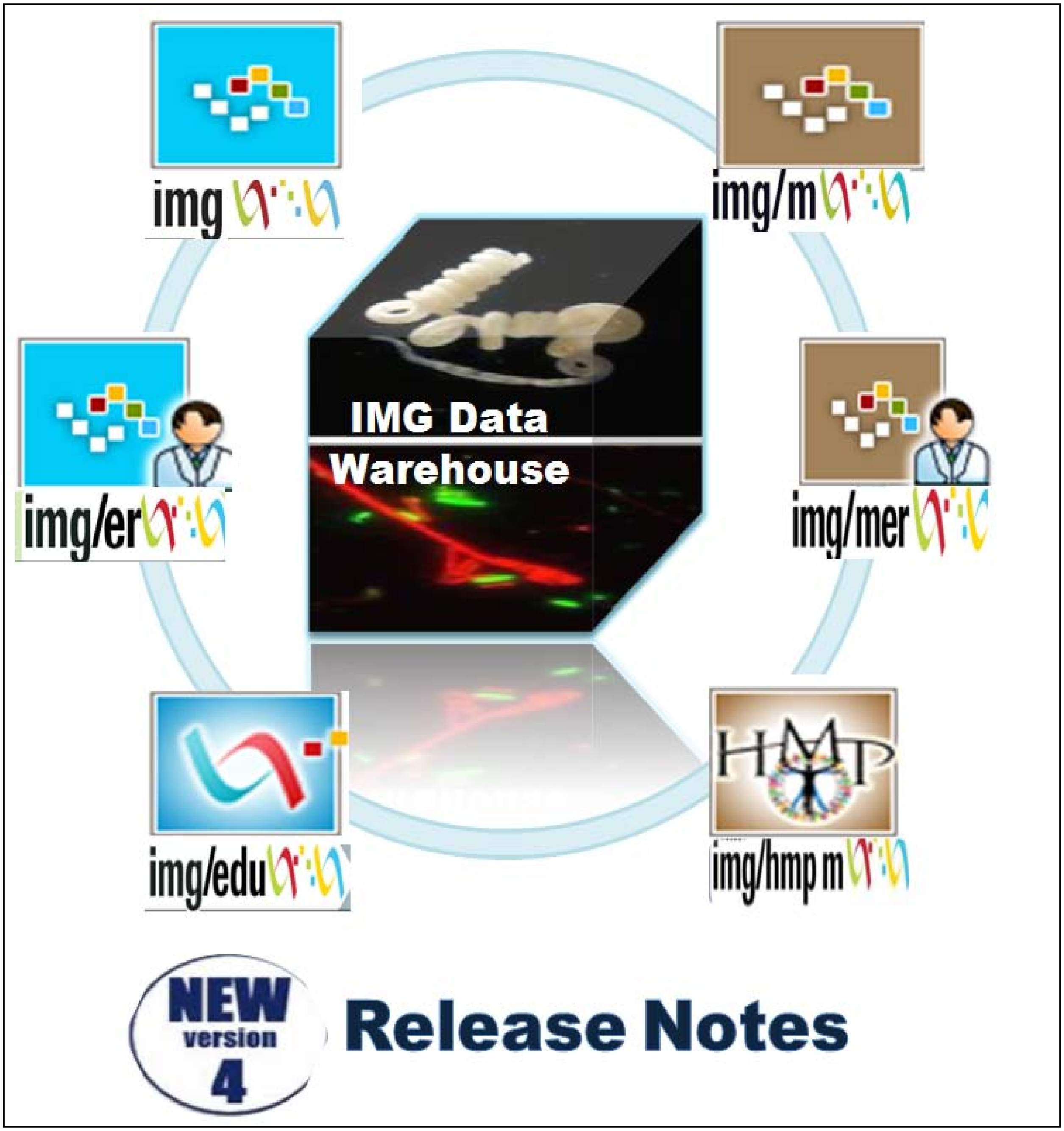

img
img/m
img/er
IMG Data Warehouse
img/mer
img/edu
HMP
img/hmp m
NEW version 4
Release Notes

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.
- 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 *Brachybacterium 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 **December 12th, 2012**, IMG had **2,670 registered users** from **63 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 **Dec 12th, 2012**, IMG contains a total of **12,039** genomes, plasmids and genome fragments, with a total of **27 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.

