Supporting Information

For

Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment

Yuanqing Chao,[1] Liping Ma,[1] Ying Yang,[1] Feng Ju,[1] Xu-Xiang Zhang,[1,2] Wei-Min Wu,[3] Tong Zhang[*,1]

[1] Environmental Biotechnology Lab, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China

[2] State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing 210046, China

[3] Department of Civil and Environmental Engineering, Stanford University, Stanford, California 94305, United States

Table and Figure Legends:

Table S1 Diversity evaluation (Chao and Shannon) of RW and TW samples at 0.03 and 0.06 cluster distances by RDP pipelines.

Table S2 The basic information of 42 metagenomic data in four ecosystems (including soil, human faeces, activated sludge, and ocean), which were applied to principal component analysis in the present study.

[*] Corresponding author. Email: zhangt@hku.hk. Phone: (852) 28598551. Fax: (852) 25595337.

Table S3 Primers used for *gshA* and *gshB* amplification by qRT-PCR**.**

Figure S1 Quality scores across all bases in Illumina reads of 5 metagenomes tested by using FastQC pipelines.

Figure S2 Repeatability evaluation of two technical duplicates (RW12-1 and RW12-2) by Scatter plot. G-test was performed by using STAMP and corrected *q*-values (*p*-values) were calculated by using Benjamini-Hochberg's FDR approach.

Figure S3 Percentage of microbial community in RW and TW at domain level classified by using all the annotation source databases on MG-RAST. The reads number which could be annotated by MG-RAST was taken as 100%.

Figure S4 Percentage of bacterial rRNA reads in RW and TW at phylum level classified by using SILVA SSU database. The reads number which annotated to phylum level was taken as 100%. The phyla, which accounted for more than 1% of total rRNA reads in either RW or TW, were shown in the figure.

Figure S5 Percentage of annotated reads in the major Level 1 subsystems (analyzed by SEED Subsystems) in RW and TW. The reads number which annotated as Level 1 subsystems was taken as 100%. The asterisks showed the significant differences between RW and TW (one asterisk: $P < 0.05$; two asterisks: $P < 0.01$).

Figure S6 Verification of metagenomic data by quantitative real-time PT-PCR analysis. Two glutathione systhesis genes, i.e. *gshA* (glutamate-cysteine ligase, EC 6.3.2.2) and *gshB* (glutathione synthase, EC 6.3.2.3), were selected for PCR amplication. RW and TW samples in 6 months were applied. The PCR products were checked by gel electrophoresis (Figure S7). The asterisks showed the significant differences between RW and TW (one asterisk: $P < 0.05$; two asterisks: $P < 0.01$).

Figure S7 Agarose gel electrophoresis of qRT-PCR products of *gshA* and *gshB* genes in RW and TW.

Figure S8 The treatment processes of the drinking water treatment plant located at Pearl River Delta area.

Figure S9 Bright field image (A) shows a fouling layer accumulated on the filter surface containing microorganisms (B) collected from DW samples. After ultrasonication, the majority of fouling layer (C) and microorganisms (D) on the filter had been effectively detached. The microorganisms on the filter (B & D) were stained by SYTO9 and visualized by CLSM using a 63× objective.

Figure S10 The procedures of creating a sub-database derived from ARDB for sorting.

Table S1 Diversity evaluation (Chao and Shannon) of 16S rRNA genes[a] in RW and TW samples at 0.03 and 0.06 cluster distances by RDP pipelines.

Samples	0.03		0.06	
	Chao	Shannon	Chao	Shannon
RW	2,306 ± 100	7.4 ± 0.021	2,226 ± 98	7.3 ± 0.024
TW	1,465 ± 257	6.0 ± 0.18	1,378 ± 250	6.0 ± 0.20
P value	0.012	0.001	0.011	0.001

[a] the 16S rRNA genes were extracted from the metagenomes according to the blast results against the SILVA SSU database;

Table S2 The basic information of 42 metagenomic data in four ecosystems (including soil, human faeces, activated sludge, and ocean), which were applied to principal component analysis in the present study.

Biomes	MG-RAST ID	bp No.	Reads No.	Location	Sequencing Method
Soil_1	4441091.3	154,475,569	138,347	Farm in Waseca County, Minnesota, United States	sanger
Soil_2	4445990.3	219,117,356	583,724	Mercury, Navada, United States	454
Soil_3	4445994.3	254,548,462	683,082	Mercury, Navada, United States	454
Soil_4	4445993.3	133,555,260	352,417	Mercury, Navada, United States	454
Soil_5	4445996.3	116,821,792	312,444	Mercury, Navada, United States	454
Soil_6	4446153.3	322,213,082	782,404	subtropical lower montane wet forest in the Luquillo forest, Puerto Rico	454
Soil_7	4443231.3	440,390,760	1,103,048	Eureka, Canada	454
Soil_8	4443232.3	413,050,687	1,024,347	Eureka, Canada	454
Soil_9	4450750.3	87,160,647	239,933	Navada, United States	454
Soil_10	4450752.3	76,860,743	233,279	Navada, United States	454
Soil_11	4451103.3	397,257,248	1,040,697	Navada, United States	454
Soil_12	4451104.3	347,578,191	998,484	Navada, United States	454
Human Faeces_1	4440943.3	40,076,128	30,198	Japan	other[a]
Human Faeces_2	4440949.3	25,941,797	16,164	Japan	other

