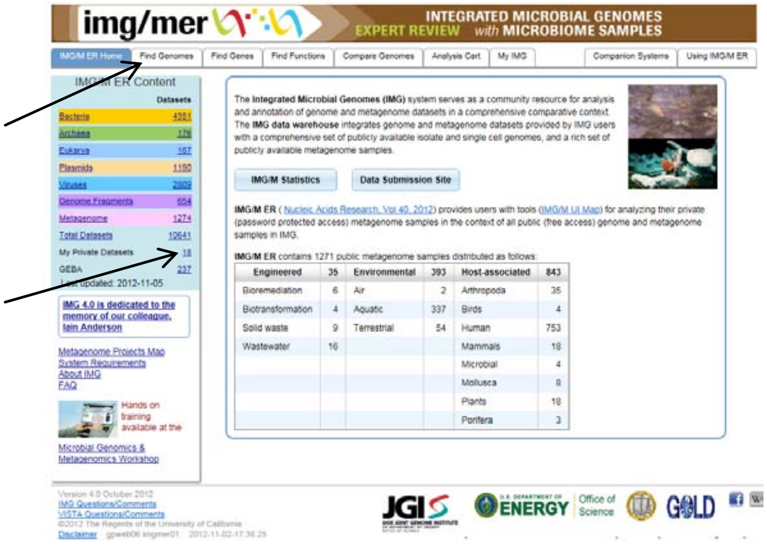
JGI Microbial Single Cell Program
Single Cell Data Decontamination

Author: Scott Clingenpeel

Despite our best efforts, it is likely that there are some contigs in your single cell genome(s) that are from contaminant organisms. Common contaminants that are known to be in the reagents we purchase are *Delftia*, *Pseudomonas*, and *Ralstonia*. Other contaminants that we commonly see are *Propionibacterium* and *Lactobacillus*. In addition, there may be contaminants from your particular sample in the form of free DNA that made it into the well along with your single cell. Although we do an automated screen of your data for the known common contaminants and this information is provided to you in the JGI Single-cell Assembly QC report, this data is not removed because this could result in the removal of legitimate, highly conserved genes from your genome. Thus, it is necessary for you to do a manual screening of your data to remove contaminant sequences. While there are no clear rules on the identification and removal of contamination (i.e. phage or horizontal gene transfer may be difficult to discriminate from contamination), we would like to provide some recommendations and guidance.

JGI warrants a 6-month review period for single cell genome(s) sequenced at the JGI, during which you can curate and clean the data set(s). After 6 months, the data will be released to the public.

When you first log into the IMG/MER system there are two ways to find your genomes.



IMG/M ER Content

	Datasets
Bacteria	4351
Archaea	176
Eukarya	167
Plasmids	1190
Viruses	2809
Genome Fragments	654
Metagenome	1274
Total Datasets	10641
My Private Datasets	18
GEBA	237

Last updated: 2012-11-05

IMG 4.0 is dedicated to the memory of our colleague, Iain Anderson

The **Integrated Microbial Genomes (IMG)** system serves as a community resource for analysis and annotation of genome and metagenome datasets in a comprehensive comparative context. The **IMG data warehouse** integrates genome and metagenome datasets provided by IMG users with a comprehensive set of publicly available isolate and single cell genomes, and a rich set of publicly available metagenome samples.

IMG/M ER (Nucleic Acids Research, Vol 40, 2012) provides users with tools (IMG/M UI Map) for analyzing their private (password protected access) metagenome samples in the context of all public (free access) genome and metagenome samples in IMG.

IMG/M ER contains 1271 public metagenome samples distributed as follows:

Engineered	35	Environmental	393	Host-associated	843
Bioremediation	6	Air	2	Arthropoda	35
Biotransformation	4	Aquatic	337	Birds	4
Solid waste	9	Terrestrial	54	Human	753
Wastewater	16			Mammals	18
				Microbial	4
				Mollusca	8
				Plants	18
				Porifera	3

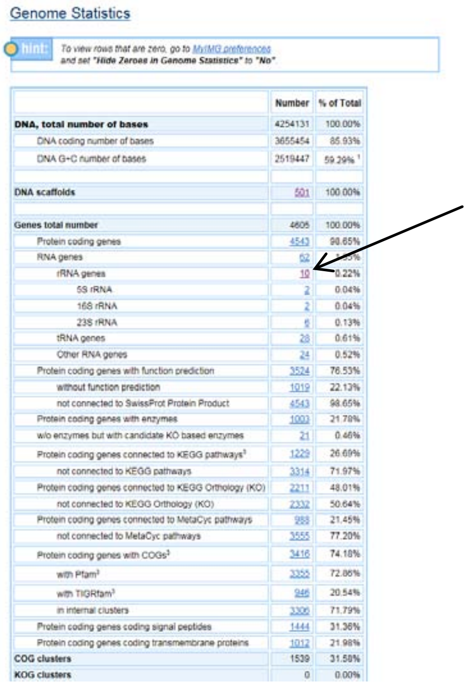
Version 4.0 October 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer gpweb06 imgmer01 2012-11-02-17:38:25

Under the Find Genomes tab you can search for any genome in IMG by a variety of criteria.

Your own genomes should be under this link.

Click on the name of the genome that you want in order to bring up the genome overview page.

A good place to start with contamination screening is to look at any ribosomal RNA sequences in your genome. When you pull up the overview of your genome and scroll down, you will see some genome statistics.

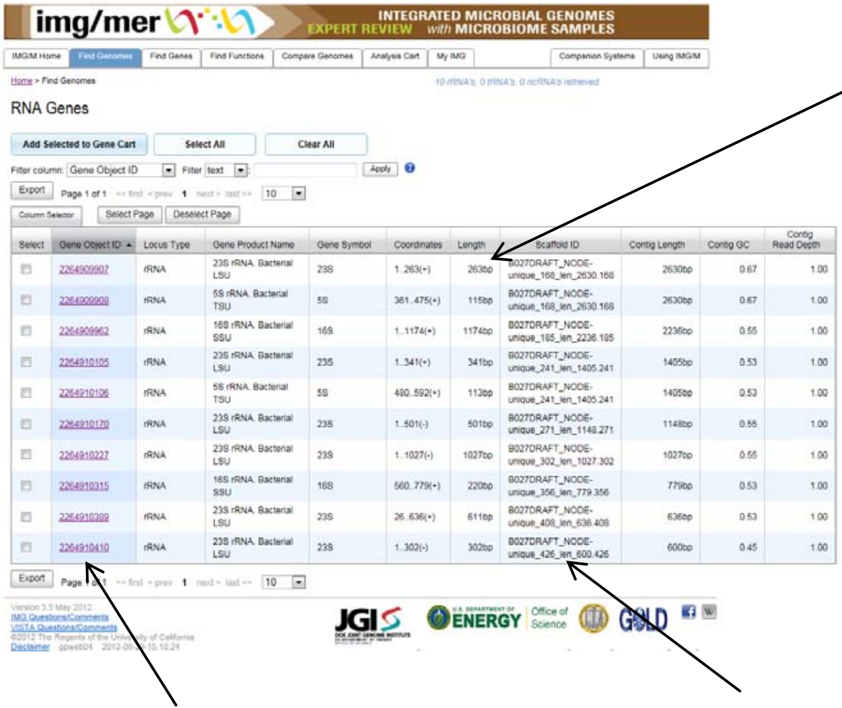


Genome Statistics

To view rows that are zero, go to MyIMG preferences and set "Hide Zeroes in Genome Statistics" to "No"

	Number	% of Total
DNA, total number of bases	4254131	100.00%
DNA coding number of bases	3655454	85.93%
DNA G+C number of bases	2519447	59.29% [1]
DNA scaffolds	501	100.00%
Genes total number	4605	100.00%
Protein coding genes	4543	98.65%
RNA genes	62	1.35%
rRNA genes	10	0.22%
5S rRNA	2	0.04%
16S rRNA	2	0.04%
23S rRNA	6	0.13%
tRNA genes	28	0.61%
Other RNA genes	24	0.52%
Protein coding genes with function prediction	3524	76.53%
without function prediction	1019	22.13%
not connected to SwissProt Protein Product	4543	98.65%
Protein coding genes with enzymes	1003	21.78%
w/o enzymes but with candidate KO based enzymes	21	0.46%
Protein coding genes connected to KEGG pathways[3]	1229	26.69%
not connected to KEGG pathways	3314	71.97%
Protein coding genes connected to KEGG Orthology (KO)	2211	48.01%
not connected to KEGG Orthology (KO)	2332	50.64%
Protein coding genes connected to MetaCyc pathways	988	21.45%
not connected to MetaCyc pathways	3555	77.20%
Protein coding genes with COGs[3]	3416	74.18%
with Pfam[3]	3355	72.86%
with TIGRfam[3]	946	20.54%
in internal clusters	3306	71.79%
Protein coding genes coding signal peptides	1444	31.36%
Protein coding genes coding transmembrane proteins	1012	21.98%
COG clusters	1539	31.58%
KOG clusters	0	0.00%

Clicking here will bring up all your rRNA genes



Home > Find Genomes

10 rRNA's, 0 tRNA's, 0 ncRNA's retrieved.

RNA Genes

Add Selected to Gene Cart
Select All
Clear All
Filter column: Gene Object ID
Filter text
Apply
Export
Page 1 of 1 << first < prev 1 next > last >> 10
Column Selector
Select Page
Deselect Page

Select	Gene Object ID ▲	Locus Type	Gene Product Name	Gene Symbol	Coordinates	Length	Scaffold ID	Contig Length	Contig GC	Contig Read Depth
□	2264909907	rRNA	23S rRNA, Bacterial LSU	23S	1..263(+)	263bp	B027DRAFT_NODE-unique_168_len_2630.168	2630bp	0.67	1.00
□	2264909908	rRNA	5S rRNA, Bacterial TSU	5S	361..475(+)	115bp	B027DRAFT_NODE-unique_168_len_2630.168	2630bp	0.67	1.00
□	2264909962	rRNA	16S rRNA, Bacterial SSU	16S	1..1174(+)	1174bp	B027DRAFT_NODE-unique_185_len_2236.185	2236bp	0.55	1.00
□	2264910105	rRNA	23S rRNA, Bacterial LSU	23S	1..341(+)	341bp	B027DRAFT_NODE-unique_241_len_1405.241	1405bp	0.53	1.00
□	2264910106	rRNA	5S rRNA, Bacterial TSU	5S	480..592(+)	113bp	B027DRAFT_NODE-unique_241_len_1405.241	1405bp	0.53	1.00
□	2264910170	rRNA	23S rRNA, Bacterial LSU	23S	1..501(-)	501bp	B027DRAFT_NODE-unique_271_len_1148.271	1148bp	0.55	1.00
□	2264910227	rRNA	23S rRNA, Bacterial LSU	23S	1..1027(-)	1027bp	B027DRAFT_NODE-unique_302_len_1027.302	1027bp	0.55	1.00
□	2264910315	rRNA	16S rRNA, Bacterial SSU	16S	560..779(+)	220bp	B027DRAFT_NODE-unique_356_len_779.356	779bp	0.53	1.00
□	2264910399	rRNA	23S rRNA, Bacterial LSU	23S	26..636(+)	611bp	B027DRAFT_NODE-unique_408_len_636.408	636bp	0.53	1.00
□	2264910410	rRNA	23S rRNA, Bacterial LSU	23S	1..302(-)	302bp	B027DRAFT_NODE-unique_426_len_600.426	600bp	0.45	1.00

Version 3.5 May 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer gpweb04 2012-09-28 15.10.24

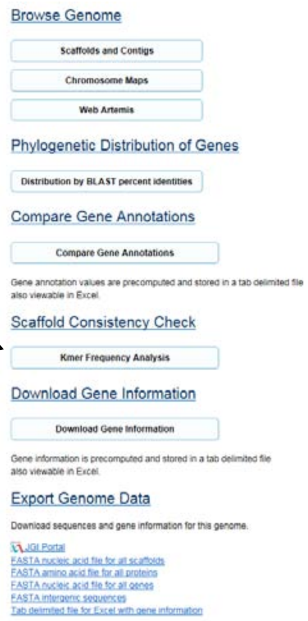
If the length of the gene is too short then it will not be phylogenetically informative.

Clicking the links in this column will allow you to retrieve the sequence of the gene.

This column tells you which scaffold each gene is located on.

BLAST these rRNA sequences to see if they come from your target genome or from a contaminant. In this example, rRNA genes from scaffolds 185, 302, and 408 match Rhizobium genes, which are the target organism. Scaffolds 168, 271, and 356 contain rRNA genes that match Caulobacter and scaffolds 241 and 426 contain rRNA genes that match Betaproteobacteria. These five scaffolds should be removed as contaminants. Note the scaffold numbers that contain rRNA genes as this will be important in the next step.

Now go back to the genome overview and scroll down further than before. Click on the button for Kmer Frequency Analysis.



Browse Genome

Scaffolds and Contigs

Chromosome Maps

Web Artemis

Phylogenetic Distribution of Genes

Distribution by BLAST percent identities

Compare Gene Annotations

Compare Gene Annotations

Gene annotation values are precomputed and stored in a tab delimited file also viewable in Excel

Scaffold Consistency Check

Kmer Frequency Analysis

Download Gene Information

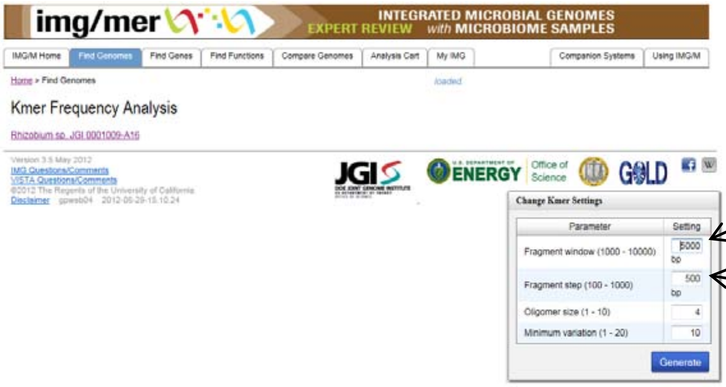
Download Gene Information

Gene information is precomputed and stored in a tab delimited file also viewable in Excel

Export Genome Data

Download sequences and gene information for this genome.

