Tropical Soil Metagenome Library Reveals Complex Microbial Assemblage

Kok-Gan Chan [1#] and Zahidah Ismail [1]

[1] Division of Genetics and Molecular Biology, Institute of Biological Sciences, Faculty of Science,

University of Malaya, Kuala Lumpur, Malaysia

\#   Author to whom correspondence should be addressed; E-Mail: kokgan@um.edu.my;

Tel.: +603-7967-5162; Fax: +603-7967-4509.

1

2 **Abstract**

3   In this work, we characterized the metagenome of a Malaysian mangrove soil sample via next

4   generation sequencing (NGS). Shotgun NGS data analysis revealed high diversity of microbes

5   from Bacteria and Archaea domains. The metabolic potential of the metagenome was

6   reconstructed using the NGS data and the SEED classification in MEGAN shows abundance of

7   virulence factor genes, implying that mangrove soil is potential reservoirs of pathogens.

8

9   Keywords: Metagenomics, Mangrove, Soil, Proteobacteria

10

Introduction

12   Mangrove forests are usually located at the tropical and subtropical latitudes. They are present at

13   the transition of land and sea which makes them susceptible to tidal change and salinity.

14   Mangrove soils generally comprise soft, muddy and anaerobic sediment with thin top layer of

15   aerobic sediment. They also function as heavy metal sink[1] which acts as natural sink and

16   filtration system. Mangrove swamps are the habitat for a diverse variety of fauna especially

17   juvenile fishes, and they also act as breeding and nursery grounds for these aquatic animals.

18   Microorganisms in the mangrove habitat play an important role in maintaining the productivity,

19   conservation and nutrients of this ecosystem. Microorganisms are involved in biogeochemical

20   cycles that supply nutrients to plants and animals[2,3]. Mangroves are rich in organic matters but

21   usually lack phosphorus and nitrogen[4-7]. Their microorganisms' activities are high because they

22   are very efficient in recycling the nutrients contained therein. Microorganisms are directly

23   involved in nitrogen fixation, phosphate solubility, photosynthesis, sulfate reduction and

24   production of other substances.

25   The mangrove environment is highly susceptible to anthropogenic effects such as pollution,

26   deforestation and human activity. These could change the dynamic mangrove ecosystem which in

27   turn affects the mangrove community and disturbs the microorganism community that maintains

28   the productivity and conservation of mangrove.

29   This study aimed to investigate the metagenome of a mangrove soil sample and their ecological

30   role via metabolic reconstruction. We used the Illumina HiSeq 2000 platform to carry out

31   shotgun metagenome next generation sequencing (NGS). This method avoided bias of PCR

32   amplification as in the case of amplicon sequencing and it enabled parallel study on both the

33 taxonomic and functional diversities. We hypothesized that the abundance and diversity of

34 microbes and their functional attributes to be similar to those of previous studies[8,9] .

Materials and Methods

36 Sampling was done on a soil sample obtained in the east coast of Peninsular Malaysia, namely

37 Rantau Abang (RA) (N04^o 54.189' E103^o 22.208'). No specific permissions were required for the

38 chosen locations and such research activities. Our work also did not involve endangered or

39 protected species. The top 5 to 20 cm of soil was collected and stored at -20°C until processing. A

40 portion of the soil sample was sent for biochemical analyses of its pH, carbon nitrogen ratio, and

41 contents of phosphorus, sulfur and heavy metals, like arsenic, cadmium, lead and mercury, as

42 described previously[10].

43 DNA extraction was carried out according to the protocol as described previously[11] with

44 modifications. Traces of plant materials were removed from the soil prior to extraction. Briefly,

45 5g of soil was added with 13.5ml of DNA extraction buffer (Tris-HCl, pH8 100mM; EDTA, pH8

46 100mM; Na_2HPO_4, pH7.8 100mM; NaCl, 1.5M; CTAB, 1% w/v), 100μl of proteinase K

47 (10mg/μl), and 200μl of lysozyme (10mg/μl). The mixture was incubated horizontally at 37°C

48 with orbital shaking (225rpm). After 30min, 0.5ml of SDS (20% w/v) was added and the mixture

49 further incubated in a 65°C water bath for 2h with gentle mixing by inverting the tube at 15min

50 intervals. The supernatant was collected by centrifugation at $6000 \times g$ for 10min. The pellet was

51 suspended in 4.5ml of DNA extraction buffer and 0.5ml of SDS (20% w/v) and vortexed for 10s

52 followed by incubation at 65°C for 10min. The supernatant was then collected by centrifugation

53 and pooled with the supernatant collected previously. Equal volume of chloroform:isoamyl

54 alcohol (24:1, vol/vol) was added to the pooled supernatant and the mixture was gently mixed by

55 inversion. The aqueous phase was transferred to a clean, sterile tube after centrifugation at 6000 ×

56 g for 10min. The chloroform:isoamyl alcohol step was repeated once. For DNA precipitation, 0.6

57 volume of cold isopropanol was added and the resultant mixture incubated at -20°C for 30min.

