

Shotgun metagenomic profiles have a high capacity to discriminate activated sludge treatment plants according to wastewater type

Federico M. Ibarbalz¹, Esteban Orellana¹, Eva L.M. Figuerola^{1,2} and Leonardo Erijman^{1,2*}

¹Instituto de Investigaciones en Ingeniería Genética y Biología Molecular (INGEBI-CONICET).
Buenos Aires, Argentina

²Departamento de Fisiología, Biología Molecular y Celular, Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. Buenos Aires, Argentina

Supplementary information

Table S1. Main features of the wastewater treatment plants analyzed in this work.

Label	Location	Sampling Date	Wastewater type	Scale	Configuration	Population served (or PE)
P1-a	Buenos Aires, Argentina	May-2009	Petrochemical (Petroleum Refinery)	Full-scale	Anoxic/Aerobic	15,000 (PE)
P1-b		May-2011				
P2-a	Buenos Aires, Argentina	Sep-2008	Petrochemical (Acrylic Polymer Synthesis)	Full-scale	Aerobic	450 (PE)
P2-a		May-2009				
T-a	Buenos Aires, Argentina	Sep-2008	Textile Dyeing	Full-scale	Aerobic	25,000 (PE)
T-b		Jun-2011				
W-a	Córdoba, Argentina	May-2010	Whey filtering	Full-scale	Aerobic	45,000 (PE)
W-b		Jun-2011				
S1-a	Buenos Aires, Argentina	Nov-2008	Municipal	Full-scale	Aerobic	3,500
S1-b		Apr-2013				
S2-a	Buenos Aires, Argentina	Oct-2012	Municipal	Full-scale	Aerobic	270,000
S2-b		Aug-2013				
P3-a	Gujarat, Western India	Oct-2013	Petrochemical (Petroleum Refinery Low TDS)	Full-scale	NA	NA
P3-b		Feb-2014				
P4-a	Gujarat, Western India	Oct-2013	Petrochemical (Petroleum Refinery High TDS)	Full-scale	NA	NA
P4-b		Feb-2014				
S3	Dachang, China	Jul-2012	Municipal	Full-scale	NA	NA
S4	Wuxi, China	Aug-2013	Municipal	Full-scale	NA	NA
S5	Zhengzhou, China	Nov-2012	Municipal	Full-scale	NA	NA
S6	Stanley, Hong Kong	Feb-2011	Municipal	Full-scale	Anoxic/Aerobic	28,000

NA: Not available

Table S2. Main operational parameters of the wastewater treatment plants analyzed in this work.

Label	ABV (m ³)	Avg DO (mg/l)	Effluent pH	MLSS (mg/l)	Temp. (°C)	Influent COD (mg/l)	OLR (kg BOD ₅ /m ³ d)	Effluent COD (mg/l)	Effluent BOD ₅ (mg/l)	Effluent NO ₃ ⁻ (mg/l)	SRT (d)	Protein (mg/L)	CHO (mg/L)	VFA (eq/m ³)
P1-a	5,430	1.9	7.1-7.6	4,850	32-37	516-837	0.44	58-120	37-49	48-68	86	<10	2	0.56
P1-b		3.9	7.1-7.7	4,425	33-37	674-901	0.51	73-134	25-41	61-70	69			
P2-a	200	8.1	7.0-7.4	4,350	24-27	439-920	0.07	47-88	17-34	33-46	70	NA	NA	NA
P2-b		8.6	7.0-7.5	3,313	22-26	475-526	0.06	34-71	10-28	41-64	53			
T-a	3,500	0.7	8.0-8.8	3,594	26-33	685-922	0.33	59-104	10-33	< 1	68	10	25	0.12
T-b		0.6	8.0-8.9	6,563	26-30	515-593	0.29	38-121	12-31	< 1	40			
W-a	5,000	0.9	7.2-7.7	8,163	24-28	3,545-5,030	0.6	32-80	15-29	37-61	28	160	75	28.0
W-b		0.8	7.2-7.8	6,531	22-28	2,598-3,883	0.63	21-112	8-34	44-56	33			
S1-a	200	3.4	7.1-7.5	4,469	17-24	463-738	0.41	54-66	18-29	NA	18	20	6	0.16
S1-b		3.1	7.1-7.5	5,002	17-22	142-546	0.38	37-58	15-25	NA	23			
S2-a	13,400	NA	7.0-7.6	2,110	18-23	212-284	1.37	33-78	12-24	9-18	7	10	8	0.12
S2-b		3.2	7.0-7.6	2,445	16-20	212-284	1.59	41-84	11-26	NA	7			
P3-a	NA	2.5	6.5-7.0	3,000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
P3-b	NA	2.5	6.5-7.0	3,000	NA	NA	NA	NA	NA	NA	NA			
P4-a	NA	2.9	7.7-8.0	4,000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
P4-b	NA	2.9	7.7-8.0	4,000	NA	NA	NA	NA	NA	NA	NA			
S3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
S4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
S5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
S6	2,716	NA	6.6-7.0	2,500	24-32	196-481	0.62	NA	NA	NA	NA	NA	NA	NA

ABV: Aeration basin volume

DO: Dissolved oxygen

MLSS: Mixed liquor suspended solids

COD: Chemical oxygen demand

OLR: Organic loading rate

SRT: Solids retention time

CHO: Carbohydrates

VFA: Volatile fatty acids

Table S3. Extraction method, summary of sequencing output and database accession numbers. Technical replicates were averaged.