2,494 genomes in IMG, with a total of **7.5 million** protein coding **genes**, are "private", that is, are owned by the users have password protected access to them. 1,008 genomes in IMG are single cells, distributed among bacterial, archaeal, and eukaryotic genomes as shown in the table below.

Public	**9,545**	**Private**	**2,494**	**Total**	**12,039**
Bacteria	4,525		2,181		6,706
Single Cell	7	Single Cell	838	Single Cell	845
Archaea	180		236		416
		Single Cell	155	Single Cell	155
Eukarya	187		27		214
		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 **Dec 12th, 2012**, IMG contains **2,215 metagenome samples,** with a total of **8.9 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.

1,282 metagenome samples in IMG, with a total of 2 billion protein coding genes, are **publicly** available, are distributed as follows:

Engineered	**39**	**Environmental**	**396**	**Host associated**	**847**
Bioremediation	6	Air	2	Annelida	4
Biotransformation	4	Aquatic	340	Arthropoda	35
Lab enrichment	2	Terrestrial	54	Birds	4
Solid waste	9			Human	753
Wastewater	18			Mammals	21
				Microbial	1
				Mollusca	8
				Plants	18
				Porifera	3

933 metagenome samples in IMG, with a total of 6.8 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)	**2,084**	39	(1.2 Mil)	**39**
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	2,700	2,714 (8.3 Mil)	**12,039**	577	(7.8 Bil)	**2,215**

* 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.

