**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**Novel method of probe design for characterising
unclassified microbial taxa in wastewater**

**TAN SHI MING
Interdisciplinary Graduate School
The Singapore Centre for Environmental Life Sciences
Engineering (SCELSE)**

2017

**Novel method of probe design for characterising
unclassified microbial taxa in wastewater**

TAN SHI MING

**Interdisciplinary Graduate School
The Singapore Centre for Environmental Life Sciences
Engineering (SCELSE)**

A thesis submitted to the Nanyang Technological University in partial
fulfilment of the requirement for the degree of Doctor of Philosophy

2017

Acknowledgements

First and foremost, I will like to express my gratitude towards my supervisor, Professor Yehuda Cohen for providing me with the academic freedom to pursue my PhD candidature in SCELSE. I thank him for the countless opportunities he has provided, the financial stability for the projects that I have pursued and his strong resolve in moulding me into an independent researcher.

Secondly, I will like to express my thanks to Dr Paul William, Mr Teo Guo Hui and Mr Ryan Lim for their tremendous help with the flow cytometry work. The flow cytometry work would not have gone smoothly without their invaluable expertise and advice. Prof Federico Lauro had proposed the use of FISH-FACS tool for the enrichment of the unclassified bacteria and I appreciate his input.

The sequencing team in SCELSE has helped with the sequencing aspect of my thesis. Special thanks go to Dr Daniela Moses whom I have consulted on the type of sequencing platform to use and Mr Alexander Putra who has efficiently handled my samples for sequencing.

Mr Larry Liew was instrumental in obtaining sludge samples from Ulu Pandan Water Reclamation Plant, and I will like to thank him for his time and effort.

Dr Xie Chao was the author of the RiboTagger software and he suggested the design of FISH probes from RiboTags. Mr Wesley Goi rendered advice in the updating of the SILVA database in the RiboTagger software. Dr Xiang Hui Liu was instrumental in attempting genomic binning using other software that was not covered in this thesis. Recovery of draft genomes from enriched samples was influenced through discussions with Dr Rohan Williams.

My sincerest appreciation goes towards Dr Muhammad Hafiz, Dr Martin Tay, Dr Rasmus Kirkegaard and Miss Krithika Arumugam for their kind patience in imparting basic bioinformatics knowledge to me. The thesis would not have been a success without the acquisition of these essential skills.

I will like to show my appreciation to Dr Nguyen Quoc Mai Phuong, Mr Muhammad Jasrie, Dr Maria Yung, Dr Lim Chun Ping and Prof Cao Bin for their invaluable advice and feedback on my thesis.

Heartfelt thanks to my friends in SCELSE: Syed Munir, Chan Siew Heng, Rosalie Chai, Choo Pei Yi, Adelia Li, Wong Jun Jie, Kelvin Chong and many more for creating a fun-filled environment to do my PhD in. The wonderful friendships forged have helped me through the tough times.

I am extremely grateful to the Environmental and Water Industry (EWI) and SCELSE for their generous scholarship and financial support during my PhD journey.

Finally, I will like to extend my utmost gratitude towards my family members: Dad, Mum and sister for their emotional support and pampering. Thank you for putting me through the wonderful education system in Singapore. Last but not least, a big thank you to my wonderful girlfriend, Julianna Ng for her unwavering faith in my abilities. Your constant encouragements have spurred me in every little way possible.

Publications

- 1.) **Shi Ming Tan**, Maria Yung, Chao Xie, Paul Hutchinson, Guo Hui Teo, Muhammad Hafiz Ismail, Martin Tay Qi Xiang, Rohan Williams, Yehuda Cohen. Design of next generation FISH probes from omics dataset for targeted visualisation and enrichment of environmental samples. Manuscript in preparation.
- 2.) **Shi Ming Tan**, Maria Yung, Chao Xie, Paul Hutchinson, Guo Hui Teo, Rohan Williams, Yehuda Cohen. Genome analysis of an unclassified and rare bacterial taxon in activated sludge recovered with next generation FISH probes. Manuscript in preparation.
- 3.) **Shi Ming Tan**, Chao Xie, Paul Hutchinson, Guo Hui Teo, Yehuda Cohen. Deciphering different spatial structures of *Haliangium* species in activated sludge with FISH and cell sorting. Manuscript in preparation.

Table of Contents

Acknowledgements.....	i
Publications.....	iii
Table of Contents.....	iv
List of Figures	x
List of Tables.....	xvi
List of Abbreviations	xviii
Abstract.....	xx
Chapter 1 Introduction	1
1.1 Background	1
1.2 Research gap.....	2
1.3 Aims and scopes.....	4
Chapter 2 Introduction	7
2.1 Activated sludge process for biological wastewater treatment	7
2.2 Unclassified microbial taxa present in floccular sludge community drives biological wastewater treatment under tropical conditions	10
2.3 Approaches to identification and classification of prokaryotes	15
2.3.1 Culture-dependent approach.....	15
2.3.2 Culture-independent approach.....	16
2.3.3 Methods of FISH probe design.....	24
2.3.4 Methods for enrichment of target bacterial taxa	29
Chapter 3 Novel method of probe design driven by next generation sequencing reads.....	35
3.1 Introduction	35

3.2	Materials and methods	37
3.2.1	Sample preparation.....	38
3.2.2	Sample fixation.....	39
3.2.3	Fluorescence in situ hybridisation (FISH)	39
3.2.3.1	Evaluation of the in silico specificity and coverage of probes	42
3.2.4	Probe dissociation curve	43
3.2.5	Sequencing of the 16S rRNA gene	44
3.2.6	Visualisation of probe accessibility site of the 16S rRNA gene	45
3.2.7	Co-localisation analyses	45
3.3	Results	46
3.3.1	Evaluation of the in silico specificity and coverage of FISH probes	46
3.3.2	In silico accessibility of 16S rRNA to RiboProbe	48
	In situ validation of RiboProbe	49
3.3.3	Determining probe stringency of RiboProbe	51
3.3.4	Hybridisation of RiboProbe with activated sludge.....	52
3.3.5	Co-localisation assays.....	54
3.3.6	Applications of RiboProbe.....	57
3.4	Discussion.....	61
3.4.1	Design of RiboProbe.....	61
3.4.2	Validation of RiboProbe	62
3.4.3	Evaluation of RiboProbe.....	63
3.4.4	Summary	66
Chapter 4	Targeted enrichment of microbial group through FISH-FACS.....	68

4.1	Introduction	68
4.2	Methods	72
4.2.1	Sample preparation.....	72
4.2.2	Fluorescence in situ hybridisation	74
4.2.3	Quantitative FISH	76
4.2.4	Fluorescence activated cell sorting (FACS)	76
4.2.5	Multiple displacement amplification (MDA).....	78
4.2.6	16S rDNA clone libraries	78
4.2.7	Phylogenetic analysis	79
4.2.8	Metagenomics sequencing	80
4.2.9	RiboTagger 16S rRNA analysis.....	81
4.2.10	16S rRNA nucleotide sequence accession numbers	81
4.3	Results	81
4.3.1	Evaluating suitability of FACS sorters for targeted enrichment.....	81
4.3.2	Eliminating contamination for downstream MDA.....	82
4.3.3	Specific sorting of target population from an axenic culture	87
4.3.4	Evaluating the effectiveness of sorting from an axenic culture.....	92
4.3.5	Specific sorting of a target taxon from the floccular sludge community.....	96
4.3.6	Evaluating the effectiveness of sorting from activated sludge samples.....	99
4.3.7	Comparing environmental specificity of FISH probes.....	105
4.3.8	Phylogeny of Thauera obtained from sorted sample	107
4.4	Discussion.....	110
4.4.1	Selection of an effective FACS sorter	110

4.4.2	Effectiveness of sorting with RiboProbe	112
4.4.3	Managing DNA contamination.....	114
4.4.4	Environmental specificity of RiboProbe	116
4.4.5	Phylogenetic analysis of Thauera	118
4.4.6	Summary	119
Chapter 5	Characterisation of unclassified bacteria taxa in activated sludge systems	120
5.1	Introduction	120
5.2	Methods	123
5.2.1	Sample preparation.....	123
5.2.2	Fluorescence in situ hybridisation	124
5.2.3	Fluorescence activated cell sorting (FACS)	126
5.2.4	Multiple displacement amplification (MDA).....	126
5.2.5	Generation of clone libraries and phylogenetic analysis	126
5.2.6	Metagenomics sequencing and analysis.....	127
5.2.7	Taxonomic assignment of ORFs	129
5.2.8	Coding potential of ORFs.....	129
5.2.9	Nucleotide sequence accession numbers.....	129
5.3	Results	130
5.3.1	Probe design.....	130
5.3.2	Visualisation of UPWRP_1.....	130
5.3.3	Calibration of probe Ribo_Unk1029_17	131
5.3.4	Sorting of UPWRP_1.....	132
5.3.5	Evaluation of cell sorting effectiveness of UPWRP_1	134

5.3.6	16S rRNA phylogeny of UPWRP_1	136
5.3.7	Metagenomic analysis of UPWRP_1	138
5.3.8	Phylogenomics analysis of UPWRP_1	143
5.3.9	Functional analysis of UPWRP_1.....	146
5.3.10	Visualisation of UPWRP_2.....	148
5.3.11	Calibration of probe Ribo_Halia1029_17.....	151
5.3.12	Cell sorting of Haliangium	151
5.3.13	Evaluation of the sorting efficiency	154
5.3.14	16S rRNA phylogeny of UPWRP_2	157
5.3.15	Design of specific FISH probe for UPWRP_2	158
5.3.16	Metagenomic analysis of UPWRP_2	160
5.3.17	Phylogenomics analysis of UPWRP_2	163
5.3.18	Functional analysis of UPWRP_2.....	165
5.3.19	Evaluation of genome completeness, contamination and assembly quality of draft genomes with other studies	166
5.4	Discussion.....	169
5.4.1	Effect of multiple displacement amplification on genomic binning	169
5.4.2	FACS sorting	172
5.4.3	Genome recovery.....	175
5.4.4	Phylogeny and phylogenomics.....	179
5.4.5	Low similarity of draft genomes to reference genomes.....	182
5.4.6	Functional analysis of novel taxa	183
Chapter 6	Conclusion and perspectives.....	184

References 189

Appendix 204

List of Figures

Figure 2-1: An illustration of the structure of an activated sludge floc.	7
Figure 2-2: Reactor configuration for biological wastewater treatment process at UPWRP.	8
Figure 2-3: Relative abundance of floccular microbial community in UPWRP as estimated by RiboTagger 16S rRNA analysis.	11
Figure 2-4: A model representing the higher-order secondary structure of prokaryotic 16S rRNA gene.	17
Figure 2-5: Outline of the steps in full-cycle rRNA approach adopted for microbial ecology studies.	19
Figure 2-6: Extraction of taxonomically-relevant RiboTags using the universal recognition profiles with RiboTagger software.	27
Figure 3-1: An illustration of FISH probe designed through the process of comparative sequence analysis.	35
Figure 3-2: Flowchart demonstrating the steps involved in validation and evaluation of RiboProbe.	38
Figure 3-3: A diagram illustrating FISH probe design from 33bp-RiboTag.	40
Figure 3-4: Evaluation of the in silico coverage and specificity of RiboTagger and canonical probes against the genus <i>Thauera</i>	47
Figure 3-5: 16S rRNA gene sequences of the genus <i>Thauera</i> in an ARB-parsimony guide tree covered by various probe combination of probes Ribo_Thau1029_17, Thau646 or both.	48
Figure 3-6: Confocal micrographs of an axenic culture of R086 co-hybridised with probes Ribo_Thau1029_17 _{Cy5} , Thau646 _{Cy3} and EUB 338 _{A488}	50
Figure 3-7: Probe dissociation curve of probe Ribo_Thau1029_17.	51
Figure 3-8: Confocal micrographs of <i>Thauera</i> in fixed samples of activated sludge.	53
Figure 3-9: Confocal micrograph of different morphologies of <i>Thauera</i> existing in UPWRP's floccular sludge.	54

Figure 3-10: Co-localisation scatterplots of Cy5 and Cy3 pixel intensities of activated sludge samples.	55
Figure 3-11: Pearson's correlation coefficient (PCC) of confocal images of activated sludge samples that were co-hybridised with probes Ribo_Thau1029_17 _{Cy5} and Thau646 _{Cy3} , or a single-labelled probe.....	56
Figure 3-12: Manders' co-localisation coefficient (MCC) of confocal images of activated sludge samples that were co-hybridised with probes Ribo_Thau1029_17 _{Cy5} and Thau646 _{Cy3}	57
Figure 3-13: Design of a second probe Ribo_Unk1009_17 on the same sequencing tag that probe Ribo_Unk1029_17 was designed from.	58
Figure 3-14: Confocal micrographs of members of an unclassified bacterial taxon in UPWRP's floccular microbial community.	59
Figure 3-15: Ethidium bromide-stained agarose gel of PCR products amplified with different primer sets.	60
Figure 4-1: An illustration of FISH-FACS for the sorting of fluorescent-labelled target cells from a mixed microbial community.	69
Figure 4-2: Flowchart demonstrating the steps involved in FISH-FACS for the enrichment of a target taxon, followed by analyses to determine the levels of enrichment and phylogeny of the target taxon.	72
Figure 4-3: Features of pCR 4-TOPO vector with its cloning site.	78
Figure 4-4: Diversity of a Sy3200-sorted sample analysed by RiboTagger 16S rRNA analysis.	83
Figure 4-5: An illustration depicting the potential hotspots of contamination in a FACS machine.	84
Figure 4-6: Ethidium bromide-stained agarose gel of negative controls that were alkaline-lysed, subjected to MDA amplification and PCR-amplified with primer set 27F/1492R to detect for contamination.....	86
Figure 4-7: Confocal micrographs demonstrating the process of sonicating dense clusters of Thauera cells from an axenic culture into a single-cell suspension for FACS sorting.	87

Figure 4-8: Flow cytometric analysis of sorting from an axenic culture of R086 hybridised with probes Ribo_Thau1029_17 _{Cy5} and EUB388 _{A488}	90
Figure 4-9: Ethidium bromide-stained agarose gel of sorted samples.	91
Figure 4-10: Cell sorting purity obtained from an initial round of sorting from an axenic culture of R086.	93
Figure 4-11: Confocal micrographs of probe-labelled cells sorted from sorting gates 1-4 after FISH-FACS of an axenic culture of R086.	94
Figure 4-12: Relative abundance of RiboTags matching to probe Ribo_Thau1029_17 and Thauera-specific OTUs in pre-sorted and sorted samples from axenic cultures of R086.	95
Figure 4-13: Distribution of RiboTags obtained from sorted samples of R086 into various categories.	96
Figure 4-14: Confocal micrographs demonstrating the process of sonicating dense clusters of activated sludge flocs into single-cell suspension for FACS sorting.	97
Figure 4-15: Flow cytometric analysis of sorting Thauera from an activated sludge sample hybridised with probes Ribo_Thau1029_17 _{Cy5} and EUB388 _{A488}	99
Figure 4-16: Cell sorting purity obtained from the initial round of sorting from activated sludge samples.	100
Figure 4-17: Factors affecting the purity of cell sorting.	101
Figure 4-18: Confocal micrographs depicting the heterogeneous distribution of Thauera in activated sludge.	102
Figure 4-19: Confocal micrographs of probe-labelled cells sorted from sorting gates 1-4 after FISH-FACS of an activated sludge sample.	103
Figure 4-20: Quantitative FISH analysis depicting the relative abundance of Thauera in pre-sorted and sorted samples from activated sludge.	103
Figure 4-21: Relative abundance of RiboTags annotated to probe Ribo_Thau1029_17 and Thauera in pre- and post-sorted samples from activated sludge samples.	105

Figure 4-22: Comparison of the environmental specificity of probes Ribo_Thau1029_17 and Thau646 for Thauera in activated sludge samples.	106
Figure 4-23: Maximum-likelihood (PhyML) phylogenetic tree depicting the 16S rRNA phylogenetic relationship of representative sequences (99% sequence identity cut-off) obtained from clone libraries of probe Ribo_Thau1029_17 sorted sample and its closest relatives.	109
Figure 5-1: Flowchart describing the processes used in the enrichment and subsequent characterisation of unclassified bacterial taxa from activated sludge samples.	123
Figure 5-2: Confocal micrographs of members of UPWRP_1 hybridised with probes Ribo_Unk1029_17 _{Cys} (red) and EUB338 _{A488} (green) in UPWRP activated sludge samples.	131
Figure 5-3: Probe dissociation curve of probe Ribo_Unk1029_17 performed on technical triplicates of UPWRP activated sludge samples.	132
Figure 5-4: Confocal micrographs of the different processes used to obtain single cell suspension of UPWRP_1, and flow cytometric plots that represented the processes.	133
Figure 5-5: Effectiveness of cell sorting and levels of enrichment for UPWRP_1 from activated sludge samples were estimated with quantitative-FISH analysis, flow cytometric analysis and RiboTagger 16S rRNA analysis.	135
Figure 5-6: Maximum-likelihood (PhyML) phylogenetic tree depicting the 16S rRNA phylogenetic relationship of representative sequences of OTUs (99% cut off) obtained from clone libraries of sorted samples hybridised with probe Ribo_Unk1029_17, and its closely related sequences in the SILVA database.	137
Figure 5-7: A differential coverage plot of contigs from co-assembly of the negative controls.	138
Figure 5-8: Decontamination of the contigs of UPWRP_1 produced three clusters of contigs.	139
Figure 5-9: Visualisation of the differential coverage binning plot of contigs from UPWRP_1... ..	141
Figure 5-10: Phylogenetic placement of genomics bins of Candidatus Shimmingles in the reference genome tree using a concatenation of 43 phylogenetic marker genes.	143
Figure 5-11: Taxonomic comparison of ORFs from Candidatus Shimmingles and Lewinella persica DSM 23188 using Megan's LCA approach.	145

Figure 5-12: Relative comparison of PEGs from the genomes of <i>Candidatus Shimmingles merlion</i> , <i>Candidatus Shimmingles singa</i> and <i>Lewinella persica</i> DSM 23188 being assigned to the highest SEED subsystem hierarchy.	146
Figure 5-13: Confocal micrographs depicting the different morphotypes of <i>Haliangium</i> cells hybridised with probes Ribo_Halia1029_17 _{Cy5} (red) and EUB338 _{A488} (green) in activated sludge samples.	149
Figure 5-14: Confocal micrographs of different clades of <i>Haliangium</i> hybridised with probes Ribo_Halia1029_17 _{Cy5} (red), HalianMix _{Cy3} (magenta) and EUB338 _{A488} (green) in UPWRP activated sludge samples.	150
Figure 5-15: 16S rRNA gene sequences of the genus <i>Haliangium</i> represented in the ARB-parsimony guide tree covered by the various probe combination of probe Ribo_Halia1029_17 or HalianMix.	150
Figure 5-16: Probe dissociation curve of probe Ribo_Halia1029_17 performed on technical triplicates of UPWRP activated sludge samples.	151
Figure 5-17: Confocal micrographs depicting the process of breaking up dense clusters of <i>Haliangium</i> into single-cell suspension for FACS sorting.	152
Figure 5-18: Different morphotypes of <i>Haliangium</i> cells captured with different sorting gates of forward versus side scatter during the second round of sorting.	153
Figure 5-19: Effectiveness of cell sorting and levels of enrichment for UPWRP_2 from activated sludge samples were estimated with: quantitative-FISH analysis, flow cytometric analysis and RiboTagger 16S rRNA analysis.	154
Figure 5-20: Relative abundance of (A) UPWRP_2 and (B) <i>Haliangium</i> species (AB286567) in samples sorted with sorting gates using high- or low-light scattering properties, plotted against the different number of sort events: 10 and 1000 events.	156
Figure 5-21: Maximum-likelihood (PhyML) phylogenetic tree depicting the 16S rRNA phylogenetic relationship of OTUs (99% cut off) obtained from clone libraries of sorted samples hybridised with probe Ribo_Halia1029_17, and its closely related sequences in the SILVA database.	158

Figure 5-22: Confocal micrographs depicting the spatial interaction of UPWRP_2 in fixed samples of activated sludge.....	159
Figure 5-23: Confocal micrographs showing that the spherical-shaped objects were produced by other Haliangium species, but not by UPWRP_2 in fixed samples of activated sludge.	160
Figure 5-24: Decontamination of the contigs of UPWRP_2 produced five clusters of contigs. ..	161
Figure 5-25: Phylogenetic placement of draft genome of UPWRP_2 in the reference genome tree using a concatenation of 43 phylogenetic marker genes.	163
Figure 5-26: Taxonomic comparison of ORFs from Haliangium clustero and Haliangium ochraceum DSM 14365 using Megan’s LCA approach.	164
Figure 5-27: Relative comparison of genes from the genomes of Haliangium clustero and Haliangium ochraceum DSM 14365 being assigned to the highest SEED subsystem hierarchy.	165
Figure 5-28: Comparison of the genome completeness and contamination of draft genomes.	167
Figure 5-29: Comparison of the assembly quality of draft genomes.....	168

List of Tables

Table 2-1: Differences in probe design between RiboTagger and comparative sequence analysis	29
Table 3-1: FISH probes and RiboTags used in this chapter	41
Table 3-2: PCR primers used for the evaluation of <i>Thauera</i>	44
Table 3-3: In silico accessibility of R086 to RiboProbes extracted from variable regions V4-V7... ..	49
Table 3-4: Homologous probe binding region of Ribo_Thau1029_17 in axenic cultures of R086 and <i>Thauera Linaloolentis</i>	52
Table 4-1: Sampling of axenic culture and activated sludge samples for biological and technical replicates.....	73
Table 4-2: FISH probes used for sorting of <i>Thauera</i> from axenic cultures and activated sludge samples	74
Table 4-3: Number of replicates processed for MiSeq genomic sequencing	80
Table 4-4: Differences between the Sy3200 (Sony) and MoFlo XDP (Beckman Coulter) sorter ...	82
Table 4-5: Actions implemented to eliminate DNA contaminants in the FISH-FACS workflow	85
Table 4-6: Sequence and relative abundance of RiboTags of the sorted sample used for phylogenetic analysis	107
Table 4-7: Single nucleotide polymorphism (A→C) observed in members of OTU2	108
Table 5-1: Biological and technical replicates used for analysis of unclassified bacterial taxa ...	124
Table 5-2: FISH probes and RiboTags used for the characterisation of UPWRP_1 and UPWRP_2	125
Table 5-3: Number of pre- and post-sorted samples processed for HiSeq genomic sequencing	127
Table 5-4: Genomic statistics for the clusters of contigs produced from the decontamination process	140
Table 5-5: Statistics for the draft genomes of UPWRP_1	142

Table 5-6: Major RiboTags present in FISH-FACS sorted samples with probe Ribo_Halia1029_17
..... 155

Table 5-7: Statistics for the draft genome of UPWRP_2..... 162

List of Abbreviations

16S rRNA	16S ribosomal RNA
AAI	Average amino acid identity
ANI	Average nucleotide identity
AOB	Ammonia-oxidizing bacteria
bp	Base pair
CTC	5-Cyano-2,3-ditolyl tetrazolium chloride
DAIME	Digital Image Analysis In Microbial Ecology
EBPR	Enhanced biological phosphorus removal
EPS	Extracellular polymeric substance
FA	Formamide
FISH	Fluorescence <i>in situ</i> hybridisation
Kbp	Kilo base pairs
LCA	Lowest common ancestor
Mbp	Mega base pairs
MCC	Manders' co-localisation coefficient
MDA	Multiple displacement amplification
NaCl	Sodium chloride
NGS	Next generation sequencing
NOB	Nitrite-oxidizing bacteria

ORFs	Open reading frames
OTU	Operational taxonomic unit
PAOs	Polyphosphate-accumulating organisms (PAOs)
PBS	Phosphate buffer saline
PCC	Pearson's correlation coefficient
PCR	Polymerase chain reaction
PEGs	Protein-encoding genes
PHAs	Poly- β -hydroxyalkanoates (PHAs)
PI	Propidium iodide
RAST	Rapid Annotation using Subsystem Technology
rRNA	Ribosomal RNA
SD	Standard deviation
SEM	Standard error of the mean
SMRT	Single molecule real time sequencing
SND	Simultaneous nitrification and denitrification
SRT	Sludge retention time
UPWRP	Ulu Pandan Water Reclamation Plant
WGA	Whole genome amplification

Abstract

Floccular microbial communities inhabiting the activated sludge biosphere of wastewater treatment plants are widely exploited for their metabolic activities in removing anthropogenic carbon and inorganic nutrients. Although numerous studies have identified key microorganisms responsible for catalysing important bioprocesses, the activated sludge system is still regarded as a 'black box' because only a small proportion of the community has been taxonomically classified. 'Black box' of the sludge biosphere was demonstrated through deep sequencing to saturation of the floccular sludge community obtained from a municipal wastewater treatment plant in Singapore. The tag sequence of 37 out of the 50 most abundant OTUs could not be taxonomically classified, and none of the draft genomes obtained through genomic analysis were close to completion. Characterising the unclassified bacteria taxa has several implications. One apparent advantage is the broadening of our knowledge on the prokaryotic tree of life, which can subsequently lead to the design of more accurate PCR primers and FISH probes.

This thesis describes the improvisation of a methodological approach that involves the targeted enrichment of unclassified bacterial taxa from floccular sludge community through FISH-FACS to obtain its "mini-metagenome" prior to downstream multiple displacement amplification (MDA), genomic sequencing and analyses. Even without access to a dedicated sterile environment and FACS machine for MDA-experiments, the FISH-FACS methodology outlined in this thesis provides a comprehensive guideline that manages DNA contamination and sterilisation of a communal FACS machine. Target cells were labelled with a new variant of fluorescence *in situ* hybridisation (FISH) probes which are termed as RiboProbes, and subsequently isolated via a highly-sensitive fluorescence-activated cell sorting (FACS) sorter. In contrast to canonical FISH probe design, RiboProbes were designed from short sequencing reads present in omics dataset (metagenomics and metatranscriptomics) that corresponded to highly variable regions (V4-V7) of the 16S rRNA. RiboProbes were affiliated with a high taxonomic resolution that allowed the differentiation of unclassified bacterial taxa down to the species level. However, truncation of the original length of the RiboProbe (33 bp) is necessary for combinatorial use with canonical FISH probes (average

length: 17-25 bp) at a standardised hybridisation temperature. FISH-FACS was initially validated on a specific *Thauera* species in an axenic culture as a reference organism, followed by activated sludge samples where high level of enrichment of *Thauera* (>97%) was reproduced across many replicates. FISH-FACS with a canonical FISH probe produced lower level of enrichment (63%) for *Thauera* due to lower specificity of the probe, which subsequently led to the inclusion of non-target bacterial taxa in the sorted samples.

FISH-FACS methodology was further extrapolated to two unclassified bacterial taxa in activated sludge, where target population was enriched through a collection of sorted samples that contained multiple events (5-1000). FISH-FACS resulted in one of the unclassified taxa (*Candidatus Shimingles*) being enriched to 99% and the other taxon (*Haliangium clustero*) being enriched to 44.14%. Level of enrichment for *Haliangium clustero* was lower than *Ca. Shimingles* due to co-sorting of closely-related *Haliangium* species into the sample. Due to the low mean read depth of sequencing ($2.56X \pm 2.97$), no RiboTags could be assigned to *Ca. Shimingles* and *Haliangium clustero* in the pre-sorted samples of activated sludge. Thus far, FISH-FACS with RiboProbe is the only method that has been used for the enrichment and genome recovery of these two unclassified taxa. No other enrichment methods have been described for these novel taxa, except for culture-dependent methods used for the isolation of *Haliangium* species from costal and terrestrial regions.

Taxonomic identification was made possible through phylogenetic assignment of full-length 16S rRNA gene sequences obtained from clone libraries. Through the 16S rRNA phylogenetic analysis, *Ca. Shimingles* was a novel bacterial genus categorised to an unclassified family of the order Sphingobacteriales, and *Haliangium clustero* was a novel bacterial species categorised to the genus *Haliangium*. Two distinct clades were observed in the 16S rRNA phylogenetic analysis of the *Ca. Shimingles*; distinct clades were in line with two draft genomes recovered through genomic binning. The two genomic bins were recovered with a size of 4.92 Mbp and 4.49 Mbp respectively, and a completeness of more than 90% and a contamination of less than 4%. The other draft

genome recovered from *Haliangium clusterio* had a size of 2.37 Mbp, with a completeness of 21.70% completeness and 0.02% contamination. A more complete genome of *Haliangium clusterio* could not be obtained due to the complications of genome binning from strain heterogeneity. Even though the sorted samples were sequenced on the same sequencing flow cell as the pre-sorted samples with the same sequencing depth, genome coverage of both the novel taxa had vastly improved with the FISH-FACS enrichment procedure. The draft genomes exhibited low average amino acid identity and average nucleotide identity to the closest reference genomes in the database, therefore demonstrating novelty of these draft genomes.

The novel *Haliangium* species formed aggregated structures that were distinct from other documented *Haliangium* species in activated sludge. Fruiting bodies of the novel *Haliangium* species were species-specific even in a mixed microbial environment. In addition, the formation of spherical myxospores was shown to be a characteristic of certain *Haliangium* species, and it was not strongly associated with the novel *Haliangium* species. The use of FISH probe to label myxospores and FACS to sort myxospores from the vegetative cells is unprecedented, and this method was first established in this thesis. Here, a reproducible and robust approach of overcoming the current limitations of conventional FISH probes design and metagenomics for the visualisation, identification and recovery of draft genomes from previously unclassified microbial taxa in wastewater is presented.

Chapter 1 Introduction

1.1 Background

With a staggering population of 5.6 million people (Singapore Department of Statistics, 2016) and a demand of 430 million gallons of water per day (Singapore Public Utility Board, 2016), the provision of clean drinking water is crucial to Singapore where there is a lack of natural water resources. One of Singapore's strategies in supplying clean drinking water is the purification of sewage water. Like most countries, Singapore's wastewater reclamation plants have integrated the activated sludge bioprocess to treat polluted wastewater containing elevated concentration of complex organic compounds and inorganic nutrients, such as phosphorus and nitrogen prior to their release into receiving water bodies. Elevated levels of inorganic nutrients are detrimental to water bodies because they have been shown to cause eutrophication, which consequently presents a public health risk and a lapse in environmental regulation (Mainstone and Parr, 2002; Seviour and Nielsen, 2010).

Optimal functioning of the activated sludge bioprocess is dependent on the metabolic activities of microbial aggregates in activated sludge flocs to drive degradation of organic compounds and sequestration of inorganic nutrients (Tchobanoglous *et al.*, 2003). Highly active flocs are circulated across chemically and physically defined compartmentalized reactors where key physicochemical parameters such as dissolved oxygen concentration, nutrient levels and sludge retention time (SRT) are routinely measured (Daims *et al.*, 2006). As such, these reactors are amenable to manipulations that can create a range of selective pressures on the floccular sludge communities with the aim of achieving a targeted process performance. For instance, the recently discovered comammox process: complete oxidation of ammonium to nitrate by a single *Nitrospira* species was achieved by supplying low concentrations of ammonium, nitrite and nitrate under hypoxic conditions (van Kessel *et al.*, 2015). Although integration of the disciplines of process engineering and wastewater microbial ecology has led to the enhancement of wastewater process

performances, the structure and function of floccular sludge communities driving wastewater is still regarded as a 'black box'.

1.2 Research gap

Many microorganisms that catalysed key biological treatment processes have been identified through a wide range of culture-independent molecular biology methods (Sanz and Köchling, 2007). However, due to limitations of these methods, the microorganisms that have been identified only represent the tip of the iceberg of what is known about the floccular sludge community. Recently, deep sequencing to saturation of the complex floccular sludge community obtained from the Ulu Pandan wastewater reclamation plant (UPWRP) in Singapore was undertaken by Kjelleberg *et al.* (unpublished). The deep sequencing to saturation approach presented a diverse and novel bacterial community driving wastewater purification, with taxonomic novel entities in the long-tail of rank-abundance curve present at low fractions of the overall biomass (Lynch and Neufeld, 2015). The degree of taxonomic novelty described was substantially greater than other floccular sludge communities in Hong Kong (Yu and Zhang, 2012) and Denmark (Albertsen *et al.*, 2012) that have been previously characterised by metagenomic sequencing.

While shotgun sequencing is an invaluable tool in profiling the community structure and understanding the function potential of the community, the lack of reference genomes resulting from metagenomic studies is limiting the progress in deciphering microbial ecology of microorganisms critical to wastewater treatment (Albertsen *et al.*, 2013b). Genomic binning has shown great promise in facilitating genome recovery from shotgun surveys, but the low relative abundance of microbial groups is often a limiting factor in genome recovery (Albertsen *et al.*, 2013a). To circumvent the pitfalls of metagenomics and binning in genome recovery, low abundance microbial groups can be enriched through the process of FISH-FACS to obtain its 'sub-metagenome' prior to genomic sequencing. FISH-FACS involves specific hybridisation of fluorescence *in situ* hybridisation (FISH) probes to the 16S ribosomal RNA (rRNA) of target taxon,

followed by isolation of probe-labelled bacterial cells with a highly-sensitive fluorescence-activated cell sorting (FACS) machine that supports prokaryotic cell sorting.

Ideally, FISH probes could be designed from highly variable region of the 16S rRNA on short sequencing reads obtained from the community shotgun survey at UPWRP, with a meaningful level of taxonomy resolution that can be used for characterising the taxonomic novel entities. However, the design of FISH probes from short sequencing reads in omics dataset has not been reported. Conventional FISH probe design necessitates the alignment of full-length 16S rRNA sequences for comparative sequence analysis. However, the gold standard of generating full-length 16S rRNA sequences for low abundance microorganisms is often limited by PCR primer and amplification bias (Hong *et al.*, 2009). Recent advancement in molecular tagging (Burke and Darling, 2016; Karst *et al.*, 2016a) and longer read length of sequencing technology (e.g. real-time sequencing technology) have bypassed the conventional approach of clone libraries and Sanger sequencing in producing full-length 16S rRNA sequences from mixed communities on next generation sequencing platforms. Although these tools are promising for probe design, no probes were designed from the full-length 16S rRNA sequences and no visual evidence of the novel taxonomic entities was provided (Burke and Darling, 2016; Karst *et al.*, 2016a; Wagner *et al.*, 2016).

Here, a novel method of primer-free FISH probe design that skips the pre-requisite for full-length 16S rRNA sequences was introduced. Besides elucidating the genomics and functional capabilities of the microbial community, the omics dataset (metagenomics and metatranscriptomics) from UPWRP was also used as input for the design of FISH probes. Bona fide FISH probes were subsequently designed from the V6 hypervariable region of the 16S rRNA from metatranscriptomics surveys, and integrated with FISH-FACS for the enrichment of the two novel bacterial taxa. Characterising the novel bacterial taxa has many benefits, and the direct implications include: (1) broadening our knowledge on the prokaryotic tree of life and (2) generating new reference genomes that can aid in taxonomic assignment of future metagenomes.

Acquiring knowledge about the identity and metabolic potential of novel organisms can potentially help to improve wastewater treatment, but this requires an in-depth study into the genomic functional potential and metabolic pathways of the novel organisms which is beyond the scope of this thesis.

1.3 Aims and scopes

The main aim of this thesis is to establish a reproducible and robust pipeline for the characterisation of targeted taxa of hitherto unclassified bacteria from the floccular sludge community. Characterisation of the unclassified taxa involved developing tools to elucidate their morphology, identity, phylogenetic affiliation, draft genomes and functional potential. A novel method of FISH probe design from omics dataset is one of the significant tools that were developed. Newly-designed FISH probes or RiboProbes would be used for the enrichment of the novel taxa through the process of FISH-FACS, and sorted cells would be processed for downstream genomic sequencing and analyses.

The thesis is organised as follows:

Chapter 2 is a literature review on the state-of-the-art molecular biology tools used to investigate microorganisms that play a vital role in wastewater treatment system.

Chapter 3 aims to demonstrate a novel method of FISH probe design from omics dataset. RiboProbe was validated on *Thauera* as a reference organism, through co-hybridisation experiments with a canonical probe that was designed through comparative sequence analysis. The validation was performed on an axenic culture, and subsequently in a mixed microbial community of activated sludge. In addition, RiboProbe was evaluated through a series of *in silico* models and experimentally with quantitative co-localisation assays. Results obtained from the validation of RiboProbe through co-hybridisation experiments and *in silico* evaluation of RiboProbe are being prepared in a manuscript, entitled: "Designing next generation FISH probes from omics dataset for targeted visualisation and enrichment".

Chapter 4 consists of two objectives. The first objective is to optimise a methodology that couples RiboProbe to flow cytometry (FISH-FACS) for selective enrichment of the target taxon. This was achieved through outlining of a detailed strategy that described: (1) configuration of a FACS sorter that supports small-particle sorting; (2) hybridisation and contamination controls and (3) sterilisation procedures to minimise contamination for downstream MDA amplification of sorted cells. The second objective is to compare the environmental specificity of RiboProbe with canonical probe for the target genus *Thauera* in activated sludge using the optimised FISH-FACS protocol. Details describing the pipeline for effective FISH-FACS sorting of *Thauera* as a positive control from an axenic culture and floccular sludge community are being prepared in a manuscript, entitled: “Designing next generation FISH probes from omics dataset for targeted visualisation and enrichment”.

Chapter 5 aims to apply the methodologies developed in the previous chapters to characterise two novel bacterial taxa present in the floccular sludge community from UPWRP. FISH-FACS was used as a tool for enrichment, followed by phylogenetic assignment using the 16S rRNA gene and a concatenation of phylogenetic marker genes, and genome recovery of the enriched population. Completeness and contamination of the draft genomes were evaluated using two different approaches: essential single-copy genes and lineage-specific marker genes. A cursory analysis of the metabolic potential of the novel taxa is also provided. In addition, FISH was used as a visualisation tool to contrast the structure of a novel species classified under the genus *Haliangium* from other well-established *Haliangium* denitrifiers present in activated sludge. Finally, microscopy, FACS sorting and RiboTagger 16S rRNA analysis were employed to investigate the ability of different *Haliangium* species to produce myxospores through sorting gates constructed with different light-scattering properties. Results from the methods describing genome recovery of the novel taxa, phylogenetic analysis of the 16S rRNA gene and phylogenomics of the draft genomes are being prepared in a manuscript, entitled: “Designing next generation FISH probes from omics dataset for targeted visualisation and enrichment”. Results from the discovery of different *Haliangium* species possessing the propensity to form different spatial structure and to

produce myxospores are being prepared in a manuscript, entitled: “Deciphering different spatial structures of *Haliangium* species in activated sludge with FISH and cell sorting”.

Chapter 6 provides a conclusion and perspective to the thesis and discusses future directions to be taken. An effort to perform a detailed functional analysis of the genomes of the two novel taxa is currently being undertaken, and the results are being prepared in a manuscript, entitled: “Genome analysis of an unclassified and rare bacterial taxon in activated sludge recovered with next generation FISH probes”.

Chapter 2 Introduction

2.1 Activated sludge process for biological wastewater treatment

Most municipal sewage treatment plants, such as the Ulu Pandan Water Reclamation Plant (UPWRP) in Singapore relies on activated sludge biotechnological process for the biological treatment of domestic sewage, prior to discharge into surface water bodies or for possible reuse in the industries (Qin *et al.*, 2009). The activated sludge process is designed with engineering tools and parameters to create an engineered ecosystem for active floccular microbial communities in an aerobic aquatic environment. Floccular microbial communities form a component of the biomass of an activated sludge floc, and metabolic capabilities of the microorganisms are responsible for the major bioprocesses that lead to purification of polluted water in wastewater operations (Eikelboom, 2000).

Alongside the study of hydrology and flow-dynamics in wastewater reactors, the study of microorganisms in activated sludge flocs has been a central theme in understanding the bioprocesses that drive biological wastewater treatment. An activated sludge floc is a suspended aggregation of microorganisms that are spatially organised in an extracellular polymeric substance (EPS) often called the matrix, which includes organic fibres, DNA, proteins and inorganic particulates from the influent wastewater (Frølund *et al.*, 1996; Laspidou and Rittmann, 2002) (Figure 2-1).

Figure 2-1: An illustration of the structure of an activated sludge floc.

Mixed microbial community forms a major element of the floc. EPS matrix is responsible for holding components of the flocs together in a spatial arrangement. Image is reprinted from Encyclopedia of Ecology, Pell M and Wörman A, Biological Wastewater Treatment Systems, 426-441, 2011, with permission from Elsevier.

UPWRP was designed for the biological elimination of organic carbon and excess of nitrogen from influent wastewater. To achieve this, the reactor configuration for biological wastewater treatment at UPWRP was designed to continuously subject floccular microbial community to alternating phases of anoxic and aerobic conditions in respective compartmentalized basins (Figure 2-2).

Figure 2-2: Reactor configuration for biological wastewater treatment process at UPWRP. Periodic exposure of activated sludge flocs to alternating phases of anoxic and aerobic conditions in different basins results in the efficient elimination of nutrients and organic carbon from influent wastewater.

The alternate cycling of the sludge in different redox conditions in different reactors creates a selective pressure that enriches for functional microbial groups in activated sludge flocs (e.g. nitrifiers) that aid in the removal of inorganic nutrients and the degradation of complex organic pollutants (Tchobanoglous *et al.*, 2003; Falcioni *et al.*, 2006; Seviour and Nielsen, 2010). Therefore, discerning the relationship between structure of the microbial community, environmental parameters that alter the structure and functional performances of bioprocesses in wastewater biological treatment processes is of paramount importance (Onuki *et al.*, 2000; Biesterfeld *et al.*; Tchobanoglous *et al.*, 2003). Structure of an activated sludge floc is defined by the composition, taxonomy, phylogeny, and three-dimensional spatial arrangement of the microorganisms in a single floc. Composition refers to the richness: number of taxa and evenness: abundance of the

taxa. Structure of the community can be further categorised into two scientific disciplines: taxonomy and phylogeny. Nomenclature, description and classification of microorganisms are established in the field of taxonomy. Microorganisms are taxonomically classified on the basis of their evolutionary relationship, and this is the field of phylogeny. Spatial distribution is the arrangement of microorganisms in relation with each other in the floc.

Wastewater treatment plants provide the perfect platform to understand the structure-function relationship of activated sludge flocs because of the well-defined ecosystem where physiochemical parameters: temperature, pH, nutrient levels and SRT are routinely measured. These metadata provide additional information to fundamental microbiological research undertaken to extend further insights into the structure-function relationship of floccular sludge communities (Daims *et al.*, 2006). The ultimate goal of investigating the structure-function relationship of activated sludge flocs is to create a framework for knowledge-based management of key microbial groups that drive these bioprocesses. For example, polyphosphate accumulating organisms (PAOs) are a functional group of organisms that catalyse the removal of phosphorus from influent wastewater through the uptake and subsequent accumulation of phosphate in their biomass as poly- β -hydroxyalkanoates (PHA). The understanding of the ecophysiology of PAOs have led to the creation of advanced configurations in wastewater treatment system: enhanced biological phosphorus removal (EBPR) plants equipped with alternating cycles of anaerobic and aerobic conditions (Mino and Satoh, 2006). EBPR plants are operated with engineering parameters that favour the growth of PAOs over other types of competing bacteria (Oehmen *et al.*, 2010). In the anaerobic phase, fermentation process produces volatile fatty acids that are assimilated by PAOs into intracellular poly- β -hydroxyalkanoates (PHAs). In the aerobic phase, oxygen serves as an electron acceptor for the oxidation of PHAs, and this results in the uptake of orthophosphate into the cells (Forbes *et al.*, 2009). EBPR is currently an accepted configuration in full-scale wastewater plants that supports the growth of PAOs for the biological elimination of phosphate from wastewater.

2.2 Unclassified microbial taxa present in floccular sludge community drives biological wastewater treatment under tropical conditions

As structure of the floccular sludge community is influenced by physicochemical variables such as temperature, structure of the floccular microbial community of a tropical wastewater treatment plant (e.g. UPWRP) is different from those described from temperate conditions. Hitherto, the only available information to date that extensively describes the floccular sludge community structure in a wastewater treatment plant based in Singapore was been obtained from whole community shotgun metagenomics and metatranscriptomics of the floccular microbial community from UPWRP (Kjelleberg *et al.*, unpublished; Law *et al.*, 2016). With advancement in sequencing technology, whole community shotgun sequencing has provided a wide array of information pertaining to the identity and community composition of wastewater microorganisms.

In the study undertaken by Kjelleberg *et al.* (unpublished), metagenomics and metatranscriptomics shotgun sequencing to complete saturation of the floccular microbial community in UPWRP were performed. This ultra-deep sequencing represented one of the deepest sequencing efforts to date, where 275 billion genomic DNA bases were sequenced and the genomic DNA saturated sequence diversity was predicted to be 95.2%. Their RNA analyses revealed the presence of up to 26,000 OTUs that corresponded to approximately 1,985 genera of bacteria, with a large fraction of the community comprising of unclassified bacterial taxa - 84.4% of the community could not be taxonomically identified using the SILVA database (Figure 2-3).

Figure 2-3: Relative abundance of floccular microbial community in UPWRP as estimated by RiboTagger 16S rRNA analysis.

Y-axis depicts OTU abundance from DNA sequencing reads of the floccular community; X-axis depicts accumulative abundance of OTUs. OTUs with taxonomic affiliation are demarcated in red; OTUs lacking a homology to 16S rRNA references are demarcated in blue. A large fraction of the floccular sludge community at UPWRP could not be taxonomically classified. Image is reproduced from Kjelleberg *et al.* (unpublished).

The ultra-deep saturated sequencing allowed the composition of the floccular microbial community in UPWRP to be better defined in a manner that exceeded the description of floccular communities in other full-scale wastewater treatment plants. For instance, a total of 275 Gbp of genomic DNA bases were sequenced by Kjelleberg *et al.* (unpublished) as compared to 2.4 Gbp that were sequenced by Ye & Zhang (2013); 57 Gbp that were sequenced by Albertsen *et al.* (2013a) and 8 Gbp that were sequenced by Albertsen *et al.* (2013b). Despite the massive sequencing efforts, there are still knowledge gaps that could be further exploited. For instance, a large fraction of the floccular sludge community could not be taxonomically assigned. This was especially evident in the most abundant community members where 37 out of 50 most abundant OTUs were taxonomically unidentified at the taxonomic ranks of species to genus. The high relative abundance of unclassified novel taxa of microorganisms suggests that biological wastewater treatment at UPWRP may largely be driven by yet unclassified/novel groups of microorganisms. From an ecological viewpoint, characterising these groups of unclassified microbial taxa would provide novel insights into their phylogenetic relationship with other microorganisms, and a potential putative genus function could be assigned to them. From the perspective of wastewater management, researching into the unclassified microbial taxa would

shed light on the mechanisms in which they drive biological wastewater treatment. This leads to a better understanding of the key activated sludge bioprocesses and could potentially create framework for better management of the uncharacterised taxa for better process performance.

Discovery of a large proportion of unclassified OTUs at UPWRP is in stark contrast with the recent study by Saunders *et al.* (2015), where majority of the top 50 genus-level OTUs of the abundant core floccular community frequently observed in most Danish wastewater treatment plants could be taxonomically annotated. Only 3 genera of the top 50 genus-level OTUs lacked homology to 16S rRNA references in the database. Although different methods were employed to observe the most abundant genus-level OTUs in both studies, both studies involved the analysis of the 16S rRNA as a consensus phylogenetic marker. Furthermore, assessment of the richness of the floccular microbial community in UPWRP challenge the claim made by Saunders *et al.* (2015) that an abundant core floccular community might exist in wastewater treatment plants globally, as many of the genus-level OTUs present in the Danish wastewater plants could also be detected in other countries (Zhang *et al.*, 2012). Majority of the top 50 annotated genus-level OTUs detected in Danish wastewater plants were present at a much lower frequency in UPWRP. Further investigations are necessary to understand the selective pressure that cause the diversity of floccular sludge community in UPWRP to differ from those present in other wastewater treatment plants.

