

 Open access • Posted Content • DOI:10.1101/2021.08.27.457951

mobileOG-db: a manually curated database of protein families mediating the life cycle of bacterial mobile genetic elements — [Source link](#)

Connor L. Brown, James Mullet, Fadi Hindi, James E. Stoll ...+6 more authors

Institutions: Virginia Tech

Published on: 27 Aug 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Related papers:

- [ACLAME: A CLAssification of Mobile genetic Elements, update 2010](#)
- [MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes.](#)
- [Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data](#)
- [Automatic and manual functional annotation in a distributed web service environment](#)
- [Gene-oriented ortholog database: a functional comparison platform for orthologous loci.](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/mobileog-db-a-manually-curated-database-of-protein-families-3fo6m0wo32>

1 mobileOG-db: a manually curated database of protein families 2 mediating the life cycle of bacterial mobile genetic elements

3 Connor L. Brown¹, James Mullet², Fadi Hindi², James E. Stoll³, Suraj Gupta¹, Minyoung Choi², Ishi
4 Keenum², Peter Vikesland², Amy Pruden^{*2}, Liqing Zhang^{*4}

5
6 ¹Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA, 24061, USA

7 ²Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA, 24061, USA

8 ³Fralin Life Science Institute, Blacksburg, VA, 24061, USA

9 ⁴Computer Science, Virginia Tech, Blacksburg, VA, 24061, USA

10
11 * To whom correspondence should be addressed. Tel: 1-540/231-6635 Fax: 540/231-7532
12 Fax: 540/231-7532; Email: apruden@vt.edu, lqzhang@vt.edu

13 14 **ABSTRACT**

15 Currently available databases of bacterial mobile genetic elements (MGEs) contain both
16 “core” and accessory MGE functional modules, the latter of which are often only transiently
17 associated with the element. The presence of these accessory genes, which are often close
18 homologs to primarily immobile genes, limits the usability of these databases for MGE annotation. To
19 overcome this limitation, we analysed 10,776,212 protein sequences derived from seven MGE
20 databases to compile a comprehensive database of 6,140 manually curated protein families that are
21 linked to the “life cycle” (integration, excision, replication/recombination/repair, transfer, and
22 stability/defense) of all major classes of bacterial MGEs. We overlay experimental information where
23 available to create a tiered annotation scheme of high-quality annotations and annotations inferred
24 exclusively through bioinformatic evidence. We additionally provide an MGE-class label for each entry
25 (e.g., plasmid, integrative element) derived from the source database, and assign a list of keywords to
26 each entry to delineate different MGE functional modules and to facilitate annotation. The resulting
27 database, mobileOG-db (for mobile orthologous groups), provides a simple and readily interpretable
28 foundation for an array of MGE- centred analyses. mobileOG-db can be accessed at
29 mobileogdb.flsi.cloud.vt.edu/, where users can browse and design, refine, and analyse custom
30 subsets of the dynamic mobilome.

31 **INTRODUCTION**

32 Prokaryotic genomes undergo frequent interactions with mobile genetic elements (MGEs)
33 including plasmids, bacteriophages, insertion sequences, and other integrative elements (IGEs).
34 These interactions can confer beneficial or detrimental properties to the organism (1), and in some
35 cases appear to have little impact on the organism at all (2). Of particular importance to public health,
36 bacterial MGEs can function as key vehicles for the proliferation of antimicrobial resistance (AMR),
37 which is now pandemic in many clinically-significant bacteria (3). Emerging efforts aimed at
38 developing frameworks for predicting the emergence of novel antibiotic resistance genes (ARGs),
39 whose products confer AMR in bacteria, have identified associations with MGEs as a key indicator of
40 a potential novel ARG (4).