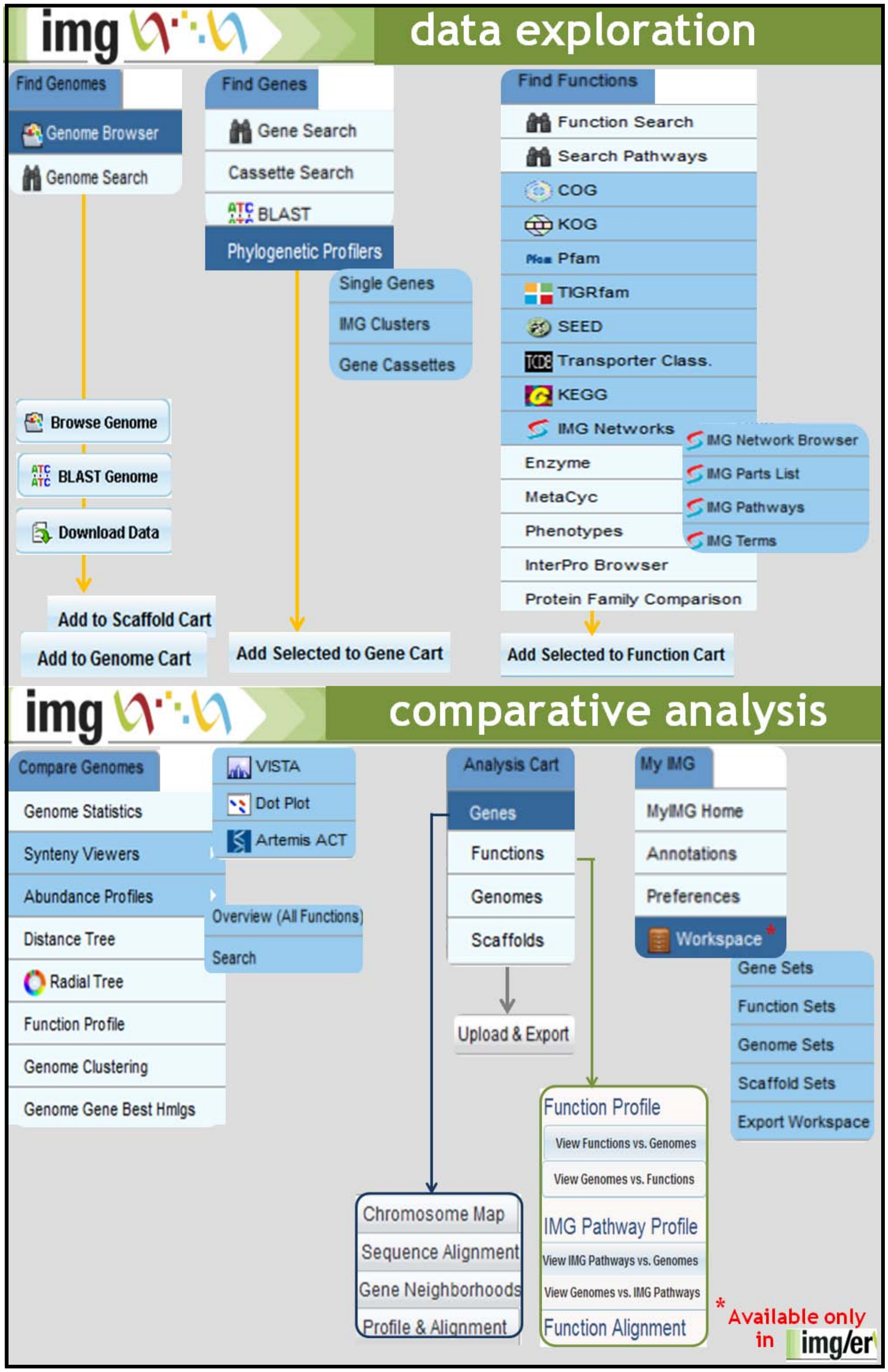


Figure 1. Overview of IMG tools for analyzing genomes in IMG and IMG ER.

