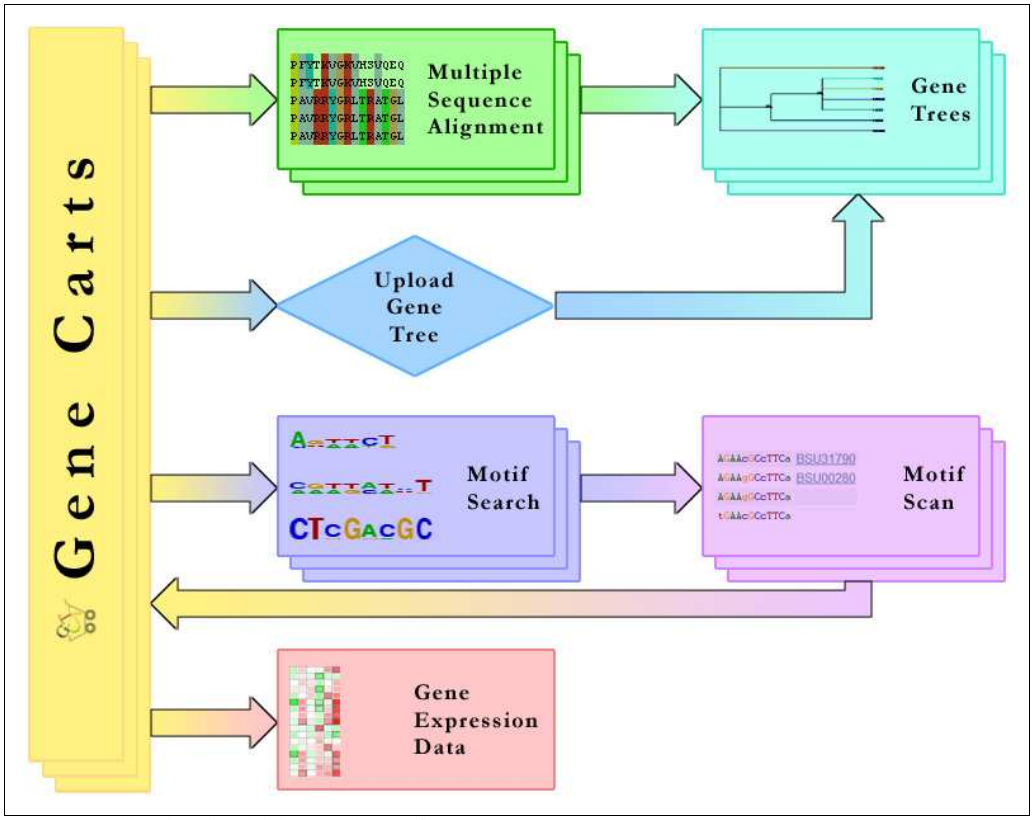


Figure 75. Workbench tools and their respective workflows

Each job submitted from the *workbench tools* is saved temporarily to your session including all of the results generated by the job. Since these job results will eventually be removed from our system to save space, it is recommended that you save job results that you do not wish to be removed once your session is terminated. Each account has

a limited amount of space and once the space is used up you will be unable to save additional job results.

You can run at most two jobs concurrently and each job can run at most for 4 hours before our system terminates it. Since our analysis server is a shared resource for all users it is highly recommended that you perform any intensive or long-running analysis offline.

Each job result is treated as a resource associated with your account, much like gene carts. Each job result can also be shared with other users or groups using the *resource access control* tool. We also provide basic tools for viewing and filtering job results, which will allow you to easily return to the results of a previously completed job, as well as to save or remove the results. These tools and their respective interfaces are described in the subsequent sections, organized by workbench tool types.

Job Status

After you submit a job, you will be redirected to a job status page that will update periodically to display the status of the job you've recently submitted. This shows your job's unique *Job Id*, the job type, its status, and the total amount of time the job has been running as of the most current update. Once the job completes you will be automatically redirected to the results page for the specific job type. Our analysis server is a shared resource therefore job performance is dependent upon the current server load and may be unusually slow when there are many concurrent users running analysis. Please be patient and refrain from resubmitting your job multiple times. If you suspect a problem, please feel free to contact us by using the **Contact Us** link present in the footer of most pages.

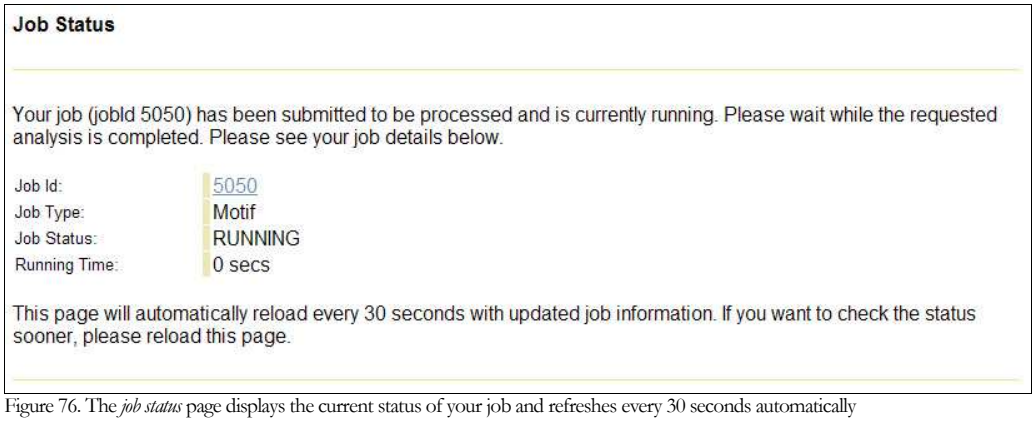


Figure 76. The *job status* page displays the current status of your job and refreshes every 30 seconds automatically

If your job failed to run or produce the expected output files, if you have exceeded the maximum number of allowed concurrent jobs, or if your job's running time exceeded the maximum allowed running time, the job status will show **FAIL** along with an error message explaining the problem. If you've specified invalid custom parameters, you will need to return to the job submission page to correct the invalid parameters before resubmitting your job.

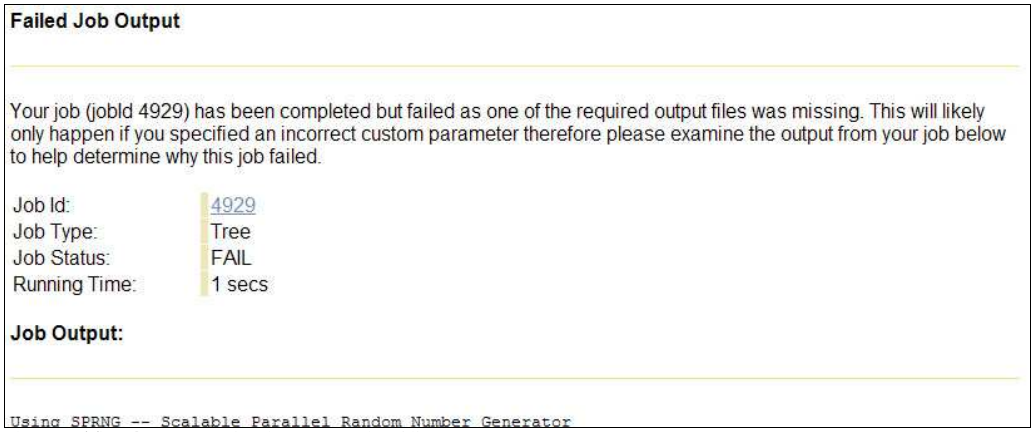


Figure 77. The *job status* page showing an execution failure

Multiple Sequence Alignments

A multiple sequence alignment shows the optimal global alignment of all gene sequences in your cart and is useful in determining or computing the relative distance of genes for phylogenetic studies. We use the MUSCLE tool to generate multiple sequence alignments. To access the *multiple sequence alignment* tool, you must first access the *gene cart viewer* for the cart containing the genes on which you wish to base your multiple sequence alignment. From the *gene cart viewer*, click on the **Create a Multiple Sequence Alignment** link.

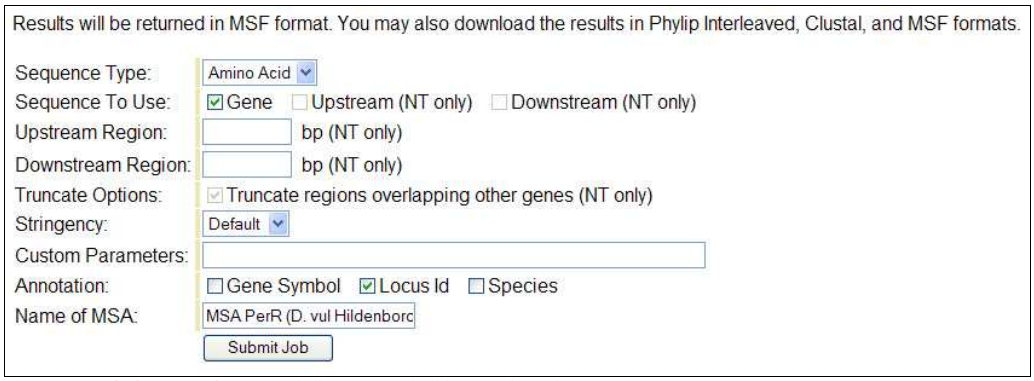


Figure 78. *Multiple sequence alignment* tool allows you to build MSAs from the genes in a cart

The *multiple sequence alignment* tool allows you to customize your multiple sequence alignment. The default parameters will compute the multiple sequence alignment for the protein sequences of the genes in the selected cart and assign a name derived from the name of the selected cart.

You may change whether the protein sequence (**Amino Acid**) or transcript sequences (**Nucleotide**) are used in the multiple sequence alignment by selecting the appropriate option from the **Sequence Type** drop-down select box. Note that the availability of some of the subsequent options depends on the selected **Sequence Type**, so some options may be grayed out and unavailable. Selecting to use the transcript sequences allows you to change the region used in building the multiple sequence alignment. By default, only the coding region is used, however you may opt to scan only the

upstream or **downstream** sequences, or any combination thereof. After selecting to use the upstream and/or downstream sequence, you must specify the number of base pairs upstream and/or downstream of the gene transcript you wish to use in the alignment. Finally, you have the option of truncating sequences that overlap with regions of other genes on the genome even if they are not present in your cart. This is useful for ensuring that upstream and/or downstream sequence does not contain part of an adjacent gene.

