1 # AGNOSTOS-DB: a resource to unlock the

2 # uncharted regions of the coding sequence space

3 Chiara Vanni[1,2], Matthew S. Schechter[1,3], Tom O. Delmont[4], A. Murat Eren[3,5], Martin

4 Steinegger[6,7], Frank Oliver Glöckner[8,9,2], Antonio Fernandez-Guerra[1,10] *

5 ## Affiliations

6 1 Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine

7 Microbiology, Celsiusstraße 1, 28359, Bremen, Germany

8 2 Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany

9 3 Department of Medicine, University of Chicago, Chicago, IL 60637, USA

10 4 Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry,

11 Université Paris-Saclay, 91057 Evry, France

12 5 Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA 02543, USA

13 6 School of Biological Sciences, Seoul National University, Seoul, 08826, South Korea

14 7 Institute of Molecular Biology and Genetics, Seoul National University, Seoul, 08826, South

15 Korea

16 8 University of Bremen, MARUM, Leobener Str. 8, 28359 Bremen, Germany

17 Life Sciences and Chemistry, Campus Ring 1, 28759 Bremen, Germany

18 9 Computing and Data Center, Helmholtz Center for Polar and Marine Research, Am

19 Handelshafen 12, 27570 Bremerhaven, Germany

20   10 Lundbeck GeoGenetics Centre, The Globe Institute, University of Copenhagen, 1350

21   Copenhagen, Denmark

22

Abstract

24   Genomes and metagenomes contain a considerable percentage of genes of unknown

25   function, which are often excluded from downstream analyses limiting our understanding of

26   the studied biological systems. To address this challenge, we developed AGNOSTOS, a

27   combined database-computational workflow resource that unifies the known and unknown

28   coding sequence space of genomes and metagenomes. Here, we present AGNOSTOS-DB,

29   an extensive database of high-quality gene clusters enriched with functional, ecological and

30   phylogenetic information. Moreover, AGNOSTOS allows integrating new data into existing

31   AGNOSTOS-DBs, maximizing the information retrievable for the genes of unknown function.

32   As a proof of concept, we provide a seed database that integrates the predicted genes from

33   marine and human metagenomes, as well as from Bacteria, Archaea, Eukarya and giant

34   viruses environmental and cultivar genomes. The seed database comprises 6,572,081 gene

35   clusters connecting 342 million genes and represents a comprehensive and scalable resource

36   for the inclusion and exploration of the unknown fraction of genomes and metagenomes.

37

Background & Summary

39   The characterization of genes of unknown function can catalyze research in basic science,

40   medicine, and biotechnology[1]. Approximately 30-35% of the genes in genomes and

41   metagenomes lack functional characterization[2]. The constant influx of newly sequenced

42   microbial and viral genomes through cultivation, single-cell genomics and metagenomics is

2

43   increasing the volume of genes of unknown function. Recently we developed AGNOSTOS[2], a

44   conceptual framework and a computational workflow that unifies the known and the unknown

45   coding sequence space (CDS-space) of genomes and metagenomes to address the

46   limitations of the current approaches. Such limitations include the lack of a systematic

47   categorization of the genes with unknown function into biologically meaningful categories, the

48   inclusion of both genomic and metagenomic data and often an overestimation of the number

49   of unknowns due to the lack of remote-homology searches. AGNOSTOS partitions the CDS-

50   space using gene clusters, and unlike the previous GC centered methods[3–6], performs an

51   extensive validation yielding high-quality biologically and phylogenetically aware functional

52   units. Further, AGNOSTOS provides a thorough characterization of the unknown space,

53   including distant-homology classification methods. Moreover, GCs sharing remote homologies

54   are aggregated in communities (GCCs)[2]. Altogether AGNOSTOS allows exploring the CDS-

55   space at different levels: from the gene similarities within a GC to the remote homologies in

56   the GCCs.