Human Faeces_3	4440944.3	39,071,077	31,237	Japan	other
Human Faeces_4	4440946.3	29,296,224	20,226	Japan	other
Human Faeces_5	4440941.3	43,259,070	36,326	Japan	other
Human Faeces_6	4440947.3	45,480,292	35,177	Japan	other
Human Faeces_7	4440948.3	46,397,089	37,296	Japan	other
Human Faeces_8	4440951.3	43,473,860	34,797	Japan	other
Human Faeces_9	4440942.3	45,906,118	36,455	Japan	other
Human Faeces_10	4440950.3	27,208,886	20,532	Japan	other
Activated sludge_1	4489206.3	2,561,651,200	25,616,512	Shatin wastewater treatment plant, Hong Kong	Illumina
Activated sludge_2	4489205.3	2,561,651,200	25,616,512	Shatin wastewater treatment plant, Hong Kong	Illumina
Activated sludge_3	4489073.3	2,561,357,900	25,613,579	Shatin wastewater treatment plant, Hong Kong	Illumina
Activated sludge_4	4489072.3	2,542,600,100	25,426,001	Shatin wastewater treatment plant, Hong Kong	Illumina
Activated sludge_5	4487666.3	2,564,795,000	25,647,950	Shatin wastewater treatment plant, Hong Kong	Illumina
Activated sludge_6	4487665.3	2,564,795,000	25,647,950	Shatin wastewater treatment plant, Hong Kong	Illumina
Activated sludge_7	4489245.3	2,560,000,000	25,600,000	Shatin wastewater treatment plant, Hong Kong	Illumina
Activated sludge_8	4487697.3	2,563,286,700	25,632,867	Shatin wastewater treatment plant, Hong Kong	Illumina
Activated sludge_9	4487700.3	2,563,286,700	25,632,867	Shatin wastewater treatment plant, Hong Kong	Illumina
Activated sludge_10	4488258.3	2,560,000,000	25,600,000	Shatin wastewater treatment plant, Hong Kong	Illumina
Ocean_1	4441148.3	54,752,102	50,609	Indian Ocean, St. Anne Island, Seychelles	other

Ocean_2	4441139.4	52,667,848	50,096	Indian Ocean, International waters between Madagascar and South Africa	other
Ocean_3	4441150.3	64,230,062	61,020	Indian Ocean	other
Ocean_4	4441155.3	62,752,349	59,813	Indian Ocean	other
Ocean_5	4441608.3	53,607,277	49,597	Indian Ocean	other
Ocean_6	4441135.3	45,710,196	46,052	Indian Ocean, Madagascar Waters, Madagascar	other
Ocean_7	4441149.3	64,223,447	60,932	Outside Seychelles, Indian Ocean, Seychelles	other
Ocean_8	4441147.3	55,638,894	52,118	Indian Ocean	other
Ocean_9	4441156.3	62,072,289	59,080	Indian Ocean	other
Ocean_10	4441134.3	53,607,277	49,597	Indian Ocean	other

[a] indicates the researchers applied other sequencing method, except for sanger, Illumina and 454 methods.

Table S3 Primers used for *gshA* and *gshB* amplification by qRT-PCR**.**

Taget genes	Primers[a]	Sequences (5'-3')
gshA	gshA_F	GGCGGCGAAGCGTATCAGAAA
	gshA_R	AATGCTTTGCCTGTTCCGCCA
gshB	gshB_F	CGTGATTGCCGAAACCCTGA
	gshB_R	GCCAGATTGCCACGGGTTTC

[a] **Reference:** Helbig, K., Grosse, C., & Nies, D. H. Cadmium toxicity in glutathione mutants of *Escherichia coli*. *J. Bacteriol.* **190**, 5439-5454 (2008).

