

IMG/GOLD Glossary

Abundance Profile provides a count of the number of genes (or gene copies) across protein family collections (eg COG, Pfam, etc) for a set of selected datasets. IMG is using a variety of different Abundance Profile Tools which are available [here](#)

Analysis Project (AP) (aka **GOLD Analysis Project**) is the informatics processing of a Sequencing Project in IMG and GOLD. It describes how the assembly and annotation of a Sequencing Project were performed. Individual submissions in IMG correspond to individual Analysis Projects. Accordingly, for each IMG project ID there is a corresponding AP ID from GOLD (GOLD AP).

Analysis Project Type (aka **GOLD Analysis Project Type**) is determined by the sequencing project type and basically describes the annotation process applied. Common examples are isolate genome analysis, metagenome analysis, metatranscriptome analysis, etc.

ANI stands for Average Nucleotide Identity, which is a measure of nucleotide-level genomic similarity between the coding regions of two genomes. More information on how IMG is using ANI is available [here](#) and in the following publication (PMID: [26150420](#))

ANI Cluster is a set of genomes that have been grouped together based on a certain ANI threshold. Each cluster has a unique ID called **ANI Cluster ID**.

ANI Cluster Type represents the type of connection the various genomes have within each cluster. If all genomes within a cluster are connected to each other with a value about a certain threshold (predetermined by large scale analysis), then the cluster type is a **Clique**. If some genomes within a cluster are connected with values below that threshold, then the Cluster type a **Clique group**.

Binning is the process of grouping reads or contigs and assigning them to operational taxonomic units. More information on the binning process used in IMG is available [here](#)

Biome represents the environmental sample selected for sequencing

Biosample is defined in GOLD and represents the original isolation place of the physical sample, from where the DNA was isolated. It is usually the description of the environment from where the sample was taken. In the case of isolate genomes where the sample isolation environment is not known, the name of the organism may be the Biosample name. Biosample is nothing more than metadata on the sequencing project. However, due to the fact that all DNA samples are coming from a Biosample and because under one Biosample, several different DNA samples may be extracted, it is forming a level above Seq Project.

Biosynthetic Cluster refers to computationally predicted biosynthetic gene clusters (BGC) for the production of secondary metabolites. Predictions are based on [antiSMASH ver 5.0](#).

Cassette or Chromosomal Cassette or Gene Cassette merely refers to co-located genes in a chromosome, that are identified as a series of CDS with intergenic distance \leq 300 nucleotides. These are only computed for isolate genomes

Contig is a consensus sequence of a continuous DNA fragment generated from a set of overlapping reads; as a result the bases, their order, as well as the length of the fragment are known with high

confidence. This is in contrast to **scaffold**, which is reconstructed based on information generated as a result of partial sequencing of DNA fragments (e. g. End-sequencing of clones), and therefore has gaps represented by Ns and has uncertain length.

Culture Type represents the type of the culture from which the organism has been obtained. The Culture type has two values, it can be either an isolate or co-culture

Draft is a type of “sequencing status” of a genome project which is at an incomplete or “draft” stage, and unlike “permanent draft” status, further sequencing improvements or gap closures may be planned.

Ecosystem classification (aka GOLD Ecosystem Classification) is a five level classification scheme describing the environment from which an environmental sample or an organism was collected. This five level hierarchical classification system was described here: [20653767](#). Ecosystem at the top describes the broader environment (Environmental, Engineered and Host-associated) and at the specific ecosystem at the bottom of this hierarchy refers to the specific feature of the environment as shown in the following example.

The five levels of ecosystem classification include: Ecosystem -> Ecosystem Category -> Ecosystem Type -> Ecosystem Subtype -> Specific Ecosystem

Example Path: Environmental -> Aquatic -> Marine -> Oceanic -> Aphotic zone

Ecotype: is a population of a species that survives as a distinct group through environmental selection and isolation and that is comparable with a taxonomic subspecies, but not yet classified as a subspecies.

Estimated Gene Copies is the average coverage (read depth) of a gene in a metagenome. It is used in computing gene abundances in order to account for different population abundance in the community. Estimated gene copies are approximated by the average contig coverage and computed as the average number of reads aligned to each base of the contig.

FASTA refers to a text-based file format representing either nucleotide or amino acid sequences. The first line in a FASTA file starts with a ">" (greater-than) symbol and holds a summary description of the sequence, often starting with a unique accession number or identifier or description of the sequence to follow. Following the initial line is the actual sequence itself in standard one-letter character string. (description adapted from [Wikipedia](#))

Finished is the status of a sequencing project when the genome sequences have less than 1 error per 100,000 base pairs and where each replicon is assembled into a single contiguous sequence with a minimal number of possible exceptions commented in the submission record. All sequences are complete and have been reviewed and edited, all known misassemblies have been resolved, and repetitive sequences have been ordered and correctly assembled. The definition is following community standards as defined [here](#)

Function Profile is an IMG tool for visualizing the abundance of protein families of interest in a (meta)genome or scaffold/contig set of interest. It essentially represents a matrix with rows representing (meta) genomes or scaffolds/contigs and columns representing protein families, whereby each cell contains the count of protein family in (meta)genome or scaffold/contig.