57   The AGNOSTOS databases (AGNOSTOS-DBs) are embedded into the AGNOSTOS

58   analytical environment and allow different levels of analysis, as shown in Figure 1. The

59   AGNOSTOS-DB provides GC sequence profiles that can be used to quickly screen new

60   datasets (*profile-search* module). Alternatively, a novel DB can be created from any

61   metagenomic and genomic data via the *DB_creation* module. Finally, a key feature of

62   AGNOSTOS is its scalability, i.e., the possibility to update existing DBs with new sequences.

63
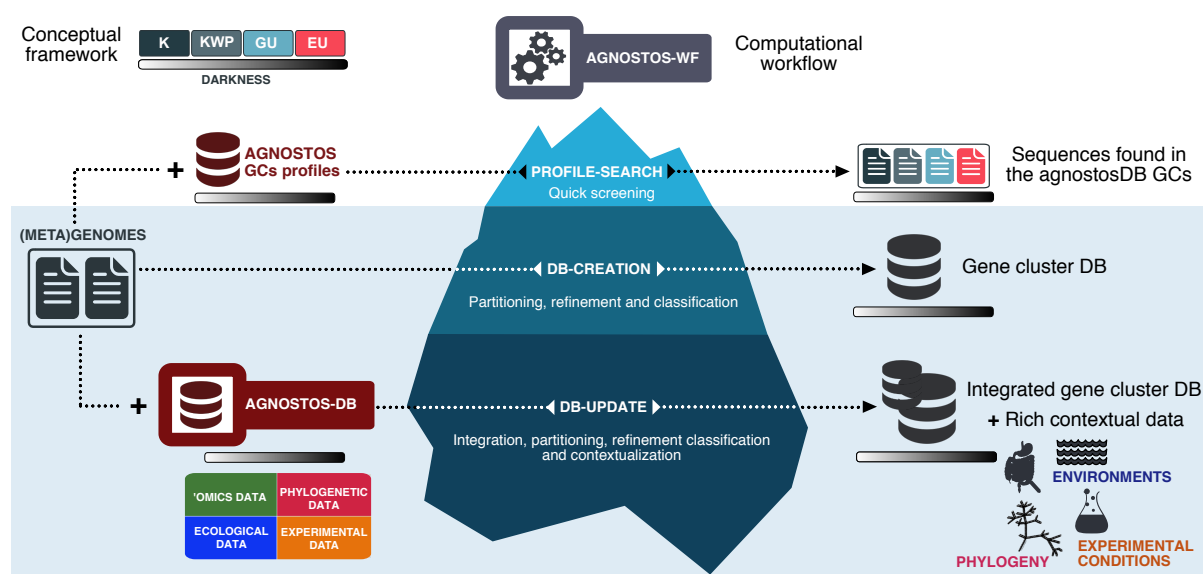
Figure 1. Schematic overview of the AGNOSTOS analytical environment (workflow and databases).

The AGNOSTOS seed-DB integrated the sequence data from 1,829 marine and human metagenomic assemblies and 28,941 bacterial and archaeal genomes, for a total of 5,287,759 GCs and 335,439,673 genes[2]. To show the flexibility and scalability provided by AGNOSTOS, here we expanded the seed-DB by integrating genomes affiliated to Eukarya and their infecting nucleocytoplasmic large DNA viruses (NCLDV) characterized mostly from metagenomic sequence data corresponding to the surface of the oceans (Gaia et al., personal communication). First, we integrated the predicted genes from 3,243 NCLDV environmental and cultivar genomes[7-9] (Gaia et al., personal communication), obtaining a database that contains 5,383,876 GCs and 336,513,365 genes (Table 1).

Table 1. Seed + NCLDV database. Number of genes, GCs, and GCCs per category.

	K	KWP	GU	EU	Total
Genes	230,972,742	32,974,245	68,898,794	3,667,584	336,513,365
GCs	1,680,863	799,647	2,678,313	225,053	5,383,876

4

| GCCs | 75,332 | 119,111 | 443,961 | 122,812 | 761,216 |

76

77 Subsequently, we integrated the predicted genes from 713 eukaryotic environmental

78 genomes[10], creating a comprehensive database of 6,572,081 GCs and 341,655,294 genes

79 (Table 2) that expand to all domains of life and viruses.

80 **Table 2.** AGNOSTOS seed + NCLDV + Eukaryotic plankton GC database. Number of genes,
81 GCs and GCCs by category.