Significant efforts have been invested into microorganisms that contribute to biogeochemical processes such as nitrogen removal and phosphorus removal in wastewater to circumvent eutrophication (Seviour *et al.*, 2003). *Nitrosomonas* has been recognized as an ammonia-oxidizing bacterial (AOB) genus that contributes significantly to the nitrification process in wastewater (Nielsen *et al.*, 2009). Despite its prominent role in the nitrification process, it was ranked 539th in relative abundance in the UPWRP's floccular sludge community. Therefore, the contribution of *Nitrosomonas* to the nitrification process in UPWRP is an enigma due to its low abundance. There lies a possibility that the nitrification process might be contributed by members of uncharacterised

taxa present in higher abundance than *Nitrosomonas*, and this warrants further investigation into the uncharacterised taxa. Quite often, it is the dominant taxa that are hypothesised to be responsible for catalysing specific activated sludge bioprocess. For instance, the discovery of high abundance (30%) of *Tetrasphaera* in several EBPR treatment plants in Denmark that actively uptake orthophosphate under aerobic conditions have challenged the putative role of *Candidatus Accumulibacter* as the key functional microbial group responsible for phosphate removal (Nguyen *et al.*, 2011).

In addition, less than 1% of the sequencing data generated from UPWRP could be mapped to annotated complete microbial genomes in the RefSeq database (Pruitt *et al.*, 2012) despite the ultra-deep saturated sequencing efforts. None of the assembled contigs could be assembled into complete genomes, and this could be attributed to the current limitations of metagenomics (Albertsen *et al.*, 2013a). Inability to perform full genome recovery from a heterogeneous community was corroborated by Law *et al.* (2016), where they could only achieve 74.4% gene recovery from a metagenome assembly of *Accumulibacter* from activated sludge. Full genome recovery still remained a limitation despite the relatively high abundance (3.44%) of *Accumulibacter* at the point of sampling, coupled with a sequencing depth that should theoretically allow full genome recovery. Genome recovery through a different approach, such as the enrichment of a target taxon to a sufficiently high abundance should permit close to full genome recovery and yield insights into its functional potential. For instance, enrichment of *Candidatus Kuenenia stuttgartiensis* in a laboratory-scale reactor to 73% of the total biomass allowed the almost complete genome (>98%) to be recovered and genes responsible for anammox to be identified (Strous *et al.*, 2006). However, it is not possible to enrich for the uncharacterised taxa through traditional enrichment systems due to a lack of knowledge about their ecophysiology.

Before venturing into the characterisation of unclassified microbial taxa, seemingly well-characterised groups of organisms should also entail further investigation given the low percentage of reads mapping to reference genomes. Another reason for further investigating seemingly well-characterised groups of organisms is the fact that many abundant genera in the floccular sludge community (e.g. *Nitrospira* and *Accumulibacter*) have shown to be involved in other bioprocesses. Despite a putative genus function being assigned to *Nitrospira* (nitrogen transformation) and *Accumulibacter* (phosphate removal), Kjelleberg *et al.* (unpublished) reported that transcriptomics analysis of the floccular sludge community showed the involvement of these genera in xenobiotic degradation.

Genome recovery and subsequent genome annotation would shed light on the potential ecophysiology of members of the floccular sludge community, yielding more functional insights than putative functional annotation at the genus level. Last but not least, there is a lack of phylogenetic information on the OTUs present in UPWRP. This was due to the short length of sequencing reads which limits the taxonomic resolution of the 16S rRNA gene and subsequent taxonomic assignment of these short reads (Schloss *et al.*, 2016). Sequencing of the full-length 16S rRNA gene will yield more phylogenetic information on the unclassified OTUs and will provide better resolution that can resolve bias associated with sequencing individual variable regions (Klindworth *et al.*, 2013). Therefore, a method complementary with deep genomic sequencing is necessary to aid future phylogenetic assignment of OTUs. A literature review on existing technologies that produce full-length 16S rRNA sequences is described in Section 2.3.3.

2.3 Approaches to identification and classification of prokaryotes

2.3.1 Culture-dependent approach

Culture-dependent approach is associated with the imperative prerequisite of isolating single-cell microorganism into axenic culture, followed by subsequent assays which aim to characterise the phenotypic, biochemical and metabolic properties of the microorganism (Sanz and Köchling, 2007). However, it has been estimated that 99% of microorganisms from environmental samples are yet to be made into axenic collections due to inadequate knowledge about their physiological requirements and symbiotic relationships with other community members; preponderance of microbial biodiversity in environmental and engineered ecosystems are represented by uncultured microorganisms (Torsvik *et al.*, 1990; Amann *et al.*, 1995; Liesack *et al.*, 1997; Rappé and Giovannoni, 2003; Acinas *et al.*, 2004; Stevenson *et al.*, 2004; Lasken and McLean, 2014). As such, investigation of the structure of microbial communities in environmental and engineered samples becomes skewed when detection of microorganisms through cultivation-dependent techniques does not depict an accurate picture of the community structure (Wagner *et al.*, 1993; Ludwig *et al.*, 2004). Inherent bias associated with cultivation-dependent methods was exemplified by the fact that in the year 1998, 65% of worldwide microbiology research was focused only on 8 bacterial species, which merely represented 3 bacterial lineages (Galvez *et al.*, 1998).

In addition, conventional cultivation-dependent techniques severely limit our understanding of the biology, population dynamics and ecological functions of microorganisms that play an active role in wastewater treatment. For instance, *Nitrobacter* species were believed to be the main nitrite-oxidizing bacterial group in activated sludge because of the easier task of isolation through cultivation-dependent techniques and its faster growth kinetics (Bock *et al.*, 1990; Sorokin *et al.*, 1998). This assumption was radically challenged when cultivation-independent techniques showed *Nitrospira* species to be more dominant in abundance and to play a more vital role in nitrite-oxidation than *Nitrobacter* species in activated sludge (Juretschko *et al.*, 1998; Daims *et al.*, 2001). Elucidating the contribution of *Nitrospira* species to nitrite-oxidation through cultivation-

dependent techniques was not because of the laborious effort of isolating an axenic culture due to its slow growth rate, sensitivity to changes in the environmental conditions and growth competition from *Nitrobacter* under laboratory conditions (Fujitani *et al.*, 2014). The example above exemplified the fact that culture-dependent methods are biased towards the proliferation of certain groups of microorganisms and this has skewed our perception of the “true” dynamics and composition of microbial communities. There was an urgent need for a paradigm shift away from cultivation-dependent techniques to truly decipher the ecology of wastewater microorganisms with minimal perturbation to the communities (Amann *et al.*, 1995; Lasken, 2012).

2.3.2 Culture-independent approach

16S rRNA gene as a phylogenetic marker

Recognizing existing limitations of culture-dependent methods, many cultivation-independent technologies evolved over the past two decades have analysed the small ribosomal subunit (SSU RNA) - the 16S rRNA gene (Woese and Fox, 1977) in particular for phylogeny and taxonomic classification of prokaryotes (Pace, 1997). Currently, it is the most widely-accepted phylogenetic marker for various reasons. Being an element of the SSU RNA, it is ubiquitous with a high copy number: an average of 10^3 to 10^5 ribosomes per cell in prokaryotic cells, thus making it easy and suitable for detection (Woese *et al.*, 1975; Sanz and Köchling, 2007). Although the 16S rRNA gene is a highly evolutionary conserved molecule, hypervariable domains are present and random mutations in the hypervariable domains reflect the evolution of prokaryotes (Woese, 1987; Janda and Abbott, 2007). The entire length of the 16S rRNA gene is approximately 1500 bp long, with differential evolution rates in the nine hypervariable regions that provide taxonomic classification of prokaryotes from the domain to the species taxonomic rank (Van de Peer *et al.*, 1996; Ludwig *et al.*, 1998b) (Figure 2-4). In addition, helices present in the secondary structure of the 16S rRNA catalyse more than 50% base pairing of the residues, and this translates to higher confidence in positional homology during comparative sequence analysis of 16S rRNA gene sequences (Bergey, 2005).

Figure 2-4: A model representing the higher-order secondary structure of prokaryotic 16S rRNA gene. Hypervariable regions are located adjacent to conserved regions. Location of hypervariable regions (V1–V9) is demarcated in the red boxes. Colour represents fragments that contain different combination of variable regions. Image is adapted from Nature Reviews Microbiology, Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences, 635-645, 2014, with permission from Nature Publishing Group.

Convenience associated with Sanger sequencing of PCR-amplified 16S rRNA genes have resulted in the large deposition of sequences in databases, where the number of 16S rRNA sequences currently exceeds 5 million (SILVA 126 Parc database, May 2016). The rate of discovery of new bacterial species is predicted to be approximately $4 \times 10^4 \pm 0.8 \times 10^4$ per year (Yarza *et al.*, 2014). Comprehensiveness of the 16S rRNA database is another alluring reason for the global acceptance of 16S rRNA gene as a phylogenetic marker. Classification of uncultured microorganisms is important as they represent the bulk of the richness in complex microbial communities. Currently, the number of sequences from uncultured microorganisms has already exceeded those obtained from axenic cultures in 16S rRNA databases. Microorganisms are phylogenetically assigned to their respective taxa groups through comparative sequence analysis of the 16S rRNA gene sequences

of newly-discovered microorganisms with curated 16S rRNA databases (Erko and Ebers, 2006). Recently, a statistical 16S rRNA gene sequence identity threshold boundary was proposed for the taxonomic ranks of cultured and uncultured prokaryotes. A threshold sequence identity of $\leq 98.7\%$ for species, $\leq 94.5\%$ for genus, $\leq 86.5\%$ for family, $\leq 82.0\%$ for order, $\leq 78.5\%$ for classes and $\leq 75.0\%$ for phyla would place two units in different taxa groups within the same taxonomic rank (Yarza *et al.*, 2014).

Full-cycle rRNA approach

Ever since 16S rRNA was widely accepted as a phylogenetic marker for the identification of microorganisms, the gold-standard strategy in exploring the microbial diversity and phylogeny of uncultured microorganisms in environmental samples was coined as the 'full-cycle rRNA approach' (Amann *et al.*, 1995). This approach involves the sequential combination of cultivation-independent molecular biology methods: (1) direct extraction of nucleic acids from environmental samples; (2) polymerase chain reaction (PCR) amplification of 16S rDNA using conserved bacterial primers; (3) construction of 16S rDNA clone libraries to separate mixed copies of 16S rDNA for subsequent Sanger sequencing; (4) Sanger sequencing of individual clones and (5) comparative 16S rDNA sequence analysis where sequences of the 16S rRNA gene can be used for construction of phylogenetic trees or design of fluorescence *in situ* hybridisation (FISH) probes (Lane *et al.*, 1985; Olsen *et al.*, 1986; Amann *et al.*, 1995; Pace, 1997; Snaird *et al.*, 1999) (Figure 2-5).

Figure 2-5: Outline of the steps in full-cycle rRNA approach adopted for microbial ecology studies.

The laborious nature of constructing numerous clones from 16S rDNA clone libraries cannot be avoided as a mixture of rDNA template sequences cannot be used for Sanger sequencing, and increasing the number of clone constructs yields a more accurate description of the diversity present in the sample. To illustrate the tedious nature of cloning, Nguyen *et al.* (2011) sequenced a total of 189 clones to determine the phylogenetic relationship of *Tetrasphaera* in EBPR wastewater plants and to design *Tetrasphaera*-specific FISH probe. In a separate study, Zhang *et al.* (2011) screened 135 clones from a clone library to analyse the bacterial composition of a quinolone-degrading bioreactor.

The main drawbacks to the full-cycle rRNA approach is the inherently tedious and costly efforts of constructing and performing paired end sequencing of hundreds of clones. Monitoring temporal dynamic shift in microbial community using this approach is difficult as it requires sequencing hundreds of clones per sample per time point (Giovannoni, 1990; Sanz and Köchling, 2007). Furthermore, minority members of the microbial community that contribute to important ecological functions might not be reflected during the sequencing of the clone libraries. Only the

more abundant organisms in the sample would be captured through the clone libraries. Low sequencing coverage of the full-cycle rRNA approach highlights the inaccuracy of profiling the microbial diversity present in the sample. To illustrate this, discrepancy in microbial community analysis of a quinolone-degrading bioreactor using two different methods: clone library analysis and pyrosequencing highlighted the difference in sequencing coverage (Zhang *et al.*, 2011). Clone library analysis identified fewer numbers of OTUs, and failed to elucidate phylotypes that were detected by pyrosequencing. Other drawbacks of the full-cycle rRNA approach include: (1) PCR amplification bias of microorganisms present in low abundance (Muyzer *et al.*, 1993; Lee *et al.*, 1996); (2) bacterial conserved primers not being able to amplify and capture the full diversity present in a given sample (Ravel, 2012; Klindworth *et al.*, 2013); (3) uneven nucleic acid extraction from different microorganisms (Picard *et al.*, 1992) and (4) differential cloning efficiency (Rainey *et al.*, 1994).

Although cloning of the 16S rRNA gene provides more precise phylogenetic information than other methods due to sequencing of the entire gene, analysis of phylogenetic relationship between microorganisms is limited to only a specific genomic locus: 16S rRNA which is not linked to other metabolic genes present in the genome (Faust and Raes, 2012). Other gene catalogues in the genome can be obtained through the use of metagenomics and one such benefit is the elucidation of other phylogenetic markers that can reveal phylogenetic relationships that might not have been previously revealed by 16S rRNA gene comparison (Podar *et al.*, 2009). For instance, a phylogenetic relationship between the phylum *Acidobacteria* and *Deltaproteobacteria* was recently established through comparative analysis of concatenated conserved marker genes (Ciccarelli *et al.*, 2006). Furthermore, the 16S rRNA gene might not always provide the best phylogenetic resolution, as demonstrated in phylogenetic studies where the *ppk1* gene is shown to be better in distinguishing the various sub-lineages of *Accumulibacter* (He *et al.*, 2007; Albertsen *et al.*, 2016).

Next generation sequencing (NGS)

Low-throughput Sanger sequencing and its association with under-representation of species richness in complex community has seen a paradigm shift to higher-throughput NGS platforms which produce shorter reads and gigabases of sequencing data at a lower cost (Youssef *et al.*, 2009). NGS reduces the need for molecular analysis of numerous clones from clone libraries and it overcomes the coverage bias associated with previous Sanger sequencing. The two most frequent methods used on NGS platforms are 16S rRNA gene amplicon sequencing and whole community shotgun surveys. 16S rRNA gene amplicon sequencing is frequently utilised for profiling the diversity of microbial communities, and it involves the sequencing of an individual variable region or a combination of variable regions (Caporaso *et al.*, 2012). However, correlating functional aspects of the community with taxonomic context is difficult in 16S rRNA gene amplicon sequencing as the focus is on specific 16S rRNA gene that have been amplified with conserved bacterial primers (Faust and Raes, 2012).

An increasing number of studies are incorporating whole community shotgun surveys: metagenomics or metatranscriptomics to characterise the ecology of the microbial community. Whole community shotgun surveys involve high-throughput shotgun sequencing of the nucleic acid that has been directly extracted from environmental samples (Riesenfeld *et al.*, 2004; Tringe and Rubin, 2005; Podar *et al.*, 2009). The output of whole community shotgun surveys is the generation of massive number of short sequencing reads – both phylogenetic marker genes and metabolic genes from multiple loci across the genomes of different cells. The advantage of unbiased genomic shotgun sequencing is the ability to: (1) obtain a comprehensive profile of the community due to high-throughput sequencing technology; (2) identify the functional potential of the community due to establishment of the phylotype-functional linkage of the members of the community and (3) perform genome recovery through the extraction of draft genomes with genomic binning – a process which groups phylogenetically-related metagenomic contigs on the basis of their differential abundance across multiple-samples (Albertsen *et al.*, 2013a) or lineage-specific signatures such as tetranucleotide frequency and GC content (Sangwan *et al.*, 2016).

For instance, draft genomes of *Candidatus Accumulibacter phosphatis*- responsible for phosphorus removal (García Martín *et al.*, 2006) and *Candidatus Nitrospira defluvii*- responsible of nitrite removal (Lücker *et al.*, 2010) were examples of taxa identified to be responsible for important bioprocesses after genome recovery.

Fluorescence *in situ* hybridisation (FISH)

Investigation of the identity, morphology, spatial arrangement and abundance of microorganisms in a microcosm are common themes that need to be addressed in microbial ecology. Since the inception of FISH as a culture-independent technique to study these common themes by DeLong *et al.* (1989), FISH has been widely used on samples from different complex ecosystems due to its high sensitivity and speed. FISH involves *in situ* hybridisation of oligonucleotide probes with typical read length of 15–25 bp, labelled with fluorophores to specific target signature sequences of 16S rRNA molecules within prokaryotic cells (DeLong *et al.*, 1989; Amann *et al.*, 1995; Wagner *et al.*, 2003; Thiele *et al.*, 2010). An advantage of FISH is the ability to identify and visualise the morphology of target group of individual cells in mixed assemblages, and to investigate the three-dimensional spatial distribution of cells with respect to its symbiotic neighbouring cells and its environment. This is performed with the use of either an epifluorescence or confocal laser scanning microscope after hybridisation of the sample with FISH probes under stringent hybridisation and washing conditions (Giovannoni *et al.*, 1988; DeLong *et al.*, 1989; Amann *et al.*, 1990b). For example, candidate division TM7 was observed to possess a multitude of morphotypes in full-scale floccular microbial communities, with the filamentous morphology being the most dominant among the TM7 division, as targeted by TM7-specific FISH probes (Hugenholtz *et al.*, 2001). In a separate study, FISH probes aided in the visualisation of *Candidatus Magnospira bakii* as a novel bacteria species, which possessed a corkscrew-shaped filamentous morphology and was often localised in the deeper segments of the floccular sludge (Snaird *et al.*, 1999).

Accurately quantifying the abundance of microorganisms is important in monitoring the proper functioning of a wastewater treatment plant that relies on the activated sludge system. For instance, the problem of sludge bulking in wastewater plant has a strong correlation to the high abundance of filamentous bacteria (De Los Reyes *et al.*, 1998; Van Der Waarde *et al.*, 1998). Using quantitative FISH to track the early stages of filamentous bacteria growth can help to circumvent the problem of sludge bulking. While numerous studies have incorporated metagenomics for quantifying purposes, the result is largely influenced by DNA extraction bias (Albertsen *et al.*, 2012) and sequencing coverage (Rodriguez-R and Konstantinidis, 2014). Quantitative FISH is independent of DNA extraction, provides a faster method to quantification - does not require library preparation and genomic sequencing - and can be used to complement quantitative metagenomic results. Quantitative FISH is frequently used to accurately monitor the population dynamics of target groups of organisms to gain insight into potential ecological functions. For example, the annamox activity was believed to be carried out by *Candidatus Kuenenia stuttgartiensis* in a laboratory-scale trickling filter biofilm reactor due to the dominating presence of the novel bacterial group under anaerobic conditions as estimated by quantitative FISH (Schmid *et al.*, 2000). Besides the ability for accurate quantification, FISH probes offer the advantage of fluorescent-labelling for subsequent isolation and characterisation of target organisms via methods such as Fluorescence Activated Cell Sorting (FACS).

The foundation of FISH probe design relies on the use of 16S rRNA molecule, which contains highly conserved and hypervariable regions (Woese, 1987; Thiele *et al.*, 2010). Highly conserved regions allow for the design of conserved primers (e.g. 27F, 1492R) that targets majority of the bacterial taxa, while the hypervariable regions allows for the classification of bacterial species (Lasken, 2012). Depending on the degree of evolutionary variability of the variable region, specificity of the FISH probes can be manipulated to the desired phylogenetic depth between the species and kingdom levels (Stahl D. A. and Amann, 1991). Published FISH probes that have been designed and widely used can be accessed online via ProbeBase (Loy *et al.*, 2003). Commonly used FISH probes for the identification of microorganisms frequently present in activated sludge can be found in the

“FISH handbook for biological wastewater treatment” (Nielsen *et al.*, 2009). *De novo* probe design is necessary for an organism whose probe sequence is not available in the FISH probe database. Hence, probe design remains a bottleneck for scientists because *a priori* information on the full-length 16S rRNA sequence of the target taxon is often lacking.

2.3.3 Methods of FISH probe design

Comparative sequence analysis

With increasing demands for FISH to be integrated into microbial ecology, software tools such as the ARB software were developed to guide users in the selection and evaluation of target sites on the 16S rRNA sequence (Ludwig *et al.*, 2004). FISH probe design is a component of the full-cycle rRNA approach, where probes are designed from comparative sequence analysis of the target taxon and its phylogenetically related 16S rDNA sequences (Figure 2-5). Taxon-specific probe designed through comparative sequence analysis target a defined group of phylogenetically related organisms (Ludwig *et al.*, 2004). This is achieved through the search for potential probe target sites in homologous regions of the 16S rDNA in target group of organisms. In addition, potential mismatches of FISH probe with non-target organisms containing similar target sites are evaluated *in silico*. Specificity of probe hybridisation is influenced by the strength of hybridisation, which in turn is affected by the position and type of mismatches to non-target organisms (Yilmaz *et al.*, 2008). For example, mismatches in the centre of the probe binding site have an enhanced destabilising effect as compared to mismatches near the extreme ends of the probe. The typical output of probe design is a list of potential FISH probes, which are ranked in terms of thermodynamic binding between the FISH probe and the target sites, with and without mismatches (McIlroy *et al.*, 2011; Yilmaz *et al.*, 2011).

An exponential expansion of the number of 16S rRNA sequences in databases has led to greater sequence diversity, and this increases the number of potential cross-hybridisations to non-target organisms for a given set of probes. Thus, probe design is proving to be more challenging in locating suitable probe target sites for an intended target taxon. To circumvent this problem, probe design can be modified to reduce probe coverage such that it searches for target sites in a

subset of the original target group. Alternatively, specificity of probe design can be reduced to encompass members that is outside of the original target taxon, but having some degree of phylogenetic relatedness to the target taxon (Ludwig *et al.*, 2004). For the latter scenario, the use of multiple independent probes with overlapping specificity in a 'multiple nested probe concept' can increase the reliability of detection for the intended target group of organisms (Rudolf and Schleifer, 2001).

Two pre-requisites are required for optimal probe design: (1) manual curation of the alignment between 16S rRNA sequences of the target taxa and its closest relatives and (2) good quality 16S rRNA sequences that are full-length (~1500 bp) or near full-length sequences (≥ 1200 bp). Full-length or near full-length sequences provide a more precise taxonomic resolution compared to short fragments generated from amplicon sequencing of the hypervariable regions, and is better at resolving closely-related species (Singer *et al.*, 2016). Therefore, FISH probes are designed with higher confidence for their intended target taxa. Furthermore, maximum sequence coverage associated with full-length or near full-length sequences allows probe designers to locate multiple potential probe sites during sequence analysis (Hugenholtz *et al.*, 2002). This provides more option of designing multiple probes hybridizing to the same target taxa, which subsequently increases the specificity for positive identification of the target taxon (Ludwig *et al.*, 1998a). Even though the concept of comparative sequence analysis was devised more than 20 years ago, it remains the gold standard of designing FISH probe and this is reflected in recent publications (Lücker *et al.*, 2014; Albertsen *et al.*, 2016; Kindaichi *et al.*, 2016).

Probe design from next generation sequencing (NGS) platforms

Construction of clone libraries and Sanger sequencing were the gold standard of producing full-length 16S rRNA sequences for comparative sequence analysis, but drawbacks of this method were the laborious nature and extravagant cost of sequencing. Amplicon sequencing is a cheaper and rapid method of surveying the microbial diversity of environmental samples using NGS platforms such as the Illumina platform (Saunders *et al.*, 2015), where specific hypervariable

region (250-300 bp) of the 16S rRNA gene are sequenced. Although the short-read approach has proven invaluable in providing taxonomic assignments, the amplicon products are not compatible for FISH probe design through comparative sequence analysis. Generating full-length 16S rRNA sequences from NGS platforms for FISH probe design was an obstacle until recent improvements in sequencing library preparation and technology. For instance, the use of molecular dual tagging during sequencing library preparation on an Illumina MiSeq platform (Burke and Darling, 2016) or single molecule real time (SMRT) sequencing technology using the latest PacBio P6/C4 chemistry (Wagner *et al.*, 2016) have generated full-length 16S rRNA sequences from a mixed microbial community. Both technologies have the potential to produce reference sequences that are underrepresented in 16S rRNA databases, although this was not shown in both studies.

One drawback to both technologies is the need for PCR-based amplification of the 16S rRNA gene with broad bacterial primers (7F, 27F, 1391R, 1510R) prior to molecular tagging or sequencing. Approximately 10% of microbial diversity are missed out by broad bacterial primers (Eloe-Fadrosh *et al.*, 2016) due to primer mismatches (Lynch and Neufeld, 2015). Recognizing the bias associated with primers, Karst *et al.* (2016a) integrated reverse transcription of full-length 16S rRNA genes with synthetic long read sequencing by molecular tagging on the Illumina sequencing platforms. As a result, 30% of full-length bacterial 16S rRNA OTUs were reported to be novel. While the cutting-edge tool mentioned above is promising for FISH probe design from primer-free full-length 16S rRNA sequences, the focus is on 16S rRNA gene as a prokaryotic phylogenetic marker; other functional genes would be completely missed during NGS sequencing. It would be a cumbersome process for users who wish to perform potential functional genomic analysis of the community in addition to profiling the microbial diversity.

This problem could be circumvented with the RiboTagger software (Xie *et al.*, 2016), which identifies and extracts taxonomically useful short rRNA reads among the massive number of sequencing reads generated from shotgun surveys. Universal recognition profiles that represent conserved regions of the 16S rRNA are used by RiboTagger for the extraction of taxonomically-

relevant sequencing reads or RiboTags from different hypervariable regions: V4, V5, V6 and V7 of the 16S rRNA gene. The universal recognition profiles have shown to exude high sensitivity (>95%) against the SILVA database and a low false positive rate against RefSeq genomic fragments. For instance, the universal recognition profile used for community profiling at UPWRP was based on the V6 hypervariable region. The universal recognition profile adjacent to the V6 region matched 98.3% of sequences in SILVA database, thus highlighting the high sensitivity of universal recognition profile used in profiling the microbial diversity in UPWRP. Universal recognition profiles are located directly adjacent to RiboTags (Figure 2-6).

Figure 2-6: Extraction of taxonomically-relevant RiboTags using the universal recognition profiles with RiboTagger software.

RiboTags could be extracted from the V4-V7 region of the 16S rRNA gene. FISH probes were subsequently designed from RiboTags. Image is adapted from XRNA, 2017.

RiboTags extracted from the hypervariable regions have a read-length of 33 bp and can be defined as a signature sequence representing a single OTU. A longer representative sequence (81-84bp) of the RiboTag can also be extracted. The output of RiboTagger provides taxonomic classification and relative abundance of members of the microbial community based on annotation with a curated 16S rRNA database (e.g. SILVA or RDP) and the abundance of the respective RiboTags. RiboTagger provides rapid 16S whole community profiling from metagenomics or metatranscriptomics dataset, and was initially used in the diversity analysis of the floccular sludge community obtained from a tropical wastewater plant. Due to the fast computational time of RiboTagger in characterising the microbial community structure, RiboTagger was also used by Tan *et al.* (2014) in identifying the top 50 most dominant OTUs in an SBR, and correlating the microbial diversity with concentrations of N-acyl-homoserine-lactone (AHLs) and granulation measurements. In another study, Feng *et al.* (2017) used RiboTagger to show that Proteobacteria and Bacteroidetes were the major bacterial groups in both floccular and granular sludge communities of SBR that were susceptible to the predatory effect of *Bdellovibrio bacteriovorus*.

No studies thus far have described the design of FISH probes from short sequencing reads in omics datasets originally intended for taxonomic profiling or studying the functional capacity. While RiboTagger produces RiboTags that resemble the characteristics of a canonical FISH probe in terms of its read length and taxonomic affiliation, RiboTagger was used only as a tool for community structure analyses in the studies of Kjelleberg *et al.* (unpublished), Tan *et al.* (2014) and Feng *et al.* (2017). FISH probe design from the tag sequence of RiboTags is a great molecular biology tool to add to microbial ecological studies because it provides more contextual information such as visualisation, in addition to genomic and functional information provided by shotgun surveys. The importance of FISH probe design from short sequencing reads is more apparent given the extensive integration of next generation sequencing strategies into microbial ecology studies. Differences in method of probe design between RiboTagger and the gold standard of comparative sequence analysis is presented in Table 2-1.

Table 2-1: Differences in probe design between RiboTagger and comparative sequence analysis

Methods of probe design		
	Comparative sequence analysis	RiboTagger
Input	Full-length (1500 bp) or near full-length sequences (≥1200 bp)	RiboTags (≤33 bp) from metagenomics or metatranscriptomics dataset
Method of probe design	Comparative sequence alignment of 16S rRNA genes from the target and non-target taxa to search for potential probe target sites	Universal recognition profiles to extract sequencing reads with that correspond to hypervariable regions of the 16S rRNA
Selection of target sites in the 16S rRNA	Users specify regions for probe selection	V4-V7 hypervariable regions

2.3.4 Methods for enrichment of target bacterial taxa

Enrichment systems

Many studies in the field of wastewater research have attempted to identify bacterial taxa responsible for a specific bioprocess through enrichment systems. Enrichment systems are designed with the intent of enriching for certain bacterial taxa and selecting against others in the community (Hess *et al.*, 2011). Fluctuations in population dynamics in wastewater bioreactors are monitored with respect to time, and correlated with manipulations to the operational parameters and performance of the measured bioprocess. For example, in an enrichment study involving the use of bioreactor with simultaneous nitrification and denitrification (SND), alternating the C/N ratio of the influent feed changed the composition of different groups of bacteria, which were identified to be ammonia-oxidizing bacteria (AOB) and the nitrite-oxidizing bacteria (NOB). As AOB and NOB depended on the composition of nitrogen loading in the influent feed to carry out their metabolic activities, changes in the influent feed altered the composition of the two functional groups of bacteria (Xia *et al.*, 2008).

In another enrichment study, the involvement of *Acinetobacter* (Mudaly *et al.*, 2001) in the removal of phosphorus was brought into doubt because the dominating presence of *Rhodocyclus* species in the floccular biomass in an EBPR reactor was highly correlated with the high removal efficiency of phosphorus from wastewater (Pijuan *et al.*, 2004; Oehmen *et al.*, 2005). The functional role of *Rhodocyclus* species in phosphorus removal was also described by Crocetti *et al.* (2000), and a correlation between phosphorus removal and the high abundance of *Rhodocyclus* species could be established.

Both case studies involved monitoring the temporal variations in the microbial community structure before and after operational perturbations, with the predominant microbial group after the perturbation process often hypothesised to be responsible for the changes in bioprocesses. Enrichment systems coupled with molecular biology techniques have yielded great success in elucidating the structure-function relationship of floccular sludge community. However, low abundant microorganisms that contribute to an ecological functional role might not be identified through bioreactor enrichment systems if there are no prior knowledge about their ecophysiology. Therefore, an alternate approach is necessary for the enrichment of low abundant organisms.

Fluorescence Activated Cell Sorting (FACS)

FACS is frequently used for high-throughput fractionation and quantification of single cells from environmental samples (Amor *et al.*, 2002; Forster *et al.*, 2002; Ziglio *et al.*, 2002; Rinke *et al.*, 2013). FACS is performed with a flow cytometer, and droplets containing fluorescently-labelled cells are separated based on their specific fluorescent signal when cells passed through a beam of laser excitation light source (Podar *et al.*, 2009). With advances in multi-parametric measurements of flow cytometry, labelling cells with multiple fluorescent dyes permits the identification and characterisation of target cells in various physiological states (Czechowska *et al.*, 2008). Besides separation of cells on the basis of fluorescent signal, cells can be separated based on their phenotypic properties: size (side scatter) and granularity (forward scatter) (Robertson and Button, 1989; Steen, 1990; Park *et al.*, 2005).

Although most flow cytometers are not able to discriminate between different types of prokaryotic cells because of similarity in morphology and cell size, recent advancements in the field of flow cytometry that has allowed the detection of prokaryotes up to 0.1 μm in size should not be ignored (Gougoulas and Shaw, 2012). Through the precise construction of sorting gates, bacterial population can be accurately discriminated from sample and machine noise at a rapid pace of up to $\sim 75,000$ cells/s (Arnold and Lannigan, 2010). In addition to FACS, other tools were developed for the isolation of targeted specific cells from the microbial community: micromanipulation, microfluidics and microdroplet (Blainey, 2013). The ability to selectively enrich for target population at a high throughput and sorting speed gives flow cytometry an advantage over other types of cell-isolation methods, hence enrichment can be achieved in a shorter time frame (Rinke *et al.*, 2014). To illustrate this point, *Nitrospira* population of a floccular sludge community could only be selectively enriched to 80% of the total microorganism population with the stable operation of a fixed bed continuous feeding bioreactor after a period of 2 years; other non-target organisms were resistant to elimination. However, a pure *Nitrospira* micro-colony could be sorted using specific morphological characteristics of the micro-colonies with a flow cytometry within a single day (Fujitani *et al.*, 2014). Another advantage of FACS is the ability to sort in an extremely small volume of sample encompassing the target cells. Sieracki *et al.* (2005) had managed to sort a single target cell with 3-10 picolitre of sample solution surrounding the cell. Sorting in a small volume is essential in reducing the amount of contaminating DNA (e.g. eDNA) that is co-sorted together with the target cells, therefore simplifying the post-hoc genome analysis process in MDA experiments.

The combination of FISH and FACS (FISH-FACS) requires the use of FISH probes to fluorescently label target cells that contained the specific complementary hybridisation site on the 16S rRNA gene, followed by FACS to isolate fluorescent-labelled target cells from the other non-labelled cells. FISH-FACS has been used for various purposes. For instance, FISH-FACS was employed by Wallner *et al.* (1995) for the quantification of different groups of bacteria in activated sludge samples. In another study, FISH-FACS was also used for the enrichment of a target population from

bioreactor sludge samples; high level of enrichment of the target population was demonstrated by direct Sanger sequencing of the 16S rDNA PCR amplicon product from the sorted sample without the prior need for cloning (Yilmaz *et al.*, 2010b). However, it would be more conclusive if a diversity analysis through metagenomic sequencing was integrated into the workflow to determine if the sorted sample does indeed contain a pure population from the sorted samples. FISH-FACS has also been used for the isolation of target populations from other ecosystems. Kalyuzhanaya *et al.* (2006) have used a combination of FISH-FACS for the efficient separation of type I (59% enrichment) and type II Methanotroph (47.5% enrichment) populations from a lake sediment samples using type-I and type-II Methanotroph specific FISH probes.

FISH-FACS is frequently employed to enrich for rare members of the community whose presence might not be detected due to preferential representation of abundant organisms by current molecular methods (Rodriguez-R and Konstantinidis, 2014). This was demonstrated by Podar *et al.* (2007) where they isolated TM7 cells which were abundant at a frequency of 0.02% in the soil community prior to cell sorting, to an abundance of 89% after FISH-FACS. *De novo* genome assembly of low abundance organism is a challenge after metagenomic sequencing of a complex microbial community because of the presence of sequencing reads that belong to other organisms (Czechowska *et al.*, 2008; Müller and Nebe-Von-Caron, 2010). In most studies, only near-complete genomes belonging to the dominant taxa can be recovered. Through FACS enrichment, genome complexity of the community is reduced and genomic analysis can be vastly improved.

In addition to the use of FACS in recovering genomes of underrepresented members of the microbial community and performing phylogenetic analysis on the 16S rRNA gene or functional genes of the enriched population, FACS has been readily employed to analyse the physiological state of different sub-population of live cells. Most microorganisms in the natural environment exist as a biofilm and gradients within the biofilm (Stewart and Franklin, 2008) are responsible for creating sub-population of cells with differential genotypes and phenotypes. Functional heterogeneity in a biofilm can be dissected with single-cell tools, such as coupling FACS with

fluorescence dyes that can discern the different physiological states of the sub-populations. Commonly used fluorescence dyes with FACS include: (1) 5-Cyano-2,3-ditolyl tetrazolium chloride (CTC) dye to differentiate between respiring and non-respiring cells (Caro *et al.*, 2007); (2) SYTO9 and Propidium iodide (PI) dyes to differentiate between dead and live cells (Berney *et al.*, 2007) and (3) Hoechst 33342 to infer proliferation activity (Achilles *et al.*, 2007). In addition, gene expression between the different sorted sub-populations can be compared through mRNA or protein expression profiling (Heine *et al.*, 2009).

There are many challenges associated with FISH-FACS for the isolation of target cells from a complex microbial community, mainly the non-specificity of cell sorting which leads to co-sorting of non-target cells. Presence of non-target cells in sorted samples is often attributed to: (1) strong adherence of non-target cells to target cells due to the complex nature of the environmental sample; (2) background electronic noise in the scatter channels of the flow cytometer that masks the presence of small non-target cells and (3) auto-fluorescence of non-target cells, organic or inorganic particulates (Wallner *et al.*, 1997; Moter & Göbel, 2000; Kalyuzhnaya *et al.*, 2006; Gougoulas & Shaw, 2012). Bacterial cells in the floccular sludge community are often embedded in an EPS matrix with other organic and inorganic matters, and cannot be readily sorted (Wallner *et al.*, 1995; Marie *et al.*, 1996). Detachment of bacterial cells from floccular aggregates to form a homogenous cell suspension is a pre-requisite for flow cytometry (Falcioni *et al.*, 2006). However, it is difficult to obtain a single cell suspension from activated sludge. Wallner *et al.* (1997) could only enrich target groups from activated sludge to purity of less than 90% due to co-sorting of non-target cells. In another study, Kim *et al.* (2010) could not isolate different clades of *Candidatus Accumulibacter*: Acc-SG1 clade, Acc-SG2, Acc-SG3 and Acc-SG4. Sorted samples of Acc-SG2, Acc-SG3 and Acc-SG4 clades always contained members of Acc-SG1 clade because of the strong aggregation of different *Candidatus Accumulibacter* clades in activated sludge flocs. Despite various strategies employed for the dispersal of floccular aggregates: (1) direct sonication; (2) filtering of large cell aggregates by 35- μm and 10- μm pore-size filter prior to cell sorting and (3)

construction of several sorting gates to filter out cell aggregates, non-target cells in sorted samples is a persistent problem.

In addition to the complexity of the sample, specificity of sorting is also influenced by the 'environmental specificity' of probes. Gougoulas *et al.* (2012) showed that in a sorted sample, 30% of the clones constructed from a clone library belonged to *Burkholderia* species, even though the FISH probe was intentionally designed for the target population: *Pseudomonas*. Comparative sequence analysis revealed the presence of probe target site in the 16S rDNA clones of *Burkholderia*, thus highlighting that low environmental specificity of FISH probe subsequently affects the specificity of FISH-FACS sorting.

Chapter 3 Novel method of probe design driven by next generation sequencing reads

3.1 Introduction

Accurate detection and quantification of organisms in complex environments are important themes in microbial ecology that can be readily achieved through optimally designed fluorescence *in situ* hybridisation (FISH) probes. An optimal probe design aims to produce FISH probes that will hybridise specifically to the ribosomal sequences of most members of the intended target taxon under stringent experimental conditions (high coverage), and with minimal cross-hybridisation with non-target taxa present in the same sample (high specificity). Hitherto, published probes have always been designed through the process of 16S rDNA comparative sequence analysis (Figure 3-1).

Figure 3-1: An illustration of FISH probe designed through the process of comparative sequence analysis. Comparative sequence analysis of 16S rDNA gene sequences using probe design software allows the visualisation and discrimination of conserved regions, or regions that are exclusive to members of the target taxon. In this example, FISH probe can be designed against organism ThuPhen3 because of its different 16S rDNA sequence (highlighted in the red box). Organisms' names are depicted on the left side of the diagram.

Current probe design software relies on comparative sequence analysis for identifying potential signature sequence sites in the targeted taxon. Quality of probe design is highly dependent on comprehensiveness of the database and the curated taxonomy of the sequences in the database (Amann and Fuchs, 2008). Therefore, a curated database such as the SILVA database containing high-quality 16S rRNA sequences of full-length (~1500 bp) or near full-length (≥1200 bp) for comparative sequence analysis is frequently used. The use of full-length or near full-length sequences is beneficial in optimal probe design because it provides a higher taxonomic resolution

that guides probe designers on potential cross-hybridisations with other non-target taxa (Srinivasan *et al.*, 2015). Furthermore, the number of potential probe target sites increases concomitantly with the length of the 16S rRNA gene, thus enabling probe designers to choose better probe candidates with higher specificity and coverage for the intended target taxon. In addition, *in silico* accessibility of the 16S rRNA secondary structure to FISH probes can be modelled with full-length or near full-length sequences (Fuchs *et al.*, 1998; Behrens *et al.*, 2003). These benefits are compelling reasons for state-of-the-art probe design software such as ARB (Ludwig *et al.*, 2004) and Decipher (Wright *et al.*, 2012) to include comparative sequence analysis using full-length or near full-length sequences as the backbone of FISH probe design.

Probe design is integrated in the full-cycle rRNA approach where PCR amplification generates 16S rDNA clone libraries (Pace *et al.*, 1985). Drawback of PCR amplification is the inability of conserved bacterial primers to map an accurate community diversity due to primer biases, therefore resulting in a skewed analysis of the diversity in the community (Ravel, 2012; Klindworth *et al.*, 2013). Furthermore, bacterial populations present in low abundance are often not characterised due to limitations in time and manpower generating clone libraries (Snaird *et al.*, 1999). Therefore, it is only possible to design FISH probes for organisms through the full-cycle rRNA approach whose: 16S rDNA sequence is amenable to amplification by PCR primers and abundance is relatively high and can be easily detected from 16S rDNA clone libraries. Inherent problem of the full-cycle rRNA approach necessitates an alternate approach to FISH probe design.

There has been rampant increase in the use of metagenomic sequencing to characterise microbial communities over the past 10 years. Most genomic approaches involving shotgun sequencing are used to gather insights into the potential metabolic functions of a community (Chistoserdova, 2014) or to perform genome recovery (Albertsen *et al.*, 2013a). However, there have been no reported attempts on FISH probe design from short sequencing reads produced from genomic sequencing. This is because designing FISH probes through the conventional approach of comparative sequence analysis is difficult due to reduction in the taxonomic resolution of the 16S

rRNA gene. Hence, an alternate method to visualise target microbial group from short sequencing reads with a meaningful taxonomic resolution was established in this study. This method relies on the use of the RiboTagger software (Xie *et al.*, 2016) for the extraction of taxonomically-relevant 16S rRNA gene from shotgun surveys. Although the RiboTagger software has been applied for surveillance of the community structure (Kjelleberg *et al.*, 2014; Tan *et al.*, 2014; Feng *et al.*, 2017), it has not been employed in FISH probe design for the subsequent visualisation and characterisation of target bacterial taxa. RiboProbe is a molecular biology tool that can be integrated easily in studies that employ whole community shotgun surveys. For better clarity on the vocabulary pertaining to RiboTags and RiboTagger FISH probes, the taxonomical entity that each 33bp-RiboTag represents is defined as an OTU and the taxonomical entity which RiboTagger FISH probes hybridise to is defined as a taxon.

In this chapter, a strategy to visualise target microbial taxa from omics dataset is presented. This approach involved the design of FISH probes from hypervariable regions of the 16S rRNA present on short sequencing reads in omics dataset. The newly designed probe, whose nomenclature will henceforth be termed as RiboProbe was initially validated using *Thauera* as a reference organism. The RiboProbe was evaluated both *in silico* and empirically with a canonical FISH probe that was designed through the process of comparative sequence alignment. Evaluation of the FISH probes was carried out with: (1) *in silico* analysis of probe specificity, coverage and accessibility of target sites to FISH probes and (2) microscopic images and quantitative co-localisation assays. In addition, the ability of RiboProbe to visualise a previously uncharacterised bacterial taxon in activated was demonstrated. Lastly, the usefulness of RiboProbe as a PCR primer for taxon-specific detection was shown.

3.2 Materials and methods

A flowchart describing the validation and evaluation of RiboProbe is presented in Figure 3-2.

Figure 3-2: Flowchart demonstrating the steps involved in validation and evaluation of RiboProbe.

RiboProbe was validated through in two stages: initial validation was performed on an axenic culture, and subsequent validation was performed on the floccular sludge community of activated sludge.

3.2.1 Sample preparation

Microorganisms and culture conditions

Two axenic cultures of *Thauera* were used: *Thauera* sp. R086 (accession number: KC252920) was used as the reference organism and it was isolated directly from UPWRP activated sludge (Tan, 2013); the other axenic culture *Thauera linaloolentis* DSM 12138 (DSMZ, Germany) was used as the non-target organism for probe Ribo_Thau1029_17 in a probe dissociation curve. *Thauera* sp. R086 was used as a positive control to minimise the potential problem of non-specific cross hybridisation, or sample auto-fluorescence that could lead to subsequent confusion in the validation of RiboProbe. *Thauera* was selected because it was ranked as one of the more dominant members (2.88% in the metatranscriptomics dataset) of the floccular community in UPWRP. Furthermore, selection of *Thauera* simplifies the evaluation process of RiboProbe because of the

availability of only one canonical FISH probe: Thau646 (Lajoie *et al.*, 2000) that is specific for *Thauera* in probeBase (Loy *et al.*, 2003). Both bacterial cultures were grown aerobically in Luria-Bertani media supplemented with 5% w/v sodium chloride (LB5), in a shaker with a rotation speed of 200 rpm at 30°C for 12 hours. Single bacterial colony was selected by streaking the bacterial culture on LB5 agar that was incubated at 30°C.

Acquisition of activated sludge samples

Activated sludge was sampled from aerobic tank “2B” of a municipal full-scale wastewater treatment plant (Ulu Pandan Water Reclamation Plant, Southworks, Singapore). 90% of the influent originates from an urban settlement, with the remainder of the influent contributed by light industries. Samples were transported to the laboratory on ice (duration of transportation is approximately 45 minutes), where they were stored at 4°C prior to sample processing.

3.2.2 Sample fixation

In situ visualisation was performed on paraformaldehyde (PFA)-fixed samples. Fixation was performed upon arrival of the samples at the laboratory. Both axenic cultures and sludge samples were fixed using the same protocol as described by Amann *et al.* (1990). Bacterial biomass was first harvested by centrifugation, followed by the rinsing of the cell pellet with phosphate buffer saline (PBS, Thermo Fisher, USA). One volume of the sample was subsequent fixed with 3 volumes of 4% paraformaldehyde (PFA) (w/v) (Sigma Aldrich, USA) for 3 hours at 4°C. After incubation, the fixed biomass was rinsed twice with PBS, followed by resuspension in an equal volume of 100% Ethanol absolute (Merck, USA) and PBS. Fixed samples were stored at –20°C.

3.2.3 Fluorescence *in situ* hybridisation (FISH)

RiboProbe design

The highly complex floccular microbial community from UPWRP was used as a proof-of-concept for RiboProbe design. Initial 16S sequence tag profiling was performed to examine the diversity of the microbial community (Figure 2-3). Profiling was performed via the RiboTagger software (Xie *et al.*, 2016), where it detects for sequence tags from V4 to V7 variable regions of the 16S rRNA. Each identified sequence tag has a default length of 33 bp, and an OTU is defined by its unique

33bp-RiboTag sequence. In this study, RiboProbes with read length of 17 bp were designed from 33bp-RiboTags extracted from the V6 region. This was performed through truncation from the 3' end of the 33bp-RiboTag (Figure 3-3) or adjustment of the -tag INT parameter in the RiboTagger software (-tag 17).

Figure 3-3: A diagram illustrating FISH probe design from 33bp-RiboTag. Truncation of the length of the 33bp-RiboTag from the 3' end resulted in a 17bp-FISH probe.

Either adapter- and quality-trimmed metagenomics or metatranscriptomics reads obtained could be used as input for the design of RiboProbe. Nomenclature of RiboProbe generally follows the nomenclature of FISH probes as described by Alm *et al.* (1996), but with the following modifications:

1. “Ribo” to be added in front of the probe name, followed by an underscore to distinguish it from canonical probes
2. Probe name as described by Alm *et al.* (1996), followed by an underscore
3. Length of the probe

FISH probes

FISH probes were commercially synthesized and purified by high-pressure liquid chromatography (Sigma-Aldrich, USA or Integrated DNA Technologies, USA). Oligonucleotide sequences, optimal hybridisation conditions and related references of the FISH probes applied in this chapter are specified in Table 3-1.