Label	DNA extraction	Sequencing technology	Number of reads	High-quality reads ¹	Protein coding regions ¹	MG-RAST ID	NCBI-SRA ID
P1-a	FastDNA Spin Kit for Soil	Illumina HiSeq1500	1.51E+07	1.30E+07 (86.6)	6.73E+06 (51.4)	4603523.3; 4603525.3	SRR2107128; SRR2107133
P1-b	FastDNA Spin Kit for Soil	Illumina HiSeq1500	1.23E+07	1.05E+07 (85.7)	5.03E+06 (47.8)	4603527.3; 4603529.3	SRR2107174; SRR2107175
P2-a	FastDNA Spin Kit for Soil	Illumina HiSeq1500	1.37E+07	1.20E+07 (87.5)	6.90E+06 (57.1)	4603531.3; 4603533.3	SRR2107171; SRR2107173
P2-a	FastDNA Spin Kit for Soil	Illumina HiSeq1500	1.35E+07	1.18E+07 (87.2)	6.71E+06 (56.9)	4603535.3; 4603537.3	SRR2107176; SRR2107177
T-a	FastDNA Spin Kit for Soil	Illumina HiSeq1500	1.05E+07	9.10E+06 (87.0)	4.83E+06 (53.0)	4603539.3; 4603541.3	SRR2107178; SRR2107179
T-b	FastDNA Spin Kit for Soil	Illumina HiSeq1500	1.25E+07	1.09E+07 (87.5)	6.66E+06 (60.5)	4603543.3; 4603545.3	SRR2107180; SRR2107181
W-a	FastDNA Spin Kit for Soil	Illumina HiSeq1500	1.56E+07	1.39E+07 (89.3)	8.44E+06 (60.5)	4603547.3; 4603549.3	SRR2107182; SRR2107183
W-b	FastDNA Spin Kit for Soil	Illumina HiSeq1500	1.25E+07	1.09E+07 (87.8)	5.66E+06 (51.6)	4603551.3; 4603553.3	SRR2107184; SRR2107185
S1-a	FastDNA Spin Kit for Soil	Illumina HiSeq1500	1.33E+07	1.23E+07 (92.3)	9.47E+06 (76.8)	4603555.3; 4603557.3	SRR2107210; SRR2107211
S1-b	FastDNA Spin Kit for Soil	Illumina HiSeq1500	1.07E+07	9.81E+06 (92.1)	7.34E+06 (74.9)	4603559.3; 4603561.3	SRR2107212; SRR2107213
S2-a	FastDNA Spin Kit for Soil	Illumina HiSeq1500	1.51E+07	1.39E+07 (92.1)	9.52E+06 (68.6)	4603563.3; 4603565.3	SRR2107215; SRR2107216
S2-b	FastDNA Spin Kit for Soil	Illumina HiSeq1500	1.51E+07	1.40E+07 (93.3)	9.01E+06 (64.2)	4603567.3; 4603569.3	SRR2107218; SRR2107219
P3-a	FastDNA Spin Kit for Soil	Illumina NextSeq500	4.42E+06	4.08E+06 (92.2)	3.09E+06 (75.8)	4660921.3	SRR1702225
P3-b	FastDNA Spin Kit for Soil	Illumina NextSeq500	4.42E+06	4.60E+06 (91.6)	3.67E+06 (79.8)	4636805.3	SRR1702234
P4-a	FastDNA Spin Kit for Soil	Illumina NextSeq500	5.28E+06	4.81E+06 (91.0)	3.15E+06 (65.6)	4660920.3	SRR1702232
P4-b	FastDNA Spin Kit for Soil	Illumina NextSeq500	5.28E+06	3.98E+06 (92.2)	3.30E+06 (82.9)	4636804.3	SRR1702233
S3	NA	Illumina HiSeq2000	1.25E+07	1.11E+07 (89.4)	9.26E+06 (83.1)	4629717.3	SRR1106701
S4	NA	Illumina HiSeq2000	1.21E+07	1.07E+07 (88.3)	8.43E+06 (78.6)	4629718.3	SRR1106709
S5	NA	Illumina HiSeq2000	6.78E+06	6.62E+06 (97.6)	5.70E+06 (86.1)	4660919.3	SRR1106873
S6	FastDNA Spin Kit for Soil	Illumina HiSeq2000	1.67E+07	1.57E+07 (93.9)	1.13E+07 (72.1)	4629725.3	SRR609293

¹In brackets, percentage relative to previous column.

Table S4. Spearman's rank correlation coefficients for multiple pairwise correlations between 11 highly-represented phyla, utilizing GC content values in all WWTPs (n = 12). Values $\rho > 0.56$ and $\rho \leq 0.05$ were highlighted. Phyla names were shortened. Actb: *Actinobacteria*; Prtb: *Proteobacteria*; Bctr: *Bacteroidetes*; Plct: *Planctomycetes*; Acdb: *Acidobacteria*; Vrcm: *Verrucomicrobia*; Clfx: *Chloroflexi*; Ntrp: *Nitrospirae*; Clrb: *Chlorobi*; Euch: *Euryarchaeota*; DcTh: *Deinococcus-Thermus*.

	Actb.	Prtb.	Bctr.	Plct.	Acdb.	Vrcm.	Clfx.	Ntrp.	Clrb.	Euch.	DcTh
Actb.	-	0.31	0.05	0.24	0	0.05	0.29	0.38	0.2	0.29	0.39
Prtb.	0.31	-	0.41	0.65	0.14	0.45	0.33	0.72	0.13	0.74	0.73
Bctr.	0.05	0.41	-	0.46	0.35	0.68	0.58	0.13	0.48	0.55	0.4
Plct.	0.24	0.65	0.46	-	0.7	0.78	0.81	0.63	0.66	0.74	0.87
Acdb.	0	0.14	0.35	0.7	-	0.85	0.76	0.35	0.93	0.45	0.37
Vrcm.	0.05	0.45	0.68	0.78	0.85	-	0.87	0.41	0.84	0.71	0.56
Clfx.	0.29	0.33	0.58	0.81	0.76	0.87	-	0.26	0.8	0.6	0.64
Ntrp.	0.38	0.72	0.13	0.63	0.35	0.41	0.26	-	0.28	0.7	0.66
Clrb.	0.2	0.13	0.48	0.66	0.93	0.84	0.8	0.28	-	0.46	0.38
Euch.	0.29	0.74	0.55	0.74	0.45	0.71	0.6	0.7	0.46	-	0.62
DcTh	0.39	0.73	0.4	0.87	0.37	0.56	0.64	0.66	0.38	0.62	-

Table S5: Correspondence Analysis (CA) and Principal Components Analysis (PCA) based on SEED subsystems (level 3), in order to retrieve the most relevant drivers of both ordinations.