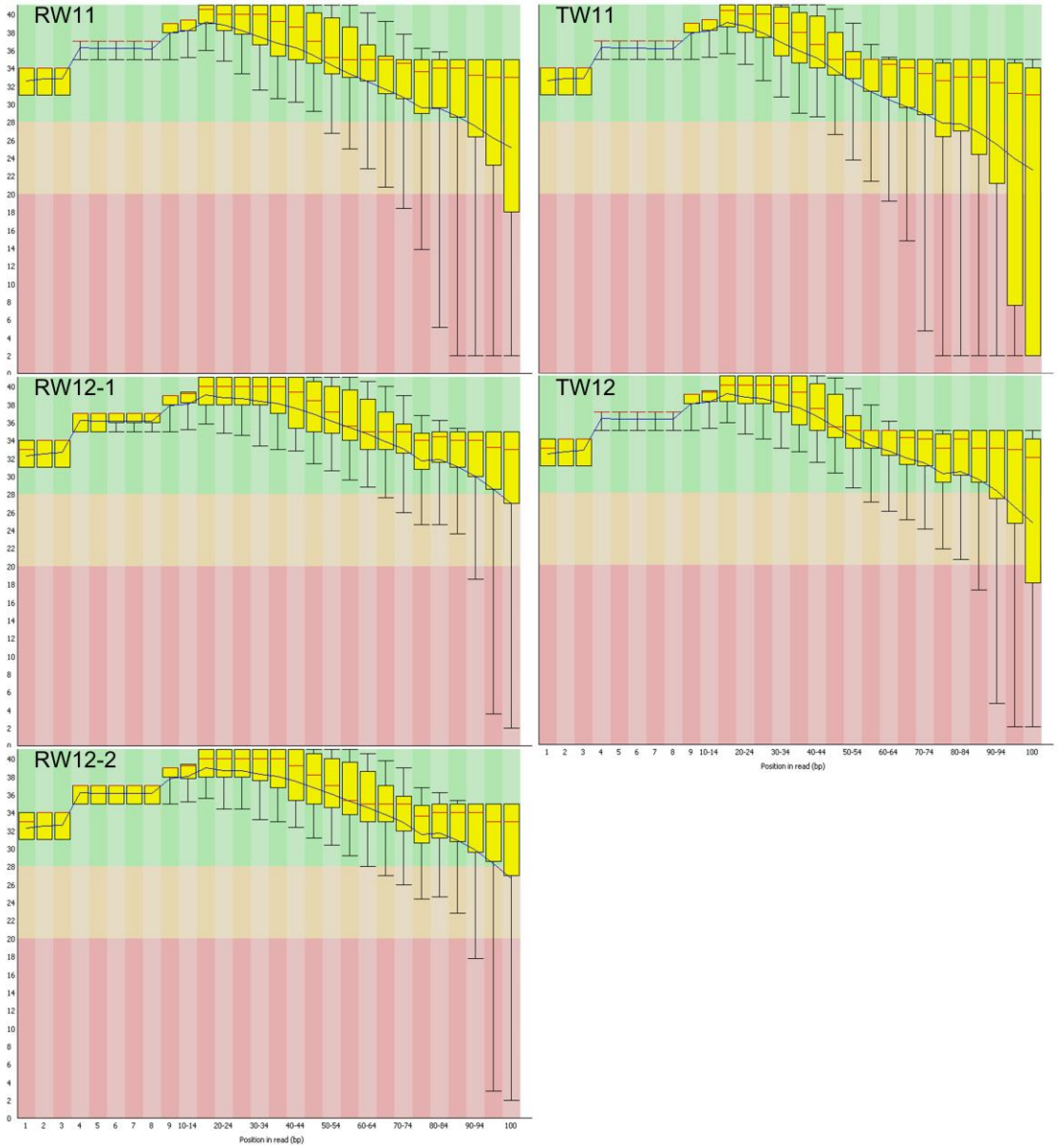


Figure S1 Quality scores across all bases in Illumina reads of 5 metagenomes tested by using FastQC pipelines.