Table 3-1: FISH probes and RiboTags used in this chapter

Probes/RiboTags	Intended target taxon	Sequence of probe (5'→3')	Formamide (%)	References
EUB338	Most bacteria	GCTGCCTCCCGTAGGAGT	0	(Amann <i>et al.</i> , 1990a)
Thau646	<i>Thauera</i>	TCTGCCGTACTCTAGCCTT	45	(Lajoie <i>et al.</i> , 2000)
Ribo_Thau1029_33	<i>Thauera</i>	GTGTTCTGGCTCCCGAAG	N.D	This study
*		GCACCCTCGCCTCTC		
Ribo_Thau1029_17	<i>Thauera</i>	GTGTTCTGGCTCCCGAA	45	This study
Ribo_Unk1029_33*	Unclassified bacterial taxon	TGCTTCGCGTCTCCGAAGA GCCGACCACCTTTC	N.D	This study
Ribo_Unk1029_17	Unclassified bacterial taxon	TGCTTCGCGTCTCCGAA	45	This study
Ribo_Unk1009_17	Unclassified bacterial taxon	CCGACCACCTTTCAGCA	N.D	This study

* Original 33bp-RiboTag or V6 sequence tag of the RiboProbe

N.D: Not determined

Slide-FISH

Fixed samples were immobilised on microscopic slides (dimensions: 8 wells with 6mm diameter, Cell-Line, USA) by air drying in the hybridisation oven (Shanke 'n' Stack, Thermo Fisher, USA) for 15 minutes. Slides were subsequently dehydrated in an ethanol (Merck, USA) dehydration series: 50%, 80% and 96% (v/v) for 3 minutes respectively. Hybridisation buffer was added to the sample. Subsequently, FISH probes were added to a final concentration of 5 ng/μL in a 'multiple nested probe concept' (Rudolf and Schleifer, 2001) according to their respective optimal hybridisation

and washing stringencies (Table 3-1) using a standard FISH protocol (Manz *et al.*, 1992). Hybridisation and washing buffer were prepared according to Appendix A-1. In a scenario where it was not possible to apply two probes in a simultaneous hybridisation due to different hybridisation stringencies, probe with the higher melting temperature was applied to the sample in the first round of hybridisation and washing; probe with the lower melting temperature was applied in the subsequent round of hybridisation and washing (Snaidr *et al.*, 1997).

Samples were examined through a confocal laser scanning microscope (Zeiss LSM 780, Germany) using x63 oil immersion objective lens. A 633-, 561- and 488-nm laser were used for the excitation of Cy5, Cy3 and FITC/A488 fluorophores respectively. Crosstalk between the different fluorophores was restricted through careful adjustment of the emission filters, as follows: Cy5: 642-695nm; Cy3: 571-615nm; FITC/A488: 500-535nm. Crosstalk was further prevented through acquisition of the individual fluorophores in individual tracks of the microscope.

eFISH analysis

Presence of FISH probes in 16S rRNA gene sequences of axenic culture of R086 was identified with an eFISH.pl script, which uses exact grep exact matching of the probe sequence (Law *et al.*, 2016).

3.2.3.1 Evaluation of the *in silico* specificity and coverage of probes

In silico coverage and specificity of probes were evaluated against the SILVA SSU Ref NR 99 database (version 123) by importing the database into ARB software (Ludwig *et al.*, 2004) and using the “multiple probe function” tool in ARB. Coverage is defined as the fraction of sequences in the target taxon with perfect complementary target site to the FISH probe relative to the total number of sequences in the target taxon.

$$\text{Coverage} = \frac{\text{Number of perfectly matched sequence in target taxon}}{\text{Total number of sequences in target taxon}} \times 100$$

Specificity is defined as the fraction of sequences in the target taxon with perfect complementary target site to the FISH probe relative to the total number of perfectly matched sequences in the database.

$$\text{Specificity} = \frac{\text{Number of perfectly matched sequence in target taxon}}{\text{Total number of perfectly matched sequence in database}} \times 100$$

3.2.4 Probe dissociation curve

Optimal hybridisation and washing condition for newly-designed RiboProbe was determined using a melting curve analysis as described by Crocetti *et al.* (2002). A melting curve was generated from a series of FISH experiments with increments in the concentration of formamide (FA) (Sigma Aldrich, USA) in the hybridisation buffers, and decrements in the concentration of sodium chloride (NaCl) (Thermo Fisher, USA) in the washing buffers. Stringency of the probe is altered by varying the concentration of FA and NaCl. Melting curve was performed with an axenic culture that showed no mismatch: R086 and an axenic culture that had a 2 bp mismatch to the probe: *Thauera linaloolentis*. Hybridisation temperature and washing temperature were maintained at 46°C and 48°C throughout the melting curve analysis. The same duration for hybridisation and washing was applied to samples at each respective FA concentration. Changes in probe stringency were proportional to the changes in fluorescence intensity of probe-labelled objects. Fluorescence intensity of probe-labelled objects for each FA concentration was measured using the same exposure time and detector settings with a confocal microscope. Multiple field-of-views of target cells (n=12) were acquired at 63x oil immersion objective for each FA concentration, and images were imported into the Digital Image Analysis In Microbial Ecology (DAIME, Austria) software where probe-labelled objects were segmented using the 'RATS-L' algorithm and mean fluorescence intensities were measured. Mean fluorescence intensity was plotted against each respective FA concentration in a melting curve. Formamide curve for a particular FISH probe was performed on technical triplicates.

3.2.5 Sequencing of the 16S rRNA gene

Axenic cultures were subjected to cell lysis and DNA was extracted with the QIAamp DNA Mini Kit (Qiagen, Germany) that was performed according to manufacturer's protocol. Near full-length 16S rRNA gene sequences were amplified by PCR primer set 27F/1492R from the extracted genomic DNA with a thermal capillary cycler (Eppendorf, Germany). Each PCR reaction mixture contained: 1 x Taq Buffer with $(\text{NH}_4)_2\text{SO}_4$; 2 mM MgCl_2 ; 0.4 mM dNTPs; 0.2 μM of each forward and reverse primer; 1.25 U of *Taq* DNA polymerase; genomic DNA as the template; DNase/RNase free water. Primer set 27F/RT_Thau was used to validate the presence of *Thauera*. PCR primers and their associated properties are presented in Table 3-2.

Table 3-2: PCR primers used for the evaluation of *Thauera*

Primers	Sequence (5'→3')	Length (bp)	GC content (%)	Annealing temperature (°C)
27F	AGAGTTTGATYMTGGCTCAG	20	40-50	55
1492R	TACCTTGTTACGACTT	16	38	55
RT_Thau	GTGTTCTGGCTCCCGAA	17	59	55

PCR amplification was performed in the thermal capillary cycler with the following settings: an initial denaturation step at 95°C for 3 minutes, followed by 30 cycles of denaturation at 95°C for 30s, annealing at 55°C for 30s, extension at 72°C for 2 minutes and a final extension step at 72°C for 10 minutes. Presence of PCR amplicon product was analysed by running the PCR products through 1% agarose gel electrophoresis (w/v) and subsequently staining with ethidium bromide. PCR products were purified using the PureLink® PCR purification kit (Life Technologies, USA) according to manufacturer's instructions and Sanger sequenced with the Applied Biosystems 3730xl DNA analyser at 1st Base (First Base, Singapore).

"Pureness" of axenic cultures was determined through two methods. The first method involved parsing the 16S rRNA gene sequence into the Lasergene's SeqMan Pro software (DNASTAR, USA) and observing for secondary peaks. The second method involved the taxonomic classification of the 16S rRNA gene sequence with the SILVA database.

Temperature-gradient PCR

Optimal annealing temperature for primer set 27F/RT_Thau was determined with a temperature-gradient PCR (Park *et al.*, 2005), with annealing temperatures ranging from 40-60°C being carried out because *in silico* annealing temperature was predicted to be 49.4°C using the web-based tool of Thermo Fisher Tm calculator (TM Calculator - ThermoFisher - <https://www.thermofisher.com/sg/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/tm-calculator.html>).

3.2.6 Visualisation of probe accessibility site of the 16S rRNA gene

16S rRNA gene sequence of R086 was imported and aligned with the ARB software; alignment was manually curated with other similar sequences imported from the SILVA SSU Ref NR 99 database (version 123). Aligned sequence of R086 was fitted into secondary structure model of *Escherichia coli* as described by Fuchs *et al.* (1998). A model of the secondary structure of the 16S rRNA gene of R086 was visualised using the 'secondary structure editor' in ARB edit (Ludwig *et al.*, 2004) and mapped with different brightness classes: I-VI as defined for the 16S rRNA structural model for *Escherichia coli*.

3.2.7 Co-localisation analyses

Post hoc co-localisation analyses: generation of scatterplots, Pearson's correlation coefficient (PCC) and Manders' co-localisation coefficient (MCC) were performed after acquisition of multiple field-of-views of images (n=22) containing *Thauera* cells at 63x oil immersion objective. Two channels, which represented the Cy5 and Cy3 fluorophores attached to probes Ribo_Thau1029_17 and Thau646 respectively, were selected for co-localisation analyses with Imaris 8.2.0 software (Bitplane). PCC and MCC values were derived from the 'co-localisation' function of the Imaris software. PCC estimates the degree of overlap between the Cy5 and Cy3 image pairs, and it has values ranging from -1 to 1: a value of 1 indicates that the image pairs are perfectly overlapped; a value of 0 indicates that the image pairs are uncorrelated; a value of -1 indicates that the overlap in the image pairs are inversely related to each other.

PCC values were obtained from the co-localisation output: 'Pearson's coefficient in ROI volume'. MCC estimates the contribution of Cy5 and Cy3 to the co-localised regions, and it has values ranging from 0 to 1: a value of 1 indicates that 100% of the channel will co-localise and a value of 0 indicates that none of the channel will co-localise. MCC values were obtained from the co-localisation output: 'Original Manders' Coefficient'.

3.3 Results

3.3.1 Evaluation of the *in silico* specificity and coverage of FISH probes

Prior to *in situ* hybridisation experiments, *in silico* coverage and specificity of RiboTagger and canonical probes were evaluated using *Thauera* as the target taxon. Canonical probe Thau646 had higher coverage (80.22%) than probe Ribo_Thau1029_17 (70.75%); 35 more members of the genus *Thauera*, out of a total of 360 sequences were targeted by the canonical probe. However, probe Thau646 had lower specificity (4.98%) than probe Ribo_Thau1029_17 (15.72%); there were 4,160 more outgroup hits for probe Thau646 (Figure 3-4). Outgroup hits are defined to be false-positive identification of organisms that are not in the intended target taxon (Amann and Fuchs, 2008). Interestingly, major outgroup phylum taxa targeted by both probes were different: probe Thau646 targeted mostly the *Betaproteobacteria* (91.72% excluding *Thauera* sequences) whereas probe Ribo_Thau1029_17 targeted mostly the *Cyanobacteria* (64.55%).

Probe evaluation was further extended to probes designed from the other variable regions: V4, V5 and V7. RiboProbes designed from the other variable regions were superior in terms of probe coverage and specificity than the canonical probe, with probes designed from the V5 region having the highest coverage and specificity (Figure 3-4).

Figure 3-4: Evaluation of the *in silico* coverage and specificity of RiboTagger and canonical probes against the genus *Thauera*.

RiboProbes were designed from variable regions: V4-V7 of the 16S rDNA gene of R086. For consistency of evaluation, all RiboProbes were truncated from their original 33bp-RiboTag to a 17bp-FISH probe.

Combinatorial use of probes Ribo_Thau1029_17 and Thau646 led to higher probe coverage (93.04%). Increase in the probe coverage can be explained by members of the genus *Thauera* being targeted by probe Ribo_Thau1029_17, but not by canonical probe Thau646 (Figure 3-5). Even with a combination of both probes, some members of the genus *Thauera* (6.94%) still could not be targeted.

Figure 3-5: 16S rRNA gene sequences of the genus *Thauera* in an ARB-parsimony guide tree covered by various probe combination of probes Ribo_Thau1029_17, Thau646 or both. Only a sub-set of the guide tree is displayed. Coverage of the different probes is colour coded and the legend of the various colour codes is located at the bottom of the figure.

3.3.2 *In silico* accessibility of 16S rRNA to RiboProbe

Secondary structure of the 16S rRNA is a factor that can influence the hybridisation of oligonucleotide probes to its target sites and the subsequent brightness of probe-labelled cells. Therefore, evaluating the 16S rRNA accessibility of target taxa to FISH probes prior to any *in situ* hybridisation experiments probe is important. This was achieved by incorporating the 16S rRNA gene sequence into the 16S rRNA accessibility model of *Escherichia coli* that was described by Fuchs *et al.* (1998), followed by mapping of 16S rRNA colour-coded accessibility map onto the sequence. Prior to modelling the accessibility model for the target taxon, there was a need to search for a suitable reference organism that would contain the RiboProbe binding site in its 16S rRNA gene.

An axenic culture of R086 was used as the reference organism. The presence of probe Ribo_Thau1029_17 binding site in R086 was determined through amplification of its 16S rRNA gene with bacterial universal primer set 27F/1492R from extracted genomic DNA and subsequent Sanger sequencing. Probes Ribo_Thau1029_17 and Thau646 binding sites were identified in the 16S rRNA gene sequence of R086 using eFISH analysis. Subsequently, the manually-curated aligned sequence of R086 was visualised in the secondary structure editor of the ARB software (Ludwig et al., 2004). Based on the *in silico* probe accessibility model, probe Ribo_Thau1029_17 was categorised as a Class III probe, with moderate probe accessibility to R086. Probe Thau646 was categorised as a Class IV probe when evaluated with the same accessibility model. RiboProbe designed from the V6 region of the 16S rRNA has better probe accessibility than the current canonical probe in targeting *Thauera*.

In addition, *in silico* accessibility of R086 to RiboProbes extracted from different variable regions: V4, V5 and V7 were tested using the same accessibility model (Table 3-3). Probes designed from the V5 and V6 region have the best accessibility; probe from the V7 region had the worst accessibility and probe from the V4 region had mixed accessibility which included a partially accessible target site (class V).

Table 3-3: *In silico* accessibility of R086 to RiboProbes extracted from variable regions V4-V7

Variable regions	Category for <i>in silico</i> accessibility
V4	Class II / V
V5	Class III
V6	Class III
V7	Class VI

***In situ* validation of RiboProbe**

Prior to *in situ* hybridisation, it was essential to check that the culture of R086 was indeed axenic.

R086 was validated to be axenic through the observation of sharp sequencing peak that corresponded to single nucleotide of the 16S rRNA gene. Additionally, the 16S rRNA gene

sequence of R086 was taxonomically assigned to the genus *Thauera* with the SILVA database. Confocal images from the co-hybridisation of probe Ribo_Thau1029_17_{Cy5} with published FISH probes, Thau646_{Cy3} and EUB338_{A488}, showed *Thauera* cells being overlapped with bright fluorescent signals from all three probes (Figure 3-6). A non-target organism, *Thauera Linaloolentis* which had 2 bp mismatch with Ribo_Thau1029_17_{Cy5} was used as a negative control. The negative control did not yield any fluorescence signals in confocal images using the same microscope settings.

Figure 3-6: Confocal micrographs of an axenic culture of R086 co-hybridised with probes Ribo_Thau1029_17_{Cy5}, Thau646_{Cy3} and EUB 338_{A488}. Cells were visualised with: (A) Cy5 filter set (red); (B) Cy3 filter set (magenta); (C) Alexa488 filter set (green) and (D) superimposition of the filters (pinkish-white). Bar: 5 µm. Magnification: 63x.

This provided evidence that RiboProbe designed from shotgun sequencing dataset was highly specific for the intended target taxon. Bright fluorescing signals emitted by probe Ribo_Thau1029_17 showed that the V6 region is an accessible region for FISH probes, and this corroborates with its *in silico* accessibility map.

3.3.3 Determining probe stringency of RiboProbe

Optimal hybridisation stringency of probe Ribo_Thau1029_17 was determined by performing a probe dissociation curve with an axenic culture of R086 as the target organism. An initial hybridisation with 10% formamide yielded a bright fluorescent signal, therefore demonstrating the good accessibility of the V6 region to probe Ribo_Thau1029_17. The probe exhibited maximum fluorescence intensity at a formamide concentration of 30% and 50%. Optimal formamide concentration for probe Ribo_Thau1029_17 was determined to be 50% at a hybridisation temperature of 46°C because it represented the highest formamide concentration before a drastic decrease in fluorescence intensity of the target cells (Figure 3-7).

Figure 3-7: Probe dissociation curve of probe Ribo_Thau1029_17.

Probe dissociation curve involved a series of FISH experiments that were performed on an axenic culture of R086 as the target organism, and *Thauera linaloolentis* as the non-target organism with 2 bp mismatches to the probe Ribo_Thau1029_17.

An organism with a weak single bp mismatch to the probe binding site is usually selected as a non-target organism for probe calibration, but such an organism was not available in axenic culture in this study. Nevertheless, *Thauera Linaloolentis* (DSMZ: 12138) was selected as the non-target organism because it contained 2 bp mismatch to the homologous probe binding site of probe Ribo_Thau1029_17 (Table 3-4). Even at the initial formamide concentration of 10%, there was negligible fluorescing signal for *Thauera Linaloolentis*. This proved that probe Ribo_Thau1029_17 is specific for R086 even under low hybridisation stringency.

Table 3-4: Homologous probe binding region of Ribo_Thau1029_17 in axenic cultures of R086 and *Thauera Linaloolentis*

Target organism	Probe target site
R086 (accession number: KC252920)	TTCGGGAGCC <u>AG</u> ACAC
<i>Thauera linaloolentis</i> (DSM 12138)	TTCGGGAGCC <u>TG</u> ACAC

Mismatched regions of the probe binding site are underlined

3.3.4 Hybridisation of RiboProbe with activated sludge

In situ difference between RiboProbe and canonical probes was further evaluated through co-hybridisation of probes Ribo_Thau1029_17_{Cy5}, Thau646_{Cy3} and EUB338_{A488} to floccular sludge community present in activated sludge. Activated sludge was used for comparison because activated sludge contained multiple species of *Thauera*, and this was shown where five unique 33bp-RiboTags were annotated to *Thauera* (Kjelleberg et al., unpublished). The truncation of the probe Ribo_Thau1029_33 into Ribo_Thau1029_17 altered the probe coverage, and the truncated probe Ribo_Thau1029_17 now hybridised to three different RiboTags of *Thauera*. In addition, presence of non-target organisms present in activated sludge that might potentially affect probe specificity would aid in the process of probe evaluation. Probe Ribo_Thau1029_17 overlapped with probe Thau646 in most labelled cells of the activated sludge. However, probe Ribo_Thau1029_17 did not overlap with probe Thau646 in some microscopy images (Figure 3-8).

Figure 3-8: Confocal micrographs of *Thauera* in fixed samples of activated sludge. Activated sludge was simultaneously hybridised with probes Ribo_Thau1029_17_{Cy5}, Thau646_{Cy3} and EUB338_{A488}. Arrow-pointed cells were hybridised with Thau646_{Cy3}, but not with Ribo_Thau1029_17_{Cy5}. Cells were visualised with: (A) Cy5 filter set (red); (B) Cy3 filter set (magenta); (C) Alexa488 filter set (green) and (D) superimposition of the filters. Bar: 5 µm. Magnification: 63x.

Objects that hybridised with probe Thau646, but not with probe Ribo_Thau1029_17 were not considered to be background artefacts because of the overlapping hybridisation with probe EUB338. Instead, these objects were identified to be bacterial cells. This group of cells had similar rod-shaped morphology with the cells that were co-hybridised with both FISH probes. Furthermore, different morphotypes of *Thauera* in UPWRP activated sludge consisting of rod and coccus morphology could be observed with probe Ribo_Thau1029_17 (Figure 3-9).

Figure 3-9: Confocal micrograph of different morphologies of *Thauera* existing in UPWRP's floccular sludge. Activated sludge sample was simultaneously hybridised with probes Ribo_Thau1029_17_{Cy5} (red) and EUB338_{A488} (green). *Thauera* appeared yellow because of the merging of probe fluorescence. Bar: 5 μ m. Magnification: 63x.

3.3.5 Co-localisation assays

The degree of co-localisation of probes Ribo_Thau1029_17- and Thau646-labelled objects would provide an insight to the *in situ* differences between the two probes. This was achieved by acquiring multiple confocal images of probe Ribo_Thau1029_17- and Thau646-labelled cells in biological replicates of activated sludge, followed by post-hoc quantitative co-localisation analysis. Two channels, representing the Cy5 and Cy3 fluorophores of probes Ribo_Thau1029_17 and Thau646 respectively were selected for co-localisation analysis. Scatterplots of the intensity of Cy5 versus the intensity of Cy3 for each pixel from paired images were plotted to identify: (1) co-localisation regions for quantitative co-localisation analysis and (2) the number of compartments in which the two probes co-localise. Single-labelled controls of Cy5 (Figure 3-10A) and Cy3 (Figure 3-10B) controls were used to gauge the level of background noise and to accurately set-up the crosshairs to identify co-localised regions.

Figure 3-10: Co-localisation scatterplots of Cy5 and Cy3 pixel intensities of activated sludge samples. Activated sludge samples were hybridised with: (A) probe Ribo_Thau1029_17_{Cy5}; (B) probe Thau646_{Cy3} and (C) both probes Ribo_Thau1029_17_{Cy5} and Thau646_{Cy3}.

A single linear relationship between the Cy5 and Cy3 pixels (Figure 3-10C) provided a qualitative indication that the ratio of cells co-labelled by Cy5 and Cy3 is similar and the co-localisation seems to reflect a single compartment where both the probes co-localise. To further accurately quantify co-localisation, Pearson's correlation coefficient (PCC) and Manders' co-localisation coefficient (MCC) were determined. PCC describes the similarity of shapes between images. PCC values of triplicate sets of samples co-hybridised with both probes had a mean value of 0.88 ± 0.027 ($n=66$, mean $\pm$ SD), whereas PCC measured negative for both sets of single-labelled controls (Figure 3-11).

Figure 3-11: Pearson's correlation coefficient (PCC) of confocal images of activated sludge samples that were co-hybridised with probes Ribo_Thau1029_17_{Cy5} and Thau646_{Cy3}, or a single-labelled probe. Samples with single-labelled probe served as negative controls.

PCC of co-hybridised samples were close to a value of 1, demonstrating that the fluorescence intensities of Cy5 and Cy3 were linearly related and a high degree of overlap existed between image pairs of the two channels. For single-labelled negative controls, PCC values were close to 0, demonstrating that the fluorescence intensities of Cy5 and Cy3 were uncorrelated and any observed co-localisation was due to random co-localisation effects. However, PCC does not yield any quantitative data about the fraction of probe that co-localised with each other. MCC analysis estimates the contribution of either the Cy5 or Cy3 channel to the co-localised regions of the image. MCC analysis was subsequently performed on the same set of microscopic images that were acquired for PCC analysis. MCC analysis indicated that $98.81 \pm 0.037\%$ ($n=66$, mean $\pm$ SD) of Cy5 co-localised to compartments associated with Cy3, whereas only $80.47 \pm 0.013\%$ ($n=66$, mean $\pm$ SD) of Cy3 co-localised to the same compartments (Figure 3-12).

Figure 3-12: Manders' co-localisation coefficient (MCC) of confocal images of activated sludge samples that were co-hybridised with probes Ribo_Thau1029_17_{Cy5} and Thau646_{Cy3}.

MCC values showed that majority of probe Ribo_Thau1029_17_{Cy5} objects co-localised with a smaller proportion of probe Thau646_{Cy3} objects in activated sludge samples. The difference in MCC values (18.3%) showed that probe Thau646_{Cy3}-labelled objects were present in compartments which lacked probe Ribo_Thau1029_17_{Cy5}.

3.3.6 Applications of RiboProbe

Visualisation of unclassified bacterial taxa

Usefulness of RiboProbe in visualizing members of bacterial taxa with no previous phylogenetic affiliations was demonstrated in this experiment. Probe Ribo_Unk1029_17 corresponded to a RiboTag that could not be taxonomically classified with the SILVA database. The RiboTag was selected because it targeted the highest abundant (0.77%) unclassified bacterial taxon in the metatranscriptomics dataset of UPWRP; dataset was profiled using 17bp-RiboTags. Majority of the unclassified bacterial taxa profiled with the 17bp-RiboTags have a relative abundance of less than 0.50% (Appendix A-2). Targeting an unclassified taxon present at a higher abundance would facilitate its detection through microscopy and subsequent FISH-FACS experiments.

As metatranscriptomics was performed in the year 2011, it was necessary to validate the unclassified taxonomy of the probe Ribo_Unk1029_17 in the year 2015 when these experiments were performed. Taxonomy of the RiboTag was verified to remain unclassified by scanning the sequence of probe Ribo_Unk1029_17 *in silico* using the 'TestProbe' tool of the SILVA database. The short length of Ribo_Unk1029_17 did not allow for the design of a second FISH probe through comparative sequence analysis. A second independent FISH was designed from the same, but longer representative sequence (81bp) of Ribo_Unk1029_17 to improve accuracy of a positive detection of the unclassified bacterial taxon. The second FISH probe annotated as Ribo_Unk1009_17 was designed 3 bp downstream of the target site of Ribo_Unk1029_17 (Figure 3-13).

Figure 3-13: Design of a second probe Ribo_Unk1009_17 on the same sequencing tag that probe Ribo_Unk1029_17 was designed from.

Probe Ribo_Unk1009_17 was labelled with a different fluorophore (FITC) at the 3' end, to minimise fluorescence quenching of the Cy5 dye that is attached at the 5' and 3' end of probe Ribo_Unk1029_17. However, artefacts were found to be associated with the hybridisation of probe Ribo_Unk1009_17_{FITC}. Therefore, the probe was used only to support detection of the unclassified bacterial taxon and not for other purposes (e.g. quantification). Co-hybridisation of probes Ribo_Unk1029_17_{Cy5}, EUB338_{Cy3} and Ribo_Unk1009_17_{FITC} yielded overlapping fluorescence signal that revealed the unclassified bacterial taxon to be a long filamentous bacteria (Figure 3-14). This was the first visual evidence that the RiboProbe can be used for the detection of an unclassified bacterial taxon.

Figure 3-14: Confocal micrographs of members of an unclassified bacterial taxon in UPWRP's floccular microbial community.

Activated sludge was simultaneously hybridised with probes Ribo_Unk1029_17_{Cy5}, EUB338_{Cy3} and Ribo_Unk1009_17_{FITC}. The unclassified filamentous bacteria was identified through the overlapped of the three probes. Objects which did not overlapped were regarded as noise. Cells were visualised with (A) Cy5 filter set (red); (B) Cy3 filter set (magenta); (C) FITC filter set (green) and (D) superimposition of the filters. Bar: 5 µm. Magnification: 63x.

Design of PCR primer from RiboProbe for detection of *Thauera*

For the next thesis chapter which involves sorting of *Thauera* from a mixed microbial community, a specific PCR primer set was developed to qualitatively determine the presence of *Thauera* in the sorted sample. Probe Ribo_Thau1029_17 was developed into a PCR reverse primer: primer RT_Thau. Oligonucleotide sequence of primer RT_Thau is identical to the sequence of probe Ribo_Thau1029_17. As primer RT_Thau targets the V6 region, it was essential to include a forward primer that targets the 5' end of the 16S rDNA. In an earlier experiment that was performed in

Section 3.3.2, PCR amplification with bacterial universal primer 27F/1492R on genomic DNA extracted from R086 yielded a PCR amplicon with a size of approximately 1500 bp (Figure 3-15A). The successful amplification of 16S rDNA highlights the suitability of primer 27F to be used as a forward primer.

Primer set of 27F/RT_Thau was evaluated on an axenic culture of R086. A temperature-gradient PCR with primer set 27F/RT_Thau yielded PCR amplicon products with a size of approximately 1000bp at a wide spectrum of temperature that ranged from 40°C to 60°C (Figure 3-15B). Since PCR amplicon products were produced at a wide range of temperature, an annealing temperature of 60°C was subsequently selected as the annealing temperature for future PCR amplification experiments involving primer set 27F/RT_Thau to minimise non-specific PCR amplification. PCR amplicon product amplified from primer set 27F/RT_Thau was sent for Sanger sequencing, and the amplicon sequence (accession number: KP941745.1) had a 100% sequence similarity to the 16S rDNA sequences of R086. This proved that RiboProbe can also be applied as a reverse primer in PCR reaction in combination with forward primer 27F to qualitatively validate the presence of the target taxon in a given sample.

Figure 3-15: Ethidium bromide-stained agarose gel of PCR products amplified with different primer sets. (A) Amplicon product of 1500 bp obtained with PCR amplification with primer set 27F/1492R; (B) temperature-gradient PCR of primer set 27F/RT_Thau with annealing temperature ranging between 40-60°C produced amplicon products of 1000 bp. -ve: negative control (nuclease-free water). L: 1 kbp ladder.

3.4 Discussion

3.4.1 Design of RiboProbe

Given the large-scale employment of high-throughput shotgun sequencing for microbial community profiling, the resulting short-length sequencing reads derived from highly variable region of the SSU gene present an opportunity for the visualisation of target taxon. In this chapter, a unique method of FISH probe design is presented and the design is conceptualised on the extraction of taxonomically-important 16S rRNA sequences from shotgun dataset, followed by the design of probes from the retrieved tag sequence. Current 16S rRNA sequence extraction software includes RiboTagger (Xie *et al.*, 2016), RiboFrame (Ramazzotti *et al.*, 2015) and SSUsearch (Guo *et al.*, 2016). RiboTagger software was eventually selected for probe design because of its faster processing time, and the use of universal recognition profile with a high sensitivity (>95%) for the extraction of RiboTag sequence.

RiboTags can be extracted from the V4-V7 regions of the 16S rRNA gene. In this study, the V6 region was selected for the design of RiboProbe for various reasons. Different variable regions of the 16S rRNA gene is expected to vary in their degree of evolutionary conservation (Fuchs *et al.*, 1998), and hence present different levels of accuracy in predicting taxonomic assignments (Yarza *et al.*, 2014). In an analysis conducted by Yarza *et al.* (2014), gene segments from the V5 and V6 regions provided better taxonomic resolution than the V4 and V7 regions. This enabled a more precise estimation of the taxa richness at the species level and prediction of unique taxa. Predicting unique taxa using the V6 region was also demonstrated by Kjelleberg *et al.* (unpublished), where RiboTags generated at the V6 region captured a large fraction of unclassified microorganisms residing in UPWRP's floccular sludge community. The hypothesis was that FISH probes could be designed from saturated sequencing dataset to further characterise organisms with no previous phylogenetic affiliation. The hypothesis was proven to be true as RiboProbe was used successfully to detect and visualise members of an unclassified bacterial taxon with filamentous morphology. Hitherto, the taxonomy of the unclassified taxon could not be deciphered due to the short read length of probe Ribo_Unk1029_17.

RiboProbe design adhered to the principles of conventional probe design: (1) probe length of 15-25 bp; (2) GC content of 50-80%; (3) absence of potential hairpin formation from self-annealing sites and (4) good *in silico* accessibility of probes to ribosomes (Pernthaler *et al.*, 2001; Thiele *et al.*, 2010). Design of RiboProbe involved the truncation of 33bp-RiboTag into a FISH probe length of 17 bp. Truncation was necessary so that RiboProbe could be used in complement with other canonical probes of similar read-length to attain stringent hybridisation at the standardised FISH hybridisation temperature of 46°C (Pernthaler *et al.*, 2001). An advantage of RiboTagger is the sample centric-design of FISH probes, as compared to conventional probe design which is based on a public database. The number of potential non-target taxa increases with a public database, and this makes probe design for an intended taxon a more difficult task.

In addition, RiboProbe design offers partial flexibility associated with comparative sequence probe design. The location and type of mismatches of probes to non-target taxa affect the design of probes (Yilmaz *et al.*, 2008). As RiboTagger extracts RiboTags from homologous regions (e.g. V6 region) of the 16S rRNA, the location and type (number and weight) of mismatches based on the relative strength of base pairings to other non-target taxa can be identified (Loy *et al.*, 2008). Competitor probes can subsequently be designed to prevent potential non-specific cross hybridisations. However, unlike other probe design software which predict potential mismatches and hybridisation efficiency, this projection is not integrated with the RiboTagger software and it requires an additional script to produce such an output.

3.4.2 Validation of RiboProbe

Co-hybridisation experiments of RiboProbe and canonical probes with the same intended target taxon was necessary for validating the concept of RiboProbe. Proof-of-concept of FISH probe design from short reads was initially validated with an axenic culture of *Thauera* (R086), which contained probe binding site for a RiboProbe: Ribo_Thau1029_17. Co-hybridisation of RiboProbe Ribo_Thau1029_17 with canonical probe Thau646 – labelled with fluorophores of different excitation and emission wavelength - showed a complete overlap in probe-labelled cells. This provided evidence that the RiboProbe matched its intended taxonomic affiliation. Specificity of

probe Ribo_Thau1029_17 was further scrutinized in a complex environmental sample through co-hybridisation with probe Thau646 in a sample of activated sludge. Pearson's correlation coefficient (PCC) values indicated 88% of the probe-labelled object overlapped, thus showing that the RiboProbe matched up to the canonical probe.

3.4.3 Evaluation of RiboProbe

RiboProbe was evaluated against the canonical probe using the genus *Thauera* as an example. The probes were evaluated on their *in silico* coverage, specificity and accessibility of the 16S rDNA of *Thauera* to probes; these criteria are commonly used for FISH probe evaluation (Pernthaler *et al.*, 2001; Loy *et al.*, 2003; Lückner *et al.*, 2007). Canonical probe Thau646 had higher probe coverage than its corresponding RiboProbe that was designed from the V6 region. This was expected because of the different approaches adopted for FISH probe design. Probe Thau646 was intentionally designed as a *Thauera* taxon specific probe through the process of comparative sequence analysis where 16S rRNA sequences of most members of the genus *Thauera* were aligned (Lajoie *et al.*, 2000), and therefore are expected to have higher *in silico coverage* for its target taxon. On the other hand, RiboProbe was not designed with the intention of detecting most members of the target taxon and multiple RiboTags can be annotated to the target taxon (Tan *et al.*, 2014). Even though probe Thau646 had a higher coverage, a fraction of *Thauera* species was missed out and this fraction could be captured by the V6-RiboProbe (Figure 3-5).

Interestingly, probes designed from the V4, V5 and V7 region have higher group coverage than probe Thau646. Probe designed from the V5 region has the highest specificity and coverage, and it presents an opportunity for users who wish to accurately quantify the abundance of *Thauera* in any examined habitats. Although the V5 probe is superior over the V6 probe, FISH probe design on the V6 region was continued in the subsequent chapters because the diversity study performed at UPWRP was based on the V6-RiboTag (Figure 2-3). Due to confidentiality and intellectual property of the raw sequencing data from UPWRP, only RiboTags that were extracted from the V6 region were available. Therefore, probe design from the other variable regions in targeting the unclassified bacterial taxa was not feasible. *In silico* coverage and specificity of the V4-V7 region

of *Thauera* could be compared due to Sanger sequencing of almost the entire 16S rRNA gene of R086 (~1400 bp).

An added advantage of RiboProbe is the higher specificity of probe to its intended target taxon. Probe Thau646 had lower specificity compared to RiboProbes designed from the V4-V7 regions, and this was probably because the probe was designed 16 years ago, where the size of the rRNA database was 345 times smaller than today (source: <http://www.arb-silva.de>) and the database contained fewer non-target sequences. Specificity of probe Thau646 has changed alongside with more comprehensive rRNA databases (Rappé and Giovannoni, 2003). Drawback of using a lower specificity probe is the propensity of the probe to hybridise to non-target taxa, whose 16S rRNA gene contain complementary binding site to the probe. Hence, quantification of *Thauera* would be skewed and biased with the use of low-specificity probe Thau646. This is especially important if probe Thau646 is applied to activated sludge, where Betaproteobacteria - the major outgroup hits of probe Thau646 - constitutes the most abundant class of microorganisms in floccular sludge community (Nielsen *et al.*, 2009; Ye and Zhang, 2013). Non-specific hybridisation to outgroup hits cannot be eradicated with the use of stringent hybridisation conditions or competitor probe (Manz *et al.*, 1992). Although specificity of RiboProbes were higher than the canonical probe, only probe designed from the V5 region had a specificity of more than 50%. Lower-than-expected specificity of the FISH probe was due to truncation of the original 33bp-RiboTag that had an original *in silico* specificity of 86.20%.

Higher-order structure of the ribosomes in target organisms is a consideration in probe design because it affects the accessibility of the probe to its target site and the subsequent brightness of probe-labelled cells (DeLong *et al.*, 1989). Accessibility of 16S rRNA of *Thauera* to the various FISH probes was evaluated using the accessibility model to *Escherichia coli*, where the colour-coded map reflects accessibility of probes to different target position of the 16S rRNA (Fuchs *et al.*, 1998). Probes designed from V5 and V6 regions have better probe accessibility than probe Thau646, and were categorised as Class III probes. Good accessibility of V6 region to probe Ribo_Thau1029_17

was reflected in confocal micrographs of bright fluorescently labelled cells in both axenic culture and activated sludge samples. Although it has been acknowledged that accessibility model is similar for microorganisms that share a close phylogenetic relationship, the model should provide hints for the selection of probe target sites with better probe accessibility. Hybridisation of probes to the V6 region of the filamentous unclassified bacterial taxon also resulted in bright fluorescently labelled cells, therefore substantiating the claims that the V6 region is a highly accessible region.

Quantitative co-localisation analyses were performed on activated sludge samples to gain insights on how RiboProbe measured up to canonical probe in an environmental sample. MCC analysis showed that approximately 18% of probe-labelled Thau646 objects did not co-localise with probe-labelled Ribo_Thau1029_17 objects. Higher *in silico* coverage or a lower *in silico* specificity of probe Thau646 were plausible reasons for the discrepancy in probe-labelled objects, and hitherto, it was not possible to decipher the discrepancy. The next thesis chapter proposes other molecular biology methods to decipher the rationale for this discrepancy. Interestingly, despite probe Ribo_Thau1029_17 having an *in silico* specificity of only 15.72%, majority of its probe-labelled cells co-localised with probe Thau646 (MCC=98.81%) in microscopic images and therefore can be identified as *Thauera* cells. The major outgroup hit of probe Ribo_Thau1029_17 was identified *in silico* to be Cyanobacteria, and MCC values obtained for probe Ribo_Thau1029_17 showed that Cyanobacteria is not prominent member of UPWRP's floccular sludge community. This was accomplished without the use of a Cyanobacteria specific probe or genomic sequencing. Cyanobacteria was also not identified in the RiboTagger analysis in the studies of Kjelleberg *et al.* (unpublished). Inconspicuous presence of Cyanobacteria could be due to the plant design of UPWRP where treatment of sewage water is shielded from direct sunlight. As Cyanobacteria relies on sunlight as an energy source for oxygenic photosynthesis (Ting *et al.*, 2002), the lack of sunlight suppressed the growth of Cyanobacteria.

3.4.4 Summary

In this chapter, a strategy of visualizing target taxon from omics dataset (metagenomics or metatranscriptomics) through FISH probe design from taxonomically-important short sequencing reads was demonstrated. The strategy relies on the use of the RiboTagger software to extract RiboTags from highly variable regions (V4-V7) of the 16S rRNA gene present in omics dataset. Subsequently, FISH probes were designed from RiboTags that represented taxa of interest. Two motivations drive the concept of RiboProbe. The first motivation aims to integrate high-throughput shotgun sequencing with the applications of FISH, such that FISH probes can be used for *in situ* quantification and visualisation of the identity and activity of individual microbial cells (Amann and Fuchs, 2008). The second motivation aims to avoid the bias of PCR amplification and generation of clone library associated with comparative sequence analysis. Although Karst *et al.* (2016a) have shown the prospect of producing full-length 16S rRNA sequence without the bias associated with primers through the adoption of synthetic long read sequencing by molecular tagging, readers are reminded that this advancement happened only in recent years (year: 2016). The RiboTagger approach was conceptualised in the year 2012 and producing full-length 16S rRNA sequences from a mixed microbial community on NGS platforms was not conceivable during that period.

A RiboProbe annotated to *Thauera* was first validated in an axenic culture, and further tested in a complex environmental ecosystem such as activated sludge. A high degree of overlap (PCC=88%) existed between probe-labelled cells of RiboTagger and canonical probe. V6-RiboProbe had higher specificity and probe accessibility than canonical probe when evaluated using *in silico* models. Probe Thau646 remains to be the only canonical probe for the detection of *Thauera* in Probe Base, but probe Thau646 was shown to be no longer reliable for the detection of *Thauera* species due to its low *in silico* specificity. RiboProbes designed from the V5 or V6 region are better probe candidates in terms of probe specificity.

Visualisation of an unclassified bacteria taxon with filamentous morphology is a first step towards demonstrating the potential of RiboProbe in characterising unclassified taxa in an ecosystem, and a direction towards better understanding the microbial community that drives wastewater purification. With results obtained from the evaluation of probes, RiboProbe can be used as alternative or in complement to existing canonical probes in future studies. RiboProbes were used for the visualisation, quantification and genome recovery in the scope of this thesis. However, application of RiboProbe can be further extended to methodologies such as MAR-FISH (Lee *et al.*, 1999) and NanoSIMS (Li *et al.*, 2008) in future studies for probing microbial activities.

Chapter 4 Targeted enrichment of microbial group through FISH-FACS

4.1 Introduction

Since its induction as a tool for cell biology, FACS sorters have primarily been used for the separation of fluorescent-tagged mammalian cells. The average size of a prokaryotic cell is approximately 1 μm , 10 times smaller than a eukaryotic cell. Sorting of prokaryotic cells based on their phenotypic properties (e.g. size, morphology) is inherently a difficult task and is highly reliant on the sensitivity of the FACS machine. However, advancements in the technology of FACS sorter such as improved detection, sorting capabilities and multi-parametric measurements have facilitated the sorting of prokaryotic cells (Czechowska *et al.*, 2008). FACS has been used in recent years for the targeted enrichment of prokaryotic cells from a spectrum of sampling sites. For instance, Irie *et al.*, (2016) had used the FACSria II sorter (BD Biosciences, USA) for enrichment of *Accumulibacter* and *Nitrospira* from activated sludge samples, while Rinke *et al.*, (2013) had use the Influx sorter (BD Biosciences, USA) for the sorting of single prokaryotic cell from nine different sampling sites (e.g. lake, bioreactor, hot spring, etc).

While FISH allows for the visualisation and quantification of target taxa using microscopy (Amann *et al.*, 1990b), it does not yield information pertaining to the functional capability of the target taxon. Genomic sequencing overcomes the limitation of FISH by providing access to the genetic information of the organism where comprehensive insights into the evolution and metabolic potential can be obtained (Podar *et al.*, 2007; Rinke *et al.*, 2013). The coupling of FISH with FACS or FISH-FACS is a technique based on FISH hybridisation and labelling of cells of the target taxon with fluorescence, followed by the isolation of fluorescent-labelled cells through cell sorting (Figure 4-1).

Figure 4-1: An illustration of FISH-FACS for the sorting of fluorescent-labelled target cells from a mixed microbial community.

(A) FISH probes hybridised to signature sequence in the 16S rRNA of target cells. (B) Subsequently, fluorescent-labelled target cells are sorted with the FACS sorter.

Initially, FISH-FACS was used for phylogenetic analysis of the 16S rRNA gene (Amann *et al.*, 1990a; Wallner *et al.*, 1997; Snaidr *et al.*, 1999; Park *et al.*, 2005; Schroeder *et al.*, 2009) or specific functional genes like the *ppk1* gene (Kim *et al.*, 2010) and *pmoA* gene (Kalyuzhnaya *et al.*, 2006) of the sorted population. More recently, FISH-FACS has been utilised for the genome sequencing of the sorted population, such as the *Candidatus Methanoperedens nitroreducens* that is responsible for the anaerobic oxidation of methane (Haroon *et al.*, 2013a). As fixation by paraformaldehyde is detrimental to genome recovery (Clingenpeel *et al.*, 2014, 2015), a fixation-free FISH-FACS protocol has been described by Yilmaz *et al.*, (2010b) to facilitate genomic sequencing of the sorted population. This protocol will be integrated into this study for the effective sorting of prokaryotic cells.

FISH-FACS has frequently been used to enrich for target taxon from a mixed microbial community to overcome the current bottleneck in metagenomics, which is the inability to retrieve draft

genomes of low abundant organisms from metagenomics despite the advancements in genomic binning technology and sequencing technology (Albertsen *et al.*, 2013a). Only genomes belonging to the dominant members of the community are recovered through metagenomic approaches (Tyson *et al.*, 2004). Enrichment of the target taxon prior to metagenomics will vastly improve the resolution of genomic analysis. However, genome recovery can only be achieved if sufficient biomass are sorted for genomic extraction. This approach is not feasible for target population present in low abundance, as it is extremely time-consuming to sort sufficient biomass (at least 10^6 cells) to meet the minimum DNA amount (nanograms to micrograms) required for current sequencing technology (Shapiro *et al.*, 2013). Nanograms of starting material are required due to low efficiency of library preparation for sequencing, which consequently leads to drop out of the genomic loci and loss of genomic information (Blainey, 2013).

Limitation of obtaining sufficient biomass for sequencing could be overcome through whole genome amplification (WGA) (Bing *et al.*, 2008). Multiple displacement amplification (MDA) is a WGA-affiliated technology that relies on *Bacillus subtilis* bacteriophage Phi29 polymerase to extend the 3' end of the random hexamer primers on genomic template in a strand displacement fashion (Dean *et al.*, 2001). Minute amounts of starting genetic material are amplified with high fidelity to genomic fragments that are kilo-base pairs in length and micrograms in quantity. High fidelity of amplification can be attributed to the 3-5' exonuclease proofreading mechanism of Phi29 polymerase (Zhang *et al.*, 2001). The success of genome recovery from FISH-FACS sorted samples was demonstrated by Podar *et al.* (2007), where 20% of the genome belonging to the uncultivated division TM7 phylum present in low frequency in soil samples (0.02% from flow cytometric analysis) was recovered.

Rinke *et al.*, (2014) had outlined a comprehensive protocol for FACS single-cell sorting of environmental microorganisms, where single cells could be isolated from the environmental sample and the genome could be amplified through multiple displacement amplification (MDA) for downstream shotgun sequencing and phylogenetic screening. While the protocol outlined by

Rinke *et al.*, (2014) had resulted in the successful amplification of 201 cells from 29 previously unexplored branches of the tree of life, the requirements for a specialized area (clean PCR hood) and equipment (Influx sorter) for single-cell MDA experiment were not attainable in our experiment due to limited resources. These limitations have made the task of genome recovery more susceptible to DNA contamination. Amplification of contaminated DNA often presents a major setback during MDA amplification of the target genome (Raghunathan *et al.*, 2005), and it further complicates downstream genomic analyses (Woyke *et al.*, 2010). In lieu of the protocol outlined by Rinke *et al.*, (2014), limitations to single-cell MDA experiment could be overcome with a more comprehensive decontamination procedure and appropriate controls to estimate the degree of contamination.

One of the aims of this chapter is to develop a robust and reproducible framework that could overcome the constraints of having limited access to a sterile environment. In addition, the lack of access to the Influx sorter as mentioned by Rinke *et al.*, (2014) has made it mandatory to explore other FACS sorters. However, not all FACS sorters are suitable for the sorting of prokaryotic cells. The effectiveness of sorting from the same sample: an axenic culture of *Thauera* using two FACS machines (MoFlo XDP versus Sy3200) with different configurations were evaluated. Although the MoFlo XDP was equipped with a larger diameter nozzle, the nano-view particle detector in the forward scatter proved to be more sensitive than the Sy3200 in the detection of prokaryotic cells.