41 While there are many tools and databases available for annotating mobile genetic elements
42 (7–10), there is at present no centralized resource for mobile genetic element “hallmark” genes which
43 can serve as the basis for annotating diverse classes of MGEs, such as is aggregated by pVOG (11)
44 for phages. However, even pVOG includes many primarily cellular genes (11, 12), which they identify
45 based on the frequency of occurrence in phages relative to the occurrence in cellular genome
46 sequences. Similarly, public databases of full length mobile genetic elements, such as ACLAME (6) or
47 ICEBerg (13), comprise both core and accessory MGE genes. While such databases are highly
48 informative, the presence of these cargo genes leads to frequent occurrences of false-positive hits
49 that confound and complicate the annotation of MGEs (14). In sum, the presence of these cargo
50 genes creates the need for extensive expertise and research to detect, analyse, and annotate diverse
51 types of mobile genetic elements in biological data. This is a key obstacle particularly for antibiotic
52 resistance research, where knowledge of the carriage of ARGs on MGEs is highly valuable towards
53 identifying mobile ARGs. (15).

54 To facilitate MGE annotation, we propose the mobile orthologous groups database and
55 webserver (mobileOG-db), an interactive resource compiling knowledge encompassing a
56 comprehensive variety of proteins mediating the essential functions of all major classes of bacterial
57 MGEs. Here we define the essential functions or “life cycle” of MGEs as (1) integration and excision
58 (IE) from one genetic locus to another; (2) replication, recombination, or repair (RRR) of element
59 nucleic acid; (3) inter-organism transfer (T); (4) element stability, transfer or defense (STD); and (5)
60 phage (P) related biological processes (e.g., genome packaging, or lysis and lysogeny). These

61 functions are essential to the persistence of MGEs as independent elements and are orchestrated by
62 an extremely diverse assortment of proteins (16) that we deem suitable as candidate “hallmarks”
63 because of the key roles that they play. Thus, the precise detection of these protein coding genes can
64 serve as the basis for the discovery, classification, and characterization of diverse MGEs in a simple
65 and intuitive way.

66 MATERIAL AND METHODS

67 Aggregation of a draft pan-mobilome and gene name assignment

68 A pan-mobilome, i.e., an extensive collection of sequences comprising diverse MGEs, was
69 created by merging sequences produced from seven publicly-available MGE-databases into a single
70 database of protein sequences: ICEBerg 2.0 (13), COMPASS (17), NCBI Plasmid RefSeq Gut Phage
71 Database (18), ISfinder (19), ACLAME (6) and immedb (20). The genomes comprising the basis for
72 the pVOG database (11) and COMPASS, a collection of exclusively nucleotide sequences, were
73 processed with prodigal (v2.6.3) (21) to generate open reading frames using the -p meta setting. The
74 final aggregated dataset included 10,776,849 sequences, or 2,649,813 sequences dereplicated at
75 97% identity and 80% query coverage (Fig. 1). The 10,776,749 proteins were searched against
76 UniProt (downloaded in the fall of 2020) using diamond blastp, with minimum identity 90% and
77 minimum query coverage of 80% cut-offs. This yielded 8,460,321 matches to UniProt. The 8,460,321
78 proteins were then used to parse a merged Bacterial, Archaeal, and viral UniProt knowledge base
79 (.dat file) with a custom script (available on the project GitHub page, [github.com/clb21565/mobileOG-](https://github.com/clb21565/mobileOG-db/scripts)
80 [db/scripts](https://github.com/clb21565/mobileOG-db/scripts)) to extract gene names yielding 110,234 gene names matched to UniProt entries; 20,979 of
81 the 110,234 gene names were unique.

82 Manual curation and annotation of the mobileOG-db

83 The 20,979 unique gene names were augmented to prepare queries for searching against the MGE
84 abstract database (Supplementary Methods, Table S1). The MGE abstract database was searched
85 using the unique queries and a resulting 8,372 gene name-abstract pairs were manually inspected by
86 at least two researchers. Sequences were manually curated and provided a functional annotation by
87 comparing the abstract text to the putative function reported within each UniProt or NCBI entry (Fig 1).
88 Because the same gene name can be attributed to multiple UniProt entries, sequence entities were
89 annotated on the basis of their 40% identity 50% query/subject coverage (mmseqs easy-clust -c 0)
90 cluster. If the cluster representative had a putative function inconsistent with the attributing abstract(s)
91 (Supplementary Methods), the sequence was reannotated with a review of literature recovered by
92 searching for the gene name and putative function in PubMed. To improve coverage, keyword
93 matches in the fasta headers with a table of MGE protein keywords (Table S2) was used as evidence
94 for inclusion in mobileOG-db. The evidence used to determine inclusion in mobileOG-db (manual
95 curation, homology, or keyword searches) is recorded in mobileOG-db. Examples of our rationale are
96 provided in Supplementary Methods. Last, sequences with matches to SwissProt entries were
97 considered a special case and were manually curated regardless of whether they were returned
98 during the abstract analysis. The gene names, queries, and the abstract database, are available at
99 the FigShare project (<https://doi.org/10.6084/m9.figshare.15170736>).