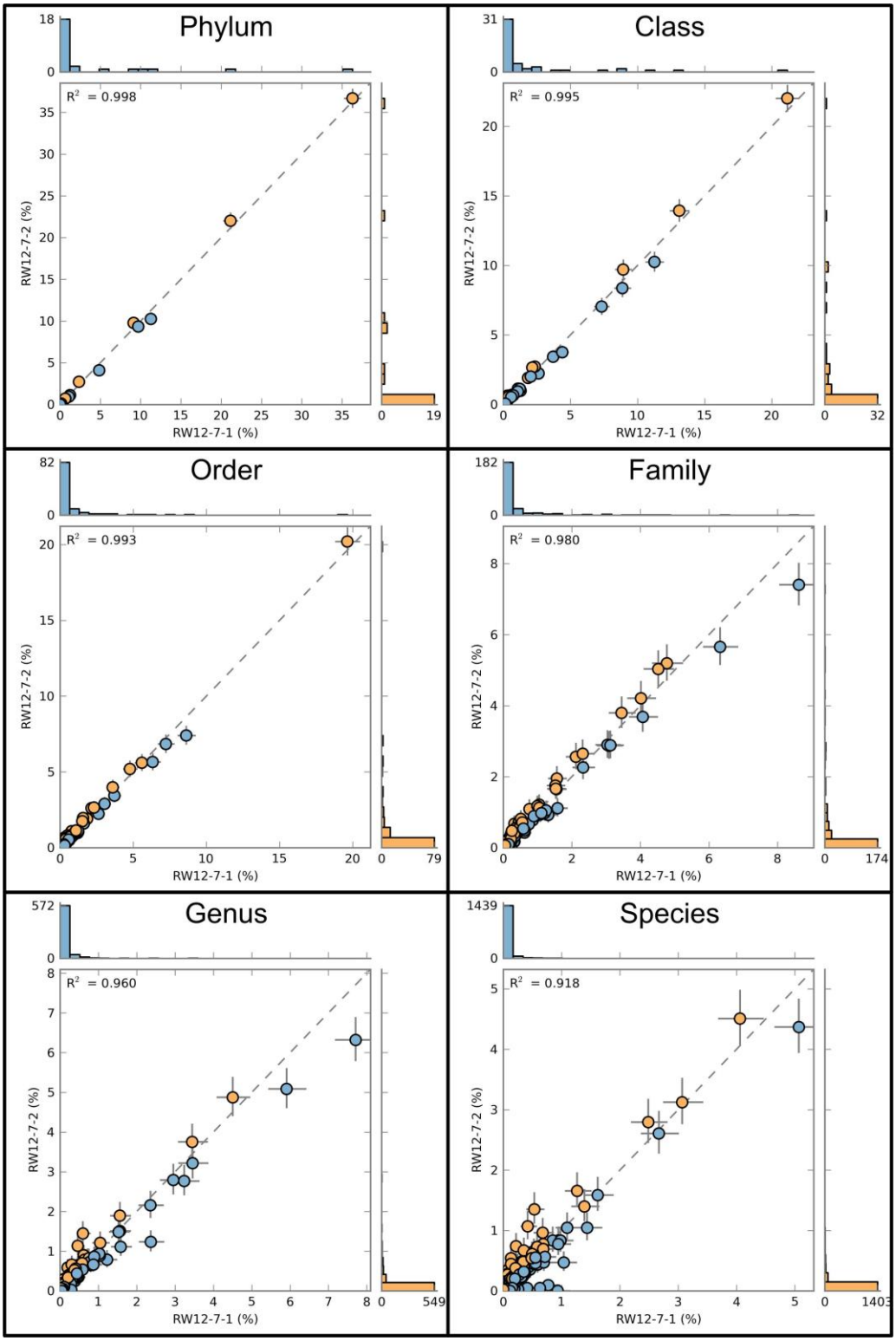


Figure S2 Repeatability evaluation of two technical duplicates (RW12-1 and RW12-2) by Scatter plot. G-test was performed by using STAMP and corrected *q*-values (*p*-values) were calculated by using Benjamini-Hochberg's FDR approach.

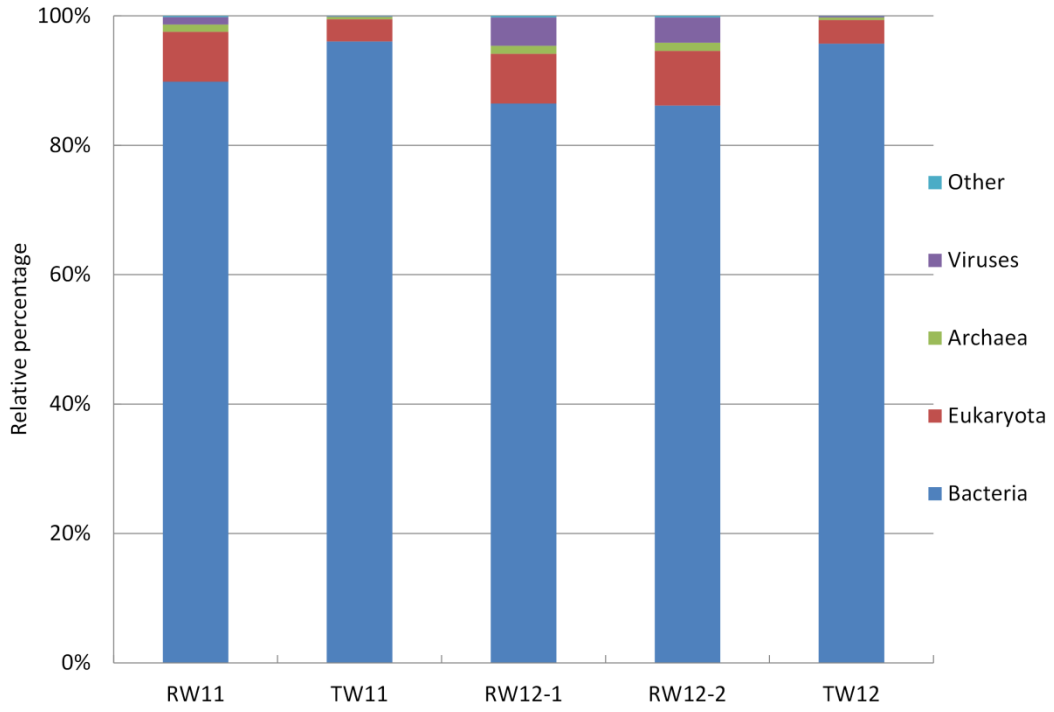


Figure S3 Percentage of microbial community in RW and TW at domain level classified by using all the annotation source databases on MG-RAST. The reads number which could be annotated by MG-RAST was taken as 100%.