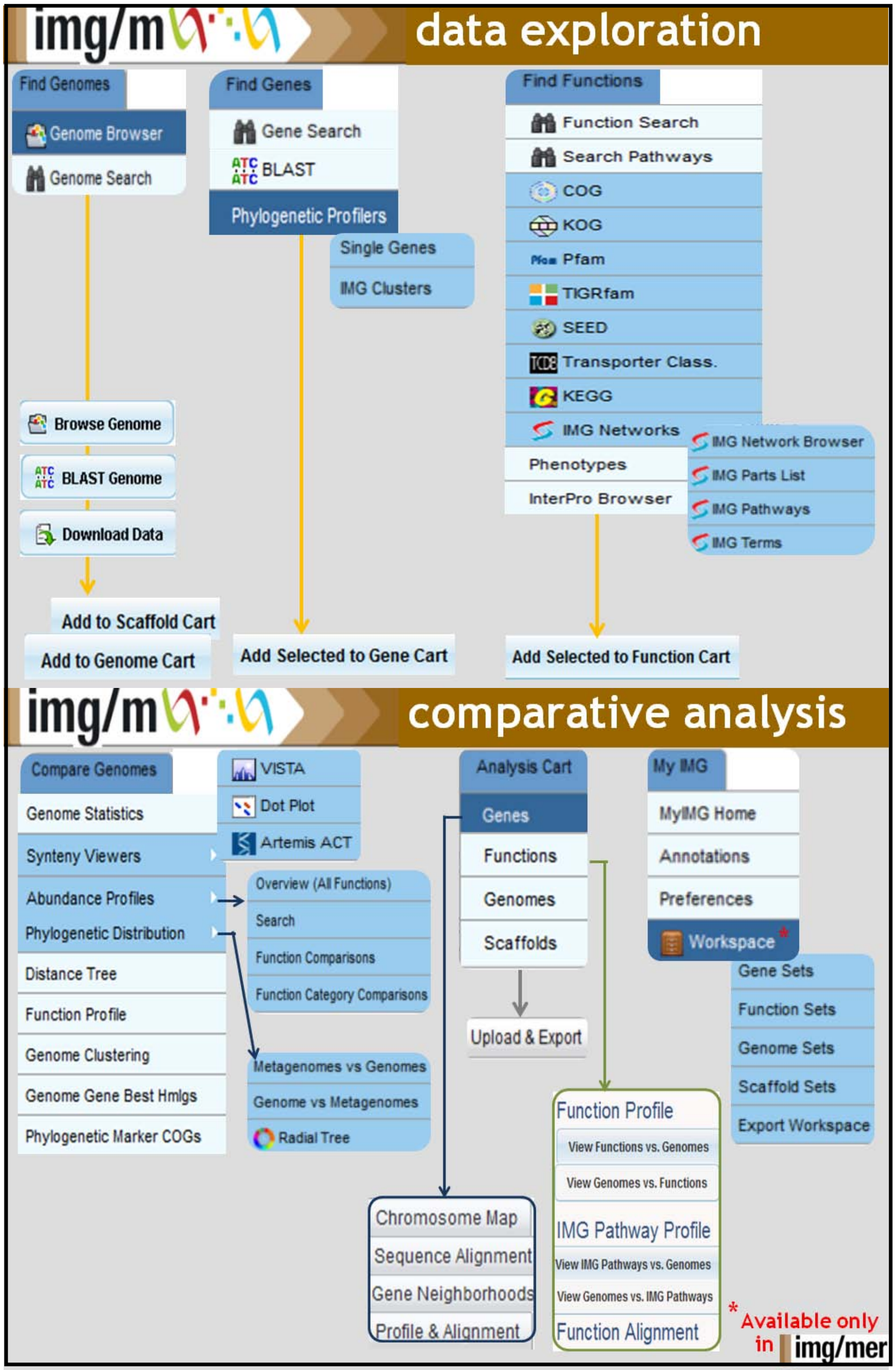
