Environmental shaping of codon usage and functional adaptation across microbial communities

Maša Roller, Vedran Lucić, István Nagy, Tina Perica & Kristian Vlahoviček

Supplementary information

TABLE OF CONTENTS

SUPPLEMENTARY FIGURES

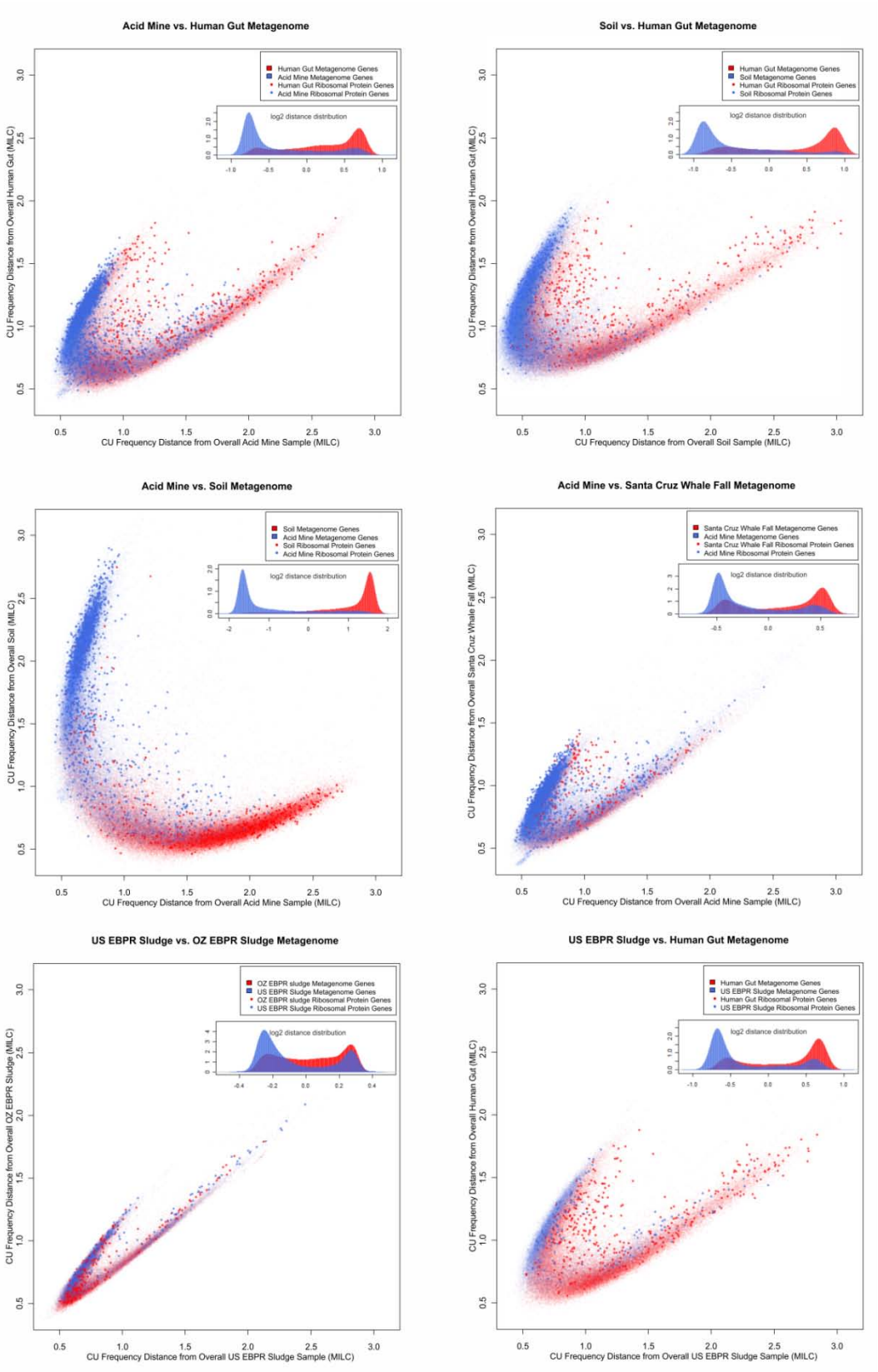

Figure S1: The distance (MILC, outlined in Methods) and distance ratios (MELP, inset for each figure) of each gene's CU frequency to overall CU frequencies of six microbial metagenomes. The separation between two compared genomes is clearly visible in the MELP distributions (inset).

