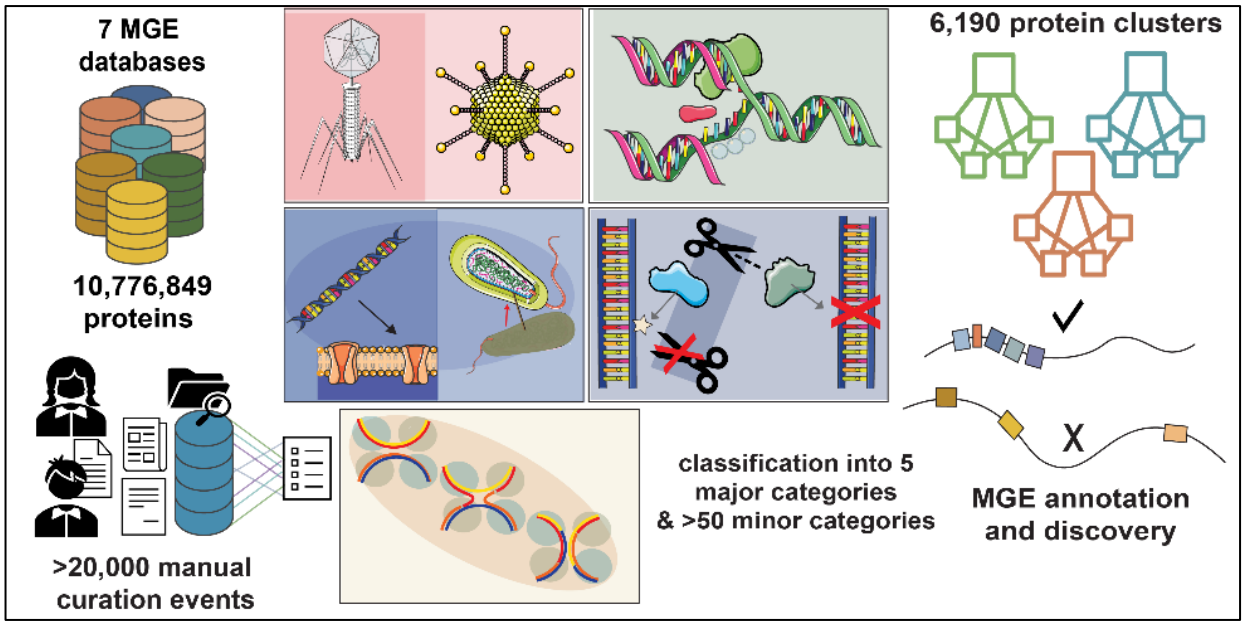
mobileOG-db | Release name: beatrix-1.6

Link to downloads: mobileOG-db.flsi-vt.edu/downloads

Description of mobileOG-db



mobileOG-db (for mobile orthologous groups) is a manually curated database of protein families mediating the integration/excision, replication/recombination/repair, stability/defense, or transfer of bacterial mobile genetic elements and phages as well as the associated transcriptional regulators of these processes. mobileOG-db was created in a process involving more than 20,000 unique manual curation events, with 1,558 references and 2,444 unique manual functional annotations at present.

Description of the mobileOG-db categories

It is helpful for annotating mobile genetic elements to delineate their genes into distinctive functional modules. We identified a core set of categories that comprise key steps of the mobile genetic element "lifestyle." These include (1) replication/recombination/repair of element nucleic acid; (2) stability/transfer/defense;

(3) transfer through conjugation or natural competence; (4) excision/integration; (5) phage related biological processes.

Usage Recommendations (for annotating full length proteins)

An example shell script and an R script used for analysis are provided (here). Recommended workflow:

1. Annotate open reading frames of contigs/scaffolds/genomes using ORF prediction software such as prodigal (recommended)
2. Search ORFs against diamond database. *ensure that complete headers are recovered from the mobileOGs by including -outfmt 6 sseqid or stitle*
3. Merge data frame with mobileOG metadata.
4. Calculate contig-element purity and choose cut-off value. *In our evaluations, we found that higher purities performed better for annotating phages, and lower purity values performed better for plasmids.*

Some caveats.

In general, it is recommended that a successful MGE detection should have multiple hits to multiple modules, all tending to agree on a consensus element type. See R scripts for tentative annotation recommendations.

Excision/integration module hits that also have an RRR tag might be housekeeping genes *xerC/xerD* or others. Likewise, replication/recombination/repair modules alone should not be interpreted as probable MGEs.

Hits to T4SS systems in the conjugation module are not necessarily indicative of MGEs. Paralogs are associated with virulence in some organisms.

Interpreting mobileOG protein headers

mobileOG headers contain helpful information about the sequence origin, mobileOG category, element type, and annotation strategy employed in recovering the sequence, in a semi-colon separated format. Users can split this header (delimiter: "|") to extract metadata associated with the protein sequence to aid annotation efforts.