Majority of studies using FISH-FACS have employed standard fluorescent-labelled oligonucleotide probes for the sorting of prokaryotic cells. The novelty of this study lies in the use of calibrated RiboProbe. As RiboProbe specific for *Thauera* had already been calibrated in the previous chapter, FISH-FACS was first applied to enrich for *Thauera* from an axenic culture of R086 as positive control, and subsequently from a floccular sludge community. The goal was to test whether RiboProbe had a higher specificity for *Thauera* than canonical probe as predicted by *in silico* analysis in the previous chapter and this was achieved through RiboTagger 16S rRNA analysis of the sorted samples.

4.2 Methods

A flowchart describing the workflow of FISH-FACS optimised for the specific sorting of a target organism from floccular sludge community and downstream phylogenetic and diversity analyses of sorted sample is presented in Figure 4-2.

Figure 4-2: Flowchart demonstrating the steps involved in FISH-FACS for the enrichment of a target taxon, followed by analyses to determine the levels of enrichment and phylogeny of the target taxon.

Reagents used for FISH-FACS and MDA amplification were purchased as molecular biology grade reagents. Equipment and reagents used for MDA purposes were DNA-free and RNA-free and have been UV-treated by the UV-crosslinker (CL-1000, UVP). MDA experiments were conducted in a UV-treated biological safety cabinet. Collectively, this helped to minimise the risk of DNA contamination.

4.2.1 Sample preparation

Sample acquisition

Both axenic cultures of R086 and activated sludge samples were used for FISH-FACS. R086 culture was identical to the sample as described in Section 3.2.1. Activated sludge samples were sampled at the identical location as described in the Section 3.2.1. A more methodological approach was undertaken in this chapter to include biological and technical replicates for various purposes:

genomic DNA extraction, quantitative-FISH analysis and cell sorting (Table 4-1). Samples were vortexed to ensure even mixing of the sample before distribution into aliquots for technical replicates. Only one technical replicate was used for FACS sorting because of the long duration (approximately 8 hours) required to perform cell sorting.

Table 4-1: Sampling of axenic culture and activated sludge samples for biological and technical replicates

Samples	Biological replicates and sampling date	Number of technical replicates		
		DNA extraction*	FISH analysis*	FACS sorting**
Axenic cultures of R086	January 29 th , 2016 (Replicate 1)	2	3	1
	February 19 th , 2016 (Replicate 2)	2	3	1
Activated sludge samples	January 21 st , 2016 (Replicate 1)	3	3	1
	March 9 th , 2016 (Replicate 2)	3	3	1
	March 23 rd , 2016 (Replicate 3)	3	3	1

*Performed on pre-sorted samples

**Refers to the number of sample used as input for FACS sorting with the Molo XDP sorter

DNA extraction from pre-sorted samples

Sludge samples were deposited in a -80°C freezer prior to cell lysis. DNA extraction was performed using the FastDNA spin kit for soil (MP Biomedicals, USA) according to manufacturer's instructions. Briefly, 2 ml aliquots of sludge were used for each DNA extraction. Sludge samples were homogenised for 4 cycles of 40s at a speed of 6.0 m/s using a bead homogeniser (MP Biomedicals FastPrep-24, USA) and DNA was eluted in pyrogen-free water. Eluted DNA was purified using genomic DNA Clean & Concentrator (Zymo Research, USA) that was performed according to manufacturer's instructions.

Sample fixation

Samples were fixed according to protocol as described in Section 3.2.2.

4.2.2 Fluorescence *in situ* hybridisation FISH probes

FISH probes and their associated properties are presented in Table 4-2.

Table 4-2: FISH probes used for sorting of *Thauera* from axenic cultures and activated sludge samples

Probes	Intended target taxon	Sequence of probe (5'→3')	Formamide (%)	Reference
EUB338	Most bacteria	GCTGCCTCCCGTAGGAGT	45	(Amann <i>et al.</i> , 1990a)
NON338	None	ACTCCTACGGGAGGCAGC	45	(Wallner <i>et al.</i> , 1993)
Thau646	<i>Thauera</i>	TCTGCCGTACTCTAGCCTT	45	(Lajoie <i>et al.</i> , 2000)
Ribo_Thau1029_17	<i>Thauera</i>	GTGTTCTGGCTCCCGAA	45	This study

Fixation-free in-solution FISH

Fixation-free in-solution FISH (Haroon *et al.*, 2013b) was performed prior to cell sorting so that FACS-sorted samples were amenable to further downstream genomic sequencing and analysis. A suspension of 100 µl of fixation-free samples was centrifuged and the biomass was subsequently rinsed twice with PBS to remove auto-fluorescence particles. Supernatant was discarded and the pellet was hybridised with 100 µl of hybridisation buffer, with FISH probes added to a final concentration of 5 ng/µl. Hybridisation and washing buffer were prepared according to Appendix A-1. Samples were hybridised overnight in the hybridisation oven at 46°C. Washing buffer was equilibrated to a temperature of 48°C at least 15 minutes before the washing procedure. After hybridisation, 500 µl of pre-warmed washing buffer was added to the hybridised samples. The sample was vortexed and then centrifuged to remove the supernatant. 500 µl of pre-warmed washing buffer was resuspended with the cell pellet and incubated at 48°C for 30 minutes. After the washing procedure, sample was centrifuged and washed with 500 µl of PBS, and the cell pellet was finally resuspended in a solution of 3 ml of PBS.

Breaking up of cellular aggregates

Cellular aggregates needed to be disrupted prior to FACS sorting, and this was achieved through sonication treatment via a probe sonicator (SONICS Vibracell VCX 750 Ultrasonic Cell Disrupter, USA) for 15 seconds with an interval of 5 seconds pulse.

Validation of hybridisation and sonication results

15 µl of pre-sonicated and sonicated samples were spotted onto microscope slides (Cell-Line, dimensions: 8 wells with 6mm diameter, USA), air dried in the hybridisation oven (Shanke 'n' Stack, Thermo Fisher, USA) for approximately 10 minutes, and subsequently mounted in Citifluor (Citifluor LTD, United Kingdom). Effectiveness of in-solution FISH hybridisation and aggregate dispersal was examined through a confocal laser scanning microscope (Zeiss LSM 780, Germany) using 63x oil immersion objective. Settings for image acquisition were as described in Section 3.2.3.

4.2.3 Quantitative FISH

Quantitative FISH analysis was performed on images obtained from probe-hybridised samples to quantify the relative abundance of the target taxon in the biomass through the following equation:

$$\text{Relative abundance of target taxon} = \frac{\text{biovolume of target taxon hybridised by specific probe}}{\text{biovolume of biomass hybridised by probe EUB338}}$$

Quantitative FISH was performed on fixed samples using the slide-FISH protocol as described in Section 3.2.3. Biovolume was calculated with the Imaris 8.2.0 software (Bitplane, Switzerland). To obtain the biovolume, multiple three-dimensional microscopic stack images along different random positions of the sample were acquired at 63x magnification. Probe-labelled cells were segmented using the 'surface segmentation' algorithm of the Imaris software. Subsequently, a filter was used to remove background noise on the segmented images by using an absolute intensity threshold value of >10.

4.2.4 Fluorescence activated cell sorting (FACS)

FACS sorting was performed only after confirmation through microscopic observation that the target cells were positively labelled by FISH probes, and that cellular aggregates were broken up. Flow sorting and cytometric analysis was initially performed on the Sy3200 cell sorter system (Sony, USA), and subsequently on the MoFlo XDP (Beckman Coulter, USA). FACS machines were equipped with ion lasers that provide a source of excitation light for the fluorophores attached to FISH probes. MoFlo XDP sorter was operated with a 100 µm nozzle and 30 psi sheath liquid pressure. A 488 nm solid state laser (100 mW) was utilised as an excitation source for the measurement of Alexa488 fluorophore and light-scattering properties; a 640 nm laser (20 mW) was used as an excitation source for the measurement of Cy5 fluorophore. MoFlo XDP sorter was equipped with a Nanoview small particle detection module (Propel Labs, Fort Collins, USA) and was calibrated with 0.4 µm and 0.8 µm beads that allowed the detection of small particles. Forward scatter was detected with the Nanoview module; side scatter was detected with a BP 488/10; Alexa 488 fluorescence was detected with BP 530/30 and Cy5 fluorescence was detected

with BP 670/40. Drop delay was calibrated using fluorescent FlowCheck pro beads (Beckman coulter) and auto drop delay wizard in Summit software (Beckman Coulter). Analysis of data on sorting days was performed with the Summit software. Post hoc flow cytometric plots were presented using the FlowJo software (LLC, USA).

Sy3200 sorter was operated with a 70 µm nozzle and 50 psi sheath liquid pressure. A 561 nm laser (75 mW) and 642 nm laser (70 mW) were utilised as an excitation source for the respective excitation of Cy3 fluorophore and Cy5 fluorophore. Light-scattering properties were measured with 830nm (45mW) and 808nm (75mW) lasers. Forward scatter was detected with a BP795/50; side scatter was detected with a LP 825; Cy3 and Cy5 fluorescence were detected with a BP 670/40. Drop delay was calibrated using fluorescent SortCal beads (Sony). Analysis of data on sorting days was performed with the WinList 3D software (Version 7.1, Verity Software House, USA). Post hoc flow cytometric plots were presented using the FlowJo software (LLC, USA).

Prior to cell sorting, FACS clean solution (BD Biosciences, USA) was passed through the fluidic lines of the cell sorter for an hour, followed by 30 minutes of flushing with DNA-free water. The same procedure was applied prior to the second round of sorting. New sterile sheath fluid (BD FACS-flow sheath fluid, USA) was used for each sorting experiment. These procedures helped in sterilising the FACS machine and minimising the entry of contaminants that would complicate downstream multiple displacement amplification (MDA) reactions.

Sorting gates were constructed to exclude cell aggregation and to capture events exhibiting high fluorescence from probe-labelled cells. Negative controls: no-probe and nonsense probes were used to estimate the level of background and non-specific binding respectively. Two-rounds of sorting were performed with the defined sorting gates to increase the specificity of sorting. Cells were initially sorted with 'high purity' mode based on the sorting gates into a sterile 1.5 ml Eppendorf tube, which was subsequently used as an input for a second-round sort with 'single cell' mode into 200 µl sterile PCR tubes. Efficiency of sorting was verified through microscopic

visualisation of the sorted cells (n=2000 events) on a confocal microscope (Zeiss LSM 780, Germany).

4.2.5 Multiple displacement amplification (MDA)

Sorted cells were directly used as templates for whole-genome amplification to meet the required minimum concentration of 20 ng/μL or a minimum total quantity of 1.5 μg for Illumina genome sequencing. Cell lysis of sorted cells and WGA was performed using MDA (REPLI-g single cell kit, Qiagen, USA) according to manufacturer's protocol. Amplified DNA was purified using ethanol precipitation as described by the QIAGEN REPLI-g single cell supplementary protocol: Purification of DNA amplified using REPLI-g kits.

4.2.6 16S rDNA clone libraries

16S rDNA clone libraries were constructed from a single-sorted, MDA-purified sample to obtain full-length 16S rDNA sequences for phylogenetic analyses. This was achieved using PCR with universal bacterial primer set 27F/1492R and *taq* polymerase. Set-up and purification of the PCR reactions were as described in Section 3.2.6. A clone library was generated using the TOPO TA cloning kit for sequencing (Invitrogen, USA) by ligating purified *Taq* polymerase-amplified PCR amplicon products into pCR 4-TOPO TA vector (Invitrogen, USA) which contained ampicillin, kanamycin markers and the lethal *ccdB* gene.

Figure 4-3: Features of pCR 4-TOPO vector with its cloning site. Image is adapted from Invitrogen (Invitrogen, 2016).

The vector was subsequently transformed into TOP 10 Electrocomp *E.coli* cells (Invitrogen, USA). Selection of recombinants was based on positive screening for colonies that propagated on LB5 agar plates which contained 50 µg/ml of ampicillin. Clones were randomly selected and the plasmids were extracted using the PureLink Quick Plasmid Miniprep Kit (Invitrogen, USA). Plasmids were sequenced with the Applied Biosystems 3730xl DNA analyser at 1st Base (First Base, Singapore) using the primer set M13F (-20) and M13R-pUC(-26). Sequencing reads were assembled into contig using a minimum 80% match criteria with the SeqMan Pro software (DNASTAR, USA). Screening of vector sequences and trimming of low-quality sequences at the end of reads were performed before the assembly process. Only clones with near full-length sequences (≥1200 bp) were selected. Presence of chimeric sequences was screened using the DECIPHER web-interface tool (Wright *et al.*, 2012).

4.2.7 Phylogenetic analysis

16S rDNA sequences of clones were clustered into OTUs using a 97% and 99% similarity cut-off value with the pick_otus.py script in QIIME (Caporaso *et al.*, 2010; Edgar, 2010). Representative sequences were selected from OTUs with the pick_rep_set.py script in QIIME. For taxonomic classification of clone sequences, sequences were submitted to the web-interface SINA 1.2.11 (Pruesse *et al.*, 2012). The “search and classify” tool was used to taxonomically classify representative sequences with the SILVA SSU Ref NR 99 database (version 123) using the Lowest common ancestor (LCA) method with a minimum sequence identity of 0.95%. Representative sequences were imported into the SILVA database using the ARB software (Ludwig *et al.*, 2004). Sequence alignment was performed with ARB edit, where the aligned sequences were eventually manually inspected with its closest neighbour to check for correct alignment. A maximum likelihood tree was calculated using the RAxML (Randomized Axelerated Maximum Likelihood) program with rapid bootstrap analysis of 1000 bootstraps.

4.2.8 Metagenomics sequencing

Metagenomic sequencing and quality-trimming of reads

Illumina TruSeq Nano DNA sample preparation protocol was used for sequencing library preparation. Genomic sequencing of MDA-amplified samples was performed on a MiSeq (Illumina, USA), producing paired-end reads with a read length of 300bp. The total number of pre- and post-sorted samples processed for genomic sequencing is presented in Table 4-3.

Table 4-3: Number of replicates processed for MiSeq genomic sequencing

	Number of samples		
	Probe	Probe	Probe Thau646
	Ribo_Thau1029_17 in axenic culture	Ribo_Thau1029_17 in activated sludge	in activated sludge
Pre-sorted samples	6	9	0
Post-sorted samples	24	31	12

Illumina TruSeq adapters were removed and quality-trimmed using a minimum Phred score of 20 and a minimum length of 30 bp from both end of the reads using the software BBDuk tools (BBMap- Bushnell B. – <http://sourceforge.net/projects/bbmap>).

Calculating sequencing coverage

Average sequencing coverage of sample was estimated using BBMap tool with the default setting (BBMap- Bushnell B. – <http://sourceforge.net/projects/bbmap>).

Extraction of 16S rRNA gene sequences from contigs

Partial 16S rRNA gene sequences present on contigs were retrieved using the rRNA.sh script from the mmgenome toolbox. Extracted 16S rRNA gene sequences were aligned and taxonomically classified using the web-based tool of SINA.

4.2.9 RiboTagger 16S rRNA analysis

Taxonomic profiling of pre-sorted and sorted samples was performed using the RiboTagger software (Xie *et al.*, 2016). Reads originating from the V6 regions of the 16S rDNA were extracted and taxonomically classified using the SILVA SSU Ref NR 99 database (version 123). Relative abundance of a taxon was derived through the following equation:

$$\text{Relative abundance of a taxon} = \frac{\text{Sum of sequence tags annotated to the taxon}}{\text{Total number of sequence tags in the sample}}$$

4.2.10 16S rRNA nucleotide sequence accession numbers

Near full-length 16S rDNA sequences (>1200 bp) obtained from clone libraries were deposited in NCBI's 16S prokaryotic ribosomal RNA database with the following accession numbers: KX914678–KX914731.

4.3 Results

4.3.1 Evaluating suitability of FACS sorters for targeted enrichment

The feasibility of using RiboProbe for targeted enrichment with flow cytometry was first established on an axenic culture of R086 as a positive control. Probes Ribo_Thau1029_17 and EUB338 were used for the hybridisation of the target cells. Two FACS machine: Sy3200 and MoFlo XDP with different system configurations were evaluated on their ability to effectively sort *Thauera* cells. The main differences between the configurations of Sy3200 and MoFlo XDP are presented in Table 4-4. Sy3200 sorter was selected for the initial sorting experiment because it contained more laser units for fluorophore excitation as compared to the MoFlo XDP.

Table 4-4: Differences between the Sy3200 (Sony) and MoFlo XDP (Beckman Coulter) sorter

	Sy3200 (Sony)	MoFlo XDP (Beckman Coulter)
Nozzle size	70 μm	100 μm
Lasers	405-, 488-, 561- and 640-nm	405-, 488- and 640-nm
Scatter	Forward and side scatter	Forward scatter equipped with a nano-view particle detector that can detect particles down to 0.2 μm in size, and side scatter
Enrichment after one round of sorting*	5.33% to 39.3%	39.3% to 99.5%

* Enrichment was measured using flow cytometric analysis

Using similar sorting gates (details of the sorting gate are presented in Section 4.3.3 for Mo-Flo XDP and Appendix A-3 for Sy3200), the Mo-Flo XDP was more sensitive than the Sy3200 in the initial detection of *Thauera* cells (Table 4-4). The Mo-Flo XDP also proved to be more efficient at enriching for *Thauera* cells after an initial round of sorting. Even though it was sorting from an axenic culture of R086, the Sy3200 sorter was not as sensitive for the sorting of prokaryotic cells. Difficulty of observing Sy3200 sorted cells under microscopy further validated the low sensitivity of Sy3200 for small-particle sorting. Therefore, subsequent cell sorting experiments were conducted on the MoFlo XDP sorter.

4.3.2 Eliminating contamination for downstream MDA

Presence of contamination in sorted samples

RiboTagger 16S rRNA analysis was performed on Sy3200 sorted sample to define the levels of enrichment for *Thauera*. RiboTagger 16S rRNA analysis revealed trends of low level of enrichment for *Thauera* and high traces of contamination in the sorted sample (Figure 4-4).

Figure 4-4: Diversity of a Sy3200-sorted sample analysed by RiboTagger 16S rRNA analysis. *Thauera* has not been enriched through the process of FISH-FACS from an axenic culture of R086, as shown by the relative abundance of RiboTags annotated to *Thauera*. High levels of contaminants represented by *Neisseriaceae* and *Malassezia* were present.

Thauera was present at an abundance of only 0.87%; *Neisseriaceae* (50.84%) and *Malassezia* (45.50%) formed the most abundant taxa in the sorted sample and were identified to be contamination due to the lack of probe Ribo_Thau1029_17 in the RiboTag sequence (Appendix A-4).

Hotspots of contamination in the FACS machine

High levels of contamination present in the sorted sample showed that sterilisation of the FACS machine was necessary prior to sorting experiments. Fluidic lines from both the sampling port and sheath tank were hypothesised to be potential hot spots of contamination.

Figure 4-5: An illustration depicting the potential hotspots of contamination in a FACS machine. Fluidic lines from the sheath- and sampling-port represented two potential sources of contamination.

Due to technical limitations, fluidic lines from the sheath tank could not be replaced with new and sterile tubing, and therefore represented a potential source of contamination. Another potential source of contamination is the fluidic lines from the sampling port because various types of samples have previously been passed through it; the FACS machines belonged to a core facility laboratory.

Processes to eliminate contaminating DNA

To circumvent the problems of contamination resulting from the FACS sorter, 20% sodium hypochlorite or bleach solution (v/v) was parsed through the fluidic lines of the sheath- and sampling-port of the FACS machine for an hour prior to cell sorting experiments. The same bleach solution was parsed through sampling port of the FACS machine for 30 minutes preceding the second round of cell sorting. DNA-free water was parsed through the fluidic lines for 30 minutes after each bleaching procedures. A list of actions that were implemented in this study to minimise contamination during sample preparation and cell sorting are listed in Table 4-5.

Table 4-5: Actions implemented to eliminate DNA contaminants in the FISH-FACS workflow

Potential sources of contaminating DNA	Recommended actions
Reagents: MDA kit, components of FISH hybridisations and washing buffer, PBS	<ul style="list-style-type: none"> - Use of molecular-grade reagents and sterile consumables which were DNA- and RNA-free - Reagents were dispensed into smaller, disposable sterile tubes in a clean BSC hood for one-time usage - Use of commercial MDA kits whose reagents have been UV-treated - MDA reactions were performed in a UV-treated BSC hood
Presence of eDNA or DNA from lysed cells in samples	- Potential DNA contaminants in samples were diluted with PBS buffer prior to cell sorting
FACS machine	- Bleaching the fluidic lines of the FACS machine with 20% sodium hypochlorite or bleach solution (v/v)
Probe sonicator tip	- Sonication in with 20% sodium hypochlorite or bleach solution (v/v)

Controls for detecting contamination

It was essential to include negative controls to validate the effectiveness of the sterilisation procedures and to detect for contaminants prior to cell sorting experiments. Ensuring minimal contamination before cell sorting is crucial because MDA amplification would be performed on sorted samples. The presence of DNA, which encompasses DNA contaminants in the samples, would be amplified together with the template DNA in MDA amplification reaction. The use of negative controls would further substantiate evidence that sequences obtained from sorted samples after MDA amplification and genomic sequencing would belong to the sorted cells. Negative controls were employed to detect contaminants originating from either the FACS machine, or the PBS buffer used for the collection of sorted cells. The negative controls were: (1)

PBS buffer in which cells would be sorted into; (2) sheath fluid and (3) PBS that have passed through the FACS machine and were collected prior to cell sorting.

Following collection of the negative controls, 16S PCR amplification with universal bacterial primer sets 27F/1492R was performed on negative controls that have been MDA-amplified to verify the presence of contamination. Absence of PCR amplicon products in the negative controls would indicate the absence of contamination (Figure 4-6). A positive control, which comprised of 100 cells of *Pseudomonas aeruginosa*, was used to demonstrate the efficiency of MDA in amplifying minute amounts of starting materials.

Figure 4-6: Ethidium bromide-stained agarose gel of negative controls that were alkaline-lysed, subjected to MDA amplification and PCR-amplified with primer set 27F/1492R to detect for contamination. Negative controls included: PBS buffer in which cells were sorted into (PBS); sheath fluid (SF) and PBS that was parsed through the sampling port (SP) and collected prior to cell sorting. L: 1 kbp ladder; +ve: 100 cells of *Pseudomonas aeruginosa*; Sorted: sorted cells

Downstream genomic sequencing would only be conducted on sorted samples if contamination was absent from the negative controls. However, the use of PCR amplification as a contamination assay has its limitations, as it can only qualitatively detect contaminants with an intact 16S rRNA gene locus that could be amplified with broad bacterial primer set 27F/1492R. Presence of contaminants in negative controls was further substantiated through genomic shotgun sequencing performed on negative controls that have undergone MDA amplification. RiboTagger

16S rRNA analysis did not detect any RiboTags, thus indicating that negative controls were relatively free from contamination. Therefore, sequences obtained from sorted samples would originate from the sorted target population.

4.3.3 Specific sorting of target population from an axenic culture

Fixation-free in-solution FISH

To develop and optimise sorting gates for the enrichment of *Thauera* from activated sludge, initial cell sorting validation was first performed on an axenic culture of R086 as a positive control. A fixation-free in-solution FISH protocol was used to label the target cells with two FISH probes conjugated with different fluorophores: probes Ribo_Thau1029_17_{Cy5} and EUB338_{A488}. Hybridisation of the probes produced a bright fluorescence signal for the target cells, but the fluorescence intensity was not evenly distributed within the cells (Figure 4-7A). Even in axenic culture, *Thauera* cells were present in aggregates that required disintegration to prevent clogging of the FACS sorter. Disintegration of cell aggregates was achieved through sonication, where the sonication duration was optimised to ensure that: (1) cell aggregates were broken up with minimal cell lysis and (2) probe-labelled cells would still emit high fluorescence intensity after the sonication process (Figure 4-7B).

Figure 4-7: Confocal micrographs demonstrating the process of sonicating dense clusters of *Thauera* cells from an axenic culture into a single-cell suspension for FACS sorting.

Thauera cells were hybridised with probes Ribo_Thau1029_17_{Cy5} (red) and EUB338_{A488} (green). *Thauera* cells appeared yellow/orange because of the merging of probe signals. *Thauera* cells visualised after: (A) in-solution FISH hybridisation and (B) probe sonication and dilution in PBS buffer. Bar: 5 μ m. Magnification: 63x.

Flow cytometric analysis

Sonicated samples of *Thauera* cells were sorted with the MoFlo XDP sorter using four types of sorting gates. Sorting gate 1 (Figure 4-8A) differentiates and sorts bacterial cells based on forward and side light-scattering properties. Sorting gate 1 was determined through comparison with a negative control of PBS (without cells) that was parsed through the MoFlo XDP. Big cell aggregates were filtered out through sorting gate 2 (Figure 4-8B) and sorting gate 3 (Figure 4-8C). Sorting gates 2 and 3 were constructed on the assumption that single cell shared a liner relationship between the forward or side scatter height and area. On the other hand, forward or side scatter height of cell aggregates will not change, but the forward or side scatter area will be larger than that of single-cells. Sorting gate 4 (Figure 4-8G) was constructed to isolate probe-labelled cells exhibiting high Cy5 and A488 fluorescence over the other non-labelled cells. Construction of sorting gate 4 was based on negative hybridisation controls: (1) no-probe control where no FISH probes were added (Figure 4-8D) and (2) non-specific controls where probe NON338 was labelled with either a Cy5 fluorophore or Alexa 488 fluorophore to estimate the levels of non-specific binding for the respective fluorophores (Figure 4-8E, F).

A clear and distinct population, with enhanced signal on the Cy5 and A488 axis appeared in sorting gate 4 after hybridisation of probes Ribo_Thau1029_17_{Cy5} and EUB338_{A488} (Figure 4-8G). Events collected in sorting gate 4 were gated above the fluorescence signal of the negative controls. No events were observed in the same sorting gate of the negative controls. Events that were not captured in sorting gate 4 represented cells that were not positively-labelled with FISH probes or background noise that resulted from the FISH-FACS procedure. Two rounds of sorting were performed on events captured in sorting gates 1-4. An initial round of sorting showed that the target population could be enriched from 39.3% to a purity of 99.5% (Figure 4-8H).

Figure 4-8: Flow cytometric analysis of sorting from an axenic culture of R086 hybridised with probes Ribo_Thau1029_17_{Cy5} and EUB388_{A488}.

An approximate ~100,000 events were collected for each FACS plot, except for the purity check. Sorting gates were outlined in black and values shown were indicative of the percentage of gated events over the total number of events. Flow cytometric analysis: (A) Sorting gate 1: sorting of bacterial cells based on forward versus side scatter; (B) sorting gate 2: filtering out cell aggregates based on side scatter area versus side scatter height; (C) sorting gate 3: filtering out cell aggregates based on forward scatter area versus forward scatter height. Sorting gate 4 was constructed to exclude events exhibiting Cy5 and A488 fluorescence signal in the negative controls: (D) no-probe control; (E) hybridisation with probe NON338_{Cy5} to control for non-specific binding for Cy5 fluorophore; (F) hybridisation with probe NON338_{A488} to control for non-specific binding for A488 fluorophore. (G) Events exhibiting Cy5 and A488 fluorescence signal above the cut-off threshold for the negative controls were collected in sorting gate 4; (H) purity of the sorted sample after an initial round of sorting.

Validating presence of target population in sorted samples

PCR-based screening was employed to validate the presence of *Thauera* in sorted samples prior to downstream library preparation and next generation sequencing, and it involved the screening of PCR amplicon products with primer set 27F/RT_Thau (Section 3.3.6). PCR products amplified with primer set 27F/RT_Thau and corresponding to a size of 1000 bp on ethidium bromide-stained agarose gel were indicative of *Thauera* present in the sorted sample. The consistent identification of *Thauera* (accession number: KP941745) in PCR amplicon products (n=27) corroborates with the findings. Presence of bacteria in a sorted sample could be determined with primer set 27F/1492R and the PCR amplicon product would correspond to a size of 1500 bp. However, the absence of PCR product obtained with primer set 27F/RT_Thau in the same sorted sample indicates the presence of contamination. For instance, cells sorted with 10, 100 and 1000 events yielded a product size of 1000- and 1500-bp when amplified with primer set 27F/RT_Thau and 27F/1492R

respectively (Figure 4-9). However, cells sorted with 1 event could only yield a product size of 1500 bp when amplified with the same primer sets, thus indicating bacterial contamination in this sample.

Figure 4-9: Ethidium bromide-stained agarose gel of sorted samples. 16S rRNA gene of the sorted samples was amplified with primer sets 27F/RT_Thau and 27F/1492R, yielding PCR amplicon products of 1000- and 1500-bp respectively. A product size of 1000- and 1500-bp indicates the presence of *Thauera* and bacteria respectively. In this example, all the sorted samples except the 1 event contained *Thauera* cells. L: 1 kbp ladder; 1000, 100, 10 and 1 event correspond to the number of events sorted into the tubes during FISH-FACS.

Bacterial contamination in the sorted sample of 1 event was eventually identified to be *Propionibacterium acnes* through Sanger sequencing. A cost-effective and upfront screening approach was adopted by only processing sorted samples that could be amplified with the primer set specific for the target taxon. Sorted samples that did not amplify with the specific primer set indicated that the target taxon was not detected in the sample.

Consistency of cell sorting

The ability of MoFlo XDP to consistently sort out small number of events was evaluated by sorting different number of events: 1, 5, 10, 100 and 1000 events per tube. A successful sorting event was defined by the ability to obtain a PCR amplicon product whose 16S rRNA taxonomic classification matches the target taxon. A positive control of R086 was used to evaluate the consistency of

sorting from the MoFlo XDP. Sorting experiments were performed on two separate days (two biological replicates).

Fifteen samples were collected on each sorting day and three technical replicates were collected from each of the following category: single-event sort; 5-events sort; 10-events sort; 100-events sort and 1000-events sort. Consistency of sorting was estimated to be $90 \pm 0.31\%$ ($n=30$, mean $\pm$ SD); sorting of cells into tubes for the different number of events was 100% consistent except for single-event sort. For single-cell sort, only 1 out of 3 tubes for biological replicate one and 2 out of 3 tubes for biological replicate two were positive for PCR amplicon products. With the inconsistency of the MoFlo XDP for single-cell sorting, 5 events would be set as the lowest number of events per tube for subsequent cell sorting experiments.

4.3.4 Evaluating the effectiveness of sorting from an axenic culture

Effectiveness of sorting from axenic cultures with the MoFlo XDP sorter was evaluated on two biological replicates that were sorted on different days using these criteria: (1) purity of the initial cell sort from flow cytometric analysis; (2) visual inspection of sorted cells through microscopy and (3) RiboTagger 16S rRNA analysis.

Purity of sorting

Flow cytometric analysis could only estimate the purity of sorting in the initial round, but not in the second round of sorting because samples would have to be processed for further downstream analysis. Purity of sorting was calculated by passing the sorted samples from the initial round of bulk sorting back into the MoFlo XDP, and calculating the percentage of events that mapped back to the sorting gates that were defined during flow cytometric analysis. Positively-labelled cells could be enriched from 49.60% to 97.80% ($n=2$) in the initial round of sorting (Figure 4-10). The reproducibility of achieving an enhanced purity for both replicates highlights the suitability of the MoFlo XDP for sorting of prokaryotic cells.

Figure 4-10: Cell sorting purity obtained from an initial round of sorting from an axenic culture of R086. R086 was hybridised with probes Ribo_Thau1029_17_{Cy5} and EUB338_{A488}. Purity was determined from flow cytometric analysis.

Microscopic visualisation of sorted samples

Quantitative FISH was not performed on sorted samples from the axenic culture because of the assumption that an axenic culture should contain only *Thauera* cells. This assumption was eventually verified to be true through RiboTagger 16S rRNA analysis performed in the later stages (Figure 4-12). Nevertheless, effectiveness of sorting was demonstrated through microscopic visualisation of probe-labelled cells in a sorted sample. Microscopic observation in at least 20 field-of-views confirmed the presence of positively-labelled single cells having a coccus morphology, with identical co-localisation in both the channels of Cy5 (red) and Alexa488 (green) filter set (Figure 4-11). This proved that sorting gates 2 and 3 helped in the removal of cellular aggregates, and sorting gate 4 was specific for the positively-labelled cells.

Figure 4-11: Confocal micrographs of probe-labelled cells sorted from sorting gates 1-4 after FISH-FACS of an axenic culture of R086.

R086 was dual hybridised with probes Ribo_Tag1029_17_Cy5 and EUB338_A488. Cells were visualised with: (A) Cy5 filter set (red); (B) Alexa488 filter set (green) and (C) overlap of both filters (yellow). Bar: 10 µm. Magnification: 63x.

RiboTagger 16S rRNA analysis

Sorted samples that produced a PCR amplicon product with primer set 27F/RT_Thau were processed for paired-end Mi-Seq sequencing. In addition, pre-sorted axenic cultures of R086 were also processed for paired-end Mi-Seq sequencing. The aims of sequencing pre- and post-sorted R086 samples were to verify that the: (1) axenic culture contained only a single taxon of *Thauera* or a single 33bp-RiboTag matching to *Thauera*, and (2) to analyse the distribution of RiboTags obtained after cell-sorting respectively. The latter aim determines the contribution of artefact

sequences catalysed by MDA reactions, assuming that the culture was indeed 'axenic' and no contamination was introduced to the sorted sample.

RiboTagger 16S rRNA analyses on pre-sorted sample revealed the presence of only one RiboTag which could be taxonomically annotated to *Thauera*, thus proving that pre-sorted samples contained only one taxon of *Thauera*. Interestingly, multiple RiboTags (n=36) with different sequences were present in sorted samples. Only one RiboTag could be taxonomically annotated, and this RiboTag had the same tag sequence present in pre-sorted samples. The other 35 RiboTags could not be taxonomically classified. RiboTag annotated to *Thauera* constituted the majority of RiboTags at $99.12\% \pm 1.25$ (n=24, mean $\pm$ SD). Although multiple RiboTags were present, $99.67\% \pm 0.99$ (n=24, mean $\pm$ SD) of total RiboTags in the sorted samples had the presence of probe Ribo_Thau1029_17 in their sequence (Figure 4-12). Collectively, these results showed that a high purity of the target population in sorted samples was achieved through effective cell sorting.

Figure 4-12: Relative abundance of RiboTags matching to probe Ribo_Thau1029_17 and *Thauera*-specific OTUs in pre-sorted and sorted samples from axenic cultures of R086.

RiboTags whose sequences: (1) did not match probe Ribo_Thau1029_17 and (2) matched probe Ribo_Thau1029_17 but were not annotated to *Thauera* were of particular interest. RiboTagger 16S rRNA analysis for the negative controls revealed the absence of RiboTags, and pre-sorted

samples revealed the presence of only one RiboTag. Therefore, the presence of other RiboTags in the sorted samples might be attributed to artefact sequences formed during MDA amplification. These RiboTags contribute to 0.88% of the total RiboTags in the sorted sample and did not have any taxonomic affiliation in the SILVA database (Figure 4-13).

Figure 4-13: Distribution of RiboTags obtained from sorted samples of R086 into various categories.

4.3.5 Specific sorting of a target taxon from the floccular sludge community

In-solution fixation-free FISH

After optimising the FISH-FACS methodology and having defined sorting gates that led to the successful enrichment of target population from an axenic culture, the same methodology was applied to cell sorting from activated sludge samples. Activated sludge samples were hybridised with probes Ribo_Thau1029_17_{Cy5} and EUB338_{A488}, and it resulted in target cells being positively labelled (Figure 4-14A). The same probe sonication method was used to break up cellular aggregates that contained the target taxon into a homogenous suspension of single cells (Figure 4-14B). Although small cellular aggregates of other bacterial cells could be observed, these aggregates did not contain the target taxon and could subsequently be filtered out using the FACS sorting gates.

Figure 4-14: Confocal micrographs demonstrating the process of sonicating dense clusters of activated sludge flocs into single-cell suspension for FACS sorting.

Thauera cells were hybridised with probes Ribo_Thau1029_17_{Cys} (red) and EUB338_{A488} (green). *Thauera* cells appeared yellow/orange because of the merging of probe signals. *Thauera* cells visualised after: (A) in-solution FISH hybridisation; (B) probe sonication and dilution in PBS buffer. Bar: 5 µm. Magnification: 63x.

Flow cytometric analysis

Identical sorting gates (sorting gates 1-4) that were previously defined in Section 4.3.3 of the thesis were applied to cell sorting from activated sludge (Figure 4-15). Similar to sorting from axenic cultures, sorting gates were correctly gated with the aid of negative controls to exclude any events appearing in the sorting gate of the negative controls. The forward versus side scatter of activated sludge samples (Figure 4-15A) appeared different from the axenic culture (Figure 4-8A) with more events appearing in the upper quadrant of the side scatter. This was expected as cells and particles of various sizes are present in activated sludge. A distinct population appeared in sorting gate 4 after the hybridisation of probes Ribo_Thau1029_17_{Cys} and EUB338_{A488}. Compared to the axenic cultures, more background noise or negatively-labelled cells could be seen on the exterior of sorting gate 4 (Figure 4-15G). Similar to previous experiments, two rounds of sorting were performed on events captured in sorting gates 1-4. An initial round of sorting showed that the targeted population could be enriched from 0.91% to 97.3% after an initial of sorting (Figure 4-15H).

Figure 4-15: Flow cytometric analysis of sorting *Thauera* from an activated sludge sample hybridised with probes Ribo_Thau1029_17_{Cy5} and EUB388_{A488}.

An approximate ~100,000 events were collected for each FACS plot, except for the purity check. Sorting gates were outlined in black and values shown indicate the percentage of gated events over the total number of events. Flow cytometric analysis: (A) Sorting gate 1: sorting of bacterial cells based on forward versus side scatter; (B) sorting gate 2: filtering out cell aggregates based on side scatter area versus side scatter height; (C) sorting gate 3: filtering out cell aggregates based on forward scatter area versus forward scatter height. Sorting gate 4 was constructed to exclude events exhibiting Cy5 and A488 fluorescence signal in the negative controls: (D) no-probe control; (E) hybridisation with probe NON338_{Cy5} to control for non-specific binding for Cy5 fluorophore; (F) hybridisation with probe NON338_{A488} to control for non-specific binding for A488 fluorophore. (G) Events exhibiting Cy5 and A488 fluorescence signal above the cut-off threshold for the negative controls were collected in sorting gate 4; (H) purity of the sorted sample after an initial round of sorting.

4.3.6 Evaluating the effectiveness of sorting from activated sludge samples

Effectiveness of sorting from a mixed microbial community in activated sludge was evaluated with the same parameters as described in Section 4.3.4. The parameters: (1) purity of the initial cell sort from flow cytometric analysis and (2) RiboTagger 16S rRNA analysis were further supplemented with quantitative FISH analysis. Quantitative FISH analysis measured the relative abundance of *Thauera* in pre- and post-sorted samples. These parameters were recorded in biological triplicates of activated sludge samples sorted on 3 different days (Table 4-1).

Purity of sorting

From an initial abundance of 1.14%, 1.39% and 0.91%, the target population could be enriched to a purity of 88.7%, 97.6% and 97.3% in the respective biological replicates after an initial round of sorting (Figure 4-16).

Figure 4-16: Cell sorting purity obtained from the initial round of sorting from activated sludge samples. Samples were hybridised with probes Ribo_Thau1029_17_{Cy5} and EUB338_{A488}. Purity was calculated from flow cytometric analysis.

Purity of sorting was influenced by two factors. The first factor involved the presence of non-target cells appearing on the exterior of sorting gate 4 (Figure 4-17A). The non-target cells appeared to be background noise or cells that could potentially be located in close proximity with the target cells. The second factor involved drifting of the 640 nm laser, which subsequently lowered the percentage of the target population being included in sorting gate 4. A lower percentage of Cy5-labelled cells were excited by the 640 nm laser and more negatively labelled-Cy5 cells could be seen accumulating in the left quadrant of the FACS plot (Figure 4-17B). Fortunately, the drifting of the 640-nm laser only occurred during the second round of sorting. Hence, an enriched target population could still be obtained due to the high purity obtained from the initial round of sorting.

Figure 4-17: Factors affecting the purity of cell sorting.

The two factors were: (A) presence of background noise or non-target cells in close proximity of the target cells as demarcated with arrows and (B) drifting of the 640-nm laser that reduced the percentage of Cy5-labelled cells in the sorting gate.

Quantitative FISH image analysis

Quantitative FISH image analysis was performed on pre- and post-sorted samples to determine the enrichment level of the target population. Total cell abundance was measured with probe EUB338. Ratio of the biovolume between probe-labelled *Ribo_Thau1029_17* and EUB338 cells represented the relative abundance of *Thauera* in the sample. Biovolume of *Thauera* in pre-sorted samples was estimated to be approximately $1.06\% \pm 0.23$ ($n=135$, mean $\pm$ SEM) (Figure 4-20A). As floccular sludge is heterogeneous in its spatial distribution of microorganisms, some microscopic images contained large clusters of *Thauera* cells (Figure 4-18A) while other images contained smaller clusters or single cells (Figure 4-18B). Therefore, standard deviation for biovolume quantification is expected to be large because of the random acquisition of images of the flocs. Consequently, standard error of the mean (SEM) instead of standard deviation was used in biovolume measurement.

Figure 4-18: Confocal micrographs depicting the heterogeneous distribution of *Thauera* in activated sludge. Activated sludge samples were hybridised with probes Ribo_Thau1029_17_{Cy5} (red) and EUB338_{A488} (green). *Thauera* cells appeared yellow because of the merging of probe signal. (A) Large and (B) smaller clusters of *Thauera* cells were present in images acquired for quantitative FISH analysis. Bar: 10 µm. Magnification: 63x.

An alternate approach was adopted to calculate the relative abundance of *Thauera* in sorted samples with the aim of reducing image acquisition and image processing time associated with 3D images. As sorted samples comprised of mostly single cells (Figure 4-19), relative abundance of *Thauera* was estimated by dividing the number of probe Ribo_Thau1029_17-labelled cells over the number of EUB338-labelled cells. This was performed by acquiring multiple 2D images of the sorted cells in random XY axis, and then using DAIME software to calculate the ratio of probe-labelled cells.

Figure 4-19: Confocal micrographs of probe-labelled cells sorted from sorting gates 1-4 after FISH-FACS of an activated sludge sample.

Activated sludge sample was hybridised with probes Ribo_Thau1029_17_{Cy5} and EUB338_{A488}. Cells were visualised with (A) Cy5 filter set (red), (B) Alexa488 filter set (green) and (C) overlap of the two filters. Bar: 10 μm. Magnification: 63x.

Thauera had been enriched up to 93x after FISH-FACS from a relative abundance of $1.06\% \pm 0.23$

($n=135$, mean $\pm$ SEM) in pre-sorted samples to $98.66\% \pm 0.67$ ($n=90$, mean $\pm$ SEM) in sorted samples (Figure 4-20B).

Figure 4-20: Quantitative FISH analysis depicting the relative abundance of *Thauera* in pre-sorted and sorted samples from activated sludge.

Relative abundance of *Thauera* in: (A) pre-sorted and (B) post-sorted samples. Quantitative FISH analysis was performed by image acquisition, followed by image analysis. Ratio of probe Ribo_Thau1029_17-labelled cells over probe EUB338-labelled cells was calculated through image processing software. Each dot represents quantitative FISH analysis performed on one confocal image.

RiboTagger 16S rRNA analysis

A total of 10 V6 sequence tags were produced from the pre-sorted samples, and none of the sequence tags matched to Ribo_Thau1029_17 or *Thauera*-specific OTUs. Only 5 RiboTags were present in pre-sorted samples: 2 RiboTags had bacterial taxonomic annotation to *Aquabacterium* and *Mycobacterium*; 1 RiboTag had a eukaryotic taxonomic annotation to *Chaetonotida* and the other 2 RiboTags did not have any taxonomic affiliation. The small number of V6 sequence tags identified from pre-sorted samples is likely due to the shallow sequencing depth: average coverage of $1.38X \pm 0.68$ (Appendix A-5).

A total of 39,134 V6 sequence tags were present in MDA-amplified sorted samples, with 76 unique OTUs identified. Three out of four OTUs could be taxonomically classified at the genus level: two OTUs ($n=38,181$ V6 sequence tags) to *Thauera* and an OTU to *Leptolyngbya* ($n=24$ V6 sequence tags). The other OTU could only be classified to the family level: Rhodocyclaceae ($n=7$ V6 sequence tags). The remaining 72 OTUs ($n=922$ V6 sequence tags) lacked taxonomical affiliations even at the lowest taxonomic level. MDA amplification performed on sorted samples was the rationale for observing more RiboTag sequences. Despite multiple OTUs present in sorted samples, high specificity of probe Ribo_Thau1029_17 for its target taxon was evidenced as $98.99\% \pm 4.07$ and $96.48\% \pm 8.69$ ($n=32$, mean $\pm$ SD) of total V6 sequence tags matched to probe Ribo_Thau1029_17 and *Thauera*-specific OTUs respectively (Figure 4-21).

Figure 4-21: Relative abundance of RiboTags annotated to probe Ribo_Thau1029_17 and *Thauera* in pre- and post-sorted samples from activated sludge samples.

4.3.7 Comparing environmental specificity of FISH probes

Environmental specificity of probe Ribo_Thau1029_17 for *Thauera* in activated sludge was evaluated in biological triplicates in the previous section. In this section, environmental specificity of probe Thau646 was compared against probe Ribo_Thau1029_17 using the same methodology. The goal is to test if the actual specificity of RiboProbe for *Thauera* in activated sludge matches the predicted *in silico* specificity as outlined in Section 3.3.1. For simplicity of comparison, only one sorted sample (replicate 3) enriched by RiboProbe Ribo_Thau1029_17 would be used. $63.01\% \pm 33.82$ ($n=12$, mean $\pm$ SD) and $97.25\% \pm 6.56$ ($n=12$, mean $\pm$ SD) of total RiboTags from probe Thau646 and Ribo_Thau1029_17 sorted samples were annotated to *Thauera*, with a statistical significant difference ($p=0.0025$) between the means of the two samples (Figure 4-22). In conclusion, probe Ribo_Thau1029_17 had a significantly higher specificity than probe Thau646 for *Thauera* species in activated sludge samples.

Figure 4-22: Comparison of the environmental specificity of probes Ribo_Thau1029_17 and Thau646 for *Thauera* in activated sludge samples. The legend denotes the different number of events sorted from the FACS machine.

For a low specificity probe like probe Thau646, sorting of 10 events was the optimal number of events that led to higher abundance of target taxon in the sorted samples. Sorting of higher number of events (100 or 1000 events) lowered the abundance of target taxon, whereas sorting of lower number of events (5 events) led to inconsistent results. In contrast, sorting with a more specific probe like probe Ribo_Thau1029_17 - with the exception of a sample of 1000 events - consistently led to high abundance of target organism in the sorted samples.

Aquabacterium and *Nitrosomonadaceae* were identified by RiboTagger 16S rRNA analysis to be the two most abundant non-target taxa to be consistently sorted with probe Thau646. Presence of probe Thau646 binding site on the 16S rRNA gene sequence of the non-target taxa was hypothesised to be the rationale for the non-specific sorting. To investigate this hypothesis, partial 16S rRNA gene sequences from non-target taxa were extracted from the assembled metagenomic contigs. Homology search has identified Thau646 binding site in the V4 region of the 16S rRNA sequence of *Aquabacterium*.

4.3.8 Phylogeny of *Thauera* obtained from sorted sample

To determine the phylogeny of *Thauera* captured by FISH-FACS with probe Ribo_Thau1029_17, 56 clones were randomly selected from a 16S rRNA gene clone library constructed from a FACS sorted sample of 1000 events. This sorted sample was particularly selected because RiboTagger 16S rRNA analysis had revealed the presence of only two RiboTags that could be taxonomically annotated to *Thauera* (Table 4-6).

Table 4-6: Sequence and relative abundance of RiboTags of the sorted sample used for phylogenetic analysis

RiboTags annotated to <i>Thauera</i>	RiboTag's relative abundance (%)
GTGTTCTGGCTCCCGAAGGCACCCTCG <u>C</u> CTCTC	96.50
GTGTTCTGGCTCCCGAAGGCACCCTCG <u>G</u> CTCTC	3.50

Probe sequence of Ribo_Thau1029_17 used for FISH-FACS is demarcated in red and SNP between the two RiboTags is underlined.