58 DNA was collected by centrifugation at 16000 × g for 20min, followed by washing with 70%

59 (v/v) ethanol and kept in -20°C for 15min. Ethanol was removed by centrifugation at top speed in

60 a table top centrifuge for 10min and the pellet was air dried aseptically. The DNA pellet was then

61 dissolved in elution buffer (Roche High Pure PCR Product Purification Kit).

62 The soil metagenomic DNA was further purified by gel elution in a 3% (w/v) low melting

63 temperature agarose electrophoresis. The metagenomic DNA was mixed with 80% (v/v) glycerol

64 and 6× loading dye and the mixture was then loaded into a well. Electrophoresis was carried out

65 at 15V for 16 to 20h. DNA was excised from the gel with a sterile blade and recovered using the

66 Qiagen Gel Extraction Kit (Venlo, Netherlands). DNA concentration and purity were determined

67 using Qubit and Nanodrop 2000c by Life Technologies, respectively. The purified DNA was then

68 subject to NGS using Illumina HiSeq 2000.

69 For taxonomic analysis, the metagenomic nucleotide sequences obtained were trimmed using

70 CLC Bio Genomic Workbench 5.5.2 (Aarhus, Denmark) at 50-nucleotide length to remove short

71 low quality reads. The trimmed data were then blasted against the NCBI Microbial database

72 (dated 22 Jan 2013) using Blastall 2.2.25 (NCBI) at the expected value of 1×10^{-20}.

73 For functional gene study, the trimmed nucleotide sequences were assembled using *de novo*

74 assembly in CLC Bio Genomic Workbench at the minimum contig length of 400 nucleotides.

75 The assembled data were extracted at coverage of ≥10%. Gene prediction was performed on the

76 extracted sequences using Prodigal 2.60[12], and each predicted gene was annotated using

77 RAPsearch 2.09[13,14] (against the NCBI NR database (dated 22 January 2013). Data obtained for

78   both taxonomic and functional distributions were analyzed in MEGAN 4.70.4[15]. Taxonomic

79   analysis was done according to the percentage identity filter to get the best sequence match. In

80   MEGAN, functional analysis was accomplished with the SEED[16] classification.

81   The data of this study are available as NCBI database accession number SRR748204.

82   **Results**

83   Biochemical Analyses

84   Table 1 shows the biochemical properties of the Rantau Abang (RA) soil sample. The pH for RA

85   sample was recorded as pH of 5.1.

86   Metagenomic Library Analysis

87   The RA sample metagenome library shows very high nucleotides sequenced after editing and

88   >500,000 contigs generated at the coverage at approximately 41 times (Table 2).

89   Microbial Taxonomic Distribution

90   A total of 98% of the reads from RA sample was assigned to the domains level by MEGAN and

91   they excluded the "No hits" reads category. The majority of the assigned reads from RA samples

92   was of the domain Bacteria while the remaining were of the domain Archaea with 78.52% reads

93   assigned to Bacteria and 21.48% to Archaea (Figure 1).

94   There were 27 phyla hits from the domain Bacteria for RA metagenome library. The phylum

95   Proteobacteria dominated other phyla in (43.72%) in the RA sample. There were 10 phyla with

96   abundance percentages of more than 1% in the RA sample. In this sample, the phyla detected

97   were Proteobacteria, Acidobacteria (17.68%), Firmicutes (13.45), Actinobacteria (4.55%),

6

98   Nitrospirae (4.22%), Planctomycetes (3.06%), Chloroflexi (2.88%), Verrucomicrobia (2.69%),

99   Spirochaetes (1.70%), Chlamydiae (1.32%) and Bacteroidetes (1.31%) (Figure 2a). In RA sample,

100   unclassified bacteria were clustered as Caldithrix, Haloplasmatales and some phototrophic

101   bacteria.

102   Forty-three classes of bacteria were detected in RA sample (Figure 2b) where the first two most

103   abundant classes were Deltaproteobacteria (19.29%) and Alphaproteobacteria (16.89%), followed

104   by Acidobacteria (16.61%) (Figure 3b). The other minor classes included Clostridia (9.32%),

105   Gammaproteobacteria (5.50%), Actinobacteria (4.58%), Nitrospira (4.24%), Bacilli (3.24%),

106   Planctomycetia (3.08%), Spirochaetia (1.71%), Betaproteobacteria (1.69%), Opitutae (1.52%),

107   Chlamydia (1.33%), Verrucomicrobiae (1.19%), Holophagae (1.17%), Anaerolineae (1.06%),

108   and Ktedonobacteria (1.02%).

109   At the genus level, *Acidobacterium* of the Acidobacteria phylum was the dominant genus in RA

110   sample. The abundance frequencies of this genus is 10.01% (Figures 2c).

111   Five   classes   of   Proteobacteria   namely   Alphaproteobacteria,   Betaproteobacteria,

112   Deltaproteobacteria, Gammaproteobacteria and Epsilonproteobacteria, were detected in RA

113   sample). Among these, Deltaproteobacteria (43.88%) was the major class, followed by

114   Alphaproteobacteria (38.43%), Gammaproteobacteria (12.51%) Betaproteobacteria (RA 3.85%)

115   and Epsilonproteobacteria (RA 1.33%) (Figure 3). The segregation of orders within

116   Deltaproteobacteria (Figure 4) showed Syntrophobacterales was the most abundant order in this

117   soil sample. At the genus level of this order *Syntrophobacter* was the most abundant genus

118   (Figures 5).