100 RESULTS

101 Catalogue of the mobile orthologous groups database

102 mobileOG-db consists of five major functional categories (P, RRR, STD, T, and IE) and
103 numerous minor categories, providing intuitive interpretation of search and filter terms. Starting with a
104 pan-mobilome of 10,776,213 proteins derived from ICEBerg, ACLAME, NCBI Plasmid RefSeq,
105 COMPASS, immedb, and ISfinder, we identified proteins performing the defining functions of phages,
106 IGEs, plasmids, and insertion sequences. Owing to the extensive curation effort, a key advance
107 achieved in the present database is its delineation of major and minor mobileOG categories that
108 compose complex elements (Fig. 2A). For example, the *Shigella flexneri* plasmid R100 is displayed
109 with different functional modules coloured by our mobileOG categories (Fig. 2A). There is a prominent
110 RRR module (including *repA*); T: conjugation module (including *finO*); and two transposons (Tn21 and
111 Tn10) dense with IE module protein-coding genes. Altogether, this first release of mobileOG-db
112 (beatrix) comprises 823,797 dereplicated proteins including over 29,000 derived directly from
113 manually curated entries; 6,140 protein clusters or families (defined as greater than 40% identical
114 over 50% of the subject and query length, see methods); 2,444 unique manual annotations, and
115 1,393 references (Fig. 2B).

116 Usage recommendations and examples

117 For detecting and classifying elements from long genomic segments (i.e., long reads or
118 assembled short reads), it is recommended that an accurate annotation consists of multiple co-
119 localized hits in close proximity, similar to the pattern-based co-localization approach leveraged by
120 ICEBerg (5) for IGE discovery. Likewise, it is noted that hits solely to RRR modules are not

121 necessarily indicative of an MGE; plasmids and phages frequently encode homologs of RRR
122 machinery (22) that are also present in exclusively cellular DNA. An additional caveat is that hits to
123 type four secretion systems may not be indicative of a MGE; paralogues of these proteins are also
124 virulence determinants in some organisms (23). A preliminary annotation pipeline has been
125 developed (Supplementary Methods, Table S1) to allow for automated element annotation
126 (Supplementary Methods). Usage of mobileOG-db enabled successful classification of up to 98.2%
127 and 99.7% of the plasmids and phages, respectively, comprising a test dataset
128 (<https://doi.org/10.6084/m9.figshare.15170736>, Table S3) of genomes in the COMPASS or the
129 GutPhage databases. Other uses might include the creation of quantitative metrics for horizontal gene
130 transfer hypothesis testing (Fig S3).

131 The mobileOG-db web portal provides a user-friendly interface for scientists across relevant
132 fields intersecting the Life Sciences to browsing and customizing datasets of MGE proteins (Fig. 2C).
133 Usage of the website allows users to select different major and minor mobileOG categories to hone
134 their experimental design or intended usage. Further, the ability to filter and select from different
135 element-types allow for the user to identify genes that occur across different element types. For
136 instance, users could select experimentally validated insertion sequence proteins that also occur on
137 plasmids, a key route for the horizontal transmission of ARGs (24).

138 **DISCUSSION**

139 The creation of mobileOG-db was motivated by a lack of an up-to-date and comprehensive
140 resource for markers of diverse classes of MGEs. Here, using a layered annotation scheme, we
141 analysed 10,776,212 MGE-encoded proteins to differentiate sequences that are anticipated to be
142 informative for annotation from those that are not defensible for this purpose from a biological
143 standpoint. Importantly, we recognize that the annotation framework implemented here cannot
144 produce a highly granular description of MGE function. However, providing such a resource for every
145 major class of bacterial MGEs is far beyond the scope of the present work and, in addition to
146 uncertainty of protein function, element-specific resources are available that form the basis for
147 mobileOG-db. Instead, mobileOG-db provides a “Swiss Army knife” that can serve as the foundation
148 for an array of analyses, which can be designed, customized, and refined using the web service.

