

## IMG 4 Data Warehouse

The **Integrated Microbial Genomes (IMG) data warehouse** integrates genome and metagenome datasets provided by IMG users with bacterial, archaeal, eukaryotic, and phage genomes from NCBI's Genbank (<http://www.ncbi.nlm.nih.gov/genbank/>) and Reference Sequence non redundant collection (<http://www.ncbi.nlm.nih.gov/RefSeq/>), and a rich set of publicly available engineered, environmental and host associated metagenome samples.

Genome and metagenome datasets are provided by IMG users using the IMG Submission system (<http://img.jgi.doe.gov/submit>), are processed using IMG's microbial genome and metagenome annotation pipelines and integrated into the IMG data warehouse using IMG's data integration pipelines. The annotation and data integration standard operating procedures are available at:

- microbial genomes: [http://img.jgi.doe.gov/w/doc/MGAandDI\\_SOP.pdf](http://img.jgi.doe.gov/w/doc/MGAandDI_SOP.pdf).
- metagenomes: <http://img.jgi.doe.gov/m/doc/MetagenomeAnnotationSOP.pdf>.

Genes of both publicly available and user provided genomes and metagenomes in IMG are characterized using several functional resources, including COG, KOG, KEGG (release 63.0, 7/2012), PFAM (version 26.0, 11/ 2011), TIGRfam (release 12.0, 2/2012), MetaCyc (release 16.1, 7/ 2012), Gene Ontology (6/ 2012), and Interpro (4/ 2012).

In addition to genome and metagenome sequence data, IMG contains data from four **protein expression studies** (two *Arthrobacter chlorophenolicus* studies, a *Cryptobacterium curtum* study, and a *Brachy bacterium faecium* study) and seven **RNASeq experiments**. Protein expression data are publicly available, while access to RNASeq data is restricted.

## Users

IMG's "public" (unrestricted access) content is available to all interested scientific users.

IMG's "private" (restricted access) content is available only to IMG **registered users** who have password protected access to their own genomes/metagenomes as well as to all public genomes and metagenomes in IMG.

IMG registered users can submit genomes or metagenome samples for inclusion into IMG at: <http://img.jgi.doe.gov/submit>.

Users can request an IMG account at: <https://img.jgi.doe.gov/request>. Users who have JGI Single Sign-On (SSO) accounts can use their JGI accounts to access IMG.

As of **November 10<sup>th</sup>, 2012**, IMG had about **2,500 registered users** from **62 countries** across North America (57% users), Europe (20% users), Asia (13% users), South America (5% users), Oceania (4.5% users), and Africa (.5% users).

## Genomes

As of **Nov 10<sup>th</sup>, 2012**, IMG contains a total of **11,997** genomes, plasmids and genome fragments, with a total of **26.6 million** protein coding **genes**. About **30%** of the genomes in IMG are sequenced at the Joint Genome Institute, while **70%** are sequenced at other centers.

**2,431** genomes in IMG, with a total of **7.4 million** protein coding **genes**, are “private”, that is, are owned by the users have password protected access to them. 1,035 genomes in IMG are single cells, distributed among bacterial, archaeal, and eukaryotic genomes as shown in the table below.

<b>Public</b>	<b>9,458</b>	<b>Private</b>	<b>2,539</b>	<b>Total</b>	<b>11,997</b>
Bacteria	4,442		2,217		6,659
Single Cell	8	Single Cell	929	Single Cell	937
Archaea	176		250		426
		Single Cell	175	Single Cell	175
Eukarya	187		22		209
		Single Cell	8	Single Cell	8
Viruses	2,809		34		2,843
Plasmids	1,190		16		1,206
Genome Fragments	654				654

## Metagenomes

As of **Nov 10<sup>th</sup>, 2012**, IMG contains **2,099 metagenome samples**, with a total of **7.9 billion** protein coding **genes**. About **75%** of the metagenome samples in IMG have been sequenced at DOE’s Joint Genome Institute, with 25% sequenced at other sequencing organizations.