JGI Portal
FASTA nucleic acid file for all scaffolds
FASTA amino acid file for all proteins
FASTA nucleic acid file for all genes
FASTA intergenic sequences
Tab delimited file for Excel with gene information



Home > Find Genomes

Kmer Frequency Analysis

Rhizobium sp. JGI 0001009-A16

Version 3.5 May 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer gpweb04 2012-05-29-15.10.24

Change Kmer Settings

Parameter	Setting
Fragment window (1000 - 10000)	5000 bp
Fragment step (100 - 1000)	500 bp
Oligomer size (1 - 10)	4
Minimum variation (1 - 20)	10

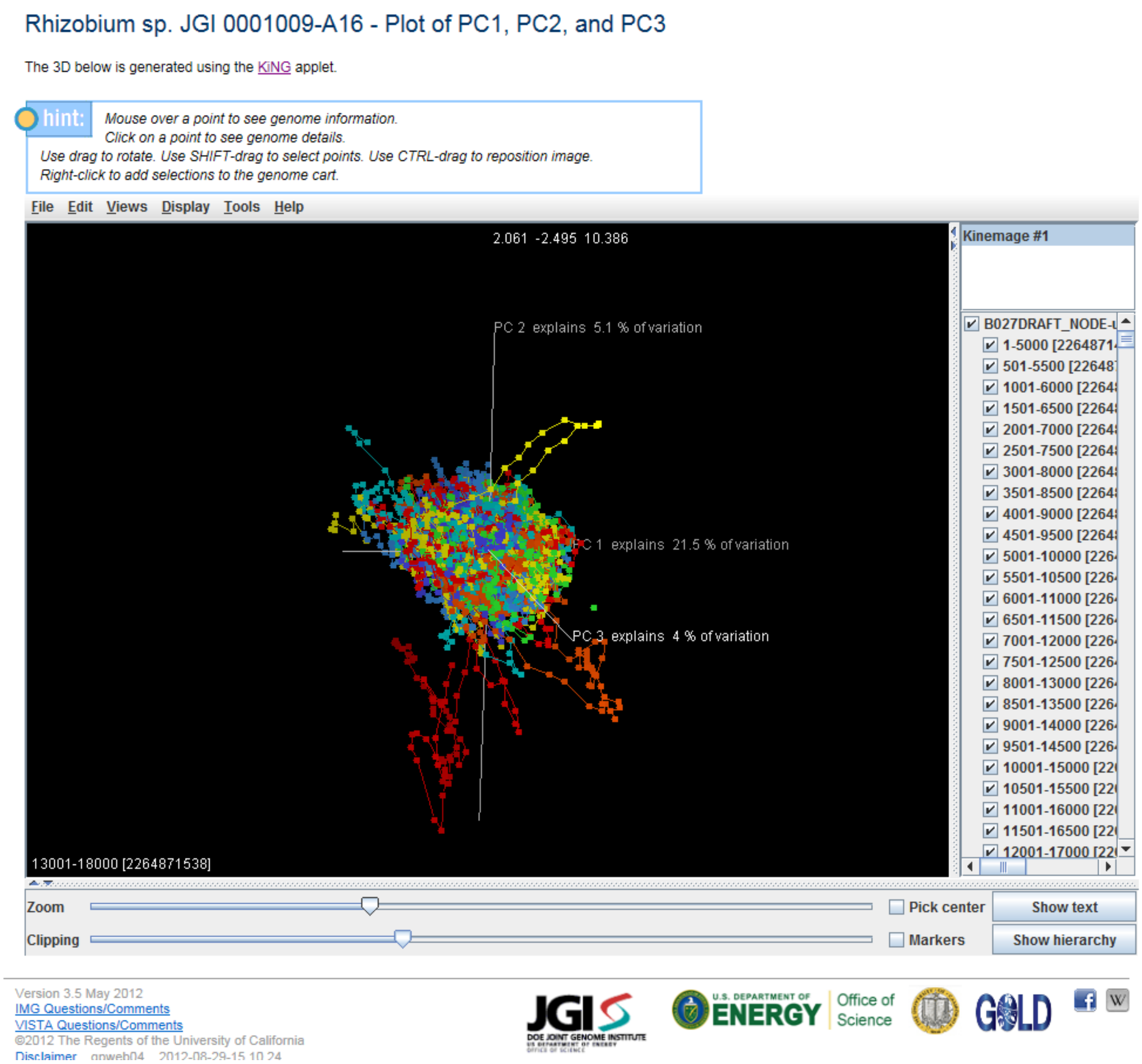
Generate

This number is the minimum size of scaffold to include in the analysis.

This number indicates that a point will be plotted for every 500bp in the scaffold.

Begin by sticking with the defaults. The larger scaffolds have more statistical power, which will produce a more defined cloud of points. Also, it is easier to get a feel for the data with the few large scaffolds than if you included all of the data. Later you will want to rerun this analysis with a smaller fragment window to include all your scaffolds in the screen.

Clicking the Generate button will produce a Kmer plot. There is both a 2D and a 3D view. We find the 3D view to be the most useful.



Rhizobium sp. JGI 0001009-A16 - Plot of PC1, PC2, and PC3

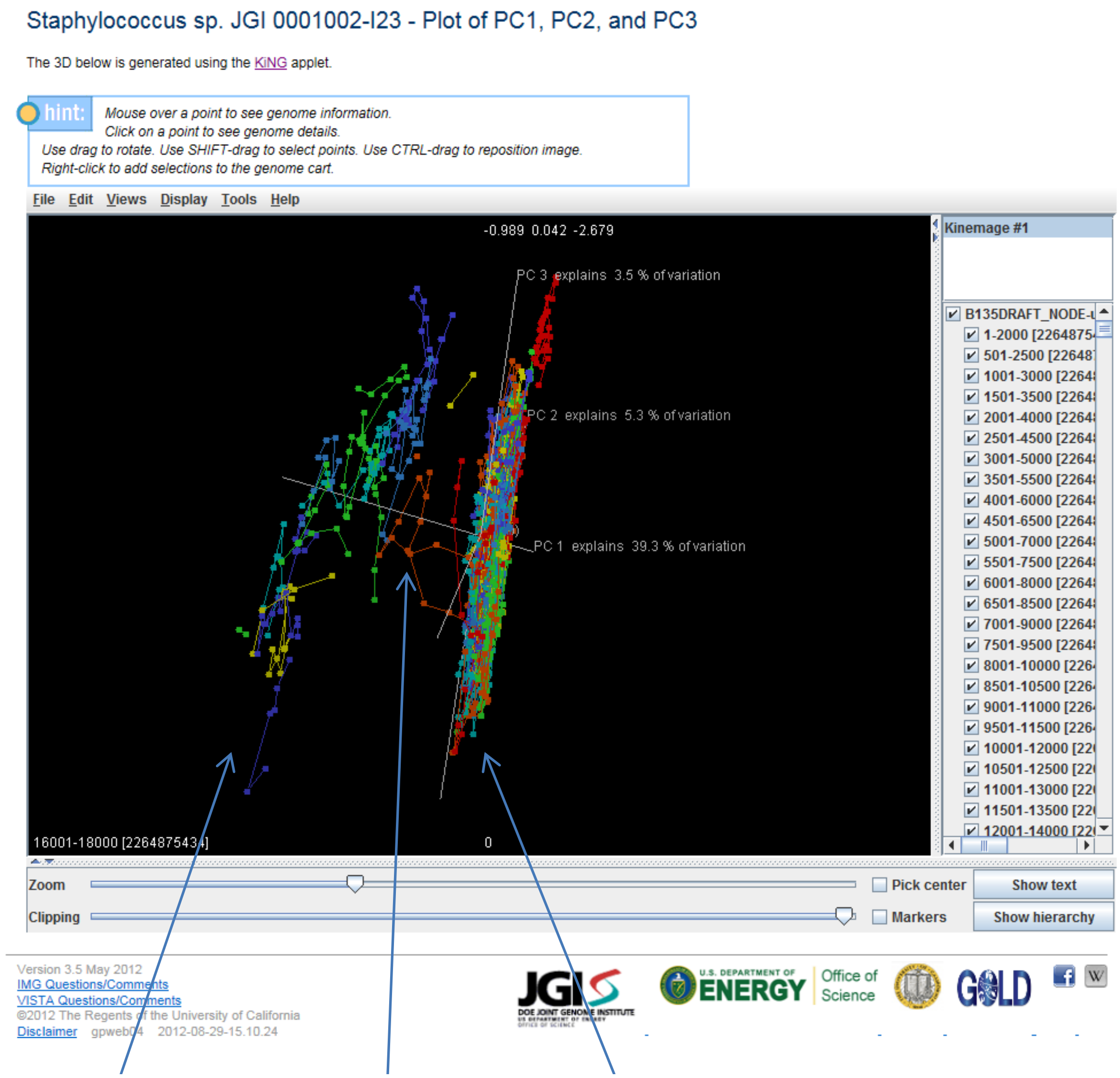
The 3D below is generated using the KiNG applet.

hint: Mouse over a point to see genome information.
Click on a point to see genome details.
Use drag to rotate. Use SHIFT-drag to select points. Use CTRL-drag to reposition image.
Right-click to add selections to the genome cart.

Version 3.5 May 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer gpweb04 2012-08-29-15 10 24

You can click and drag on the image to rotate it in three dimensions. First, look at the percent of variation explained by each principal component. If the percentages are all small (<~5%) then you have a very clean genome and the outliers are unlikely to be a problem.

Below is a fairly contaminated genome. Most points are in a large mass which is our target genome, but there is a distinct cloud of contaminant scaffolds to the left of the main cloud. By clicking on any of the points in the plot it will open a separate window of that scaffold.



Staphylococcus sp. JGI 0001002-I23 - Plot of PC1, PC2, and PC3

The 3D below is generated using the KiNG applet.

hint: Mouse over a point to see genome information.
Click on a point to see genome details.
Use drag to rotate. Use SHIFT-drag to select points. Use CTRL-drag to reposition image.
Right-click to add selections to the genome cart.

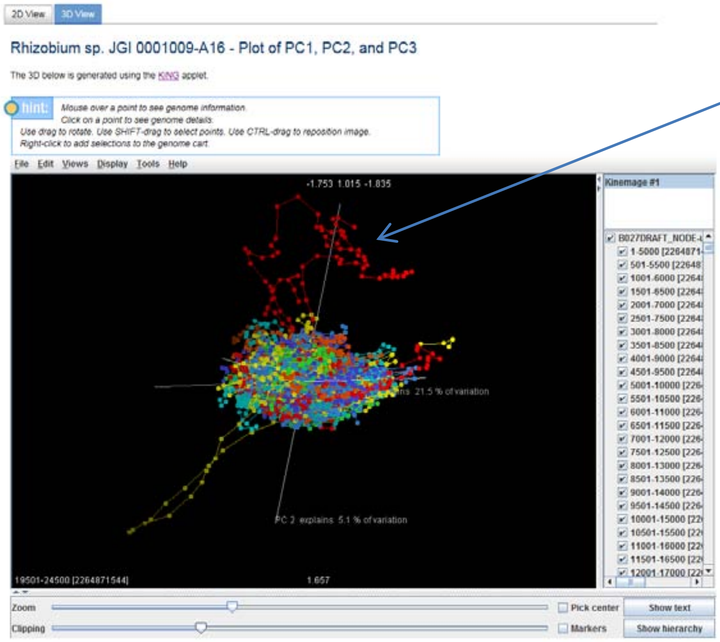
Version 3.5 May 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer gpweb04 2012-08-29-15.10.24

Potential Contaminants

Orange Scaffold

Target Organism

Note the Orange scaffold. This one starts in the main cloud, extends into the contaminant zone, and then returns to the main cloud. Upon examining this scaffold, we find that the region that extends out from the main cloud contains rRNA genes that match the target organism. Ribosomal RNA genes often contain a different GC content from the rest of the genome and thus will plot outside the main cloud of your target genome. Scaffolds that extend from the main genome cloud can also contain other interesting features.