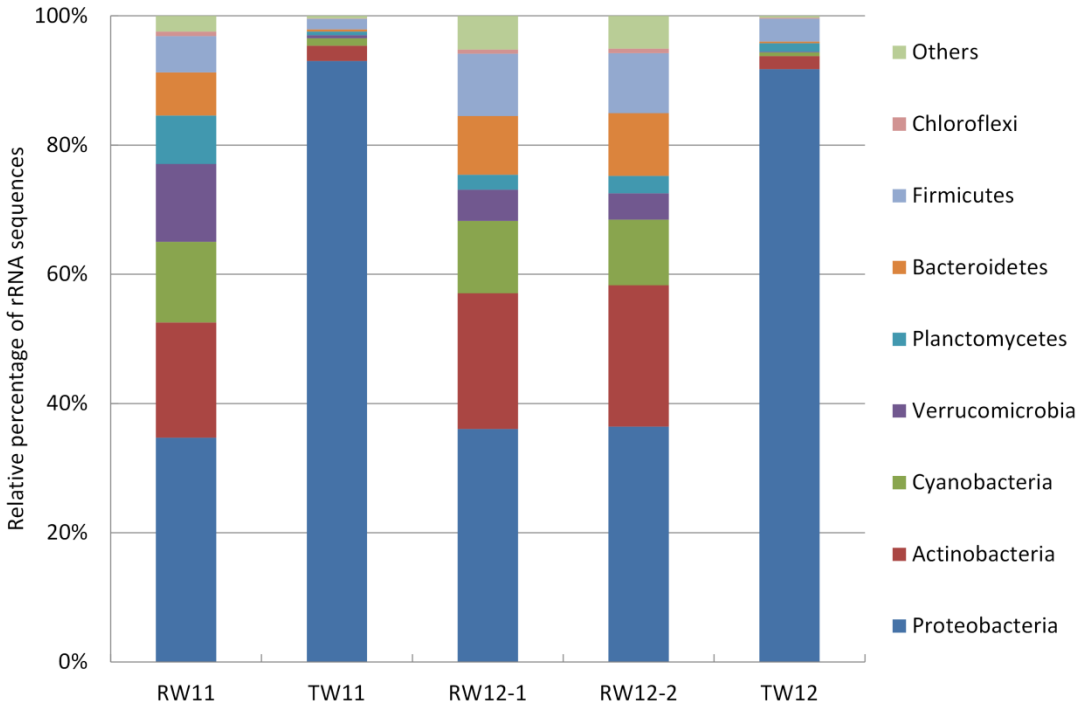


Figure S4 Percentage of bacterial rRNA reads in RW and TW at phylum level classified by using SILVA SSU database. The reads number which annotated to phylum level was taken as 100%. The phyla, which accounted for more than 1% of total rRNA reads in either RW or TW, were shown in the figure.