119   The RA sample showed present of archaea but only phyla Crenarchaeota and Euyarchaeota were

120   detected. In the RA sample, Crenarchaeota (63.78%) was present at a higher percentage as

121   compared to Euyarchaeota (36.22%) (Figure 6). A total of eight classes of archaea were detected

122   in both soil samples and among them, Thermoprotei (RA 63.78%) and Methanomicrobia (RA

123   17.85%) were the two dominant classes. Other minor archaea classes present in the RA sample

124   were Thermococci (5.48%), Methanococci (5.35%), Thermoplasmata (3.52%), Methanobacteria

125   (2.17%), Archaeoglobi (1.36%) and Halobacteria (0.47%).

126   The taxonomic diversity for the domains Bacteria and Archaea was estimated at the genus level

127   using the Shannon-Weaver diversity index, H', in MEGAN and the H' value RA samples was

128   7.765.

129   Metabolic Functional Analysis via Reconstruction of Metagenome Library

130   The gene anthology was derived from the SEED classifications. Using this approach, the most

131   abundant gene detected in the RA sample was associated with carbohydrate metabolism (12.97%).

132   The second most abundant genes in the RA sample were associated with protein metabolism

133   (9.89%), virulence (9.53%), respiration (8.39%) and amino acids and their derivatives (8.19%)

134   (Figure 7).

135   **Discussion**

136   The application of the NGS enables the study of microbial diversity and function in metagenomes

137   without the need of culturing bacteria thus bypassing the growing of fastidious bacteria, which

138   are often unculturable, on laboratory media. However, this method depends on the reliability of

139   NGS data generated.

8

140   Most bacteria detected in the RA metagenomes were either anaerobic or facultative anaerobic.

141   They were predominantly from the domain Bacteria. However, in the RA metagenome, a

142   significant number of bacteria belonging to the domain Archaea were detected and the percentage

143   of archaeal abundance (21.48%) is significantly higher than the percentage values reported for

144   other soil metagenomes[17]. Crenarchaeota is known to be present at high sulfur content

145   environment and for its ability to utilize sulfur[18]. The high percentage of this phylum detected in

146   the RA sample is consistent with the high concentration of sulfur in the RA sample.

147   The bacterial diversity detected in the RA sample conforms to the common soil bacteria present

148   in other types of soil in other geographical locations[19]. In our RA sample, Proteobacteria remains

149   the most abundant bacteria and it comprised five different classes. However, the distribution of

150   Proteobacteria classes in this study differs from the distributions of Proteobacteria classes

151   reported for other mangrove habitats[8,20]. In contrast to other mangrove metagenomes reported to

152   date, both the Malaysian mangrove metagenome possessed Deltaproteobacteria as the dominant

153   class of Proteobacteria. Even though Proteobacteria was the dominant phylum in RA

154   metagenome, *Acidobacterium* of the Acidobacteria phylum was the most abundant genus in our

155   soil sample.

156   The presence of the high frequency of genes associated with carbohydrate metabolism in our soil

157   metagenome analysis is not surprising because these genes are commonly detected in abundance

158   in most studies of soil metagenomes[8,21]. However, the presence of the high frequency of

159   virulence factor genes in our soil metagenomes is unusual because they are not commonly

160   reported[8,22]. This leads to the speculation that mangrove soil is a potential reservoir of pathogenic

161   bacteria but further work is required to verify this finding.

162 In our RA soil metagenome, antibiotic and toxic compound resistance genes were also detected

163 frequently. Their abundance may be related to the high percentages of Actinobacteria, which are

164 known to produce a myriad of antibacterial compounds, and Deltaproteobacteria, members of

165 which are known to be resistant to heavy metals and to oxidize heavy metals to their benign

166 forms[23], in our two tropical mangrove soil samples. We also obtained a high hit rate of the stress

167 response gene in our soil metagenomic libraries, suggesting that the tropical mangrove

168 environment is harsh for microorganisms. This is most probably due to polluted marine waters

169 and high salinity and low aeration available in the muddy mangrove soil.

170 The biochemical tests showed considerably high amount of phosphorus and sulfur in our soil

171 sample. However, the gene ontology analysis revealed that the genes associated with the

172 metabolism of compounds containing phosphorus and sulfur were of relatively low abundance,

173 suggesting that these compounds may mainly be involved in redox reactions in electron transport

174 but not in microbial metabolism. This also implies that sulfur and phosphorus compounds exist in

175 stable forms as mangroves are sink for inorganic compounds.

176 Although Actinobacteria and Firmicutes were two of the detected sub-dominant phyla, our

177 analysis showed low frequency of spore producing bacteria such as *Bacillus* and *Streptomyces.*

178 This may explain why the genes for sporulation and dormancy were not detected in our soil

179 metagenomic DNA.