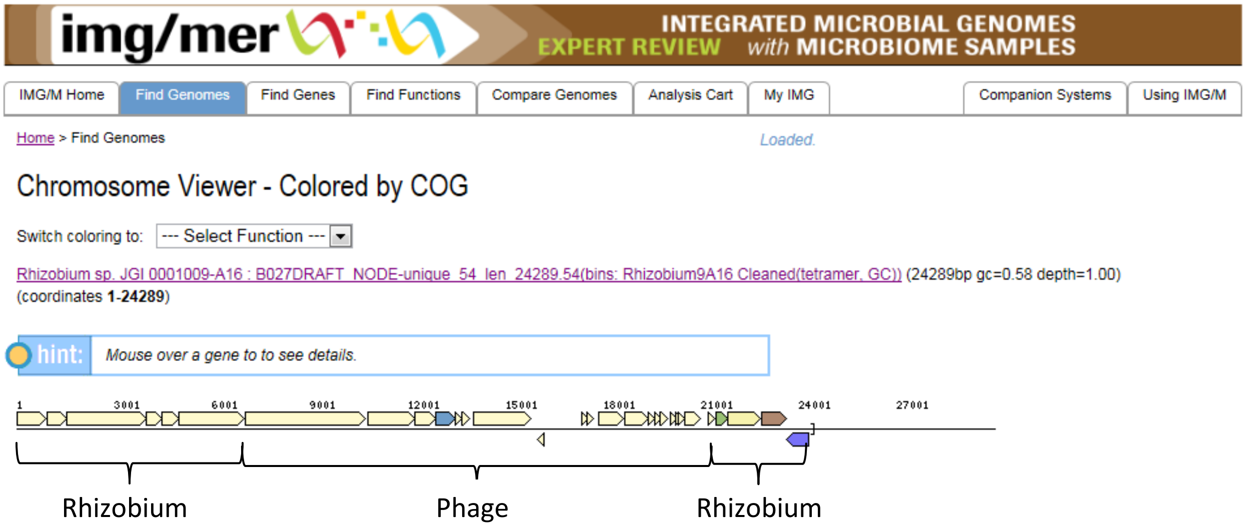
Rhizobium sp. JGI 0001009-A16 - Plot of PC1, PC2, and PC3

The 3D below is generated using the KING applet.

hint: *Mouse over a point to see genome information.*
Click on a point to see genome details.
Use drag to rotate. Use SHIFT-drag to select points. Use CTRL-drag to reposition image.
Right-click to add selections to the genome cart.

This red scaffold has points in the main cloud but extends well out.

Clicking on the points in this scaffold opens a separate window with more detail on the scaffold shown below.



Home > Find Genomes

Chromosome Viewer - Colored by COG

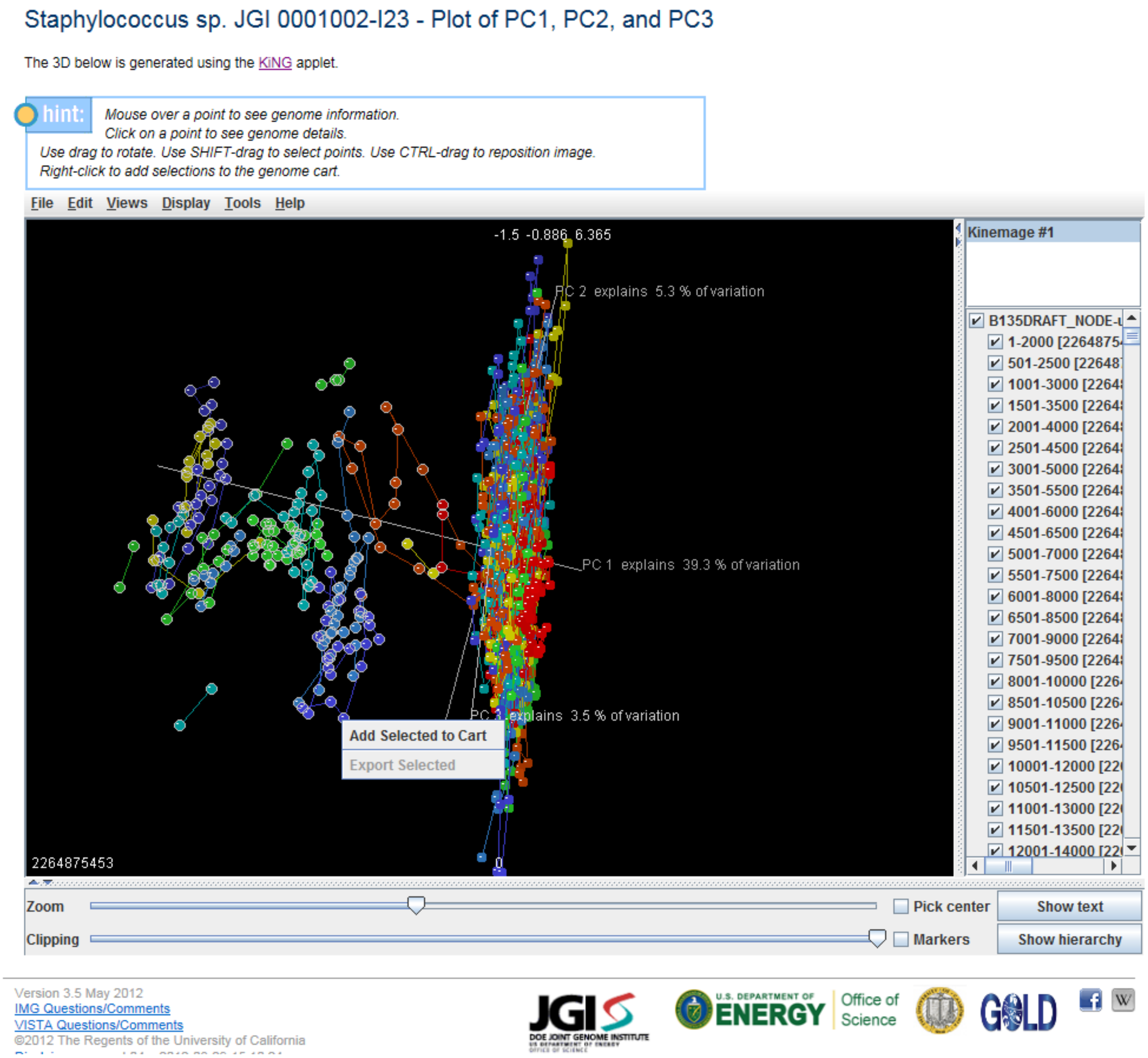
Switch coloring to: --- Select Function ---

Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_54_len_24289.54(bins: Rhizobium9A16 Cleaned(tetramer, GC)) (24289bp gc=0.58 depth=1.00)
(coordinates **1-24289**)

hint: *Mouse over a gene to to see details.*

This scaffold is from a Rhizobium single cell and it contains 30 predicted genes. By doing a BLASTx search on each of the genes we found that the ends have high matches to proteins from various Rhizobium species. However, the genes in the middle, which caused the scaffold to stick out from the main cloud in the plot, have best matches to phage proteins. This cell appears to be infected with a lysogenic phage.

When a scaffold is suspicious based on the plot you need to verify if it is a contaminant by doing BLAST searches. Left-clicking a point on the plot will open a separate window for the scaffold. Right clicking a point will add that scaffold to your scaffold cart. By holding shift while you click and drag you can select multiple scaffolds and then right click to add them to the scaffold cart.



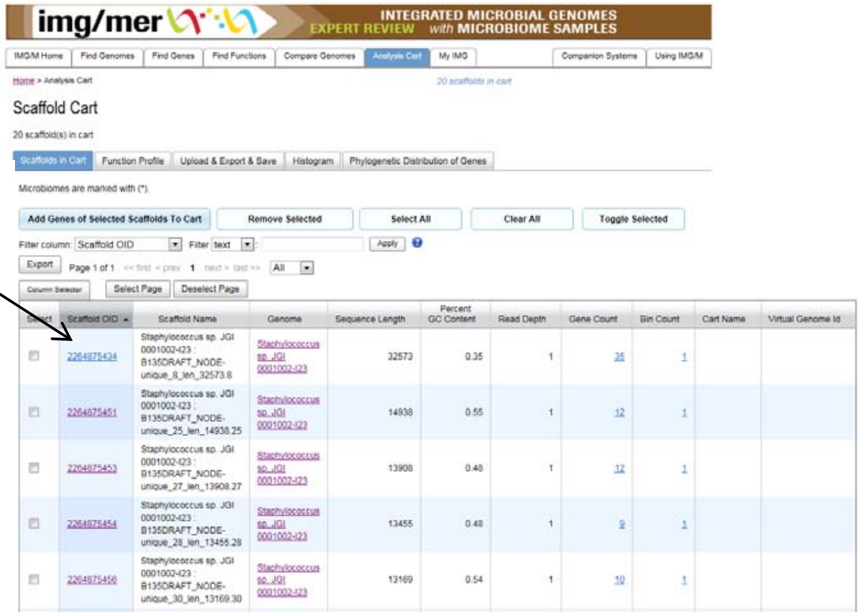
Staphylococcus sp. JGI 0001002-I23 - Plot of PC1, PC2, and PC3

The 3D below is generated using the KiNG applet.

hint: *Mouse over a point to see genome information.*
Click on a point to see genome details.
Use drag to rotate. Use SHIFT-drag to select points. Use CTRL-drag to reposition image.
Right-click to add selections to the genome cart.

Version 3.5 May 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California

From the scaffold cart select a scaffold.



Scaffold Cart

20 scaffold(s) in cart

Microbiomes are marked with (*).

Filter column: Scaffold OID Filter text

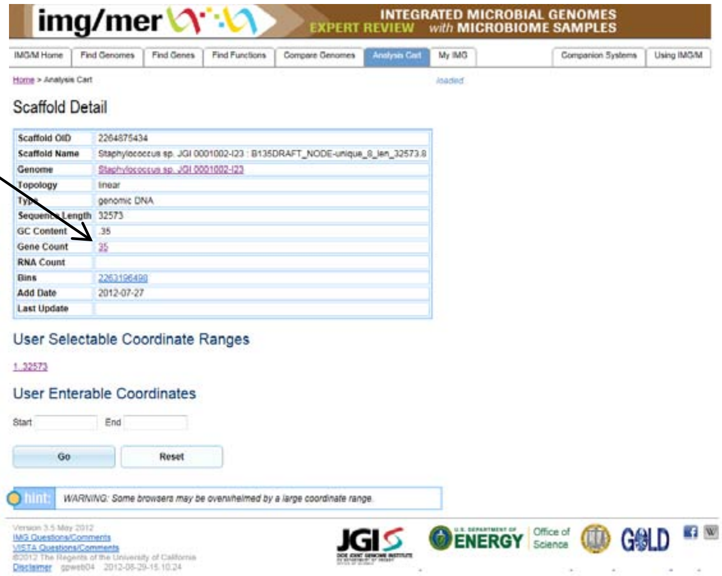
Page 1 of 1 << first < prev 1 next > last >> All

Select	Scaffold OID	Scaffold Name	Genome	Sequence Length	Percent GC Content	Read Depth	Gene Count	Bin Count	Cart Name	Virtual Genome Id
□	2264875434	Staphylococcus sp. JGI 0001002-I23 : B135DRAFT_NODE-unique_8_len_32573.8	Staphylococcus sp. JGI 0001002-I23	32573	0.35	1	35	1		
□	2264875451	Staphylococcus sp. JGI 0001002-I23 : B135DRAFT_NODE-unique_25_len_14938.25	Staphylococcus sp. JGI 0001002-I23	14938	0.55	1	12	1		
□	2264875453	Staphylococcus sp. JGI 0001002-I23 : B135DRAFT_NODE-unique_27_len_13908.27	Staphylococcus sp. JGI 0001002-I23	13908	0.48	1	12	1		
□	2264875454	Staphylococcus sp. JGI 0001002-I23 : B135DRAFT_NODE-unique_28_len_13455.28	Staphylococcus sp. JGI 0001002-I23	13455	0.48	1	9	1		
□	2264875456	Staphylococcus sp. JGI 0001002-I23 : B135DRAFT_NODE-unique_30_len_13169.30	Staphylococcus sp. JGI 0001002-I23	13169	0.54	1	10	1		

Click here to go to the scaffold.

Click here to get a list of genes on the scaffold. From there you can select each gene to get its sequence for BLASTing.

Click here to get graphical depiction of the scaffold (see below).



Scaffold Detail

Scaffold OID	2264875434
Scaffold Name	Staphylococcus sp. JGI 0001002-I23 : B135DRAFT_NODE-unique_8_len_32573.8
Genome	Staphylococcus sp. JGI 0001002-I23
Topology	linear
Type	genomic DNA
Sequence Length	32573
GC Content	.35
Gene Count	35
RNA Count	
Bins	2263196490
Add Date	2012-07-27
Last Update	

User Selectable Coordinate Ranges

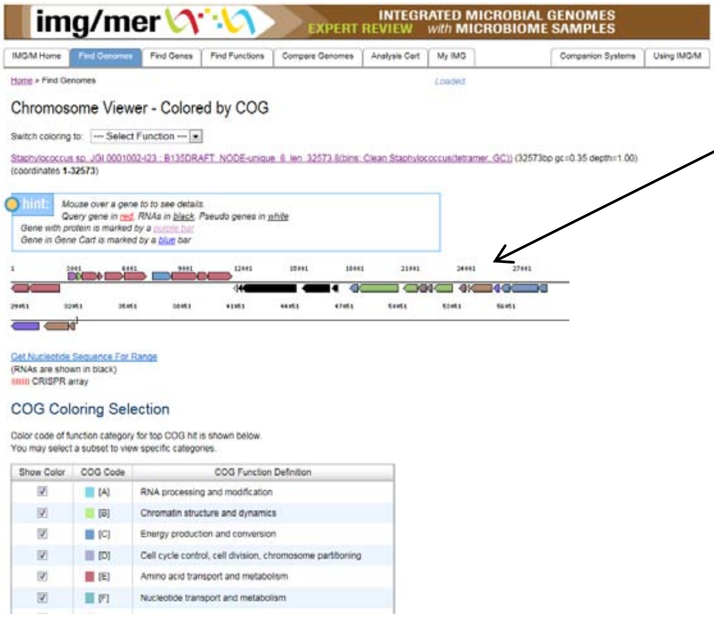
1..32573

User Enterable Coordinates

Start End

WARNING: Some browsers may be overwhelmed by a large coordinate range.