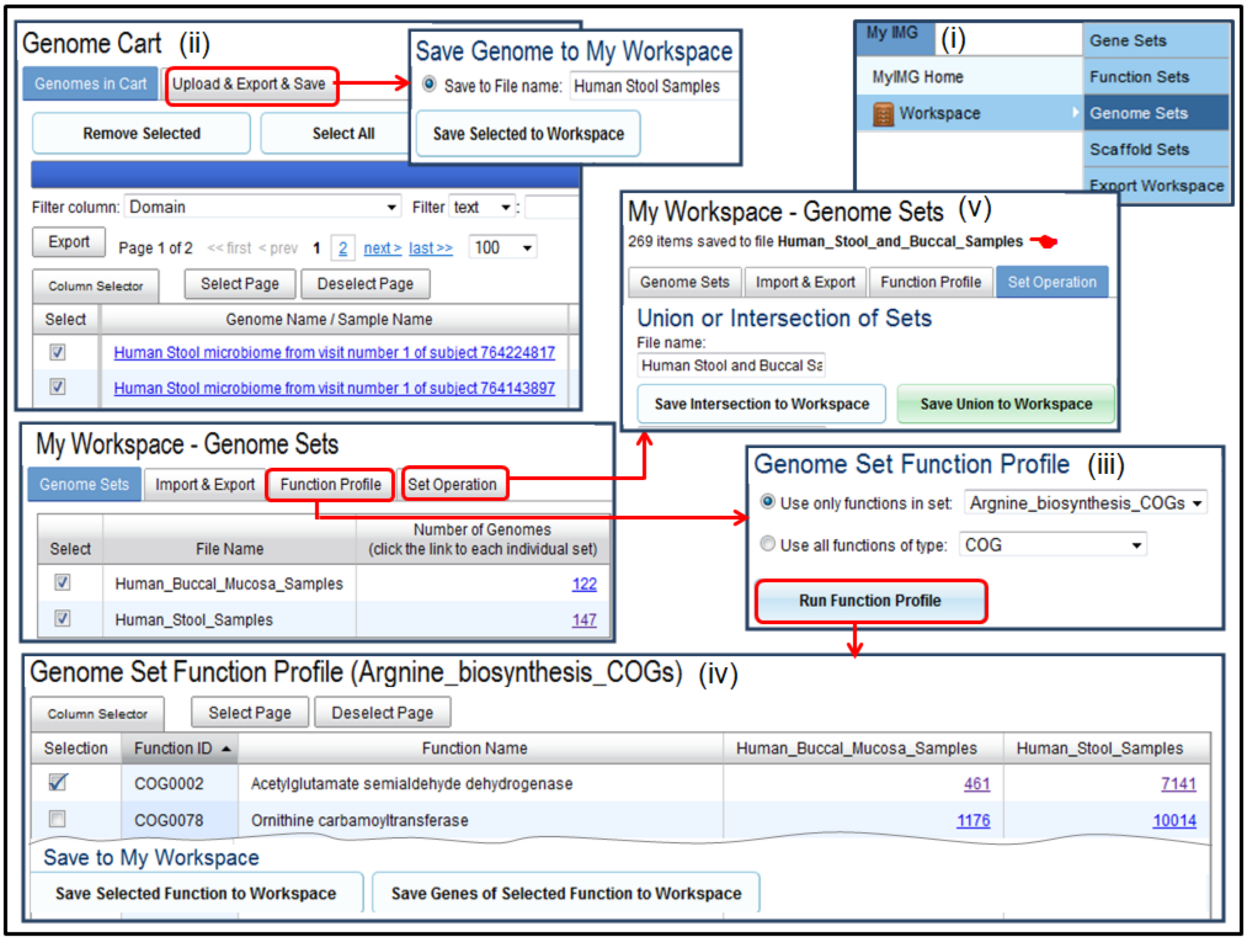