Out of the 56 random clones, 2 sequences comprising of partial full-length sequences (<1200 bp) were rejected and the remaining 54 sequences were selected for phylogenetic analysis. *De novo* OTU picking (Caporaso *et al.*, 2010; Edgar, 2010) was performed on the 54 clone sequences at 97% and 99% sequence identity, which corresponded to species and sub-species level OTUs (Stackebrandt and Goebel, 1994; Větrovský and Baldrian, 2013). Clustering of 54 clones into 1 OTU at 97% sequence identity gave an indication that clone sequences obtained were highly similar with each other. Since near full-length sequences (>1200 bp) obtained from the clones provide a higher taxonomic resolution, clone sequences were eventually clustered at 99% sequence identity as a proxy to sub-species level resolution.

50 clone sequences were clustered into OTU 1, and 4 clone sequences were clustered into OTU 2. Clones 8 and 49 were selected as representative sequence for OTU 1 and 2 respectively. Both representative sequences were assigned to the genus *Thauera* using the SINA tool (Pruesse *et al.*, 2012), with the representative sequences having a sequence similarity of 99.46-99.87% to its closest neighbours in the SILVA database. Closest neighbours to the representative sequences in the phylogenetic tree were observed to be from the activated sludge ecosystem. Both

representative sequences possessed the 33bp-RiboTag sequence that represented the major *Thauera* taxon in sorted sample. However, probe Thau646 binding site was absent in representative sequence OTU2_clone49. Other members of OTU2 also revealed the absence of probe Thau646 binding site. A homology search of probe Thau646 binding site of members of OTU2 using the ARB software revealed a single nucleotide polymorphism (SNP) at the start of the probe binding site (Table 4-7).

Table 4-7: Single nucleotide polymorphism (A→C) observed in members of OTU2

OTUs	Homology of probe Thau646 binding site	No of clones
OTU1	<u>A</u> AGGCTAGAGTACGGCAGA	50
OTU2	<u>C</u> AGGCTAGAGTACGGCAGA	4

SNP in probe Thau646 binding site is underlined.

Although both representative sequences exhibited high similarity in their 16S rRNA gene sequences (98.73%), OTU2 formed a separate clade away from OTU1 and this was supported by a high bootstrap value of >90% (Figure 4-23). The short branch length of the representative sequences showed that *Thauera* species present in the sorted samples do not deviate phylogenetically from its closest relatives in the SILVA database.

Figure 4-23: Maximum-likelihood (PhyML) phylogenetic tree depicting the 16S rRNA phylogenetic relationship of representative sequences (99% sequence identity cut-off) obtained from clone libraries of probe Ribo_Thau1029_17 sorted sample and its closest relatives.

Only near full-length sequences (≥ 1200 bp) were used for phylogenetic analysis. 16S rRNA gene sequences of representative sequences are demarcated in red. Closely related sequences were obtained from the SILVA SSU Ref NR 99 database (version 123). Members of the genus *Fusobacterium* were used as the outgroup. Bootstrap values were calculated from 1000 bootstrap analysis and only bootstrap values $> 50\%$ are displayed. Branches with low bootstrap values $\leq 50\%$ have been multifurcated. The scale bar represents substitutions per nucleotide base. Legend of the various bootstrap values is located at the upper left-hand corner of the diagram.

4.4 Discussion

Targeted enrichment of specific taxon using RiboProbe is inherently a challenging task. Multiple factors influence the purity of sorted samples obtained from FISH-FACS sorting: (1) sensitivity of FACS sorter in detecting prokaryotic cells; (2) specificity of FISH probes for its target taxon and labelling them with a fluorescence signal that can be distinguished from the background noise; (3) presence of contaminants introduced during the sorting process; (4) laser drifting during FISH-FACS sorting; (5) ability of FACS sorter to accurately and consistently sort cells of interest into a small volume of buffer in a collection tube; (6) successful lysis of the isolated cells and (7) high-fold amplification of genomic material from lysed cells. The goal of this chapter is to incorporate RiboProbe into a reproducible and robust pipeline for an effective and contamination-free targeted cell sorting.

4.4.1 Selection of an effective FACS sorter

While many studies have described FACS to perform cell sorting, it is still incoherent how certain FACS machines are selected over others for the detection of small particles. An easy approach is to adhere with FACS machines described in papers that have supported the sorting of prokaryotic cells. However, this is often difficult to achieve because purchasing a FACS machine is a pricey affair (unit cost: approximately SGD \$500,000.00), and access to a dedicated FACS sorter for prokaryotic cell sorting is often limited due to the innate fear of cross-contamination with eukaryotic cell sorting – even if proper decontamination procedures are employed. Two FACS machines (Sy3200 versus Mo Flo XDP) equipped with different configurations were compared on their sensitivity in detecting bacterial cells from an axenic culture through flow cytometric analysis. This is an important step in describing an effective FISH-FACS methodology for the selective enrichment of the target taxon.

Even though the MoFlo XDP was equipped with a larger nozzle size of 100 μm , it was evident that the small size particle detector in the forward scatter proved to be a component critical for the detection of prokaryotic cells as observed by the purity obtained after an initial round of sorting. While Rinke *et al.*, (2014) mentioned that prokaryotic cells could be sorted with other types of cell

sorters, this study showed that not all cell sorters were suitable for high-purity prokaryotic cell sorting, and users who want to perform small particle sorting are encouraged to use a FACS machine equipped with a small size particle detector in the forward scatter. A detailed investigation into the configurations of FACS sorters in studies which had success in prokaryotic cell sorting showed that these machines were also equipped with a small size particle detector that led to good flow cytometric detection and subsequently optimal sorting purity (Yilmaz *et al.*, 2010b; Gougoulas and Shaw, 2012). Both FACS sorter of Beckman Coulter (MoFlo XDP) and Becton Dickinson (Influx) mentioned in the studies were additionally equipped with the small size particle detector. In addition, the MoFlo XDP was calibrated with 0.4 μm and 0.88 μm fluorescent beads. This further extends the sorting capability of FACS to other small particles that fall within the size range of the beads.

Single-cell sorting was not pursued due to the inconsistency of obtaining bacterial amplicon products with the MoFlo XDP. Single-cell sorting inconsistencies were also presented in another study where only two out of eight single-cell sorts resulted in an amplification product (Kvist *et al.*, 2007). In another study, less than 1% success rate was achieved in single-cell sorting of TM6 cells (McLean *et al.*, 2013). There are two plausible explanations for the absence of amplicon product with single-cell sort. First is the lack of template for MDA reaction, where a cell deposited in a single event is present in picolitre-scale of sheath fluid, and they often tend to stick to the side of the walls of the tube. Due to the small volume, sheath fluid evaporates rapidly and the cell remains embedded on the side of the tube even after centrifugation. Second is due to representation bias catalysed by MDA reaction where the 16S rRNA gene is not amplified, and hence a lack of amplicon product. Representation bias has been shown to be more pronounced in single-cell MDA experiments.

4.4.2 Effectiveness of sorting with RiboProbe

The main novelty of this study compared to other FISH-FACS studies was the application of RiboProbe for the specific sorting of a target taxon. One advantage of RiboProbe over canonical probe is the higher *in silico* specificity as shown in the previous chapter, but hitherto has not been proven experimentally. To prove it experimentally, RiboTagger and canonical probes were applied in separate FISH-FACS experiments and tested on their ability to enrich for *Thauera* cells from a mixed microbial community in activated sludge. FISH-FACS of a target taxon from activated sludge is extremely challenging given the presence of auto-fluorescent particles and high abundance of non-target cells (Wallner *et al.*, 1997). Therefore, optimisation of the FACS machine to correctly define the sorting gates for the target taxon, with the appropriate negative controls was first established on an axenic culture as a positive control, and subsequently on activated sludge samples.

Prior to FACS sorting, the target taxon has to be labelled with FISH probes, and this was achieved using a fixation-free in-solution FISH protocol as described by Haroon *et al.* (2013) where *Thauera* cells in axenic cultures and activated sludge samples could be positively identified through confocal microscopy. This protocol was adopted because previous studies have shown that samples treated with preservatives (PFA) commonly used in fixation of samples for FISH were not suitable for MDA amplification. This is due to protein and DNA cross-linkages that inhibit phi29 polymerase from accessing the DNA (Clingenpeel *et al.*, 2014). In addition, it was observed through our study that cell aggregates were harder to break up through probe sonication and fixed sorted samples produced a lower DNA yield (data not presented).

It was essential that target cells were hybridised with FISH probes, labelled with different fluorophores to different regions of the 16S rRNA in a 'multiple probe approach' to increase the confidence of accurately identifying the target taxon (Amann *et al.*, 1990a). The 'multiple probe approach' can also be translated to cell sorting because the FACS sorter contains many laser units and detectors to handle multi-parametric measurements. Microscopic observation of probe-labelled cells from axenic cultures showed that the fluorescent signal intensities were not

homogenously distributed throughout the cell population, and this resulted in some cells with displaying weaker fluorescence intensity (Figure 4-7). This was assumed to be due to insufficient concentration of probes used that led to some cells not being fully hybridised with the probes (Zwirgmaier *et al.*, 2003). In contrast, *Thauera* cells in activated sludge samples were bright fluorescently labelled and could be easily distinguished from the background noise and non-target cells (Figure 4-14). Brightness of probe-labelled cells has been associated with the number of ribosomes that the target organism possessed, which in turn is correlated to the physiological state of the cells (Gifford *et al.*, 2013). This suggests that *Thauera* plays an active role in wastewater treatment at UPWRP. This is verified by the fact that *Thauera* is a dominant member of the UPWRP floccular sludge community as determined by quantitative FISH analysis ($1.06 \pm 0.23\%$), and previous studies have shown *Thauera* to degrade aromatic compounds under denitrifying conditions in wastewater plants (Jiang *et al.*, 2012).

The successful hybridisation of target cells using fixation-free in-solution FISH protocol with RiboProbe was also reflected in flow cytometric analysis. FACS scatterplots of Cy5 versus Alexa488 revealed a distinct target population that was observed above the fluorescence intensity thresholds of the negative controls, which were used to check for auto-fluorescence and non-specific probe binding. Fluorophores have been shown to differ in their propensity to bind non-specifically to the same substrates (Zanetti-Domingues *et al.*, 2013). Therefore, it was essential to control for non-specific binding using nonsense probe NON338 attached with Cy5 and Alexa488 fluorophores. As the probe sequence of NON338 is the reverse complement of probe EUB338, probes EUB338 and NON338 should not be co-hybridise together so as to prevent the formation of primer dimers.

Previous studies have used either flow cytometric analysis or quantitative-FISH analysis to determine the efficacy of cell sorting (Wallner *et al.*, 1997; Yilmaz *et al.*, 2010b; Haroon *et al.*, 2013b; Lee *et al.*, 2015). Both analyses provide a general estimate of the levels of enrichment obtained through cell sorting, but neither of the analyses quantifies the level of enrichment of the

target taxon. False-positive labelling of non-target taxa by FISH probes cannot be discriminated with both analyses. These limitations are usually overcome by identifying conserved phylogenetic marker genes such as the 16S rRNA gene through clone libraries (Podar *et al.*, 2007; Gougoulas and Shaw, 2012) or quantitative-PCR (Bruder *et al.*, 2016). While the methods mentioned above could help in quantifying the relative abundance of target taxon in the sorted samples, these methods are limited by primer bias. These limitations were overcome using the primer-free RiboTagger 16S rRNA survey to determine the purity of sorted cells. Although a low mean read depth of sequencing was used as estimated in pre-sorted samples (Appendix A-5), MDA amplification could overcome the low sequencing depth and RiboTagger analysis could be used to profile the diversity captured in sorted samples. Collectively, the analyses used in this study are more comprehensive in determining the efficacy of cell sorting and enrichment levels of a target population than other studies performed so far.

4.4.3 Managing DNA contamination

The management of DNA contamination is important in MDA-amplification experiments because of the propensity of MDA to amplify minute quantity of DNA templates, including DNA contaminants. Subsequently, contamination would complicate downstream analysis of genomes of the target taxa (Woyke *et al.*, 2010). The high abundance of *Neisseriaceae*, *Malassezia* and *Propionibacterium* in Sy3200-sorted samples showed how easy it was for sorted samples to be contaminated. RiboTags assigned to these contaminants made up more than half the sorted population even though sorting was performed on an axenic culture (Figure 4-4). The contaminants were phylogenetically different from *Thauera*: *Neisseriaceae* is categorised under a different order; *Malassezia* is categorised under the Fungi Kingdom and *Propionibacterium* is categorised under a different phylum.

Minimising contamination is a challenging task, as there are numerous potential hotspots of contamination which include: (1) the sample (e.g. eDNA); (2) sample handling; (3) laboratory reagents and (4) cell sorting process. Difficulties encountered in minimising contamination were reflected in many studies (Binga *et al.*, 2008; Blainey, 2013). To ensure that MDA reactions were

performed only on genomic template of target cells, DNA contamination was effectively minimised through a multitude of methods which have been adopted and summarized in Table 4-5. Rinke *et al.*, (2014) have described a comprehensive protocol for the sorting of individual prokaryotic cells from environmental samples, which resulted in the successful amplification of 201 uncultivated microbial cells from 29 previously undiscovered branches of the tree of life (Rinke *et al.*, 2013). However, access to a dedicated sterile environment (e.g. clean UV-treated PCR hood, clean room with positive-pressure airflow) and FACS machine for MDA experiments as recommended by Rinke *et al.*, (2014) was limited in our experiments. Managing DNA contamination proved to be challenging, but a robust experimental design which encompassed three types of negative controls were used to overcome the limitations mentioned above. The negative controls aided in: (1) qualitative detection of bacterial contamination; (2) evaluation of the effectiveness of the measures taken to minimise contamination and (3) strengthening of the hypothesis that amplified genomic fragments obtained from MDA reactions originated from the sorted cells.

The negative controls included in this study were more comprehensive than the study of Rinke *et al.*, (2014) where only a single negative control was employed. The negative control used by Rinke *et al.*, (2014) was TE buffer in which a single cell would be sorted into. Without a dedicated sterile FACS machine, it was necessary to employ additional negative controls: (1) sheath fluid that had passed through the fluidic lines and (2) PBS that had passed through the sampling lines of the FACS machine prior to cell sorting. Another difference between this study and Rinke *et al.*, (2014) is the use of sodium hypochlorite to bleach both the fluidic lines of the sheath- and sampling-port (Figure 4-5). As Rinke *et al.*, (2014) had a dedicated FACS sorter, sodium hypochlorite was used only for bleaching of the fluidic lines of the sheath port. In this study, bleaching of the sampling-port was necessary because the MoFlo XDP was a shared equipment used for the sorting of other specimens.

The current gold standard for detecting contaminating DNA is to screen for bacterial 16S rDNA amplicon product. However, this method is limited as it cannot detect: (1) contamination whose 16S rDNA sequence cannot be amplified by broad bacterial primers 27F/1492R or (2) contamination that belonged to the Archaea or Eukaryote domain of life. Although no 16S rRNA amplicon products were obtained from the negative controls, bands were observed with the genomic DNA ScreenTape (Agilent, USA) – an indication of genomic fragments. Additional RiboTagger 16S rRNA analysis performed on the negative controls did not yield RiboTags, therefore suggesting the controls were relatively free from contamination. Therefore, the bands observed in the genomic DNA ScreenTape probably originated from artefact sequences, which are often by-products of MDA amplification (Lasken and Stockwell, 2007).

4.4.4 Environmental specificity of RiboProbe

In the previous chapter, quantitative co-localisation analysis of probes Ribo_Thau1029_17 and Thau646 in activated sludge showed that approximately 18% of cells labelled by probe Thau646 did not overlap with probe Ribo_Thau1029_17. This group of cells has been hypothesised to represent non-*Thauera* cells. This hypothesis was validated by hybridizing activated sludge samples with either probe Ribo_Thau1029_17 or Thau646. Subsequently, probe-labelled cells were sorted in separate FISH-FACS experiments, and the sorted cells were subjected to MDA amplification and metagenomic shotgun sequencing. The aim was to determine the ‘environmental specificity’ of the probe, or the bacterial diversity captured with the FISH probe in the activated sludge ecosystem (Gougoulis and Shaw, 2012). RiboTagger 16S rRNA analysis revealed that 63% and 97% of RiboTags could be annotated to *Thauera* from probe Thau646 and Ribo_Thau1029_17 sorted samples respectively. 30% of RiboTags from probe Thau646 sorted sample could be annotated to other non-target organisms such as *Aquabacterium* and *Comamonadaceae*. No RiboTags from probe Ribo_Thau1029_17 sorted samples could be annotated to other organisms with established taxonomy. The remaining RiboTags were assumed to either represent novel organism which were co-sorted together with *Thauera* cells, or artefacts induced by primer dimer formation from MDA amplification reactions. The high percentage of

non-target cells in probe Thau646 sorted samples strengthened the hypothesis that cells which did not overlap in microscopic images - as calculated by MCC analysis - belonged to non-target taxa.

Presence of probe Thau646 binding site in the 16S rRNA gene of distantly-related non-target cells was hypothesised to be responsible for the high percentage of non-target cells in the sorted sample. Probe Thau646 binding site on non-target taxa could not be identified with RiboTagger analysis because the RiboTagger software only extracts 16S rRNA gene from the highly variable regions. To test the hypothesis, near full-length 16S rRNA gene sequence were extracted from assembled contigs of probe Thau646 sorted samples. Probe Thau646 binding site could be located in the 16S rRNA gene of *Aquabacterium*. This result corroborates with the findings from RiboTagger analysis that majority of non-target organisms in the sorted sampling were annotated to *Aquabacterium*.

These results showed that probe Ribo_Thau1029_17 has a higher environmental specificity than probe Thau646 for *Thauera* in activated sludge samples. Despite having an *in silico* specificity of only 16.68%, *Thauera* could be enriched to an average of 96.58% based on RiboTagger analysis. These findings were in agreement with Gougoulis and Shaw (2012), whom had defined the environmental specificity of FISH probes to be more vital than it's *in silico* specificity. Non-specific taxa predicted to be targeted *in silico* by probe Ribo_Thau1029_17 did not exist or were present in extremely low abundance relative to the *Thauera* population in UPWRP's floccular sludge community.

The use of a curated ecosystem-centric 16S rRNA database specific for the floccular sludge community would benefit future FISH probe design, as more accurate information pertaining to the *in silico* specificity and coverage can be assigned to FISH probes. Currently, such database does not exist, but there is the Microbial Database for Activated Sludge (MiDAS) field guide that was established to improve microbial ecology of activated sludge by correlating the identity, taxonomy and functional aspects of microorganisms involved in critical wastewater treatment processes

(McIlroy *et al.*, 2015). MiDAS field guide provides a shared platform to provide access into established wastewater microorganisms, but a genomic database would greatly enhance future work that requires identification or comparative genomic analysis of newly discovered microorganisms.

4.4.5 Phylogenetic analysis of *Thauera*

Two RiboTags were annotated to *Thauera* from probe Ribo_Thau1029_17 sorted samples. The most abundant RiboTag annotated to *Thauera* (96.50%) represented the original 33bp-RiboTag from which probe Ribo_Thau1029_17 was initially designed from. The next RiboTag annotated to *Thauera* was present at a lower abundance of 2.77% and the RiboTag sequences differed by a single nucleotide polymorphism (SNP). Hitherto, it was unclear how organisms represented with the RiboTag would be placed phylogenetically based on its 16S rRNA gene sequence. An insight into the phylogenetic relationship of *Thauera* beyond the taxonomic resolution of 33bp-RiboTag was obtained through the gold standard of generating clone library from a probe Ribo_Thau1029_17 sorted sample. The lower abundant RiboTag annotated to *Thauera* was not present in the clone sequences. This is probably due to the insufficient number of clones generated. Clustering of OTUs at 99% sequence identity resulted in two OTUs: OTU1 was the dominant OTU which consisted of 50 clones; members of OTU2 did not contain probe Thau646 binding site in its 16S rRNA gene sequence. SNP (A→C) observed in the homology binding site of probe Thau646 in the four members of OTU2 suggest that the SNP was unlikely due to artefacts introduced by PCR or sequencing errors. Micro-diversity of closely related *Thauera* species seems to be a more plausible explanation for the observed SNP.

In conclusion, phylogenetic placement of the two OTUs into different clades could not have been elucidated with the limited resolution of the 33bp-RiboTag. This is in agreement with studies that have shown that accurate phylogenetic placement of bacteria in the tree of life requires near full-length 16S rDNA sequences (Yarza *et al.*, 2014). Generation of clone library still remains to be the gold standard in generating full-length sequences even though 16S rRNA sequences can be extracted from *de novo* assemblies, as demonstrated with the non-target taxa in probe Thau646

sorted samples. However, the presence of multiple copies of 16S rRNA gene and conserved regions in the 16S rRNA gene complicates the process of *de novo* assembly (Albertsen *et al.*, 2013a). This limitation was exemplified in this study where only one 16S rRNA gene sequence could be extracted from the sorted sample that produced two different OTUs of *Thauera* from clone libraries. Finally, the absence of probe Thau646 binding site in some of the clone sequences highlights the need to constantly evaluate the *in silico* specificity and coverage of probes – even for published probes (Loy *et al.*, 2008).

4.4.6 Summary

In this chapter, a strategy for the specific enrichment of a target organism from a mixed microbial community in activated sludge was demonstrated. This was achieved through the coupling of fluorescently-labelled RiboProbe with a high-sensitive FACS machine that supports the sorting of prokaryotic cells (FISH-FACS). FISH-FACS resulted in highly-enriched, fixative-free samples of *Thauera* from both axenic cultures ($99.12\% \pm 1.25$) and activated sludge ($96.48\% \pm 8.69$) that were further amenable to downstream genomic sequencing and analyses. *Thauera* was chosen as a reference organism to demonstrate the targeted sorting with highly specific RiboProbe, but the same strategy can be extrapolated to other bacterial taxa and this would be demonstrated in the subsequent chapter.

Chapter 5 Characterisation of unclassified bacteria taxa in activated sludge systems

5.1 Introduction

Significant efforts have been invested into expanding the phylogeny of the prokaryotic tree of life, which was primarily achieved through the discovery of organisms that possessed novel full-length SSU rRNA gene sequences (<97% similarity) compared to curated 16S rRNA databases. According to a study performed by Yarza *et al.* (2014), the estimated rate of newly detected species has been estimated to be approximately 4×10^4 per year. There are many benefits with expanding the 16S rRNA database. For instance, higher quality FISH probes and PCR primers could be designed with better precision in probe/primer coverage and specificity to increase the confidence of *in situ* hybridisation and PCR experiments (Karst *et al.*, 2016a). Before the advent of metagenomics, the only information that could be concluded from novel organisms was the phylogenetic relationship with other closely related organisms on the basis of the full-length 16S rRNA gene sequence, and visualisation of its morphology and spatial interaction with its neighbour from design of FISH probes.

With recent advancements in metagenomics sequencing and bioinformatics tools, obtaining draft genomes from novel organisms is now a possibility and subsequent genomic analysis would provide insights into the functional potential of the novel organisms. For example, members of the novel phylum Hyd24-12 were discovered to play a role in acidogenesis and fermentation in anaerobic digesters through genomic analysis of draft Hyd24-12 genomes (Kirkegaard *et al.*, 2016). In a separate study, novel *Nitrospira* species involved in the commamox process were uncovered through the discovery of a complete complement of genes on their draft genome that were involved in ammonia oxidation: ammonia monooxygenase and hydroxylamine dehydrogenase and nitrite oxidation: nitrite oxidoreductase (van Kessel *et al.*, 2015).

From a wastewater perspective, most novel organisms that played an important role in wastewater ecology were often discovered due to their dominating abundance in the microbial community. High abundance of the novel organisms could be attributed to the inherent selective

pressure of wastewater treatment systems, or it could be due to the effects of manipulation on the operating parameters of enrichment system. Quite often, the discovery of novel organisms is an accidental process where operating conditions of enrichment system were intentionally designed to select for a different target group of organisms. For instance, the role of *Candidatus Halomonas phosphatis* (Nguyen *et al.*, 2012) and *Tetrasphaera* (Nguyen *et al.*, 2011) as putative polyphosphate accumulating organisms (PAOs) were unravelled in full-scale EBPR plants because their abundance was substantially higher than that of the well-characterised PAO organism: *Candidatus Accumulibacter phosphatis*. In another study, *Candidatus Propionivibrio aalborgensis* was accidentally discovered as a novel glycogen accumulating organism (GAOs) in a laboratory SBR operated to enrich for PAOs (Albertsen *et al.*, 2016). Accidental enrichment of novel organisms often led to the discovery of novel organisms that either played an important role in the functioning of the intended bioprocess (e.g. *Tetrasphaera* in the removal of phosphate), or inhibiting the intended bioprocess (e.g. *Candidatus Propionivibrio aalborgensis* as a GAO that competes with PAOs). An advantage associated with this approach is that the role of the novel organism has a direct impact on the functioning of the bioprocess because of its relative high abundance, and correlating the identity of the organism to its function is an easier task due to the ease of obtaining its draft genome. This approach is termed as 'targeted metagenomics' (Hess *et al.*, 2011) and the process typically involves shotgun sequencing and genomic binning - a bioinformatics analysis that clusters together phylogenetically-related genomic contigs (Tyson *et al.*, 2004).

On the other hand, organisms that play an important ecological role in the ecosystem but are present at low abundance (<1% relative abundance) might be missed out by targeted metagenomics. For example, *Nitrosomonas* has been previously been shown to be an important ammonia-oxidizer in wastewater treatment systems (Nielsen *et al.*, 2009), but it was ranked 539th on the basis of 16S-V6 rDNA abundance in UPWRP's floccular sludge community. Due to the current limitations of metagenomics analysis and genomic binning, recovery of genomes to study the potential function for low abundant organism is difficult (Albertsen *et al.*, 2013a). 'Targeted

enrichment' has been proposed as a solution to enrich for specific target population, so as to obtain a 'sub-metagenome' prior to genomic sequencing and analysis (Sekar *et al.*, 2004). Ecological questions on low abundant novel organisms cannot be answered through traditional enrichment system of bioreactors due to inadequate knowledge on their ecophysiology. Therefore, targeted enrichment through FISH-FACS methodology that was optimised in the previous chapter would be applied for enrichment of the novel organisms. Activated sludge has been previously been shown to harbour taxonomically unclassified microorganisms and this was demonstrated through the high richness of unclassified taxa present in UPWRP's floccular sludge community. In this context, both the identity and functions of the unclassified taxa remained elusive because of the limited resolution of the RiboTag with a short-read length of 33bp.

In this chapter, two unclassified bacterial taxa were selected for characterisation based on their high abundance (0.77% and 2.35%) in UPWRP's floccular sludge community at the point of ultra-deep sequencing to saturation. Pseudonyms for the unclassified bacterial taxa are as followed: UPWRP_1 and UPWRP_2 until the establishment of their 16S rRNA phylogenetic relationship in Sections 5.3.6 and 5.3.14. Relative abundances mentioned above correspond to UluPandan_1 and UluPandan_2 respectively. Morphology and spatial interaction of the unclassified bacterial taxa in floccular sludge would be microscopically analysed through FISH experiments. As an initial step towards identifying the organisms, FISH-FACS method was applied to enrich for these novel organisms from the floccular sludge community. The gold standard of obtaining full-length 16S rRNA gene sequences was performed on enriched, sorted samples to determine the phylogenetic relationship of the unclassified bacterial taxa. Additionally, draft genomes of the organisms were generated from the enriched samples through the process of *de novo* assembly and genomic binning. Obtaining the draft genome is a prelude to understanding the metabolic functions of these organisms and a succinct genome annotation will be performed to shed light on the potential metabolic function of these organisms. The flow of the thesis is organised as follows: findings pertaining to UPWRP_1 would be described in Sections 5.3.2–5.3.9, followed by the findings of UPWRP_2 as described in Sections 5.3.10–5.3.18. Section 5.3.19 would be dedicated

to the comparison of genome completeness, genome contamination and quality of assembly between the draft genomes generated from this study and single-cell genomics.

5.2 Methods

A flowchart describing the processes involved in characterisation of the unclassified bacterial taxa is presented in Figure 5-1.

Figure 5-1: Flowchart describing the processes used in the enrichment and subsequent characterisation of unclassified bacterial taxa from activated sludge samples.

5.2.1 Sample preparation

Sample acquisition

Activated sludge samples were sampled at the identical location as described in the Section 3.2.1.

Biological and technical replicates used for quantitative FISH analysis, genomic DNA extraction and cell sorting are presented in Table 5-1.

Table 5-1: Biological and technical replicates used for analysis of unclassified bacterial taxa

Samples	Biological replicates and date of sampling	Number of technical replicates		
		DNA extraction*	FISH analysis *	FACS sorting**
UPWRP_1	August 5 th , 2016 (Replicate 1)	3	3	1
	August 11 th , 2016 (Replicate 2)	3	3	1
	August 23 rd , 2016 (Replicate 1)	3	3	1
UPWRP_2	August 29 th , 2016 (Replicate 2)	3	3	1

*Performed on pre-sorted samples

**Refers to the number of sample used as input for FACS sorting with the MoFlo XDP sorter

DNA extraction and sample fixation

DNA extraction and sample fixation were performed according to protocols as described in Section 4.2.1.

5.2.2 Fluorescence in situ hybridisation

Fixation-free in-solution FISH protocol as described in Section 4.2.2 was applied for FISH-FACS.

Slide-FISH protocol as described in Section 4.2.2 was applied for *in situ* visualisation and quantification. UPWRP sludge samples were used for generating probe dissociative curves for newly-designed probes: Ribo_Unk1029_17 and Ribo_Halia1029_17, and the protocol has been adopted from Section 3.2.5.

FISH probes

FISH probes used for the characterisation of UPWRP_1 and UPWRP_2 are presented in Table 5-2.

Table 5-2: FISH probes and RiboTags used for the characterisation of UPWRP_1 and UPWRP_2

Probes	Intended target	Sequence of probe	Formamide	Reference
	taxon	(5'→3')	(%)	
EUB338	Most bacteria	GCTGCCTCCCGTAGGAGT	0	(Amann <i>et al.</i> , 1990a)
NON338	None	ACTCCTACGGGAGGCAGC	45	(Wallner <i>et al.</i> , 1993)
Halian2 [^]	<i>Haliangium</i>	CCGACTTCTAGAGCAACTG A	25	(McIlroy <i>et al.</i> , 2014)
Halian3 [^]	<i>Haliangium</i>	CCAGTCACTCTTTAGGCGG C	25	(McIlroy <i>et al.</i> , 2014)
Ribo_Unk1029 _33*	UPWRP_1	TGCTTCGCGTCTCCGAAGA GCCGACCACCTTTC	N.D	This study
Ribo_Unk1029 _17	UPWRP_1	TGCTTCGCGTCTCCGAA	50	This study
Ribo_Halia1029 _33*	UPWRP_2	TCTCACTCGCTCCCGAAGG CACCCCGACATCTC	N.D	This study
Ribo_Halia1029 _17	UPWRP_2	TCTCACTCGCTCCCGAA	40	This study
Halia183	UPWRP_2	GAAATCCGGAAACCTCACA GAC	40 [#]	This study

[^]: Probes are components of the HalianMix

* Original 33bp-RiboTag or V6 sequence tag of the RiboProbe

[#]: Optimal formamide concentration was not empirically determined

N.D: Not determined

Probe Halia183 was designed using the 'Probe Design' tool of the ARB software that comprised of sequences from the SILVA SSU Ref NR 99 (version 123) database. Representative sequence obtained from the clone library of probe Ribo_Halia1029_17 sorted sample was used as the target sequence for the design of probe Halia183.

Breaking up of cell aggregates

Two different methods were adopted for disrupting the cell aggregates. Probe sonication method with the same sonication duration and power setting as described in Section 4.2.2 was used to disrupt cell aggregates of UPWRP_2. Passing the sample repeatedly through a fine syringe needle (26G x ½", Terumo, Japan) was used to break up aggregates of UPWRP_1.

5.2.3 Fluorescence activated cell sorting (FACS)

FACS sorting was performed according to protocols as described in Section 4.2.4.

5.2.4 Multiple displacement amplification (MDA)

Multiple displacement amplification was performing according to protocols as described in Section 4.2.5.

5.2.5 Generation of clone libraries and phylogenetic analysis

Generation of clone libraries was performed according to protocols as described in Section 4.2.6.

Phylogenetic analysis was performed according to protocols as described in Section 4.2.7, where the differences were: (1) clustering of clones into OTUs with 99% sequence similarity, and (2) taxonomical classification of representative sequences were searched with a minimum sequence identity of 80% and 90% for UPWRP_1 and UPWRP_2 respectively using the web-based SINA tool.

Pairwise sequence analysis of 16S rRNA gene sequences

Similarity between 16S rRNA gene sequences was calculated using pairwise sequence analysis with the stand-alone version of the SINA software (Pruesse *et al.*, 2012).

5.2.6 Metagenomics sequencing and analysis

Metagenomic sequencing and trimming of reads

Illumina TruSeq Nano DNA sample preparation protocol was used for sequencing library preparation. Genomic sequencing of MDA-amplified samples was performed on a HiSeq 2500 (Illumina, USA), producing paired-end reads with a read length of 250 bp. The number of pre- and post-sorted samples processed for genomic sequencing is presented in Table 5-3. Adapter- and quality-trimming of reads were performed as described in Section 4.2.8.

Table 5-3: Number of pre- and post-sorted samples processed for HiSeq genomic sequencing

Number of samples		
	UPWRP_1	UPWRP_2
Pre-sorted samples	6	6
Post-sorted samples	16	16

Calculating sequencing coverage

Average sequencing coverage was estimated according to protocols as described in Section 4.2.8.

RiboTagger 16S rRNA analysis

RiboTagger 16S rRNA analysis was performed according to protocols as described in Section 4.2.9.

Metagenomic assembly, contamination screening and genomic binning

Quality-trimmed reads were assembled into contigs using the SPADES software (Nurk *et al.*, 2013) with the following settings: --sc and --careful. Contigs with a length of less than 1 kbp were removed using the 'anvi-script-reformat-fasta' script (Eren *et al.*, 2015). Remaining contigs were screened for contaminants using the ACDC software (Lux *et al.*, 2016) with the default settings. Contaminants were detected using a reference-free methodology that encompassed tetramer frequencies clustering, non-linear dimensionality reduction of oligonucleotide sequences using Barnes-Hut approximation (BH-SNE) (van der Maaten, 2014) and further clustering algorithms using the DIP-statistic test (Hartigan and Hartigan, 1985; Kalogeratos and Likas, 2012) or CC-statistic test (Von Luxburg, 2007) with machine learning techniques. Subsequently, clusters of

contigs were exported from the ACDC software, and completeness and contamination level of the genomes were estimated.

Metagenomic binning was performed on the extracted cluster from the ACDC software using the MetaBat software (Kang *et al.*, 2015) with the following parameters: -B 200 and –superspecific. Prior to binning, a sorted BAM file from individual samples was generated from the assemblies with BBDuk tools (BBMap- Bushnell B. – <http://sourceforge.net/projects/bbmap>). Draft genomes recovered from binning were examined for completeness and contamination with CheckM. Visualisation of differential coverage binning was performed using the mmgenome toolbox (Karst *et al.*, 2016b).

Estimating completeness and contamination levels of genomes

Completeness and contamination of the draft genomes were assessed with CheckM (Parks *et al.*, 2015), which estimates the proportion of lineage-specific marker genes. Completeness and contamination of draft genomes were further validated by measuring the fraction of essential single-copy genes conserved in 95% of bacteria (Dupont *et al.*, 2012) using the Anvi'o software (Eren *et al.*, 2015).

Phylogenomic analysis of draft genomes

Draft genomes were phylogenetically assigned using a concatenated set of 43 phylogenetic conserved marker genes and placed in the reference genome tree that contained a set of 2,502 complete genomes and 3,604 draft genomes from the IMG database using the CheckM software (Parks *et al.*, 2015). Phylogenetic relationship of the draft genomes in the reference genome tree was presented using the ARB software (Ludwig *et al.*, 2004).

Computing genome similarity

Similarity between draft genomes of the unclassified bacterial taxa and its closest neighbour was calculated using average nucleotide identity (ANI) and amino acid identity (AAI). ANI was predicted with Pairwise ANIb calculation with Jspecies (Richter *et al.*, 2016). AAI was calculated with the AAI calculator (Rodriguez-R and Konstantinidis, 2016) using a minimum identity of 20% and a

minimum alignment score of 50. Output from the 'two-way AAI' was used to estimate average AAI between ORFs of the two genomes in comparison.

5.2.7 Taxonomic assignment of ORFs

Open reading frames (ORFs) were predicted and translated from the draft genomes of the unclassified bacterial taxa and its closest neighbour using Prokka (Seemann, 2014) with the default parameters. ORFs were annotated against the NCBI NR database (downloaded on November 2016) using the DIAMOND software (Buchfink *et al.*, 2015) with a stringent e-value threshold of 10^{-6} and a minimum percentage identity of 30% (Pearson, 2013). Taxonomic assignment of the ORFs were conducted by parsing the output of Diamond into Megan 6.0 (Huson *et al.*, 2016) with the following modification to the default LCA parameters: min support=1.0%, followed by the mapping of reads with NCBI taxonomy using the 'prot_acc2tax' mapping file.

5.2.8 Coding potential of ORFs

ORFs were submitted to the RAST (Rapid Annotation using Subsystem Technology) server (Aziz *et al.*, 2008) for genome annotation using the manually-curated SEED subsystem-based approach (Overbeek *et al.*, 2005). The "function-based comparison" tool of RAST was used to compare genomic features of the draft genomes.

5.2.9 Nucleotide sequence accession numbers

Near full-length 16S rDNA sequences obtained from the clone libraries of UPWRP_1 and UPWRP_2 were deposited in NCBI's 16S prokaryotic ribosomal RNA database with the following accession numbers: KX907544–KX907601 and KX954221–KX954279 respectively. Draft genomes of *Candidatus Shimingles merlion*, *Candidatus Shimingles singa* and *Haliangium clusterio* were deposited in NCBI under project IDs: PRJNA359419, PRJNA380517 and PRJNA359471 respectively.

5.3 Results

5.3.1 Probe design

RiboProbes were designed for the detection and subsequent characterisation of two unclassified bacterial taxa: UPWRP_1 and UPWRP_2 present in UPWRP's floccular sludge community. The two unclassified taxa were selected based on their high abundance in saturated metagenomic dataset (year: 2011) as described by Kjelleberg *et al.* (unpublished). 33bp-RiboTags representing the unclassified bacterial taxa could not be taxonomically classified with the SILVA database using the RiboTagger software and 'TestProbe' tool of the SILVA database. Truncation of the length of the 33bp-RiboTag: Ribo_Unk1029_33 that was annotated to UPWRP_1 into probe Ribo_Unk1029_17 (Figure 3-3) showed that the FISH probe could not be affiliated with any taxonomy using the 'TestProbe' tool of the SILVA database. Therefore, the bacterial population targeted by probe Ribo_Unk1029_17 is considered a novel taxon with a high degree of certainty.

Truncation of the 33bp-RiboTag: Ribo_Halia1029_33 that was annotated to UPWRP_2 into probe Ribo_Halia1029_17 showed that in addition to UPWRP_2, three additional *Haliangium* species were targeted *in silico* in the SILVA database (Appendix A-6). Ribo_Halia1029_33 represents a novel OTU of the genus *Haliangium*, of the family Haliangiaceae, of the order Deltaproteobacteria and of the phylum Proteobacteria.

UPWRP_1

5.3.2 Visualisation of UPWRP_1

Members of UPWRP_1 hybridised by probe Ribo_Unk1029_17_{Cy5} were observed to possess a filamentous morphology whose dimensions varied: the longest filament observed was 52 µm x 0.85 µm and the shortest was 4.7 µm x 0.74 µm. Aggregates of filamentous bacteria could occasionally be visualised (Figure 5-2A) but most members existed as single cells (Figure 5-2B).

Figure 5-2: Confocal micrographs of members of UPWRP_1 hybridised with probes Ribo_Unk1029_17_{cys} (red) and EUB338_{A488} (green) in UPWRP activated sludge samples.

Members of UPWRP_1 appeared orange because of the merging of probe signals. (A) Aggregates and (B) single cell of UPWRP_1 in a fixed sample of activated sludge. Bar: 5 µm. Magnification: 63x.

Quantitative FISH performed on two biological replicates of UPWRP activated sludge samples showed that UPWRP_1 constituted between 0.50-0.70% of the total bacterial biomass (Figure 5-5A).

5.3.3 Calibration of probe Ribo_Unk1029_17

Melting curves of probe Ribo_Unk1029_17 were obtained from technical triplicates of UPWRP sludge because no pure culture of UPWRP_1 existed. Members of UPWRP_1 were observed to be embedded heterogeneously in different depths of the flocs, which in turn affected the brightness of target cells. Therefore, the brightness of target cells at a specific formamide concentration would vary considerably (Figure 5-3). Despite variations in mean fluorescence intensity of the target cells between technical replicates, the highest formamide concentration before a large dip in mean fluorescence intensity was between 50–55%. An optimal formamide concentration of 50% was eventually selected for subsequent hybridisation experiments with probe Ribo_Unk1029_17 because artefacts that did not resemble the morphology of UPWRP_1 appeared at a formamide concentration of 55%.

Figure 5-3: Probe dissociation curve of probe Ribo_Unk1029_17 performed on technical triplicates of UPWRP activated sludge samples.

An optimal formamide concentration of 50% was estimated for Ribo_Unk1029_17.

5.3.4 Sorting of UPWRP_1

Prior to cell sorting, floccular aggregates have to be broken up to a suspension of mostly single cells or smaller aggregates. The probe sonication method described in Section 4.2.2 was initially applied, but it resulted in the filamentous bacteria being sheared into very small filaments that could not be easily detected under confocal microscopy and flow cytometric analysis (Figure 5-4A). Sorting gate constructed on probe fluorescence intensities of probes Ribo_Unk1029_17_{Cy5} and EUB338_{A488} did not reveal any distinct population when compared with the negative controls (Figure 5-4B). Thus, members of UPWRP_1 could not be sorted using the sonication method.

To overcome this problem, an alternative approach of breaking the floccular aggregates by passing the hybridised sludge samples through a fine syringe needle (26G x ½") repeatedly was adopted. Microscopic observation showed that hybridised cells that passed through the syringe method had longer filaments, and they were observed to be present at high abundance as compared to the sonication method (Figure 5-4C). Flow cytometric analysis of the sludge sample broken up by the syringe method showed a distinct population, thus enabling the subsequent sorting of the

target cells (Figure 5-4D). The entire flow cytometric plots of UPWRP_1 are presented in Appendix A-7.

Figure 5-4: Confocal micrographs of the different processes used to obtain single cell suspension of UPWRP_1, and flow cytometric plots that represented the processes.

Members of UPWRP_1 have been hybridised with probes Ribo_Unk1029_17_{Cy5} (red) and EUB338_{A488} (green) in UPWRP activated sludge samples. Members of UPWRP_1 appeared yellow because of the merging of probe signals and are demarcated by an arrow. (A) Confocal microscopy of UPWRP_1 in a sonicated sample of sludge; (B) flow cytometric scatterplot (A488 vs Cy5) of sonicated sludge sample; (C) confocal microscopy of UPWRP_1 in a sludge sample that has been broken up by passing the flocs repeatedly through a 26G x ½" syringe needle; (D) flow cytometric scatterplot (A488 vs Cy5) of sludge sample disrupted by the syringe method needle. Bar: 5 μm. Magnification: 63x.

5.3.5 Evaluation of cell sorting effectiveness of UPWRP_1

The effectiveness of cell sorting was evaluated with the methods described in Section 4.3.6.

Quantitative-FISH estimated that the target population was enriched with an enrichment factor of 161x, from an abundance of $0.60 \pm 0.070\%$ ($n=140$, mean $\pm$ SEM) to $96.53 \pm 2.45\%$ ($n=24$, mean $\pm$ SEM; Figure 5-5A). Fewer microscope field-of-views ($n=24$) were taken of post-sorted samples because of the time constraint encountered during the sorting of the low abundant UPWRP_1. Based on the defined sorting gates in flow cytometric analysis, an initial round of sorting enhanced the abundance of the target population from 0.033% to 41.90% ($n=2$; Figure 5-5B). Non-target events present outside of the sorting gate were responsible for lowering the purity of cell sort.

Despite the low cell purity obtained after an initial round of sorting, RiboTagger 16S rRNA analysis showed that RiboTags annotating to probe Ribo_Unk1029_17 and UPWRP_1 could be enriched from 0% (before sorting, $n=6$) to $99.9\% \pm 0.277$ ($n=16$, mean $\pm$ SD) and $99.56\% \pm 0.819$ ($n=16$, mean $\pm$ SD) respectively (Figure 5-5C). A total of 15 RiboTags and 6,317 RiboTag sequences were produced in post-sorted samples. No RiboTags were found to be annotated to probe Ribo_Unk1029_17 and UPWRP_1 in pre-sorted samples due to its low abundance and an insufficient sequencing depth used: an average coverage of $2.56X \pm 2.97$ (Appendix A-5).

Figure 5-5: Effectiveness of cell sorting and levels of enrichment for UPWRP_1 from activated sludge samples were estimated with quantitative-FISH analysis, flow cytometric analysis and RiboTagger 16S rRNA analysis.

(A) Quantitative FISH analysis depicting the relative abundance of UPWRP_1 in biological duplicates of activated sludge samples in pre-sorted and sorted samples. Quantitative FISH analysis was performed by acquiring confocal images, followed by biovolume image analysis. Ratio of probe Ribo_Unk1029_17-labelled cells over probe EUB338-labelled cells was calculated through image processing software. Each dot represents quantitative FISH analysis performed on one confocal image. (B) Cell sorting purity obtained after an initial round of sorting was calculated from flow cytometric analysis. (C) Relative abundance of RiboTags annotated to Ribo_Unk1029_17 and UPWRP_1 in pre- and post-sorted samples.

5.3.6 16S rRNA phylogeny of UPWRP_1

The 16S rRNA phylogenetic relationship of the novel bacterial taxon was determined through the construction of a clone library from a sorted sample, where 100% of V6 sequence tags matched UPWRP_1. Out of 60 random selected clones, 2 clone sequences comprising of partial full-length sequences (<1200 bp) were rejected and the remaining clone sequences were selected for phylogenetic analysis. *De novo* OTU picking was performed on the clones at 99% sequencing identity. The representative sequence was selected from the most abundant sequence of the OTU. *De novo* OTU picking of sequenced 16S rRNA library clones at 99% sequence similarity resulted in 4 representative OTUs: OTU1_Clone36, OTU2_Clone31, OTU3_Clone9 and OTU4_Clone14. The four OTUs were subdivided into 2 “clades”, with 16S rRNA sequences having sequence similarity ranging from 98.11-98.97%.

Using the SILVA taxonomy, representative sequences were proposed to be affiliated to the order Sphingobacteriales, of the class Sphingobacteriia and of the phylum Bacteroidetes. No further taxonomical information could be inferred on the family or genus level. A review of the closest neighbours corroborate with the unassigned nomenclature at the family level. The closest cultured isolates to the 4 OTUs was *Portibacter lacus* (accession no: AB675658), with a sequence similarity of 82.12-82.52%. The highest sequence identity to published sequences in the SILVA database was between 90.63–91.11%. Phylogenetic analysis of the 16S rDNA sequences showed that the representative sequences formed an independent monophyletic clade that clustered outside its closest relative: Sphingobacteriales clone BT-62 (accession number: KP411858) that was obtained from a sequencing batch reactor (Figure 5-6). No other sequences were present in the monophyletic clade which was supported by the high bootstrap value of >90%.

Figure 5-6: Maximum-likelihood (PhyML) phylogenetic tree depicting the 16S rRNA phylogenetic relationship of representative sequences of OTUs (99% cut off) obtained from clone libraries of sorted samples hybridised with probe Ribo_Unk1029_17, and its closely related sequences in the SILVA database. Only near full-length sequences (≥ 1200 bp) were selected for phylogenetic analysis. Representative sequences of OTUs are demarcated in red, and sequences of the closest cultured isolates are demarcated in blue. Closely related sequences were obtained from the SILVA Ref NR 99 database (version 123). Members of the genus *Chlorobium* were used as the outgroup. Bootstrap values were calculated from 1000 bootstrap analyses and only bootstrap values over 50% were displayed. Branches with low bootstrap values $\leq 50\%$ have been multifurcated. The scale bar represents substitutions per nucleotide base. Legend of the various bootstrap values is located at the upper left-hand corner of the diagram.