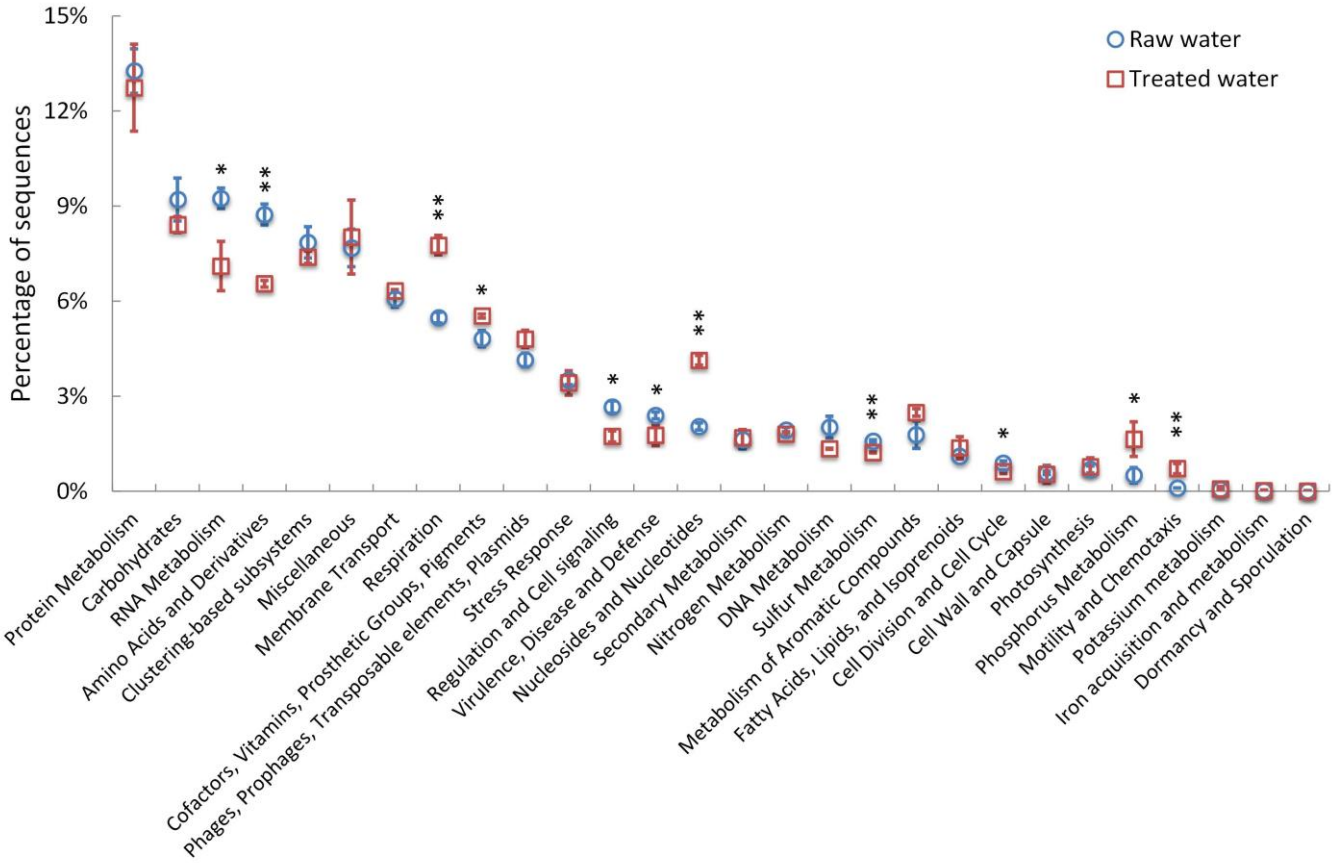


Figure S5 Percentage of annotated reads in the major Level 1 subsystems (analyzed by SEED Subsystems) in RW and TW. The reads number which annotated as Level 1 subsystems was taken as 100%. The asterisks showed the significant differences between RW and TW (one asterisk: $P < 0.05$; two asterisks: $P < 0.01$).