Figure 2. Overview of IMG tools for analyzing metagenomes in IMG/M and IMG/M ER.

Workspace

Workspace tools were released on **November 10th 2012**, and are available to IMG r**egistered 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.



Filter column: Domain Filter text

Select	Genome Name / Sample Name
☑	Human Stool microbiome from visit number 1 of subject 764224817
☑	Human Stool microbiome from visit number 1 of subject 764143897

● Save to File name: Human Stool Samples

269 items saved to file **Human_Stool_and_Buccal_Samples**

File name: Human Stool and Buccal Sa

Select	File Name	Number of Genomes (click the link to each individual set)
☑	Human_Buccal_Mucosa_Samples	122
☑	Human_Stool_Samples	147

● Use only functions in set: Argnine_biosynthesis_COGs

○ Use all functions of type: COG

Selection	Function ID ▲	Function Name	Human_Buccal_Mucosa_Samples	Human_Stool_Samples
☑	COG0002	Acetylglutamate semialdehyde dehydrogenase	461	7141
□	COG0078	Ornithine carbamoyltransferase	1176	10014

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).

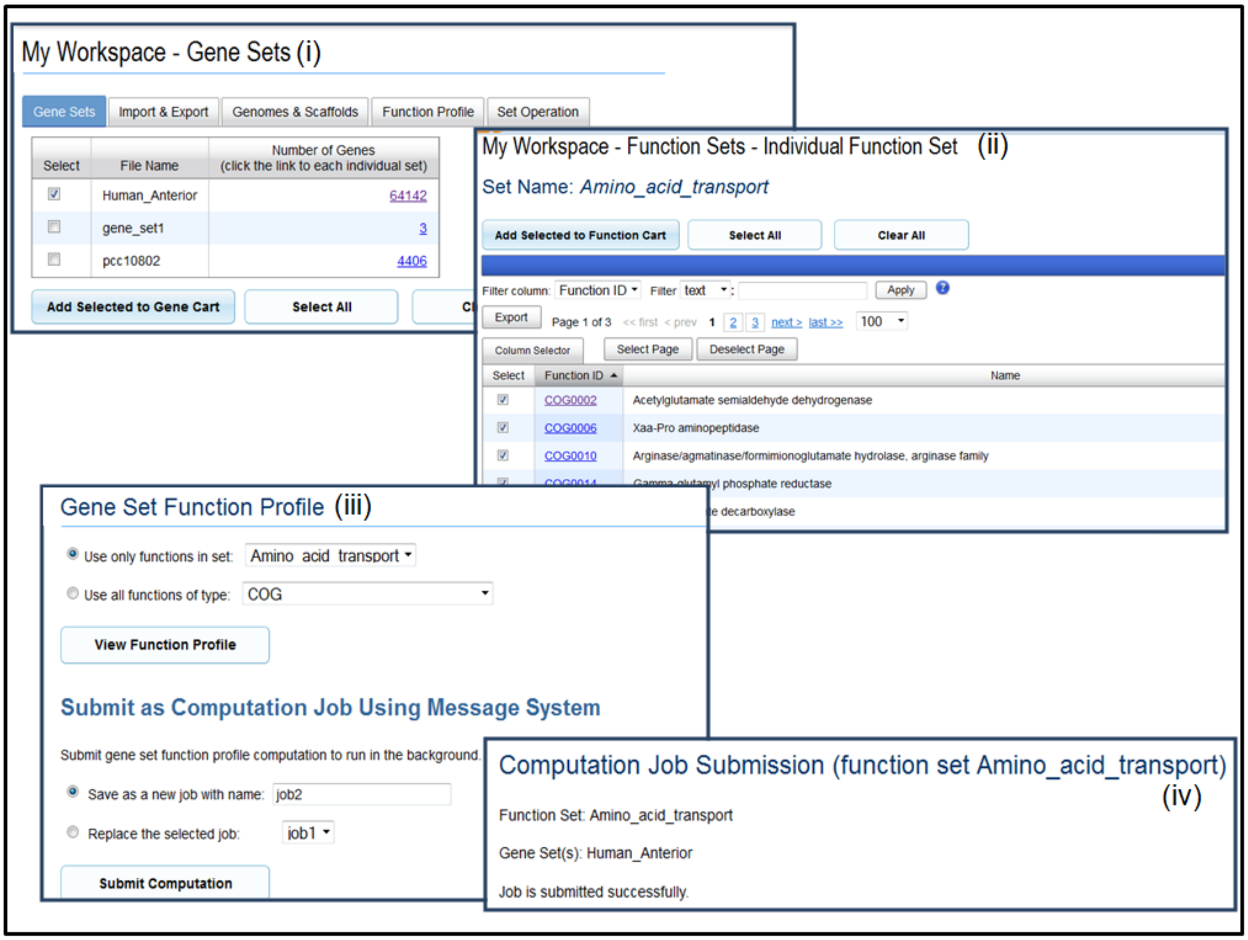
Sets of genes, functions, genomes/metagenome samples and scaffolds can be exported from (downloaded) or imported (uploaded) into IMG's workspace, and can be involved in set based functional profiles, as illustrated in Figure 3(iii) where two sets of metagenome samples are compared in terms of a predefined 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.

Background computations

A mechanism for performing analysis tasks employing **background** (off line) **computations** was released on **December 13th 2012**, and is available to IMG **registered users**. This mechanism targets analyses that involve very large sets (e.g., millions) of genes or a large number (e.g., hundreds) of scaffolds and that would take a long time (e.g., tens of minutes) to complete. Such computations tend to time out in interactive mode and therefore are preferable to be carried out in background (off line).

Users can use the background computation mechanism for the **Workspace** "**Gene Function Profile**", "**Scaffold Function Profile**", and "**Scaffold Phylogenetic Distribution**" analysis tools, as illustrated in Figures 4 and 5 below.



My Workspace - Gene Sets (i)

Select	File Name	Number of Genes (click the link to each individual set)
☑	Human_Anterior	64142
□	gene_set1	3
□	pcc10802	4406

My Workspace - Function Sets - Individual Function Set (ii)

Set Name: *Amino_acid_transport*

Filter column: Function ID ▾ Filter text ▾ : Apply

Select	Function ID ▲	Name
☑	COG0002	Acetylglutamate semialdehyde dehydrogenase
☑	COG0006	Xaa-Pro aminopeptidase
☑	COG0010	Arginase/agmatinase/formimionoglutamate hydrolase, arginase family
☑	COG0014	Gamma-glutamyl phosphate reductase
		e decarboxylase

Gene Set Function Profile (iii)

● Use only functions in set: Amino_acid_transport ▾

○ Use all functions of type: COG ▾

Submit as Computation Job Using Message System

Submit gene set function profile computation to run in the background.