Version 3.5 May 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer gpweb04 2012-08-29-15.10.24



Home > Find Genomes

Chromosome Viewer - Colored by COG

Switch coloring to: --- Select Function ---

Staphylococcus sp. JGI 0001002-I23 : B135DRAFT_NODE-unique_8_len_32573.8(bins: Clean Staphylococcus(tetramer, GC)) (32573bp gc=0.35 depth=1.00)
(coordinates **1-32573**)

hint: *Mouse over a gene to to see details.*
Query gene in red. RNAs in black. Pseudo genes in white
Gene with protein is marked by a purple bar
Gene in Gene Cart is marked by a blue bar

Get Nucleotide Sequence For Range
(RNAs are shown in black)
IIIIII CRISPR array

COG Coloring Selection

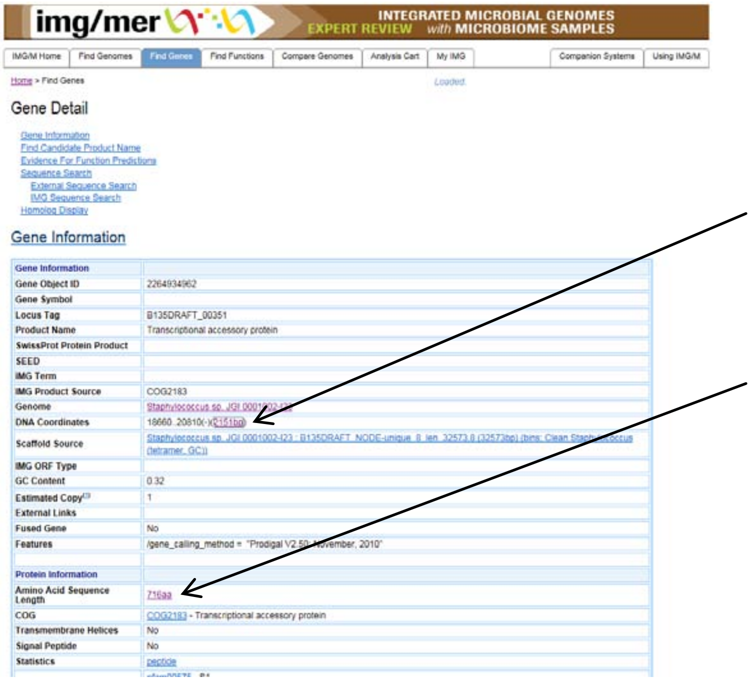
Color code of function category for top COG hit is shown below.
You may select a subset to view specific categories.

Show Color	COG Code	COG Function Definition
☑	[A]	RNA processing and modification
☑	[B]	Chromatin structure and dynamics
☑	[C]	Energy production and conversion
☑	[D]	Cell cycle control, cell division, chromosome partitioning
☑	[E]	Amino acid transport and metabolism
☑	[F]	Nucleotide transport and metabolism

The genes are color coded according to their COG category (scroll down for key).

Mouse over the genes to get a summary of their annotation.

Click on a gene to open the Gene Detail page (see below).



Home > Find Genes

Gene Detail

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

Gene Information

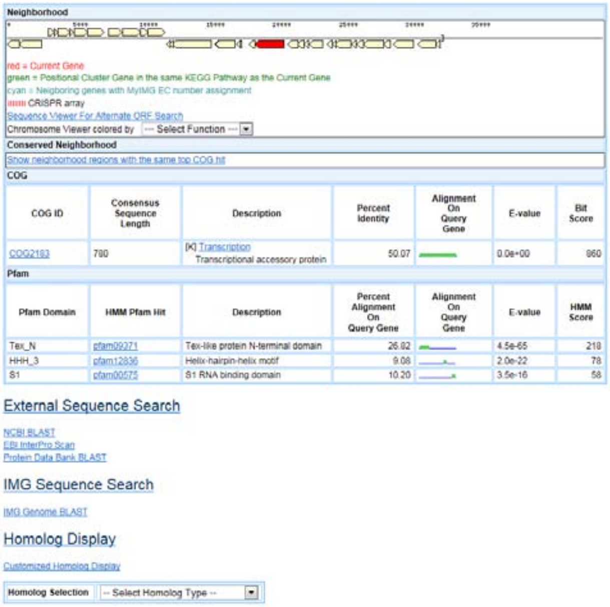
Gene Information	
Gene Object ID	2264934962
Gene Symbol	
Locus Tag	B135DRAFT_00351
Product Name	Transcriptional accessory protein
SwissProt Protein Product	
SEED	
IMG Term	
IMG Product Source	COG2183
Genome	Staphylococcus sp. JGI 0001002-I23
DNA Coordinates	18660..20810(-)(2151bp)
Scaffold Source	Staphylococcus sp. JGI 0001002-I23 : B135DRAFT_NODE-unique_8_len_32573.8 (32573bp) (bins: Clean Staphylococcus (tetramer, GC))
IMG ORF Type	
GC Content	0.32
Estimated Copy[?]	1
External Links	
Fused Gene	No
Features	/gene_calling_method = "Prodigal V2.50: November, 2010"
Protein Information	
Amino Acid Sequence Length	716aa
COG	COG2183 - Transcriptional accessory protein
Transmembrane Helices	No
Signal Peptide	No
Statistics	peptide

Click here to get the nucleotide sequence of the gene in fasta format.

Click here to get the amino acid sequence of the gene in fasta format.

If you scroll down to the bottom of the page there is a link to BLAST your gene's sequence.

Click here to do a blastp on your gene.



Neighborhood

red = Current Gene
green = Positional Cluster Gene in the same KEGG Pathway as the Current Gene
cyan = Neigboring genes with MyIMG EC number assignment
IIIIII CRISPR array
Sequence Viewer For Alternate ORF Search
Chromosome Viewer colored by --- Select Function ---

Conserved Neighborhood

Show neighborhood regions with the same top COG hit

COG

COG ID	Consensus Sequence Length	Description	Percent Identity	Alignment On Query Gene	E-value	Bit Score
COG2183	780	[K] Transcription Transcriptional accessory protein	50.07		0.0e+00	860

Pfam

Pfam Domain	HMM Pfam Hit	Description	Percent Alignment On Query Gene	Alignment On Query Gene	E-value	HMM Score
Tex_N	pfam09371	Tex-like protein N-terminal domain	26.82		4.5e-65	218
HHH_3	pfam12836	Helix-hairpin-helix motif	9.08		2.0e-22	78
S1	pfam00575	S1 RNA binding domain	10.20		3.5e-16	58

External Sequence Search

NCBI BLAST
EBI InterPro Scan
Protein Data Bank BLAST

IMG Sequence Search

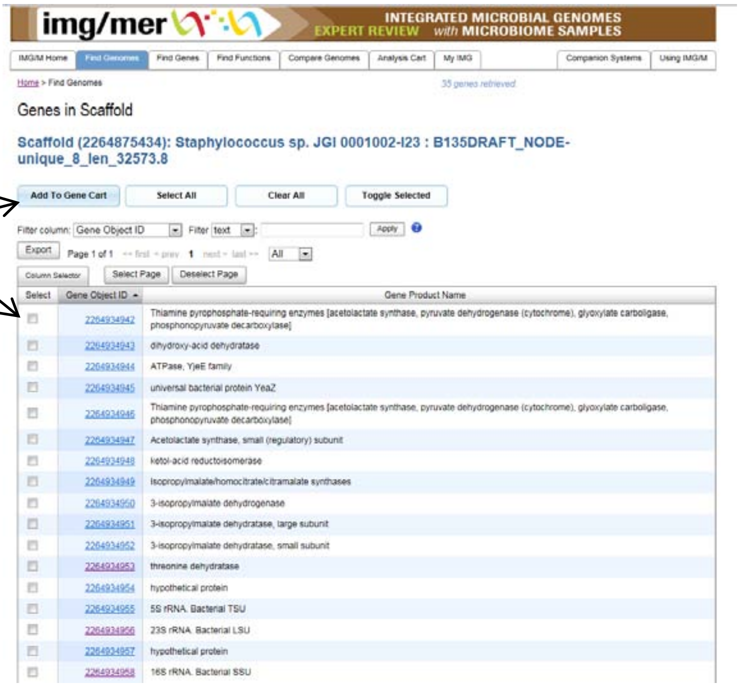
IMG Genome BLAST

Homolog Display

Customized Homolog Display

If you don't want to BLAST the genes individually, then go the Scaffold Detail page and click to get a list of genes from that scaffold.

Select the genes you want and then add them to the gene cart.



Home > Find Genomes

35 genes retrieved

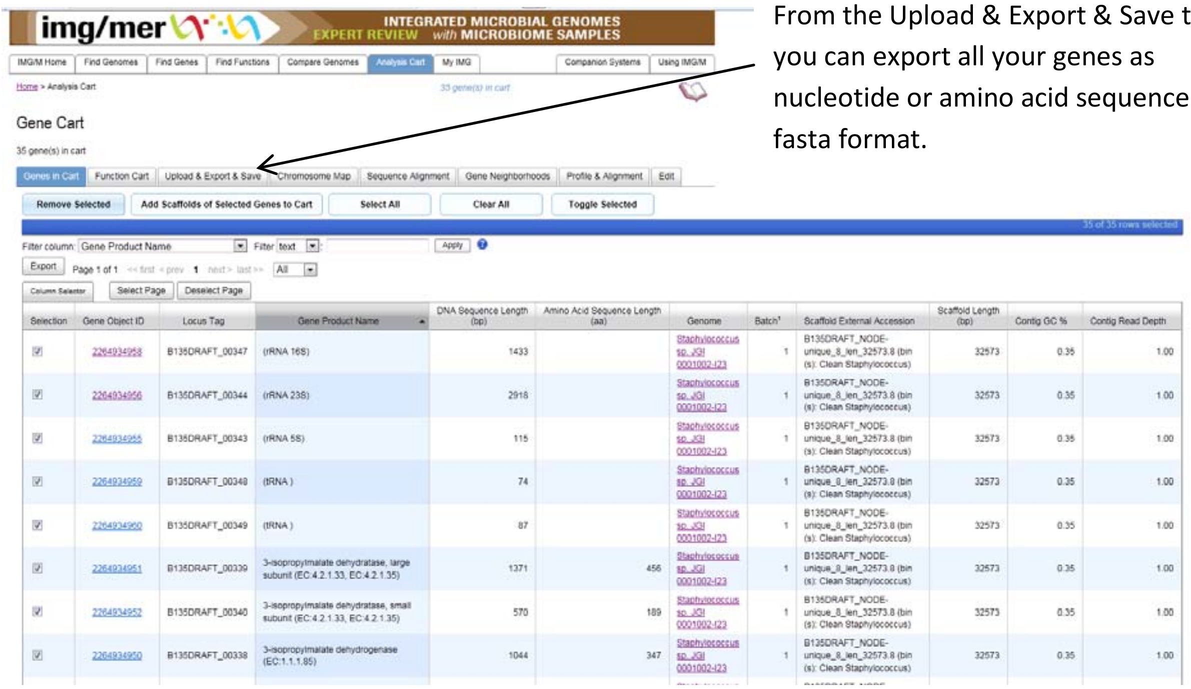
Genes in Scaffold

Scaffold (2264875434): Staphylococcus sp. JGI 0001002-I23 : B135DRAFT_NODE-unique_8_len_32573.8

Filter column: Gene Object ID Filter text :

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

Select	Gene Object ID	Gene Product Name
□	2264934942	Thiamine pyrophosphate-requiring enzymes [acetolactate synthase, pyruvate dehydrogenase (cytochrome), glyoxylate carboligase, phosphonopyruvate decarboxylase]
□	2264934943	dihydroxy-acid dehydratase
□	2264934944	ATPase, YjeE family
□	2264934945	universal bacterial protein YeaZ
□	2264934946	Thiamine pyrophosphate-requiring enzymes [acetolactate synthase, pyruvate dehydrogenase (cytochrome), glyoxylate carboligase, phosphonopyruvate decarboxylase]
□	2264934947	Acetolactate synthase, small (regulatory) subunit
□	2264934948	ketol-acid reductoisomerase
□	2264934949	Isopropylmalate/homocitrate/citramalate synthases
□	2264934950	3-isopropylmalate dehydrogenase
□	2264934951	3-isopropylmalate dehydratase, large subunit
□	2264934952	3-isopropylmalate dehydratase, small subunit
□	2264934953	threonine dehydratase
□	2264934954	hypothetical protein
□	2264934955	5S rRNA. Bacterial TSU
□	2264934956	23S rRNA. Bacterial LSU
□	2264934957	hypothetical protein
□	2264934958	16S rRNA. Bacterial SSU