**1,271** metagenome samples in IMG, with a total of 2 billion protein coding genes, are **publicly** available, are distributed as follows:

<b>Engineered</b>	<b>35</b>	<b>Environmental</b>	<b>393</b>	<b>Host associated</b>	<b>843</b>
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

**825** metagenome samples in IMG, with a total of **5.9 billion** protein coding **genes**, are “private”, that is, are owned by the users have password protected access to them.

## Content History

Year	Genomes Added [B+A+E All (Genes)] Total			Metagenomes Added (Genes) Total		
2006	500	2,084 (1.7 Mil)	<b>2,084</b>	39	(1.2 Mil)	<b>39</b>
2007	335	1,191 (1.5 Mil)	3,275	21	(1.5 Mil)	60
2008	356	805 (1.4 Mil)	4,080	64	(1.4 Mil)	124
2009	722	1,131 (2.9 Mil)	5,211	109	(12 Mil)	233
2010	1,037	1,476 (3.4 Mil)	6,687	304	(573 Mil)	537
2011	1,710	2,659 (7.9 Mil)	9,346	1,101*	(976 Mil)	1,638
2012	2,637	2,651 (7.9 Mil)	<b>11,997</b>	462	(6.9 Bil)	<b>2,100</b>
Nov 10						

\* Includes 748 samples from the Human Microbiome Project.

## IMG 4 Analysis Tools

Microbial genome and metagenome specific user interfaces provide access to different subsets of the IMG data warehouse and analysis toolkits:

- Analysis tools for **microbial genomes** are summarized in Figure 1 and are available via:
  - **IMG** (<http://img.jgi.doe.gov>) provides support for the analysis of publicly available isolate genomes in the context of all publicly available genomes in the IMG data warehouse;
  - **IMG ER** (<http://img.jgi.doe.gov/er>) provides registered IMG users with tools for the “expert review” analysis and curation of their private (password protected) genomes in the context of all publicly available genomes and metagenomes in the IMG data warehouse.
- Analysis tools for **metagenomes** are summarized in Figure 2 and are available via:
  - **IMG/M** (<http://img.jgi.doe.gov/m>) provides support for the analysis of publicly available metagenome samples in the context of all publicly available genomes and metagenomes in the IMG data warehouse;
  - **IMG/M-ER** (<http://img.jgi.doe.gov/mer>) provides registered IMG users with tools for the “expert review” analysis and curation of their private (password protected) metagenome samples in the context of all publicly available genomes and metagenomes in the IMG data warehouse.

The IMG 4 analysis toolkits preserve in general the functionality provided by earlier versions of IMG, with the development effort focused on adapting the tools to the new IMG data warehouse and on handling substantially larger metagenome datasets.

In order to handle a rapidly growing number of **metagenome datasets** of increasing size (hundred million to billion genes, including unassembled reads), the IMG’s user interface has been changed as follows:

1. **Genes** of metagenome samples **are no longer associated** with TIGRfam, Transporter Classification, Signal peptides, and Transmembrane proteins (see annotation standard operating procedure for metagenomes), and therefore these annotations are not available in the **Gene Detail** pages for metagenome genes, nor in the **Metagenome Statistics** section of the **Metagenome Detail** pages.
2. The following tools that were provided as part of the **Metagenome Detail** page are no longer supported: **Compare Gene Annotations**, **Web Artemis**, **Find Candidate Product Name**, and **Find Candidate Enzyme**.

An important extension of the IMG analysis toolkits is the addition of user specific “**workspace**” capabilities, further discussed below.