Relative loading, CA1 (37 %)¹	Subsystem
1.000	Petrobactin-mediated iron uptake system
0.802	D-allose utilization
0.796	Siderophore [Alcaligin-like]
0.768	Betaine biosynthesis from glycine
0.763	Siderophore Enterobactin
0.743	<i>Salmonella</i> invasion locus
0.726	Energy-conserving hydrogenase (ferredoxin) Ech
0.660	Membrane bound hydrogenases
0.655	Archaeal membrane bound hydrogenases
0.648	(GlcNAc) ₂ Catabolic Operon
-1.000	alpha-Fimbriae
-0.849	<i>Streptococcus agalactiae</i> hemolysin operon
-0.849	Phage Ea cluster
-0.817	Archease
-0.764	Lysozyme inhibitors
-0.741	PFGI-1-like cluster 2
-0.734	PFGI-1-like cluster 1
-0.680	Acetophenone carboxylase 1
-0.644	Biofilm Adhesin Biosynthesis
-0.637	Predicted mycobacterial monooxygenase
Relative loading, PCA1 (52 %)¹	Subsystem
1.000	Sugar utilization in Thermotogales
0.349	COG3533
0.332	Universal GTPases
0.267	Maltose and Maltodextrin Utilization
0.264	Lactose and Galactose Uptake and Utilization
0.255	D-Galacturonate and D-Glucuronate Utilization
0.254	TCA Cycle
0.216	Translation elongation factors bacterial
0.212	Galactosylceramide and Sulfatide metabolism
0.212	DNA-replication
-1.000	Fatty acid degradation regulons
-0.852	Fatty acid metabolism cluster
-0.849	n-Phenylalkanoic acid degradation
-0.767	Acetyl-CoA fermentation to Butyrate
-0.704	Polyhydroxybutyrate metabolism
-0.670	Isoleucine degradation
-0.653	ABC transporter branched-chain amino acid (TC 3.A.1.4.1)
-0.568	Biotin biosynthesis
-0.549	Butanol Biosynthesis
-0.495	Flagellum

1: In brackets, proportion of variance explained by main axis.

Table S6. Permutation test to evaluate relevant process variables in their fit to a PCA ordination based on SEED subsystems (level 1). P-values are based on 1000 permutations.

VECTORS				
	PC1	PC2	r ²	p
DO	-0.9137	0.4063	0.8153	0.002
pH	0.9872	-0.1594	0.4200	0.080
MLSS	0.9002	0.4356	0.4300	0.075
T	-0.6226	-0.7825	0.1171	0.562
SRT	-0.9046	0.4264	0.2095	0.347

FACTORS		
Centroids	PC1	PC2
P1	-0.6243	-0.3024
P2	-0.8605	0.8515
S1	0.0774	-0.6336
S2	-0.1050	-0.3760
T	0.5903	-0.3235
W	0.9221	0.7841

Goodness of fit		
	r ²	p
Site	0.9384	0.001

Table S7: Assembly statistics for recovered genomic sequences, obtained from activated sludge by differential coverage binning of two temporally replicated metagenomes. The score refers to the bi-directional best-hit classification.

ID	Size (bp)	Number of scaffolds	Mean scaffold length (bp)	Maximum scaffold length (bp)	GC (%)	Number of essential genes	Nr. of unique essential genes	Relative abundance (%)	Assembled scaffolds ¹	Bi-directional best hit classification	Score
P1GA1	3.56E+06	93	3.83E+04	1.84E+05	72.7	78	77	2.0	a	<i>Leptothrix cholodni</i>	507
										<i>Methylibium petroleiphilum</i> PM1	495
P1GB1	2.31E+06	455	5.08E+03	5.73E+04	69.2	95	93	0.8	a	<i>Nitrosococcus watsoni</i>	504
										<i>Alkalilimnicola ehrlichii</i> MLHE-1	487
P1GC1	3.82E+06	382	1.00E+04	9.47E+04	66.6	102	100	2.2	a	<i>Herpetosiphon aurantiacus</i>	513
										<i>Roseiflexus</i> sp. RS-1	409
P1GE1	3.06E+06	372	8.21E+03	8.99E+04	60.5	96	90	1.6	a	<i>Anaerolinea thermophila</i>	534
P1GG1	2.83E+06	20	1.42E+05	7.88E+05	57.3	97	94	3.6	a	<i>Desulfuromonas</i> sp.	545
P1GI1	2.15E+06	220	9.79E+03	1.39E+05	65.6	85	84	3.3	a	<i>Alkalilimnicola ehrlichii</i>	521
										<i>Methylococcus capsulatus</i> str. Bath	493
P1GK1	3.31E+06	193	1.72E+04	8.93E+04	69.5	86	84	2.5	a	<i>Polaromonas naphthalenivorans</i>	526
										<i>Alicyclophilus denitrificans</i>	505
P1GL1	2.31E+06	53	43617.6	3.56E+05	65.6	89	87	1.1	a	<i>Thiobacillus denitrificans</i>	522
P1GL2	4.32E+06	46	93855.0	6.07E+05	68.2	85	85	2.2	a	<i>Thioalkalivibrio</i> sp.	540
P1GL3	3.65E+06	44	83064.4	7.95E+05	62.8	110	105	2.1	a	<i>Hyphomicrobium denitrificans</i>	534
P1GL4	2.60E+06	141	2.32E+04	2.28E+05	68.3	88	85	1.5	a	<i>Azoarcus</i> sp.	525
										<i>Aromatoleum aromaticum</i>	513
										<i>Sideroxydans lithotrophicus</i>	496
										<i>Ignavibacterium album</i>	548
P1GP1	3.39E+06	34	9.98E+04	3.56E+05	33.4	104	103	4.9	a	<i>Mycobacterium vanbaaleni</i>	541
P1GQ1	5.63E+06	165	3.41E+04	2.07E+05	68.0	99	96	2.4	b	<i>Solibacter usitatus</i>	533
P2GA1	3.37E+06	50	6.74E+04	3.13E+05	54.0	105	101	7.0	a	<i>Acidobacterium</i> sp.	523
										<i>Methylibium petroleiphilum</i>	513
P2GC1	3.94E+06	56	7.03E+04	3.34E+05	69.1	92	90	4.8	a	<i>Leptothrix cholodni</i>	463
										<i>Methylobacterium nodulans</i>	520
P2GD1	4.55E+06	161	2.83E+04	7.58E+05	66.3	109	104	4.2	a	<i>Acidovorax</i> sp.	525
P2GE1	4.12E+06	119	3.46E+04	1.43E+05	65.9	105	104	3.0	a	<i>Alicyclophilus denitrificans</i> BC	490