Gene Cart

35 gene(s) in cart

Filter column: Gene Product Name Filter text

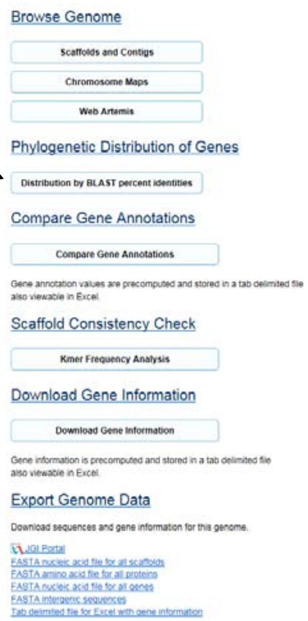
Selection	Gene Object ID	Locus Tag	Gene Product Name	DNA Sequence Length (bp)	Amino Acid Sequence Length (aa)	Genome	Batch[1]	Scaffold External Accession	Scaffold Length (bp)	Contig GC %	Contig Read Depth
☑	2264934958	B135DRAFT_00347	(rRNA 16S)	1433		Staphylococcus sp. JGI 0001002-I23	1	B135DRAFT_NODE-unique_8_len_32573.8 (bin (s): Clean Staphylococcus)	32573	0.35	1.00
☑	2264934956	B135DRAFT_00344	(rRNA 23S)	2918		Staphylococcus sp. JGI 0001002-I23	1	B135DRAFT_NODE-unique_8_len_32573.8 (bin (s): Clean Staphylococcus)	32573	0.35	1.00
☑	2264934955	B135DRAFT_00343	(rRNA 5S)	115		Staphylococcus sp. JGI 0001002-I23	1	B135DRAFT_NODE-unique_8_len_32573.8 (bin (s): Clean Staphylococcus)	32573	0.35	1.00
☑	2264934959	B135DRAFT_00348	(tRNA)	74		Staphylococcus sp. JGI 0001002-I23	1	B135DRAFT_NODE-unique_8_len_32573.8 (bin (s): Clean Staphylococcus)	32573	0.35	1.00
☑	2264934960	B135DRAFT_00349	(tRNA)	87		Staphylococcus sp. JGI 0001002-I23	1	B135DRAFT_NODE-unique_8_len_32573.8 (bin (s): Clean Staphylococcus)	32573	0.35	1.00
☑	2264934951	B135DRAFT_00339	3-isopropylmalate dehydratase, large subunit (EC:4.2.1.33, EC:4.2.1.35)	1371	456	Staphylococcus sp. JGI 0001002-I23	1	B135DRAFT_NODE-unique_8_len_32573.8 (bin (s): Clean Staphylococcus)	32573	0.35	1.00
☑	2264934952	B135DRAFT_00340	3-isopropylmalate dehydratase, small subunit (EC:4.2.1.33, EC:4.2.1.35)	570	189	Staphylococcus sp. JGI 0001002-I23	1	B135DRAFT_NODE-unique_8_len_32573.8 (bin (s): Clean Staphylococcus)	32573	0.35	1.00
☑	2264934950	B135DRAFT_00338	3-isopropylmalate dehydrogenase (EC:1.1.1.85)	1044	347	Staphylococcus sp. JGI 0001002-I23	1	B135DRAFT_NODE-unique_8_len_32573.8 (bin (s): Clean Staphylococcus)	32573	0.35	1.00

From the Upload & Export & Save tab you can export all your genes as nucleotide or amino acid sequences in fasta format.

Because of limitations with the databases, you should BLAST multiple genes from each scaffold and take the consensus taxonomy for all of them before deciding whether a particular scaffold is contaminant or not.

Although we primarily rely on the Kmer plot to identify suspicious scaffolds, there are other tools that you may find useful. First go back to the genome overview and scroll down

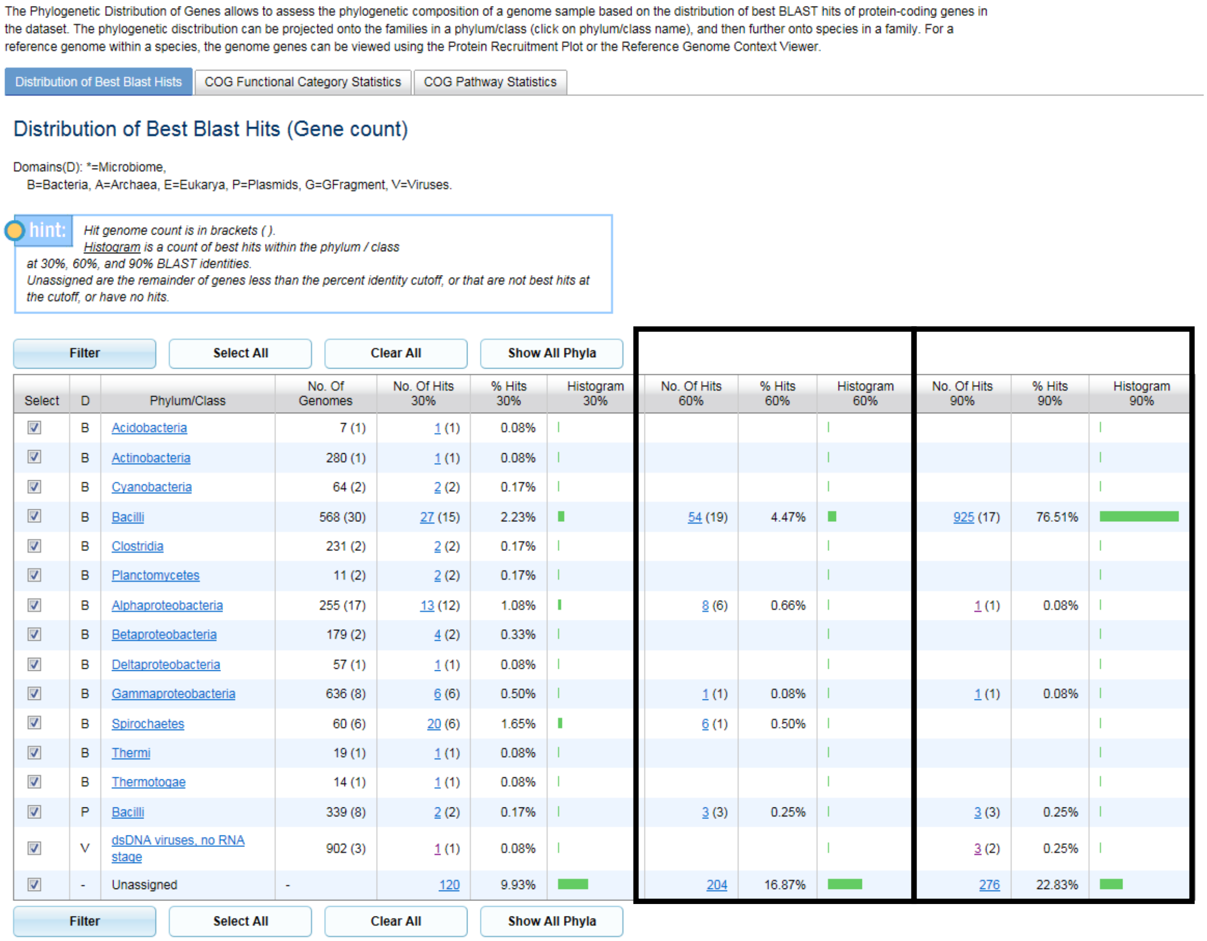
Click this button to get a phylogenetic distribution of your genes based on BLAST.


Browse Genome
Scaffolds and Contigs
Chromosome Maps
Web Artemis
Phylogenetic Distribution of Genes
Distribution by BLAST percent identities
Compare Gene Annotations
Compare Gene Annotations
Scaffold Consistency Check
Kmer Frequency Analysis
Download Gene Information
Download Gene Information
Export Genome Data
JGI Portal
FASTA nucleic acid file for all scaffolds
FASTA amino acid file for all proteins
FASTA nucleic acid file for all genes
FASTA intergenic sequences
Tab delimited file for Excel with gene information

Gene annotation values are precomputed and stored in a tab delimited file also viewable in Excel.

Gene information is precomputed and stored in a tab delimited file also viewable in Excel.

Download sequences and gene information for this genome.



The Phylogenetic Distribution of Genes allows to assess the phylogenetic composition of a genome sample based on the distribution of best BLAST hits of protein-coding genes in the dataset. The phylogenetic discribution can be projected onto the families in a phylum/class (click on phylum/class name), and then further onto species in a family. For a reference genome within a species, the genome genes can be viewed using the Protein Recruitment Plot or the Reference Genome Context Viewer.

Distribution of Best Blast Hists
COG Functional Category Statistics
COG Pathway Statistics

Distribution of Best Blast Hits (Gene count)

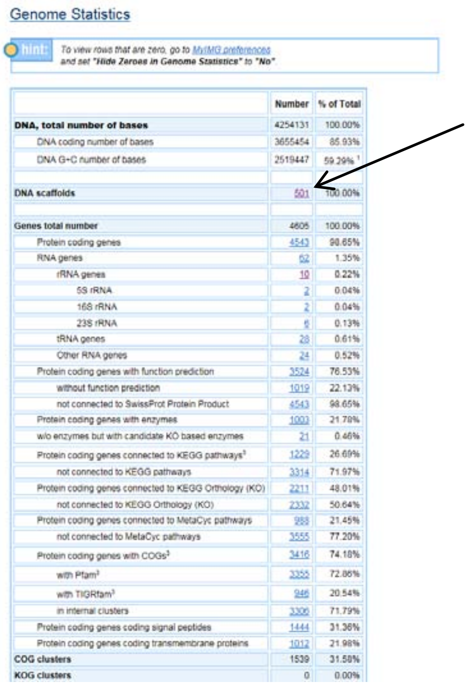
Domains(D): *=Microbiome,
B=Bacteria, A=Archaea, E=Eukarya, P=Plasmids, G=GFragment, V=Viruses.

hint: *Hit genome count is in brackets ( ).*
Histogram is a count of best hits within the phylum / class at 30%, 60%, and 90% BLAST identities.
Unassigned are the remainder of genes less than the percent identity cutoff, or that are not best hits at the cutoff, or have no hits.

Select	D	Phylum/Class	No. Of Genomes	No. Of Hits 30%	% Hits 30%	Histogram 30%	No. Of Hits 60%	% Hits 60%	Histogram 60%	No. Of Hits 90%	% Hits 90%	Histogram 90%
☑	B	Acidobacteria	7 (1)	1 (1)	0.08%	▏			▏			▏
☑	B	Actinobacteria	280 (1)	1 (1)	0.08%	▏			▏			▏
☑	B	Cyanobacteria	64 (2)	2 (2)	0.17%	▏			▏			▏
☑	B	Bacilli	568 (30)	27 (15)	2.23%	▮	54 (19)	4.47%	▮	925 (17)	76.51%	▬
☑	B	Clostridia	231 (2)	2 (2)	0.17%	▏			▏			▏
☑	B	Planctomycetes	11 (2)	2 (2)	0.17%	▏			▏			▏
☑	B	Alphaproteobacteria	255 (17)	13 (12)	1.08%	▮	8 (6)	0.66%	▏	1 (1)	0.08%	▏
☑	B	Betaproteobacteria	179 (2)	4 (2)	0.33%	▏			▏			▏
☑	B	Deltaproteobacteria	57 (1)	1 (1)	0.08%	▏			▏			▏
☑	B	Gammaproteobacteria	636 (8)	6 (6)	0.50%	▏	1 (1)	0.08%	▏	1 (1)	0.08%	▏
☑	B	Spirochaetes	60 (6)	20 (6)	1.65%	▮	6 (1)	0.50%	▏			▏
☑	B	Thermi	19 (1)	1 (1)	0.08%	▏			▏			▏
☑	B	Thermotogae	14 (1)	1 (1)	0.08%	▏			▏			▏
☑	P	Bacilli	339 (8)	2 (2)	0.17%	▏	3 (3)	0.25%	▏	3 (3)	0.25%	▏
☑	V	dsDNA viruses, no RNA stage	902 (3)	1 (1)	0.08%	▏			▏	3 (2)	0.25%	▏
☑	-	Unassigned	-	120	9.93%	▬	204	16.87%	▬	276	22.83%	▬

We mainly focus on hits that are 60% or 90% identity. This particular genome is in the Bacilli and it is good to see that most of the hits are either Bacilli or Unassigned. However, you can see that there are a few hits for Alpha- and Gammaproteobacteria, Spirochetes, and dsDNA viruses. Clicking on the blue number in the "No. Of Hits" column will take you to a list of the genes that hit that phylum.

Another useful tool shows a histogram of the GC content of the scaffolds. First, go back to the genome overview. Near the top where you found the rRNA genes there is the number of scaffolds you have.



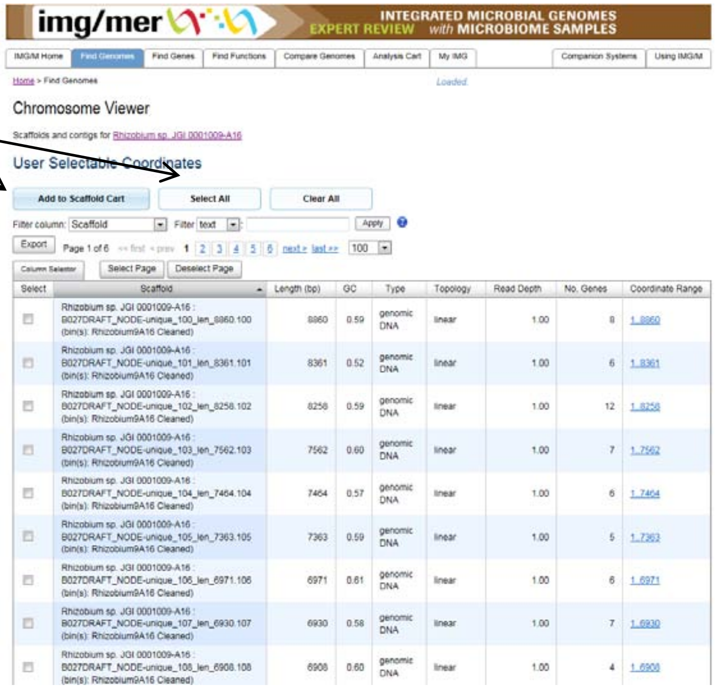
Genome Statistics

hint: *To view rows that are zero, go to MyIMG preferences and set "Hide Zeroes in Genome Statistics" to "No"*

	Number	% of Total
DNA, total number of bases	4254131	100.00%
DNA coding number of bases	3655454	85.93%
DNA G+C number of bases	2519447	59.29% [1]
DNA scaffolds	501	100.00%
Genes total number	4605	100.00%
Protein coding genes	4543	98.65%
RNA genes	62	1.35%
rRNA genes	10	0.22%
5S rRNA	2	0.04%
16S rRNA	2	0.04%
23S rRNA	6	0.13%
tRNA genes	28	0.61%
Other RNA genes	24	0.52%
Protein coding genes with function prediction	3524	76.53%
without function prediction	1019	22.13%
not connected to SwissProt Protein Product	4543	98.65%
Protein coding genes with enzymes	1003	21.78%
w/o enzymes but with candidate KO based enzymes	21	0.46%
Protein coding genes connected to KEGG pathways[3]	1229	26.69%
not connected to KEGG pathways	3314	71.97%
Protein coding genes connected to KEGG Orthology (KO)	2211	48.01%
not connected to KEGG Orthology (KO)	2332	50.64%
Protein coding genes connected to MetaCyc pathways	988	21.45%
not connected to MetaCyc pathways	3555	77.20%
Protein coding genes with COGs[3]	3416	74.18%
with Pfam[3]	3355	72.86%
with TIGRfam[3]	946	20.54%
in internal clusters	3306	71.79%
Protein coding genes coding signal peptides	1444	31.36%
Protein coding genes coding transmembrane proteins	1012	21.98%
COG clusters	1539	31.58%
KOG clusters	0	0.00%

Clicking here will bring up all your scaffolds

Select all your scaffolds and then add them to the Scaffold Cart.