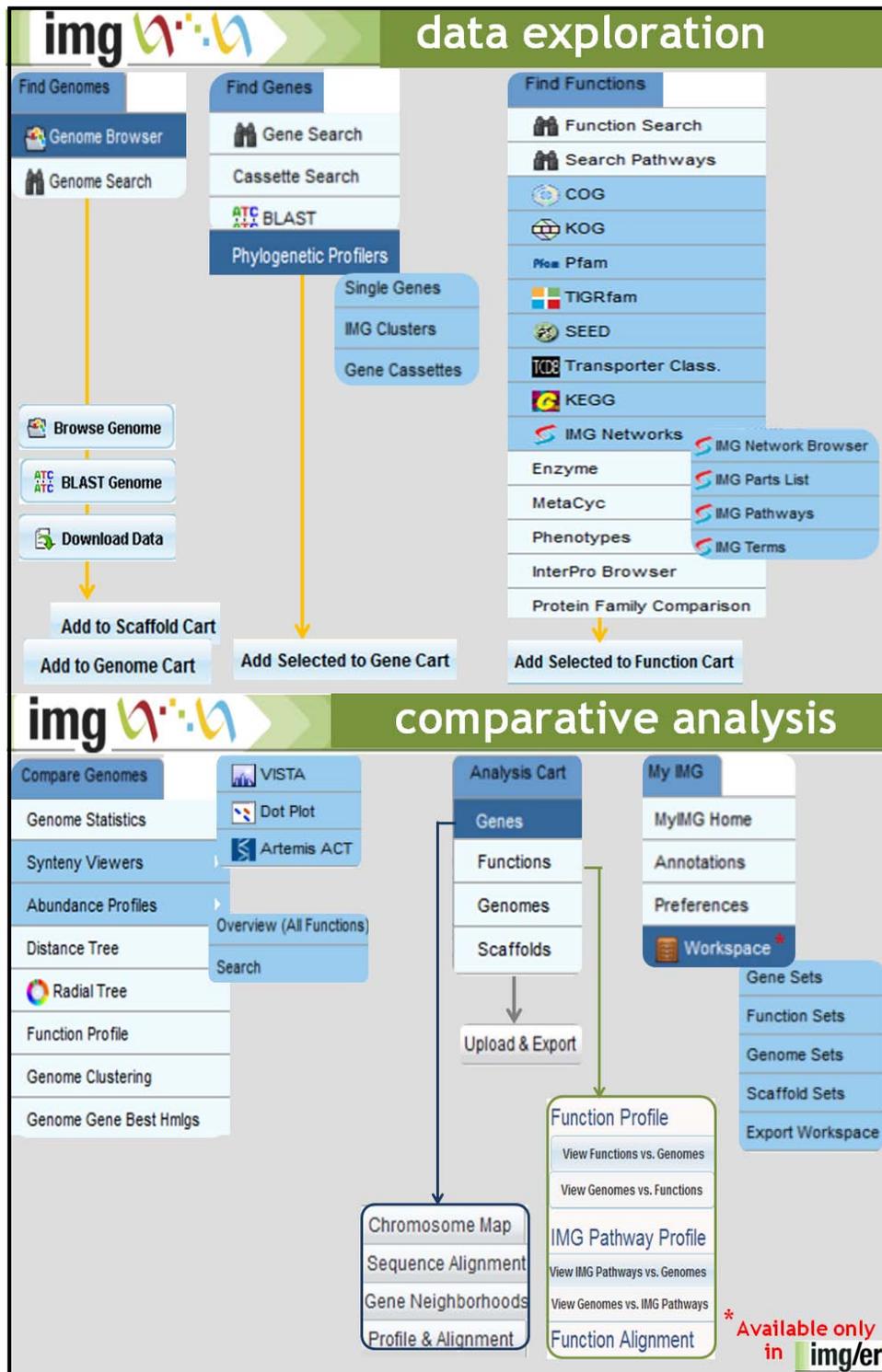


Figure 1. Overview of IMG tools for analyzing genomes in IMG and IMG ER.