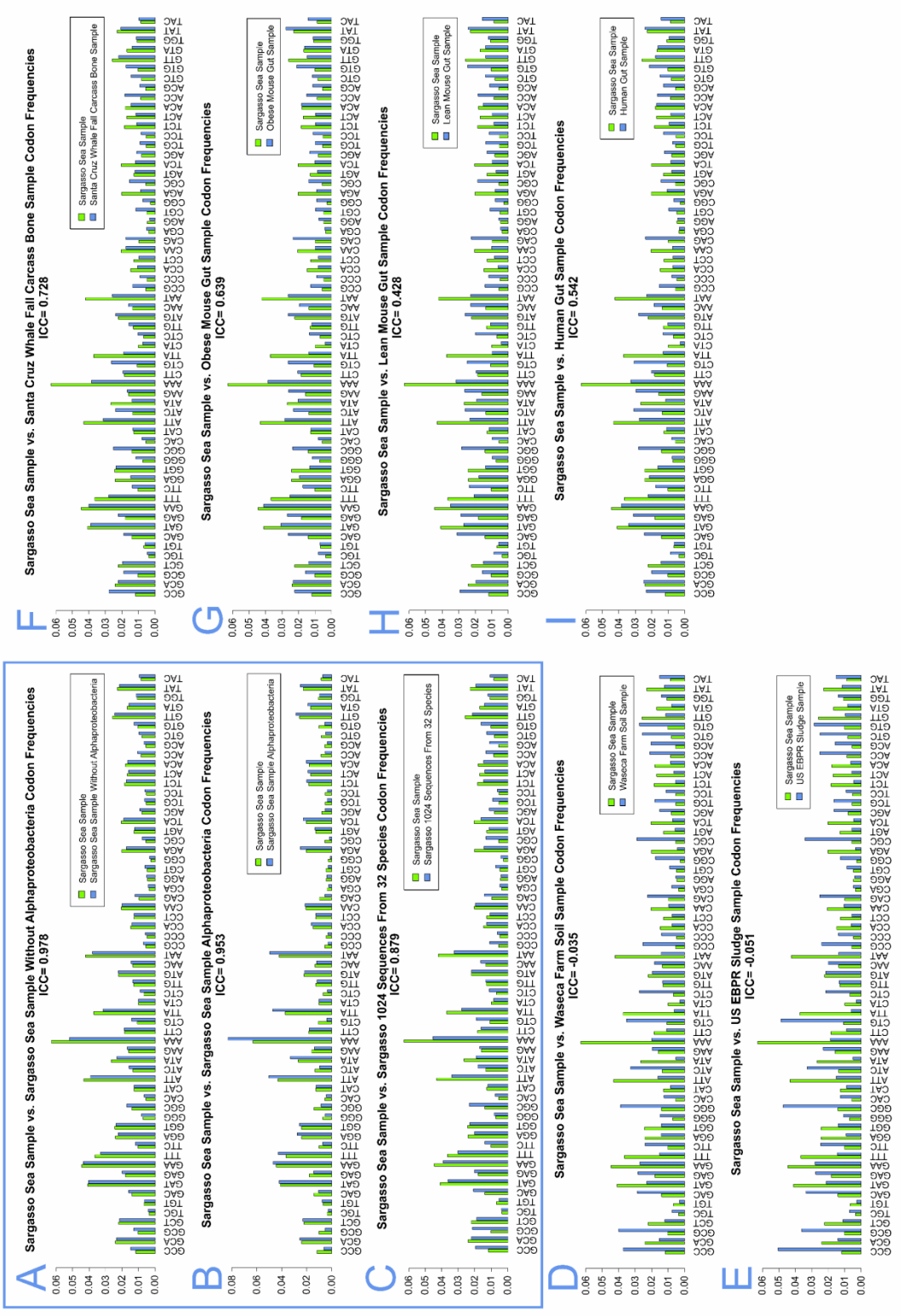

Figure S2: The frequency of all synonymous codons normalised per amino acid in the whole Sargasso Sea sample (N= 155,865,864) compared to A) Sargasso removed for all sequences belonging to the Alphaproteobacteria class (N=10,0594,880), B) Sargasso Alphaproteobacteria only (N=55,270,984), C) Sargasso sample where equal sequence samples per represented phyla were taken (n=32 sequences from 32 phyla, N=1024); D) Waseca soil (N= 16,224,742), E) US EBPR sludge (N= 4,639,388), F) Santa Cruz whale fall carcass bone (N= 4,730,048), G) obese mouse gut (N= 1,510,482), H) lean mouse gut (N= 2,444,470) and I) human gut (N= 11,410,113). The intraclass correlation coefficient (ICC, see Methods) for each comparison is also shown.

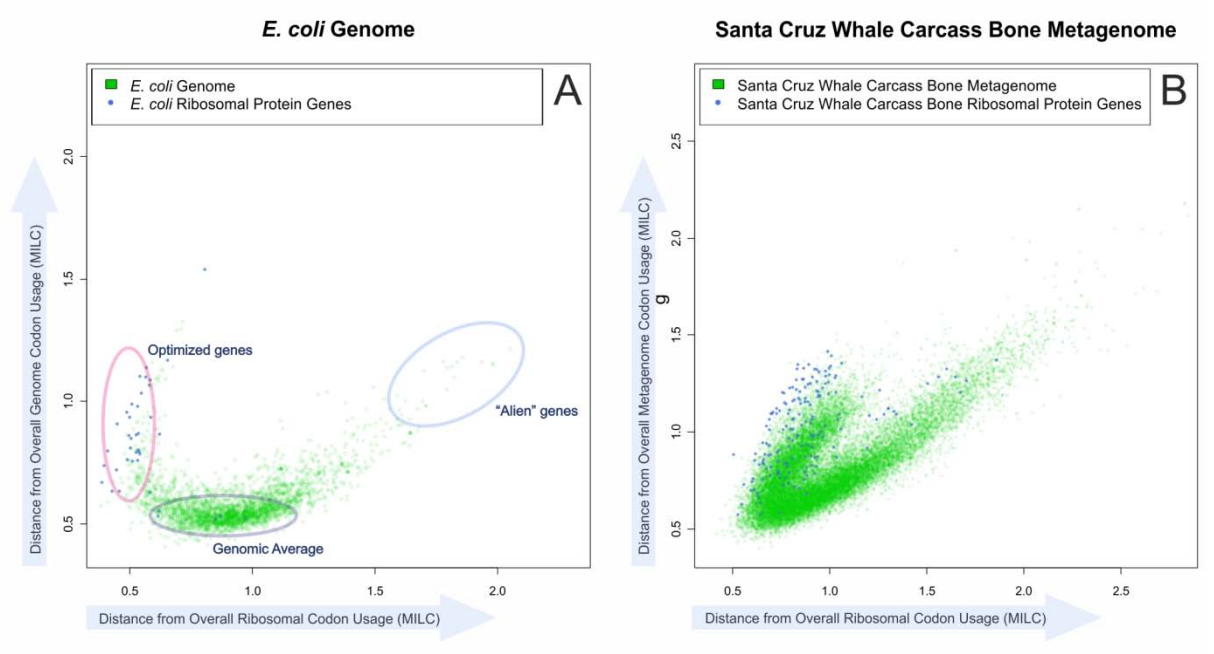

Figure S3: Metagenomes show codon usage distribution similar to single genomes. The distance of each gene's codon usage (CU) frequency form the overall CU of the (meta)genome and ribosomal reference set, displayed as a Karlin B-plot for A) a single microbial genome (*Escherichia coli*, N=4,358) and B) a metagenome (whale carcass, N=33,422). The metagenome shows the same characteristic distribution as the genome with ribosomal genes closer to the CU of the ribosomal set than the overall CU of the whole (meta)genome. MELP – the measure of expression is derived by dividing the gene's distance to the whole genome with that of the distance to the ribosomal protein CU.