● Save as a new job with name: job2

○ Replace the selected job: job1 ▾

Computation Job Submission (function set Amino_acid_transport) (iv)

Function Set: Amino_acid_transport

Gene Set(s): Human_Anterior

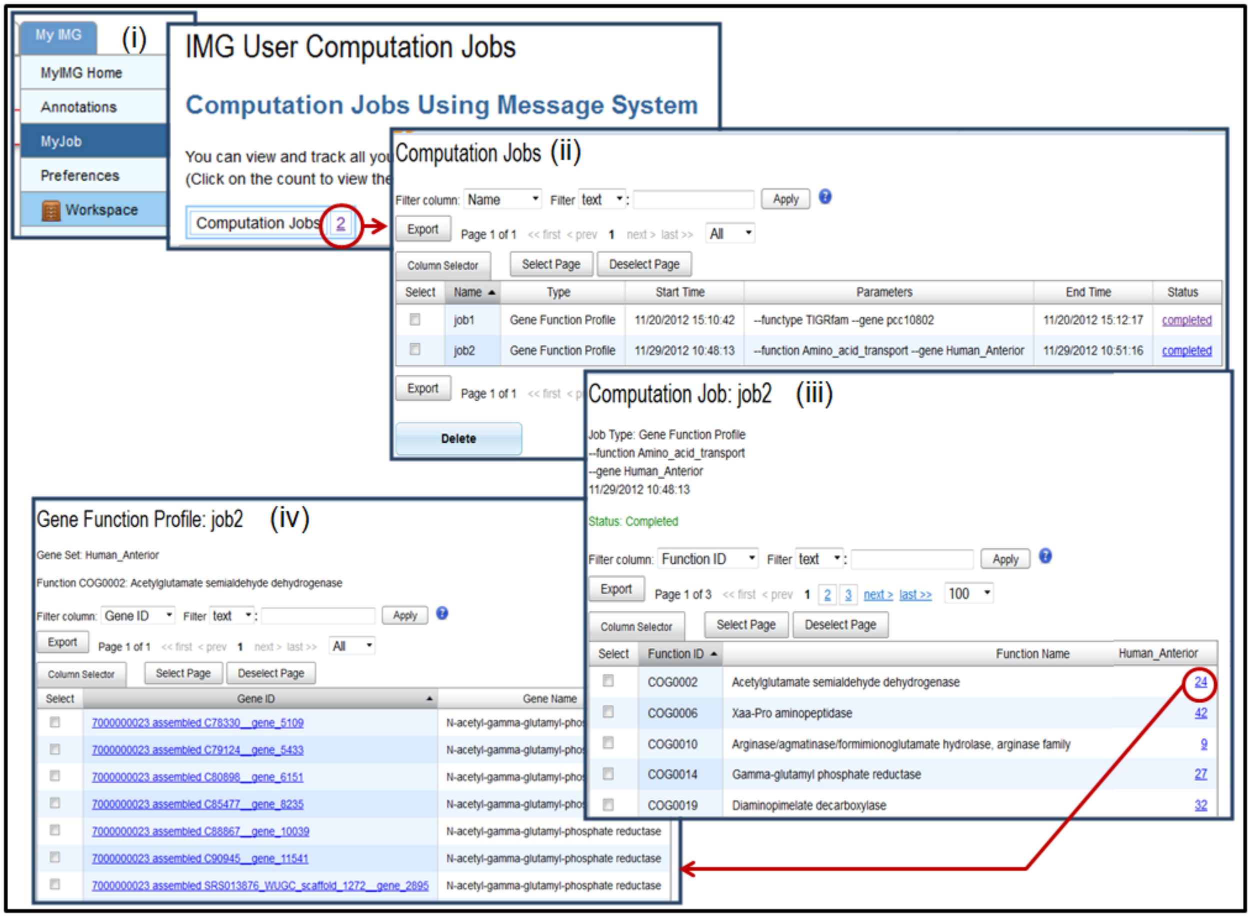
Job is submitted successfully.