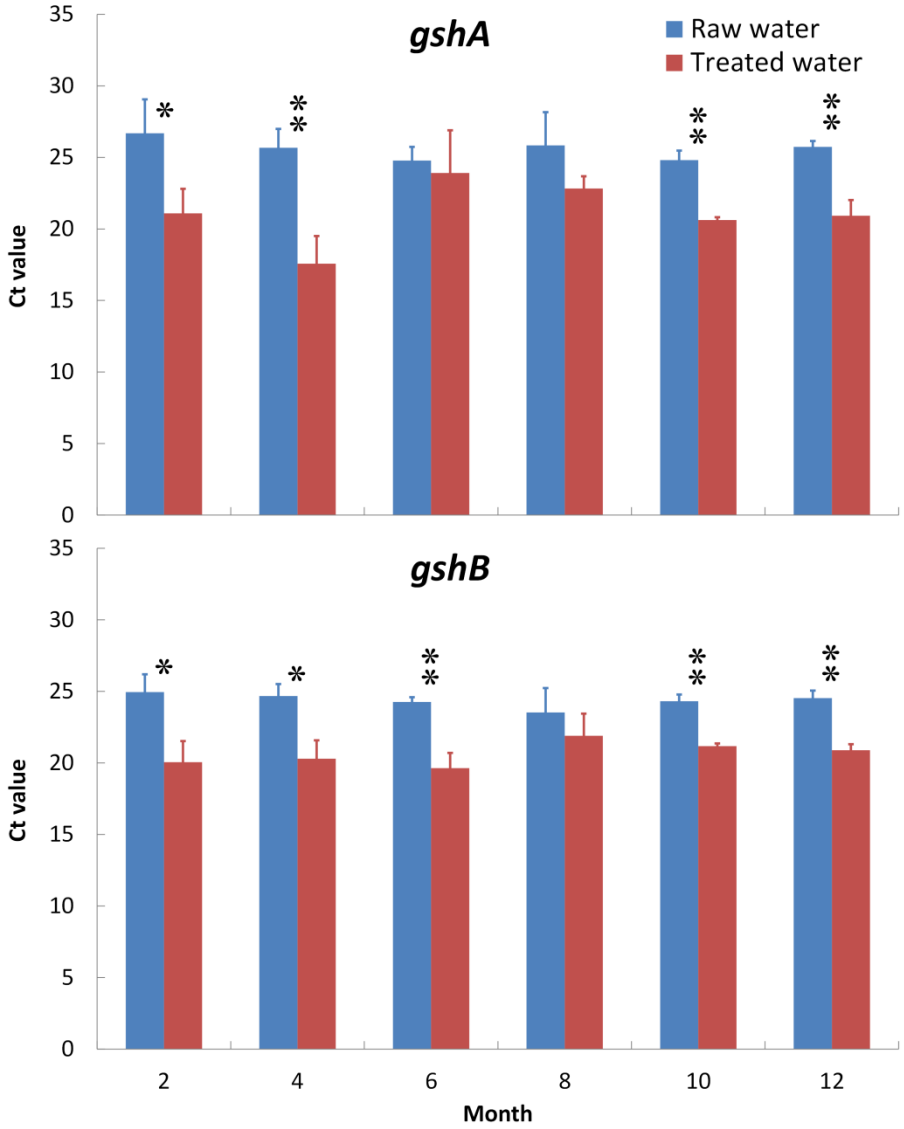
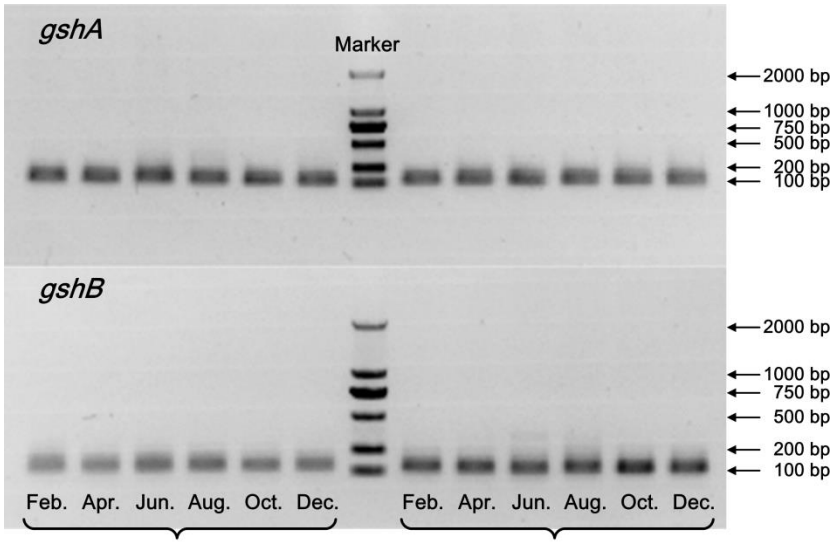


Figure S6 Verification of metagenomic data by quantitative real-time PT-PCR analysis. Two glutathione systhesis genes, i.e. *gshA* (glutamate-cysteine ligase, EC 6.3.2.2) and *gshB* (glutathione synthase, EC 6.3.2.3), were selected for PCR amplication. RW and TW samples in 6 months were applied. The PCR products were checked by gel electrophoresis (Figure S7). The asterisks showed the significant differences between RW and TW (one asterisk: $P < 0.05$; two asterisks: $P < 0.01$).

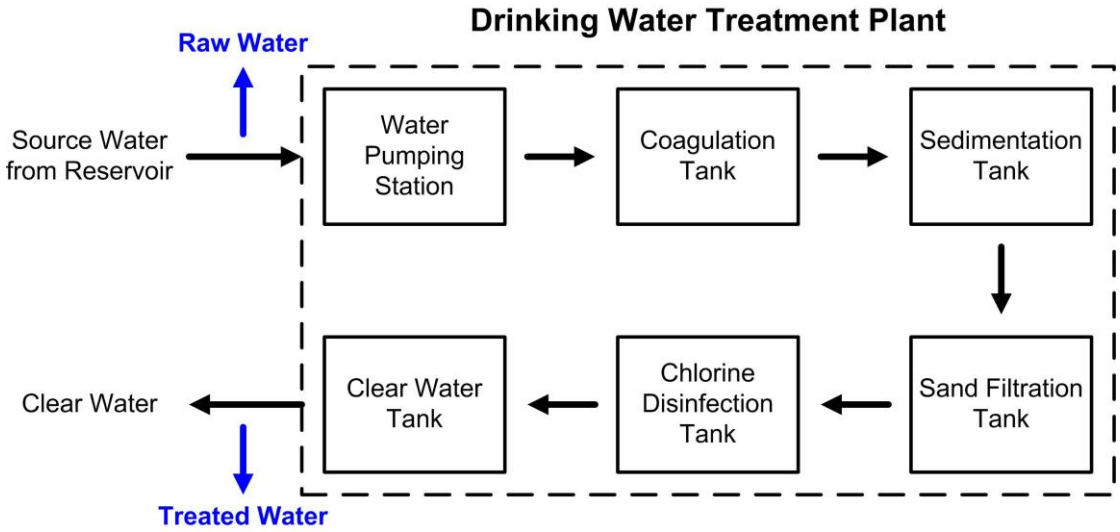


111

112 **Figure S7** Agarose gel electrophoresis of qRT-PCR products of *gshA* and *gshB* genes

113 in RW and TW.