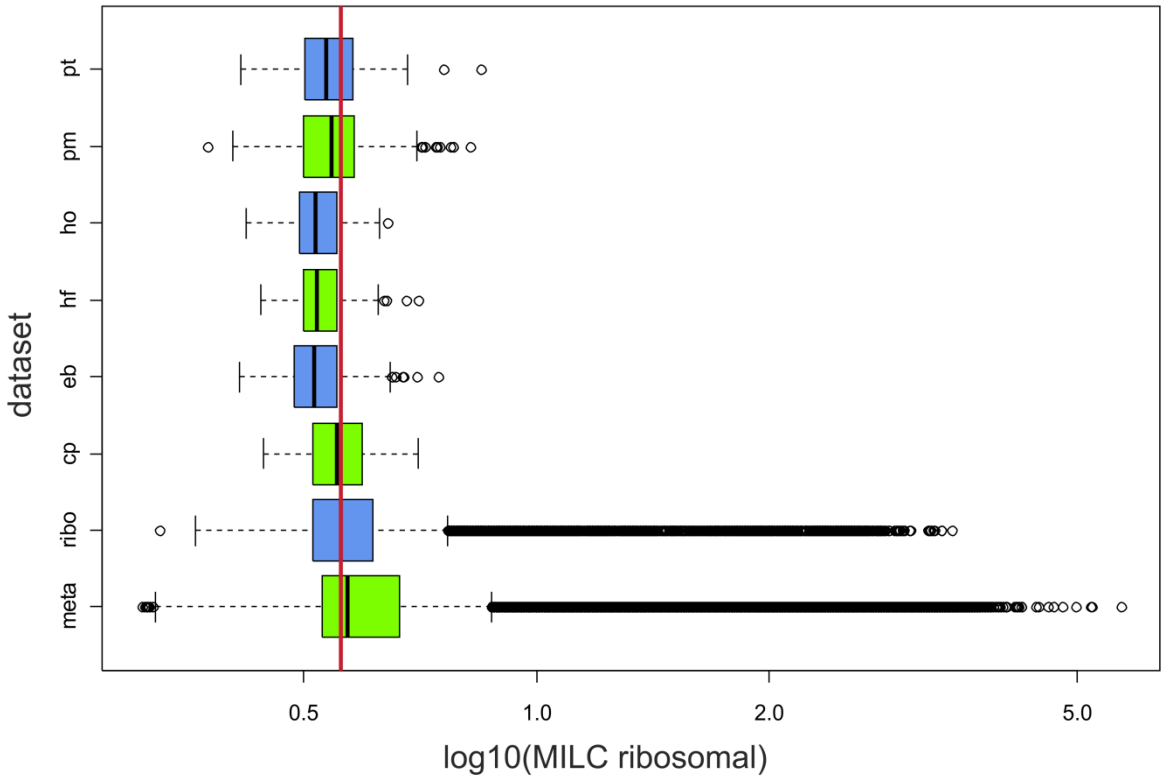

Figure S4: The distribution of CU distances (MILC) from the ribosomal reference set of all genes in the Sargasso Sea metagenome (meta, N=688,539), all ribosomal protein genes (ribo, N=1,3049), and 6 most abundant species: *Candidatus pelagibacter sp.* HTCC7211 (cp, N=214), uncultured marine gamma proteobacterium EBAC20E09 (eb, N=122), uncultured marine microorganism HF4000_005D21 80 (hf, N=80), uncultured marine alpha proteobacterium HOT2C01 (ho, N=70), *Prochlorococcus marinus* (pm, N=347) and *Psychroflexus torquis* (pt, N=181). The red line marks median distance of all ribosomal genes from the ribosomal reference set.

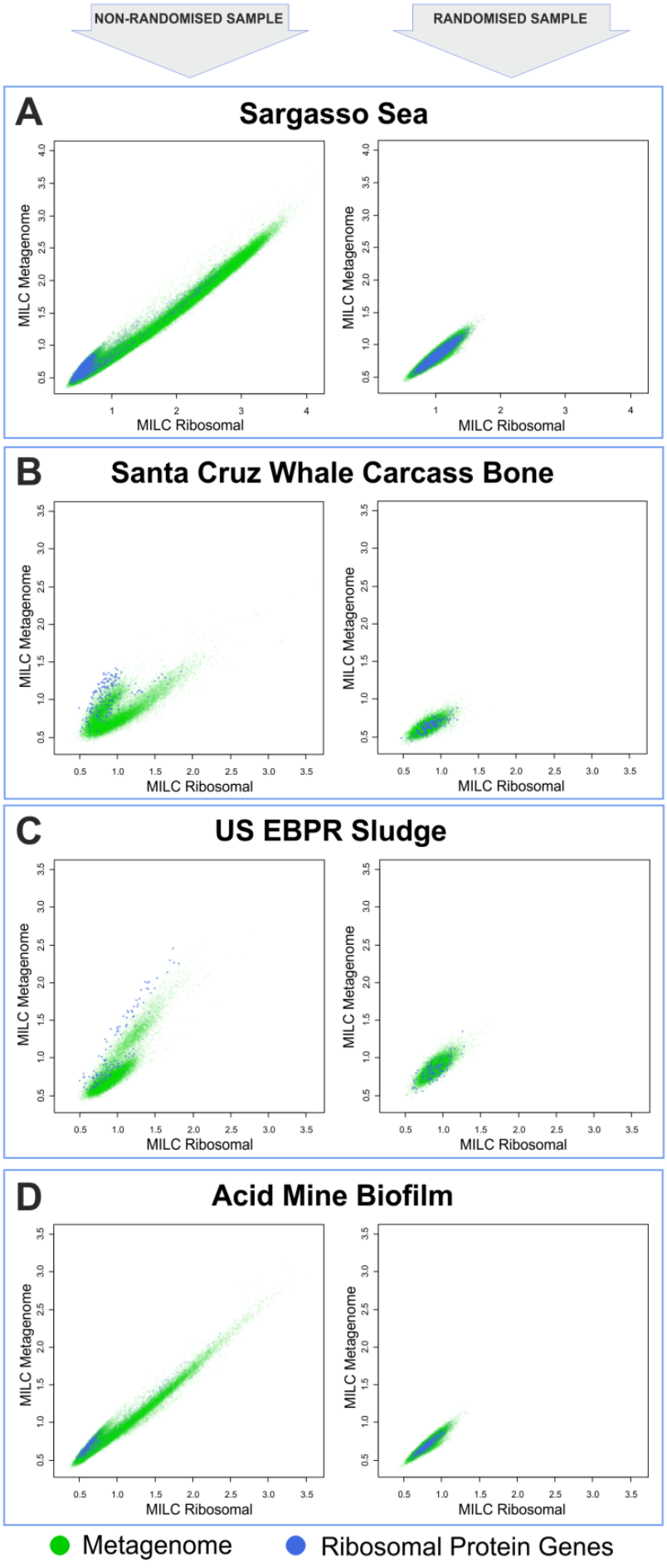

Figure S5. The distance of each gene's codon usage (CU) frequency form the overall CU of the metagenome and ribosomal reference set, displayed as a Karlin B-plot for non-randomised (left panel) and randomised (right panel) metagenomes for A) the Sargasso Sea (N=688,539), B) Santa Cruz whale carcass bone (N=33,422), C) US EBPR sludge (N=20,175) and D) acid mine biofilm (N=79,257) samples. When the amino acid content of a metagenome is kept constant but codons randomly chosen, the metagenome loses its characteristic shape.