¹The label indicates the sample from which the assembled scaffolds were used to map the metagenomic reads (see Fig. S1).

Table S7: Continued.

ID	Size (bp)	Number of scaffolds	Mean scaffold length (bp)	Maximum scaffold length (bp)	GC (%)	Number of essential genes	Nr. of unique essential genes	Relative abundance (%)	Assembled scaffolds ¹	Bi-directional best hit classification	Score
P2GI1	3.87E+06	54	7.16E+04	3.21E+05	62.0	76	75	1.3	a	<i>Comamonas sp.</i>	519
P2GJ1	3.54E+06	79	4.48E+04	3.13E+05	54.0	107	103	10.1	b	<i>Solibacter usitatus</i> Ellin6076	531
P2GK1	1.74E+06	521	3.33E+03	2.00E+04	65.1	82	78	1.9	b	<i>Clavibacter michiganensis</i> subsp. <i>sepedonicus</i>	513
P2GL1	4.25E+06	62	6.85E+04	3.26E+05	69.1	108	105	1.8	b	<i>Methylibium petroleiphilum</i> PM1	520
										<i>Leptothrix cholodni</i> SP-6	506
P2GM1	2.70E+06	340	7.96E+03	4.49E+04	33.0	83	82	1.0	b	<i>Ignavibacterium album</i> JCM 16511	549
P2GN1	2.98E+06	415	7.17E+03	6.80E+04	67.7	80	76	1.1	b	<i>Planctomyces limnophilus</i> DSM 3776	534
P2GP1	4.99E+06	78	6.39E+04	2.88E+05	69.9	91	90	2.6	b	<i>Leptothrix cholodni</i> SP-6	505
P2GQ1	6.00E+06	64	9.38E+04	3.90E+05	63.8	70	69	3.8	b	<i>Bradyrhizobium sp.</i> BTAi1	520
WGF1	3.19E+06	378	8.44E+03	7.96E+04	70.0	86	75	1.6	a	<i>Propionibacterium acnes</i>	511
WGG1	2.69E+06	316	8.52E+03	5.78E+04	38.0	85	82	0.7	a	<i>Flavobacteriales bacterium</i>	502
WGG2	2.58E+06	599	4.31E+03	4.21E+04	71.8	80	79	0.7	a	<i>Leptothrix cholodni</i>	513
WGH1	7.28E+06	530	1.37E+04	8.69E+04	56.2	108	104	2.2	a	<i>Chloroflexus sp.</i>	530
WGI1	3.68E+06	235	1.56E+04	1.07E+05	68.3	75	74	1.4	a	<i>Propionibacterium acnes</i>	517
WGL1	3.24E+06	30	1.08E+05	3.71E+05	64.9	92	91	2.4	a	<i>Paracoccus denitrifican</i>	515
WGM1	3.65E+06	24	1.52E+05	4.41E+05	39.1	104	103	1.5	a	<i>Ignavibacterium album</i>	515
WGN1	3.92E+06	328	1.19E+04	1.56E+05	70.6	99	94	5.0	b	<i>Propionibacterium acnes</i>	510
WGO1	4.15E+06	168	2.47E+04	1.60E+05	69.6	106	104	8.9	b	<i>Dinoroseobacter shibae</i>	500
WGP1	3.79E+06	56	6.77E+04	2.50E+05	70.0	104	97	5.4	b	<i>Propionibacterium sp.</i>	508
WGQ1	5.46E+06	75	7.29E+04	2.97E+05	69.7	105	104	8.3	b	<i>Magnetospirillum sp.</i>	516
WGR1	3.76E+06	219	1.72E+04	8.48E+04	68.9	107	104	9.3	b	<i>Propionibacterium acnes</i>	532
WGS1	2.76E+06	51	5.41E+04	2.08E+05	36.8	96	95	7.4	b	<i>Pedobacter sp.</i>	507
										<i>Sphingobacterium spiritivorum</i> ATCC 33300	488
TGC1	4.27E+06	259	1.65E+04	1.06E+05	66.4	103	100	3.1	a	<i>Sphingopyxis alaskensis</i> RB2256	541
TGD1	3.31E+06	125	2.65E+04	1.72E+05	54.0	106	105	3.6	a	<i>Chitinophaga pinensis</i> DSM 2588	536

¹The label indicates the sample from which the assembled scaffolds were used to map the metagenomic reads (see Fig. S1).

Table S7: Continued.