180 **<u>Conclusion</u>**

181 This study demonstrated the high level of microbial diversity in mangrove swamps compared to

182 the limited vegetation that is able to survive in this environment. The high abundance of members

183 of Deltaproteobacteria and heavy metal and toxic compound resistance genes indicates that

184 microorganisms have potential for bioremediation of heavy metals. The differences in the

185 distribution of microorganisms compared with other previous studies on mangrove soils are most

186 likely due to the different geographical locations. To the best of our knowledge, this is the first

187 study of microbial diversity of mangrove soil in Malaysia using the NGS metagenomic approach.

188 More mangrove soil samples collected from different locations in Malaysia are required to be

189 analysed by this approach before a more defined conclusion on the microbiome of Malaysian

190 mangrove soils and its functional genes can be reached.

191

192 **<u>Acknowledgements</u>**

197

References

[1] Tama NFY, Wong YS., Retention and distribution of heavy metals in mangrove soils receiving wastewater. *Environ Pollut*, 1996, **94**(3): 283-291.

[2] Holguin G, Vazquez P, Bashan Y., The role of sediment microorganisms in the productivity, conservation, and rehabilitation of mangrove ecosystems: an overview. *Biol Fert Soils*, 2001, **33**:265–278.

[3] Sahoo K, Dhal NK., Potential microbial diversity in mangrove ecosystems: a review. *Indian J Mar Sci*, 2009, **38**(2): 249-256.

[4] Sengupta A, Chaudhuri S., Ecology of heterotrophic denitrogen fixation in the rhizosphere of mangrove plant community at the Ganges river estuary in India. *Oecologia*, 1991, **87**: 560-564.

[5] Holguin G, Guzman MA, Bashan Y., Two new nitrogen-fixing bacteria from the rhizosphere of mangrove trees: their isolation, identification and in vitro interaction with rhizosphere *Staphylococcus* sp. *FEMS Microbiol Ecol*, 1992, **101**: 207-216.

[6] Alongi DM, Christoffersen P, Tirendi F., The influence of forest type on microbial nutrient relationships in tropical mangrove sediments. *J Exp Mar Biol Ecol*, 1993, **171**(2): 201-223.

[7] Vazquez P, Holguin G, Puente ME, Lopez-Cortes A, Bashan Y., Phosphate-solubilizing microorganisms associated with the rhizosphere of mangroves in semiarid coastal lagoon. *Bio Fertil Soils*, 2000, **30**: 460-468.

[8] Andreote FD, Jiménez DJ, Chaves D, Dias ACF, Luvizotto DM, Dini-Andreote F, Fasanella CC, Lopez MV, Baena S, Taketani RG, Melo IS., The microbiome of Brazilian mangrove sediments as revealed by metagenomics. *PLoS ONE*, 2012, doi: 10.1371/journal.pone.0038600.

13

221   [9] Xu Z, Hansen MA, Hansen LH, Jacquiod S, Sørensen SJ., Bioinformatic Approaches Reveal

222   Metagenomic Characterization of Soil Microbial Community. *PLoS ONE*, 2014, 9(4):

223   e93445. doi: 10.1371/journal.pone.0093445

224   [10] Paul EA., Soil Microbiology, Ecology and Biochemistry (Third Edition). *Elsevier Inc.*, 2007,

225   pp 389-430.

226   [11] Zhou J, Bruns MA, Tiedje JM., DNA recovery from soils of diverse composition. *Appl*

227   *Environ Microbiol*, 1996, **62**(2): 316-322.

228   [12] Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC., Gene and translation initiation site

229   prediction in metagenomic sequences. *Bioinformatics*, 2012, **28**(17): 2223-2230.

230   [13] Ye Y, Choi JH, Tang H., RAPSearch: a fast protein similarity search tool for short reads.

231   *BMC Bioinformatics*, 2011, doi:10.1186/1471-2105-12-159.

232   [14] Zhao Y, Tang H, Ye Y., RAPSearch2: a fast and memory-efficient protein similarity search

233   tool for next generation sequencing data. *Bioinformatics*, 2012, **28**(1): 125-126.

234   [15] Huson DH, Mitra S, Weber N, Ruscheweyh HJ, Weber N, Schuster SC., Integrative analysis

235   of environmental sequences using MEGAN4. *Genome Res*, 2011, **21**: 1552-1560.

236   [16] Disz T, Akhter S, Cuevas D, Olson R, Overbeek R, Vonstein V Stevens R, Edwards RA.,

237   Accessing the SEED genome databases via web services API: tools for programmers. *BMC*

238   *Bioinformatics*, 2010, doi:10.1186/1471-2105-11-319.

239   [17] Bates ST, Berg-Lyons D, Caporaso JG, Walters WA, Knight R, Fierer N., Examining the

240   global distribution of dominant archaeal populations in soil. *ISME J*, 2011, **5**(5): 908-917.

241   [18] Hou W, Wang S, Dong H, Jiang H, Briggs BR, Peacock JP, Huang Q, Huang L, Wu G, Zhi

242   X, Li W, Dodsworth JA, Hedlund BP, Zhang C, Hartnett HE, Dijkstra P, Hungate BA., A

243   comprehensive census of microbial diversity in hot springs of Tengchong, Yunnan Province

244   China   using   16S   RNA   gene   pyrosequencing.   *PLoS   ONE*,   2013,   doi:

245   10.1371/journal.pone.0053350.