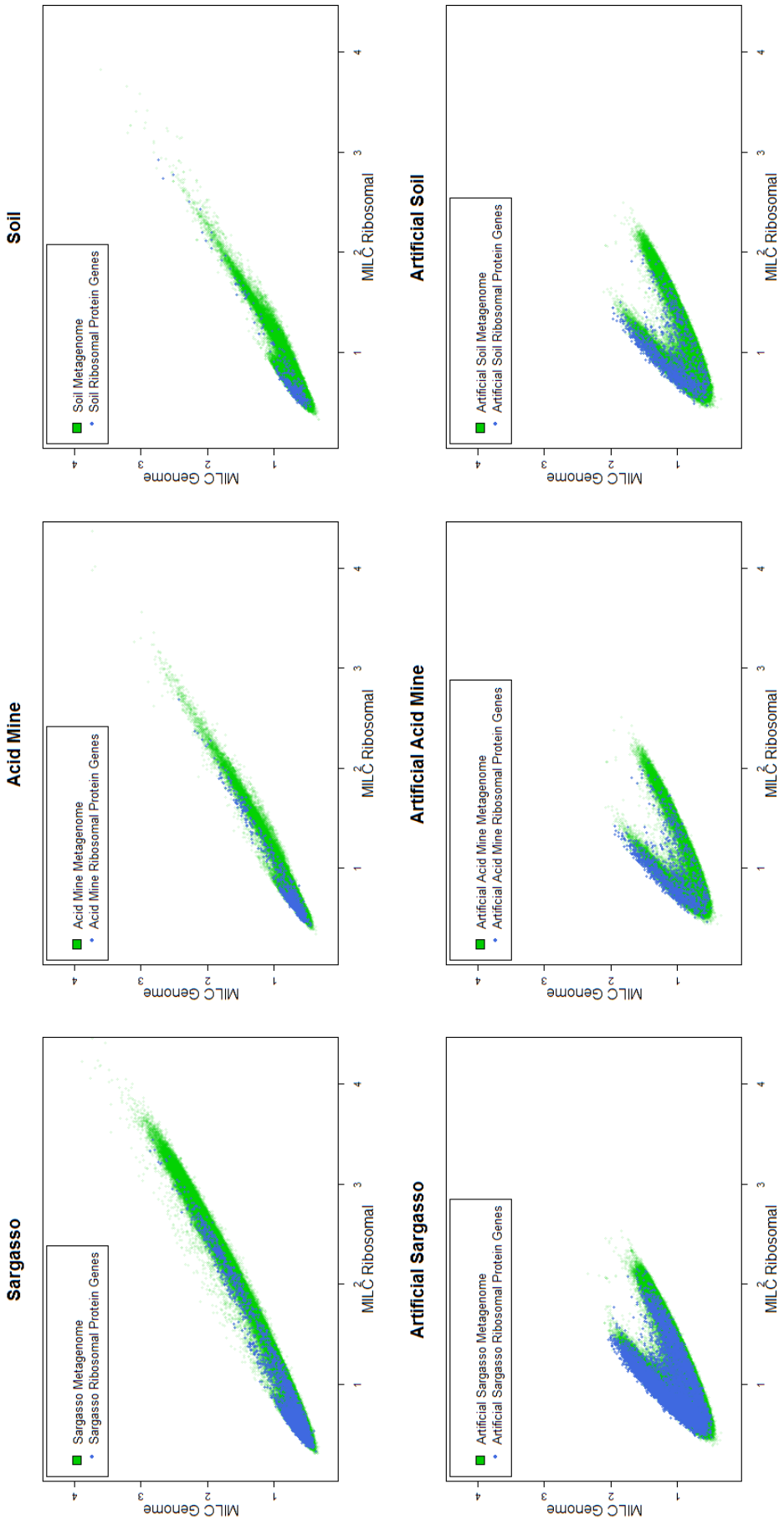

Figure S6: Construction of artificial metagenomes. Artificial metagenomes were constructed by randomly selecting equivalent number of orthologous sequences for each COG group in the real metagenome from the NCBI bacterial genomes dataset. CU distance plots were prepared according to the Methods in the main manuscript. Distance ratios of genes to two axes (the MELP value) were calculated and the results are presented for all metagenomes in Figure S7.