	K	KWP	GU	EU	Total
Genes	233,533,089	33,654,589	69,853,347	4,614,269	341,655,294
GCs	2,051,910	1,007,932	2,927,264	584,975	6,572,081
GCCs	243,664	181,759	517,242	224,913	1,167,578

82

83 The AGNOSTOS GCs proved to be relatively stable after integrating new data. The GCs

84 updated with new genes showed a decrease in the intra-cluster average similarity[2] of 7% after

85 the first integration and 11% after the second. The overall intra-cluster average similarity

86 values decreased by 1% with the two integrations (Table 3).

87 **Table 3.** Changes in the GCs median similarity values after the integration of NCLDV and

88 plankton eukaryotic genes. (The values are filtered for clusters with more than 2 members.)

Median-similarity	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Seed-DB	0.328	0.678	0.817	0.802	0.945	1
+ NCLDV	0.3	0.677	0.816	0.801	0.945	1

+ EUK	0.3	0.667	0.807	0.795	0.942	1

89

90   In general, the integration of new data led to the reduction of the original set of singletons

91   (Table 4), suggesting that many of these sequences are not just artifacts or the results of

92   sequencing and assembly errors. However, 241,669 NCLDV and 4,264,489 eukaryotic genes

93   were added to the singletons set (Table 4), increasing the total number of singletons by 17%,

94   manifesting the novelty added when we integrate poorly characterized and phylogenetically

95   diverse groups such as the NCLDVs and the plankton eukaryotes.

96   **Table 4.** Changes in the number of singletons as a consequence of the integration of NCLDVs

97   and plankton eukaryotic sequences.

| Seed-DB	Seed-DB+NCLDV	Seed-DB+NCLDV+EUK
24,977,524	25,181,853	29,328,400

98

Seed-DB		
24,977,524	Seed-DB+NCLDV	
- 37,340 + 241,669 =	25,181,853	Seed-DB+NCLDV+EUK
	- 117,942 + 4,264,489 =	29,328,400

99

100   Moreover, after the two integrations, we observed an increase of 23% in unknown GCs (187%

101   for the EU GCs alone) compared to the seed-DB. These novel GCs found in the context of

102   NCLDV and eukaryotic genomes offer a starting point for targeted analyses of the unknown

103   fraction. Ultimately, as shown in Figure 2, we found most GCs unique to the prokaryotic or

6

104   eukaryotic plankton fraction where more than half of the GCs in these two sets belong to the

105   unknown fraction (Figure 2). We also identified 5,220 GCs shared by all datasets and 2,854

106   GCs shared between the NCLDVs and the plankton eukaryotes, in both cases mainly

107   members of the known fraction (Figure 2). This last set constitutes a valuable resource to

108   better understand the relationship between the NCLDVs and their eukaryotic hosts[11].