149 Looking towards the future, the delineation of MGE functional modules and conserved protein
150 families could potentially support probabilistic methods for clustering, annotating, and classifying
151 MGEs. Such frameworks could also provide a basis for other analyses leveraging compositional and
152 structural features of the elements to quantitatively estimate potential host-linkages, cargo, and the
153 potential for transmission to clinically relevant pathogens. These analyses are being developed for
154 inclusion in a future release of mobileOG-db and show promise for harnessing large scale genomic
155 data for predictive public health insights.

156 **DATA AVAILABILITY**

157 mobileOG-db is available at mobileogdb.flsi.cloud.vt.edu/, where users can browse, filter, search, and
158 download customized datasets and references. Scripts used in the text mining analysis and two
159 example pipelines using R or Python are available at [https://github.com/clb21565/mobileOG-](https://github.com/clb21565/mobileOG-db/scripts)
160 [db/scripts](https://github.com/clb21565/mobileOG-db/scripts).

161 **SUPPLEMENTARY DATA**

162 Supplementary Data are available at NAR online and the manuscript FigShare repository
163 <https://doi.org/10.6084/m9.figshare.15170736>.

164 **ACKNOWLEDGEMENT**

165 We would like to directly express our appreciation for the work and expertise that went into designing
166 the databases making up mobileOG-db. The authors acknowledge the Advanced Research
167 Computing at Virginia Tech for providing computational resources.

168 **FUNDING**

169 This study was supported by NSF PIRE (PI Vikesland) Award 1545756, NSF CI4WARS (PI
170 Zhang) Award 2004751, USDA National Institute of Food and Agriculture competitive Grant 2017-
171 68003-26498, Water Research Foundation Project 4961, the Genetics, Bioinformatics, and
172 Computational Biology Interdisciplinary Graduate Education Program (IGEP), the Virginia Tech
173 Sustainable NanoTechnology IGEP, NanoEarth, Fralin Life Sciences Institute, the Virginia Tech Open
174 Access Support Fund, and the Virginia Tech ICTAS Center for Science and Engineering of the
175 Exposome.