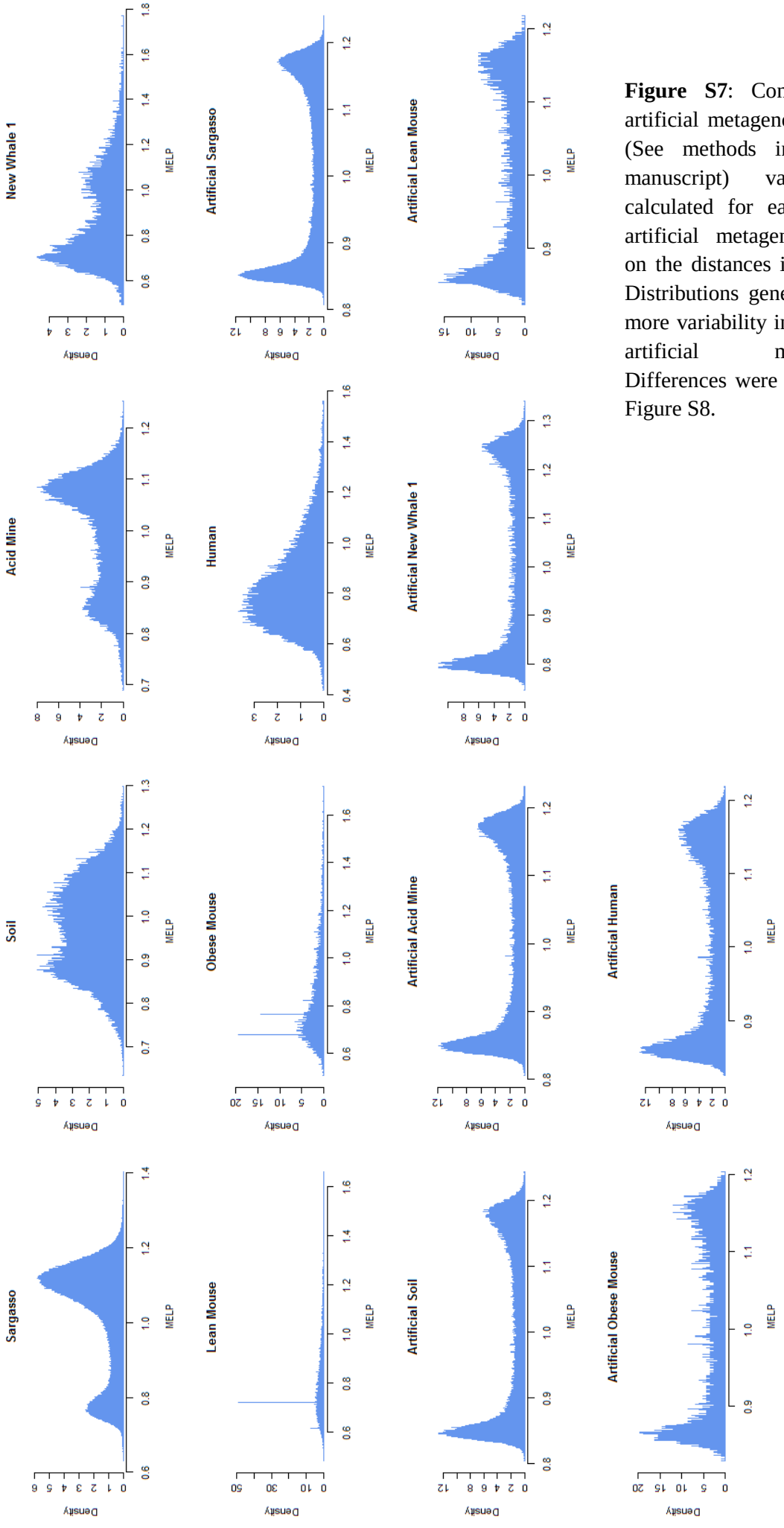

Figure S7: Construction of artificial metagenomes. MELP (See methods in the main manuscript) values were calculated for each real and artificial metagenome, based on the distances in Figure S6. Distributions generally exhibit more variability in real than in artificial metagenomes. Differences were quantified in Figure S8.

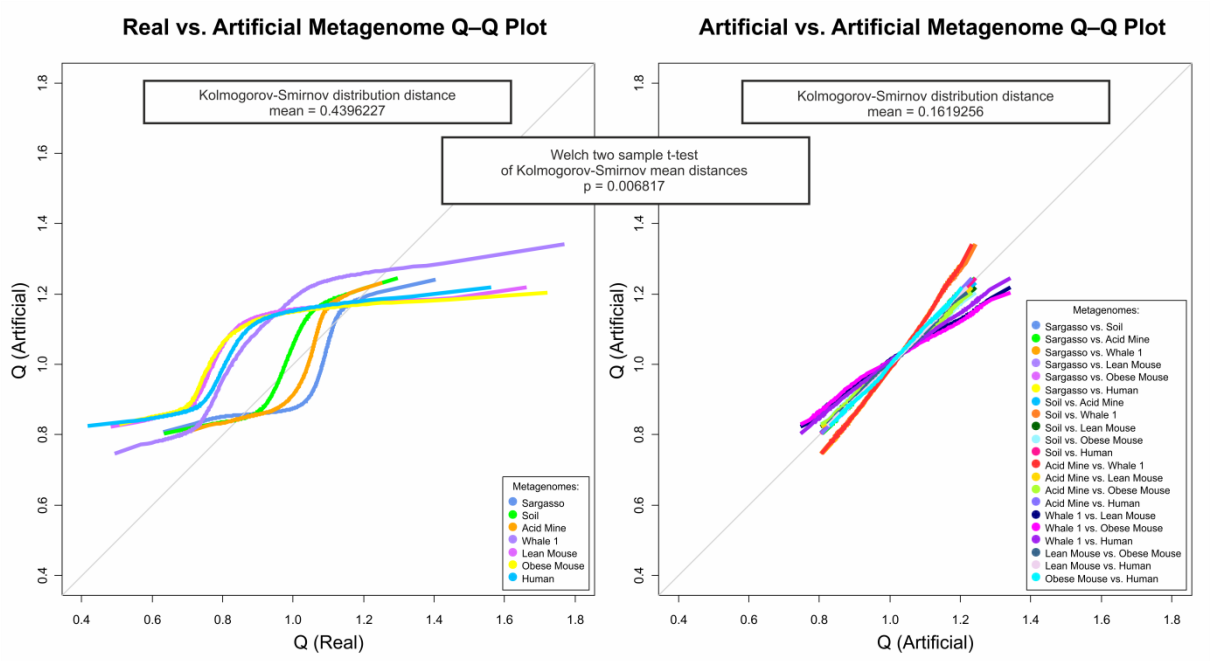

Figure S8: Construction of the artificial metagenomes from the NCBI bacterial genome datasets. Real metagenomes were decomposed into respective COG functional categories and the artificial metagenomes were generated by sampling the equivalent number of orthologous sequences for each COG group of the real metagenome from the bacterial genomes section of the NCBI, regardless of the phyletic composition. MELP distributions were calculated for each metagenome and the distributions (shown in Figure S7) were compared on quantile-quantile plots and evaluated statistically. On overall, the real metagenomes exhibit statistically significant difference in CU distribution variability, while the artificially generated metagenomes tend to adopt more similar and uniform codon usage distribution.