246   [19] Janssen PH., Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S

247   rRNA genes. *Appl Environ Microbiol*, 2006, **72**(3): 1719-1728.

248   [20] dos Santos HF, Cury JC, do Carmo FL, dos Santos AL, Tiedje J, van Elsas JD, Rosado AS,

249   Peixoto RS., Mangrove bacterial diversity and the impact of oil contamination revealed by

250   pyrosequencing: bacterial   proxies   for   oil   pollution.   *PLoS   ONE*,   2011,   doi:

251   10.1371/journal.pone.0016943

252   [21] Jeffries TC, Seymour JR, Gilbert JA, Dinsdale EA, Newton K, et al., Substrate Type

253   Determines Metagenomic Profiles from Diverse Chemical Habitats. *PLoS ONE*, 2011, *6(9):*

254   *e25173. doi: 10.1371/journal.pone.0025173*

255   [22] Ghosh A, Dey N, Bera A, Tiwari A, Sathyaniranjan KB, Chakrabarti K, Chattopadhyay D.,

256   Culture independent molecular analysis of bacterial communities in the mangrove sediment

257   of Sundarban, India. *Saline Syst*, 2010, **6**: 1 doi: 10.1186/1746-1448-6-1.

258   [23] Blöthe M, Akob DM, Kostka JE, Göschel K, Drake HL, Küsel K., pH gradient-induced

259   heterogeneity of Fe(III)-reducing microorganisms in coal mining-associated lake sediments.

260   *Appl Environ Microbiol*, 2008, **74**(4): 1019-1029.

261

262

263

Parameter	Unit	RA
Arsenic (As)	mg/kg	ND(<0.5)
Cadmium (Cd)	mg/kg	0.10
Lead (Pb)	mg/kg	25.34
Mercury (Hg)	mg/kg	ND(<0.05)
pH (10% w/w)	-	5.1
Phosphorus	mg/kg	449.70
Sulfur	mg/kg	1374.52
Carbon:Nitrogen	%	5.19:0.31

264 **Table 1**. **Results of biochemical analyses of the soil sample**. ND: not detected.

265

266

Parameters	RA
Total no. of sequences (bp)	24 227 393 584
Total no. of quality sequences (bp)	20 907 568 942
Average quality read length (bp)	96.81
Total no. of contigs (>400bp)	552 541
Total no. of contigs length (bp)	502 888 532
Coverage	41.57
Percentage of hits against NCBI 16S Microbial database (%)	0.03
Total CDS assigned to SEED categories (%)	34.71

267 **Table 2**. **Reads statistics.** The numbers of reads were generated by Illumina HiSeq 2000. The

268 RA sample showed good quality of reads in term of number, length and number of contigs

269 generated.

270

17

271   <u>Titles and legends to figures</u>

272   **Figure 1 Bacteria and Archaea.** The majority of the reads were assigned to the domain Bacteria

273   by MEGAN. The number of reads detected for archea was significantly higher than previously

274   reported.

275   **Figure 2a The segregation of phyla in the domain Bacteria**. Proteobacteria was the dominant

276   phylum in the samples with almost half of the reads assigned to this particular phylum.

277   **Figure 2b Classes of bacteria.** Classes of bacteria that were present in the RA sample at more

278   than 1%.

279   **Figure 2c Bacterial genera.** The genera shown are those that had more than 1% reads. Bacterial

280   genera detected in RA sample in which *Acidobacterium* was the most abundant genus.

281   **Figure 3 Classes of Proteobacteria.** Deltaproteobacteria was the most abundant class in RA

282   samples.

283   **Figure 4 Order level of Deltaproteobacteria.** Syntrophobacterales were the most abundant

284   order.

285   **Figure 5 The most abundant genera in Syntrophobacterales.** *Syntrophobacter* was the most

286   abundant genus.

287   **Figure 6 Distribution of archaeal phyla in the sample.** RA sample possessed high percentage

288   of Crenarchaeota.

289   **Figure 7 Gene ontology**. The SEED classification by MEGAN.