Header Entry	Meaning
mobileOG_00000828|	*mobileOG ID*
8|	*gene name*
T1PVL4|	*best hit accession ID*
phage|	*major mobileOG category*
structural|	*minor mobileOG category*
Multiple|	*source database(s)*
Manual	*evidence type*

Example R code to separate the header field would be:

```
separate(df, Header, into=c("mobileOG ID",
"Gene Name", "Best Hit ID", "Major mobileOG
Category", "Minor mobileOG Category", "Source
Database(s)", "Evidence"), sep="\\|")
```

Database File Descriptions and Content

mobileOG-db-beatrix-1.* | These files (.faa, .csv) are the core of the web interface for mobileOG-db. Includes homologs + manually curated sequences.
mobileOG-db-beatrix-1.*.All | These files are the complete collection of sequences (keyword search-recovered sequences, homologs, and manually curated sequences).
mobileOG-db-beatrix-1.*.MC | metadata + sequences for manually curated entries
mobileOG-db-beatrix-1.*.KW | metadata + sequences for sequences recovered from keyword searches

mobileOG-db-beatrix-1.* --> upload to create website
mobileOG-db-beatrix-1.*.All --> available as standalone download on downloads page
mobileOG-db-beatrix-1.*.MC --> available as standalone download on downloads page
mobileOG-db-beatrix-1.*.KW --> available as standalone download on downloads page

Table 1. mobileOG-db metadata content and descriptions

mobileOG Entry Name	Unique entry ID for each mobileOG entry. Sequence is directly derived from a mobile genetic element in one of the eight databases
Best Hit ID	The UniProt or NCBI identifier for the best hit
mobileOG Cluster	The cluster representative for the sequence cluster containing this mobileOG entry
Name	Manually curated gene name ascribed to entry
Manual Annotation	Annotation or description of literature-based evidence for gene function
Major mobileOG Category	Major category manually selected for this entry (in the case of

	Evidence=Manual)
Minor mobileOG Categories	Minor category manually selected for this entry (in the case of Evidence=Manual)
Reference(s)	DOI for publication associated with manually curated entry
Evidence	Evidence used for inclusion (manual, homology, or keyword search)
Taxonomy	Taxonomy of best hit
Database	Database of origin
Multiple (YN)	Binary (Y/N) as to whether this sequence is found in multiple databases
Total number of identical sequences	Total number of identical sequences
mobileOG fasta Header	Header of the fasta sequence for that entry
Amino Acid Sequence	Amino acid sequence of the entry
ACLAME	Number of entries occurring in each of the databases
PlasmidRefSeq	
COMPASS	
pVOG	
GPD	
immedb	
ICE	
IME	
CIME	
AICE	
ISFinder	
Taxonomic lineage (ALL)	Full taxonomic lineage of best hit

Citing Us

mobileOG-db: a manually curated database of protein families mediating the life cycle of bacterial mobile genetic elements Connor L. Brown, James Mullet, Fadi Hindi, James E. Stoll, Suraj Gupta, Minyoung Choi, Ishi Keenum, Peter Vikesland, Amy Pruden, Liqing Zhang bioRxiv 2021.08.27.457951; doi: https://doi.org/10.1101/2021.08.27.457951

Acknowledgements

To create mobileOG-db, we analyzed 10,776,212 proteins derived from 8 different databases of complete MGE sequences. The references to these databases are below.

1. **ICEBerg (ICEs, AICEs, CIMEs, IMEs):**

Meng Liu, Xiaobin Li, Yingzhou Xie, Dexi Bi, Jingyong Sun, Jun Li, Cui Tai, Zixin Deng, Hong-Yu Ou, ICEberg 2.0: an updated database of bacterial integrative and conjugative elements, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019,

Pages D660– D665, https://doi.org/10.1093/nar/gky1123

Downloaded with: wget
https://bioinfo-mml.sjtu.edu.cn/ICEberg2/download/ICE_aa_all.fas

2. **ACLAME (various):**

Leplae R, Lima-Mendez G, Toussaint A. ACLAME: a CLAssification of Mobile genetic Elements, update 2010. Nucleic Acids Res. 2010 Jan;38(Database issue):D57-61. doi: 10.1093/nar/gkp938. Epub 2009 Nov 23. PMID: 19933762; PMCID: PMC2808911.
Downloaded from website

3. **GutPhage Database (Bacteriophages derived from human gut metagenomes):**

Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. Massive expansion of human gut bacteriophage diversity. Cell. 2021 Feb 18;184(4):1098-1109.e9. doi: 10.1016/j.cell.2021.01.029. PMID: 33606979; PMCID: PMC7895897.

Downloaded with: wget
http://ftp.ebi.ac.uk/pub/databases/metagenomics/genome_sets/gut_phage_database/

4. **Prokaryotic viral orthologous groups (pVOG) (a collection of prokaryotic virus protein HMMs):**

Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. 2017;45(D1):D491-D498. doi:10.1093/nar/gkw975

Downloaded from NCBI ftp site

5. **COMPASS (plasmids):**

Douarre PE, Mallet L, Radomski N, Felten A, Mistou MY. Analysis of COMPASS, a New Comprehensive Plasmid Database Revealed Prevalence of Multireplicon and Extensive Diversity of IncF Plasmids. Front Microbiol. 2020;11:483. Published 2020 Mar 24. doi:10.3389/fmicb.2020.00483

Downloaded with:

```
git clone https://github.com/itsmeludo/COMPASS.git
./COMPASS/src/uncompress_COMPASS.sh
```

6. **NCBI Plasmid RefSeq (plasmids):**

O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res.
2016;44(D1):D733-D745. doi:10.1093/nar/gkv1189

Downloaded with: wget https://ftp.ncbi.nlm.nih.gov/refseq/release/plasmid/plasmid.nonredundant_protein..protein.faa.gz*

7. **Immedbs (integrative elements)**

Jiang,X., Hall,A.B., Xavier,R.J. and Alm,E.J. (2019) Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. PLoS One, 14, e0223680.

Downloaded with: wget http://immedb.gutfun.org/data/Data1_MGE_sequences.fasta

8. **ISfinder (insertion sequences):**

Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D32-6. doi: 10.1093/nar/gkj014. PMID: 16381877; PMCID: PMC1347377.

Downloaded with: wget https://github.com/thanhleviet/ISfinder-sequences/blob/master/IS.faa