Fused Protein (or Gene Fusion) is a protein consisting of at least two domains that are encoded by separate genes that have been joined so that they are transcribed and translated as a single unit, producing a single polypeptide

Genome Fragments are DNA sequences representing inserts in plasmids, cosmids, fosmids and other vectors that were typically generated from efforts to sequence biosynthetic gene clusters for secondary metabolites. They were imported from NCBI and believed to be experimentally validated or characterized.

GOLD Classification is a Habitat based ecosystem classification of Biosamples or organisms organized in a five-tiered scheme as described [here](#)

GOLD Sequencing Project ID is a unique database identifier assigned to every sequencing project.

GOLD Sequencing Quality represents community-defined categories of standards that better reflect the quality of the genome sequence, based on our understanding of the technologies, available assemblers, and efforts to improve upon drafted genomes. The values are based on the [Chain et al.](#) publication.

GPTS Proposal ID is a unique legacy JGI proposal ID assigned to old JGI sequencing projects.

Habitat: Natural environment of an organism or biosample; place that is natural for the life and growth of an organism or a general description of the place where a biosample was collected from. E.g. Wetland, Human skin etc..

Homologs in IMG merely refers to a similar sequence (matched by BLAST or some other pairwise alignment method) and does not necessarily imply shared ancestry.

High Quality is a quality metric IMG assigns to genomes based on the following rules (all of which must be met):

- Genome is not a MAG or unscreened SAG (I guess we've t add MEGs to it soon)
- GOLD phylogeny is neither 'UNCLASSIFIED' nor 'UNCLASSIFIED-ARCHAEAL' nor 'UNCLASSIFIED-BACTERIAL'
- The coding density is between 70% and 100%
- Genes per million bases is either between 700 and 1400
- Sequences per million bases is not greater than 300

HMP stands for the Human Microbiome Project

HMP ID is a unique ID assigned to a dataset from the HMP's Data Analysis and Coordination Center ([DACC](#)) project catalogue.

IMG Submission ID is a unique ID a dataset receives when submitted to the IMG annotation pipeline

IMG Terms/Pathways/Parts/Network: refers to IMG specific functional curation efforts as described here: [23424620](#)

ITS Proposal ID is a unique ID assigned to all the proposals approved for sequencing at the JGI

JGI Genome Portal: is a centralized resource at the JGI for data download and can be accessed [here](#)

JGI Project ID / ITS SP ID are both unique IDs generated from the JGI production sequencing pipeline, and are available only for JGI sequenced projects

Locus Type in IMG is a subset of “feature keys” defined by the International Nucleotide Sequence Database Collaboration (INSDC, <http://www.insdc.org/>) used to describe regions of nucleotide sequence performing a certain biological function, affecting or resulting from the expression of a certain biological function, affecting replication, etc. The subset available in IMG includes CDS (protein coding sequence), rRNA, tRNA, miscRNA.

MAGs are genomes that have been reconstructed through assembly and binning from metagenomes (standing for Metagenome Assembled Genomes).

Metadata: Supplementary data linked to DNA sequences that provide information in a standard and searchable way such as organismal or bulk environmental sample description. Metadata fields and content displayed in IMG are adapted from the Genomes Online Database (GOLD).

Metagenomics: The study of genetic material isolated directly from environmental samples, such as water, soil or sediments, may also be referred to as environmental genomics, ecogenomics or community genomics.

Metatranscriptomics: The study of the expressed portion of genomes, mRNAs, isolated directly from an environmental sample that may be transcribed into cDNAs for high-throughput sequencing.

Metagenome Bins are scaffold sets grouped based on IMG’s Metagenome binning process. More information is available [here](#)

Metagenome - Cell Enrichment is a draft metagenome assembly derived from a cell enrichment (> 1 cell) sample. A cell enrichment is generally obtained by physical separation of a biologically relevant unit, such as microcolonies. Due to the low biomass for cell enrichments, the extracted DNA is typically amplified using whole-genome amplification prior to sequencing.

Metagenome - Single Particle Sort is a draft genome or metagenome assembly derived from a single particle isolated via flow cytometry. A single particle sort can consist of a single cell or an aggregate of multiple cells, not necessarily of the same phylogenetic background. The extracted DNA is amplified using whole-genome amplification prior to sequencing. No amplicon-based 16S rRNA gene information is available for single particle sorts

Mixed Analysis Group is a user created group composed of isolate genomes and metagenome scaffold bins for the purposes of statistical analysis only. Mixed analysis groups can be created and accessed within the workspace only.