290

Percentage of assigned reads to domain
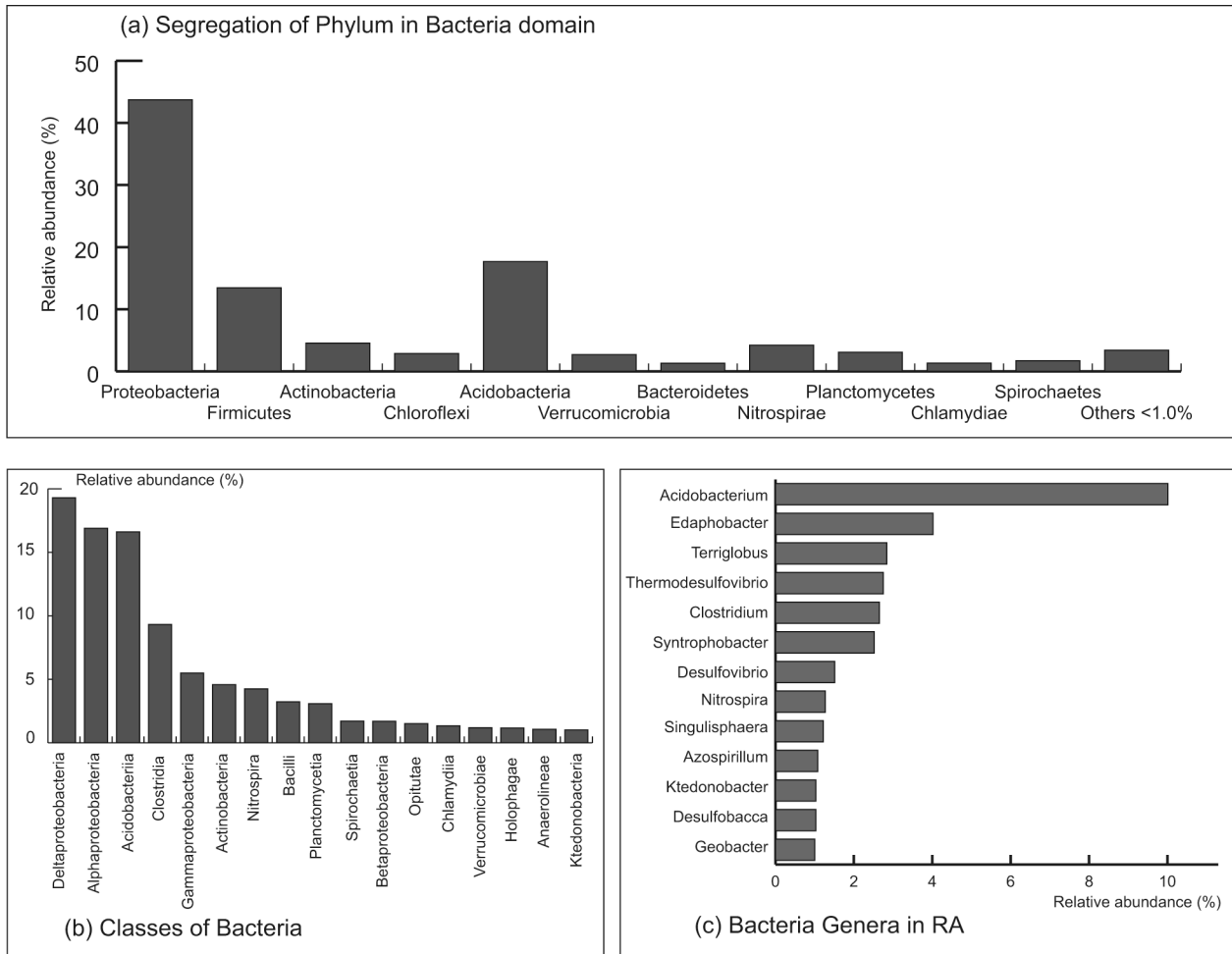