Figure 4. Background computations for analysis involving large sets of genes and functions.

For example, a user that has a gene set with 64,000+ human anterior nares microbiome genes, as illustrated in Figure 4(i), may be interested in examining the profile of this gene set across a large number (e.g., 287) of COG and Pfam functions related to amino acid transport and metaboilism, as illustrated in Figure 4(ii). After selecting the "Human_Anterior" gene set in the "Gene Set" page of the Workspace, "Function Profile" is computed using function set

"Amino_acid_transport", as illustrated in Figure 4(iii). For employing the background computation mechanism, the user needs to enter or select a job name and then select "Submit Computation" instead of selecting "View Function Profile" which is employed for interactive analysis. Once the background computation is submitted, the user is informed as illustrated in Figure 4(iv).

The status of a background computation is available via the "MyJob" part of the "MyIMG" menu option, as illustrated in Figure 5(i) and 5(ii).



IMG User Computation Jobs

Computation Jobs Using Message System

You can view and track all yo
(Click on the count to view the

Computation Jobs 2

Computation Jobs

Filter column: Name Filter text :

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

Select	Name	Type	Start Time	Parameters	End Time	Status
□	job1	Gene Function Profile	11/20/2012 15:10:42	--functype TIGRfam --gene pcc10802	11/20/2012 15:12:17	completed
□	job2	Gene Function Profile	11/29/2012 10:48:13	--function Amino_acid_transport --gene Human_Anterior	11/29/2012 10:51:16	completed

Page 1 of 1 << first < p

Computation Job: job2

Job Type: Gene Function Profile
--function Amino_acid_transport
--gene Human_Anterior
11/29/2012 10:48:13

Status: Completed

Filter column: Function ID Filter text :

Page 1 of 3 << first < prev 1 2 3 next > last >> 100

Select	Function ID	Function Name	Human_Anterior
□	COG0002	Acetylglutamate semialdehyde dehydrogenase	24
□	COG0006	Xaa-Pro aminopeptidase	42
□	COG0010	Arginase/agmatinase/formimionoglutamate hydrolase, arginase family	9
□	COG0014	Gamma-glutamyl phosphate reductase	27
□	COG0019	Diaminopimelate decarboxylase	32

Gene Function Profile: job2

Gene Set: Human_Anterior

Function COG0002: Acetylglutamate semialdehyde dehydrogenase

Filter column: Gene ID Filter text :

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

Select	Gene ID	Gene Name
□	7000000023 assembled C78330__gene_5109	N-acetyl-gamma-glutamyl-pho
□	7000000023 assembled C79124__gene_5433	N-acetyl-gamma-glutamyl-pho
□	7000000023 assembled C80898__gene_6151	N-acetyl-gamma-glutamyl-pho
□	7000000023 assembled C85477__gene_8235	N-acetyl-gamma-glutamyl-pho
□	7000000023 assembled C88867__gene_10039	N-acetyl-gamma-glutamyl-phosphate reductase
□	7000000023 assembled C90945__gene_11541	N-acetyl-gamma-glutamyl-phosphate reductase
□	7000000023 assembled SRS013876_WUGC_scaffold_1272__gene_2895	N-acetyl-gamma-glutamyl-phosphate reductase

Figure 5. Examining the status of background computations and associated analysis results.

When a computation (e.g., job 2) is "completed", links are provided for accessing the analysis results, as illustrated in Figure 5(iii), and associated details, as illustrated in Figure 5(iv).

The results of background computations are saved until users either explicitly delete the jobs using the "Delete" option in the "Computation Jobs" page (see Figure 5(ii)) , or override them with new jobs using the "Replace the selected job" option.