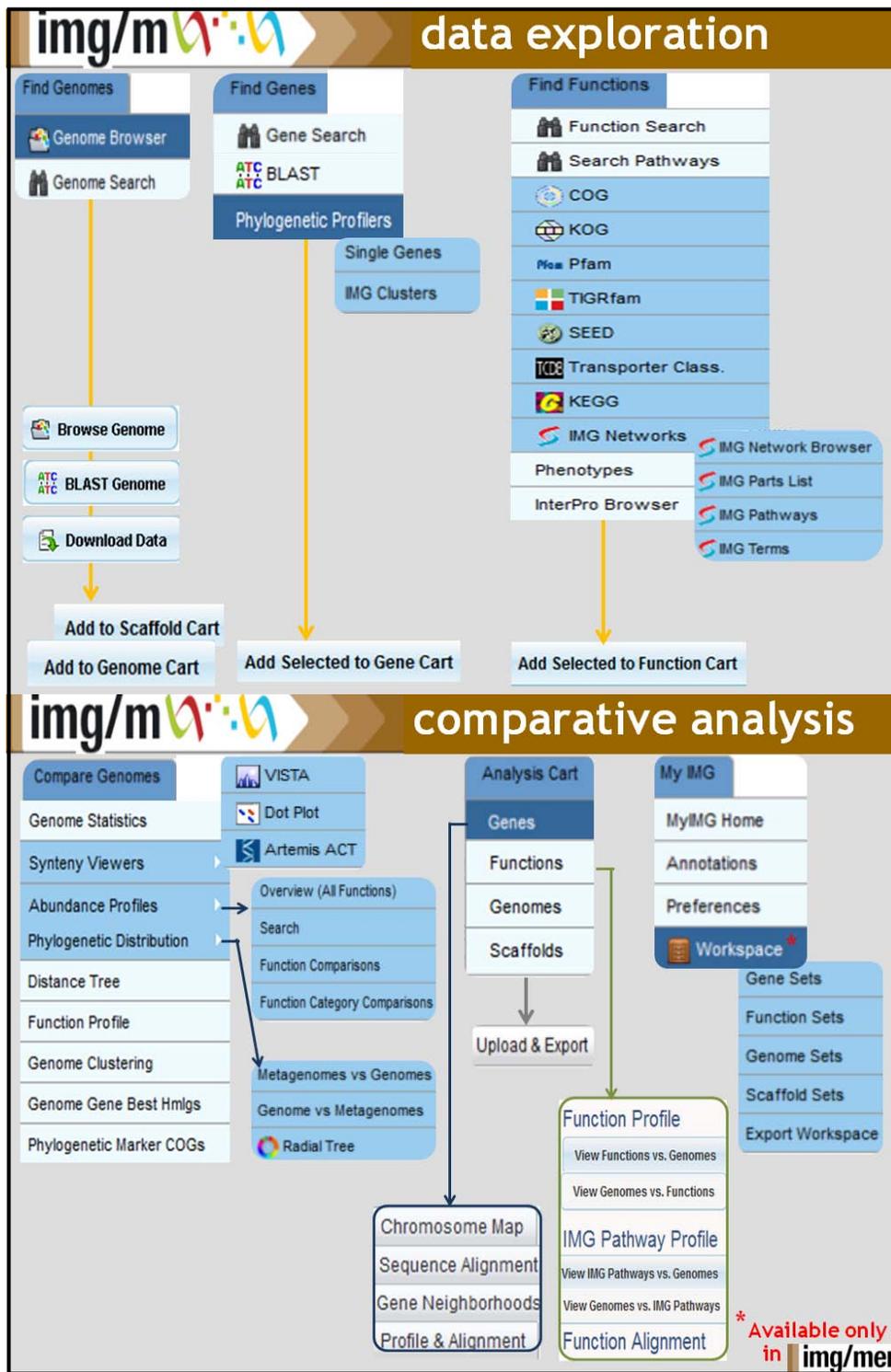


Figure 2. Overview of IMG tools for analyzing metagenomes in IMG/M and IMG/M ER.