109
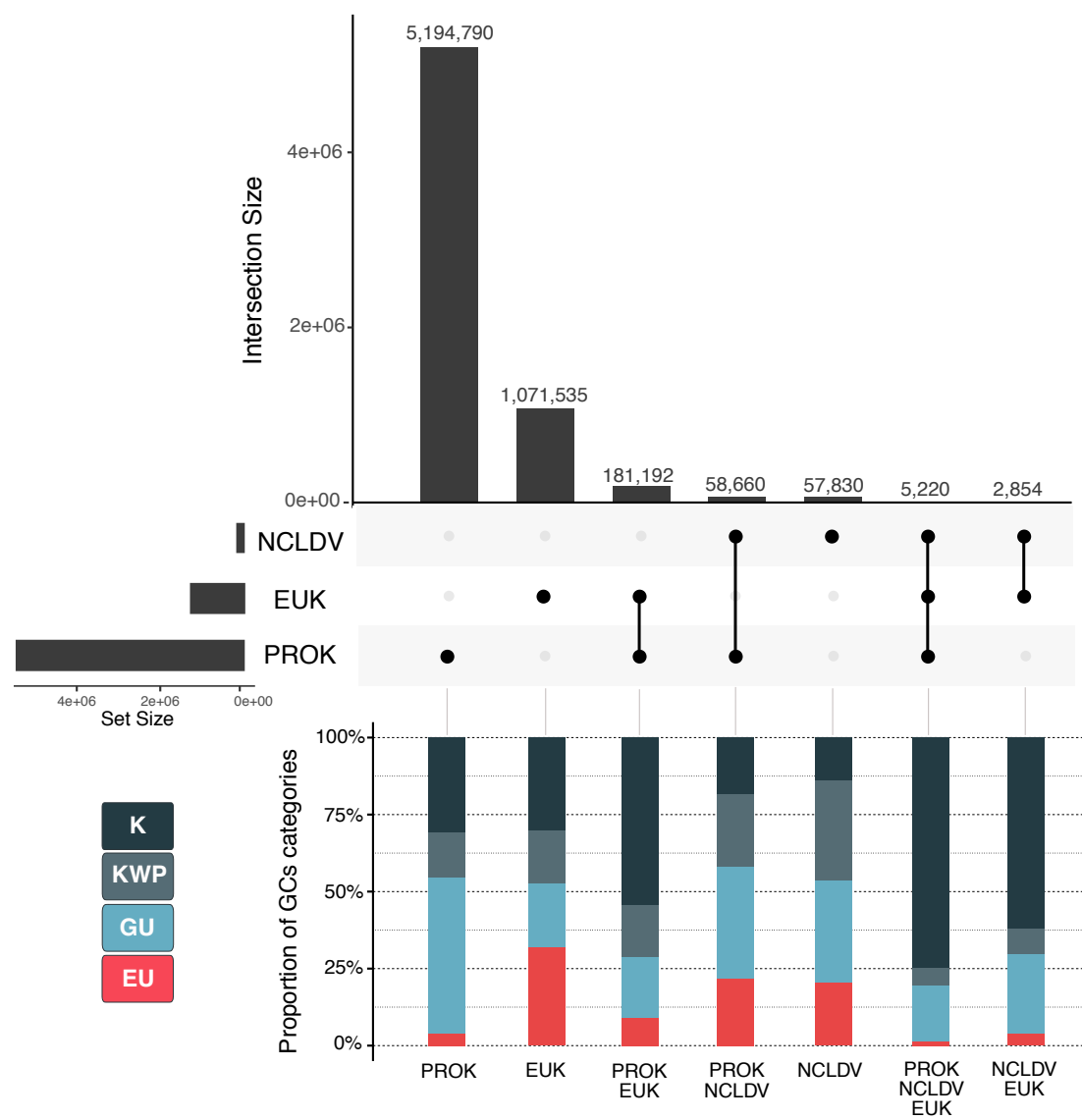
110   **Figure 2.** GC distribution in the different datasets and proportion of categories per set.

111   NCLDV=giant virus GCs, EUK=plankton eukaryotic GCs and PROK=seed-DB archaea and

112   bacterial GCs.

7

113

114 The seed-DB and the databases resulting from the integration of NCLDV and plankton

115 eukaryotes genomes are publicly available on Figshare (Data Citation 1-3). The workflow code

116 is available on GitHub (https://github.com/functional-dark-side/agnostos-wf).

117

Methods

118

Computational workflow

119

120 We implemented the AGNOSTOS using the Snakemake workflow management system[12] to

121 easily process large datasets in a reproducible manner. The workflow is publicly available in

122 a GitHub repository at https://github.com/functional-dark-side/agnostos-wf (version 1.0, used

123 in this paper, is archived at https://doi.org/10.5281/zenodo.4557847). The workflow provides

124 three different modules (Fig. 1): The *DB-creation* module, which creates a de-novo GC

125 dataset, taking as input metagenomic contigs or predicted genes in *fastA* format. The output

126 is a dataset of characterized GCs, partitioned in four categories (Fig. 1). The output of the *DB-*

127 *creation* module is already formatted to be updated with new data via the *DB-update* module.

128 The *DB-update* module allows the continuous integration of new sequence data into an

129 existing GC database. The new sequences are integrated into existing clusters if similar

130 enough or clustered in new GCs if not. The final output consists of the updated GC database

131 and related files containing contextual data about the GCs.