5.3.7 Metagenomic analysis of UPWRP_1

Metagenomic assembly

Although 16 sorted samples were collected from the FISH-FACS experiments, only 4 samples were selected for co-assembly into contigs due to computer memory constraints of using the SPAdes software. From an initial metagenomic co-assembly constructed from 4 sorted samples containing 23,169,652 quality-trimmed, paired-end reads, 1,302 contigs were generated with a total length of 12.7 Mbp, an N50 contig size of 32,430 bp and the longest contig with a length of 145,517 bp (Appendix A-8).

Quality-control processing of genomic sequences

Some of the contigs in the co-assembly were hypothesised to be sequencing artefacts produced from MDA amplification or contaminants introduced into the sample. To verify this hypothesis, contigs from a co-assembly of the three types of negative controls used for detecting contaminations during the FISH-FACS procedure (Section 4.3.2) were visualised in a differential coverage plot (Figure 5-7).

Figure 5-7: A differential coverage plot of contigs from co-assembly of the negative controls. Contigs are represented by circles and coloured in accordance to their GC content. Contigs did not cluster together based on their coverage profile and GC content.

Bacterial DNA contaminants were absent in the negative controls and this was verified through the absence of 16S rDNA amplicon products from PCR amplification with conserved bacterial primer set and absence of RiboTags from RiboTagger 16S rRNA analysis. The co-assembly produced 112 contigs with a total length of 251,856 bp, N50 of 2,462 bp and the largest contig with a length of 6,776 bp (Appendix A-9). Contigs did not cluster together on the basis of their coverage or GC content. Even though proper management of DNA contamination was established during the FISH-FACS sorting process, the presence of contigs in the negative controls warrants the need for quality-control processing and decontamination of contigs.

Given the novelty of UPWRP_1 from its 16S rRNA sequence, an unsupervised machine learning reference-free method that integrates dimensionality reduction (BH-SNE) and clustering algorithms using tetramer frequencies and the DIP-statistical clustering was employed on the co-assembly of contigs from the sorted samples. Three clusters of contigs were identified (Figure 5-8) and the contigs were exported and evaluated with CheckM for contamination and completeness (Table 5-4).

Figure 5-8: Decontamination of the contigs of UPWRP_1 produced three clusters of contigs. Decontamination was performed with the ACDC software which employs tetramer frequencies and DIP statistics, followed by visualisation with BH-SNE. Clusters of contigs were represented by different colours.

Table 5-4: Genomic statistics for the clusters of contigs produced from the decontamination process

	Cluster of contigs		
	Blue	Green	Purple
Completeness (%)	2.82	98.28	0
Contamination (%)	0.16	94.48	0
GC (%)	60.15	45.9	35.55
Number of contigs	139	768	395
N50 (bp)	2,206	35,511	3,980
Longest contig (bp)	8,915	145,517	52,361
Taxonomy*	k__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__Myxococcales	k__Bacteria; p__Bacteroidetes; c__Sphingobacteriia; o__Sphingobacteriales	NA

* Represents different taxonomic levels. K: kingdom; p: phylum; c: class; o: order

NA: Not available

Two clusters: blue and green could be taxonomically classified using a conserved set of 43 phylogenetic markers (Parks, Imelfort et al. 2015). Taxonomy of contigs from the blue cluster matched another multiplexed sample in the same sequencing flow cell (UPWRP_2). The cross-contamination probably originated from sequencing library preparation or bulk amplification of sequencing libraries (Kircher *et al.*, 2012), rather than contaminants introduced during FISH-FACS. Contigs from the purple cluster could not be taxonomically classified and contigs were hypothesised to be artefacts of MDA: primer dimer or chimeric sequences (Binga, Lasken et al. 2008). This hypothesis was further substantiated with the analysis of a set of 111 essential single-copy genes (Dupont, Rusch et al. 2012), where none of the essential single-copy genes were present in the cluster.

In contrast, the green cluster contained multiple genomes and this was shown through the concept of essential single-copy genes. Out of a total of 207 essential single-copy genes that were documented, 104 were observed to be unique essential single-copy genes. The total number of

essential single copy genes (n=207) exceeded the number of essential single copy-genes (n=111) conserved in 95% of bacteria. The presence of duplicated genes (n=103) provided further evidence of multiple genomes in the co-assemblies. Presence of multiple genomes was further substantiated with the CheckM software that showed a contamination of 94.48%.

As the MetaBat binning software does not provide visualisation of the contigs, multiple genomes in the green cluster were visualised with the mmgenome software. Two clusters of contigs could be visualised in a differential coverage plot using the set of sorted samples: UPWRP0508161KB and UPWRP1108161KB (Figure 5-9A). These samples were selected because they provided the best clustering of contigs in the differential coverage plot. The plot also showed the assignment of most of the essential single-copy genes to two separate clusters (Figure 5-9B).

Figure 5-9: Visualisation of the differential coverage binning plot of contigs from UPWRP_1. (A) Two clusters of contigs representing different genomic bins could be visualised; (B) duplicate sets of essential single-copy genes were present on two separate clusters; red lines connected duplicated sets of essential single-copy genes. Each scaffold is represented by a circle that is scaled by their length, and coloured based on the taxonomic classification of essential single-copy genes. Only scaffolds of >3000 bp were displayed. Contigs were assembled from 4 MDA-amplified, sorted samples obtained after FISH-FACS. The differential coverage plot was visualised with the mmgenome software.

Genomic binning of UPWRP_1

Genomes in the green cluster were de-convoluted using genomic binning with Metabat, which employs an empirical probabilistic distance of genome abundance/contig coverage and tetranucleotide frequency to separate genomes. Genomic binning performed solely based on tetranucleotide frequency led to poor deconvolution of the genomes, as it produced a single

genomic bin with similar level of contamination as the green cluster of contigs. When genome abundance/contig coverage was factored into the binning process, two genomic bins with a similar GC content of ~46%, more than 90% completeness and less than 4% contamination were recovered. The two genomic bins were designated as genomic bin A and B of UPWRP_1 (Table 5-5). Nomenclature of genomic bins A and B of UPWRP_1 will henceforth be assigned as *Candidatus Shimingles merlion* and *Candidatus Shimingles singa* respectively.

Table 5-5: Statistics for the draft genomes of UPWRP_1

	Draft genomes	
	Bin A	Bin B
	<i>Candidatus shimingles merlion</i>	<i>Candidatus shimingles singa</i>
Number of contigs	178	173
Genome size (Mbp)	4.92	4.49
GC content (%)	46.02	45.69
N50 of contigs (bp)	38,494	38,959
Largest contig (bp)	145,517	106,171
Number of essential single-copy genes	76/111	92/111
Number of unique single-copy genes	73	90
Genome completeness	90.28	91.46
Genome contamination	3.94	2.96
Number of unique phylogenetic markers	19/43	29/43
Number of ORFs	3,576	3,339
Number of tRNA	40	36

Number of rRNA	2	NA
Phylogenetic placement in reference genome tree	k__Bacteria; p__Bacteroidetes; c__Sphingobacteriia; o__Sphingobacteriales; f__Saprospiraceae (sister lineage)	

NA: Not available

5.3.8 Phylogenomics analysis of UPWRP_1

Using a concatenated set of 43 phylogenetic conserved marker genes with CheckM, both sets of genomic bins were phylogenetically classified to the order Shingobacteriales, of the class Sphingobacteriia and of the phylum Bacteroidetes. The closest genome to *Ca. Shimmingles* was identified to be *Lewinella persica* DSM 23188 (IMG 2515154070) from the sister lineage genus *Lewinella*, under the family Saprospiraceae (Figure 5-10). Based on the phylogenomics analysis, the genomic bins can be categorised as a novel genus under a novel family. *Ca. Shimmingles merlion* and *Lewinella persica* DSM 23188 shared a 16S rRNA sequence similarity of 81.95%. However, 16S rRNA similarity between *Ca. Shimmingles merlion* and *Ca. Shimmingles singa* could not be estimated because a 16S rRNA sequence could not be extracted from *Ca. Shimmingles singa*.

Figure 5-10: Phylogenetic placement of genomics bins of *Candidatus Shimmingles* in the reference genome tree using a concatenation of 43 phylogenetic marker genes.

A set of 2,502 complete genomes and 3,604 draft genomes from IMG database were used in the reference genome tree. Genomic bins of *Ca. Shimmingles* are demarcated in red. Closest relative to the genomic bins was *Lewinella persica* DSM 23188 (IMG 2515154070). Closest outgroup was assigned as UID152 and it corresponded to different classes under the phylum Bacteroidetes.

Genome comparison with average nucleotide identity (ANI) and amino acid identity (AAI)

Ca. Shimingles merlion and *Ca. Shimingles singa* shared a genome similarity of 63.21% and 63.32% average nucleotide identity (ANI) with *Lewinella persica* DSM 23188 respectively. Both draft genomes shared an ANI of approximately 84% with each other. A total of 3,576 ORFs and 3,339 ORFs were identified on 178 contigs and 173 contigs of *Ca. Shimingles merlion* and *Ca. Shimingles singa* respectively. As *Lewinella persica* DSM 23188 was identified to be the closest sister lineage, ORFs were predicted using the same approach and only 5,313 ORFs were identified on 85 contigs. Despite being identified as the closest genome to UPWRP_1, *Lewinella persica* DSM 23188 shared an average amino acid identity (AAI) of only 42.70% and 43.26% to *Ca. Shimingles merlion* and *Ca. Shimingles singa* respectively. The highest AAI of *Lewinella persica* DSM 23188 to both the genomic bins was 88.54%, thus showing that the ORFs between *Lewinella persica* DSM 23188 and *Ca. Shimingles* did not share a high degree of similarity. Draft genomes of *Ca. Shimingles merlion* and *Ca. Shimingles singa* shared an average AAI of 86.89% and the highest AAI was 100%. This indicated that there were identical ORFs shared between both genomes.

Taxonomic analysis of ORFs

It is important to note that taxonomic classification of ORFs was established using the NCBI taxonomy and not the SILVA taxonomy as described for 16S rRNA phylogenetic analysis. Taxonomic description of the lineage of *Lewinella persica* DSM 23188 using both the NCBI and SILVA taxonomy is presented in Appendix A-10. A total of 82.97% of ORFs from *Ca. Shimingles merlion*, 84.01% of ORFs from *Ca. Shimingles singa* and 95.80% of ORFs from *Lewinella persica* DSM 23188 could be assigned a putative taxonomy and function against the NCBI NR database using the LCA settings described in Section 5.2.7. 99.23% of ORFs from *Lewinella persica* DSM 23188 could be correctly assigned to the class Saprospira (equivalent SILVA taxonomy would be Sphingobacteriia). Within the class Saprospira, 99.92% of ORFs could be assigned to the order Lewinellaceae and genus *Lewinella*.

In contrast, ORFs from *Ca. Shimingles merlion* and *Ca. Shimingles singa* were taxonomically classified not only to the phylum Bacteroidetes (76.17% and 76.83%), but to other bacterial phyla: Proteobacteria (2.02% and 2.00%) and Terrabacteria group (2.73% and 2.60%). ORFs were further differentiated into six different classes within the phylum Bacteroidetes, with majority of the ORFs assigned to Cytophagia (10.01% and 9.73%), Flavobacteriia (4.62% and 4.49%), Sphingobacteriia (3.54% and 3.85%), Chitinophagia (2.86% and 3.03%) and Saprospiria (2.73% and 2.78%) (Figure 5-11).

Figure 5-11: Taxonomic comparison of ORFs from *Candidatus Shimingles* and *Lewinella persica* DSM 23188 using Megan's LCA approach.

ORFs were searched against the NCBI's NR database and taxonomically mapped to NCBI taxonomy. A node represents a taxon and size of the node is proportional to the absolute number of assigned reads. ORFs belonging to *Ca. Shimingles* and *Lewinella persica* DSM 23188 are demarcated in different colours, and the legend is presented on the upper left hand corner of the diagram. Only nodes with more than 1% of total number of assigned reads are presented.

5.3.9 Functional analysis of UPWRP_1

Draft genomes of *Ca. Shimingles merlion* and *Ca. Shimingles singa*, together with its closest relative, *Lewinella persica* DSM 23188 were submitted to the RAST server for genome annotation.

A relative comparison of the assignment of protein-encoding genes (PEGs) to the highest SEED subsystem hierarchy is presented in Figure 5-12.

Figure 5-12: Relative comparison of PEGs from the genomes of *Candidatus Shimingles merlion*, *Candidatus Shimingles singa* and *Lewinella persica* DSM 23188 being assigned to the highest SEED subsystem hierarchy. Legend of the colour-codes representing the three genomes is located at the lower-right corner of the image.

Ca. Shimingles merlion contained 4,485 protein-encoding genes (PEGs), and 24% of the PEGs could be categorised under 136 subsystems. Majority of the PEGs were assigned to the subsystem categories: virulence, disease and defence (18.18%); cofactors, vitamins, prosthetic groups, pigments (13.39%) and protein metabolism (10.58%). A total of 4,071 PEGs were present in *Ca. Shimingles singa*, where 26% of PEGs could be assigned to 146 subsystems. Similarly, a high proportion of PEGs were assigned to the subsystem categories: virulence, disease and defence (13.23%); cofactors, vitamins, prosthetic groups, pigments (11.93%) and carbohydrates (9.64%).

Based on PEGs common to both *Ca. Shimingles merlion* and *Ca. Shimingles singa*, it can be inferred that *Ca. Shimingles* possess a central carbohydrate metabolism (subsystem: carbohydrates) that include glycolysis, gluconeogenesis, pentose phosphate pathway and pyruvate metabolism. In addition to a central carbon metabolism, the presence of butyrate metabolism genes suggests

that fatty acid β -oxidation provides another source of energy to *Ca. Shimingles*. The presence of glutamate dehydrogenases genes and glutamine synthetases allow *Ca. Shimingles* to perform ammonium assimilation, which provides a source of Nitrogen. The presence of a comprehensive set of NADH ubiquinone oxidoreductase enzymes (subunits: A-N) and V-type ATP synthase subunit A (subunits: A-E, I, K) in *Ca. Shimingles* suggest that the organism respire aerobically. Additionally, PEGs involved in DNA repair, DNA replication, folate biosynthesis, vitamin biosynthesis (e.g. Vitamin B6), fatty acids biosynthesis and protection from oxidative stress were reported. The subsystem that contained the highest number of PEGs (n=71, 108 respectively for *Ca. Shimingles merlion* and *Ca. Shimingles singa*) belonged to Listeria surface proteins: Internalin-like proteins, therefore suggesting that *Ca. Shimingles* play an important role in virulence. The assignment of genes to Thiol-activated cytolysin corroborates with the potential pathogenicity of *Ca. Shimingles*. PEGs assigned to the subsystem: KDO2-Lipid A biosynthesis or the synthesis of gram-negative cell wall components indicate that *Ca. Shimingles* is gram-negative. No PEGs affiliated to sporulation or motility (e.g. flagella-related proteins or type IV pillus) were discovered in *Ca. Shimingles*.

More PEGs (n=5,682) were present on the genome of *Lewinella persica* DSM 23188, where 27% of the PEGs could be categorised into 179 subsystems. Unlike *Ca. Shimingles*, only 4.70% of PEGs were assigned to the subsystem category of virulence, disease and defence. PEGs that were unique to *Ca. Shimingles* but not to *Lewinella persica* DSM 23188 appeared to be genes involved in housekeeping (e.g. Biotin biosynthesis and ATP synthase subunits) and are presented in Appendix A-11. CRISPRs-associated genes that serve as a prokaryotic defence machinery (Makarova *et al.*, 2012) were present exclusively in the genome of *Ca. Shimingles singa*. Interestingly, there were genes of *Ca. Shimingles* which were resistance to toxic compounds and were assigned to the subsystems: resistance to arsenic (arsenical pump-driving ATPase and arsenical-resistance protein ACR3) and copper tolerance (periplasmic divalent cation tolerance protein *CutA*).

In addition to having a central carbohydrate metabolism such as the Entner-Doudoroff pathway and TCA cycle, *Lewinella persica* DSM 23188 is able to ferment lactate under conditions of low oxygen concentration. Genes involved in fermentation were not detected in *Ca. Shimmingles*. The proteorhodopsin genes present exclusively to *Lewinella persica* DSM 23188 suggest that this organism is capable of generating energy through light-driven proton pumps (DeLong and Bèjà, 2010). The three genomes did not contain PEGs in the following subsystem categories: Arabinose sensor and transport module; central metabolism; iron acquisition and metabolism; metabolite damage and its repair or mitigation; metabolisms of aromatic compounds; motility and chemotaxis; phages, prophages, transposable elements; secondary metabolism and virulence.

UPWRP_2

5.3.10 Visualisation of UPWRP_2

Probe Ribo_Halia1029_17 labelled cells formed dense cell clusters in fixed biomass of activated sludge; individual *Haliangium* cells were seldom observed under confocal microscopy. It was difficult to elucidate the morphology of the cells from dense cell clusters due to oversaturation of FISH probe signals. However, in one particular confocal image out of a total of 140 images, different morphologies of *Haliangium* cells could be observed interacting in close proximity within an aggregate (Figure 5-13A). The observed morphology included rod, coccus and spherical morphology. When in-solution FISH protocol was applied to unfixed sludge samples, aggregates of *Haliangium* broke up and different morphologies of *Haliangium* cells could be clearly observed (Figure 5-13B). Probe-labelled cells with the spherical morphology was observed to be larger than the coccus- or rod-shaped cells; the spherical morphology had a diameter of up to 2 µm. Hitherto, morphology of UPWRP_2 could not be deciphered as other members of *Haliangium* were targeted *in silico* by probe Ribo_Halia1029_17.

Figure 5-13: Confocal micrographs depicting the different morphotypes of *Haliangium* cells hybridised with probes Ribo_Halia1029_17_{Cy5} (red) and EUB338_{A488} (green) in activated sludge samples.

(A) *Haliangium* cells in a fixed and (B) unfixed sample of activated sludge where in-solution FISH protocol was applied. *Haliangium* cells appeared yellow because of the merging of probe signals. Arrows demarcate the different morphotypes of *Haliangium*. Bar: 5 μm. Magnification: 63x.

Furthermore, the spatial arrangement of *Haliangium* cells labelled by probe Ribo_Halia1029_17 differed from the *Haliangium* cells labelled with the canonical FISH probes for *Haliangium*: HalianMix probes. HalianMix-labelled cells do not form dense cell clusters, and they predominantly exist as single cells or small clusters (Figure 5-14).

Figure 5-14: Confocal micrographs of different clades of *Haliangium* hybridised with probes Ribo_Halia1029_17_{Cy5} (red), HalianMix_{Cy3} (magenta) and EUB338_{A488} (green) in UPWRP activated sludge samples.

Haliangium cells labelled by probe Ribo_Halia1029_17_{Cy5} formed dense cell clusters (yellow), while *Haliangium* cells labelled by HalianMix existed mostly as single cells or small clusters (purple). Different clades of *Haliangium* were demarcated by arrows. Bar: 5 µm. Magnification: 63x.

Probe-labelled cells from the RiboTagger and canonical probes did not overlap, therefore implying that the probes targeted different clades within the genus *Haliangium*. This hypothesis was verified to be true through *in silico* probe matching of probes Ribo_Halia1029_17 and HalianMix in the ARB-parsimony guide tree of the genus *Haliangium* (Figure 5-15). Species matched by the probes were mutually exclusive; species matched by Ribo_Halia1029_17 formed a monophyletic clade that did not possess probes HalianMix binding site.

Figure 5-15: 16S rRNA gene sequences of the genus *Haliangium* represented in the ARB-parsimony guide tree covered by the various probe combination of probe Ribo_Halia1029_17 or HalianMix.

The two set of probes do not hybridise to the same species and are mutually exclusive. Different taxa of *Haliangium* are targeted by the two sets of probes. Legend of the various probes used is located at the bottom right-hand corner of the diagram.

5.3.11 Calibration of probe Ribo_Halia1029_17

Melting curve of probe Ribo_Halia1029_17 was performed on technical replicates of UPWRP sludge. The optimal formamide concentration was determined to be 40%, as it was the highest formamide concentration before the huge dip in mean fluorescence intensity of probe-labelled cells (Figure 5-16).

Figure 5-16: Probe dissociation curve of probe Ribo_Halia1029_17 performed on technical triplicates of UPWRP activated sludge samples.

An optimal formamide concentration of 40% was estimated for Ribo_Unk1029_17.

5.3.12 Cell sorting of *Haliangium*

Although dense clusters of *Haliangium* cells (Figure 5-17A) were broken into individual cells after the process of fixation-free in-solution FISH hybridisation and washing, aggregates of other bacterial cells remained (Figure 5-17B). Bacterial cell aggregates were further broken up through probe sonication, with the FISH-labelled cells still fluorescing brightly after the sonication process (Figure 5-17C).

Figure 5-17: Confocal micrographs depicting the process of breaking up dense clusters of *Haliangium* into single-cell suspension for FACS sorting.

Haliangium cells were hybridised with probes Ribo_Halia1029_17_{Cy5} (red) and EUB338_{A488} (green). *Haliangium* cells appeared yellow because of the merging of probe signals. *Haliangium* cells were visualised in: (A) fixed samples; (B) fixation-free in-solution suspension and (C) probe sonicated fixation-free in-solution suspension of activated sludge sample. Bar: 5 μm. Magnification: 63x.

Similar sorting gates that were used for the sorting of *Thauera* and UPWRP_1 were applied to the sorting of UPWRP_2 in the 1st biological replicate (Appendix A-12). During the sorting of the 2nd biological replicate, sorting gate of side scatter versus forward scatter was adjusted in the second round of sort to capture different morphology of bacterial cells observed in Figure 5-13.

Sorting gate in Figure 5-18A was constructed to capture smaller bacterial cells associated with low light-scattering properties, and microscopic observation showed that the sorted cells corresponded to the rod and coccus morphotypes (Figure 5-18B). Sorting gate in Figure 5-18C was designed to capture larger bacterial cells with high light-scattering properties, and microscopic observation showed that the sorted cells corresponded to the spherical morphology (Figure 5-18D).

Figure 5-18: Different morphotypes of *Haliangium* cells captured with different sorting gates of forward versus side scatter during the second round of sorting. (A) Sorting of bacterial cells with low light-scattering properties led to (B) rod and coccus shaped cells. (C) Sorting of bacterial cells with high-light scattering properties led to (D) spherical shaped cells.

5.3.13 Evaluation of the sorting efficiency

Both quantitative FISH analysis and purity of sorting that was estimated from FACS analysis showed an enrichment of more than 90-fold. However, RiboTagger 16S rRNA analysis did not reflect the same levels of enrichment as only 44.14% of V6 sequence tags could be annotated to UPWRP_2 (Figure 5-19). The sorted samples produced a total of 86 RiboTags and 31,874 V6 sequence tags.

Figure 5-19: Effectiveness of cell sorting and levels of enrichment for UPWRP_2 from activated sludge samples were estimated with: quantitative-FISH analysis, flow cytometric analysis and RiboTagger 16S rRNA analysis.

(A) Quantitative FISH analysis depicting the relative abundance of UPWRP_2 in biological duplicates of activated sludge samples in pre- and post-sorted samples. Quantitative FISH analysis was performed by acquiring confocal images, followed by image analysis. Ratio of probe Ribo_Halia1029_17-labelled cells over probe EUB338-labelled cells was calculated through image processing software. Each dot represents quantitative FISH analysis performed on one confocal image. (B) Cell sorting purity obtained after an initial round of sorting was calculated from flow cytometric analysis. (C) Relative abundance of RiboTags annotated to Ribo_Halia1029_17, UPWRP_2 and other taxa in pre- and post-sorted samples.

Lower-than-expected level of enrichment for UPWRP_2 is consistent with co-sorting of another taxon which possessed the binding site for probe Ribo_Halia1029_17. This taxon was enriched through the process with a 14-fold enrichment, and it was annotated with a RiboTag that shared a similar sequence with UPWRP_2 (Table 5-6). RiboTag of this organism could be annotated to a *Haliangium* species (accession number: AB286567) using the 'TestProbe' tool of the SILVA database. Both UPWRP_2 and *Haliangium* species (accession number: AB286567) shared a 16S rRNA gene sequence similarity of 93.11%.

Table 5-6: Major RiboTags present in FISH-FACS sorted samples with probe Ribo_Halia1029_17

RiboTags	Average relative abundance in sorted samples (%)	Annotation
TCTCACTCGCTCCGAAGGCACCCC <u>GACAT</u> CTC	44.14	UPWRP_2
TCTCACTCGCTCCGAAGGCACCCC <u>ACCGT</u> TTC	44.86	<i>Haliangium</i> species (accession no: AB286567)
GTGAACCGACCCCAAAAGAGGCACCCATCTC	0.069	Propionibacterium
Other RiboTags	10.93	No annotation

Sequence of probe Ribo_Halia1029_17 has been highlighted in red. Sequence dissimilarities between the two RiboTags of *Haliangium* have been underlined.

Haliangium species co-sorted with UPWRP_2 constituted an average of 44.86% of the total RiboTags in sorted samples. However, the relative abundance of *Haliangium* species and UPWRP_2 differed from sample to sample. To better understand the rationale between the differential abundance of UPWRP_2 and *Haliangium* species in sorted samples, the two species were hypothesised to possess different morphology which were observed in pre-sorted samples (Figure 5-13). Sorting gates were constructed to separate the different morphotypes during the second round of sorting of biological replicate 2 (Figure 5-18).

The hypothesis was tested by plotting the abundance of UPWRP_2 in the sorted samples obtained from the two different sorting gates against the number of events collected from the cell sort (Figure 5-20A).

Figure 5-20: Relative abundance of (A) UPWRP_2 and (B) *Haliangium* species (AB286567) in samples sorted with sorting gates using high- or low-light scattering properties, plotted against the different number of sort events: 10 and 1000 events.

Relative abundance of UPWRP_2 was higher in samples sorted with a low-light scattering gate than samples sorted with a high-light scattering gate, with 10 events being the optimal number of events for specific cell sorting. Relative abundance of *Haliangium* species (accession number: AB286567) was higher in samples sorted with a high-light scattering gate, but it was also detected in samples sorted with a low-light scattering gate (Figure 5-20B). It was not possible to test the p-value between low scatter and the high scatter for different sort events because of the small sample size ($n=2$ for each sort category). Sorted samples with 1000 events contained a mixture of UPWRP_2 and *Haliangium* species (accession number: AB286567).

Approximately 11% of the RiboTags could not be annotated. 56 out of the 70 non-annotated RiboTags have long representative sequences that were closely related (>98% similarity) with UPWRP_1. This was achieved through multiple sequence analyses of the long representative sequence of UPWRP_1 and the non-annotated RiboTags; the long representative sequence was extracted by RiboTagger software. However, it should be noted that pairwise analyses were

performed with the representative sequence of a read length of only 81-84 bp long and is therefore not very informative without the full-length 16S rRNA sequences.

5.3.14 16S rRNA phylogeny of UPWRP_2

The same methods used to determine the 16S rRNA phylogeny of UPWRP_1 were applied here with the aim of resolving the 16S rRNA phylogenetic relationship of UPWRP_2. 60 clones were randomly selected from a clone library generated from a sorted sample where 100% of RiboTags were annotated to UPWRP_2. Out of the 60 sequences, 1 clone sequence comprising of partial full-length sequences were rejected, and the remaining 59 clone sequences were selected for phylogenetic analysis. *De novo* OTU picking resulted in just one OTU, where the most abundant sequence was picked as the representative sequence (accession number: KX954253). The representative sequence formed a monophyletic clade with its closest neighbour: *Haliangium* clone 0102 (accession number: AB286332) with a sequence similarity of 98.37% and the phylogenetic placement was supported by a high bootstrap value of >90% (Figure 5-21). Similar to UPWRP_2, *Haliangium* clone 0102 was isolated from activated sludge.

The representative sequence is taxonomically classified to the genus *Haliangium*, of the family Haliangiaceae, of the order Myxococcales, of the class Deltaproteobacteria and of the phylum Proteobacteria. Four different *Haliangium* species, including UPWRP_2 were covered *in silico* by probe Ribo_Halia1029_17 in the SILVA database. Surprisingly, one of the *Haliangium* species (accession number: EU734997) had a sequence similarity of 86.06% to the representative sequence. Hence, it was not captured in the phylogenetic tree of UPWRP_2. None of the closest neighbours had a draft genome. *Haliangium ochraceum* DSM 14365 (accession number: CP001804) was the only *Haliangium* species with a draft genome, but the 16S rRNA gene had a sequence homology of 86.09% to UPWRP_2.

Figure 5-21: Maximum-likelihood (PhyML) phylogenetic tree depicting the 16S rRNA phylogenetic relationship of OTUs (99% cut off) obtained from clone libraries of sorted samples hybridised with probe Ribo_Halia1029_17, and its closely related sequences in the SILVA database.

Only near full-length sequences (≥ 1200 bp) were used for phylogenetic analysis. Representative sequence of OTU is demarcated in red. Coverage of probe Ribo_Halia1029_17 is reflected with the brackets. Closely related sequences were obtained from the SILVA SSU Ref NR 99 database (version 123). Members of the genus *Chlorobium* were used as the out-group. Bootstrap values were calculated from 1000 bootstrap analysis and only bootstrap values over 50% were displayed. Branches with low bootstrap values $\leq 50\%$ have been multifurcated. Scale bar represents substitutions per nucleotide base. Legend of the various bootstrap values is located at the upper left-hand corner of the diagram.

5.3.15 Design of specific FISH probe for UPWRP_2

FISH probe Halia183 was designed through the process of comparative sequence analysis to

specifically target UPWRP_2. The motivation driving the design of probe Halia183 was to

determine the spatial interaction of UPWRP_2 with the other *Haliangium* species that were

targeted by probe Ribo_Halia1029_17. It was of interest to observe if the different *Haliangium*

species exist together in tight cell clusters as observed in Figure 5-13. Through co-hybridisation of probes Halia183_{Cy3}, Ribo_Halia1029_17_{Cy5} and EUB338_{A488}, UPWRP_2 was observed never to co-aggregate with the other *Haliangium* species in at least 50 confocal images (Figure 5-22A). UPWRP_2 had the propensity to form individual cell clusters that were devoid of other *Haliangium* cells. However, cell clusters of UPWRP_2 were observed to form interactions with other bacterial cell clumps (Figure 5-22B). Hitherto, it was not possible to decipher the identity of the interacting cell cluster.

Figure 5-22: Confocal micrographs depicting the spatial interaction of UPWRP_2 in fixed samples of activated sludge.

UPWRP_2 cells were co-hybridised with probes Halia183_{Cy3} (magenta), Ribo_Halia1029_17_{Cy5} (red) and EUB338_{A488} (green) and they appeared pinkish-white because of the merging of all probe signals. Other *Haliangium* species appeared yellow. UPWRP_2 cells were observed: (A) not to interact with the other *Haliangium* species labelled by probe Ribo_Halia1029_17_{Cy5}, (B) but with other bacterial cells in the floccular sludge community. Bar: 5 µm. Magnification: 63x.

Probe Halia183 was also used to verify the findings from FACS sorted samples, where sorting using the high-light scattering properties produced mostly spherical-shaped objects (Figure 5-18D) that were correlated to the presence of *Haliangium* species (accession number: AB286567), but not UPWRP_2 (Figure 5-18B). In-solution FISH experiment with co-hybridisation of probes Halia183_{Cy3}, Ribo_Halia1029_17_{Cy5} and EUB338_{A488} showed that the spherical-shaped objects were observed to be hybridised with probe Ribo_Halia1029_17_{Cy5}, but not with probe Halia183_{Cy3} in at least 20 field-of-views (Figure 5-23). This observation suggests that the spherical-shaped objects were not produced by UPWRP_2, but by other *Haliangium* species covered by probe Ribo_Halia1029_17.

Figure 5-23: Confocal micrographs showing that the spherical-shaped objects were produced by other *Haliangium* species, but not by UPWRP_2 in fixed samples of activated sludge. Activated sludge sample was co-hybridised with probes Halia183_{Cy3} (magenta), Ribo_Halia1029_17_{Cy5} (red) and EUB338_{A488} (green). Spherical-shaped objects appeared yellow because of the merging of probe signals from probes Ribo_Halia1029_17_{Cy5} and EUB338_{A488}. Other rod-shaped cells not hybridised with probe Halia183 were also reflected in yellow. Rod- and coccus-shaped cells hybridised with all three probes appeared pink/purple because of the merging of probe signals. No spherical-shaped objects were hybridised with probe Halia183_{Cy3}. Bar: 5 μm. Magnification: 63x.

5.3.16 Metagenomic analysis of UPWRP_2

Metagenomic assembly

The genomic bins obtained through genomic binning of the co-assembly of contigs from multiple samples sorted with probe Ribo_Halia1029_17 were associated with high level of contamination. To overcome this problem, a single sorted sample whose RiboTagger analysis revealed that 100% of RiboTags were annotated to UPWRP_2 was used to generate the draft genome of UPWRP_1. *De novo* assembly of the single sorted sample containing 14,000,892 trimmed reads produced 474 contigs with an N50 of 12,467 bp. The longest contig length was 55,295 bp and the genome size was 2.42 Mbp. The total number of essential single-copy genes was equivalent to the number of unique essential single-copy genes (n=41), thus indicating the presence of only one genome. Additionally, CheckM validated the presence of one genome that had a genome completeness of 21.70% and a contamination of 0.02%.

Quality-control of contigs

Although the assembly displayed an extremely low contamination (0.02%), decontamination and quality-control were performed on the contigs using the ACDC software because MDA has the propensity to produce non-specific artefact sequences (Zhang *et al.*, 2006; Binga *et al.*, 2008). The same unsupervised detection analysis that was used in the decontamination of contigs of *Ca. Shimingles* was applied here and CC-statistical clustering produced five clusters of contigs (Figure 5-24). The largest cluster (red contig) was assumed to contain contigs from UPWRP_2. The contigs were exported and evaluated with CheckM for contamination and completeness (Table 5-7).

Figure 5-24: Decontamination of the contigs of UPWRP_2 produced five clusters of contigs. Decontamination was performed with the ACDC software which employs tetramer frequencies and CC statistics, followed by visualisation with BH-SNE. Clusters of contigs were represented by different colours.

Table 5-7: Statistics for the draft genome of UPWRP_2

Draft genome of UPWRP_2	
Number of contigs	462
Genome size (Mbp)	2.37
GC content (%)	63.72
N50 of contigs (bp)	12,548
Longest contig (bp)	55,295
Number of essential single-copy genes	41 / 111
Number of unique single-copy genes	41
Genome completeness	21.70
Genome contamination	0.02
Number of unique phylogenetic markers	29/43
Number of ORFs	2,081
Number of tRNA	7
Number of rRNA	3
Phylogenetic placement in reference genome tree	k__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__Myxococcales; f__Kofleriaceae; g__Haliangium

Although 12 contigs were removed through decontamination, completeness and contamination of the genome remained the same as shown through the fraction of essential single-copy genes and CheckM. The other 4 clusters did not contain any essential single-copy genes or conserved phylogenetic marker genes and were therefore assumed to be sequencing artefacts.

5.3.17 Phylogenomics analysis of UPWRP_2

Using a set of 43 phylogenetic conserved marker genes with CheckM, draft genome of UPWRP_2 was phylogenetically classified to the genus *Haliangium*, of the family Kofleriaceae, of the order of Myxococcales, of the class Deltaproteobacteria and of the phylum Proteobacteria (Figure 5-25). The closest genome was identified to be *Haliangium ochraceum* DSM 14365 (IMG 646311933). Nomenclature of the draft genome of UPWRP_2 will henceforth be assigned as *Haliangium clustero*.

Figure 5-25: Phylogenetic placement of draft genome of UPWRP_2 in the reference genome tree using a concatenation of 43 phylogenetic marker genes.

A set of 2,502 complete genomes and 3,604 draft genomes from IMG database were used in the reference genome tree. Genomic bin is demarcated in red. Closest relative to the genomic bins was *Haliangium ochraceum* DSM 14365 (IMG 646311933). Closest outgroup corresponded to different genera under the family Geobacteraceae, of the order Desulfuromonadales and of the class Deltaproteobacteria.

Genome comparison with average nucleotide identity (ANI) and amino acid identity (AAI)

Haliangium clustero shared an average ANI of approximately 65% with *Haliangium ochraceum* DSM 14365. A total of 2,081 ORFs were identified on 462 contigs from the draft genome of *Haliangium clustero* and 6,845 ORFs were identified on one contiguous chromosome of the closest sister lineage: *Haliangium ochraceum* DSM 14365. *Haliangium clustero* and *Haliangium ochraceum* DSM 14365 shared an average ANI of 38.90%. The highest AAI was 90.20% and only 12 ORFs have an AAI value of more than 75%. This indicated that no identical ORFs were shared between the two genomes. The small degree of similarity observed between the two genomes

must be interpreted with caution because only 21.70% of the draft genome of *Haliangium clustero* was recovered and an incomplete genome influences the average AAI.

Taxonomic analysis of ORFs

The NCBI and SILVA taxonomy of *Haliangium* differs and is presented in Appendix A-13. A total of 55.98% of ORFs from *Haliangium clustero* and 99.74% of ORFs from *Haliangium ochraceum* DSM 14365 could be assigned a putative taxonomy and function against the NCBI NR database. 99.97% of ORFs from *Haliangium ochraceum* DSM 14365 could be correctly assigned to the genus *Haliangium*. However, only 3.61% of ORFs from *Haliangium clustero* could be classified as *Haliangium* (Figure 5-26). Majority of the ORFs were assigned to different phyla: Alphaproteobacteria (2.66%), Betaproteobacteria (3.26%) and Gammaproteobacteria (2.40%). Within the phylogenetic lineage of *Haliangium*, ORFs were assigned to different orders: Cystobacterineae (7.47%) and Sorangineae (11.59%), and different family: Nannocystaceae (1.63%).

Figure 5-26: Taxonomic comparison of ORFs from *Haliangium clustero* and *Haliangium ochraceum* DSM 14365 using Megan's LCA approach.

ORFs were searched against the NCBI's NR database and taxonomically mapped to NCBI taxonomy. A node represents a taxon, and size of the node is proportional to the absolute number of assigned reads. ORFs belonging to *Haliangium clustero* and *Haliangium ochraceum* DSM 14365 are demarcated in different

colours, and the legend of the colour code is presented on the upper left hand corner of the diagram. Only nodes with more than 1% of total number of assigned reads are presented.

5.3.18 Functional analysis of UPWRP_2

Even with a genome completeness of 21.70%, the same functional analysis and comparison that was performed in Section 5.3.8 was applied to *Haliangium clustero* to gain an insight into its functional potential. A relative comparison of the assignment of genes to the highest SEED subsystem hierarchy from *Haliangium clustero* and *Haliangium ochraceum* DSM 14365 is presented in Figure 5-27.

Figure 5-27: Relative comparison of genes from the genomes of *Haliangium clustero* and *Haliangium ochraceum* DSM 14365 being assigned to the highest SEED subsystem hierarchy.

Legend of the colour-codes representing the two genomes is located at the lower-right corner of the image.

Haliangium clustero produced 1,858 PEGs, and only 9% of PEGs could be categorised into 31 subsystems. In contrast, *Haliangium ochraceum* DSM 14365 had 7,029 PEGs, where 32% of PEGs could be assigned to 198 subsystems. Majority of PEGs of *Haliangium clustero* were assigned to the categories: protein metabolism (15.56%); virulence, disease and defence (13.33%); carbohydrates (11.11%); cofactors, vitamins, prosthetic groups, pigments (11.11%); membrane transport (11.11%); RNA metabolism (11.11%) and respiration (11.11%). On the other hand, majority of PEGs from *Haliangium ochraceum* DSM 14365 were assigned to different categories: amino acids and derivatives (12.56%); protein metabolism (11.68%); fatty acids, lipids and isoprenoids (11.31%). Only PEGs that played a role in pyruvate metabolism was present in the

central carbon metabolism. The presence of c-type cytochromes suggests that *Haliangium clusterio* respire aerobically. Copper-translocating P-type ATPase had the highest number of PEG assignment, thus suggesting that *Haliangium clusterio* is capable of copper homeostasis (González *et al.*, 2008).

PEGs that were unique to *Haliangium clusterio* are presented in Appendix A-14. Among the common housekeeping genes such as synthesis of tRNA subunits (subsystem: protein metabolism) and folate biosynthesis (subsystem: cofactors, vitamins, prosthetic groups, pigments), interesting genes unique to *Haliangium clusterio* were associated with cellular invasion (*Listeria* surface proteins: Internalin-like proteins) and resistance to toxic compounds (chromate transport protein ChrA). The presence of sigma-B stress-response cluster in *Haliangium clusterio* demonstrates the ability of the organism to regulate its metabolism in response to environmental stresses upon entry into early stationary phase (Binnie *et al.*, 1986). Interesting genes that were absent from *Haliangium clusterio* but present in the complete genome of *Haliangium ochraceum* DSM 14365 included the Type IV pilus for twitching motility and stress-response genes involved in heat-shock, osmotic stress and oxidative stress.

5.3.19 Evaluation of genome completeness, contamination and assembly quality of draft genomes with other studies

Genome completeness, genome contamination and assembly quality of the draft genomes obtained from this thesis were compared with studies that have performed mini-metagenomics (Podar *et al.*, 2007; McLean *et al.*, 2013) and single-cell genomics (SAGs; Rinke *et al.*, 2013). In the studies mentioned above, genome completeness was measured using the concept of essential single-copy genes. However, genome contamination was not measured. A box-and-whisker graph was constructed to observe the distribution of genome completeness from 201 single-amplified genomes (Rinke *et al.*, 2013). Five genomes representing the minimum, 25th percentile, median, 75th percentile and maximum completeness were cherry-picked for comparison of genome completeness and contamination (Appendix A-15). For a standardise comparison of genome

completeness and contamination, draft genomes constructed from the studies mentioned above were re-evaluated with the CheckM software and the results are presented in Figure 5-28A.

Figure 5-28: Comparison of the genome completeness and contamination of draft genomes. CheckM was used for the evaluation of genome completeness and contamination. Legend of the various genomes and its corresponding reference is located on the right side of the image.

Genome completeness of the SAGs varied considerably from 6.90% (minimum) to 92.31% (maximum), with a median value of 27.73% and 75th percentile of the SAGs only having a value of 50.56%. The highest genome completion (92.31%) of the SAGs was similar to the genome completeness of *Ca. Shimmingles merlion* (90.28%) and *Ca. Shimmingles singa* (91.46%). Genome completeness of *Haliangium clustero* (21.70%) was similar to TM7_GTL1 genome (22.08%), which was coincidentally also assembled from a FISH-FACS sorted sample of 5 events. However, genome completeness of *Haliangium clustero* was lower than 25th percentile (26.72%) of the SAGs.

An evaluation of the contamination levels of the SAGs showed that one of the SAG genome: AAA011-L6 had a contamination of 2.25% even though proper quality-control and decontamination procedures were performed. Similarly, TM7_GTL1 genome had a genome contamination value of 2.56% even after computational binning to remove an isolate that was co-sorted with the target cells. Genome contamination of the other SAGs ranged between 0-0.38%

and genome contamination of *Haliangium clustero* (0.02%) falls within the range. Genome contamination of *Ca. Shimingles merlion* (3.94%) and *Ca. Shimingles singa* (2.96%) were higher than the other draft genomes in comparison.

Assembly quality was quantified by measuring the number of contigs generated per genome. A box-and-whisker graph was constructed to observe the distribution of the number of contigs generated from each of the 201 single-amplified genomes (Rinke *et al.*, 2013), and 5 genomes representing the minimum, 25th percentile, median, 75th percentile and maximum number of contigs generated were cherry-picked for comparison (Appendix A-16). These genomes were compared with the genomes generated from this thesis and the studies of Podar *et al.*, (2007) and McLean *et al.*, (2013) (Figure 5-29).

Figure 5-29: Comparison of the assembly quality of draft genomes. Assembly quality was quantified by measuring the number of contigs generated per genome. Legend of the various genomes and its corresponding reference is located on the right side of the image.

The number of contigs that *Ca. Shimingles merlion* (n=178) and *Ca. Shimingles singa* (n=173) generated were slightly lower than the maximum number of contigs generated by SAGs (n=190), but higher than 75th percentile of the SAGs (n=84). *Haliangium clustero* generated the largest number of contigs (n=462) among the other genomes in comparison.

5.4 Discussion

5.4.1 Effect of multiple displacement amplification on genomic binning

The presence of duplicated essential single-copy genes in the co-assembly necessitates the use of genomic binning to separate two genomic bins of UPWRP_1. The principle of binning is based on the categorization of metagenomic contigs that share similar characteristics into discrete unit or genomic bins (Kunin *et al.*, 2008). Due to the novelty of the genomes as predicted from the average ANI and ANI with its closest neighbour, an unsupervised binning approach using tetranucleotide frequency and contig co-abundance across different samples was adopted. Tetranucleotide frequency-only binning was not effective in the deconvolution of mixed genomes as evidenced from high genome contamination and duplicated presence of essential single-copy genes. This result corroborates with the finding of Wrighton *et al.*, (2012) which showed that the use of tetranucleotide frequency as discriminative sequence signatures to separate taxonomically-related taxa resulted in genomic bins having high level of contamination as indicated by the presence of multiple copies of the essential single-copy genes. Genomic binning with differential coverage binning and tetranucleotide frequency resulted in two draft genomes with high completeness (>90%) and low contamination (<4%).

The concept of differential coverage binning relies on the differential abundance of the target taxon across multiple samples, where changes in abundance of the target taxon influence the coverage of contigs that is reflected in a coverage plot (Nielsen *et al.*, 2014). In most studies which have incorporated differential coverage binning, differential abundance of the target taxon across multiple samples was achieved either through time-series experiments (Sharon *et al.*, 2013), different DNA extraction protocols (Albertsen *et al.*, 2013a) or samples taken from different sampling locations (Kirkegaard *et al.*, 2016). Here, a different approach was used to obtain differential coverage of contigs between samples, and this was achieved through multiple displacement amplification (MDA). MDA has been shown to produce amplification bias which influences the sequencing coverage of contigs (Rodrigue *et al.*, 2009; Lasken, 2012). Therefore,

contigs representing the same homologous locus of a genome is expected to have different abundance/coverage in replicates because the amplification effect is stochastic and coverage profile of an amplified genome is independent even for clonal population (Podar *et al.*, 2009; Yilmaz *et al.*, 2010a).

It has previously been reported that contig coverage varied extensively for an MDA-amplified sample that contained 5 events (Podar *et al.*, 2007), and this would affect the accuracy of differential coverage binning as contigs from the same taxon would not cluster together. Sequencing bias could be reduced through the co-assembly of multiple samples that contained higher number of sorted events. Co-assembly will overlap regions of the genome with low sequencing coverage with regions with high sequencing coverage, therefore enhancing the overall coverage of the draft genome (Blainey, 2013). As shown in this thesis, sorted samples with higher number of collected events (n=1000) produced two distinct clusters of contigs in differential coverage plot (Figure 5-9). Unique essential single-copy genes split between the two clusters of contigs was indicative of two distinct genomic bins (Figure 5-9). In retrospect, samples with smaller number of events (n=5 or 10) resulted in some contigs having extremely low coverage and it resulted in a poorer clustering of contigs (Appendix A-17). Here, it is demonstrated that samples with higher number of events produced a better resolution for clustering of contigs in the differential coverage plot (Figure 5-9). However, it must be acknowledged that the work performed in this thesis used computational binning tools that were developed for metagenomics, and future work should emphasize on improving the binning resolution of MDA-amplified samples.