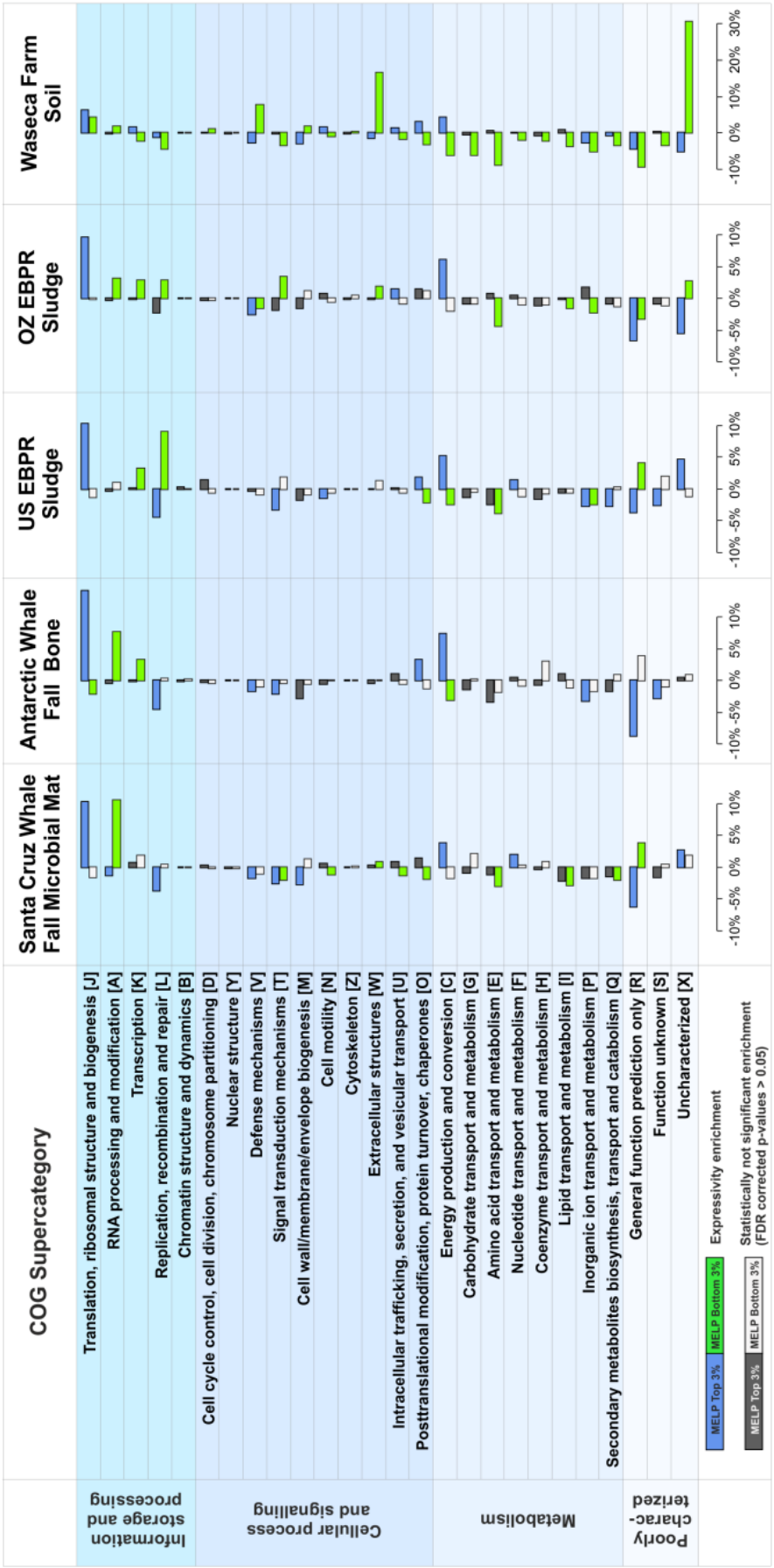

Figure S9: Comparison of gene enrichment for each functional COG supercategory. The top and low 3% of genes by gene expressivity values are shown for 4 metagenomes: Santa Cruz whale fall microbial mat (N=40,916), Antarctica whale fall bone (N=30,503), US (N=20,175) and OZ EBPR sludge (N=29,754).

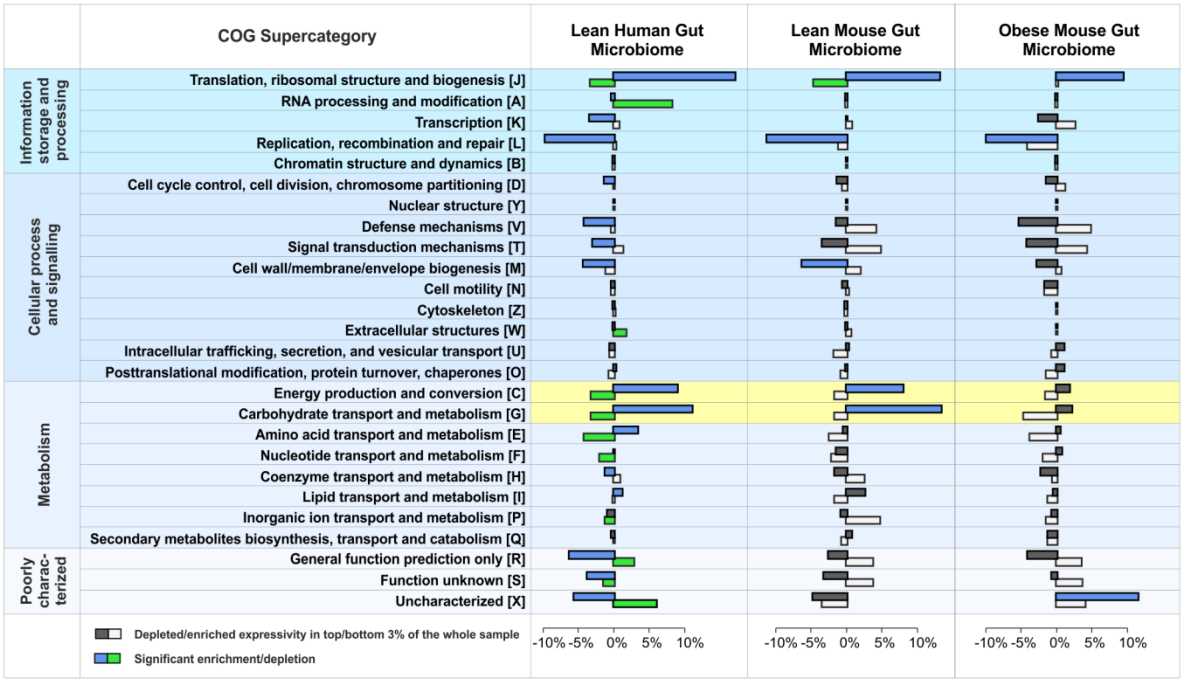

Figure S10: Lean vs. obese gut microbiomes. Comparison of gene enrichment for each functional COG supercategory. The top and low 3% of genes by gene expressivity values are shown for all 3 gut metagenomes: lean human (N=47,765), lean (N=4,955) and obese mouse (N=4,058). Categories highlighted in yellow show the loss of optimisation for two metabolic functions in obese mouse fauna.

Figure S11: Comparison between enrichment profiles in real (A) and artificially assembled metagenomes (B). Artificial metagenomes were assembled from a random selection of NCBI bacterial genomes by maintaining constant function distribution of genes (COG categories) for the original metagenome and with the same total number of genes: (artificial) Sargasso Sea N=688,539, (artificial) acid mine biofilm N=79,257 and (artificial) Santa Cruz whale fall bone N=33,422 . The enrichment patterns visible in real metagenomes are substantially diminished in artificially assembled metagenomes.