Home > Find Genomes

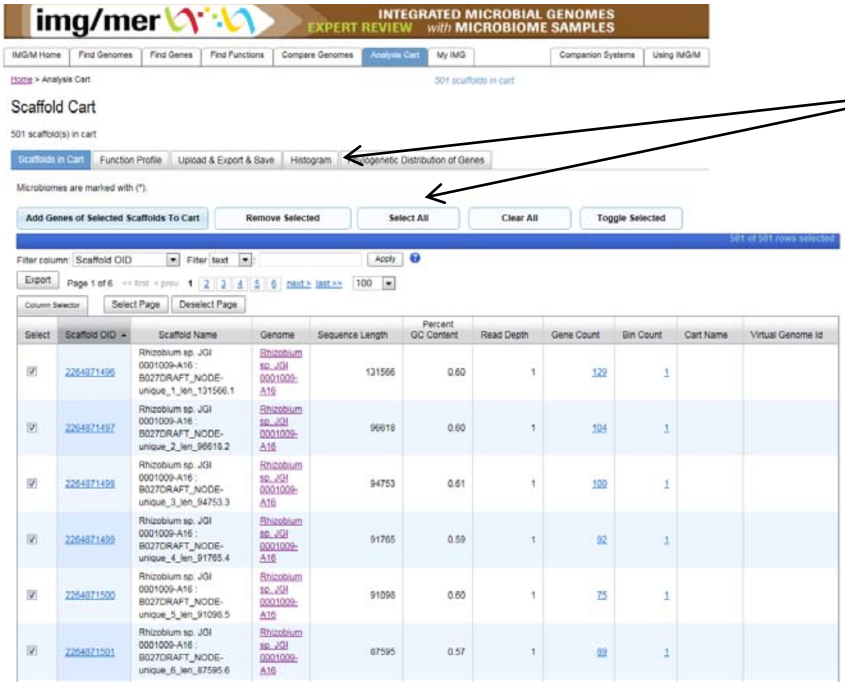
Loaded.

Chromosome Viewer

Scaffolds and contigs for Rhizobium sp. JGI 0001009-A16

User Selectable Coordinates

Select	Scaffold	Length (bp)	GC	Type	Topology	Read Depth	No. Genes	Coordinate Range
□	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_100_len_8860.100 (bin(s): Rhizobium9A16 Cleaned)	8860	0.59	genomic DNA	linear	1.00	8	1..8860
□	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_101_len_8361.101 (bin(s): Rhizobium9A16 Cleaned)	8361	0.52	genomic DNA	linear	1.00	6	1..8361
□	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_102_len_8258.102 (bin(s): Rhizobium9A16 Cleaned)	8258	0.59	genomic DNA	linear	1.00	12	1..8258
□	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_103_len_7562.103 (bin(s): Rhizobium9A16 Cleaned)	7562	0.60	genomic DNA	linear	1.00	7	1..7562
□	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_104_len_7464.104 (bin(s): Rhizobium9A16 Cleaned)	7464	0.57	genomic DNA	linear	1.00	6	1..7464
□	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_105_len_7363.105 (bin(s): Rhizobium9A16 Cleaned)	7363	0.59	genomic DNA	linear	1.00	5	1..7363
□	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_106_len_6971.106 (bin(s): Rhizobium9A16 Cleaned)	6971	0.61	genomic DNA	linear	1.00	6	1..6971
□	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_107_len_6930.107 (bin(s): Rhizobium9A16 Cleaned)	6930	0.58	genomic DNA	linear	1.00	7	1..6930
□	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_108_len_6908.108 (bin(s): Rhizobium9A16 Cleaned)	6908	0.60	genomic DNA	linear	1.00	4	1..6908



Home > Analysis Cart 501 scaffolds in cart

Scaffold Cart

501 scaffold(s) in cart

Microbiomes are marked with (*).

501 of 501 rows selected

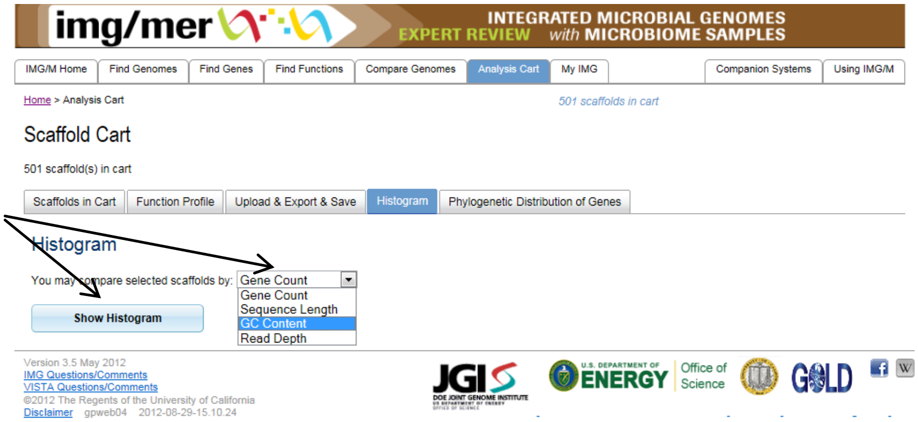
Filter column: Scaffold OID Filter text Apply

Export Page 1 of 6 << first < prev 1 2 3 4 5 6 next > last >> 100

Select	Scaffold OID	Scaffold Name	Genome	Sequence Length	Percent GC Content	Read Depth	Gene Count	Bin Count	Cart Name	Virtual Genome Id
☑	2264871496	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_1_len_131566.1	Rhizobium sp. JGI 0001009-A16	131566	0.60	1	129	1		
☑	2264871497	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_2_len_96618.2	Rhizobium sp. JGI 0001009-A16	96618	0.60	1	104	1		
☑	2264871498	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_3_len_94753.3	Rhizobium sp. JGI 0001009-A16	94753	0.61	1	100	1		
☑	2264871499	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_4_len_91765.4	Rhizobium sp. JGI 0001009-A16	91765	0.59	1	92	1		
☑	2264871500	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_5_len_91098.5	Rhizobium sp. JGI 0001009-A16	91098	0.60	1	75	1		
☑	2264871501	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_6_len_87595.6	Rhizobium sp. JGI 0001009-A16	87595	0.57	1	89	1		

Select all your scaffolds and then go to the Histogram tab.

Select GC Content from drop down menu then click Show Histogram.



Home > Analysis Cart 501 scaffolds in cart

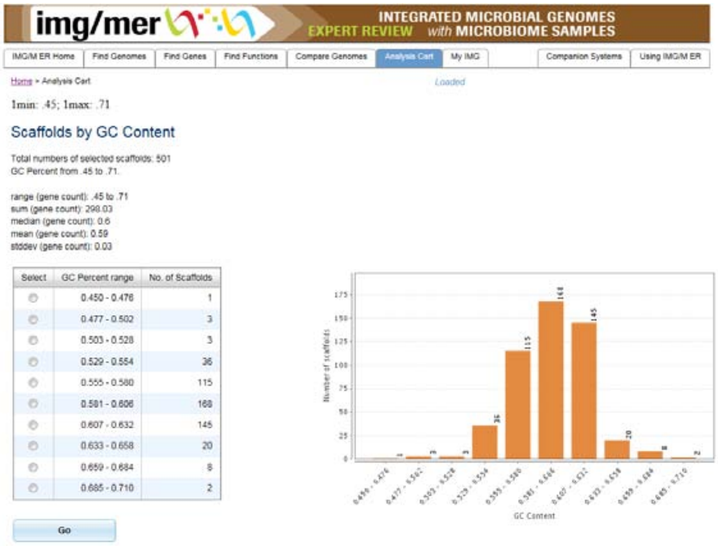
Scaffold Cart

501 scaffold(s) in cart

Histogram

You may compare selected scaffolds by: Gene Count

Version 3.5 May 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer gpweb04 2012-08-29-15.10.24



Home > Analysis Cart Loaded

1min: .45; 1max: .71

Scaffolds by GC Content

Total numbers of selected scaffolds: 501
GC Percent from .45 to .71.

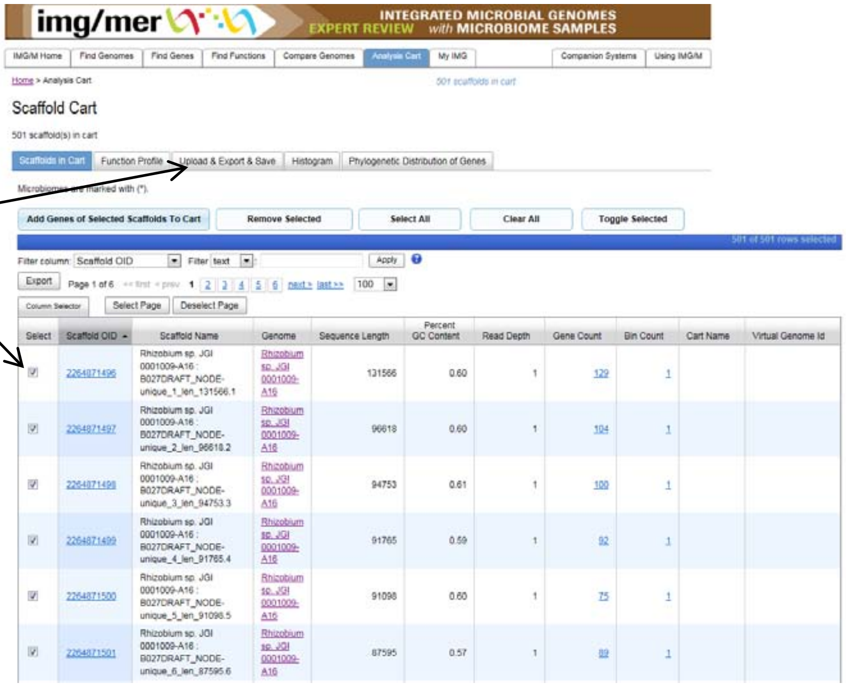
range (gene count): .45 to .71
sum (gene count): 298.03
median (gene count): 0.6
mean (gene count): 0.59
stddev (gene count): 0.03

Select	GC Percent range	No. of Scaffolds
○	0.450 - 0.476	1
○	0.477 - 0.502	3
○	0.503 - 0.528	3
○	0.529 - 0.554	36
○	0.555 - 0.580	115
○	0.581 - 0.606	168
○	0.607 - 0.632	145
○	0.633 - 0.658	20
○	0.659 - 0.684	8
○	0.685 - 0.710	2

For a clean genome, you should have a single peak and all scaffolds should be within ~10% to either side of the center. For the above genome I would investigate the four scaffolds with the lowest GC content and the 2 with the highest. Note that the histogram always gives you 10 bars/bins no matter how wide the spread in GC content is.

Once you have identified which scaffolds are contaminants, you need to remove them from your dataset. The IMG system does not have a direct way to remove data so the process involves selecting all the scaffolds that you want to keep and re-uploading them to IMG to replace the contaminated dataset.

One thing that can help with this process are scaffold sets. First, go to the Scaffold Cart.



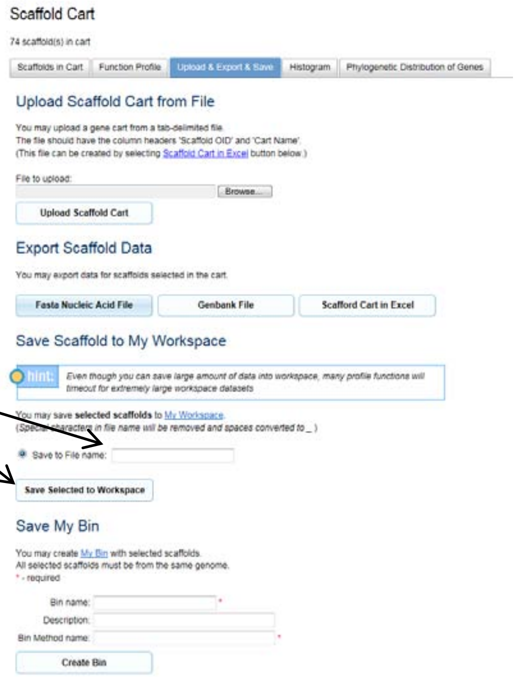
Scaffold Cart

501 scaffold(s) in cart

Microbiomes are marked with (*).