114



115

116 **Figure S8** The treatment processes of the drinking water treatment plant located at

117 Pearl River Delta area.

118

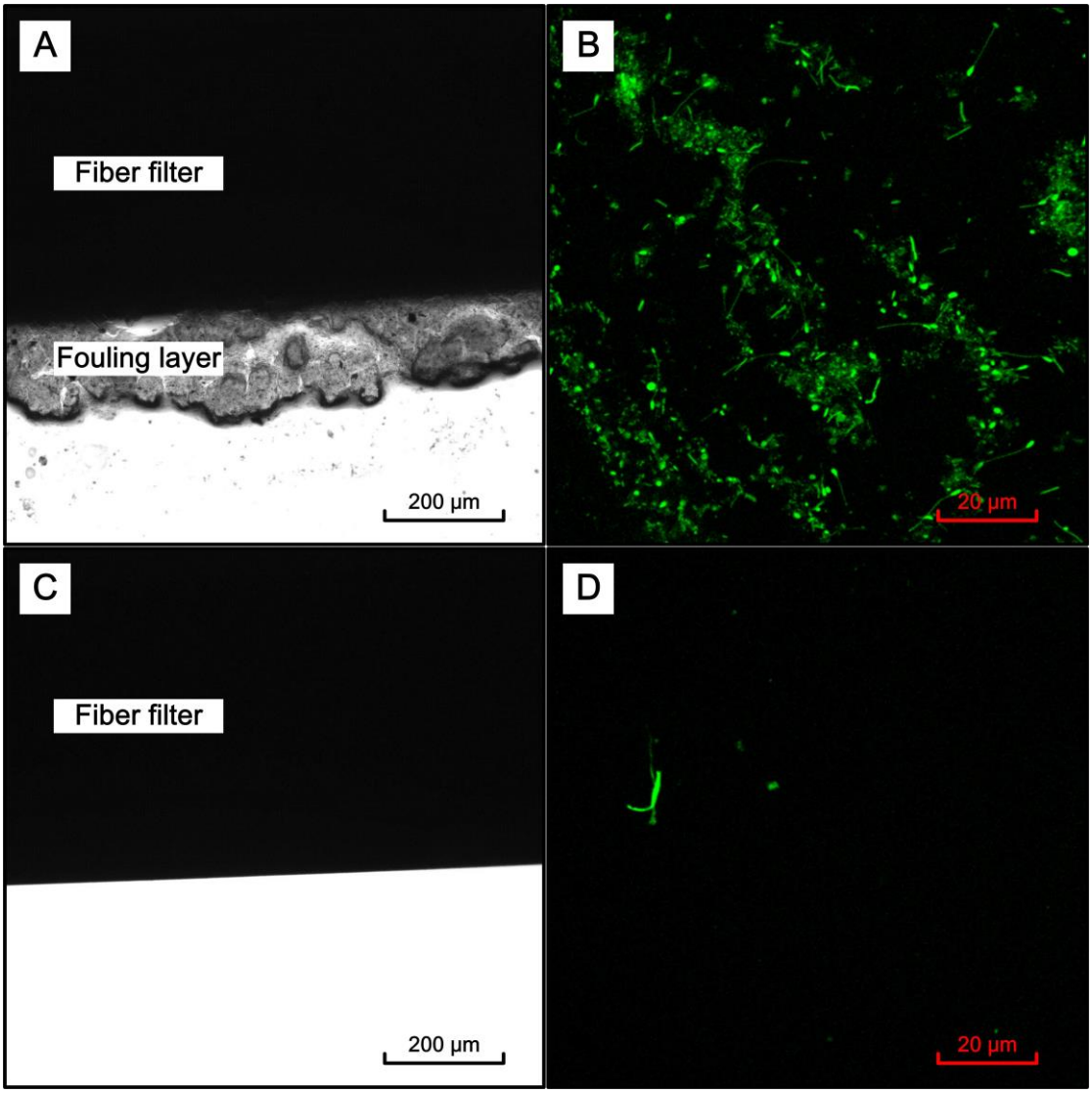


Figure S9 Bright field image (A) shows a fouling layer accumulated on the filter surface containing microorganisms (B) collected from DW samples. After ultrasonication, the majority of fouling layer (C) and microorganisms (D) on the filter had been effectively detached. The microorganisms on the filter (B & D) were stained by SYTO9 and visualized by CLSM using a 63× objective.

Figure S10 The procedures of creating a sub-database derived from ARDB for sorting.