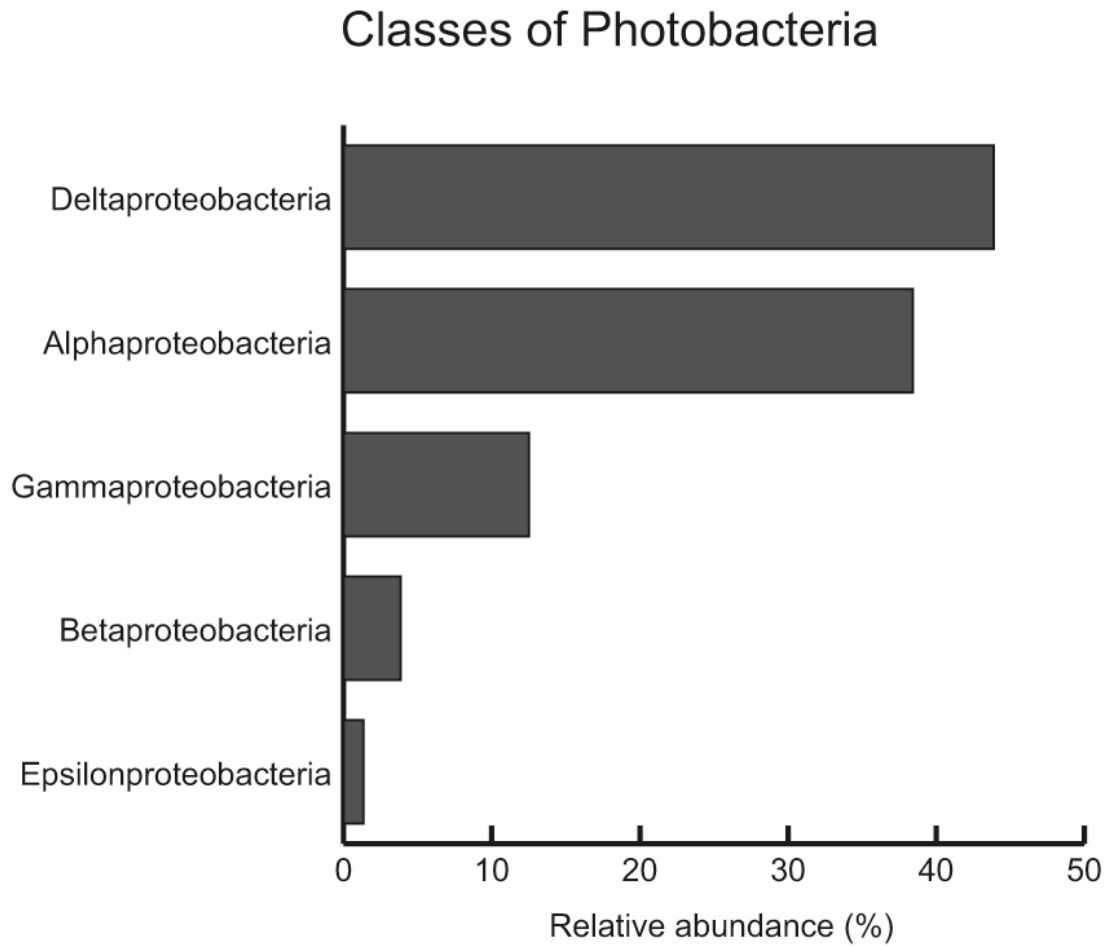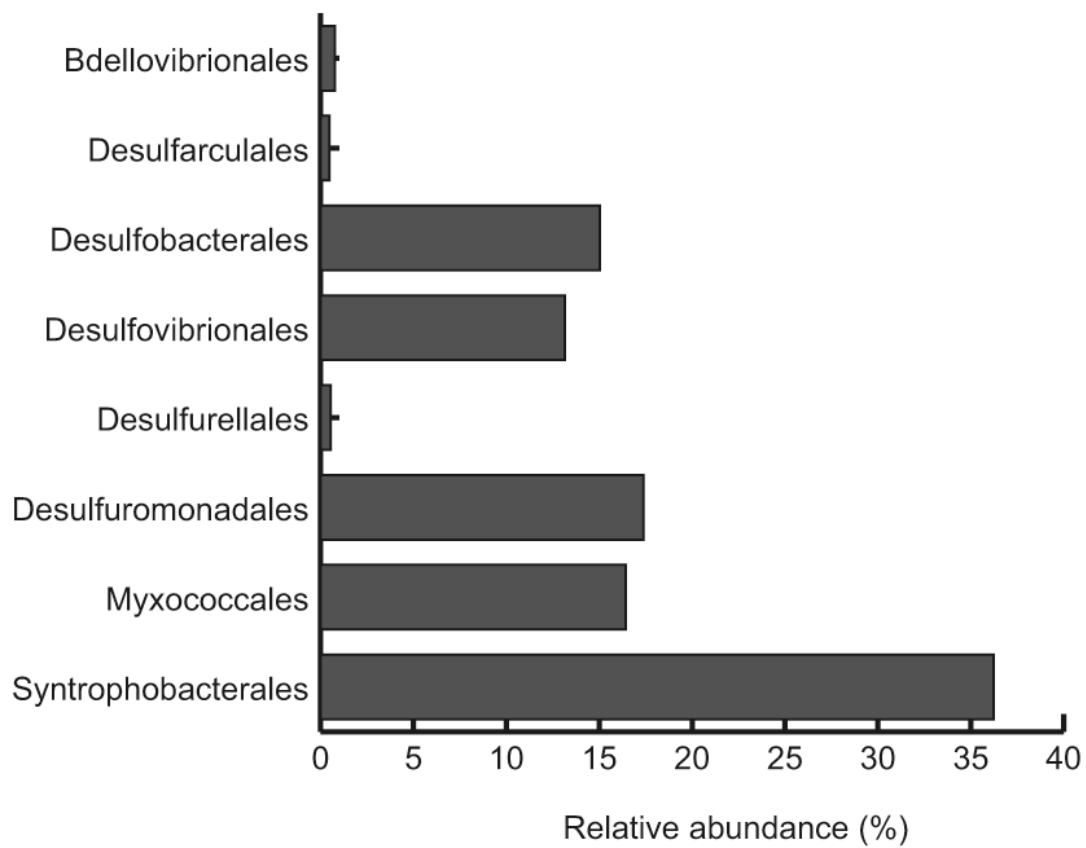