ID	Size (bp)	Number of scaffolds	Mean scaffold length (bp)	Maximum scaffold length (bp)	GC (%)	Number of essential genes	Nr. of unique essential genes	Relative abundance (%)	Assembled scaffolds ¹	Bi-directional best hit classification	Score
TGE1	4.44E+06	393	1.13E+04	7.65E+04	48.8	100	99	5.4	a	<i>Spirosoma linguale</i> DSM 74	508
										<i>Robiginitalea biformata</i> HTCC2501	496
										<i>Chitinophaga pinensis</i> DSM 2588	490
TGG1	2.77E+06	376	7.36E+03	6.27E+04	66.4	87	87	1.4	a	<i>Delftia acidovorans</i> SPH-1	523
										<i>Acidovorax</i> sp. JS42	523
TGH1	3.21E+06	102	3.15E+04	2.07E+05	39.0	100	99	1.4	a	<i>Desulfuromonas</i> sp.	522
TGI1	4.16E+06	299	1.39E+04	7.01E+04	71.1	100	99	4.0	a	<i>Blastopirellula marina</i> DSM 3645	528
										<i>Planctomyces limnophilus</i> DSM 3776	488
TGI2	3.47E+06	139	2.50E+04	1.44E+05	63.6	103	103	3.8	a	<i>Chthoniobacter flavus</i> Ellin428	516
TGJ1	1.76E+06	21	8.40E+04	2.55E+05	40.7	103	102	2.6	a	<i>Geobacter</i> sp. M18	506
										<i>Desulfuromonas</i> sp.	500
										<i>Anaeromyxobacter dehalogenans</i> 2CP-1	478
TGK1	3.63E+06	504	7.21E+03	8.18E+04	59.3	102	101	8.2	b	<i>Opitutus terrae</i> PB90-1	540
										<i>Akkermansia muciniphila</i> ATCC BAA-835	498
										<i>Chthoniobacter flavus</i> Ellin428	490
TGL1	5.15E+06	99	5.20E+04	3.03E+05	50.0	102	101	7.1	b	<i>Pirellula</i> sp. 1	535
										<i>Rhodopirellula baltica</i> SH 1	523
										<i>Blastopirellula marina</i> DSM 3645	513
TGM1	3.63E+06	129	2.81E+04	1.53E+05	64.5	100	100	15.3	b	<i>Chthoniobacter flavus</i> Ellin428	506
										<i>Pedosphaera parvula</i> Ellin514	455
TGO1	3.81E+06	33	1.15E+05	3.67E+05	38.3	101	98	4.5	b	<i>Pedobacter</i> sp.	544
TGP1	3.13E+06	25	1.25E+05	4.14E+05	53.4	104	102	9.8	b	<i>Solibacter usitatus</i> Ellin6076	509
										<i>Acidobacterium capsulatum</i> ATCC 51196	477
TGQ1	2.51E+06	680	3.69E+03	3.31E+04	52.6	88	86	7.1	b	<i>Methylacidiphilum infernorum</i> V4	507
TGR1	3.23E+06	609	5.30E+03	3.44E+04	69.8	69	66	2.5	b	<i>Chloroflexus</i> sp.	529

¹The label indicates the sample from which the assembled scaffolds were used to map the metagenomic reads (see Fig. S1).

Table S7: Continued.

ID	Size (bp)	Number of scaffolds	Mean scaffold length (bp)	Maximum scaffold length (bp)	GC (%)	Number of essential genes	Nr. of unique essential genes	Relative abundance (%)	Assembled scaffolds ¹	Bi-directional best hit classification	Score
S1GB1	3.40E+06	1567	2.17E+03	1.00E+04	69.0	85	84	0.8	a	<i>Acidimicrobium ferrooxidans</i> DSM 10331	537
S1GD1	2.46E+06	416	5.91E+03	6.10E+04	62.2	101	99	0.9	a	<i>Conexibacter woesei</i> DSM 14684	533
S1GE1	9.84E+05	18	5.47E+04	2.39E+05	36.8	83	81	0.5	a	<i>Thermobaculum terrenum</i> ATCC BAA-798	516
										<i>Sphaerobacter thermophilus</i> DSM 20745	513
S1GG1	2.17E+06	30	7.24E+04	1.56E+05	38.1	82	82	1.4	a	<i>Pedobacter saltans</i> DSM 12145	518
S1GH1	4.17E+06	807	5.17E+03	3.02E+04	69.3	98	97	4.2	b	<i>Acidimicrobium ferrooxidans</i> DSM 10331	528
S1GJ1	1.05E+06	115	9.11E+03	5.08E+04	35.3	69	68	0.6	b	<i>Chloroflexus aggregans</i> DSM 9485	510
										<i>Dictyoglomus thermophilum H-6-12</i>	482
S2GB1	2.95E+06	766	3.85E+03	5.44E+04	41.7	80	80	2.0	a	<i>Chitinophaga pinensis</i> DSM 2588	538
S2GE1	2.44E+06	746	3.28E+03	3.44E+04	69.0	71	69	1.1	a	<i>Janibacter sp.</i> HTCC2649	515
S2GF1	3.56E+06	590	6.03E+03	6.87E+04	42.0	85	81	1.4	a	<i>Pedobacter saltans</i> DSM 12145	508
S2GF2	1.57E+06	279	5.62E+03	2.91E+04	62.2	74	70	0.7	a	<i>Acidovorax sp.</i>	528
										<i>Alicyclophilus denitrificans</i> BC	522
S2GG1	2.90E+06	123	2.36E+04	1.40E+05	58.0	103	100	2.2	a	<i>Sphingopyxis alaskensis</i> RB2256	509
										<i>Novosphingobium aromaticivorans</i>	472
S2GI1	3.88E+06	124	3.13E+04	2.39E+05	45.0	107	106	5.9	a	<i>Thioalkalivibrio sp.</i> HL-EbGR7	532
										<i>Thiothrix sp.</i>	460
S2GJ1	2.89E+06	776	3.73E+03	1.81E+04	71.7	104	100	0.7	a	<i>Rhodococcus jostii</i> RHA1	504
S2GJ2	2.87E+06	1236	2.32E+03	1.51E+04	47.2	76	72	0.6	a	<i>Pedobacter heparinus</i> DSM 2366	523
S2GJ3	2.43E+06	717	3.39E+03	2.84E+04	58.5	106	85	0.5	a	<i>Kribbella flavida</i> DSM 17836	536

¹The label indicates the sample from which the assembled scaffolds were used to map the metagenomic reads (see Fig. S1).