132 Moreover, the workflow provides a *profile-search* module that enables the screening of

133 existing GC databases using sensitive profile-search methods. The sequence-profile search

134 is performed using the *search* program of the MMseqs2[13] software (parameters: *-e 1e-20 --*

135 *cov-mode 2 -c 0.6*). This module takes in input the GC MMseqs2 profiles, a tabular file with

136 the information about the GC functional categories and the sequences to search against the

137 profiles, in *fastA* format. The search results are then filtered to keep hits within the 90% of the

138 best *log(e-value)* and retrieve the queries consensus GC annotation via a quorum majority

139 voting approach.

140 The required software tools and programs are installed via Conda or via the provided

141 installation script. When the existing database is the seed-DB, or an updated GC database

142 originated from the seed-DB, the *DB-update* final rule links the new integrated GCs with the

143 ecological, phylogenetic, and experimental metadata obtained from the Vanni et al. analyses[2]

144 (Fig. 1).

145 To build the GC databases presented here, we ran the workflow in the de.NBI Cloud

146 (https://www.denbi.de/cloud)[14], using a cluster setup with 10 nodes of 28 cores and 252G of

147 memory each. The cluster was built using BiBiGrid (https://github.com/BiBiServ/bibigrid) and

148 it is using SLURM[15] for job scheduling and cluster management.

149 **Data collection**

150 To build the seed-DB, we combined a set of 1,829 metagenomes from five major metagenomic

151 surveys of the ocean and human microbiome with 28,941 archaea and bacterial genomes[2].

152 The data sources per project are specified in Table 5-A, and the number of genes predicted

153 from each project in Table 5-B. The metagenomic contextual data is found in one SQLite

154 (version 3.25.0) (https://www.sqlite.org/index.html) database "contextual_data.db", which is

155 available in the "agnostosDB_dbf02445-20200519_environmental" folder (download:

156 https://ndownloader.figshare.com/files/23066879) within the seed-DB dataset deposited in

157 Figshare (https://doi.org/10.6084/m9.figshare.12459056, Data Citation 1)

158

159 **Table 5.** AGNOSTOS seed-DB source datasets.

160   (A) Metagenomic and genomic project data sources.

Dataset	Reference	Raw data	Contextual data
TARA	Sunagawa et al.[16]	ENA	PANGAEA
Malaspina	Duarte et al.[17]	Duarte et al.[17]	Duarte et al.[17]
OSD	Kopf et al.[18]	ENA	PANGAEA
GOS	Rush et al.[19]	NCBI	iMICROBE
HMP	Lloyd-Price et al.[20]	HMP portal	HMP portal
GTDB	Parks et al.[21]	Annotree [22]	Annotree [22]

161   (B) Number of genomes, metagenomes and predicted genes per project.

Dataset	(Meta)genomes	Predicted genes
TARA	242	111,903,261
Malaspina	116	20,574,033
OSD	145	7,015,383
GOS	80	20,068,580
HMP	1,246	162,687,295
GTDB-r86	28,941	93,723,190

162

163   We expanded the seed-DB by integrating 3,243 cultivar and environmental genomes from

164   nucleocytoplasmic large DNA viruses (NCLDVs). This collection combines the environmental

165   genomes generated by Moniruzzaman et al.[7], Schultz et al.[8], and by the TARA Oceans

10

166   Consortium (Gaia et al., personal communication, dataset unpublished) as well as 235

167   reference genomes collected from GenBank[9] (Table 6).

168   The NCLDV dataset was passed to AGNOSTOS in the form of predicted genes (amino acids).

169   **Table 6.** The NCLDV genomes dataset

Dataset	Genomes	Predicted genes
Moniruzzaman	501	631,572
Schultz	1,762	181,068
TARA	745	193,855
GenBank	235	99,599

170

171   Additionally, we added to the seed+NCLDV-DB, 10,207,450 eukaryotic plankton predicted

172   genes, obtained from the combination of 683 environmental genomes and 30 single amplified

173   genomes[10]. The plankton eukaryotic genes were predicted by Delmont et al. using a

174   combination of three approaches: a homology-based search against protein reference

175   databases (Uniref90 + METdb), the mapping of the metatranscriptomics assemblies to the

176   environmental and single-amplified genomes, and an ab-initio gene prediction using *gmove*

177   (http://www.genoscope.cns.fr/externe/gmove/) [10].