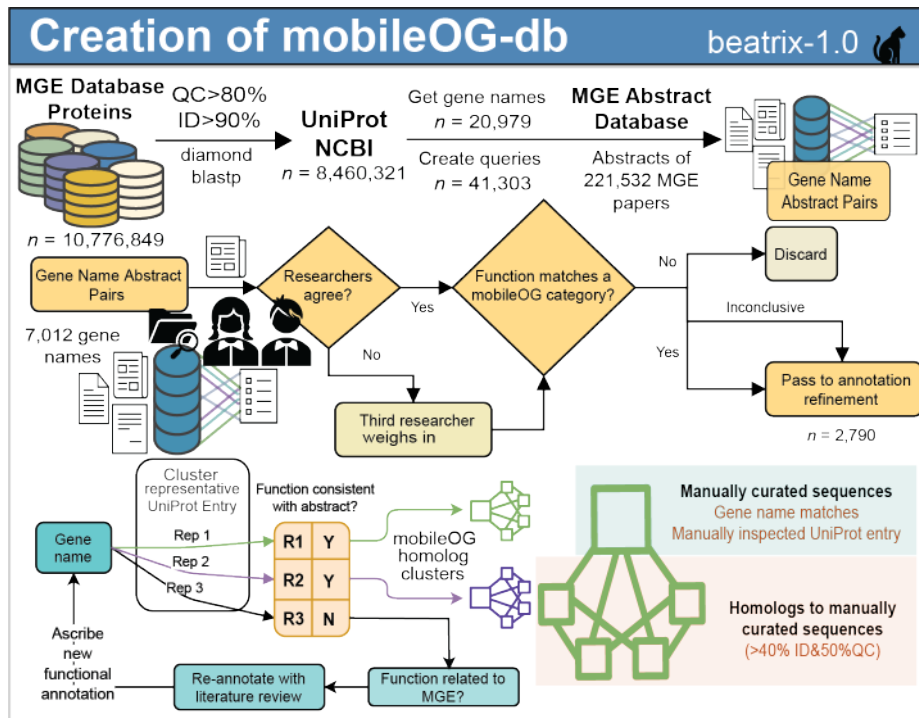
176 **CONFLICT OF INTEREST**

177 The authors report no conflicts of interest.

178 **REFERENCES**

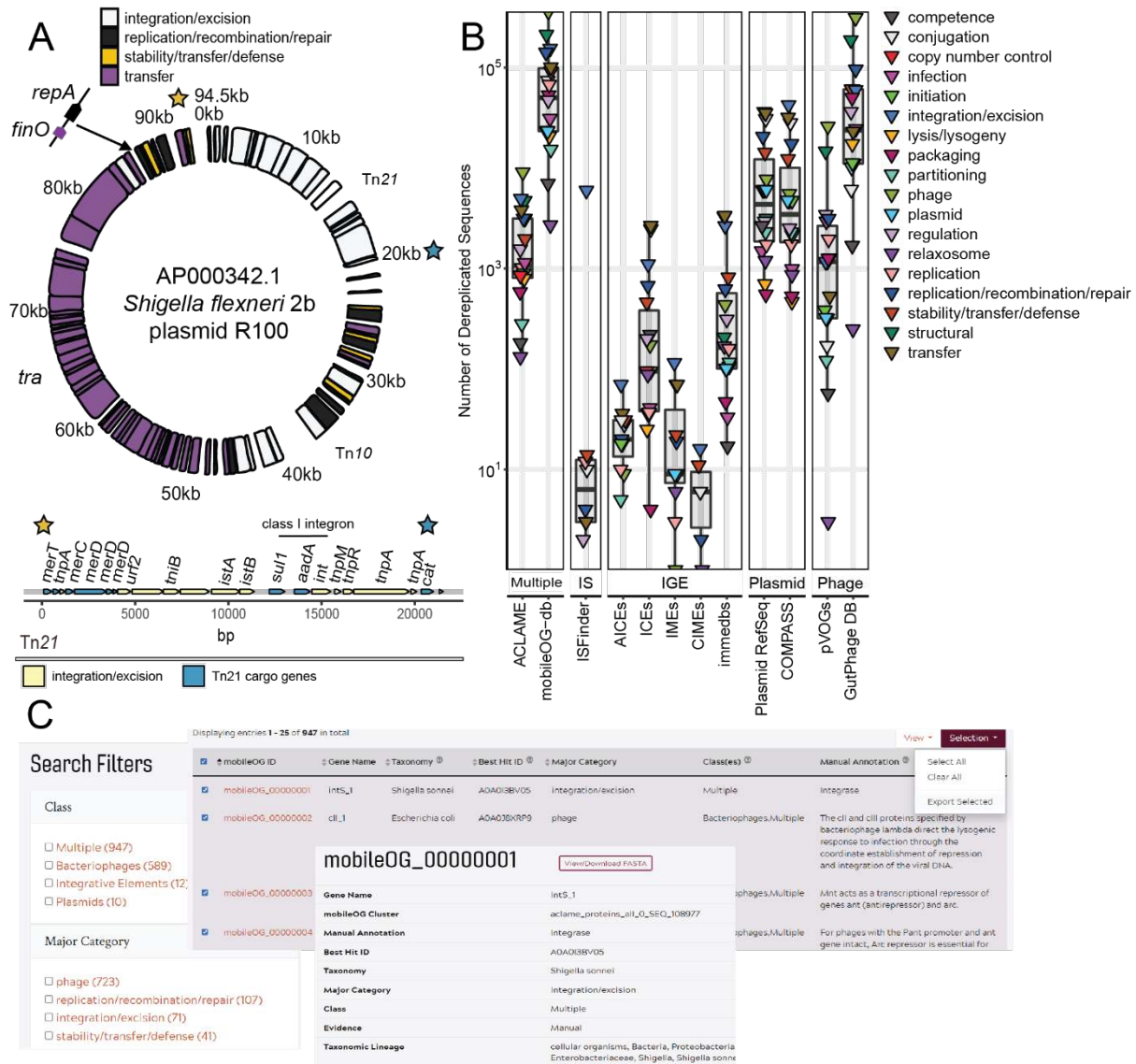
179 1. Rankin, D.J., Rocha, E.P.C. and Brown, S.P. (2011) What traits are carried on mobile genetic
180 elements, and why. *Heredity (Edinb)*, **106**, 1–10.