We have experimentally determined the optimal stringency settings that work well in most cases. These will be used unless you select **Custom** from the **Stringency** option. By doing this, you will be allowed to specify custom tool parameters to pass to MUSCLE, however you must be familiar with these options as specifying an invalid option will cause your job to fail.

The **Annotation** option allows you to specify which attributes to display in the results. By default only the **Locus Id** is selected.

Finally, you can optionally specify a name for your multiple sequence alignment. This is useful in locating the results at a later time.

To submit the job, click on the **Submit Job** button and the corresponding *job status* page will be displayed. Once completed, you will automatically be redirected to the *multiple sequence alignment result* viewer. Please see the *job status* section of this document for assistance if your job fails.

Multiple Sequence Alignment Results

The *multiple sequence alignment result* viewer contains several sections, each of which will be described in greater detail in this section. In general, this viewer allows you to view the results of your multiple sequence alignment in graphical or raw MSF format, to generate or upload a related gene tree, to view and change the job's access control, or to view other related jobs.

View Results

This section provides tools for viewing your multiple sequence alignment. If you have a Java™ runtime environment (JRE) you can use the graphical multiple sequence alignment viewer *JalView* by clicking on the **JalView** button. If you do not see this button it is likely you do not have a JRE installed or your JRE must be updated to a more current version. For assistance with *JalView*, please see the documentation located on their website.

REFERENCES

JalView
http://www.jalview.org

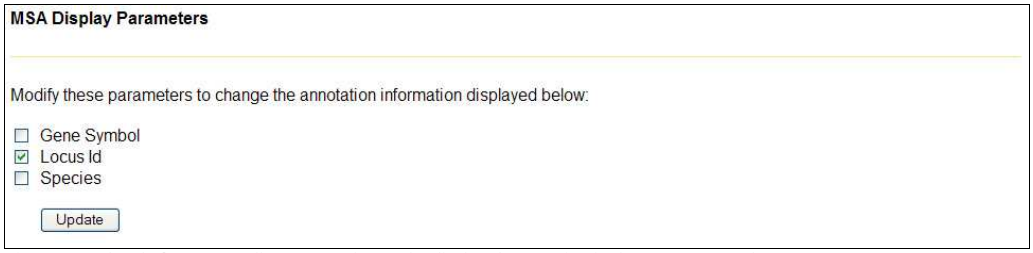


Figure 79. *MSA display properties* allow you to change the displayed annotation in the raw MSA results

Below the **JalView** button you can view the raw multiple sequence alignment in MSF format, as well as download the raw alignment in a number of different formats. To configure the attributes that are displayed in the raw alignment, check the desired boxes under the *MSA Display Parameters* heading. By default, only the **Locus Id** will be selected. If you change the display parameters, you must click the **Update** button to refresh the page with the new settings.

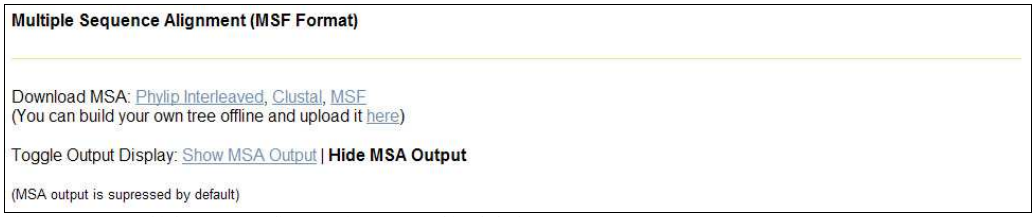


Figure 80. Raw multiple sequence alignment is suppressed by default

The raw alignment is displayed in the section below the *MSA Display Parameters* labeled *Multiple Sequence Alignment (MSF Format)* and is hidden by default. You can download the raw alignment in **Phylip Interleaved**, **Clustal,** and **MSF** formats by clicking on the corresponding link following the **Download MSA** label. We offer gene tree building tools, however you may have a specific tool optimized for the genes you selected so the download option allows you to download the multiple sequence alignment from which you can generate your own gene tree offline and upload it to MicrobesOnline for viewing. If you choose to generate a tree offline and wish to associate it with the multiple sequence alignment you just performed, you can click on the **upload it here** link below the download options or scroll down to the *Generate a Phylogenetic Tree* section of the results view page.

Figure 81. Raw multiple sequence alignment is shown

As mentioned previously, the raw alignment is hidden by default to conserve space on the results page. If you wish to view the raw alignment, click on the **Show MSA Output** link and the raw alignment should appear below. You can re-hide the raw alignment at any time by clicking on the **Hide MSA Output** link. When viewing the raw alignment, the selected gene attributes also provide a hyperlink which will open the *Locus Information* tool for the selected gene.

Generate a Phylogenetic Tree

This section allows you to generate a gene tree using the multiple sequence alignment you just generated or by uploading and associating a gene tree you built offline. If you wish to upload a tree rather than generate one using our workbench, click on the **Upload Tree** link under the **Method** label. For more information on building or uploading gene trees, please see the *gene trees* section.

Save Job Results

Once a job has completed, its results are only temporarily associated with your account on MicrobesOnline and is subject to removal in as little as one hour. If you wish to access the results at a later time without re-running the analysis then you should save the results permanently by clicking on the *Save My Results* link.

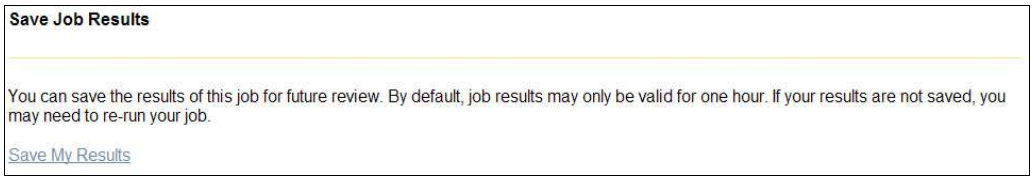


Figure 82. Save your job results so that you can access them later without re-running the analysis

Each account is allowed a fixed amount of space for storing job results. You will be unable to save additional job results once you've exceeded the allotted amount.

Current Resource Permissions

By default your job results can only be viewed, modified, or deleted using your account however if you wish to share certain job results with other users or groups you may do so by clicking on the **Add New Rights** button. If you're already given other users or groups access to a particular resource and you wish to change their level of access you can add or remove permissions as appropriate, then click on the **Update Rights** button. For more information on permissions, see the *resource access control* section.

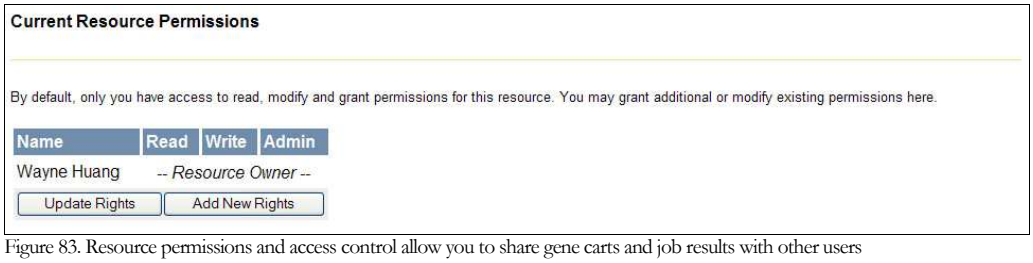


Figure 83. Resource permissions and access control allow you to share gene carts and job results with other users

Other Related Resources

This section provides links for viewing lists of subsets of related job results. You can view other multiple sequence alignments created using the same cart as the multiple

sequence alignment you are viewing by clicking on the **View other MSAs created from the same gene cart** link. You can view a list of tree job results created from multiple sequence alignments created from the same cart as the multiple sequence alignment you are viewing by clicking on the **View trees created from the same gene cart** link. You can also view a list of all your multiple sequence alignment and gene tree results by clicking on the **View all MSAs** or **View all trees** links, respectively.

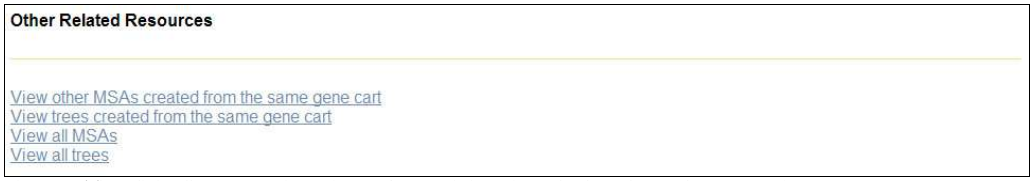


Figure 84. These links allow you to view subsets of your existing job results that relate to the current job

Gene Trees

MicrobesOnline provides various tree building tools to allow you to build gene trees from multiple sequence alignments. We also provide an upload facility that allows you to upload a custom built gene tree. The facility to build or upload gene trees is integrated into the *multiple sequence alignment results* viewer. The facility to upload custom gene trees is also available from the *bioinformatics workbench* section of the *gene cart summary* and *gene cart viewer*, accessible by clicking on the **Upload a tree to view** link.

Building a Gene Tree from a Multiple Sequence Alignment

The primary method is to use the MicrobesOnline workbench to generate a gene tree from your multiple sequence alignment. If you wish to upload a tree you generated offline instead of building a tree with the workbench, click on the **Upload Tree** link under the **Method** label.