RA
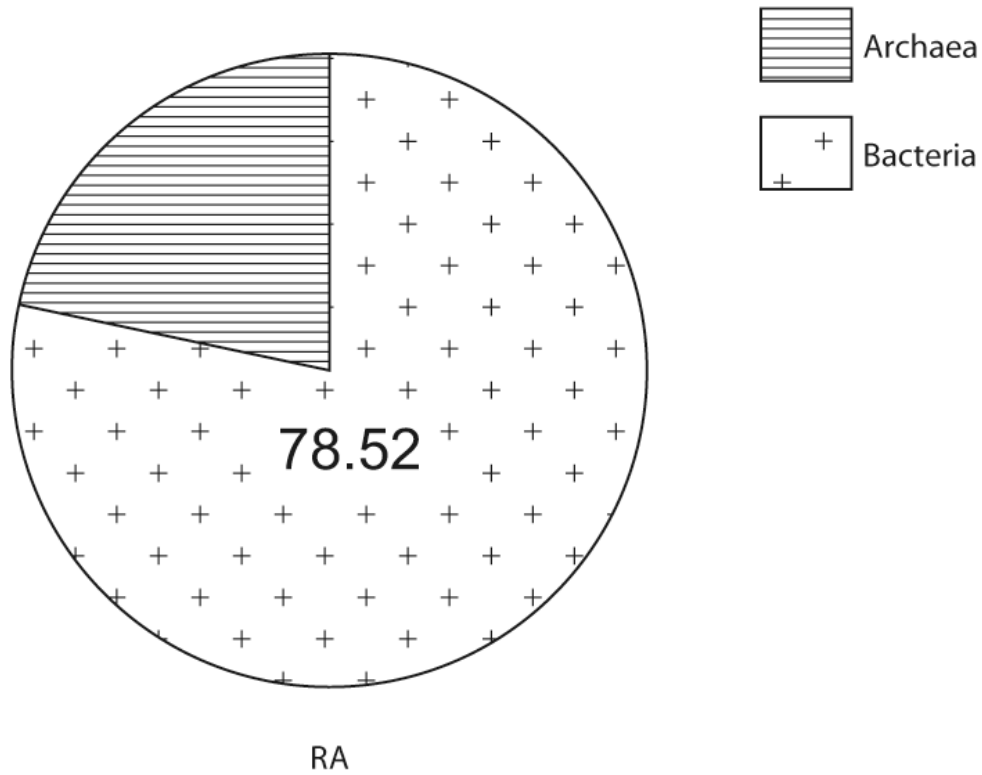
291

292   Figure 1

293

294
295

296   Figure 2

297

Classes of Photobacteria

298

299   Figure 3

300

301

302   Figure 4

303

Genus level of Syntrophobacterales

304

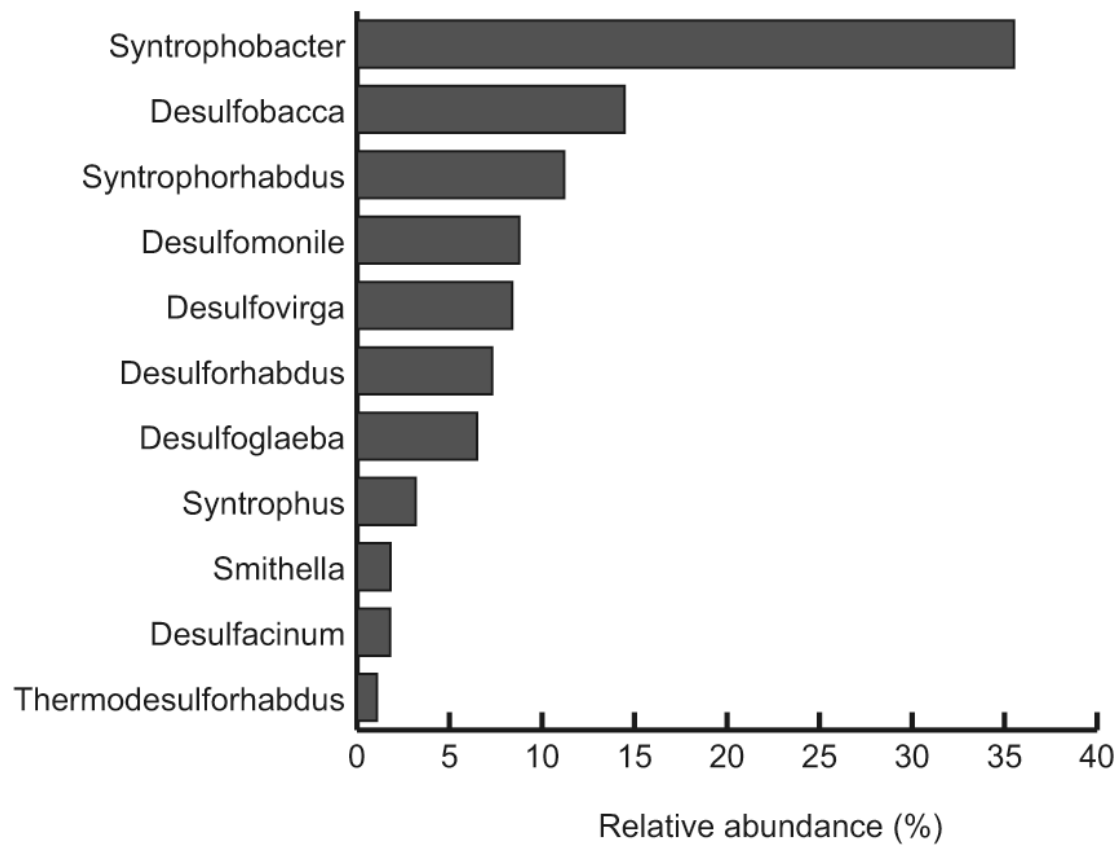305   Figure 5

306

Percentage Distribution of Archaea Phylum

307

308   Figure 6

309

Gene Ontology

311   Figure 7