- 181 2. Berg, O.G. and Kurland, C.G. (2002) Evolution of microbial genomes: Sequence acquisition and
182 loss. *Mol. Biol. Evol.*, **19**, 2265–2276.
- 183 3. Partridge, S.R., Kwong, S.M., Firth, N. and Jensen, S.O. (2018) Mobile genetic elements associated
184 with antimicrobial resistance. *Clin. Microbiol. Rev.*, **31**.
- 185 4. Ellabaan, M.M.H., Munck, C., Porse, A., Imamovic, L. and Sommer, M.O.A. (2021) Forecasting the
186 dissemination of antibiotic resistance genes across bacterial genomes. *Nat. Commun.*, **12**, 1–10.
- 187 5. Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z. and Ou, H.Y. (2019) ICEberg 2.0: An
188 updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res.*, **47**,
189 D660–D665.
- 190 6. Leplae, R., Lima-Mendez, G. and Toussaint, A. (2009) ACLAME: A CLAssification of mobile genetic
191 elements, update 2010. *Nucleic Acids Res.*, **38**.
- 192 7. Roux, S., Enault, F., Hurwitz, B.L. and Sullivan, M.B. (2015) VirSorter: Mining viral signal from
193 microbial genomic data. *PeerJ*, **2015**, e985.
- 194 8. Krawczyk, P.S., Lipinski, L. and Dziembowski, A. (2018) PlasFlow: predicting plasmid sequences in
195 metagenomic data using genome signatures. *Nucleic Acids Res.*, **46**, e35.
- 196 9. Orlek, A., Stoesser, N., Anjum, M.F., Doumith, M., Ellington, M.J., Peto, T., Crook, D., Woodford, N.,
197 Sarah Walker, A., Phan, H., *et al.* (2017) Plasmid classification in an era of whole-genome
198 sequencing: Application in studies of antibiotic resistance epidemiology. *Front. Microbiol.*, **8**, 1–
199 10.
- 200 10. Carr, V.R., Shkoporov, A., Hill, C., Mullany, P. and Moyes, D.L. (2021) Probing the Mobilome:
201 Discoveries in the Dynamic Microbiome. *Trends Microbiol.*, **29**, 158–170.
- 202 11. Grazziotin, A.L., Koonin, E. V. and Kristensen, D.M. (2017) Prokaryotic Virus Orthologous Groups
203 (pVOGs): A resource for comparative genomics and protein family annotation. *Nucleic Acids*
204 *Res.*, **45**, D491–D498.
- 205 12. Pfeifer, E., Moura De Sousa, J.A., Touchon, M. and Rocha, E.P.C. (2021) Bacteria have numerous
206 distinctive groups of phage-plasmids with conserved phage and variable plasmid gene
207 repertoires. *Nucleic Acids Res.*, **49**, 2655–2673.
- 208 13. Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z. and Ou, H.Y. (2019) ICEberg 2.0: An
209 updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res.*, **47**,
210 D660–D665.
- 211 14. Slizovskiy, I.B., Mukherjee, K., Dean, C.J., Boucher, C. and Noyes, N.R. (2020) Mobilization of
212 Antibiotic Resistance: Are Current Approaches for Colocalizing Resistomes and Mobilomes
213 Useful? *Front. Microbiol.*, **11**, 1376.
- 214 15. Partridge, S.R., Kwong, S.M., Firth, N. and Jensen, S.O. (2018) Mobile genetic elements associated
215 with antimicrobial resistance. *Clin. Microbiol. Rev.*, **31**.
- 216 16. Craig, N.L. (2015) A Moveable feast: An introduction to mobile DNA. In *Mobile DNA III*. Wiley, pp.
217 3–39.
- 218 17. Douarre, P.E., Mallet, L., Radomski, N., Felten, A. and Mistou, M.Y. (2020) Analysis of COMPASS, a
219 New Comprehensive Plasmid Database Revealed Prevalence of Multireplicon and Extensive
220 Diversity of IncF Plasmids. *Front. Microbiol.*, **11**, 483.
- 221 18. Camarillo-Guerrero, L.F., Almeida, A., Rangel-Pineros, G., Finn, R.D. and Lawley, T.D. (2021)
222 Massive expansion of human gut bacteriophage diversity. *Cell*, **184**, 1098–1109.e9.
- 223 19. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. and Chandler, M. (2006) ISfinder: the reference
224 centre for bacterial insertion sequences. *Nucleic Acids Res.*, **34**.
- 225 20. Jiang, X., Hall, A.B., Xavier, R.J. and Alm, E.J. (2019) Comprehensive analysis of chromosomal
226 mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools.
227 *PLoS One*, **14**, e0223680.
- 228 21. Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J. (2010) Prodigal:
229 Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**,
230 119.
- 231 22. Chen, S.H., Byrne, R.T., Wood, E.A. and Cox, M.M. (2015) Escherichia coli radD(yejH) gene: A
232 novel function involved in radiation resistance and double-strand break repair. *Mol. Microbiol.*,
233 **95**, 754–768.
- 234 23. Costa, T.R.D., Harb, L., Khara, P., Zeng, L., Hu, B. and Christie, P.J. (2021) Type IV secretion
235 systems: Advances in structure, function, and activation. *Mol. Microbiol.*, **115**, 436–452.
- 236 24. Che, Y., Yang, Y., Xu, X., Brinda, K., Polz, M.F., Hanage, W.P. and Zhang, T. (2021) Conjugative
237 plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial
238 resistance genes. *Proc. Natl. Acad. Sci. U. S. A.*, **118**.
- 239