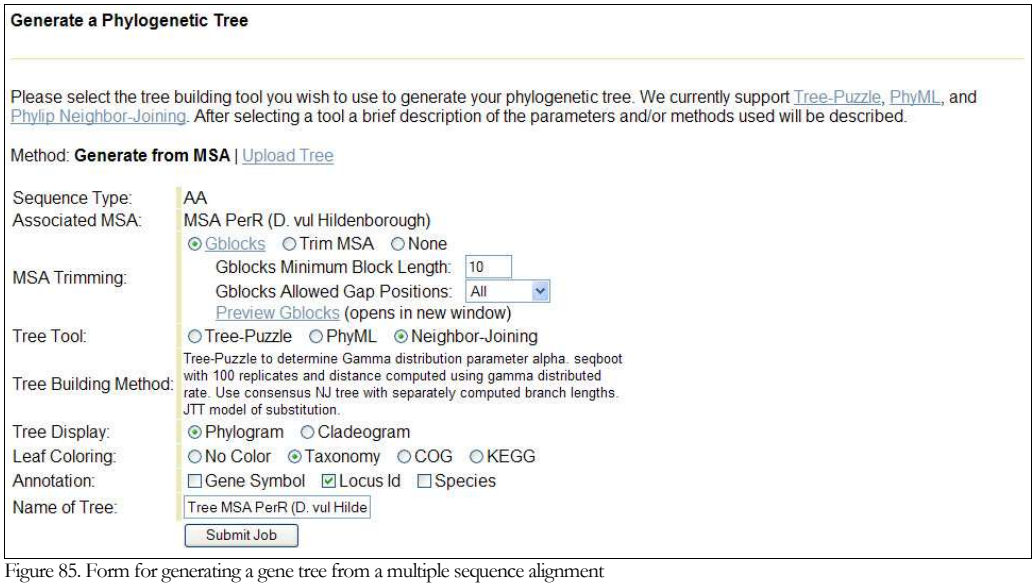


Figure 85. Form for generating a gene tree from a multiple sequence alignment

The form includes basic information about the multiple sequence alignment. The **Sequence Type** is **AA** or **NA** corresponding to whether amino acids (protein) or nucleic acids (transcript) were used, respectively. In addition, the tree you build will be associated with this multiple sequence alignment.

The first set of options allow you to specify if desired, the method by which your multiple sequence alignment should be trimmed prior to the tree-building step. The default is to use Gblocks, a tool that intelligently removes positions from a multiple sequence alignment using a number of different criteria. For information about Gblocks, please click on the **Gblocks** link.

Gblocks has a number of parameters that will allow you to fine tune trimming of your multiple sequence alignment. The **Gblocks Minimum Block Length** parameter specifies the minimum size of any trimmed block. Blocks less than this length are omitted when building the final trimmed sequence. Setting this value too high may result in unusual truncation of your multiple sequence alignment. The **Gblocks Allowed Gap Positions** parameter specifies how to treat positions in your multiple sequence alignment that contain one or more gap characters. The default setting, **All**, allows any number of gaps to appear in positions in the final trimmed alignment. Setting the parameter to **None** disallows any positions that contain one or more gap characters, while it to **Half**, allows positions in the final trimmed alignment to contain gaps so long as more than half of the sequences contain a non-gap character at this position. Finally, you may preview the final trimmed alignment before proceeding with the tree building process by clicking on the **Preview Gblocks** link. This link will open a new window showing how Gblocks created the final trimmed alignment.

The **Trim MSA** option uses simply removes all positions that contain one or more gap characters. This works well when the sequences used in the multiple sequence alignment are highly conserved, but may result in unusual truncation of your multiple sequence alignment if the sequences are distant. In this case, it is recommended that you use the **Gblocks** trimming method, or no trimming (**None**).

Selecting **None** as the **MSA Trimming** option disables trimming of your multiple sequence alignment prior to tree building.

The tree building workbench supports three different tree building tools. **Tree-Puzzle** is generally accurate and very fast for small trees containing less than 30 sequences. When selecting **Tree-Puzzle** as the desired tree tool you can also select the rate type used to generate the tree. **4 rate categories** is slightly more accurate but is a couple times slower than the default, **Uniform rates**. **PhyML** is generally the most accurate of the three tools however it does not provide bootstrap values and therefore it is difficult to assess the correctness of the generated tree. **Neighbor-Joining** is the slowest method and has comparable accuracy to **Tree-Puzzle**, however it provides bootstrap values, which can help you assess the quality of the tree. **Tree-Puzzle** provides puzzle values, which are similar to bootstrap values and can help you assess the quality of the tree. The default method is **Neighbor-Joining**.

Beneath the **Tree Tool** selection we present a brief description of the method and other parameters that are used with the selected tool. If you are unable to generate a satisfactory gene tree using our tools, please download your multiple sequence alignment, generate the tree offline using a tool and parameters of your choice, and upload the resultant tree using the *upload tree interface*. Your tree must use VIMSS ids as sequence identifiers in order to be compatible with MicrobesOnline.

The **Tree Display** option sets the type of tree that will be rendered. A ***Phylogram*** is a tree whose branch lengths are proportional to the estimated amount of evolutionary change whereas a **Cladogram**'s branch lengths are not proportional to this change. The default is **Phylogram**.

Leaf Coloring allows you to select the criteria to use when coloring the leaf node annotations. By default, leaf node annotations are colored by **Taxonomy** of the source organism. Selecting **No Color** will skip coloring of leaf node annotations, while selecting **COG** will color leaf node annotations according to each gene's COG assignments. Similarly, selecting **KEGG** will color leaf node annotations according to EC number assignments. The **Leaf Coloring** setting can be changed after your tree is generated to recolor the tree without rerunning the analysis.

The **Annotation** selection boxes allow you to specify which attributes of each gene to include in the leaf node annotations. By default, only the **Locus Id** is included.

Finally, you may assign a name to your tree or accept the default name which is generated from the name of the parent multiple sequence alignment.

To submit your tree job, click on the **Submit Job** button. This will redirect you to the *job status* page for your tree job and you will automatically be redirected to the *tree results* viewer when your job has completed. Please see the *job status* section for assistance with troubleshooting job failures.

Uploading a Gene Tree

MicrobesOnline allows you to upload your own gene tree in a variety of formats. If accessing the *upload tree interface* outside the context of a saved multiple sequence alignment, you may select the associated multiple sequence alignment from among your saved multiple sequence alignments. You must use VIMSS ids as sequence identifiers within your tree in order to be compatible with MicrobesOnline.

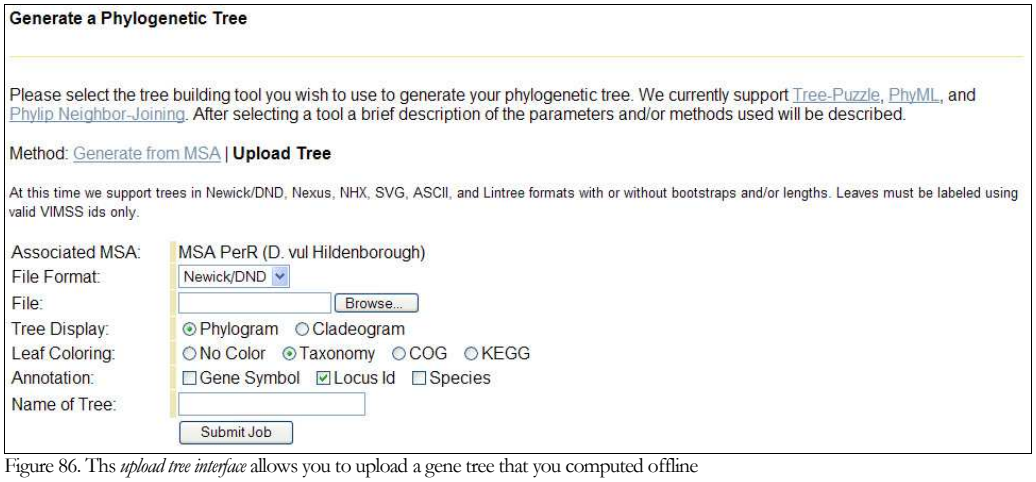


Figure 86. Ths *upload tree interface* allows you to upload a gene tree that you computed offline

From the **File Format** drop-down selection box, select the format of the tree you wish to upload. The default value, **Newick/DND**, is the most popular format produced by various tree building programs however a number of different formats are also supported. Please contact us if you are using a file format unsupported by our *upload tree interface* by clicking on the **Contact Us** link present in the footer of most pages.

Next, select the tree file from your local computer by clicking on the **Browse** button.

The remaining attributes are identical to those found when building a gene tree with our workbench. The **Tree Display** attribute selects the type of rendered to tree to show on the *tree results* viewer. A **Phylogram** has branch lengths proportional to the estimated amount of evolutionary change whereas the **Cladogram** branch lengths are merely a function of the topology.

Leaf Coloring controls how leaf node annotations are colored. The default, **Taxonomy**, colors leaf node annotations according to the taxonomy of the source organism. Selecting **No Color** will omit coloring of the leaf node annotations. Selecting **COG** will color leaf node annotations according to each gene's COG assignments. Similarly, selecting **KEGG** will color leaf node annotations according to each gene's EC assignments.

The **Annotation** selection boxes allow you to specify which of the gene's attributes are displayed in the leaf node annotations. By default, only the **Locus Id** is checked.

You may optionally assign a name to the tree file you are uploading.

To submit the job, click on the **Submit Job** button. As processing a tree file is generally very fast when compared to building the tree itself, the *tree results* viewer should show almost instantaneously however in the event of any delays you will be redirected to the *job status* page.

Gene Tree Results

This tool allows you to view the rendered gene tree and associated annotation according to the settings used when building or uploading your gene tree.

Figure 87. The display settings change how the resultant tree is rendered

The top **Display** section displays a summary of the parameters used to build or upload your gene tree, along with options to change the rendering of the tree. You must refresh the gene tree viewer by clicking on the **Update** button to apply your changes.

The **Display Outgroup** select box shows a list of all genes by their corresponding VIMSS locus id included in your gene tree. You may select one or more genes to use as an outgroup in rendering your gene tree. To select more than one gene hold down the **Ctrl** key on a PC or the **Command** key on a Mac, then use your mouse to click on the genes you wish to include.

The **Tree Display**, **Leaf Coloring**, and **Annotation** options work identically to those found in the *upload tree* and *multiple sequence alignment results* interfaces, described previously.

Finally, to apply your changes you must click on the **Update** button.

The rendered gene tree will appear below the **Update** button and above the **Save Job Results** section.