Select	Scaffold OID	Scaffold Name	Genome	Sequence Length	Percent GC Content	Read Depth	Gene Count	Bin Count	Cart Name	Virtual Genome Id
☑	2264871496	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_1_len_131566.1	Rhizobium sp. JGI 0001009-A16	131566	0.60	1	129	1		
☑	2264871497	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_2_len_96618.2	Rhizobium sp. JGI 0001009-A16	96618	0.60	1	104	1		
☑	2264871498	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_3_len_94753.3	Rhizobium sp. JGI 0001009-A16	94753	0.61	1	100	1		
☑	2264871499	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_4_len_91765.4	Rhizobium sp. JGI 0001009-A16	91765	0.59	1	92	1		
☑	2264871500	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_5_len_91098.5	Rhizobium sp. JGI 0001009-A16	91098	0.60	1	75	1		
☑	2264871501	Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_6_len_87595.6	Rhizobium sp. JGI 0001009-A16	87595	0.57	1	89	1		

Select the scaffolds you want in the set and then go to the Upload & Export & Save tab.



Scaffold Cart

74 scaffold(s) in cart

Upload Scaffold Cart from File

You may upload a gene cart from a tab-delimited file.
The file should have the column headers 'Scaffold OID' and 'Cart Name'.
(This file can be created by selecting Scaffold Cart in Excel button below.)

File to upload:

Export Scaffold Data

You may export data for scaffolds selected in the cart.

Save Scaffold to My Workspace

Even though you can save large amount of data into workspace, many profile functions will timeout for extremely large workspace datasets

You may save **selected scaffolds** to My Workspace.
(Special characters in file name will be removed and spaces converted to _)

● Save to File name:

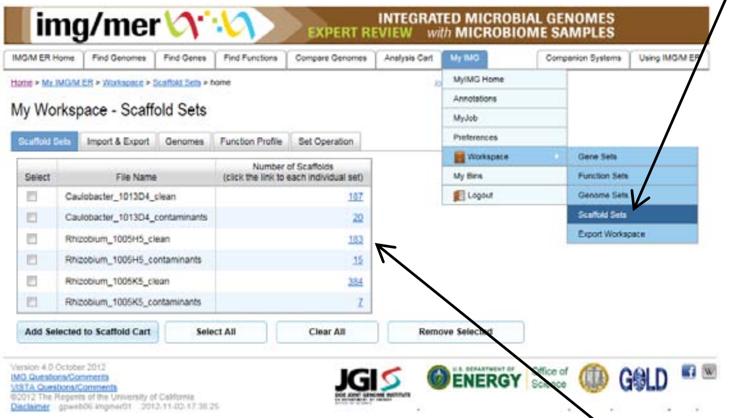
Save My Bin

You may create My Bin with selected scaffolds.
All selected scaffolds must be from the same genome.
* - required

Bin name: *
Description:
Bin Method name: *

Name the set and then press the Save Selected to Workspace button.

You can examine the sets that you have created by going to Scaffold Sets under Workspace in the My IMG tab.



Home > My IMG/M ER > Workspace > Scaffold Sets > home

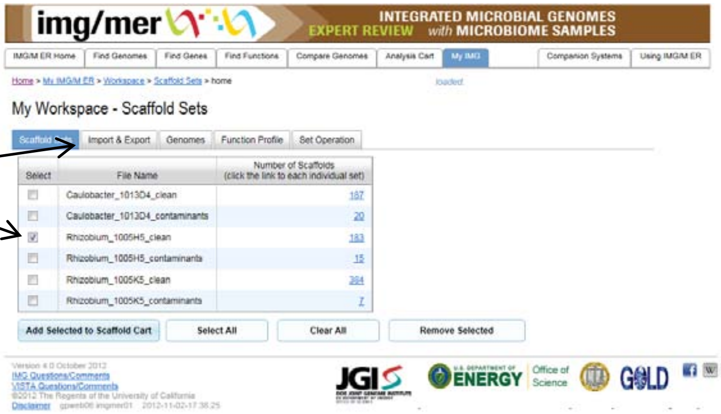
My Workspace - Scaffold Sets

Select	File Name	Number of Scaffolds (click the link to each individual set)
□	Caulobacter_1013D4_clean	187
□	Caulobacter_1013D4_contaminants	20
□	Rhizobium_1005H5_clean	183
□	Rhizobium_1005H5_contaminants	15
□	Rhizobium_1005K5_clean	384
□	Rhizobium_1005K5_contaminants	7

Version 4.0 October 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer gpweb06 imgmer01 2012-11-02-17:38:25

From here you can view the scaffolds in the sets that you have created and add them to the scaffold cart.

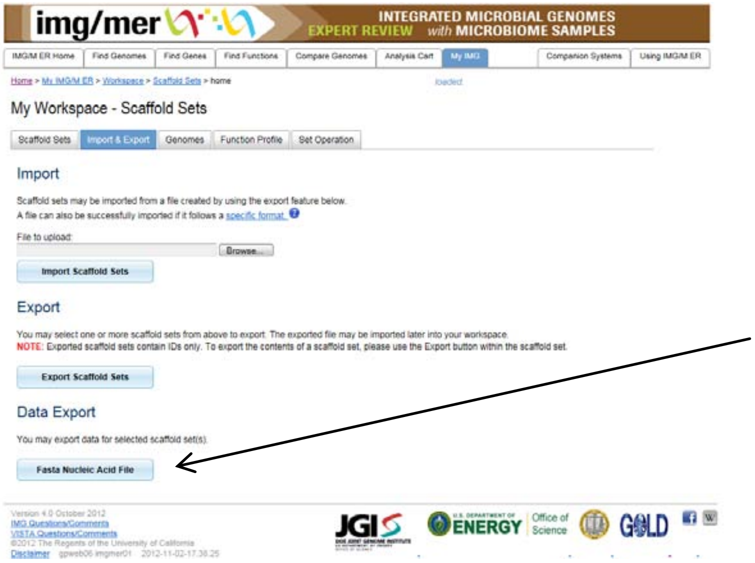
You can easily get the contents of a scaffold set by selecting the set and going to the Import & Export tab.



My Workspace - Scaffold Sets

Select	File Name	Number of Scaffolds (click the link to each individual set)
□	Caulobacter_1013D4_clean	187
□	Caulobacter_1013D4_contaminants	20
☑	Rhizobium_1005H5_clean	183
□	Rhizobium_1005H5_contaminants	15
□	Rhizobium_1005K5_clean	384
□	Rhizobium_1005K5_contaminants	7

Version 4.0 October 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer gpweb06 imgmer01 2012-11-02-17:38:25



My Workspace - Scaffold Sets

Import

Scaffold sets may be imported from a file created by using the export feature below.
A file can also be successfully imported if it follows a specific format.

File to upload:

Export

You may select one or more scaffold sets from above to export. The exported file may be imported later into your workspace.
NOTE: Exported scaffold sets contain IDs only. To export the contents of a scaffold set, please use the Export button within the scaffold set.

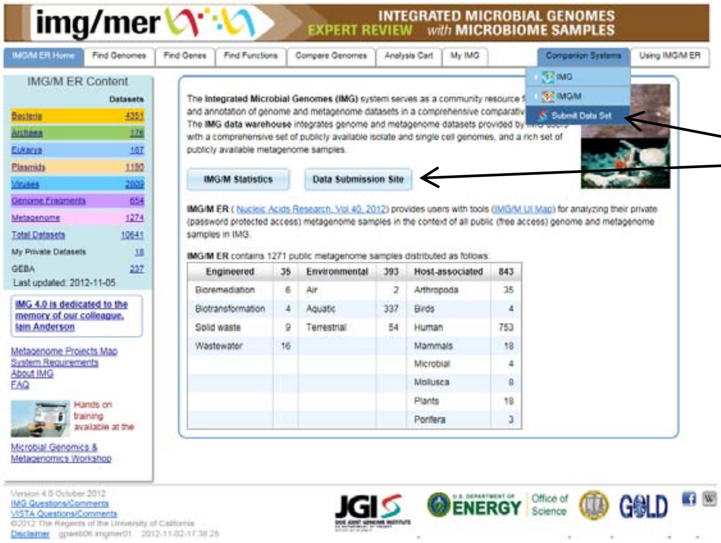
Data Export

You may export data for selected scaffold set(s).

Version 4.0 October 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer gpweb06 imgmer01 2012-11-02-17:38:25

Here you can export the contents of your scaffold set as a fasta file that you save to your computer.

Once you have a file of just the clean scaffolds from your genome, you need to re-upload the data. This will overwrite the existing data so if you want to save the sequences from your contaminant scaffolds for future use you will have to save a fasta file of those scaffolds on your own computer.



IMG/M ER Content

	Datasets
Bacteria	4351
Archaea	176
Eukarya	167
Plasmids	1180
Viruses	2809
Genome Fragments	654
Metagenome	1274
Total Datasets	10641
My Private Datasets	18
GEBA	237

Last updated: 2012-11-05

IMG 4.0 is dedicated to the memory of our colleague, Iain Anderson

Metagenome Projects Map
System Requirements
About IMG
FAQ

Microbial Genomics & Metagenomics Workshop

The **Integrated Microbial Genomes (IMG)** system serves as a community resource f
and annotation of genome and metagenome datasets in a comprehensive comparativ
The **IMG data warehouse** integrates genome and metagenome datasets provided by
with a comprehensive set of publicly available isolate and single cell genomes, and a rich set of publicly available metagenome samples.

IMG/M ER (Nucleic Acids Research, Vol 40, 2012) provides users with tools (IMG/M UI Map) for analyzing their private (password protected access) metagenome samples in the context of all public (free access) genome and metagenome samples in IMG.

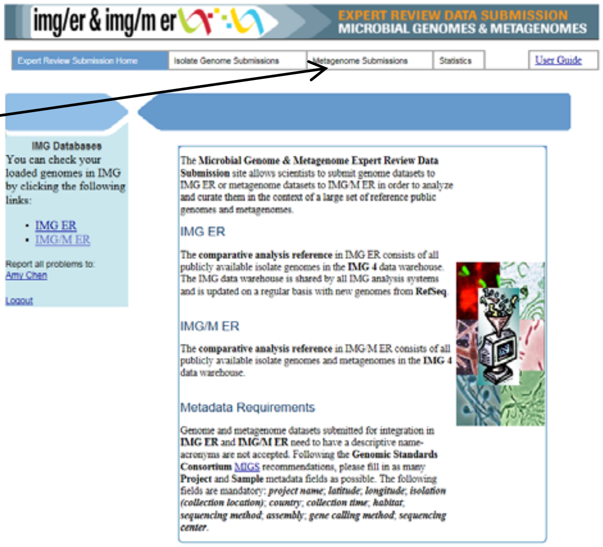
IMG/M ER contains 1271 public metagenome samples distributed as follows:

Engineered	35	Environmental	393	Host-associated	843
Bioremediation	6	Air	2	Arthropoda	35
Biotransformation	4	Aquatic	337	Birds	4
Solid waste	9	Terrestrial	54	Human	753
Wastewater	16			Mammals	18
				Microbial	4
				Mollusca	8
				Plants	18
				Porifera	3

Version 4.0 October 2012
IMG Questions/Comments
VISTA Questions/Comments
©2012 The Regents of the University of California
Disclaimer gpweb06 imgmer01 2012-11-02-17:38:25

Select Submit Data Set from the Companion Systems tab from anywhere in IMG/MER, or click the Data Submission Site button on the home page.

Select Metagenome Submissions.


img/er & img/m er
EXPERT REVIEW DATA SUBMISSION
MICROBIAL GENOMES & METAGENOMES
Expert Review Submission Home
Isolate Genome Submissions
Metagenome Submissions
Statistics
User Guide

IMG Databases

You can check your loaded genomes in IMG by clicking the following links:

- IMG ER
- IMG/M ER

Report all problems to: Amy Chen

Logout

The **Microbial Genome & Metagenome Expert Review Data Submission** site allows scientists to submit genome datasets to IMG ER or metagenome datasets to IMG/M ER in order to analyze and curate them in the context of a large set of reference public genomes and metagenomes.

IMG ER

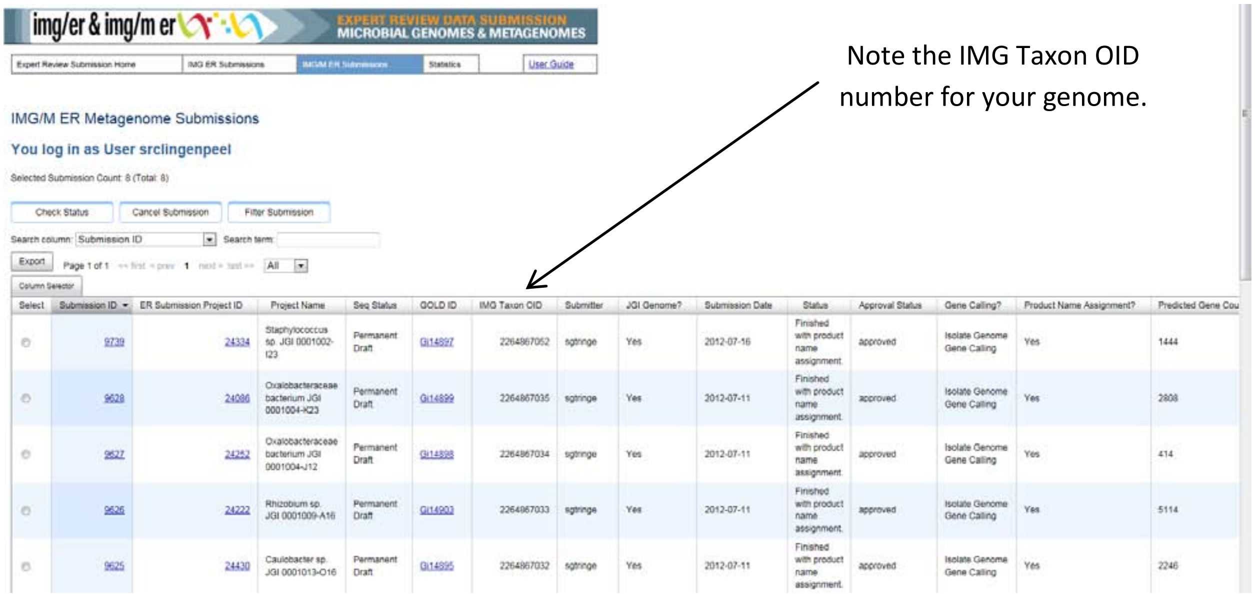
The **comparative analysis reference** in IMG ER consists of all publicly available isolate genomes in the **IMG 4** data warehouse. The IMG data warehouse is shared by all IMG analysis systems and is updated on a regular basis with new genomes from **RefSeq**.