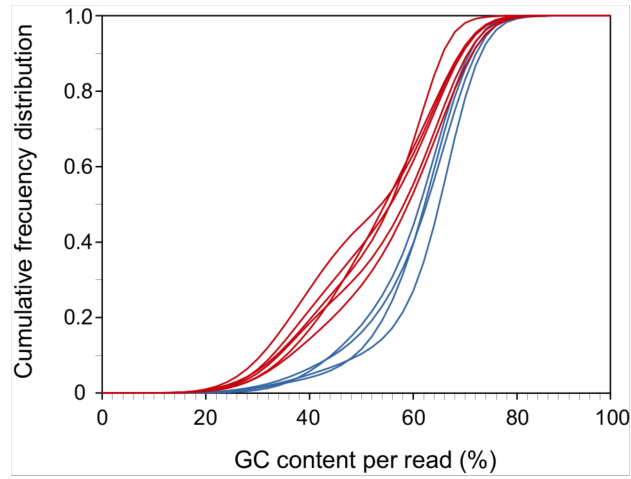


Figure S1. Empirical cumulative frequency distribution for GC content per sequence (%) of metagenomic data sets from petrochemical (blue) and municipal (red) WWTPs.

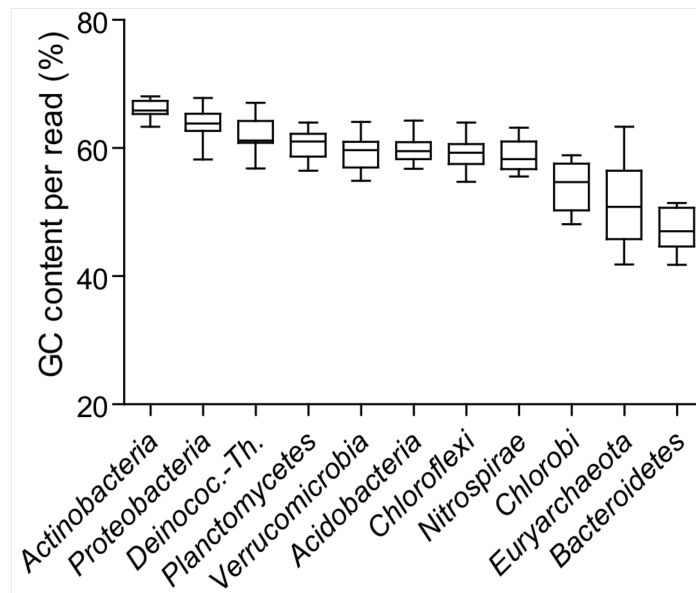


Figure S2. The GC-content was calculated for the reads belonging to the 11 highly-represented phyla and sorted by a decreasing median.

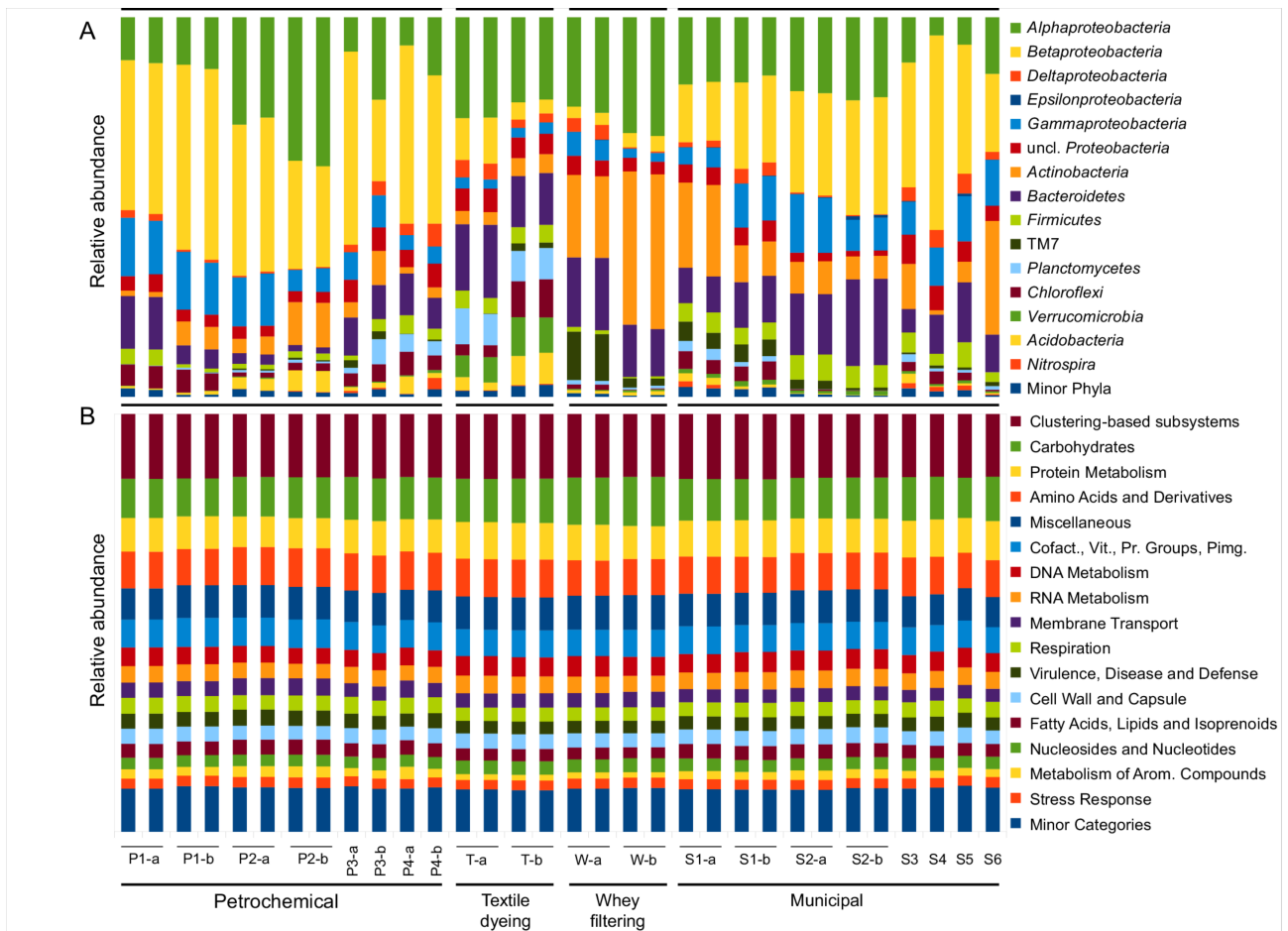


Figure S3. Variation of taxonomic and functional profiles across samples, analyzed using bacterial phyla (with *Proteobacteria* divided in classes) and SEED subsystems (level 1), respectively.