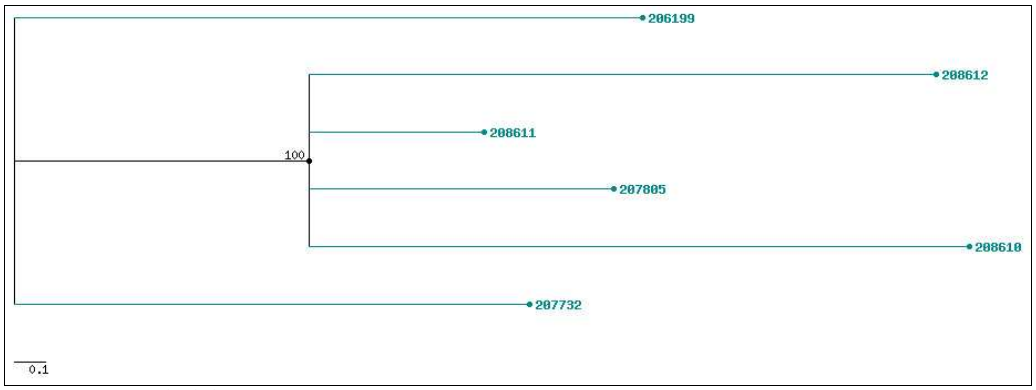


Figure 88. An example phylogram tree

Save Job Results

This section is shown on every job result viewing tool. Please see the s*ave job results* section under the *multiple sequence alignment results* section for more information.

Current Resource Permissions

This section is shown on every job result viewing tool. Please see the *current resource permissions* section under the *multiple sequence alignment results* section for more information.

Other Related Resources

This section provides a number of links to view other related tools and to view other job results that are related to the tree you are viewing.

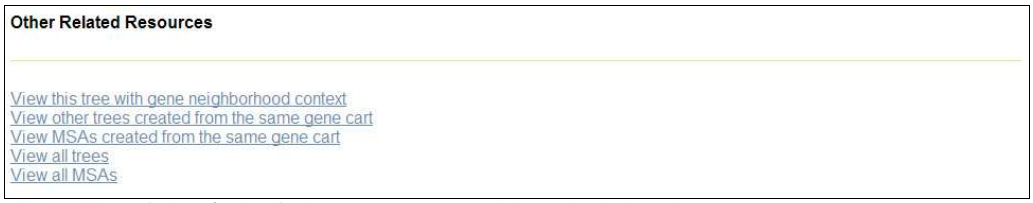


Figure 89. *Tree results* viewer's *other related resources* section contains links to related job results

To view your gene tree along with the context of each gene's neighborhood, click on the **View this tree with gene neighborhood context** link. This link will redirect you to the *cart tree browser*, a tool derived from the *Tree Browser*. The user interface has some subtle differences but contains mostly the same navigation and rendering options as the *Tree Browser*. Please see the *cart tree browser* section for more information.

The following two links are only present for trees built using MicrobesOnline or for uploaded trees that were associated with an existing multiple sequence alignment job from your account. The first, **View other trees created from the same gene cart**, will show a summary of all tree jobs built from the same gene cart. Similarly, you can view a summary of all multiple sequence alignment jobs built from the same gene cart by clicking on the **View MSAs created from the same gene cart** link.

To view all tree jobs, click on **View all trees**. To view all multiple sequence alignments, click on **View all MSAs**.

Cart Tree Browser

The *cart tree browser* allows you to view the genes in a gene tree with their respective gene neighborhoods, or contexts. Each gene in the gene tree is shown centered in the **gene context** labeled in green with the corresponding **gene tree** shown to the left.

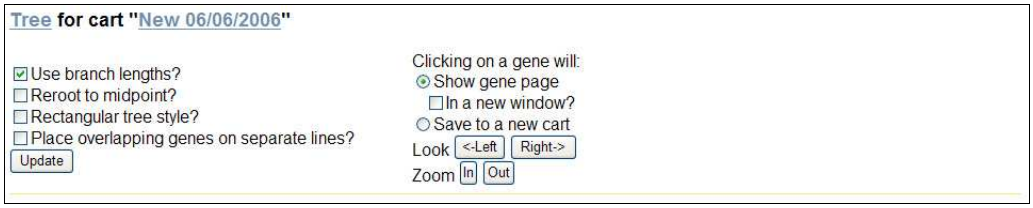


Figure 90. Display and navigation options in the *cart tree browser* control how the tree and gene context are rendered and allow you to navigate the browser

At the top of the *cart tree browser*, the link **Tree** will redirect you to the reference *gene tree* and the link labeled with the parent gene cart's name will redirect you to the *gene cart viewer*.

The four checkbox options on the left are the display options. They control how the gene tree and gene context are rendered. By default, only the **Use branch lengths** option is selected. This option controls whether branches in the gene tree are rendered proportional to their computed lengths representing relative distance. The **Reroot to midpoint** option will redraw the gene tree rooted to its midpoint, which is useful when there is no known biological outgroup. The **Rectangular tree style** option will render all connecting lines between nodes using only vertical and horizontal straight lines, whereas the default mode will draw lines at any angles to connect nodes. In the **gene context**, all genes in a track are rendered on a single line, even if they are overlapping. Checking the **Please overlapping genes on separate lines** option will prevent overlapping genes from being rendered on the same line. If you make changes to any of these display settings, you must click the **Update** button to apply your changes.

The options under the **Clicking on a gene will** heading allow you to control what action is taken when you click on a gene in the **gene context**. By default, you will be redirected to the *Locus Information* tool for the selected gene. Checking the box **In a new window** applies only applies when **Show gene page** is the selected option and will cause the *Locus Information* tool to open in a new window. Select the **Save to a new cart** option if you would like to add genes to your *session gene cart*.

The next set of buttons is for **gene context** navigation. The **Look** buttons allow you to scroll the **gene context** left or right by approximately 33% of the current viewable range, which by default is 10,000 bp. Note that changing the **Zoom** will affect the amount by which the **Look** buttons shift the display.

The **Zoom** buttons allow you to change the zoom of the **gene context** in or out 1.5x the current viewable range. Thus clicking the **Out** button once will change the default viewable range from 10,000 bp to 15,000 bp.

As with the *Ortholog Browser* and *Tree Browser*, colors of putative orthologs are determined dynamically each time the image is redrawn. Therefore it is likely that shifting the view or changing the zoom will change the specific color of a set of putative orthologs.

The primary display is divided into two components. The left side shows the selected **gene tree** and the right side shows the corresponding **gene context**.

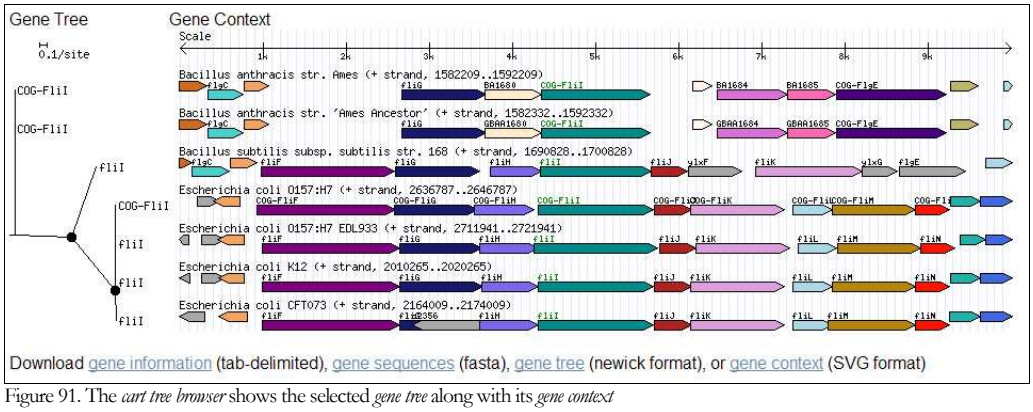


Figure 91. The *cart tree browser* shows the selected *gene tree* along with its *gene context*

In the displayed gene tree, you can move your mouse over any internal node and the bootstrap or support value will be displayed in a popup. The leaf nodes of the gene tree are labeled with the gene's name, symbol, or in some cases its COG-derived name. Clicking on the leaf node labels will result in the action selected under **Clicking on a gene will**.

In the **gene context**, each displayed gene can be clicked and will result in the desired action as specified by the **Clicking on a gene will** setting. Clicking on the genome name will redirect you to the *Genome Information* tool for the selected genome.

Below the main display are a series of links for downloading related data. The **gene information** link allows you to export basic information about each gene, such as the VIMSS id, name, description, and taxonomy information of the source genome. The **gene sequences** link will export the protein sequences of the genes present in your gene tree. The **gene tree** link will export the gene tree in *Newick/DND* format and the **gene context** can be downloaded in a high resolution scalable vector graphics (SVG) format file for use in publications by clicking on the **gene context** link.

Motif Searches

MicrobesOnline provides a facility for identifying motifs in selected regions in and/or around the genes in a gene cart. The *motif search* tool is divided into two steps. First, you must select the genes to search in. Then you select a motif search tool and configure its parameters. At present, we support *AlignACE*, *MEME*, and *Weeder*, with *MEME* being the default search tool.

Selecting Sequences for Motif Search

Figure 92. Interface for selecting or adding sequences for a *motif search*

By default, all sequences in the selected gene cart are selected for inclusion in the motif search. The table shows each gene along with basic information about the gene. The checkbox to the left of each gene's VIMSS id indicates whether the gene is to be included in the motif search. By default the genes will be displayed in the order in which they were added, however you can change the sorting by clicking on the table headers. Clicking once will result in ascending sort of the selected column and clicking again will change the sort to descending order.

Each gene's VIMSS id is a hyperlink that will redirect you to the *Locus Information* tool for the selected gene. Each gene's source organism name is a hyperlink that will redirect you to the *Genome Information* tool for the selected genome. If a COG assignment exists for a gene, the COG id is displayed and is a hyperlink to the *Find Genes* tool showing all genes in the source organism assigned to the specified COG cluster. Similarly, if an EC assignment exists for a gene, the EC number is displayed and is a hyperlink to the *Find Genes* tool showing all genes in the source organism with the specified EC assignment.