Besides amplification bias, MDA is also associated with representation bias that results in an uneven distribution of gene loci in the genome. Representation bias of the genome has been attributed to random hexamer primers annealing to random positions in the genome, where limited number of initial annealing events of the primers to the genome lead to certain regions of the genome being amplified for an average of 10-20 kbp before dissociation of the MDA phi29

polymerase (Blanco *et al.*, 1989). Uneven genome coverage resulting from MDA often leads to fragmented genes which might not be assembled into larger scaffolds during *de novo* assembly. The presence of fragmented genes was evident in the differential coverage plot of UPWRP_1, where contigs of various sizes were present in the genomic bins. Unfortunately, fragmented genes often lead to incomplete genomes where the average genome completeness achieved from single-cell genomic is estimated to be 40% (Rinke *et al.*, 2013). Representation bias of the genome has been correlated with the total number of genomic templates in the sorted sample (Lasken, 2012). Representation bias is more pronounced in samples with a low copy number of genomic templates – especially in single-cell genomics, whereas the bias is reduced in samples with a higher copy number of genomic templates (Raghunathan *et al.*, 2005; Neufeld *et al.*, 2008). Representation bias of genome - quantified by measuring the number of essential single-copy genes - varied with the number of sorted events in the samples (Appendix A-18). Stochastic representation bias can be evened out with multi-cell MDA reaction because of the presence of multiple copies of genomic loci.

Here, two approaches were implemented to reduce representation bias. The first approach involved the use of sorted samples that contained a high number of events (n=1000). Multiple closely-related cells provide a wide plethora of numerous potential annealing sites for random hexamer primers to initiate strand synthesis (Blainey and Quake, 2011). Multi-event sorted samples also increase the probability that sequencing reads are present in regions of other cells where there is a breakage in the genomic template of one cell. The second approach involved co-assembly of multiple enriched samples, where the representation bias can be averaged out as shown for both genomic bins of UPWRP_1 having a genome completeness of more than 90%. However, it is important to note that increasing the number of samples for co-assembly increases the risk of genetic heterogeneity, especially if the sorted cells are not closely-related and the polymorphism is not captured at the rRNA level (Whitaker and Banfield, 2006).

5.4.2 FACS sorting UPWRP_1

Enrichment for UPWRP_1 with probe Ribo_Unk1029_17 was successful as RiboTagger 16S rRNA analysis showed high level of enrichment for the target population. The high level of enrichment could be attributed to two reasons: (1) proper dispersal of cellular aggregates containing the target taxon and (2) a specific FISH probe with a high affinity for UPWRP_1. Most literatures recommend the use of sonication for the homogenisation of sludge samples (Yilmaz *et al.*, 2010b) prior to FISH-FACS. However, members of UPWRP_1 with a filamentous morphology have shown to be highly susceptible to cellular fragmentation upon sonication (Figure 5-4). The problem was circumvented by adopting an alternative approach of breaking up the cellular aggregates and it involved the repeated passaging of sludge samples through a fine needle syringe. Although aggregates of other bacterial cells were still visible under confocal microscopy, two-round of sorting as recommended by Haroon *et al.* (2013) removed most of the non-target cellular aggregates.

Cell purity from the initial round of sorting as estimated by flow cytometric analysis was lower than expected. This was due to an increased background noise of the scatterplot of forward versus side scatter of samples homogenised with the syringe method. Passaging samples through the syringe produced more background events than samples that were homogenised by sonication. The syringe homogenisation method produced a higher concentration of smaller-sized particles that led to higher background noise, which was subsequently detected by the FACS sorter. Consequently, threshold of the forward scatter had to be increased (Appendix A-7C) to further exclude background events. Although threshold for the forward scatter has been set, background events that were excluded appears “invisible” to the FACS sorter and subsequently will be sorted into the bulk sort. This explains the lower cell purity for the initial bulk sorting (Figure 5-5B).

An additional round of sorting improved the cell purity of the final sorted samples, which was measured with RiboTagger 16S rRNA analysis. The target population was sorted into a bulk solution of PBS after the first round of sorting, where dilution of the target population resulted in

decreased proximity between the target and the non-target population and hence a higher cell purity (Wallner *et al.*, 1997). Although the same forward scatter threshold was used for the second round of sorting, gating on the high fluorescence intensity of the probe-labelled cells proved sufficient in obtaining a high purity of the target population.

UPWRP_2

FISH-FACS sorting with probe Ribo_Halia1029_17 did not match the level of enrichment that was achieved with probes Ribo_Thau1029_17 and Ribo_Unk1029_17. Lower level of enrichment was due to co-sorting of another *Haliangium* species (accession no: AB286567) that was present at an equally high abundance (44.86%) as UPWRP_2 (44.14%). The presence of probe Ribo_Halia1029_17 binding site in the 16S rRNA sequence of both taxa was the rationale for co-sorting of the *Haliangium* species (accession no: AB286567) with UPWRP_2. At the point of deep sequencing performed at UPWRP in 2011, UPWRP_2 was present at an abundance of 2.35% and it was ranked 6th in the metatranscriptomics dataset. On the other hand, the abundance of *Haliangium* species (accession no: AB286567) was considered to be negligible as the V6 sequence tag was absent in the community profiling.

Higher abundance of *Haliangium* species over UPWRP_2 in FISH-FACS sorted samples points to a shift in the population dynamics of UPWRP's floccular sludge community, where the abundance of *Haliangium* species (accession no: AB286567) is as dominant as UPWRP_2. The inability to track changes in population dynamics was a result of the lack of consistent sampling frequency and sequencing efforts at UPWRP; design of probe Ribo_Halia1029_17 was based on saturated metagenomics results obtained 5 years ago. If there was *a priori* knowledge of the dominance of the *Haliangium* species (accession no: AB286567), a longer FISH probe could have been designed to suppress binding of probe Ribo_Halia1029_17 to the other *Haliangium* species. Using the 'TestProbe' tool of SILVA, extending the length of probe Ribo_Halia1029_17 to a length of 26 bp altered its *in silico* specificity and non-target taxa are no longer covered by the probe (data not presented).

Gating on the lower light-scattering properties of cells with the FACS machine increased the probability of sorting for UPWRP_2. Although 3 different morphotypes of *Haliangium* were observed under microscopy, cells sorted with lower light-scattering properties were observed to possess either the cocci- or rod-shaped morphology. Both UPWRP_2 and *Haliangium* species (accession no: AB286567) could be sorted with the low light-scattering sorting gate (Figure 5-20). Morphology of the rod-shaped cells highly resembled vegetative cells of *Haliangium* (Ivanova *et al.*, 2010), and vegetative cells were visually observed for both UPWRP_2 and *Haliangium* species (accession no: AB286567) (Figure 5-23).

Gating on the high light-scattering properties of cells led to the sorting of larger objects (2 μm) with spherical morphology, where majority of the sorted cells (78%) were taxonomically classified as *Haliangium* species (accession no: AB286567). In contrast, approximately 4% of the population from the same set of sorted samples was taxonomically classified as UPWRP_2. Dimension, morphology and microscopic images of the spherical object resembled myxospores of *Haliangium* described in the literatures (Hartzell *et al.*, 2001). Myxospores are usually formed from vegetative cells in response to limiting nutrient conditions or harsh environmental conditions (Shimkets and Seale, 1975). Although myxospores are known to be recalcitrant to heat and radiation (Hartzell *et al.*, 2001), *Haliangium* myxospores could be easily penetrated by FISH probe using a fixation-free FISH protocol (Figure 5-23). Additionally, *Haliangium* myxospores could be lysed with the alkaline lysis protocol and the genetic material was amenable to downstream genomic sequencing and RiboTagger analysis.

Different groups of myxobacteria varied in their adaptation fitness and social behaviour as shown in the study of Zhang *et al.* (2005), where some myxobacterial groups were equipped with the propensity to form myxospores and fruiting bodies while other groups lacked such capability. It appears that the formation of myxospores is a behaviour highly associated with *Haliangium* species (accession no: AB286567), but not with UPWRP_2 even though both species were sampled from the same environment at the point of sampling. This was eventually verified with a FISH

probe designed specifically for UPWRP_2: Halia183 which showed that the spherical objects were produced by other *Haliangium* species, but not by UPWRP_2. These comments are based on the assumptions that myxospores formed by all *Haliangium* species: (1) have similar spherical morphology and dimensions; (2) could be sorted using the high-scatter gate defined in this chapter and (3) were equally susceptible to alkaline lysis and DNA extraction.

5.4.3 Genome recovery

A mini-metagenomics approach was adopted for genome recovery of the unclassified bacterial taxa because of the higher throughput of sorting and higher success rate of genome amplification than single-cell genomics. The workflow for both mini-metagenomics and single-cell genomics are similar: (1) isolation of single cells; (2) cell lysis and whole genome amplification; (3) screening for 16S rRNA phylogenetic marker genes; (4) whole genome sequencing; (5) genome assembly; (6) quality-control processing and decontamination of contigs and (7) genome annotation. Although FACS is commonly used for the separation of individual cells from environmental samples in mini-metagenomics (Podar *et al.*, 2007; McLean *et al.*, 2013) and single-cell genomics (Woyke *et al.*, 2009; Rinke *et al.*, 2013, 2014), staining of the individual cells for subsequent isolation differed in the studies. In this thesis, a targeted approach of isolating single cells of the unclassified taxa through FISH-FACS was adopted. While FISH-FACS was also integrated by Podar *et al.*, (2007) for the isolation of TM7 cells from the soil, the 16S rRNA-targeted FISH probe used was a canonical FISH probe that was specific against members of the TM7 phylum. In contrast, this thesis incorporates newly-designed FISH probes against members of unclassified taxa whose 16S rRNA sequences were not present in the SILVA database.

Another strategy for the isolation of single cells involves staining members of the microbial community with a general nucleic acid stain such as SYBR Green and sorting the stained sample based on enhanced fluorescence of the stained cells. As the general nucleic acid stain can be applied to all cell types, this strategy has been applied successfully for the isolation of single-cells belonging to nine diverse ecosystems and 29 uncharted branches of the tree-of-life (Rinke *et al.*, 2013, 2014). A general staining strategy overcomes the limitations of applying the standard FISH

technique to environmental samples, such as high background fluorescence that mask the fluorescence of target cells and low fluorescence of the target cells due to slow-growing bacteria with low ribosomal content (Morita, 1998). However, cell sorting of cells labelled with a general stain is not a targeted approach and it requires cherry-picking of cells that have been positively-amplified with whole genome amplification. Therefore, genome recovery of rare members of the community using the cherry-picking method has a low throughput. For example, given that UPWRP_1 is present at a low relative abundance (0.50-0.70%) as quantified by quantitative FISH, there is a chance that 7 out of 1000 randomly selected cells would be UPWRP_1. A large pool of resources and manpower is required for single-cell genomics, although automation with a liquid-handling system is now a possibility. The lack of resources and access to an automated system means that the general staining is not a feasible strategy in this thesis. Although single-cell sorting could be applied with FISH-FACS, it was not considered due to the low consistency of successful genome amplification as demonstrated in the sorting of *Thauera* from the axenic culture (Section 4.3.3) and the studies of Raghunathan *et al.*, (2005) and Rinke *et al.*, (2013).

The aim of single-cell genomics is to construct complete genomes with minimal heterogeneity, whereas the aim of mini-metagenomics is to recover a “pan-genome” (Tettelin *et al.*, 2008). Due to the truncation of RiboProbe and evolutionary conservation of the 16S rRNA molecule (Woese, 1987), multiple genotypes are often present in FISH-FACS sorted samples. Mini-metagenomics necessitate the use of genomic binning computational tools to de-convolute the mixture of genomes after genomic assembly, as shown with the use of MetaBat to separate two genomic bins of UPWRP_1. Genomic binning is not required for single-cell genomics as only a single cell is analysed. While genomic binning was successfully applied to UPWRP_1, the same level of success could not be achieved with binning a co-assembly of closely-related *Haliangium* species. A plausible explanation for the successful binning of UPWRP_1 was that the two draft genomes were genetically distinct – both draft genomes shared an average ANI and AAI of 84% and 86.89% respectively. Whereas for genomic binning of *Haliangium*, closely-related genotypes present in the sorted sample complicate genomic assembly and the downstream genomic binning process.

Single-cell genomics avoid the potential issue of genomic assembly of multiple genotypes, and it only requires computational removal of contaminated contigs after the genomic assembly process. Even though single-cell genomics is based on sequencing of a single-cell, one SAG (e.g. AAA011-L6) from the study of Rinke *et al.*, (2013) had a genome contamination of 2.25% even after the decontamination process. Despite intense efforts in the removal of possible contaminants from the draft genomes through various computational means (Tennessen *et al.*, 2015; Lux *et al.*, 2016), low level of contamination should be expected with MDA-amplification experiments.

Draft genomes from this thesis were more fragmented than other studies, as quantified by the number of contigs generated per draft genome, even though the SPAdes assembler software was used. SPAdes assembly algorithm is specifically catered for assembly for MDA data because of the ability of the assembler to handle wide variations in read coverage and MDA-catalysed chimeric reads (Bankevich *et al.*, 2012). Fragmentation of assembly is often attributed to the effects of genome repeats that complicate the genome assembly process (Pop and Salzberg, 2008) and quality of Phi29 polymerase (Rinke *et al.*, 2014). In his study, a low mean read sequencing depth of sequencing was probably responsible for the higher fragmentation of draft genomes of the novel taxa (Chakraborty *et al.*, 2015).

UPWRP_1

The presence of two clades were supported by the two draft genomes having a similar GC-content of approximately 46%, an indication of the two bins belonging to the same genus (Lightfield *et al.*, 2011). Bin B had a slightly higher genome completeness and lower genome contamination than bin A. Both genomic bins A and B could be classified as a draft genome with near completeness and low contamination according to the genome quality classification scheme as proposed by Parks *et al.* (2015). Genome completeness of both draft genomes were similar to the highest genome completion attained by SAG (Rinke *et al.*, 2013). Co-assembly of multiple samples containing identical cells have been shown to improve genome completeness (Dodsworth *et al.*,

2013). While only four samples were used for co-assembly due to a limitation in computer memory, genome completeness of UPWRP_1 could have improved if all sixteen sorted samples were used for genome assembly. Draft genomes of UPWRP_1 have a genome contamination of less than 4% and this is probably due to contigs having an uneven coverage as a consequence of MDA-amplification, therefore resulting in genomic bins that are harder to separate with a good resolution.

UPWRP_2

A partial draft genome of UPWRP_2 with a completeness of 21.70% was recovered, and this finding is in agreement with other studies that have showed that genomic binning does not always lead to the successful separation of microbial genomes (Wang *et al.*, 2012; Albertsen *et al.*, 2013a). In this study, UPWRP_2 could not be binned due to the co-sorting of UPWRP_2, *Haliangium* species (accession no: AB286567) and other closely-related taxa that could not be annotated in the SILVA database. Approximately 11% of RiboTags in the sorted samples could not be annotated and they represented either chimeric artefacts or novel species co-sorted together with the target cells. The percentage of non-annotated RiboTags in the sorted samples of *Haliangium* was higher than the sorted samples from *Thauera* (3.42%) or UPWRP_1 (0.50%).

Co-sorting of a high abundance of non-annotated taxa had led to 'strain heterogeneity' and exacerbated the problem of *de novo* assembly (Luo *et al.*, 2012). Consequently, *de novo* assembly results in the assembly of chimeric sequences, and chimeric genomes are difficult to de-convolute using binning software (Treangen and Salzberg, 2012). From the problem of strain heterogeneity, it can be inferred that majority of the non-annotated RiboTags were representatives of closely related *Haliangium* species which were co-sorted together with UPWRP_2. This was eventually validated through multiple sequence alignment of the long representative sequences of UPWRP_2 and selected non-annotated RiboTags, where 56 out of 70 non-annotated RiboTags yielded high sequence similarity (>98%) to UPWRP_2. This finding suggests that many of the *Haliangium* species present in the activated sludge ecosystem are not represented in the SILVA database, and

this has severe implications for probe and primer design. The activated sludge biosphere presents an untapped source for exploring the diversity of the genus *Haliangium*. Strain heterogeneity is currently a limitation in the field of metagenomics and genomic binning. Despite this limitation, a draft genome of UPWRP_2 was produced from a single sorted sample of 5 events, where RiboTagger analysis showed that 100% of RiboTags were annotated to UPWRP_2. Genome binning was not applied to the assembly of UPWRP_2 because the total number of essential single-copy genes was identical to the number of unique essential single-copy genes, thus indicating the presence of only one genome. Representation bias catalysed by MDA was observed as only 21.70% of genome completeness could be achieved, and this was in line with genome completeness of other draft genomes that were amplified with MDA from other studies (Podar *et al.*, 2007; Rinke *et al.*, 2013). A more complete genome of UPWRP_2 could be obtained with the use of specific probe Halia183 in FISH-FACS in future studies.

5.4.4 Phylogeny and phylogenomics UPWRP_1

Phylogenetic assignment using a concatenation of 43 conserved phylogenetic marker genes coincides with the 16S rRNA phylogenetic analysis of the two draft genomes on the taxonomic ranks of phylum, order and class. Both phylogenetic analyses agreed that UPWRP_1 should be assigned to the order Sphingobacteriales, of the class Sphingobacteriia, and of the phylum Bacteroidetes. The taxonomic ranks of family and genus were unresolved using 16S rRNA phylogeny. Concatenation of multiple marker genes yields a greater phylogenetic resolution than a single gene analysis (Szollosi *et al.*, 2012), and UPWRP_1 was proposed to be taxonomically classified under the Saprospiraceae family using the 43 phylogenetic marker genes. Although *Lewinella persica* DSM 23188 has been proposed to be the closest genome on the species level, the 16S rRNA sequence similarity between *Lewinella persica* DSM 23188 and *Ca. Shimingles merlion* was only 81.95%. Therefore, UPWRP_1 is unlikely to be classified under the genus *Lewinella*, and this conclusion is supported by UPWRP_1 and *Lewinella persica* DSM 23188 having an ANI of approximately 63%.

The extraction of two genomic bins corroborates with the 16S rRNA phylogenetic analysis of UPWRP_1, where two different clades were supported by a high bootstrap value of more than 90%. UPWRP_1 shared a 16S rRNA sequence similarity of approximately 90% with its closest neighbours. In accordance to the taxonomic classification boundary proposed by Yarza *et al.* (2014), members of UPWRP_1 have been proposed to be novel bacteria species under a novel genus; nomenclature of the novel genus and species will be as followed:

- '*Candidatus Shimmingles*'; novel genus
- '*Candidatus Shimmingles merlion*'; novel species that represents genomic bin A
- '*Candidatus Shimmingles singa*'; novel species that represents genomic bin B

UPWRP_2

Phylogenetic and phylogenomics analyses showed that UPWRP_2 should be assigned to the genus *Haliangium*, of the family Haliangiaceae, of the order Myxococcales, of the class Deltaproteobacteria and of the phylum Proteobacteria. As UPWRP_2 has a 16S rRNA sequence similarity of 98.37% to the closest neighbour, UPWRP_2 has been proposed to be a novel bacteria species under the genus *Haliangium* in accordance to the taxonomic classification boundary proposed by Yarza *et al.* (2014); nomenclature of the novel species is designated as:

- '*Haliangium clustero*'; novel species of *Haliangium*

Different clades of *Haliangium* produced different spatial structures as shown with a mixture of HalianMix and Ribo_Halia1029_17 probes (Figure 5-14). Structure of *Haliangium clustero* did not resemble the other *Haliangium* species that played an important role in denitrification in wastewater treatment systems (McIlroy *et al.*, 2014). Through the hybridisation of probe Halia183, it was observed that *Haliangium clustero* formed aggregates that did not include other *Haliangium* species targeted by the RiboTagger and HalianMix probes. In contrast to what has been described for *Haliangium* in wastewater treatment systems, *Haliangium* species isolated from other environmental sources such as the coastal region have been reported to form fruiting bodies in

axenic cultures. The fruiting bodies highly resemble the aggregated structure of *Haliangium* species in fixed samples of activated sludge (Fudou *et al.*, 2002; Iizuka *et al.*, 2003). Social behaviour for the formation of fruiting bodies has been well-documented for *Haliangium* in axenic cultures (Zhang *et al.*, 2005). This study has further established that the formation of fruiting bodies appears to be species-specific even in a mixed community setting, as the fruiting bodies only contained one species of *Haliangium* – at least in the case study of *Haliangium clusterio*.

Interestingly, *Haliangium*-related 16S rRNA gene sequences have been extracted from various sampling sites that are associated with a high salt concentration: coastal regions (Iizuka *et al.*, 2003) and seaweeds (Fudou *et al.*, 2002). However, the aerobic tank of the wastewater treatment plant from where *Haliangium clusterio* was observed in does not provide such a halophilic environment. Furthermore, the development of fruiting bodies in axenic cultures requires a solid surface and how *Haliangium clusterio* formed fruiting bodies in the suspended liquor of the activated sludge environment is worth pursuing in future studies. In addition, the different structural form of *Haliangium* presents a perfect opportunity to decipher the difference in genetic content between vegetative cells and myxospores. The anti-fungal compound haliangicin has been isolated from *Haliangium ochraceum* SMP-2 and has been shown to disrupt the fungal mitochondrial respiratory chains (Fudou *et al.*, 2001). Future studies should dictate the extraction of a more complete genome of *Haliangium clusterio* with probe Halia183 to determine if *Haliangium clusterio* has the genomic potential to produce any novel secondary metabolites with antifungal or antibacterial properties. Given the high abundance of *Haliangium clusterio* in UPWRP wastewater plant, activated sludge might prove to be a potential untapped source for the production of antifungal or antibacterial products.

5.4.5 Low similarity of draft genomes to reference genomes

Only a small percentage of ORFs (<4%) predicted from the genomes of the unclassified taxa could be correctly assigned to its taxonomic path. There are two plausible explanations for this phenomenon: (1) draft genomes contained genomic fragments with origin from non-target cells and (2) lack of reference genomes in the NCBI NR database. The former scenario is unlikely due to stringent quality controls – both in the management of contamination during FISH-FACS and decontamination of genomic contigs – that ensure that sorted cells were identified to be the taxon of interest with high confidence. In addition, RiboTagger 16S rRNA analysis showed high levels of enrichment (>99%) for the targeted taxon. Furthermore, low level of contamination of the draft genomes was established with the use of essential single-copy gene markers and lineage-specific marker genes.

Therefore, the latter scenario of having a lack of reference genomes in the database is the more probable rationale for the low similarity of draft genomes to the reference database. Genome comparison of the ORFs of *Ca. Shimingles* and *Haliangium clusterio* yielded an average AAI of less than 39% to the genome of its closest lineage. The lack of closely-related reference genomes in databases has shown to produce unreliable taxonomic assignment of ORFs. In a study by Albertsen, Saunders *et al.* (2013), the deliberate removal of specific genomes from the database resulted in a complex taxonomic analysis, where ORFs were assigned to different taxonomic groups and only less than 30% of ORFs could be assigned to the correct taxon. Output of their taxonomic analysis resembled the fragmented taxonomic assignment of ORFs in this thesis, thus showing that draft genomes generated for the unclassified taxa are relatively novel and there is a lack of reference genomes covering their lineage. For instance, only 3.61% of ORFs could be correctly assigned to the genus *Haliangium* because of the presence of only one reference genome in the genus.

5.4.6 Functional analysis of novel taxa

The novel taxa assigned a major proportion of their PEGs (>13%) to the subsystem category: virulence, disease and defence and this would help them to compete with other organisms in the activated sludge environment. Secreted putative Internalin-like proteins were the major virulence genes expressed by the novel taxa and these proteins are responsible for bacterial entry into host cells (Bierne and Cossart, 2007). UPWRP was designed to treat influent wastewater which comprised 90% of domestic waste and 10% of commercial and industrial waste (Lau, 2012). The novel taxa were equipped to survive potential toxic compounds present in the waste influent by harbouring PEGs that were resistant to arsenic, zinc, copper and chromium. These heavy metal compounds are frequently present in industrial wastewater (Barakat, 2011). Sigma-B stress-response cluster present in *Haliangium clusterio* suggest that the organism do not sporulate in response to environmental stress such as glucose limitation, as Sigma-B factors are often produced as an alternative response to sporulation (Boylan *et al.*, 1993). The discovery of sigma-B stress response cluster corroborates with our observation that myxospores were not observed during microscopic visualisation of *Haliangium clusterio* (Figure 5-23). *Haliangium ochraceum* DSM 14365 is known to develop myxospores (Ivanova *et al.*, 2010). The presence of sporulation-associated gene which encodes for peptidyl-tRNA hydrolase (Menez *et al.*, 2002) verifies the sporulation capability of *Haliangium ochraceum* DSM 14365. The lack of myxospores correlates with the absence of sporulation-associated gene in *Haliangium clusterio*.

Chapter 6 Conclusion and perspectives

A strategy for visualizing and characterising members of previously unclassified bacterial taxa inhabiting the activated sludge ecosystem of a tropical wastewater treatment plant has been successfully demonstrated. This strategy was achieved through a novel method of FISH probe design from short sequencing reads in omics dataset (metatranscriptomics or metagenomics). The design of RiboProbe differs from the gold standard of FISH probe design performed through comparative sequence analysis because it skips the need for full-length 16S rRNA sequences. RiboProbes are designed directly from short sequencing reads extracted from the V6 region of the SSU 16S rRNA gene. Quite often, *de novo* probe design is necessary when published probes do not exist for a target population, but difficulties in obtaining near full-length 16S rRNA sequences (≥ 1200 bp) for conventional probe design is a bottleneck for most probe developers. Clearly, RiboProbe design can circumvent this problem and it will be interesting to extend its applications to other ecosystems, where the only available information is obtained from whole community shotgun surveys. The V6 region was selected because it provides a meaningful taxonomic resolution that can distinguish between closely related species, as shown in the case study of *Haliangium clusterio* which shared a 16S rRNA sequence similarity of 98.37% to its closest neighbour. Furthermore, V6-FISH probe was developed to complement the community profiling performed at UPWRP, which was analysed using V6-Ribotags for diversity analysis. With V6-RiboProbes conferring bright fluorescence signal to at least three different phylogenetically distant bacterial taxa (*Thauera*, *Shimingles*, *Haliangium*) in *in situ* hybridisation experiments, the V6 region is a promising candidate for future FISH probe design.

Besides the use of RiboProbe for visualisation, newly designed probes were further extended to methodologies that aided in the characterisation of two unclassified bacterial taxa in the floccular sludge community – an achievement that could not be accomplished due to limitations associated with the current tools of canonical probe design and metagenomics. Successful development of the new variant of FISH probe has important microbial ecological implications for activated sludge wastewater treatment because a large fraction of the floccular sludge community consists of

unclassified bacteria taxa. To further advance wastewater treatment, it is necessary to identify the unclassified bacteria taxa and subsequently understand their ecological contribution to wastewater purification. As a proof of concept, two unclassified taxa pseudonym as UPWRP_1 and UPWRP_2, present in high abundance at the point of deep sequencing at UPWRP were selected for characterisation. RiboProbes were applied for the visualisation of the unclassified bacterial taxa's morphology and spatial interaction with its neighbouring cells. One of the major discoveries through FISH visualisation is the discovery of species-specific fruiting bodies of UPWRP_2, an observation that was not associated with other *Haliangium* species discovered in wastewater. Furthermore, the use of FISH probes to hybridise myxospores of *Haliangium* species is unprecedented, and it showed that UPWRP_2 was unlikely to produce spherical myxospores as compared to other *Haliangium* species - despite being sampled from an identical location in the wastewater plant.

Even though a 33bp-RiboTag represents an OTU (Xie *et al.*, 2016), RiboProbe design requires truncation of the 33bp-RiboTag to fit into the melting curve profile of other canonical FISH probes which have a typical read-length of 17–25 bp (Thiele *et al.*, 2010). Consequently, intended specificity of the 33bp-RiboTag is altered and this was reflected in the case study of *Haliangium clustero*. Consequently, many closely-related taxa that contained the complementary binding site of the truncated FISH probe are included into the sorted samples. Genomic divergence in the sorted samples would further complicate downstream genomic analyses – especially in genomic assembly of repetitive regions in closely-related strains (Treangen and Salzberg, 2012). Future research should attempt the hybridisation of RiboProbe at its original length of 33 bp to retain its original intended specificity, with a concomitant increase in hybridisation temperature to minimise non-specific hybridisation.

A FISH-FACS methodology coupled with RiboProbe was optimised with the aim of recovering draft genomes from the unclassified taxa while keeping DNA contamination to a minimum. DNA contamination is a persistent problem associated with MDA experiments and the strategy for

managing DNA contamination in this thesis can be categorised into ‘pre-sequencing’ and ‘post-sequencing’. In the ‘pre-sequencing’ approach, the introduction of potential contamination into the sample is minimised through implementation of: (1) a comprehensive set of controls to check for DNA contamination; (2) an extensive sterilisation of the FACS machine; (3) careful sample handling and (4) use of ultraclean reagents. After MDA-amplification and genomic sequencing of the sorted samples, a computational method was adopted in the ‘post-sequencing’ approach to decontaminate draft genomes by removing genomic fragments that do not belong to the target taxon. This was achieved through a reference-free unsupervised machine learning tool that uses non-linear dimensionality reduction of tetramer frequencies and statistical clustering to sieve out the contaminant sequences.

Even with a low mean read depth of sequencing ($2.56X \pm 2.97$), draft genomes of the unclassified bacterial taxa were recovered from the enriched samples through bioinformatics analyses. Without the enrichment process, no draft genomes or RiboTags could be assigned to the unclassified taxa in the pre-sorted samples that were sequenced at the same mean read depth. Identities of the unclassified taxa were elucidated through its phylogeny with its closest neighbour from sequence analysis of full-length 16S rRNA gene, and the identities were further substantiated through placement of draft genomes in the reference genome tree using conserved phylogenetic marker genes. UPWRP_1 was identified to be members of a novel genus which has been named as *Ca. Shimingles*, and the two clades in the novel genus as *Ca. Shimingles merlion* and *Ca. Shimingles singa*. *Ca. Shimingles* shared approximately 90% 16S rRNA gene similarity to its closest neighbour in the SILVA database. UPWRP_2 was identified to be a novel species under genus *Haliangium* and has been named as ‘*Haliangium clustero*’. *Haliangium clustero* shared a 98.37% sequence similarity to its closest neighbour in the 16S rRNA database.

Obtaining draft genomes of the unclassified taxa has important implications because it provides a solid reference for phylogenetic assignment of genomic fragments from future metagenomic dataset (Albertsen *et al.*, 2013b). Draft genomes also serve as a prelude to deciphering the

ecological functions of the unclassified bacteria in wastewater treatment through genome annotation. A succinct genome annotation performed for both *Ca. Shimingles* and *Haliangium cluster* showed that the organisms are likely to perform carbohydrate metabolism under aerobic conditions, and they are equipped with virulence genes and genes that provide resistance to toxic compounds. The next direction of this project is a more detailed genome annotation and metabolic reconstruction, which would be performed through:

- (1) Identifying open reading frames (ORFs) from the draft genomes and assigning the ORFs a KEGG Orthology (KO) identifier by mapping the ORFs to the KEGG database. This would predict the potential function of the novel taxa.
- (2) Mapping of mRNAs - from the transcriptomics study performed at UPWRP – to the draft genomes. This would identify the functional genes expressed by the novel taxa.

While a detailed genome annotation could provide conjectures about the potential metabolic capabilities of the novel taxa, this information would have been better supplemented with physiochemical parameter measurements taken at the point of sampling. Two draft genomes were obtained for *Ca. Shimingles* through the multi-prong approach of multiple displacement amplification (MDA), genomic sequencing and binning. *De novo* co-assembly of multiple enriched samples with high number of sorted events (n=1000) averaged out the amplification and representation bias catalysed by MDA (Blanco *et al.*, 1989; Tringe and Rubin, 2005), and high levels of genome completeness could be obtained for the genomic bins of *Ca. Shimingles*. Unfortunately, a high-quality draft genome could not be obtained for *Haliangium cluster* due to the complication of strain heterogeneity during *de novo* assembly with closely-related *Haliangium* species in the sorted samples. Complication of strain heterogeneity can be avoided in the future studies through design of a longer and more specific FISH probes, or the design of multiple probes from the long representative sequence of RiboTags.

MDA is a double-edged sword that can aid in differential coverage binning due to its inherent amplification bias across multiple samples, and the first case study of using differential coverage binning on MDA-amplified samples to separate two different clades of a novel genus is presented. MDA is useful in reconstructing draft genomes only if the sorted samples contain a highly-enriched target population like *Candidatus Shimingles*. In contrast, MDA produces contigs with different coverage such that contigs whose coverage profile varied significantly might not be captured in the genomic bins. Low yields of genomic material obtained from FISH-FACS warrants the need for MDA unless sufficient number of sorted cells ($>10^6$) are collected. Sorting of $>10^6$ cells is a process hampered by the time and access to the FACS machine. Nevertheless, FISH-FACS with highly specific RiboProbe without downstream MDA should be attempted in the future to infer any difference in sequence analysis.

In the case study of *Haliangium clusterio*, importance of constantly evaluating environmental specificity of the FISH probe against a sample-centric genomic database was reinforced (Karst *et al.*, 2016a). Activated sludge biosphere of UPWRP harbours a large diversity of *Haliangium* species, as revealed by the presence of several closely-related RiboTags to *Haliangium clusterio* that were not observed in the SILVA database. The untapped diversity of *Haliangium* can be further explored to produce probes and primers against *Haliangium* with better accuracy.

Characterisation of the unclassified bacterial taxa in this thesis has surpassed their initial description in the omics study performed at UPWRP, where the only information available about the unclassified taxa was the 33bp-RiboTag. Methods developed in this thesis have yielded more information and have led to the visualisation, identification and recovery of draft genomes spanning Mbp for the unclassified taxa. This thesis outlines a novel methodology for future researchers who are interested in performing targeted characterisation of a specific taxon in their omics dataset, especially in the scenario where no canonical probe or full-length 16S rRNA sequences are available for FISH probe design.

References

- Achilles J, Stahl F, Harms H, Müller S. (2007). Isolation of intact RNA from cytometrically sorted *Saccharomyces cerevisiae* for the analysis of intrapopulation diversity of gene expression. *Nat Protoc* **2**: 2203–2211.
- Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, *et al.* (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**: 551–554.
- Albertsen M, Hansen LBS, Saunders AM, Nielsen PH, Nielsen KL. (2012). A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal. *ISME J* **6**: 1094–1106.
- Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. (2013a). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**: 533–538.
- Albertsen M, McIlroy SJ, Stokholm-Bjerregaard M, Karst SM, Nielsen PH. (2016). ‘Candidatus *Propionivibrio aalborgensis*’: A novel glycogen accumulating organism abundant in full-scale enhanced biological phosphorus removal plants. *Front Microbiol* **7**: 1033.
- Albertsen M, Saunders AM, Nielsen KL, Nielsen PH. (2013b). Metagenomes obtained by ‘deep sequencing’ - What do they tell about the enhanced biological phosphorus removal communities? *Water Sci Technol* **68**: 1959–1968.
- Amann RI, Binder BJ, Olson RJ, Chisholm SW, Devereux R, Stahl DA. (1990a). Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl Envir Microbiol* **56**: 1919–1925.
- Amann RI, Fuchs BM. (2008). Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nat Rev Microbiol* **6**: 339–348.
- Amann RI, Krumholz L, Stahl DA. (1990b). Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. *J Bacteriol* **172**: 762–770.
- Amann RI, Ludwig W, Schleifer KH. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* **59**: 143–169.
- Amor K Ben, Breeuwer P, Verbaarschot P, Rombouts FM, Akkermans ADL, De Vos WM, *et al.* (2002). Multiparametric flow cytometry and cell sorting for the assessment of viable, injured, and dead bifidobacterium cells during bile salt stress. *Appl Environ Microbiol* **68**: 5209–5216.
- Arnold LW, Lannigan J. (2010). Practical issues in high-speed cell sorting. *Curr Protoc Cytom* **1.24.1-1.24.30**.
- Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, *et al.* (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**: 75.
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al.* (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19**: 455–477.
- Barakat MA. (2011). New trends in removing heavy metals from industrial wastewater. *Arab J Chem* **4**: 361–377.
- Behrens S, Ruhland C, Inacio J, Huber H, Fonseca A, Spencer-Martins I, *et al.* (2003). In Situ Accessibility of Small-Subunit rRNA of Members of the Domains Bacteria, Archaea, and Eucarya to Cy3-Labeled Oligonucleotide Probes. *Appl Environ Microbiol* **69**: 1748–1758.
- Bergey DH. (2005). *Bergey’s Manual of Systematic Bacteriology - Vol 2: The Proteobacteria, Part C - The Alpha-, Beta-, Delta- and Epsilonproteobacteria*. Springer US.

- Berney M, Hammes F, Bosshard F, Weilenmann H-U, Egli T. (2007). Assessment and interpretation of bacterial viability by using the LIVE/DEAD BacLight Kit in combination with flow cytometry. *Appl Environ Microbiol* **73**: 3283–90.
- Bierne H, Cossart P. (2007). *Listeria monocytogenes* Surface Proteins: from Genome Predictions to Function. *Microbiol Mol Biol Rev* **71**: 377–397.
- Biesterfeld S, Figueroa L, Hernandez M, Russell P. Quantification of nitrifying bacterial populations in a full-scale nitrifying trickling filter using fluorescent in situ hybridization. *Water Environ Res* **73**: 329–338.
- Binga EK, Lasken RS, Neufeld JD. (2008). Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J* **2**: 233–241.
- Binnie C, Lampe M, Losick R. (1986). Gene encoding the sigma 37 species of RNA polymerase sigma factor from *Bacillus subtilis*. *Proc Natl Acad Sci U S A* **83**: 5943–7.
- Blainey PC. (2013). The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol Rev* **37**: 407–427.
- Blainey PC, Quake SR. (2011). Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Res* **39**: e19.
- Blanco L, Bernad A, Lázaro JM, Martín G, Garmendia C, Salas M. (1989). Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J Biol Chem* **264**: 8935–8940.
- Bock E, Koops HP, Moller UC, Rudert M. (1990). A new facultatively nitrite oxidizing bacterium, *Nitrobacter vulgaris* sp. nov. *Arch Microbiol* **153**: 105–110.
- Boylan SA, Redfield AR, Brody MS, Price CW. (1993). Stress-induced activation of the Sigma B transcription factor of *Bacillus subtilis*. *J Bacteriol* **175**: 7931–7937.
- Bruder LM, Dörkes M, Fuchs BM, Ludwig W, Liebl W. (2016). Flow cytometric sorting of fecal bacteria after in situ hybridization with polynucleotide probes. *Syst Appl Microbiol* **39**: 464–475.
- Buchfink B, Xie C, Huson DH. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59–60.
- Burke CM, Darling AE. (2016). A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq. *PeerJ* **4**: e2492.
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.
- Caporaso JG, Lauber CL, Walters W a, Berg-Lyons D, Huntley J, Fierer N, *et al.* (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**: 1621–1624.
- Caro A, Gros O, Got P, De Wit R, Troussellier M. (2007). Characterization of the population of the sulfur-oxidizing symbiont of *Codakia orbicularis* (Bivalvia, Lucinidae) by single-cell analyses. *Appl Environ Microbiol* **73**: 2101–2109.
- Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. (2015). A practical guide to de novo genome assembly using long reads. e-pub ahead of print, doi: 10.1101/029306.
- Chistoserdova L. (2014). Is metagenomics resolving identification of functions in microbial communities? *Microb Biotechnol* **7**: 1–4.
- Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**: 1283–1287.

- Clingenpeel S, Clum A, Schwientek P, Rinke C, Woyke T. (2015). Reconstructing each cell's genome within complex microbial communities-dream or reality? *Front Microbiol* **6**. e-pub ahead of print, doi: 10.3389/fmicb.2014.00771.
- Clingenpeel S, Schwientek P, Hugenholtz P, Woyke T. (2014). Effects of sample treatments on genome recovery via single-cell genomics. *ISME J* **8**: 2546–2549.
- Czechowska K, Johnson DR, van der Meer JR. (2008). Use of flow cytometric methods for single-cell analysis in environmental microbiology. *Curr Opin Microbiol* **11**: 205–212.
- Daims H, Nielsen JL, Nielsen PH, Schleifer KH, Wagner M. (2001). In Situ Characterization of Nitrospira-Like Nitrite-Oxidizing Bacteria Active in Wastewater Treatment Plants. *Appl Environ Microbiol* **67**: 5273–5284.
- Daims H, Taylor MW, Wagner M. (2006). Wastewater treatment: a model system for microbial ecology. *Trends Biotechnol* **24**: 483–489.
- Dean FB, Nelson JR, Giesler TL, Lasken RS. (2001). Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**: 1095–1099.
- DeLong EF, Béjà O. (2010). The light-driven proton pump proteorhodopsin enhances bacterial survival during tough times. *PLoS Biol* **8**. e-pub ahead of print, doi: 10.1371/journal.pbio.1000359.
- DeLong EF, Wickham GS, Pace NR. (1989). Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science* **243**: 1360–1363.
- Dodsworth JA, Blainey PC, Murugapiran SK, Swingley WD, Ross CA, Tringe SG, *et al.* (2013). Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat Commun* **4**: 1854.
- Dupont CL, Rusch DB, Yooseph S, Lombardo M-J, Richter RA, Valas R, *et al.* (2012). Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* **6**: 1186–1199.
- Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.
- Eikelboom DH. (2000). Process Control of Activated Sludge Plants by Microscopic Investigation. IWA Publishing.
- Eloe-Fadrosch EA, Ivanova NN, Woyke T, Kyrpides NC. (2016). Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat Microbiol* **1**: 15032.
- Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, *et al.* (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**: e1319.
- Erko S, Ebers J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiol Today* **33**: 152–155.
- Falcioni T, Manti A, Boi P, Canonico B, Balsamo M, Papa S. (2006). Comparison of disruption procedures for enumeration of activated sludge floc bacteria by flow cytometry. *Cytom Part B - Clin Cytom* **70**: 149–153.
- Faust K, Raes J. (2012). Microbial interactions: from networks to models. *Nat Rev Microbiol* **10**: 538–550.
- Feng S, Tan CH, Constancias F, Kohli GS, Cohen Y, Rice SA. (2017). Predation by Bdellovibrio bacteriovorus significantly reduces viability and alters the microbial community composition of activated sludge flocs and granules. *FEMS Microbiol Ecol* **93**: 359–69.
- Forbes CM, O'Leary ND, Dobson AD, Marchesi JR. (2009). The contribution of 'omic'-based approaches to the study of enhanced biological phosphorus removal microbiology: Minireview.

FEMS Microbiol Ecol **69**: 1–15.

Forster S, Snape JR, Lappin-Scott HM, Porter J. (2002). Simultaneous fluorescent gram staining and activity assessment of activated sludge bacteria. *Appl Environ Microbiol* **68**: 4772–4779.

Frølund B, Palmgren R, Keiding K, Nielsen PH. (1996). Extraction of extracellular polymers from activated sludge using a cation exchange resin. *Water Res* **30**: 1749–1758.

Fuchs BM, Wallner G, Beisker W, Schwippl I, Ludwig W, Amann R. (1998). Flow cytometric analysis of the in situ accessibility of Escherichia coli 16S rRNA for fluorescently labeled oligonucleotide probes. *Appl Environ Microbiol* **64**: 4973–4982.

Fudou R, Iizuka T, Sato S, Ando T, Shimba N, Yamanaka S. (2001). Haliangicin, a novel antifungal metabolite produced by a marine myxobacterium. 2. Isolation and structural elucidation. *J Antibiot (Tokyo)* **54**: 153–156.

Fudou R, Jojima Y, Iizuka T, Yamanaka S. (2002). Haliangium ochraceum gen. nov., sp. nov. and Haliangium tepidum sp. nov.: novel moderately halophilic myxobacteria isolated from coastal saline environments. *J Gen Appl Microbiol* **48**: 109–116.

Fujitani H, Ushiki N, Tsuneda S, Aoi Y. (2014). Isolation of sublineage I Nitrospira by a novel cultivation strategy. *Environ Microbiol* **16**: 3030–3040.

Galvez A, Maqueda M, Martinez-Bueno M, Valdivia E. (1998). Publication rates reveal trends in microbiological research. *ASM News* **64**: 269–275.

García Martín H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, *et al.* (2006). Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24**: 1263–1269.

Gifford SM, Sharma S, Booth M, Moran MA. (2013). Expression patterns reveal niche diversification in a marine microbial assemblage. *ISME J* **7**: 281–298.

Giovannoni SJ. (1990). Genetic diversity in Sargasso sea bacterioplankton. *Nature* **345**: 183–187.

Giovannoni SJ, DeLong EF, Olsen GJ, Pace NR. (1988). Phylogenetic group-specific oligodeoxynucleotide probes for identification of single microbial cells. *J Bacteriol* **170**: 720–726.

González M, Reyes-Jara A, Suazo M, Jo WJJ, Vulpe C. (2008). Expression of copper-related genes in response to copper load. In: Vol. 88. *American Journal of Clinical Nutrition*. e-pub ahead of print, doi: 88/3/830S [pii].

Gougoulas C, Shaw LJ. (2012). Evaluation of the environmental specificity of Fluorescence In Situ Hybridization (FISH) using Fluorescence-Activated Cell Sorting (FACS) of probe (PSE1284)-positive cells extracted from rhizosphere soil. *Syst Appl Microbiol* **35**: 533–540.

Guo J, Cole JR, Zhang Q, Brown CT, Tiedje JM. (2016). Microbial Community Analysis with Ribosomal Gene Fragments from Shotgun Metagenomes. *Appl Environ Microbiol* **82**: 157–166.

Haroon MF, Hu S, Shi Y, Imelfort M, Keller J, Hugenholtz P, *et al.* (2013a). Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. *Nature* **500**: 567–570.

Haroon MF, Skennerton CT, Steen JA, Lachner N, Hugenholtz P, Tyson GW. (2013b). Chapter One – In-Solution Fluorescence In Situ Hybridization and Fluorescence-Activated Cell Sorting for Single Cell and Population Genome Recovery. In: Vol. 531. *Methods in Enzymology*. Academic Press, pp 3–19.

Hartigan J a., Hartigan PM. (1985). The Dip Test of Unimodality. *Ann Stat* **13**: 70–84.

Hartzell PLL, White DJJ, Hartzell PLL, White DJJ. (2001). Myxospores. In: *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd: UK. e-pub ahead of print, doi: 10.1038/npg.els.0000307.

