FIGURES

Figure name	Description
Figure_1	Conceptual framework and computational workflow
Figure_2	Overview and evaluation of the workflow to aggregate gene clusters in communities. (A) Methodology overview (B) Numbers of cluster communities per functional category (C) Communities validation based on proteorhodopsin phylogeny (D) Communities validation based on ribosomal proteins, comparing all vs subset containing high quality clusters
Figure_3	Extent of the Known and Unknown coding sequence space (A) Proportion of genes per cluster category (B) Collector curves
Figure_4	Environmental distribution of the unknown coding sequence space (A) Proportion of number of genes / gene abundances per cluster category and biome (B) Relationship between the ratio of GU and EU in HMP samples (C) Relationship between the ratio of GU and EU in TARA samples (D) Distribution of the gene cluster and gene cluster communities based on Levin's niche breadth index
Figure_5	Phylogenomic exploration of the unknown coding sequence space in Bacteria (A) Number of lineage-specific gene clusters per taxonomic level (B) Gene clusters phylogenetic conservation (C) Numbers of non-specific, specific and prophage gene clusters (D) Distribution of bacterial phyla in the Known-Unknown space (E) Results from the integration of the TARA OM-RGC-v2
Figure_6	Mutant phenotypes and the unknown coding sequence space (A) Selected fitness experiment results and selection of the gene cluster GU_19737823 (B) GU_19737823 distribution in metagenomes (C) GU_19737823 community membership (D) Phylogeny of the community and genomic neighborhood of the genes in GU_19737823
Supp. Fig. 1	Overview of the computational workflow
Supp. Fig. 2	(A) Numerical summary of the processed datasets (B) Overlap between the environmental and genomic datasets
Supp. Fig. 3	Proportion of complete genes per cluster (Broken-stick model)

Supp. Fig. 4	Collector curves for the known and unknown coding sequence space (A) gene cluster level for TARA metagenomes considering the viral fraction (B) gene cluster communities level for metagenomes and genomes
Supp. Fig. 5	Collector curves for the known and unknown coding sequence space in metagenomes (A) and genomes(B), excluding the singletons
Supp. Fig. 6	Proportion of gene cluster categories per biome
Supp. Fig. 7	HMP outlier samples enriched in (A) crAssphages (B) papillomaviruses (HPV)
Supplementary Note Figures	
Supp. Fig. 2-1	Proportion of outlier genes per metagenomic gene cluster
Supp. Fig. 4-1	Proportion of outlier genomic genes identified within each genomic gene cluster
Supp. Fig. 7-1	Radar plots used to determine the best MCL inflation value for the partitioning of the metagenomic K into cluster communities (A) Metagenomic dataset (B) Genomic dataset
Supp. Fig. 9-1	Cluster pairs distribution based on the metrics used to weight the gene cluster HMM-HMM homology network (A) HHblits-Score/Aligned-columns (Vanni et al.) (B) maximum(HHblits-probability x coverage) (Méheust et al.)
Supp. Fig. 9-2	Test of the metrics used to weight the gene cluster HMM-HMM homology network (A) Correlation between the Méheust et al. metric and the HHblits-probability (B) Correlation between the Vanni et al. metric and the HHblits-probability (C) Correlation between the Vanni et al. and the Méheust et al. metrics
Supp. Fig. 9-3	Number of communities within ribosomal protein families generated by Méheust et al. 2019 and by Vanni et al. 2020
Supp. Fig. 10-1	EU mapping on TARA MAGs results.
Supp. Fig. 12-1	Coverage of external datasets
Supp. Fig. 13-1	Phylogenomic exploration of the unknown coding sequence space in Archaea. (A) Number of lineage-specific gene clusters per taxonomic level (B) Gene clusters phylogenetic conservation

	(C) Numbers of non-specific, specific and prophage gene clusters (D) Distribution of archaeal phyla in the Known-Unknown space
Supp. Fig. 14-1	Patescibacteria metagenomic lineage specific clusters (A) Proportion of lineage specific clusters in the metagenomes, distributed within the Patescibacteria phylogeny (B) Metagenomic lineage specific clusters in the class Gracilibacteria.

TABLES

Table name	Description
Supp. Table 1	Number of metagenomic clusters and genes after the validation and refinement steps
Supp. Table 2	MG + GTDB high quality (HQ) subset of gene clusters
Supp. Table 3	Mean proportion of complete genes per cluster in the four functional categories
Supp. Table 5	MG + GTDB gene clusters summary statistics
Supp. Table 4	KWP high quality gene clusters (GCs) distribution in the COG groups
Supp. Table 6	Metagenomic input dataset numbers, and gene completion
Supp. Table 7	Proportion of genes in each cluster category
Supp. Table 8	List of HMP outlier samples
Supp. Table 9	EU distribution in MAGs and occurrence in the environment based on the Levin's Niche Breadth index
Supp. Table 10	Number of phylogenetic conserved and lineage-specific GCs in the GTDB bacterial phylogeny
Supp. Table 11	Gene clusters in the GU community GU_g_21103
Supp. Table 12	Lineage-specific clusters of unknown function within Patescibacteria
Supp. Table 13	List of filtered samples used in metagenomic analyses.
Supp. Table 14	List of terms commonly used to define proteins of unknown function in public databases
Supplementary Note Tables	
Supp. Table 1-1	Singletons and small GCs Pfam annotations
Supp. Table 1-2	Number of singletons and small GCs per functional category

Supp. Table 2-1	Number of spurious, shadow and outlier genes in the metagenomic clusters
Supp. Table 2-2	Metagenomic gene cluster validation results
Supp. Table 2-3	Metagenomic gene cluster refinement results step by step
Supp. Table 3-1	Metagenomic gene clusters classification steps (A) Results from the search against the UniRef90 database (B) Results from the search against the and the NCBI nr databases (C) Classification of the Pfam annoptated GCs: consensus DAs
Supp. Table 3-2	Metagenomic gene cluster remote homology refinement steps
Supp. Table 4-1	Genomic genes integration in the metagenomic dataset
Supp. Table 4-2	Genomic gene cluster validation results
Supp. Table 4-3	Spurious, shadow and outlier genes in the genomic cluster
Supp. Table 4-4	Genomic gene clusters classification steps (A) Results from the search against the UniRef90 database (B) Results from the search against the and the NCBI nr databases (C) Classification of the Pfam annotated GCs based on the consensus DAs.
Supp. Table 4-5	Genomic cluster category refinement steps
Supp. Table 4-6	Genomic high quality (HQ) gene clusters
Supp. Table 4-7	MG + GTDB seed database (communities, clusters and genes)
Supp. Table 5-1	Overview of genomic genes found homologous to metagenomic genes.
Supp. Table 6-1	Number of GCs annotated to the DPD per functional category
Supp. Table 7-1	Number of gene clusters, cluster communities and reduction rate shown by functional category for the (A) Metagenomic dataset (B) Genomic dataset
Supp. Table 7-1	Measures of similarity between the community inference proposed in this paper, the one used in Méheust et al. and the “ground truth” represented by the ribosomal protein families
Supp. Table 8-1	Results of viral PRs alignment with Needham at al. viral PRs
Supp. Table 9-2	Minimum slope values for the collector curves
Supp. Table 11-1	Number of genomic singletons per functional category

Supp. Table 13-1	Number of phylogenetic conserved and lineage-specific GCs in the GTDB archaeal phylogeny
Supp. Table 14-1	Number of lineage specific clusters within the Patescibacteria phylum divided by cluster categories