Above the gene table we provide links to assist you in selecting genes of interest to consider in your motif search. The **Enable full description listing** will show the full gene **description** and descriptions associated with each **COG** cluster id and **EC** number assigned to a gene in the table. You can unselect all genes by clicking on the **Clear Selected Genes** link, or alternatively you can select all genes by clicking on the **Select All Genes** link.

If you alter the default selection of genes, you must click on the **Apply Changes** button to save the changes. To reset the changes you've made to the last saved state, click on the **Reset** button.

You may add additional genes using the **Locus Ids** text entry box. Specify one or more genes comma-separated by their VIMSS id. If you would like the selected gene cart updated to include the specified list of genes, check the **Add these sequences to the current cart** checkbox. Finally, to apply the changes click on the **Add Sequences** button.

Motif Search Common Parameters

The first three parameters are common to all available motif search tools.

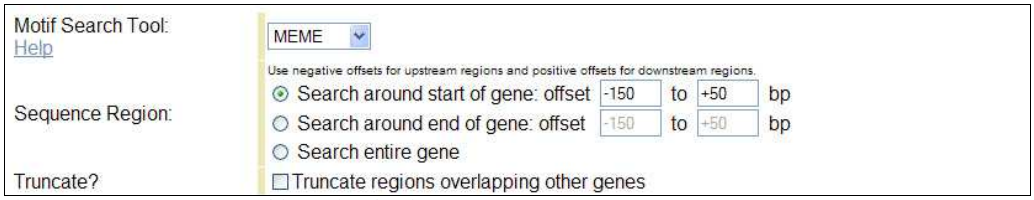


Figure 93. Common parameters of the *motif search* tool

Motif Search Tool allows you to select the desired motif search tool. By default, *MEME* is selected but we also support *AlignACE* and *Weeder*.

Sequence Region allows you to define which regions relative to the selected genes will be searched for motifs. By default the region around the start of the gene according to the specified offsets will be searched. To specify offsets, use a negative value to indicate a position upstream of the reference position (start of gene or end of gene) and a positive number to refer to a position downstream of the reference position. The default [-150, +50] specifies a region of 201 bp including 150 bp upstream of the start of the gene and 50 bp downstream from the start of the gene. To search around the end of the gene, select the second option **Search around end of gene**. To limit the search to the gene, select the third option **Search entire gene**.

Truncate specifies whether overlapping regions should be truncated. For example if two selected genes are close to each other, it's possible for the region around the start of the downstream gene to include part of the sequence of the upstream gene. Enabling this option would truncate the region of the downstream gene such that it does not include any of the sequence of the upstream gene.

The **Name of Motif Search** option is present regardless of the selected motif search tool. This option appears at the bottom of the list of parameters, just above the **Submit Job** button and allows you to name the *motif search* job. To submit your job you must click on the **Submit Job** button.

AlignACE Search Parameters

AlignACE uses Gibbs sampling to find likely motifs. It uses an iterative masking approach to find additional motifs after the best scoring motif has been identified. MicrobesOnline uses the GC content of the input sequences as the expected background GC content.

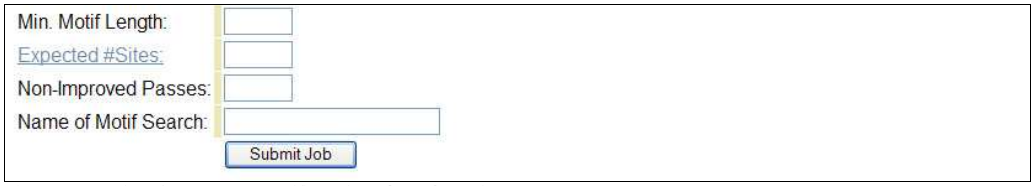


Figure 94. *Motif search* parameters specific to the *AlignACE* tool

The **Min Motif Length** parameter accepts an integer value corresponding to the number of base pairs of the minimum motif *AlignACE* will consider.

The **Expected #Sites** corresponds to the number of expected occurrences of a motif in the input sequences. By default, *AlignACE* assumes 10 however here you can set a new value.

The number of **Non-Improved Passes** affects the running time of *AlignACE*. A small number of passes will result in a faster running time but it is possible to miss higher scoring motifs. An extremely large number will result in a much slower running time but may not necessarily result in higher scoring motifs. The default value is 200.

MEME Search Parameters

MEME uses the expectation-maximization algorithm to quickly find motifs and automatically chooses the most statistically significant motif, and the most significant length for that motif.

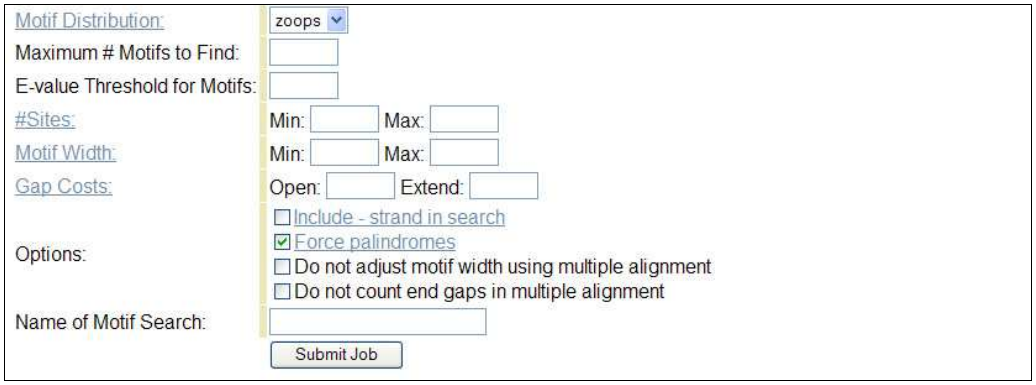


Figure 95. *Motif search* parameters specific to the *MEME* tool

The **Motif Distribution** controls the method *MEME* uses to determine whether detected motifs may be shared. The default, **zoops**, or *zero or one per sequence* assumes that each input sequence may contain at most one occurrence of each motif. This option prevents *MEME* from excluding motifs that may be missing from some of the input sequences, however it is about twice as slow and is slightly less sensitive to weak motifs compared to the **oops** model, or *one per sequence*. This model assumes that each input sequence will contain exactly one occurrence of each detected motif, and is also the fastest and most sensitive however motifs missing from some input sequences may result in those motifs being noisier. The third model, **anr**, or *any number of repetitions*, assumes each input sequence may contain any number of non-overlapping occurrences of each motif and is useful if you suspect that motifs may repeat multiple times within a

single input sequence. In this case, the detected motifs will be more accurate compared to using the other two models. This is the slowest method and is slightly less sensitive to weak motifs which do not repeat in a single input sequence. For more information, please see ↳ `http://meme.sdsc.edu/meme/help_distribution.html`.

The **Maximum # Motifs to Find** option limits the total number of motifs *MEME* will identify and return. Using an excessively large value may result in your job taking an extremely long time. Likewise, reducing this value may decrease your job's running time.

The **E-value Threshold for Motifs** requires all returned motifs to have an e-value less than or equal to the specified threshold. Using a value that is too low may result in *MEME* returning no motifs if the motifs are weak, however using a value that is too large may result in many insignificant motifs being returned.

The **# Sites** parameter controls the minimum and maximum number of sites to which a detected motif should match. By default, *MEME* expects $\sqrt{n}$ to n sites, where n is the number of input sequences.

The **Motif Width** parameter determines the minimum and maximum length of motifs *MEME* should return. By default, *MEME* searches for a wide range of motif lengths and returns motifs with lengths that represent the most statistically significant.

The **Gap Costs** parameters allow you to control the penalties used when *MEME* is searching for similarities. The **Open** value corresponds to the penalty assigned to initiate a gap and the **Extend** value corresponds to the additional penalty assigned to extend a gap. The default gap values are 11 to open and 1 to extend, but you can change these values.

The **Options** checkboxes can be used to change some of *MEME*'s behavior when identifying motifs. By default, *MEME* will not search the reverse complement of the input sequence however checking the **Include – strand in search** will cause *MEME* to also scan the reverse complement of each sequence. By default, MicrobesOnline uses *MEME* to search for palindromic motifs because most bacterial transcription factors bind DNA as dimers, and most of these dimers bind in opposite orientations. When searching for palindromes by checking the **Force palindromes** option, it doesn't make sense to also include the reverse complement of each sequence as either strand will result in a hit to the motif. However, if you are not searching for palindromic motifs, you should search both strands. The **Do not adjust motif width using multiple alignment** option controls whether *MEME* should trim motifs using a multiple sequence alignment. By default *MEME* will trim motifs, however checking this option will disable this behavior. By default, *MEME* will penalize for end-gaps when performing a multiple alignment, however checking the **Do not count end gaps in multiple alignment** will disable this behavior.

Weeder Parameters

Weeder uses exhaustive enumeration to find sequences that are frequently present, in almost exact copies, in the input sequences and is generally the fastest of the three available tools.

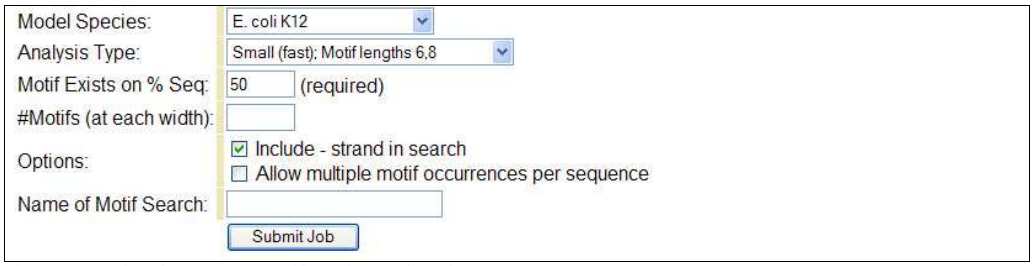


Figure 96. *Motif search* parameters specific to the *Weeder* tool

The **Model Species** provides *Weeder* with information about the frequency of various *k*-mers it should expect when searching for motifs in your sequences. You should select the species that is closest to those on which your input sequences are found.