IMG/M ER

The **comparative analysis reference** in IMG/M ER consists of all publicly available isolate genomes and metagenomes in the **IMG 4** data warehouse.

Metadata Requirements

Genome and metagenome datasets submitted for integration in **IMG ER** and **IMG/M ER** need to have a descriptive name-acronyms are not accepted. Following the **Genomic Standards Consortium** MIGS recommendations, please fill in as many **Project** and **Sample** metadata fields as possible. The following fields are mandatory: ***project name***; ***latitude***; ***longitude***; ***isolation (collection location)***; ***country***; ***collection time***; ***habitat***; ***sequencing method***; ***assembly***; ***gene calling method***; ***sequencing center***.


img/er & img/m er
EXPERT REVIEW DATA SUBMISSION
MICROBIAL GENOMES & METAGENOMES
Expert Review Submission Home
IMG ER Submissions
IMG/M ER Submissions
Statistics
User Guide

IMG/M ER Metagenome Submissions

You log in as User srclingenpeel

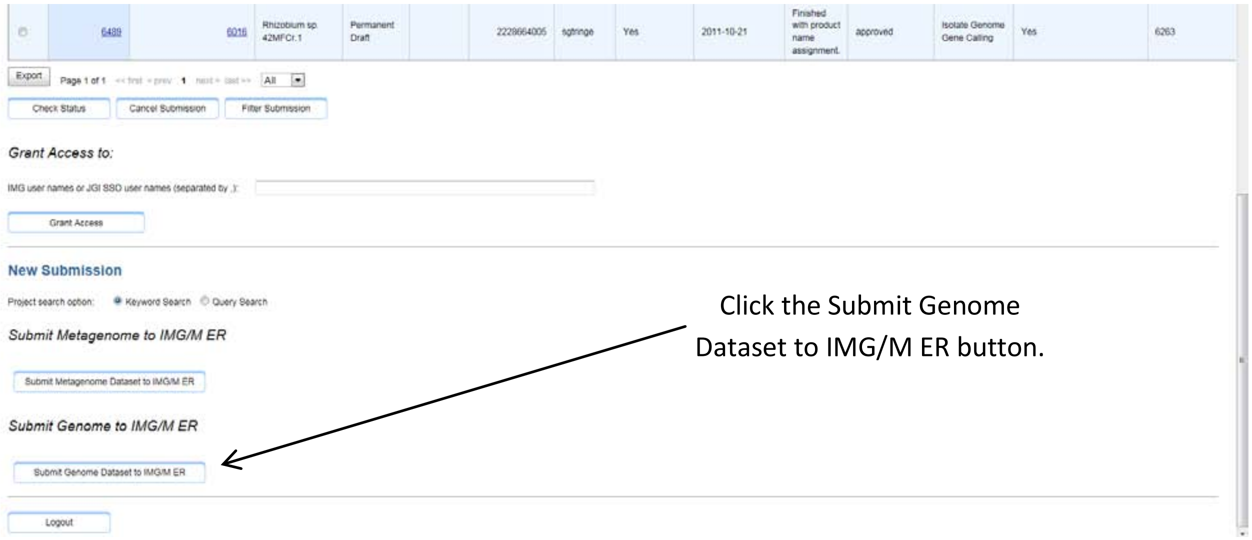
Selected Submission Count: 8 (Total: 8)

Search column: Submission ID Search term:

Select	Submission ID	ER Submission Project ID	Project Name	Seq Status	GOLD ID	IMG Taxon OID	Submitter	JGI Genome?	Submission Date	Status	Approval Status	Gene Calling?	Product Name Assignment?	Predicted Gene Cou
○	9739	24334	Staphylococcus sp. JGI 0001002-I23	Permanent Draft	Gi14897	2264867052	sgtringe	Yes	2012-07-16	Finished with product name assignment.	approved	Isolate Genome Gene Calling	Yes	1444
○	9628	24086	Oxalobacteraceae bacterium JGI 0001004-K23	Permanent Draft	Gi14899	2264867035	sgtringe	Yes	2012-07-11	Finished with product name assignment.	approved	Isolate Genome Gene Calling	Yes	2808
○	9627	24252	Oxalobacteraceae bacterium JGI 0001004-J12	Permanent Draft	Gi14898	2264867034	sgtringe	Yes	2012-07-11	Finished with product name assignment.	approved	Isolate Genome Gene Calling	Yes	414
○	9626	24222	Rhizobium sp. JGI 0001009-A16	Permanent Draft	Gi14903	2264867033	sgtringe	Yes	2012-07-11	Finished with product name assignment.	approved	Isolate Genome Gene Calling	Yes	5114
○	9625	24430	Caulobacter sp. JGI 0001013-O16	Permanent Draft	Gi14895	2264867032	sgtringe	Yes	2012-07-11	Finished with product name assignment.	approved	Isolate Genome Gene Calling	Yes	2246

Note the IMG Taxon OID number for your genome.

Scroll to the bottom of this page.



○	6489	6016	Rhizobium sp. 42MFCr.1	Permanent Draft		2228664005	sgtringe	Yes	2011-10-21	Finished with product name assignment.	approved	Isolate Genome Gene Calling	Yes	6263

Export
Page 1 of 1 << first < prev 1 next > last >> All
Check Status
Cancel Submission
Filter Submission

Grant Access to:

IMG user names or JGI SSO user names (separated by ,):

New Submission

Project search option: ● Keyword Search ○ Query Search

Submit Metagenome to IMG/M ER

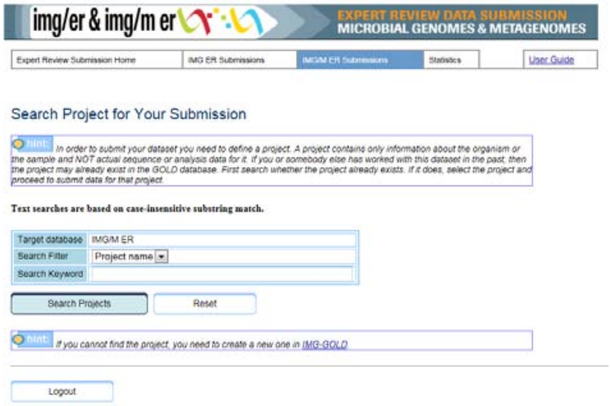
Submit Metagenome Dataset to IMG/M ER

Submit Genome to IMG/M ER

Submit Genome Dataset to IMG/M ER
Logout

Click the Submit Genome Dataset to IMG/M ER button.

Find your genome's project.



Search Project for Your Submission

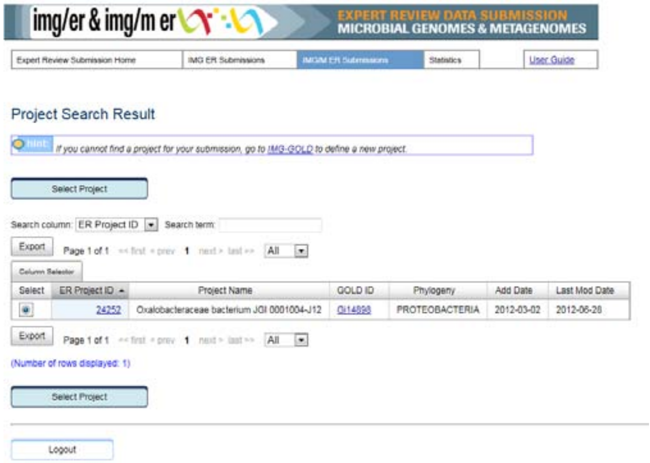
hint: *In order to submit your dataset you need to define a project. A project contains only information about the organism or the sample and NOT actual sequence or analysis data for it. If you or somebody else has worked with this dataset in the past, then the project may already exist in the GOLD database. First search whether the project already exists. If it does, select the project and proceed to submit data for that project.*

Text searches are based on case-insensitive substring match.

Target database	IMG/M ER
Search Filter	Project name ▾
Search Keyword	

hint: *If you cannot find the project, you need to create a new one in IMG-GOLD*

Select your genome's project.



Project Search Result

hint: *If you cannot find a project for your submission, go to IMG-GOLD to define a new project.*

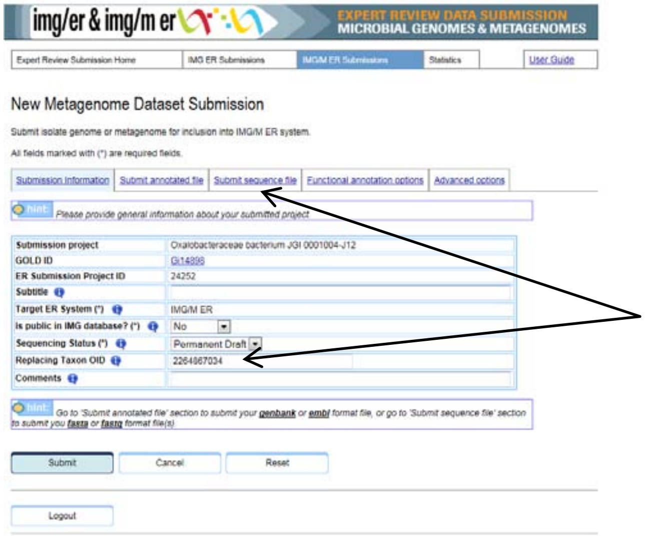
Search column: ER Project ID ▾ Search term:

Page 1 of 1 All ▾

Select	ER Project ID ▲	Project Name	GOLD ID	Phylogeny	Add Date	Last Mod Date
●	24252	Oxalobacteraceae bacterium JGI 0001004-J12	Gi14898	PROTEOBACTERIA	2012-03-02	2012-06-28

Page 1 of 1 All ▾

(Number of rows displayed: 1)



New Metagenome Dataset Submission

Submit isolate genome or metagenome for inclusion into IMG/M ER system.

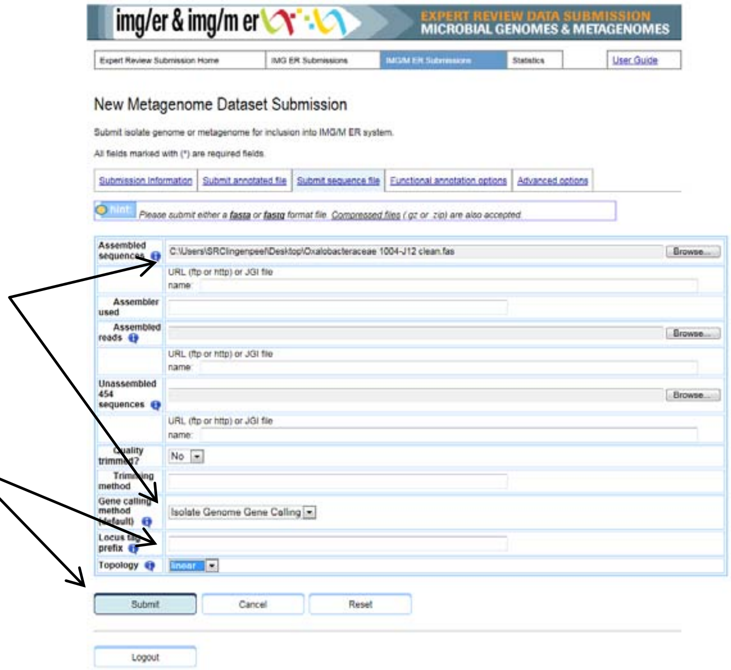
All fields marked with (*) are required fields.

hint: *Please provide general information about your submitted project*

Submission project	Oxalobacteraceae bacterium JGI 0001004-J12
GOLD ID	Gi14896
ER Submission Project ID	24252
Subtitle	
Target ER System (*)	IMG/M ER
Is public in IMG database? (*)	No ▾
Sequencing Status (*)	Permanent Draft ▾
Replacing Taxon OID	2264867034
Comments	

hint: *Go to 'Submit annotated file' section to submit your **genbank** or **embl** format file, or go to 'Submit sequence file' section to submit you **fasta** or **fastq** format file(s)*

Enter the Taxon OID then click on the Submit sequence file tab.



New Metagenome Dataset Submission

Submit isolate genome or metagenome for inclusion into IMG/M ER system.

All fields marked with (*) are required fields.

*Please submit either a **fasta** or **fastq** format file. Compressed files (gz or zip) are also accepted.*

Assembled sequences	C:\Users\SRClingenpeel\Desktop\Oxalobacteraceae 1004-J12 clean.fas
	URL (ftp or http) or JGI file name:
Assembler used	
Assembled reads	
	URL (ftp or http) or JGI file name:
Unassembled 454 sequences	
	URL (ftp or http) or JGI file name:
Quality trimmed?	No
Trimming method	
Gene calling method (default)	Isolate Genome Gene Calling
Locus tag prefix	
Topology	linear

Enter your fasta file and change the gene caller to Isolate Genome Gene Calling.

Enter a Locus tag prefix then hit the submit button.

Congratulations! Once this is loaded into IMG you will have a single cell genome to analyze that is free of contamination sequences.