Organism is an individual living entity. It can be plant, fungus, microbe etc

Organism Type refers to the origins of the organism and can be any of the following terms: Natural, Genetically modified, Hybrid, and Synthesized

Obsolete Genome: is a genome that has been removed from IMG's database due to either an error that was identified on the sequence or due to replacement from a new version. Obsolete genomes may still be available for downloading from JGI's Genome Portal.

Ortholog: is identified within IMG is based on reciprocal best BLAST hit, which may be used to conclude identical or equivalent function when comparing genes from two isolate genomes.

Paralog is identified in IMG based on BLAST hits within the same genome.

Phenotype denotes the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment. The term covers the organism's morphology or physical form and structure, its developmental processes, its biochemical and physiological properties, its behavior, and the products of behavior.

Permanent Draft is a status of a genome project which is at a Draft stage and no other sequencing improvements or gap closures are planned.

ProPortal is an IMG DataMart focused on the analysis of Prochlorococcus (and related) species datasets. More information is available [here](#)

Proportal Clade is a taxonomy assignment for Cyanobacterial lineages only, as specified by ProPortal developers. Proportal clade is available through the IMG-Proportal system:
<https://img.jgi.doe.gov/cgi-bin/proportal/main.cgi>

SAG stands for Single Amplified Genome. SAGs are obtained through the use of Whole Genome Amplification (WGA) methods on Single Cells.

Scaffold is a portion of the genome sequence reconstructed from partially sequenced DNA fragments, such as end-sequencing of clones or and paired-end short reads or optical mapping and various technologies for linking of the adjacent DNA fragments, such as Hi-C or 10X Genomics technology . . Scaffolds are composed of contigs and gaps.

Scaffold Lineage is an IMG term describing the predicted taxonomic affiliation of a scaffold computed based on the last common ancestor of LAST hits (against IMG-NR isolates database) of the genes on the scaffold, where at least 50% of the genes have LAST hits against the database.

Scaffold Read Depth (also referred to as scaffold/contig average coverage) is the average number of reads aligned to each base of the contig. Read depth is calculated by aligning the reads to the contig and can be extracted from the output of read aligner, such as bmap or Bowtie.

Scaffold Set is a group of scaffolds that IMG users can select from any project and save them on their workspace

Sequencing Project Sequencing Project is the individual organism or sample that is targeted for sequencing. An individual genome project may be composed of more than one sequencing reactions and/or sequencing technologies. A sequencing project may be an isolate genome, or a Metagenome sample, or a transcriptome, or a metatranscriptome, or a 16S survey, etc. From a single Biosample,

multiple different sequencing projects may be performed. For JGI projects, one sequencing project must always be correlated with a single SPID

Sequencing Status is pertinent to isolate genomes and can be specified by the data submitter. Options are Finished, Draft or Permanent draft (see above for individual descriptions). For metagenomes, SAGs, MAGs, and all other product types, the status is set to “permanent draft”.

Sequencing Strategy: refers to any of the following materials or approaches used for DNA/RNA sequencing: DNA synthesis, Genome fragments, Metagenome, Metatranscriptome, Plasmid, Resequencing, Transcriptome, Transposon Mutagenesis Sequencing, Whole Genome Sequencing.

Single cell (“screened” or “unscreened”) refers to Single Amplified Genome (SAGs) that are obtained using Whole Genome Amplification (WGA) methods on Single Cells. Assembled sequences may be subject to further processing such as screening for contaminant sequences by comparing against a reference database of common contaminants, and removing these sequences that don't belong to the targeted single cell. “Unscreened” implies this contaminant screening was not performed.

Single Particle Sort (see Metagenome- Single Particle Sort above)

Specimen (aka **GOLD Specimen**) refers to the sequencing material source either an Organism or Biome

Study Name (aka **GOLD Study Name**) is an overarching project name encompassing a list of sequencing projects (isolates, metagenomes, SAGs, etc.) that are part of the original research proposal. E.g., HMP study, GEBA study.

Submission Type flag is used to indicate which dataset should be included in the IMG reference data set. It can be either primary or reanalysis. Only primary datasets get included in the IMG reference dataset.

Type Strain is typically an alphanumeric string designating type strain status of an isolate genome. Type strain is the strain which was used when the species was first described, and is typically deposited and retrievable from service culture collections like DSMZ, ATCC, etc.

Uncultured Type denotes how an uncultured organism was obtained. This applies both for real uncultured organisms as well as virtual organisms of metagenomic origin. The Uncultured type can be Single Cell, Pooled Single Cells, Population enrichment, Metagenomic etc.

WGS: stands for Whole Genome Sequencing.

Workspace is only available in IMG/MER or IMG/ABC portals. Users can save contents of genome-, gene-, function-, or scaffold carts for later use in Workspace and even share datasets privately with other IMG users. Additional functionality is available for datasets stored in workspace (as opposed to carts) such as larger computation requests (like BLAST) or [statistical analysis](#) of genome sets or groups.