The **Analysis Type** option specifies the type of search you wish to perform and the respective motif lengths considered. The more motif lengths and longer motifs that are considered, generally the slower your job will run.

The **Motif Exists on % Seq** option specifies the minimum percentage of the input sequences which must have a detected motif for it to be reported. The default value is 50%. Increasing this to a very high percentage may result in no motifs being reported. Likewise, a very small percentage may result in many insignificant motifs being reported.

The **# Motifs (at each width)** parameter controls the maximum number of motifs to detect and report for each motif length considered, as determined by the selected **Analysis Type**. Using a small value will speed up the analysis but may result in a reduction of reported significant motifs. Using a large value will increase the running time of your job and may result in many insignificant motifs, however you should consider increasing this value if a weak motif you are searching for is not reported with the default setting.

Under the **Options** parameters, the **Include – strand in search** will cause *Weeder* to also search the reverse complement of each input sequence. This is enabled by default and disabling this feature will speed up your search, but may result in some motifs not being reported. The **Allow multiple motif occurrences per sequence** option will allow motifs that occur at multiple sites on a single input sequence to be reported, otherwise the first occurrence is the only one reported.

Motif Search Results

Depending on the selected motif search tool and parameters used, and the number and length of input sequences, your motif search job may take between 5 seconds and 10 minutes to complete. Due to a restriction placed on the running time of jobs, your job

may be terminated prior to completion. If this is the case you should adjust the parameters to reduce the search time.

If your search criteria do not produce any motifs, you will see **No motifs were detected with the selected tool and search criteria**. You will be given the option of viewing the raw output of the tool by clicking on the **raw output** link and will also be presented with *other related resources*, described below. If your search didn't produce any results you should try relaxing some of the constraints as described in the tool parameters' sections above.

If one or more motifs were detected and reported, you will see the *motif search results* viewer described in this section. The results are divided into two primary sections with the standard *save job results*, *current resource permissions*, and *other related resources* sections that are present on all workbench results pages, the first two of which are not described as they are identical in all cases. For more information on these two sections, please see their respective sections under the *multiple sequence alignment results* section.

The first section of the *motif search results* includes a summary of all detected motifs.

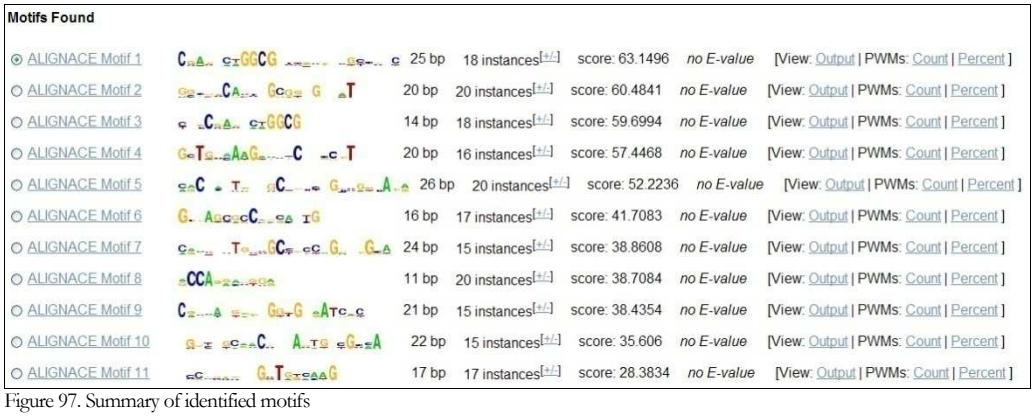
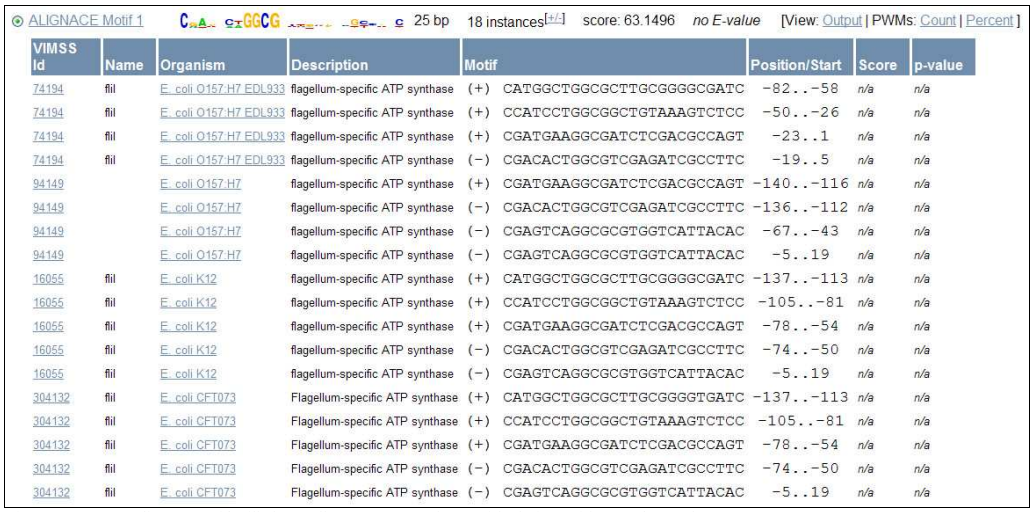


Figure 97. Summary of identified motifs

Motifs are displayed in tabular format and each motif is named with the selected motif search tool and numbered sequentially according to the order reported by the selected tool. Each motif is also shown graphically with its corresponding *sequence logo*. For more information on *sequence logos* visit `http://weblogo.berkeley.edu`. Next the length of the motif is displayed in base pairs along with the number of occurrences. Each tool generates different statistical measures of significance and if reported by the tool this information is included in the motif summary. Scores and E-values are generally not comparable across different tools. Finally, each motif presents a series of links for viewing the raw **Output** and for accessing the various *position weight matrices* (PWMs). All tools produce a **Count** PWM, which is simply the number of occurrences of each base for each position of the motif, and a **Percent** PWM, which is similar to the count PWM, but shows the percentage instead. *MEME* additionally computes a **Score** PWM showing the scoring metric for each base pair in each position of the motif.

The radio button shown to the left of each motif allows you to select a motif to use to scan other regions in a genome for additional sites. This function is described in greater detail in the *motif scans* section.

Clicking on the motif name or the **+/-** link next to the number of instances will expand that motif to show detailed information about each occurrence representing the motif.



ALIGNACE Motif 1 25 bp 18 instances[+/-] score: 63.1496 *no E-value* [View: Output | PWMs: Count | Percent]

VIMSS Id	Name	Organism	Description	Motif	Position/Start	Score	p-value
74194	fil	E. coli O157:H7 EDL933	flagellum-specific ATP synthase	(+) CATGGCTGGCGCTTGCGGGGCGATC	-82..-58	n/a	n/a
74194	fil	E. coli O157:H7 EDL933	flagellum-specific ATP synthase	(+) CCATCCTGGCGGCTGTAAAGTCTCC	-50..-26	n/a	n/a
74194	fil	E. coli O157:H7 EDL933	flagellum-specific ATP synthase	(+) CGATGAAGGCGATCTCGACGCCAGT	-23..1	n/a	n/a
74194	fil	E. coli O157:H7 EDL933	flagellum-specific ATP synthase	(-) CGACACTGGCGTCGAGATCGCCTTC	-19..5	n/a	n/a
94149		E. coli O157:H7	flagellum-specific ATP synthase	(+) CGATGAAGGCGATCTCGACGCCAGT	-140..-116	n/a	n/a
94149		E. coli O157:H7	flagellum-specific ATP synthase	(-) CGACACTGGCGTCGAGATCGCCTTC	-136..-112	n/a	n/a
94149		E. coli O157:H7	flagellum-specific ATP synthase	(-) CGAGTCAGGCGCGTGGTCATTACAC	-67..-43	n/a	n/a
94149		E. coli O157:H7	flagellum-specific ATP synthase	(-) CGAGTCAGGCGCGTGGTCATTACAC	-5..19	n/a	n/a
16055	fil	E. coli K12	flagellum-specific ATP synthase	(+) CATGGCTGGCGCTTGCGGGGCGATC	-137..-113	n/a	n/a
16055	fil	E. coli K12	flagellum-specific ATP synthase	(+) CCATCCTGGCGGCTGTAAAGTCTCC	-105..-81	n/a	n/a
16055	fil	E. coli K12	flagellum-specific ATP synthase	(+) CGATGAAGGCGATCTCGACGCCAGT	-78..-54	n/a	n/a
16055	fil	E. coli K12	flagellum-specific ATP synthase	(-) CGACACTGGCGTCGAGATCGCCTTC	-74..-50	n/a	n/a
16055	fil	E. coli K12	flagellum-specific ATP synthase	(-) CGAGTCAGGCGCGTGGTCATTACAC	-5..19	n/a	n/a
304132	fil	E. coli CFT073	Flagellum-specific ATP synthase	(+) CATGGCTGGCGCTTGCGGGGTGATC	-137..-113	n/a	n/a
304132	fil	E. coli CFT073	Flagellum-specific ATP synthase	(+) CCATCCTGGCGGCTGTAAAGTCTCC	-105..-81	n/a	n/a
304132	fil	E. coli CFT073	Flagellum-specific ATP synthase	(+) CGATGAAGGCGATCTCGACGCCAGT	-78..-54	n/a	n/a
304132	fil	E. coli CFT073	Flagellum-specific ATP synthase	(-) CGACACTGGCGTCGAGATCGCCTTC	-74..-50	n/a	n/a
304132	fil	E. coli CFT073	Flagellum-specific ATP synthase	(-) CGAGTCAGGCGCGTGGTCATTACAC	-5..19	n/a	n/a

Figure 98. Expanding a motif will show all occurrences along with their position and actual sequence

Each occurrence, or site, shows the reference gene used to create the input sequence on which the site was found, along with its name or gene symbol if defined, source organism, and description. Clicking on the reference gene's VIMSS id will redirect you to the *Locus Information* tool for the indicated gene. Similarly, clicking on the source organism name will redirect you to the *Genome Information* tool for the indicated organism. The actual motif and strand are shown in the **Motif** column followed by the offets relative to the start or end of the reference gene depending on the selected **Sequence Region**. Two additional columns, **Score** and **p-value**, show measures of statistical significance however not all tools report both values and some do not report statistical significance on a per-site basis, such as *AlignACE*.