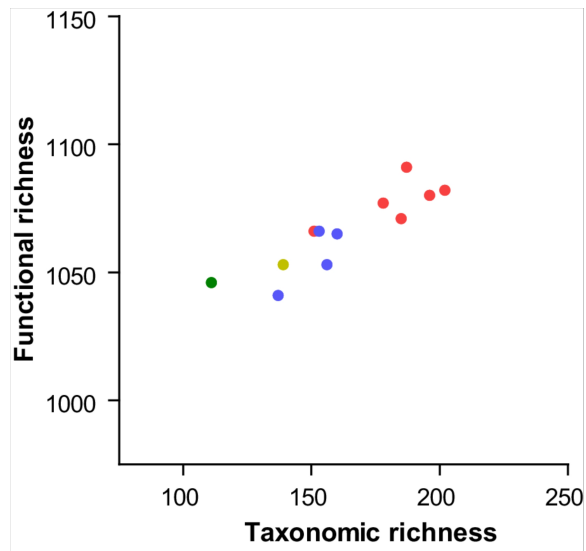


Figure S4. Relationship between the observed functional and taxonomic richness of 12 independent WWTP communities. Functional diversity is represented by the number of SEED subsystems (level 3). Taxonomic diversity was estimated from reads corresponding to 16S rRNA gene fragments classified into genera. Technical replicates and data from the two different sampling times were averaged for each WWTP.

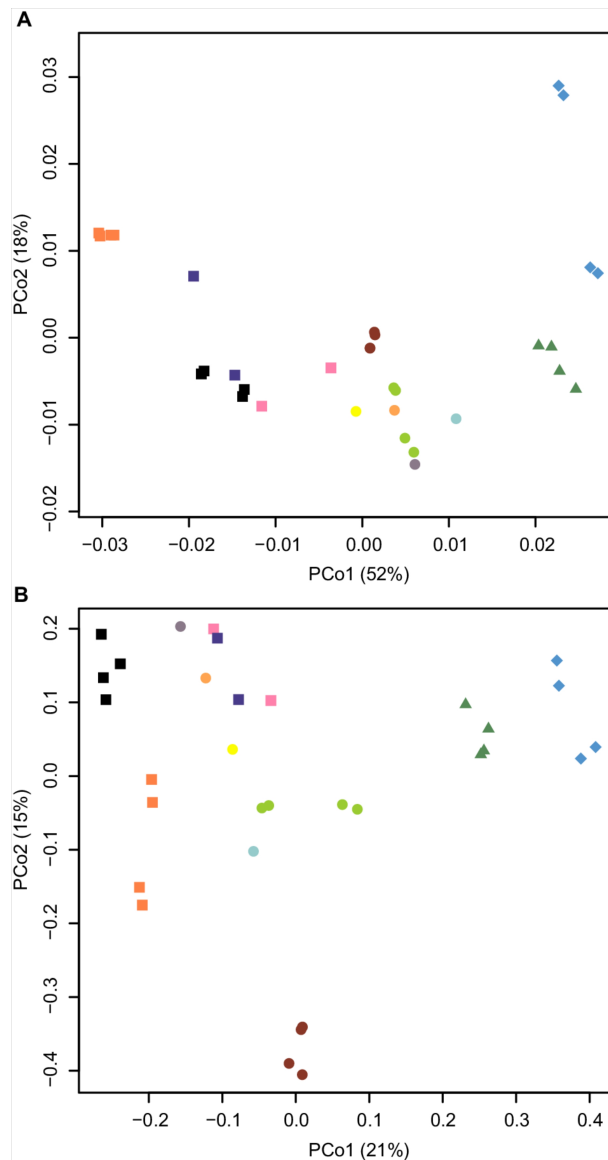


Figure S5. Principal Coordinates Analysis based on Bray-Curtis dissimilarity matrix performed for (A) SEED subsystems at level 1 and (B) bacterial genera identified from 16S rRNA shotgun fragments in municipal WWTP (circles), petrochemical WWTP (squares), textile dyeing WWTP (triangles) and whey processing WWTP (diamonds). Colors differentiate WWTPs. Technical replicates and data from the two different sampling times were included separately.