178   Both database integrations were performed using the *DB-update* module of the workflow.

179   **Update of a gene cluster dataset**

180   AGNOSTOS updates an existing GC database using the *clusterupdate* module of MMseqs2[23]

181   using the same clustering parameters described in Vanni et al.[2]. The output of the clustering

11

182　update is then parsed to retrieve three GC datasets: (1) The "original" GCs, a set containing

183　the original GCs without any similarity to the new sequences. (2) The "shared" set of GCs with

184　mixed, original and new, sequences. (3) The "new" GCs built from the new sequences without

185　any similarity to the existing GCs. The "new" GCs and the shared GCs that were originally

186　singletons, are processed through all the workflow steps until the GCC inference.

187　*Results of the integration and gene cluster contextualization.* At the end of the *DB-update*

188　module, the three GC sets, together with their additional information, are re-combined in a

189　single GC dataset. In case the original GC database originated from the seed-DB[2], the final

190　GC dataset can be decorated with the ecological and phylogenetic contextual data described

191　in Vanni et al.[2]. This information is included in the online version of the seed-DB (Data Citation

192　1), and can be used at the end of the *DB-update* module to further characterize the updated

193　GCs (Fig. 1).

194

Data Records

195

196　The　　　AGNOSTOS　　　seed-DB　　　is　　　available　　　on　　　Figshare　　　at

197　https://doi.org/10.6084/m9.figshare.12459056 (Data Citation 1). We base encoded the seed

198　version of the database as agnostosDB_dbf02445-20200519, which resulted from using the

199　*crc32* hash algorithm to encode the word "seed", followed by the release date (YYYYMMDD).

200　Future integrations of (meta)genomic data into the seed-DB will be named by adding the new

201　dataset name in the *crc32* encoding command followed by the new release date. The seed-

202　DB is a dataset of microbial GCs containing 40,423,549 GCs with 415,971,742 genes

203　predicted from 28,941 bacterial and archaeal genomes and 1,749 metagenomes.

204　The　　　　　　　　　　　　　　　　　　"agnostosDB_dbf02445-20200519_original-data"

205　(https://ndownloader.figshare.com/files/23066858)　contains　the　list　of　metagenomes,

206   genomes, contigs and predicted genes per metagenomic project and found in GTDB. The

207   original seed-DB GC MMseqs2 database files are available in "agnostosDB_dbf02445-

208   20200519_mmseqs_clustering"   (https://ndownloader.figshare.com/files/23066651).   The

209   distribution of the GCs and genes in the four categories can be found in Supplementary Figure

210   2 of Vanni et al.[2]. The list of refined GCs, their categories and their genes are available in the

211   TSV file: "cluster_ids_categ_genes.tsv.gz". The TSV file "HQ-clusters.tsv.gz" contains the IDs

212   of the set of high-quality GCs, containing mostly complete genes. The GC HH-suite database

213   files   are   contained   in   the   folder   "agnostosDB_dbf02445-20200519_hh-suite-db"

214   (https://ndownloader.figshare.com/files/23064476). The HH-suite database can be used to

215   screen the GCs in search for remote-homologies, using HHblits[24] (see Usage notes). The GC

216   MMseqs2   profiles   used   by   the   *profile-search*   module   are   stored   in   the   folder

217   "agnostosDB_dbf02445-20200519_mmseqs_profiles"

218   (https://ndownloader.figshare.com/files/23066963).

219   All other files needed by the *DB-update* and the *profile-search* modules are also available in

220   the database Figshare folder.

221   The GC database expanded with the NCLDV MAGs was encoded as agnostosDB_a42ac58a-

222   20200715 and it is available on Figshare at https://doi.org/10.6084/m9.figshare.13251743

223   (Data citation 2).

224   The database agnostosDB_4eab867d-20201104 combining the seed-DB GCs with the

225   NCLDV   and   the   eukaryotic   plankton   GCs   is   available   on   Figshare   at

226   https://doi.org/10.6084/m9.figshare.13264769 (Data citation 3).

227   The file organization of the two updated GCs datasets mirrors that of the seed-DB.

228

13

Technical Validation

Gene cluster database validations

The quality control steps in AGNOSTOS thoroughly validate the AGNOSTOS-DB GCs. These steps include the identification and removal of eventual spurious genes, the validation of the GCs internal homogeneity, both in terms of Pfam protein domain annotations and sequence composition, and a remote-homology refinement of the GCs classification in categories, which aimed to avoid an overestimation of the real number of uncharacterized genes[2,25].