Scan for Hits with Selected Motif

This form allows you to perform *motif scans*, or to search a selected genome for additional sites of the selected motif. This can help identify other genes that are potentially regulated by the same mechanism as one or more genes used to detect the motif. This is described in greater detail in the *motif scans* section.

Other Related Resources

This section contains links for viewing associated *motif search results*. Clicking on the **View other Motifs associated with this cart** link will show all *motif search results* built using the same gene cart as the current *motif search result*. Clicking on the **View all Motifs** link will show all *motif search results* built using genes from any of your saved carts.

Motif Scans

This tool is present from the *motif search results* viewer as its input depends in part on motifs identified using the *motif search* tool. For motifs generated using *MEME* or *Weeder*, we use *Patser* to scan genomes for sites matching the selected motif. For *AlignACE* motifs, we use *ScanACE*. Most of the parameters are the same, except as noted below.

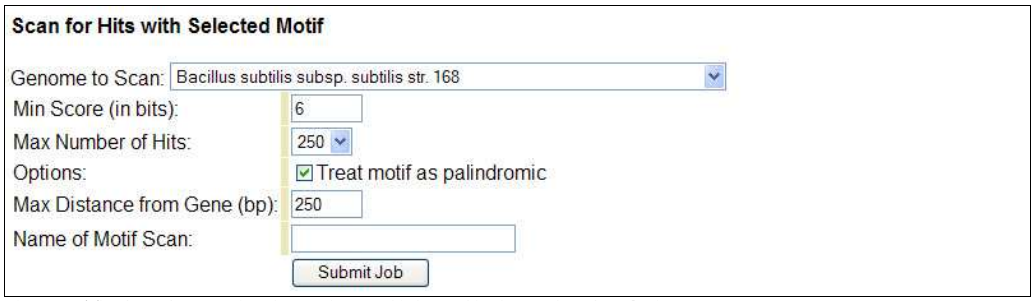


Figure 99. The form for accessing the *motif scan* tool is located in the *motif search results* viewer

First, you must select one of the detected motifs by clicking on the radio button to the left of its name. Then select the name of the genome you wish to scan under the **Genome to Scan** drop-down box.

The **Min Score (in bits)** determines the minimum bit-score threshold to consider a potential hit a site. Setting this value too low may result in spurious hits while setting the value too high may result in only near-perfect matches to the selected motif. If you used *AlignACE* to find motifs, this option is replaced by the **Threshold (std devs below average)**. In this case, a higher value means less significant hits will be considered as potential sites.

The **Max Number of Hits** allows you to limit the number of hits that will be returned. Try reducing this value if the selected motif generates too many hits, or try increasing the **Min Score (in bits)**.

The **Treat motif as palindromic** option should be checked if the identified motif is a palindrome as this eliminates the need to scan the reverse complement of the selected genome's sequence. However, if the identified motif is not a palindrome and this option is checked, we will not report hits on the reverse complement strand.

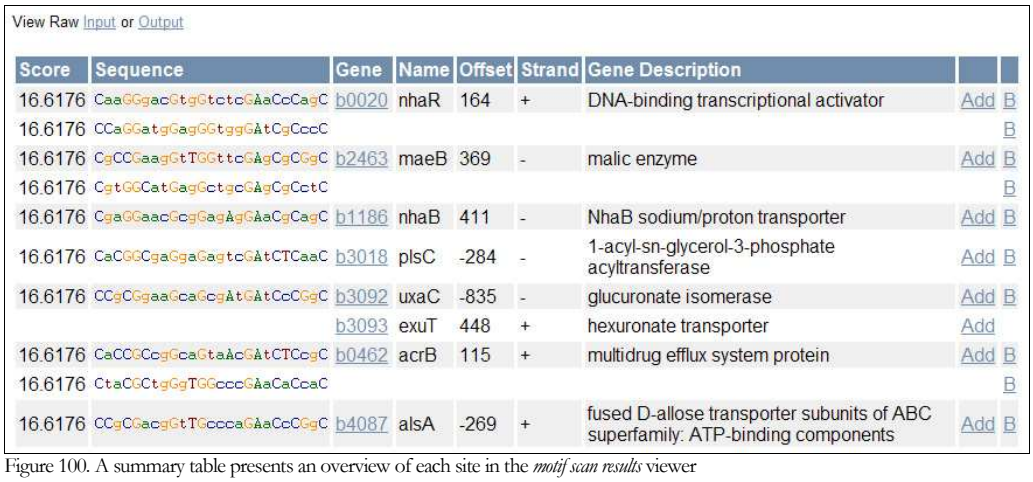
The **Max Distance from Gene (bp)** specifies the maximum distance a site can be from a gene to be reported.

Finally, you can assign a name to your *motif scan* job so that you can locate it later.

To start your *motif scan* job, click on the **Submit Job** button.

Motif Scan Results

Once your *motif scan* job has completed, you will be redirected to the *motif scan results* viewer, which will show a summary of all identified hits using the selected motif and genome, and tool parameters. If no sites were found, you will see **We were unable to locate any genes with the criteria specified for the motif scan**, and will have the option of viewing the raw **input** and **output** files.



View Raw Input or Output

Score	Sequence	Gene	Name	Offset	Strand	Gene Description		
16.6176	CaaGGgacGtgGtctcGAaCcCagC	b0020	nhaR	164	+	DNA-binding transcriptional activator	Add	B
16.6176	CCaGGatgGagGGtggGAtCgCccC							B
16.6176	CgCCGaagGtTGGttcGAgCgCGgC	b2463	maeB	369	-	malic enzyme	Add	B
16.6176	CgtGGCatGagGctgcGAgCgCctC							B
16.6176	CgaGGaacGcgGagAgGAaCgCagC	b1186	nhaB	411	-	NhaB sodium/proton transporter	Add	B
16.6176	CaCGGCgaGgaGagtcGAtCTCaaC	b3018	plsC	-284	-	1-acyl-sn-glycerol-3-phosphate acyltransferase	Add	B
16.6176	CCgCGgaaGcaGcgAtGAtCcCGgC	b3092	uxaC	-835	-	glucuronate isomerase	Add	B
		b3093	exuT	448	+	hexuronate transporter	Add	
16.6176	CaCCGCcgGcaGtaAcGAtCTCcgC	b0462	acrB	115	+	multidrug efflux system protein	Add	B
16.6176	CtaCGCtgGgTGGcccGAaCaCcaC							B
16.6176	CCgCGacgGtTGcccaGAaCcCGgC	b4087	alsA	-269	+	fused D-allose transporter subunits of ABC superfamily: ATP-binding components	Add	B

Figure 100. A summary table presents an overview of each site in the *motif scan results* viewer

You can view the raw input file containing the *position weight matrix* used to scan the selected genome by clicking on the **Input** link. The **Output** link allows you to view the raw results of the *motif scan* job. This can often provide useful information if a particular expected site is missing.

The first column if using *ScanACE* and the first two columns if using *Patser* give metrics of the statistical significance of the site. In addition to the bit-score of the hit, *Patser* will also return an E-value.

The **Sequence** column shows the actual hit sequence colored by nucleotide as in the sequence logo. Positions matching the logo are displayed in upper-case.

The **Gene**, **Name**, **Offset**, **Strand**, and **Gene Description** columns describe the gene closest to the site if it falls within the threshold specified as a parameter to the *motif scan* job. **Gene** contains the gene's systematic name or locus tag and is also a hyperlink to the *Locus Information* tool for the indicated gene. The **Name** column shows the gene's name or symbol and the **Offset** and **Strand** indicate the position of the site with respect to the reference gene. The **Gene Description** contains a description of the gene.

Some sites are found near two or more genes, such as the site associated with the gene **b3092** and **b3093** from Figure 100. In these cases, the bit-score and/or e-value, and sequence are only displayed for the first gene.

Each hit associated with a gene also has an **Add** link which will add the gene to your current *session gene cart*. Each hit also contains a **B** link which allows you to view the hit in context of surrounding genomic features in the *Ortholog Browser*.

Cart Expression Viewer

This tool is similar to the *gene expression viewer* except that it shows only experimental results for the genes in the selected *gene cart*. Please see the *gene expression viewer* section for more information on navigating and interpreting the heatmap and associated links.

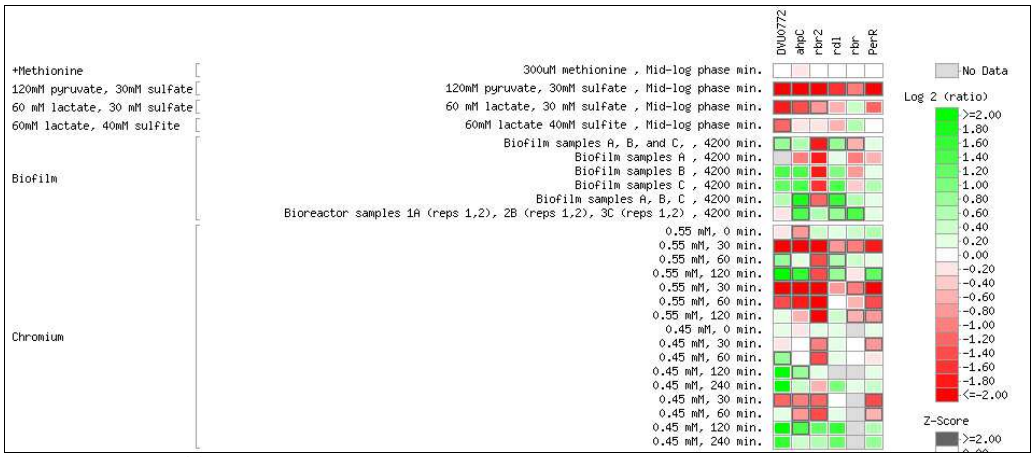


Figure 101. Heatmap of the *cart expression viewer*

Job List Summary

All of the **View all …** links present throughout the *bioinformatics workbench* link to the *job list summary* for the indicated job type showing the subset of jobs indicated. Each job type will display slightly different information in its *job list summary* and these differences are described below for each job type.