## Workspace

**Workspace** tools are available only to IMG registered users as part of the **MyIMG** toolkits, as illustrated in Figure 3(i). These tools allow users to specify, manage and analyze sets of genes, functions, genomes or metagenome samples, and scaffolds.

The figure illustrates the workflow for using workspace tools to specify and analyze sets of metagenome samples. It consists of several interconnected panels:

- Genome Cart (ii):** Shows a list of genome sets with an 'Upload & Export & Save' button highlighted in red. A 'Save Genome to My Workspace' dialog is open, showing 'Save to File name: Human Stool Samples' and a 'Save Selected to Workspace' button.
- My IMG (i):** A navigation menu on the right with 'Workspace' selected, showing options for Gene Sets, Function Sets, Genome Sets, Scaffold Sets, and Export Workspace.
- My Workspace - Genome Sets (v):** Shows '269 items saved to file Human\_Stool\_and\_Buccal\_Samples'. It includes 'Union or Intersection of Sets' options with 'Save Intersection to Workspace' and 'Save Union to Workspace' buttons.
- My Workspace - Genome Sets:** A table listing saved genome sets:
 

Select	File Name	Number of Genomes (click the link to each individual set)
<input checked="" type="checkbox"/>	Human_Buccal_Mucosa_Samples	<a href="#">122</a>
<input checked="" type="checkbox"/>	Human_Stool_Samples	<a href="#">147</a>
- Genome Set Function Profile (iii):** Shows 'Use only functions in set: Arginine\_biosynthesis\_COGs' selected. A 'Run Function Profile' button is highlighted in red.
- Genome Set Function Profile (Arginine\_biosynthesis\_COGs) (iv):** A table showing the results of the function profile:
 

Selection	Function ID	Function Name	Human_Buccal_Mucosa_Samples	Human_Stool_Samples
<input checked="" type="checkbox"/>	COG0002	Acetylglutamate semialdehyde dehydrogenase	<a href="#">461</a>	<a href="#">7141</a>
<input type="checkbox"/>	COG0078	Ornithine carbamoyltransferase	<a href="#">1176</a>	<a href="#">10014</a>

**Figure 3.** Using workspace tools to specify and analyze sets of metagenome samples.

Sets of genes, functions, genomes/metagenome samples and scaffolds can be specified using the Gene, Function, Genome, and Scaffold Cart, respectively. For example, two sets of metagenome samples are first specified using the Genome Cart and then saved as named files into a user specific workspace, as illustrated in Figure 3(ii).

Sets of genes, functions, genomes/metagenome samples and scaffolds can be exported from (downloaded) or imported (uploaded) into IMG’s workspace, and can be involved in set based functional profiles, as illustrated in Figure 3(iii) where two sets of metagenome samples are compared in terms of a predefined set of (*Arginine biosynthesis*) COG functions. The function profile result shown in Figure 3(iv) displays the number of genes associated with a specific function (COG) in the function set, across all the samples in the set of metagenome samples. The genes associated with a specific function can be used to specify a new set of genes in the user’s workspace, as shown in the bottom part of Figure 3(iv). Set operations (intersection,

union) can be applied on sets of genes, functions, genomes and scaffolds, as illustrated in Figure 3(v) where union is applied on two sets of metagenome samples in order to create a new set of samples.

The workspace tools can be used for specifying metagenome or genome **bins** consisting of subsets of scaffolds. For single cell genomes, typically scaffolds are screened for potential contamination, with scaffold sets used for separating “contaminated” scaffolds from “clean” scaffolds (for details, see: <https://img.jgi.doe.gov/er/doc/SingleCellDataDecontamination.pdf>). For metagenomes, scaffold sets are used for specifying individual genomes isolated from the microbial community.