Computational workflow validations

We followed the FAIR guiding principles[26] to create a reproducible and reusable workflow. We implemented AGNOSTOS using the Snakemake workflow management system[12] and is available as an open source software (https://doi.org/10.5281/zenodo.4557847). The majority of the dependencies can be installed within a Conda environment[27,28] integrated in the workflow. We provide an installation script to install the software not available in Conda. The workflow can be tested using the provided test datasets for the *DB-creation* and *DB-update* modules of AGNOSTOS. The test datasets are intended to validate the installation and help the user with an easy example of the workflow functioning and extensive outputs. For the testing dataset, we selected three TARA Ocean surface samples from the Indian Ocean (IO) (TARA_038, TARA_039 and TARA_041), all within a size fraction of 0.1-0.22 micro-m (giant viruses, archaea and bacteria enriched fraction) and we randomly chose 5K contigs from each sample. The test dataset is available for download from Figshare (https://doi.org/10.6084/m9.figshare.12630581, Data Citation 4).

14

Usage Notes

251

252   The workflow is accompanied by detailed usage instructions and description of the output files

253   available in the GitHub repository (https://github.com/functional-dark-side/agnostos-wf).

254   The GC profiles and the tables with the GC categories used by the *profile-search* module can

255   be downloaded from their respective AGNOSTOS-DB Figshare folders.

256   Any other MMseqs2 profile database can be used for the search. For example, to screen only

257   the seed-DB high-quality GCs, we can use their identifiers from the TSV file

258   "HQ_clusters.tsv.gz", to subset the full GC profile database with the MMseqs2 *createsubdb*

259   module (see the MMseqs2 user guide available at https://mmseqs.com/latest/userguide.pdf).

260   By default, the *DB-update* module of the workflow will download and use the seed-DB as the

261   existing GC database. Any other GC database built with the *DB-creation* module, or containing

262   the same output file structure, can be used by specifying the different database location in the

263   *DB-update* configuration file.

264   The output of the *DB-update* and *DB-creation* modules can be analyzed using anvi'o[29], by

265   integrating all the information gathered for the GCs as "functions" in the anvi'o contig database.

266   An introduction on how to integrate AGNOSTOS products in the analyses with anvi'o is

267   available in the form of a tutorial at https://merenlab.org/agnostos-tutorial. The tutorial

268   describes the integration of the infant gut dataset (IGD) from Sharon et al.[30] in AGNOSTOS

269   (workflow and databases) and how to use it in combination with anvi'o to explore genes of

270   unknown function. Within anvi'o, the GCs can be explored in the context of assembled contigs

271   allowing the identification of contigs enriched in uncharacterized genes or the exploration of

272   the genomic context of GCs of interest. Furthermore, the AGNOSTOS GCs can also be added

273   as "functions" to the external and internal genomes storages used by anvi'o to calculate a

15

274   pangenome. The pangenome can then be investigated to identify core and genome-specific

275   GCs.

Code Availability

277   The workflow is publicly available and maintained on a GitHub repository

278   (https://github.com/functional-dark-side/agnostos-wf). The workflow has been tested in an

279   HPC cluster setup with at least 4 nodes of 28 cores and 252 G of memory each, which uses

280   SLURM[15] as Grid Batch Scheduler. The workflow programs and software contained in the

281   Conda environment are listed in a .yml file at https://github.com/functional-dark-side/agnostos-

282   wf/blob/master/envs/workflow.yml. The other programs are listed in the bash installation script

283   at https://github.com/functional-dark-side/agnostos-wf/blob/master/installation_script.sh.

284

Data citation

286   1. Vanni, Chiara; Fernandez-Guerra, Antonio (2020): agnostosDB_dbf02445-20200519.

287   figshare. Dataset. https://doi.org/10.6084/m9.figshare.12459056

288   2. Vanni, Chiara; Fernandez-Guerra, Antonio (2020): agnostosDB_a42ac58a-20200715.