There are currently four different job result types: multiple sequence alignments, phylogenetic (or gene) trees, motifs, and motif scans. The *job list summary* is divided into two sections. The first lists the jobs that match the subset criteria indicated by the **View all …** link. The second section, labeled *job actions*, lists relevant links for managing the jobs in the indicated subset.

Job ID	Cart	MSA Name	Sequence	Status	Generated	Saved	Actions
5041	New 06/06/2006	MSA New 06/06/2006	AA/gene	DONE	05-Mar-2008 13:51	no	View | Delete | ACL
4928	PerR (D. vul Hildenborough)	MSA PerR (D. vul Hildenborough)	NA/Gene	DONE	27-Feb-2008 13:29	no	View | Delete | ACL
4906	PerR (D. vul Hildenborough)	MSA PerR (D. vul Hildenborough)	AA/gene	DONE	27-Feb-2008 10:04	no	View | Delete | ACL
4905	PerR (D. vul Hildenborough)	MSA PerR (D. vul Hildenborough)	AA/gene	DONE	27-Feb-2008 10:03	no	View | Delete | ACL
4902	New 06/06/2006	MSA New 06/06/2006	AA/gene	DONE	26-Feb-2008 13:19	no	View | Delete | ACL
2895	New 06/06/2006		AA/gene	DONE	06-Jun-2006 15:57	no	View | Delete | ACL

Figure 102. List of jobs from the *job list summary*

Several columns in the *job list summary* are always displayed regardless of job type. These include the **Job ID**, name of the job's associated gene **Cart**, the name of the job (**MSA Name** in the above example), the job's execution **Status**, the date the job results were **Generated**, whether or not the job has been **Saved**, and **Actions**.

Most of these labels in the table header are clickable to sort the list of jobs. Only ascending sort is supported at this time.

In addition, job type-specific parameters are often displayed. For example, for multiple sequence alignment job results, the list includes the **Sequence** region and type used to build the multiple sequence alignment. For gene trees, this would include the tree building method and parameters used as well as any sequence trimming.

If a job does not have an associated gene cart, **n/a** is shown in place of the cart's name. Clicking on the gene cart name will redirect you to the *gene cart viewer* for the selected gene cart.

The **Action** links apply to the associated job indicated in the same row. The **View** link will redirect you to the *job results viewer* for the selected job. Clicking the **Delete** link will remove the indicated job but will not remove dependent jobs. For example, removing a multiple sequence alignment will not remove gene trees created from the deleted multiple sequence alignment. Deleting a job is permanent and once deleted cannot be undeleted. There is no confirmation for deleting job results. Finally, clicking on the **ACL** link will redirect you to the *add resource permissions* interface for the indicated job result.

In the second **Job Actions** section, a number of bulk job deletion links allow you to remove large subsets of jobs as indicated. These links do not ask for confirmation so take care in using these actions. If you are viewing a subset of job results, such as those associated with a specific cart, you will also see a **View All …** link which will allow you to see all jobs.

Resource Access Control

All your gene carts and derived job results are considered "resources". By default, resources you create can only be accessed using your account however you may choose to grant access to another user or group. To do this, you must use the *resource access control* interface. There are two versions of this interface. The first is present on all job result viewers and gives you a quick interface for changing *existing* permissions only. The second interface allows more detailed assignment of rights.

Current Resource Permissions

The *current resource permissions* are displayed on every job result, beneath the *save job results* section. This tool provides a quick overview of all existing permissions on a selected resource and is accessible by viewing the job results for the desired resource.

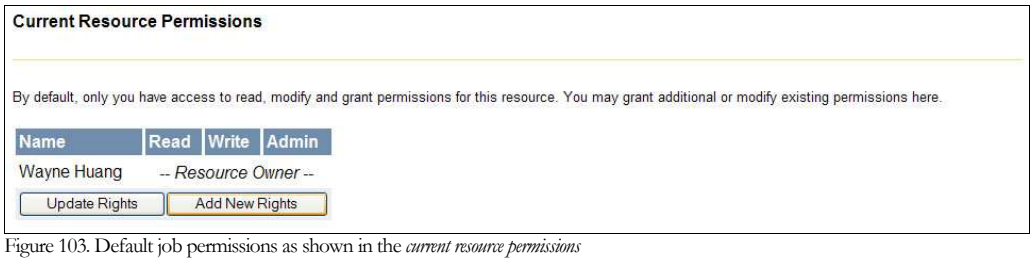


Figure 103. Default job permissions as shown in the *current resource permissions*

By default, newly created job results are only accessible by the job result's owner. The job owner's name will appear along with the text **– Resource Owner –**. In this case, the **Update Rights** button, which is meant to quickly change existing permissions, has no function, however clicking on the **Add New Rights** button will allow new permissions to be added using the *add resource permissions* interface.

When a particular resource has additional permissions, the *current resource permissions* will show the other users' or groups' names along with their respective permissions.

Figure 104. *Current resource permissions* showing additional rights assigned to users and groups

In this case, you may change the permissions for any of the listed users or groups by checking or unchecking the appropriate checkboxes. **Read** access allows another user or group to view the results only, while **Write** access allows them to change display parameters of the selected job. Assigning **Admin** rights allows the designated user or

group to remove the results as well as to grant additional permissions to other users or groups. **Admin** rights always imply both **Read** and **Write**.

To apply the changes, click on the **Update Rights** button.

Add Resource Permissions

This tool allows you to modify and/or add new permissions to any resource and is divided into four sections. To access this tool, click on the **Add New Rights** button displayed under *current resource permissions* from any job results page, or by clicking the **ACL** link from the *gene cart summary* or from any *job list summary* page.

Regardless of the resource type, a summary will show the name of the resource and its type. Clicking on the resource name will allow you to view the details of the resource, whether it is a gene cart or a job result.

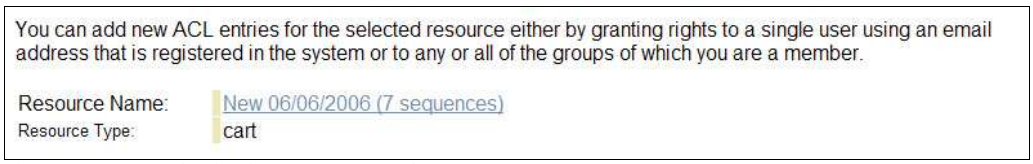


Figure 105. Summary showing the resource name and type

The *current resource permissions* interface shown on all job results pages is displayed again and you can make identical adjustments to existing permissions as described above.

The next section, labeled *grant rights to individual user*, allows you to add access to the current resource for an individual user by specifying his or her email address and the permissions they should have. By default **Read** is assigned since it does not make sense to grant rights without specifying any permissions. You may optionally assign **Write** or **Admin** access as you desire. To add the new permissions you must click on the **Grant User Rights** button.

Figure 106. Interface for granting rights to an individual user

The last section, labeled *grant rights to groups*, allows you to add access to the current resource for a group of users. Groups are defined by VIMSS administrators so you should contact us by clicking on the **Contact Us** link in the footer of most pages if you would like to create a new group. Select the **Group Name** drop down box and select

the group for which you wish to add access. Next, select the rights you wish to give to the selected group and click on the **Grant Group Rights** button to save your changes.

Figure 107. Interface for granting rights to a group

To remove permissions, see the *current source permissions* section. Unchecking **Read**, **Write**, and **Admin** access for a listed user or group and clicking on **Update Rights** will remove access for the indicated user or group.

Account Settings

This page allows you to modify the information you used to register your account, including your name, organization, email address and preferences, and password. Each registered account in MicrobesOnline must have a unique email address, therefore it is not possibly to register multiple accounts using a shared group email address.

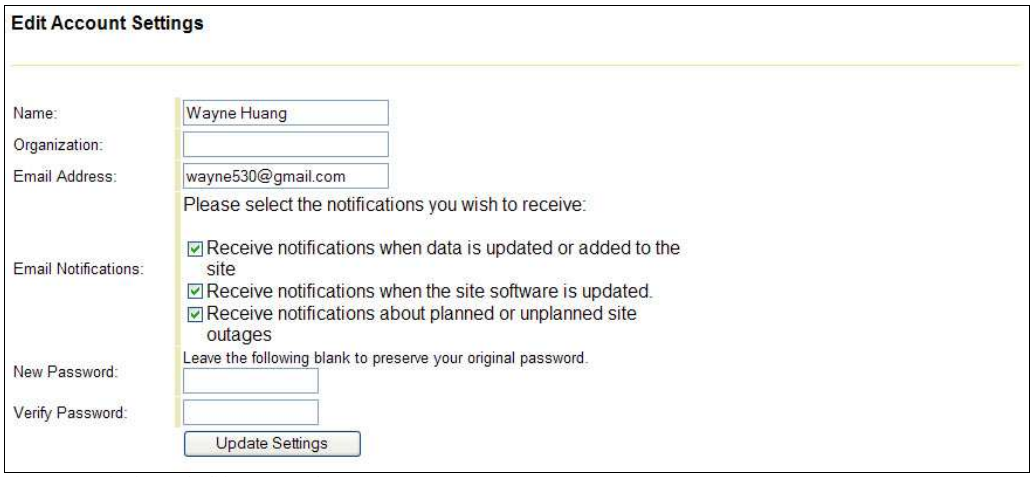


Figure 108. Viewing and editing *Account Settings*

The MicrobesOnline team will send periodic emails to notify our users of important information regarding the site. This information is divided into three categories: data updates, software updates, and site outages. Select the checkboxes corresponding to the categories of notifications you wish to receive. You can change these settings later by returning to the *account settings* page.

Finally, if you wish to change your password you may do so by specifying a new password and verifying it. We require the verification so that you don't accidentally lock yourself out of your account because you made a typo while specifying your new password. If you do not wish to change your password, leave both boxes empty.

Click on the *Update Settings* button to save any changes to your account profile.

Index

No index entries found.