240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277

Figure 1. Construction of the mobile orthologous groups database (mobileOG-db). Publicly- available MGE database were downloaded, and their contents mapped to the UniProt TrEMBL/SwissProt knowledge base. Gene names were searched against a filtered database of MGE-related abstracts. 7,012 gene name pairs were then manually inspected by at least two researchers to determine whether the identified gene encoded a protein with a role in one of the target mobileOG categories (replication/recombination/repair, stability/transfer/defence, integration/excision, phage, transfer). A total of 2,790 manually- curated gene names were passed to annotation refinement, where names were paired with UniProt/NCBI entries and associated metadata. To reduce the number of manual curation events needed, we selected one representative sequence for each cluster (40% identity over 50% of reference length using mmseqs2) with a given gene name and compared their database-derived putative function with literature descriptions of the proteins recovered from the abstract analysis. If the UniProt/NCBI entry did not support a link between the gene name and function, the protein was annotated with a literature review.



278
 279
 280
 281
 282
 283
 284
 285
 286
 287
 288
 289
 290
 291

Figure 2. Overview and examples of the content within mobileOG-db. (A) *Shigella flexneri* plasmid R100, among the first conjugative multidrug resistance plasmids to be identified carries Tn21, a resistance geneharbouring transposon (initial and terminal repeats indicated by gold and blue stars, respectively). Usage of the mobileOG-db positively identifies AP000342.1 as a conjugative plasmid enriched with integration/excision module proteins highlighting the carriage of integrative elements IS1353, IS1326, IS1b, and a class 1 integron. In the bottom panel, Tn21 is examined in greater detail; Tn21 cargo genes (blue) are not included in mobileOG-db. (B) mobileOG-db contains a substantially improved diversity of sequences compared to presently available databases, covering a wide-array of MGE functional modules. Abbreviations: IS: Insertion sequences recovered from ISfinder.; IGE: integrative elements recovered from ICEberg. (C) the mobileOG-db webserver provides a user-friendly platform for interactive annotation database customization for scientists across the life sciences. Researchers can search, browse, filter, and download customized datasets tailored to their research questions.