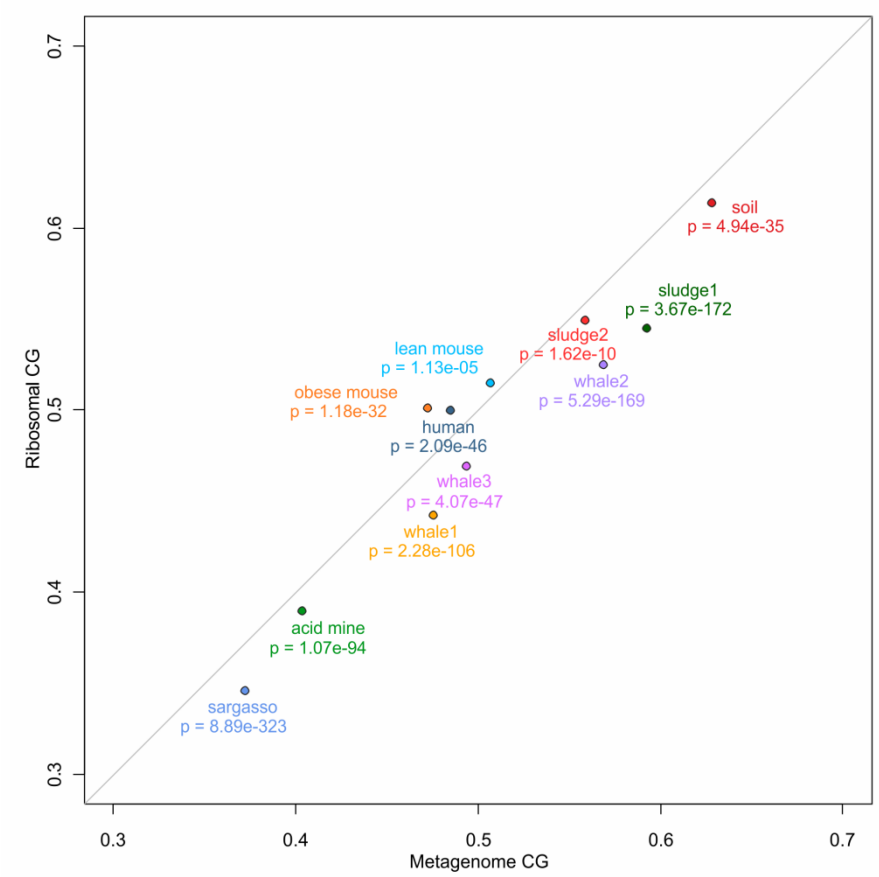

Figure S12: Comparison of the fraction of GC content in ribosomal genes and genes of whole metagenomes. The resulting fractions were tested with the binomial test for each metagenome. All metagenomes show significant differences in GC content between ribosomal genes and all genes of the metagenome.

SUPPLEMENTARY TABLES

Table SI: Names of metagenomes used in this project, their NCBI Project IDs and references for the original sequencing projects.
* In-house assembly of trace data.
†Metagenomes downloaded preassembled.

Metagenome	NCBI Project ID	Reference
†Global Ocean Sampling Expedition Metagenome, the Sargasso Sea version 1	13694	(1)
†Waseca County Farm Soil Metagenome	13699	(2)
†Whale Fall Metagenomes	13700	
*5-Way (CG) Acid Mine Drainage Biofilm Metagenome	13696	(3)
*Human Distal Gut Biome	16729	(4)
*Lean Mouse 1 Gut Metagenome	17391	(5)
*Obese Mouse 1 Gut Metagenome	17397	
*US EBPR Sludge Metagenome	17657	(6)
*OZ EBPR Sludge Metagenome	17659	

Table SII: Number and length of metagenomic sequences used in the study, number of ORFs assigned through homology to the STRING/COG database and using the 3-nearest neighbour consensus rule and the number of ORFs used in MILC and MELP calculations.

	Total number of sequences used	Total length of sequences used (bp)	Number of ORFs assigned
Sargasso Sea	100,1987	585,970,413	688,539
Whale fall Santa Cruz bone	28,151	29,783,639	33,422
Whale fall Santa Cruz microbial mat	29,934	30,862,440	40,916
Whale fall Antarctic bone	26,232	28,868,943	30,503
Waseca farm soil	139,341	144,897,582	88,696
Acid mine biofilm	154,736	206,216,325	79,257
US EBPR sludge	36,222	41,320,987	20,175
OZ EBPR sludge	63,760	81,189,385	29,754
Human gut	79,613	85,004,190	47,765
Lean mouse gut	822,800	65,615,892	4,955
Obese mouse gut	573,519	62,721,309	4,058

Table SIII: Genomes of *P. acnes* and *R. palustris* strains used in this project and references for the original sequencing projects.

Genome	NCBI Reference Sequence	Note	Reference
P. acnes KPA171202	NC_006085.1	Type IB	(7)
P. acnes 6609	CP002815	Type IB	(8)
P. acnes 51318	-	Type IB	in preparation
P. acnes PRP60	-	Type IA	
P. acnes 226	-	Type IA	
P. acnes 434	-	Type IA	
P. acnes PRP38	AIJP00000000	Type IC	(9)
P. acnes ATCC11828	CP003084	Type II	(10)
P. acnes PRP47	-	Type II	in preparation
P. acnes 35934	-	Type II	
P. acnes 9880	-	NC*	
P. acnes 33810	-	NC*	
P. acnes 440671	-	NC*	
R. palustris BisA53	NC_008435.1	none	(11)
R. palustris BisB18	NC_007925	none	
R. palustris BisB5	NC_007958	none	
R. palustris HaA2	NC_007778	none	
R. palustris CGA009	chromosome: NC_005296 plasmid: NC_005297	none	(12)
R. palustris TIE-1	NC_011004	none	none

*not clustered

Table SIV: (Additional File): List of species, the number of genes and their total codon counts per metagenome for species present in at least two metagenomes for species with at least 2,000 codons per metagenome classified with the MEtaGenome Analyzer (MEGAN).

Table SV (Additional File): Phylogenetic clade counts of the ribosomal reference set of the Sargasso Sea metagenome and the whole metagenome classified with the MEtaGenome Analyzer (MEGAN).

Table SVI: The number of genes per COG supercategory in the 6 *R. palustris* strains that fall within 10% of COGs with the smallest variation of MILC median and those that fall within 10% of COGs within the largest variation. The difference of counts per COG category is tested with the binomial test with FDR correction for greater occurrence in the tight 10% than the whole set and for greater occurrence in the wide 10% than the whole sample. COG supercategories with p values below 0.05, marked in yellow, show significant difference in count for the whole set of genes regardless of COG MILC median.