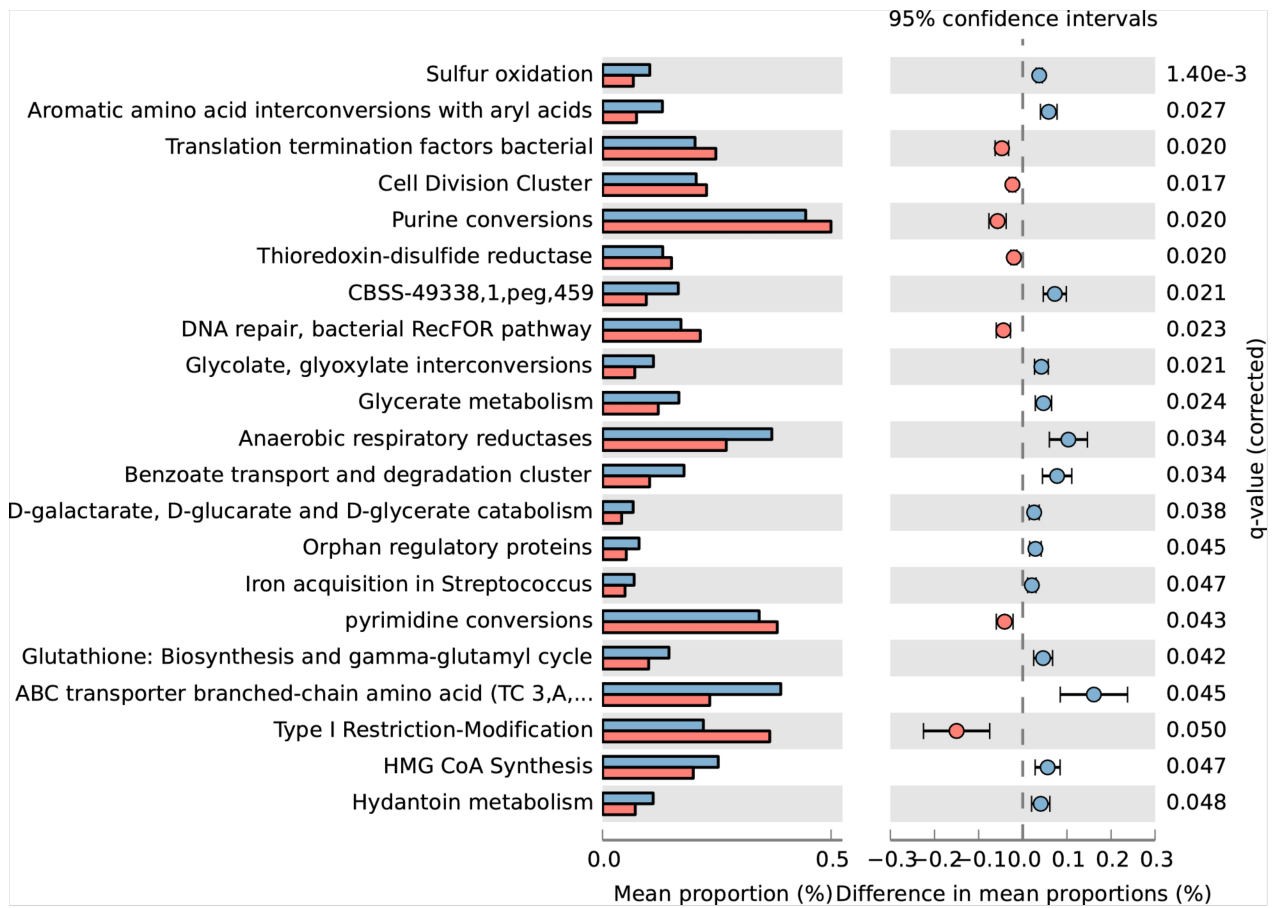


Figure S6. Functional traits comparison between the metagenomes of petrochemical (blue bars, $n = 4$) and municipal (red bars, $n = 6$) activated sludge at level 3 of the SEED hierarchy. Technical replicates and data from the two different sampling times were averaged for each WWTP, and only traits with abundance $> 0.01\%$ were considered. A t-test, with Benjamini-Hochberg FDR multiple test correction, revealed features with significant differences ($p < 0.05$, effect size > 0.02). Only those traits that showed significant differences are shown.

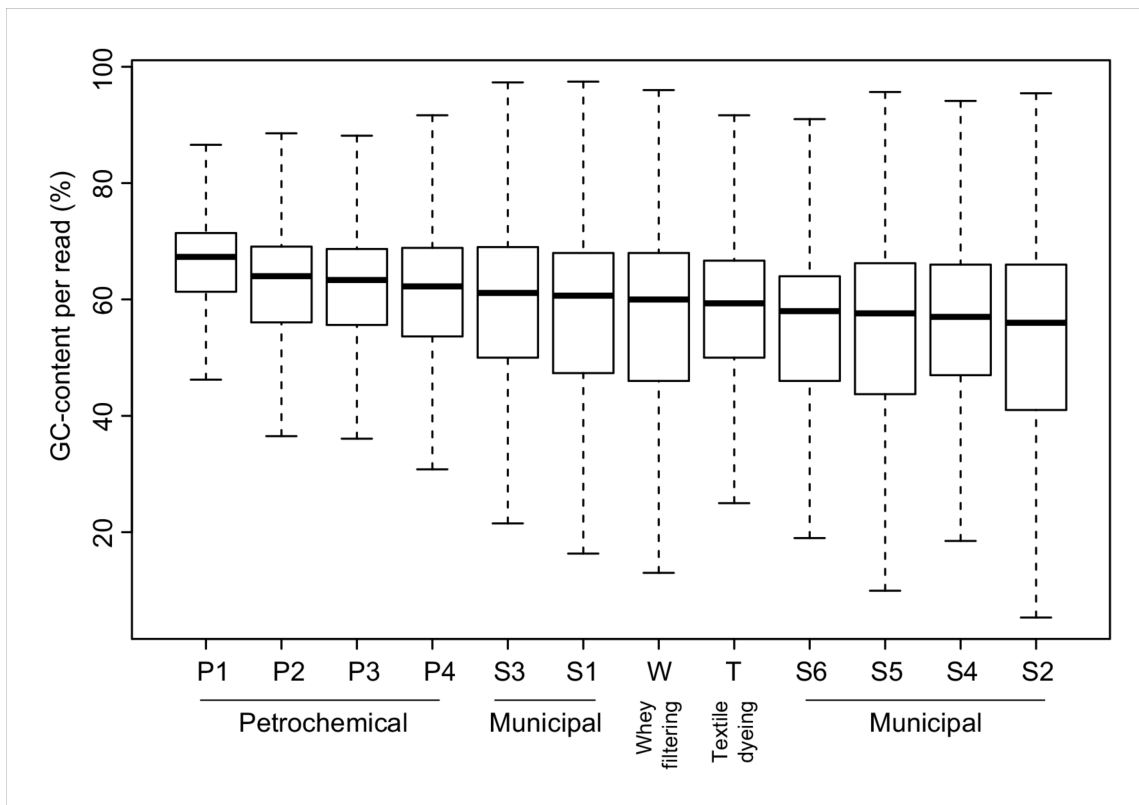


Figure S7. The GC-content was calculated for the reads belonging to each WWTP and sorted by a decreasing median.