289   figshare. Dataset. https://doi.org/10.6084/m9.figshare.13251743

290   3. Vanni, Chiara; Fernandez-Guerra, Antonio (2020): agnostosDB_4eab867d-20201104.

291   figshare. Dataset. https://doi.org/10.6084/m9.figshare.13264769

292   4. Vanni, Chiara (2020): agnostos-wf test dataset. figshare. Dataset.

293   https://doi.org/10.6084/m9.figshare.12630581

294

References

1. Zhang, X. *et al.* Assignment of function to a domain of unknown function: DUF1537 is a new kinase family in catabolic pathways for acid sugars. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E4161-9 (2016).

2. Vanni, C. *et al.* Light into the darkness: Unifying the known and unknown coding sequence space in microbiome analyses. *bioRxiv* 2020.06.30.180448 (2020) doi:10.1101/2020.06.30.180448.

3. Yooseph, S. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS Biol.* **5**, 1–35 (2007).

4. Hurwitz, B. L. & Sullivan, M. B. The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. *PLoS One* **8**, (2013).

5. Brum, J. R. *et al.* Illuminating structural proteins in viral "dark matter" with metaproteomics. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2436–2441 (2016).

6. Wyman, S. K., Avila-Herrera, A., Nayfach, S. & Pollard, K. S. A most wanted list of conserved microbial protein families with no known domains. *PLoS One* **13**, e0205749 (2018).

7. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat. Commun.* **11**, 1710 (2020).

8. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).

317   9.   Dennis Benson, G. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L.
318   GenBank. *Nucleic Acids Res.* **36**, 25–30 (2008).

319   10. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic
320   plankton lineages revealed by genome-resolved metagenomics. (2020)
321   doi:10.1101/2020.10.15.341214.

322   11. Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans
323   microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013).

324   12. Köster, J. Reproducible data analysis with Snakemake. *F1000Res.* **7**, (2018).

325   13. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat.*
326   *Commun.* **9**, 2542 (2018).

327   14. Belmann, P. *et al.* de.NBI Cloud federation through ELIXIR AAI. *F1000Res.* **8**, 842 (2019).

328   15. Yoo, A. B., Jette, M. A. & Grondona, M. SLURM: Simple Linux Utility for Resource
329   Management. *Job Scheduling Strategies for Parallel Processing* 44–60 (2003)
330   doi:10.1007/10968987_3.

331   16. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean
332   microbiome. *Science* **348**, 1261359 (2015).

333   17. Duarte, C. M. Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation
334   Expedition. *Limnol. Oceanog. Bull.* **24**, 11–14 (2015).

335   18. Kopf, A. *et al.* The ocean sampling day consortium. *Gigascience* **4**, 27 (2015).

336   19. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic
337   through Eastern Tropical Pacific. *PLoS Biol.* **5**, 1–34 (2007).

338   20. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome

339   Project. *Nature* **550**, 61–66 (2017).

340   21. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny

341   substantially revises the tree of life. *Nat. Biotechnol.* **36**, (2018).

342   22. Mendler, K. *et al.* AnnoTree: visualization and exploration of a functionally annotated

343   microbial tree of life. *Nucleic Acids Res.* **47**, 4442–4448 (2019).

344   23. Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for

345   the analysis of massive data sets. *Nat. Biotechnol.* **advance on**, (2017).

346   24. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: Lightning-fast iterative protein

347   sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).

348   25. Jaroszewski, L. *et al.* Exploration of Uncharted Regions of the Protein Universe. *PLoS*

349   *Biol.* **7**, e1000205–e1000205 (2009).

350   26. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and

351   stewardship. *Sci Data* **3**, 160018 (2016).

352   27. conda-forge   community.   *The   conda-forge   Project: Community-based   Software*

353   *Distribution   Built   on   the   conda   Package   Format   and   Ecosystem.*   (2015).

354   doi:10.5281/zenodo.4774217.

355   28. Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the

356   life sciences. *Nat. Methods* **15**, 475–476 (2018).

357   29. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat*

358   *Microbiol* **6**, 3–6 (2021).

359   30. Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts in bacterial

360   species, strains, and phage during infant gut colonization. *Genome Res.* **23**, 111–120

361   (2013).