	COG supercategory	number of genes in whole sample	number of genes with tight MILC	FDR corrected p value for tight vs. all	number of genes with wide MILC	FDR corrected p value for wide vs. all
Information storage and processing	[J] Translation, ribosomal structure and biogenesis	967	282	0.00	58	1.00
	[A] RNA processing and modification	0	0	1.00	0	1.00
	[K] Transcription	1396	79	1.00	188	0.03
	[L] Replication, recombination and repair	900	95	0.29	111	0.51
	[B] Chromatin structure and dynamics	10	0	1.00	0	1.00
Cellular processes and signalling	[D] Cell cycle control, cell division, chromosome partitioning	164	12	1.00	26	0.18
	[Y] Nuclear structure	0	0	1.00	0	1.00
	[V] Defence mechanisms	378	10	1.00	174	0.00
	[T] Signal transduction mechanisms	1655	48	1.00	124	1.00
	[M] Cell wall/membrane/envelope biogenesis	1311	57	1.00	189	0.00
	[N] Cell motility	586	39	1.00	49	1.00
	[Z] Cytoskeleton	3	3	0.01	0	1.00
	[W] Extracellular structures	0	0	1.00	0	1.00
	[U] Intracellular trafficking, secretion, and vesicular transport	448	81	0.00	44	1.00
	[O] Posttranslational modification, protein turnover, chaperones	965	120	0.00	61	1.00
Metabolism	[C] Energy production and conversion	1768	163	1.00	157	1.00
	[G] Carbohydrate transport and metabolism	1085	51	1.00	149	0.03
	[E] Amino acid transport and metabolism	2338	104	1.00	81	1.00
	[F] Nucleotide transport and metabolism	401	61	0.00	35	1.00
	[H] Coenzyme transport and metabolism	1031	76	1.00	57	1.00
	[I] Lipid transport and metabolism	1438	40	1.00	283	0.00
	[P] Inorganic ion transport and metabolism	1525	76	1.00	110	1.00
	[Q] Secondary metabolites biosynthesis, transport and catabolism	901	24	1.00	262	0.00
Poorly characterized	[R] General function prediction only	2872	286	0.27	447	0.00
	[S] Function unknown	1929	499	0.00	107	1.00
	[X] Uncharacterised	0	0	1.00	0	1.00

Table SVII: The number of genes per COG supercategory in the 12 *P. acnes* that fall within 10% of COGs with the smallest variation of MILC median and those that fall within 10% of COGs within the largest variation. The

difference of counts per COG category is tested with the binomial test with FDR correction for greater occurrence in the tight 10% than the whole set and for greater occurrence in the wide 10% than the whole sample. COG supercategories with p values below 0.05, marked in yellow, show significant difference in count for the whole set of genes regardless of COG MILC median.

	COG supercategory	number of genes in whole sample	number of genes with tight MILC	FDR corrected p-value for tight vs. all	number of genes with wide MILC	FDR corrected p-value for wide vs. all
Information storage and processing	[J] Translation, ribosomal structure and biogenesis	1167	339	0.00	54	1.00
	[A] RNA processing and modification	10	0	1.00	0	1.00
	[K] Transcription	1113	87	1.00	132	0.33
	[L] Replication, recombination and repair	750	51	1.00	86	0.80
	[B] Chromatin structure and dynamics	0	0	1.00	0	1.00
Cellular processes and signalling	[D] Cell cycle control, cell division, chromosome partitioning	138	0	1.00	0	1.00
	[Y] Nuclear structure	0	0	1.00	0	1.00
	[V] Defense mechanisms	269	3	1.00	12	1.00
	[T] Signal transduction mechanisms	468	36	1.00	100	0.00
	[M] Cell wall/membrane/envelope biogenesis	704	27	1.00	0	1.00
	[N] Cell motility	0	0	1.00	0	1.00
	[Z] Cytoskeleton	0	0	1.00	0	1.00
	[W] Extracellular structures	0	0	1.00	0	1.00
	[U] Intracellular trafficking, secretion, and vesicular transport	186	39	0.00	0	1.00
	[O] Posttranslational modification, protein turnover, chaperones	578	65	0.56	35	1.00
Metabolism	[C] Energy production and conversion	1104	71	1.00	66	1.00
	[G] Carbohydrate transport and metabolism	1846	68	1.00	537	0.00
	[E] Amino acid transport and metabolism	1454	61	1.00	103	1.00
	[F] Nucleotide transport and metabolism	652	66	1.00	0	1.00
	[H] Coenzyme transport and metabolism	984	83	1.00	52	1.00
	[I] Lipid transport and metabolism	461	22	1.00	101	0.00
	[P] Inorganic ion transport and metabolism	896	78	1.00	121	0.01
	[Q] Secondary metabolites biosynthesis, transport and catabolism	156	9	1.00	19	0.90
Poorly characterized	[R] General function prediction only	1538	103	1.00	78	1.00
	[S] Function unknown	962	248	0.00	106	0.90
	[X] Uncharacterized	0	0	1.00	0	1.00

References

1. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D.Y., Paulsen, I., Nelson, K.E., Nelson, W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66-74.
2. Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., Podar, M., Short, J.M., Mathur, E.J., Detter, J.C. *et al.* (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554-557.
3. Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S. and Banfield, J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37-43.
4. Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M. and Nelson, K.E. (2006) Metagenomic analysis of the human distal gut microbiome. *Science*, **312**, 1355-1359.
5. Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R. and Gordon, J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027-1031.
6. Martin, H.G., Ivanova, N., Kunin, V., Warnecke, F., Barry, K.W., McHardy, A.C., Yeates, C., He, S.M., Salamov, A.A., Szeto, E. *et al.* (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol.*, **24**, 1263-1269.
7. Bruggemann, H., Henne, A., Hoster, F., Liesegang, H., Wiezer, A., Strittmatter, A., Hujer, S., Durre, P. and Gottschalk, G. (2004) The complete genome sequence of Propionibacterium acnes, a commensal of human skin. *Science*, **305**, 671-673.
8. Hunyadkurti, J., Feltoti, Z., Horvath, B., Nagymihaly, M., Voros, A., McDowell, A., Patrick, S., Urban, E. and Nagy, I. (2011) Complete genome sequence of Propionibacterium acnes type IB strain 6609. *J Bacteriol*, **193**, 4561-4562.
9. McDowell, A., Hunyadkurti, J., Horvath, B., Voros, A., Barnard, E., Patrick, S. and Nagy, I. (2012) Draft genome sequence of an antibiotic-resistant Propionibacterium acnes strain, PRP-38, from the novel type IC cluster. *J Bacteriol*, **194**, 3260-3261.
10. Horvath, B., Hunyadkurti, J., Voros, A., Fekete, C., Urban, E., Kemeny, L. and Nagy, I. (2012) Genome sequence of Propionibacterium acnes type II strain ATCC 11828. *J Bacteriol*, **194**, 202-203.
11. Oda, Y., Larimer, F.W., Chain, P.S.G., Malfatti, S., Shin, M.V., Vergez, L.M., Hauser, L., Land, M.L., Braatsch, S., Beatty, J.T. *et al.* (2008) Multiple genome sequences reveal adaptations of a phototrophic bacterium to sediment microenvironments. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 18543-18548.
12. Larimer, F.W., Chain, P., Hauser, L., Lamerdin, J., Malfatti, S., Do, L., Land, M.L., Pelletier, D.A., Beatty, J.T., Lang, A.S. *et al.* (2004) Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris. *Nat. Biotechnol.*, **22**, 55-61.