- He S, Gall DL, McMahon KD. (2007). 'Candidatus accumulibacter' population structure in enhanced biological phosphorus removal sludges as revealed by polyphosphate kinase genes. *Appl Environ Microbiol* **73**: 5865–5874.
- Heine F, Stahl F, Sträuber H, Wiacek C, Benndorf D, Repenning C, *et al.* (2009). Prediction of flocculation ability of brewing yeast inoculates by flow cytometry, proteome analysis, and mRNA profiling. In: Vol. 75. *Cytometry Part A*. pp 140–147.
- Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H, Schroth G, *et al.* (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**: 463–467.
- Hong S, Bunge J, Leslin C, Jeon S, Epstein SS. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J* **3**: 1365–1373.
- Hugenholtz P, Tyson GW, Blackall LL. (2002). Design and evaluation of 16S rRNA-targeted oligonucleotide probes for fluorescence in situ hybridization. *Methods Mol Biol* **179**: 29–42.
- Hugenholtz P, Tyson GW, Webb RI, Wagner AM, Blackall LL. (2001). Investigation of candidate division TM7, a recently recognized major lineage of the domain Bacteria with no known pure-culture representatives. *Appl Environ Microbiol* **67**: 411–419.
- Huson DH, Beier S, Flade I, Górski A, El-Hadidi M, Mitra S, *et al.* (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput Biol* **12**: e1004957.
- Iizuka T, Jojima Y, Fudou R, Tokura M, Hiraishi A, Yamanaka S. (2003). *Enhygromyxa salina* gen. nov., sp. nov., a Slightly Halophilic Myxobacterium Isolated from the Coastal Areas of Japan. *Syst Appl Microbiol* **26**: 189–196.
- Invitrogen. (2016). TOPO TA Cloning Kit for Sequencing. <https://www.thermofisher.com/order/catalog/product/K457502>.
- Irie K, Fujitani H, Tsuneda S. (2016). Physical enrichment of uncultured *Accumulibacter* and *Nitrospira* from activated sludge by unlabeled cell sorting technique. *J Biosci Bioeng* **122**: 475–481.
- Ivanova N, Daum C, Lang E, Abt B, Kopitz M, Saunders E, *et al.* (2010). Complete genome sequence of *Haliangium ochraceum* type strain (SMP-2). *Stand Genomic Sci* **2**: 96–106.
- Janda JM, Abbott SL. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol* **45**: 2761–2764.
- Jiang K, Sanseverino J, Chauhan A, Lucas S, Copeland A, Lapidus A, *et al.* (2012). Complete genome sequence of *Thauera aminoaromatica* strain MZ1T. *Stand Genomic Sci* **6**: 325–335.
- Juretschko S, Timmermann G, Schmid M, Schleifer KH, Pommerening-Röser A, Koops HP, *et al.* (1998). Combined molecular and conventional analyses of nitrifying bacterium diversity in activated sludge: *Nitrosococcus mobilis* and *Nitrospira*-like bacteria as dominant populations. *Appl Environ Microbiol* **64**: 3042–3051.
- Kalogeratos A, Likas A. (2012). Dip-means: An incremental clustering method for estimating the number of clusters. In: Vol. 3. *Advances in Neural Information Processing Systems*.
- Kalyuzhnaya MG, Zabinsky R, Bowerman S, Baker DR, Lidstrom ME, Chistoserdova L. (2006). Fluorescence in situ hybridization-flow cytometry-cell sorting-based method for separation and enrichment of type I and type II methanotroph populations. *Appl Environ Microbiol* **72**: 4293–4301.
- Kang DD, Froula J, Egan R, Wang Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**: e1165.
- Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. (2016a). Thousands

- of primer-free, high-quality, full-length SSU rRNA sequences from all domains of life. e-pub ahead of print, doi: 10.1101/070771.
- Karst SM, Kirkegaard RH, Albertsen M. (2016b). mmgenome: a toolbox for reproducible genome extraction from metagenomes. *bioRxiv*. e-pub ahead of print, doi: 10.1101/059121.
- van Kessel MAHJ, Speth DR, Albertsen M, Nielsen PH, Op den Camp HJM, Kartal B, *et al.* (2015). Complete nitrification by a single microorganism. *Nature* **528**: 555–559.
- Kim JM, Lee HJ, Kim SY, Song JJ, Park W, Jeon CO. (2010). Analysis of the fine-scale population structure of ‘*Candidatus accumulibacter phosphatis*’ in enhanced biological phosphorus removal sludge, using fluorescence in situ hybridization and flow cytometric sorting. *Appl Environ Microbiol* **76**: 3825–35.
- Kindaichi T, Yamaoka S, Uehara R, Ozaki N, Ohashi A, Albertsen M, *et al.* (2016). Phylogenetic diversity and ecophysiology of Candidate phylum Saccharibacteria in activated sludge. *FEMS Microbiol Ecol* **92**: 1–11.
- Kircher M, Sawyer S, Meyer M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **40**. e-pub ahead of print, doi: 10.1093/nar/gkr771.
- Kirkegaard RH, Dueholm MS, McIlroy SJ, Nierychlo M, Karst SM, Albertsen M, *et al.* (2016). Genomic insights into members of the candidate phylum Hyd24-12 common in mesophilic anaerobic digesters. *ISME J* **10**: 2352–2364.
- Kjelleberg S, Xie C, Zhao F, Williams RB., Huson DH, Drautz DI, *et al.* (2014). Metagenomic and Metatranscriptomics of a microbial community driving wastewater purification. *Unpubl Manuscr.*
- Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, *et al.* (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**: e1.
- Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. (2008). A bioinformatician’s guide to metagenomics. *Microbiol Mol Biol Rev* **72**: 557–578.
- Kvist T, Ahring BKK, Lasken RSS, Westermann P. (2007). Specific single-cell isolation and genomic amplification of uncultured microorganisms. *Appl Microbiol Biotechnol* **74**: 926–935.
- Lajoie CA, Layton AC, Gregory IR, Sayler GS, Taylor DE, Meyers AJ. (2000). Zoogeal clusters and sludge dewatering potential in an industrial activated-sludge wastewater treatment plant. *Water Environ Res* **72**: 56–64.
- Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A* **82**: 6955–6959.
- Lasken RS. (2012). Genomic sequencing of uncultured microorganisms from single cells. *Nat Rev Microbiol* **10**: 631–640.
- Lasken RS, McLean JS. (2014). Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat Rev Genet* **15**: 577–584.
- Lasken RS, Stockwell TB. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol* **7**: 19.
- Lapidou CS, Rittmann BE. (2002). A unified theory for extracellular polymeric substances, soluble microbial products, and active and inert biomass. *Water Res* **36**: 2711–2720.
- Lau CL. (2012). Overview of UPWRP. Public Utilities Board.
- Law Y, Kirkegaard RH, Cokro AA, Liu X, Arumugam K, Xie C, *et al.* (2016). Integrative microbial community analysis reveals full-scale enhanced biological phosphorus removal under tropical

conditions. *Sci Rep* **6**: 25719.

Lee DH, Zo YG, Kim SJ. (1996). Nonradioactive method to study genetic profiles of natural bacterial communities by PCR-single-strand-conformation polymorphism. *Appl Environ Microbiol* **62**: 3112–3120.

Lee N, Nielsen PH, Andreasen KH, Juretschko S, Nielsen JL, Schleifer KH, *et al.* (1999). Combination of fluorescent in situ hybridization and microautoradiography-a new tool for structure-function analyses in microbial ecology. *Appl Environ Microbiol* **65**: 1289–1297.

Lee PKH, Men Y, Wang S, He J, Alvarez-Cohen L. (2015). Development of a fluorescence-activated cell sorting method coupled with whole genome amplification to analyze minority and trace dehalococcoides genomes in microbial communities. *Environ Sci Technol* **49**: 1585–1593.

Li T, Wu T Di, Mazéas L, Toffin L, Guerquin-Kern JL, Leblon G, *et al.* (2008). Simultaneous analysis of microbial identity and function using NanoSIMS. *Environ Microbiol* **10**: 580–588.

Liesack W, Janssen PH, Rainey FA, Ward-Rainey NL, Stackebrandt E, Elsas JD van, *et al.* (1997). Microbial diversity in soil: the need for a combined approach using molecular and cultivation techniques. 375–439.

Lightfield J, Fram NR, Ely B. (2011). Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLoS One* **6**: e17677.

De Los Reyes FL, Oerther DB, De Los Reyes MF, Hernandez M, Raskin L. (1998). Characterization of filamentous foaming in activated sludge systems using oligonucleotide hybridization probes and antibody probes. In: Vol. 37. *Water Science and Technology*. pp 485–493.

Loy A, Arnold R, Tischler P, Rattei T, Wagner M, Horn M. (2008). ProbeCheck - A central resource for evaluating oligonucleotide probe coverage and specificity. *Environ Microbiol* **10**: 2894–2898.

Loy A, Horn M, Wagner M. (2003). probeBase: an online resource for rRNA-targeted oligonucleotide probes. *Nucleic Acids Res* **31**: 514–516.

Lücker S, Schwarz J, Gruber-Dorninger C, Spieck E, Wagner M, Daims H. (2014). Nitrotoga-like bacteria are previously unrecognized key nitrite oxidizers in full-scale wastewater treatment plants. *ISME J* **9**: 708–720.

Lücker S, Steger D, Kjeldsen KU, MacGregor BJ, Wagner M, Loy A. (2007). Improved 16S rRNA-targeted probe set for analysis of sulfate-reducing bacteria by fluorescence in situ hybridization. *J Microbiol Methods* **69**: 523–528.

Lücker S, Wagner M, Maixner F, Pelletier E, Koch H, Vacherie B, *et al.* (2010). A Nitrospira metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proc Natl Acad Sci U S A* **107**: 13479–13484.

Ludwig W, Amann R, Martinez-Romero E, Schönhuber W, Bauer S, Neef A, *et al.* (1998a). rRNA based identification and detection systems for rhizobia and other bacteria. *Plant Soil* **204**: 1–19.

Ludwig W, Strunk O, Klugbauer S, Klugbauer N, Weizenegger M, Neumaier J, *et al.* (1998b). Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* **19**: 554–568.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadukumar, *et al.* (2004). ARB: A software environment for sequence data. *Nucleic Acids Res* **32**: 1363–71.

Luo C, Tsementzi D, Kyrpides NCC, Konstantinidis KT. (2012). Individual genome assembly from complex community short-read metagenomic datasets. *ISME J* **6**: 898–901.

Lux M, Krüger J, Rinke C, Maus I, Schlüter A, Woyke T, *et al.* (2016). acdc - Automated Contamination Detection and Confidence estimation for single-cell genome data. *BMC Bioinformatics* **17**: 543.

- Von Luxburg U. (2007). A tutorial on spectral clustering. *Stat Comput* **17**: 395–416.
- Lynch MDJ, Neufeld JD. (2015). Ecology and exploration of the rare biosphere. *Nat Rev Microbiol* **13**: 217–229.
- van der Maaten L. (2014). Accelerating t-SNE using Tree-Based Algorithms. *J Mach Learn Res* **15**: 3221–3245.
- Mainstone CP, Parr W. (2002). Phosphorus in rivers - Ecology and management. *Sci Total Environ* **282–283**: 25–47.
- Makarova KS, Brouns SJJ, Horvath P, Sas DF, Wolf YI. (2012). Evolution and classification of the CRISPR-Cas systems. *Nat Rev ...* **9**: 467–477.
- Manz W, Amann R, Ludwig W, Wagner M, Schleifer K-H. (1992). Phylogenetic Oligodeoxynucleotide Probes for the Major Subclasses of Proteobacteria: Problems and Solutions. *Syst Appl Microbiol* **15**: 593–600.
- Marie D, Vaulot D, Partensky F. (1996). Application of the novel nucleic acid dyes YOYO-1, YO-PRO-1, and PicoGreen for flow cytometric analysis of marine prokaryotes. *Appl Environ Microbiol* **62**: 1649–1655.
- Mcllroy SJ, Saunders AM, Albertsen M, Nierychlo M, Mcllroy B, Hansen AA, *et al.* (2015). MiDAS: The field guide to the microbes of activated sludge. *Database (Oxford)* **2015**: bav062.
- Mcllroy SJ, Starnawska A, Starnawski P, Saunders AM, Nierychlo M, Nielsen PH, *et al.* (2014). Identification of active denitrifiers in full-scale nutrient removal wastewater treatment systems. *Environ Microbiol* **18**: 50–64.
- Mcllroy SJ, Tillett D, Petrovski S, Seviour RJ. (2011). Non-target sites with single nucleotide insertions or deletions are frequently found in 16S rRNA sequences and can lead to false positives in fluorescence in situ hybridization (FISH). *Environ Microbiol* **13**: 33–47.
- McLean JS, Lombardo MJ, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, *et al.* (2013). Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proc Natl Acad Sci U S A* **110**: E2390-2399.
- Menez J, Buckingham RH, De Zamaroczy M, Campelli CK. (2002). Peptidyl-tRNA hydrolase in *Bacillus subtilis*, encoded by spoVC, is essential to vegetative growth, whereas the homologous enzyme in *Saccharomyces cerevisiae* is dispensable. *Mol Microbiol* **45**: 123–129.
- Mino T, Satoh H. (2006). Wastewater genomics. *Nat Biotechnol* **24**: 1229–1230.
- Morita RYY. (1998). Bacteria in oligotrophic environments. Starvation-Survival lifestyle. *Limnol Ocean* **43**: 1021–1022.
- Moter A, Göbel UB. (2000). Fluorescence in situ hybridization (FISH) for direct visualization of microorganisms. *J Microbiol Methods* **41**: 85–112.
- Mudaly DD, Atkinson BW, Bux F. (2001). 16S rRNA in situ probing for the determination of the family level community structure implicated in enhanced biological nutrient removal. *Water Sci Technol* **43**: 91–8.
- Müller S, Nebe-Von-Caron G. (2010). Functional single-cell analyses: Flow cytometry and cell sorting of microbial populations and communities. *FEMS Microbiol Rev* **34**: 554–587.
- Muyzer G, De Waal EC, Uitterlinden AG. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol* **59**: 695–700.
- Neufeld JD, Chen Y, Dumont MG, Murrell JC. (2008). Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. *Environ Microbiol* **10**:

1526–1535.

Nguyen HTT, Le VQ, Hansen AA, Nielsen JL, Nielsen PH. (2011). High diversity and abundance of putative polyphosphate-accumulating Tetrasphaera-related bacteria in activated sludge systems. *FEMS Microbiol Ecol* **76**: 256–267.

Nguyen HTT, Nielsen JL, Nielsen PH. (2012). ‘Candidatus Halomonas phosphatis’, a novel polyphosphate-accumulating organism in full-scale enhanced biological phosphorus removal plants. *Environ Microbiol* **14**: 2826–2837.

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, *et al.* (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotech* **32**: 822–828.

Nielsen PH, Lemmer H, Daims H. (2009). FISH Handbook for Biological Wastewater Treatment. IWA Publishing.

Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, *et al.* (2013). Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol* **20**: 714–737.

Oehmen A, Carvalho G, Lopez-Vazquez CM, van Loosdrecht MCM, Reis MAM. (2010). Incorporating microbial ecology into the metabolic modelling of polyphosphate accumulating organisms and glycogen accumulating organisms. *Water Res* **44**: 4992–5004.

Oehmen A, Zeng RJ, Yuan Z, Keller J. (2005). Anaerobic metabolism of propionate by polyphosphate-accumulating organisms in enhanced biological phosphorus removal systems. *Biotechnol Bioeng* **91**: 43–53.

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA. (1986). Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* **40**: 337–365.

Onuki M, Satoh N, Mino T, Matsuo T. (2000). Application of molecular methods to microbial community analysis of activated sludge. *Water Sci Technol* **42**: 17–22.

Overbeek R, Begley T, Butler RM, Choudhuri J V., Chuang HY, Cohoon M, *et al.* (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33**: 5691–5702.

Pace N. (1997). A molecular view of microbial diversity and the biosphere. *Science* **276**: 734–740.

Pace N, Stahl D, Lane D, Olsen G. (1985). Analyzing natural microbial populations by rRNA sequences. *ASM Am Soc Microbiol News* **51**: 4–12.

Park H-S, Schumacher R, Kilbane JJ. (2005). New method to characterize microbial diversity using flow cytometry. *J Ind Microbiol Biotechnol* **32**: 94–102.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043–1055.

Pearson WR. (2013). An introduction to sequence similarity (‘homology’) searching. *Curr Protoc Bioinforma* **3**. e-pub ahead of print, doi: 10.1002/0471250953.bi0301s42.

Van de Peer Y, Chapelle S, De Wachter R. (1996). A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Res* **24**: 3381–3391.

Pernthaler J, Glöckner FO, Schönhuber W, Amann R. (2001). Fluorescence in situ hybridization with rRNA-targeted oligonucleotide probes. *Methods Microbiol* **30**: 207–226.

Picard C, Ponsonnet C, Paget E, Nesme X, Simonet P. (1992). Detection and enumeration of bacteria in soil by direct DNA extraction and polymerase chain reaction. *Appl Environ Microbiol*

58: 2717–2722.

Pijuan M, Saunders AM, Guisasola A, Baeza JA, Casas C, Blackall LL. (2004). Enhanced biological phosphorus removal in a sequencing batch reactor using propionate as the sole carbon source. *Biotechnol Bioeng* **85**: 56–67.

Podar M, Abulencia CB, Walcher M, Hutchison D, Zengler K, Garcia JA, *et al.* (2007). Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl Environ Microbiol* **73**: 3205–3214.

Podar M, Keller M, Hugenholtz P. (2009). Single Cell Whole Genome Amplification of Uncultivated Organisms. *Genome* 83–99.

Pop M, Salzberg SL. (2008). Bioinformatics challenges of new sequencing technology. *Trends Genet* **24**: 142–149.

Pruesse E, Peplies J, Glöckner FO. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**: 1823–1829.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. (2012). NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res* **40**: D130–D135.

Qin J-J, Oo MH, Tao G, Kekre KA, Hashimoto T. (2009). Pilot study of a submerged membrane bioreactor for water reclamation. *Water Sci Technol* **60**: 3269–74.

Raghunathan A, Ferguson HR, Bornarth CJ, Song W, Driscoll M, Lasken RS. (2005). Genomic DNA amplification from a single bacterium. *Appl Environ Microbiol* **71**: 3342–3347.

Rainey FA, Ward N, Sly LI, Stackebrandt E. (1994). Dependence on the taxon composition of clone libraries for PCR amplified, naturally occurring 16S rDNA, on the primer pair and the cloning system used. *Experientia* **50**: 796–797.

Ramazzotti M, Berná L, Donati C, Cavalieri D. (2015). riboFrame: An improved method for microbial taxonomy profiling from non-targeted metagenomics. *Front Genet* **6**: 329.

Rappé MS, Giovannoni SJ. (2003). The uncultured microbial majority. *Annu Rev Microbiol* **57**: 369–394.

Ravel J. (2012). Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS One* **7**: e39315.

Richter M, Rosselló-Móra R, Oliver Glöckner F, Peplies J. (2016). JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* **32**: 929–931.

Riesenfeld CS, Schloss PD, Handelsman J. (2004). METAGENOMICS: Genomic Analysis of Microbial Communities. **38**: 525–552.

Rinke C, Lee J, Nath N, Goudeau D, Thompson B, Poulton N, *et al.* (2014). Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat Protoc* **9**: 1038–1048.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Robertson BR, Button DK. (1989). Characterizing aquatic bacteria according to population, cell size, and apparent DNA content by flow cytometry. *Cytometry* **10**: 70–76.

Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW. (2009). Whole genome amplification and de novo assembly of single bacterial cells. *PLoS One* **4**: e6864.

Rodriguez-R LM, Konstantinidis KT. (2014). Estimating coverage in metagenomic data sets and why

it matters. *ISME J* **8**: 2349–2351.

Rodriguez-R LM, Konstantinidis KT. (2016). The enveomics collection : a toolbox for specialized analyses of microbial genomes and metagenomes. *Peer J Prepr.* e-pub ahead of print, doi: 10.7287/PEERJ.PREPRINTS.1900V1.

Rudolf A, Schleifer K-H. (2001). Nucleic Acid Probes and Their Application in Environmental Microbiology. In: *Bergey's Manual of Systematic Bacteriology*. Springer US, pp 67–82.

Sangwan N, Xia F, Gilbert JA. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**: 8.

Sanz JL, Köchling T. (2007). Molecular biology techniques used in wastewater treatment: An overview. *Process Biochem* **42**: 119–133.

Saunders AM, Albertsen M, Vollesen J, Nielsen PH. (2015). The activated sludge ecosystem contains a core community of abundant organisms. *ISME J* 1–10.

Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK. (2016). Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ* **4**: e1869.

Schmid M, Twachtmann U, Klein M, Strous M, Juretschko S, Jetten M, *et al.* (2000). Molecular evidence for genus level diversity of bacteria capable of catalyzing anaerobic ammonium oxidation. *Syst Appl Microbiol* **23**: 93–106.

Schroeder S, Petrovski S, Campbell B, McIlroy S, Seviour R. (2009). Phylogeny and in situ identification of a novel gammaproteobacterium in activated sludge. *FEMS Microbiol Lett* **297**: 157–163.

Seemann T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068–2069.

Sekar R, Fuchs BM, Amann R, Pernthaler J. (2004). Flow sorting of marine bacterioplankton after fluorescence in situ hybridization. *Appl Environ Microbiol* **70**: 6210–6219.

Seviour R, Mino T, Onuki M. (2003). The microbiology of biological phosphorus removal in activated sludge systems. *FEMS Microbiol Rev* **27**: 99–127.

Seviour R, Nielsen PH. (2010). Microbial Ecology of Activated Sludge. *Microb Ecol* **13**: 257–261.

Shapiro E, Biezuner T, Linnarsson S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* **14**: 618–630.

Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* **23**: 111–120.

Shimkets L, Seale TW. (1975). Fruiting body formation and myxospore differentiation and germination in *Myxococcus xanthus* viewed by scanning electron microscopy. *J Bacteriol* **121**: 711–720.

Singapore Department of Statistics. (2016). Statistics Singapore - List of Themes. <http://www.singstat.gov.sg/statistics/latest-data#16>.

Singapore Public Utility Board. (2016). Four national taps of Singapore's water supply. <https://www.pub.gov.sg/watersupply/fournationaltaps>.

Singer E, Bushnell B, Coleman-Derr D, Bowman B, Bowers RM, Levy A, *et al.* (2016). High-resolution phylogenetic microbial community profiling. *ISME J* 1–13.

Snaird J, Amann R, Huber I, Ludwig W, Schleifer KH. (1997). Phylogenetic analysis and in situ identification of bacteria in activated sludge. *Appl Environ Microbiol* **63**: 2884–2896.

Snaird J, Fuchs B, Wallner G, Wagner M, Schleifer K-H, Amann R. (1999). Phylogeny and in situ

- identification of a morphologically conspicuous bacterium, *Candidatus Magnospira bakii*, present at very low frequency in activated sludge. *Environ Microbiol* **1**: 125–135.
- Sorokin DY, Muyzer G, Brinkhoff T, Kuenen JG, Jetten MSM. (1998). Isolation and characterization of a novel facultatively alkaliphilic *Nitrobacter* species, *N. alkalicus* sp. nov. *Arch Microbiol* **170**: 345–352.
- Srinivasan R, Karaoz U, Volegova M, MacKichan J, Kato-Maeda M, Miller S, *et al.* (2015). Use of 16S rRNA gene for identification of a broad range of clinically relevant bacterial pathogens. *PLoS One* **10**: e0117617.
- Stackebrandt E, Goebel BM. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int J Syst Bacteriol* **44**: 846–849.
- Stahl D. A., Amann R. (1991). Development and application of nucleic acid probes in bacterial systematics. In: *Nucleic Acid Techniques in Bacterial Systematics*. John Wiley & Sons, Ltd, pp 205–248.
- Steen HB. (1990). Light scattering measurement in an arc lamp-based flow cytometer. *Cytometry* **11**: 223–230.
- Stevenson BS, Eichorst SA, Wertz JT, Schmidt TM, Breznak JA. (2004). New strategies for cultivation and detection of previously uncultured microbes. *Appl Environ Microbiol* **70**: 4748–4755.
- Stewart PS, Franklin MJ. (2008). Physiological heterogeneity in biofilms. *Nat Rev Microbiol* **6**: 199–210.
- Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, Taylor MW, *et al.* (2006). Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440**: 790–794.
- Szallasi GJJ, Boussau B, Abby SSS, Tannier E, Daubin V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci* **109**: 17513–17518.
- Tan CH. (2013). Quorum sensing signalling and activated sludge microbial communities. Nanyang Technological University.
- Tan CH, Koh KS, Xie C, Tay M, Zhou Y, Williams R, *et al.* (2014). The role of quorum sensing signalling in EPS production and the assembly of a sludge community into aerobic granules. *ISME J* **8**: 1186–97.
- Tchobanoglous G, Burton FL, Stensel HD. (2003). Wastewater Engineering: Treatment and Reuse. McGraw-Hill Education.
- Tennessen K, Andersen E, Clingenpeel S, Rinke C, Lundberg DS, Han J, *et al.* (2015). ProDeGe: a computational protocol for fully automated decontamination of genomes. *ISME J* **10**: 269–272.
- Tettelin H, Riley D, Cattuto C, Medini D. (2008). Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* **11**: 472–477.
- Thiele S, Fuchs B, Amann R. (2010). Identification of microorganisms using the ribosomal RNA approach and fluorescence in situ hybridization. In: *Treatise on water science*. Academic Press, pp 171–189.
- Ting CS, Roca G, King J, Chisholm SW. (2002). Cyanobacterial photosynthesis in the oceans: The origins and significance of divergent light-harvesting strategies. *Trends Microbiol* **10**: 134–142.
- Torsvik V, Goksoyr J, Daae FL. (1990). High diversity in DNA of soil bacteria. *Appl Environ Microbiol* **56**: 782–787.

- Treangen TJ, Salzberg SL. (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat Rev Genet* **13**: 36–46.
- Tringe SG, Rubin EM. (2005). Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* **6**: 805–814.
- Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.
- Větrovský T, Baldrian P. (2013). The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS One* **8**: e57923.
- Van Der Waarde JJ, Geurkink B, Henssen M, Heijnen G. (1998). Detection of filamentous and nitrifying bacteria in activated sludge with 16S rRNA probes. In: Vol. 37. *Water Science and Technology*. pp 475–479.
- Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J, *et al.* (2016). Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiol* **16**: 274.
- Wagner M, Amann R, Lemmer H, Schleifer KH. (1993). Probing activated sludge with oligonucleotides specific for proteobacteria: Inadequacy of culture-dependent methods for describing microbial community structure. *Appl Environ Microbiol* **59**: 1520–1525.
- Wagner M, Horn M, Daims H. (2003). Fluorescence in situ hybridisation for the identification and characterisation of prokaryotes. *Curr Opin Microbiol* **6**: 302–309.
- Wallner G, Amann R, Beisker W. (1993). Optimizing fluorescent in situ hybridization with rRNA-targeted oligonucleotide probes for flow cytometric identification of microorganisms. *Cytometry* **14**: 136–143.
- Wallner G, Erhart R, Amann R. (1995). Flow cytometric analysis of activated sludge with rRNA-targeted probes. *Appl Environ Microbiol* **61**: 1859–1866.
- Wallner G, Fuchs B, Spring S, Beisker W, Amann R. (1997). Flow sorting of microorganisms for molecular analysis. *Appl Environ Microbiol* **63**: 4223–4231.
- Wang Y, Leung HCM, Yiu SM, Chin FYL. (2012). Metacluster 5.0: A two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics* **28**: i356–i362.
- Whitaker RJ, Banfield JF. (2006). Population genomics in natural microbial communities. *Trends Ecol Evol* **21**: 508–516.
- Woese CR. (1987). Bacterial Evolution. *Microbiology* **51**: 221–271.
- Woese CR, Fox GE. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74**: 5088–5090.
- Woese CR, Fox GE, ZABLEN L, UCHIDA T, BONEN L, PECHMAN K, *et al.* (1975). Conservation of primary structure in 16S ribosomal RNA. *Nature* **254**: 83–86.
- Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, Schackwitz W, *et al.* (2010). One bacterial cell, one complete genome. *PLoS One* **5**: e10314.
- Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H, *et al.* (2009). Assembling the marine metagenome, one cell at a time. *PLoS One* **4**. e-pub ahead of print, doi: 10.1371/journal.pone.0005299.
- Wright ES, Yilmaz LS, Noguera DR. (2012). DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Appl Environ Microbiol* **78**: 717–725.
- Wrighton KCC, Thomas BCC, Sharon I, Miller CSS, Castelle CJJ, VerBerkmoes NCC, *et al.* (2012).

- Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. *Science* (80-) **337**: 1661–1665.
- Xia S, Li J, Wang R. (2008). Nitrogen removal performance and microbial community structure dynamics response to carbon nitrogen ratio in a compact suspended carrier biofilm reactor. *Ecol Eng* **32**: 256–262.
- Xie C, Goi CLW, Huson DH, Little PFR, Williams RBH. (2016). RiboTagger: fast and unbiased 16S/18S profiling using whole community shotgun metagenomic or metatranscriptome surveys. *BMC Bioinformatics* **17**: 277–282.
- XRNA. (2017). E. coli 16S rRNA. http://rna.ucsc.edu/rnacenter/xrna/xrna_gallery.html.
- Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, *et al.* (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* **12**: 635–645.
- Ye L, Zhang T. (2013). Bacterial communities in different sections of a municipal wastewater treatment plant revealed by 16S rDNA 454 pyrosequencing. *Appl Microbiol Biotechnol* **97**: 2681–2690.
- Yilmaz LS, Bergsven LI, Noguera DR. (2008). Systematic evaluation of single mismatch stability predictors for fluorescence in situ hybridization. *Environ Microbiol* **10**: 2872–2885.
- Yilmaz LS, Parnerkar S, Noguera DR. (2011). mathFISH, a web tool that uses thermodynamics-based mathematical models for in silico evaluation of oligonucleotide probes for fluorescence in situ hybridization. *Appl Environ Microbiol* **77**: 1118–1122.
- Yilmaz S, Allgaier M, Hugenholtz P. (2010a). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat Publ Gr* **7**: 943–944.
- Yilmaz S, Haroon MF, Rabkin BA, Tyson GW, Hugenholtz P. (2010b). Fixation-free fluorescence in situ hybridization for targeted enrichment of microbial populations. *ISME J* **4**: 1352–1356.
- Youssef N, Sheik CS, Krumholz LR, Najar FZ, Roe BA, Elshahed MS. (2009). Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl Environ Microbiol* **75**: 5227–5236.
- Yu K, Zhang T. (2012). Metagenomic and metatranscriptomic analysis of microbial community structure and gene expression of activated sludge. *PLoS One* **7**: e38183.
- Zanetti-Domingues LC, Tynan CJ, Rolfe DJ, Clarke DT, Martin-Fernandez M. (2013). Hydrophobic Fluorescent Probes Introduce Artifacts into Single Molecule Tracking Experiments Due to Non-Specific Binding. *PLoS One* **8**: e74200.
- Zhang D, Brandwein M, Hsuih T, Li HB. (2001). Ramification amplification: A novel isothermal DNA amplification method. *Mol Diagnosis* **6**: 141–150.
- Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, *et al.* (2006). Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* **24**: 680–686.
- Zhang T, Shao M-F, Ye L. (2012). 454 Pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. *ISME J* **6**: 1137–1147.
- Zhang X, Yue S, Zhong H, Hua W, Chen R, Cao Y, *et al.* (2011). A diverse bacterial community in an anoxic quinoline-degrading bioreactor determined by using pyrosequencing and clone library analysis. *Appl Microbiol Biotechnol* **91**: 425–434.
- Zhang YQ, Li YZ, Wang B, Wu ZH, Zhang CY, Gong X, *et al.* (2005). Characteristics and living patterns of marine myxobacterial isolates. *Appl Environ Microbiol* **71**: 3331–3336.

- Ziglio G, Andreottola G, Barbesti S, Boschetti G, Bruni L, Foladori P, *et al.* (2002). Assessment of activated sludge viability with flow cytometry. *Water Res* **36**: 460–468.
- Zwirgmaier K, Ludwig W, Schleifer K-H. (2003). Improved fluorescence in situ hybridization of individual microbial cells using polynucleotide probes: the network hypothesis. *Syst Appl Microbiol* **26**: 327–337.

Appendix

Appendix A-1: Components of the hybridisation and washing buffers used in FISH experiments

Appendix A-1.1: Components of hybridisation buffer used in FISH experiments

Components of hybridisation buffer	Volume (μ L)	Final concentration in hybridisation buffer
5M NaCl	180	900 mM
1M Tris/HCL	20	20 mM
Formamide	Volume and concentration of formamide is dependent on probe stringency	
10% SDS Page	1	0.01%
MilliQ-water	Top up to a final volume of 1 ml	

Final volume of hybridisation buffer is 1ml. Volume of formamide to be added is dependent on the formamide concentration determined from melting curve analysis of FISH probes.

Appendix A-1.2: Components of washing buffer using in FISH experiments

Components of washing buffer	Volume (mL)	Final concentration in washing buffer
5M NaCl	Volume of NaCl is dependent on [formamide] used in hybridisation experiments	
1M Tris/HCL	1	20 mM
0.5M EDTA*	0.5	5 mM
MilliQ-water	Top up to a final volume of 50 ml	

* EDTA is only added if [formamide] is $\geq 20\%$.

Final volume of washing buffer is 50ml. Volume of formamide to add is dependent on the formamide concentration determined from melting curve analysis of FISH probes.

Appendix A-1.3: Correlation between the concentration of formamide in hybridisation buffer and the concentration of NaCl in washing buffer

[Formamide] in hybridisation buffer	[NaCl] in washing buffer (M)	Volume of NaCl in 50ml of washing buffer (μL)
0	0.900	9000
5	0.636	6300
10	0.450	4500
15	0.318	3180
20	0.225	2150
25	0.159	1490
30	0.112	1020
35	0.080	700
40	0.056	460
45	0.040	300
50	0.028	180
55	0.020	100
60	0.014	40
65	-	-
70	-	-

Appendix A-2: Rank, relative abundance and sequence of RiboTags of the top ten most abundant unclassified taxa residing in UPWRP's floccular sludge community that was profiled with 17bp-RiboTags at the V6 region. This is different from the community analysis in Figure 2-3 that was profiled with 33bp-RiboTags.

Rank	Abundance in UPWRP's floccular sludge community	RiboTag
23 rd	0.77	TGCTTCGCGTCTCCGAA
24 th	0.77	GTGCATGCCTCCTTGCG
38 th	0.48	GTGTACGCTGCTTGTAG
43 rd	0.34	GTGCAAGCTACCCTTGC
49 th	0.32	TCTCACCGGCTCCCGAA
57 th	0.27	TGCTTTGTGTCCTATTA
59 th	0.25	GATACACGCTCTCTTGC
60 th	0.25	TCTCACTCGCTCCTTAC
83 rd	0.19	TGCACGCGACTGGTTGC
88 th	0.18	TGTGCCCCGCCATTGCT

Appendix A-3: Flow cytometric analysis of sorting with a Sy3200 sorter from an axenic culture of R086, hybridised with probes Ribo_Thau1029_17_{Cy5} and EUB388_{Cy3}.

An approximate ~100 000 events were collected for each FACS plot, except for the purity check. Sorting gates were outlined in black and values shown indicate the percentage of gated events over the total number of events. Flow cytometric analysis: (A) Sorting of bacterial cells using their light-scattering properties with a plot of forward versus side scatter; (B) no-probe control was constructed to control for background noise; (C) events exhibiting Cy5 and Cy3 fluorescence signal above the cut-off threshold for the negative control were collected; (D) purity of the sorted sample after an initial round of sorting.

Appendix A-4: RiboTag sequence, taxonomic classification and relative abundance of the top 4 most abundant taxa captured in Sy3200-sorted sample.

RiboTag sequence	Taxonomic classification	Relative abundance (%)
GTGTTACGGTTCCCGAAGGCACTTTTTATCTC	<i>Neisseriaceae</i>	50.8
CCCATAGAATCAAGAAAGAGCTATCAATCTGTC	<i>Malassezia</i>	45.5
GTGAACCAGCCCCAAAAGAGGCGCACCCATCTC	<i>Propionibacterium</i>	2.61
GTGTTCTGGCTCCCGAA GGCACCTCGCCTCTC	<i>Thauera</i>	0.87

Sequence of probe Ribo_Thau1029_17 sequence is highlighted in red.

Appendix A-5: Statistics of the sequencing coverage in pre-sorted samples

Probe-targeted	Samples	Average coverage	Standard deviation
taxon			
<i>Thauera</i>	Replicate 1	1.43	0.74
	Replicate 2	1.35	0.65
	Replicate 3	1.39	0.68
	Replicate 4	1.45	0.77
	Replicate 5	1.39	0.69
	Replicate 6	1.39	0.68
	Replicate 7	1.4	0.71
	Replicate 8	1.31	0.58
	Replicate 9	1.34	0.61
UPWRP_1	Replicate 1	2.33	2.42
	Replicate 2	2.48	2.73
	Replicate 3	2.55	2.89
	Replicate 4	2.48	2.91
	Replicate 5	2.41	2.61
	Replicate 6	2.6	2.97
UPWRP_2	Replicate 1	2.73	3.27
	Replicate 2	2.67	3.19
	Replicate 3	2.54	2.94
	Replicate 4	2.76	3.47
	Replicate 5	2.46	2.85
	Replicate 6	2.75	3.4

Appendix A-6: *In silico* coverage of probe Ribo_Halia1029_17 that was predicted by the 'TestProbe' tool of the SILVA SSU Ref NR 99 database (version 123)

Probe sequence	Accession number of	Taxonomic	Number of
	target organisms	classification	mismatches to probe
TCTCACTCGCTCCCGAA	AB286332	<i>Haliangium</i>	0
	AB286567		
	EU734997

Appendix A-7: Flow cytometric analysis of sorting UPWRP_1 from activated sludge samples hybridised with probes Ribo_Unk1029_17_{Cy5} and EUB388_{A488}.

An approximate ~100 000 events were collected for each FACS plot, except for the purity check. Sorting gates were outlined in black and values shown indicate the percentage of gated events over the total number of events. Flow cytometric analysis: (A) Sorting gate 1: sorting of bacterial cells using their light-scattering properties with a plot of forward versus side scatter; (B) sorting gate 2: filtering out cell aggregates based on side scatter area versus side scatter height; (C) sorting gate 3: filtering out cell aggregates based on forward scatter area versus forward scatter height. Sorting gate 4 was constructed to exclude events exhibiting Cy5 and A488 fluorescence signal in the negative controls: (D) no-probe control; (E) hybridisation with probe NON338_{Cy5} to estimate non-specific binding for Cy5 fluorophore; (F) hybridisation with probe NON338_{A488} to estimate non-specific binding for A488 fluorophore. (G) Events exhibiting Cy5 and A488 fluorescence signal above the cut-off threshold for the negative controls were collected in sorting gate 4; (H) purity of the sorted sample after an initial round of sorting.

Appendix A-8: Statistics for *de novo* assembly of 4 sorted samples of UPWRP_1

Contig measurements	Values
N50 (bp)	32,430
Minimum (bp)	1,000
Maximum (bp)	145,517
GC content (%)	45.18
Count	1,302

Appendix A-9: Statistics for *de novo* assembly of negative controls comprising of sheath fluid, sheath pass and PBS

Contig measurements	Values
N50	2,462
Minimum	1,000
Maximum	6,776
Average	2,249
Count	112

Appendix A-10: Taxonomic classification of *Lewinella* using the NCBI and SILVA taxonomy

	NCBI taxonomy	SILVA taxonomy
No rank	FCB group	-
No rank	Bacteroidetes/Chlorobi group	-
Phylum	Bacteroidetes	
Class	Saprospiria	Sphingobacteriia
Order	Saprospirales	Sphingobacteriales
Family	Lewinellaceae	Saprospiraceae
Genus	<i>Lewinella</i>	

Appendix A-11: Genes that are unique to *Candidatus Shinglingles* but not to its closest neighbour, *Lewinella persica* DSM 23188

Genome	Category	Subcategory	Subsystem	Role
S	Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Polyamine Metabolism	Agmatinase (EC 3.5.3.11)
S	Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Polyamine Metabolism	Putrescine transport ATP- binding protein PotA (TC 3.A.1.11.1)
S	Amino Acids and Derivatives	Lysine, threonine, methionine, and cysteine	Methionine Biosynthesis	Cystathionine gamma- synthase (EC 2.5.1.48)
M	Carbo- hydrates	Central carbohydrate metabolism	Glycolysis and Gluconeogenesis	Glucose-6-phosphate isomerase, archaeal II (EC 5.3.1.9)
M, S	Carbo- hydrates	Central carbohydrate metabolism	Glycolysis and Gluconeogenesis	Fructose-bisphosphate aldolase class I (EC 4.1.2.13)
M, S	Carbo- hydrates	Central carbohydrate metabolism	Glycolysis and Gluconeogenesis	Pyruvate,phosphate dikinase (EC 2.7.9.1)
M, S	Carbo- hydrates	Central carbohydrate metabolism	Pyruvate metabolism I: anaplerotic reactions, PEP	Phosphoenolpyruvate carboxykinase [GTP] (EC 4.1.1.32)

Pyruvate				
M, S	Carbo- hydrates	Central carbohydrate metabolism	metabolism II: acetyl-CoA, acetogenesis from pyruvate	Acylphosphate phosphohydrolase (EC 3.6.1.7), putative
M, S	Clustering- based subsystems	no subcategory	DNA replication cluster 1	Zn-ribbon-containing, possibly RNA-binding protein and truncated derivatives
S	Clustering- based subsystems	no subcategory	Conserved gene cluster associated with Met-tRNA formyltransferase	Protein serine/threonine phosphatase PrpC, regulation of stationary phase
S	Clustering- based subsystems	no subcategory	Conserved gene cluster associated with Met-tRNA formyltransferase	Serine/threonine protein kinase PrkC, regulator of stationary phase
S	Cofactors, Vitamins, Prosthetic Groups, Pigments	Biotin	Biotin biosynthesis	ATPase component BioM of energizing module of biotin ECF transporter
M	Cofactors, Vitamins, Prosthetic Groups, Pigments	Coenzyme A	Coenzyme A Biosynthesis	2-dehydropantoate 2- reductase (EC 1.1.1.169)

M, S	Cofactors, Vitamins, Prosthetic Groups, Pigments	Biotin	Biotin biosynthesis	Substrate-specific component BioY of biotin ECF transporter
M, S	Cofactors, Vitamins, Prosthetic Groups, Pigments	Folate and pterines	Molybdenum cofactor biosynthesis	Molybdenum cofactor biosynthesis protein MoaB
M	Cofactors, Vitamins, Prosthetic Groups, Pigments	Folate and pterines	Pterin synthesis related cluster	Sepiapterin reductase (EC 1.1.1.153)
S	DNA Metabolism	CRISPs	CRISPRs	CRISPR repeat RNA endoribonuclease Cas6
S	DNA Metabolism	CRISPs	CRISPRs	CRISPR-associated RecB family exonuclease Cas4a
S	DNA Metabolism	CRISPs	CRISPRs	CRISPR-associated helicase Cas3
S	DNA Metabolism	CRISPs	CRISPRs	CRISPR-associated protein Cas2
M	DNA Metabolism	DNA repair	DNA repair, bacterial	Alkylated DNA repair protein AlkB

M	DNA Metabolism	DNA repair	DNA repair, bacterial	Methyl-directed repair DNA adenine methylase (EC 2.1.1.72)
M, S	DNA Metabolism	CRISPs	CRISPRs	CRISPR-associated protein Cas1
M, S	DNA Metabolism	DNA repair	DNA Repair Base Excision	Single-stranded-DNA-specific exonuclease RecJ (EC 3.1.-.-)
M, S	DNA Metabolism	DNA repair	DNA repair, bacterial UvrD and related helicases	ATP-dependent DNA helicase UvrD/PcrA, proteobacterial paralog
M	Fatty Acids, Lipids, and Isoprenoids	Fatty acids	Fatty Acid Biosynthesis FASII	Enoyl-[acyl-carrier-protein] reductase [FMN] (EC 1.3.1.9)
S	Fatty Acids, Lipids, and Isoprenoids	Phospholipids	Glycerolipid and Glycerophospholip id Metabolism in Bacteria	CDP-diacylglycerol--glycerol-3- phosphate 3- phosphatidyltransferase (EC 2.7.8.5)
S	Fatty Acids, Lipids, and Isoprenoids	Phospholipids	Glycerolipid and Glycerophospholip id Metabolism in Bacteria	Phosphatidylglycerophosphat ase A (EC 3.1.3.27)
M, S	Fatty Acids, Lipids, and Isoprenoids	Fatty acids	Fatty Acid Biosynthesis FASII	3-oxoacyl-[acyl-carrier- protein] synthase, KASIII (EC 2.3.1.41)

M, S	Fatty Acids, Lipids, and Isoprenoids	Isoprenoids	Isoprenoid Biosynthesis: Interconversions	Isopentenyl-diphosphate delta-isomerase (EC 5.3.3.2)
M, S	Fatty Acids, Lipids, and Isoprenoids	Phospholipids	Cardiolipin synthesis	Cardiolipin synthetase (EC 2.7.8.-)
M	Membrane Transport	Cation transporters	Magnesium transport	Mg(2+) transport ATPase protein C
M	Membrane Transport	Cation transporters	Magnesium transport	Mg(2+) transport ATPase, P- type (EC 3.6.3.2)
			Broadly	
M, S	Misc- ellaneous	no subcategory	distributed proteins not in subsystems	UPF0028 protein YchK
M, S	Nucleosides and Nucleotides	Detoxification	Nudix proteins (nucleoside triphosphate hydrolases)	5-methyl-dCTP pyrophosphohydrolase (EC 3.6.1.-)
	Phages, Prophages, Trans- posable elements, Plasmids	Pathogenicity islands	Listeria Pathogenicity Island LIPI-1 extended	Thiol-activated cytolysin

S	Protein Metabolism	Protein processing and modification	Ribosomal protein S5p acylation	Ribosomal-protein-S5p- alanine acetyltransferase
M	RNA Metabolism	RNA processing and modification	RNA processing orphans	RNA ligase
M	Regulation and Cell signaling	Programmed Cell Death and Toxin- antitoxin Systems	Phd-Doc, YdcE- YdcD toxin- antitoxin (programmed cell death) systems	Death on curing protein, Doc toxin
M, S	RNA Metabolism	Transcription	Transcription initiation, bacterial sigma factors	RNA polymerase sigma factor SigW
M	Respiration	Electron accepting reactions	Biogenesis of cbb3-type cytochrome c oxidases	Type cbb3 cytochrome oxidase biogenesis protein CcoS, involved in heme b insertion
S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH ubiquinone oxidoreductase chain A (EC 1.6.5.3)
S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain B (EC 1.6.5.3)

S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain C (EC 1.6.5.3)
S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain D (EC 1.6.5.3)
S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain E (EC 1.6.5.3)
S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain F (EC 1.6.5.3)
S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain G (EC 1.6.5.3)
S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain H (EC 1.6.5.3)
S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain I (EC 1.6.5.3)

S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain J (EC 1.6.5.3)
S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain K (EC 1.6.5.3)
S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain L (EC 1.6.5.3)
S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain M (EC 1.6.5.3)
S	Respiration	Electron donating reactions	NADH ubiquinone oxidoreductase	NADH-ubiquinone oxidoreductase chain N (EC 1.6.5.3)
S	Respiration	no subcategory	Biogenesis of c-type cytochromes	Thiol:disulfide oxidoreductase related to ResA
M, S	Respiration	ATP synthases	V-Type ATP synthase	V-type ATP synthase subunit A (EC 3.6.3.14)
M, S	Respiration	ATP synthases	V-Type ATP synthase	V-type ATP synthase subunit B (EC 3.6.3.14)
M, S	Respiration	ATP synthases	V-Type ATP synthase	V-type ATP synthase subunit C (EC 3.6.3.14)

M, S	Respiration	ATP synthases	V-Type ATP synthase	V-type ATP synthase subunit D (EC 3.6.3.14)
M, S	Respiration	ATP synthases	V-Type ATP synthase	V-type ATP synthase subunit E (EC 3.6.3.14)
M, S	Respiration	ATP synthases	V-Type ATP synthase	V-type ATP synthase subunit I (EC 3.6.3.14)
M, S	Respiration	ATP synthases	V-Type ATP synthase	V-type ATP synthase subunit K (EC 3.6.3.14)
S	Virulence, Disease and Defence	Resistance to antibiotics and toxic compounds	Arsenic resistance	Arsenical pump-driving ATPase (EC 3.6.3.16)
S	Virulence, Disease and Defence	Resistance to antibiotics and toxic compounds	Arsenic resistance	Arsenical-resistance protein ACR3
S	Virulence, Disease and Defence	Resistance to antibiotics and toxic compounds	Copper homeostasis: copper tolerance	Periplasmic divalent cation tolerance protein CutA

M: *Candidatus Shinglingles singa*

S: *Candidatus Shinglingles merlion*

Appendix A-12: Flow cytometric analysis of sorting UPWRP_2 from activated sludge samples, hybridised with probes Ribo_Halia1029_17_{Cy5} and EUB388_{A488}.

An approximate ~100 000 events were collected for each FACS plot, except for the purity check. Sorting gates were outlined in black and values shown indicate the percentage of gated events over the total number of events. Flow cytometric analysis: (A) Sorting gate 1: sorting of bacterial cells using their light-scattering properties with a plot of forward versus side scatter; (B) sorting gate 2: filtering out cell aggregates based on side scatter area versus side scatter height; (C) sorting gate 3: filtering out cell aggregates based on forward scatter area versus forward scatter height. Sorting gate 4 was constructed to exclude events exhibiting Cy5 and A488 fluorescence signal in the negative controls: (D) no-probe control; (E) hybridisation with probe NON338_{Cy5} to estimate non-specific binding for Cy5 fluorophore; (F) hybridisation with probe NON338_{A488} to estimate non-specific binding for A488 fluorophore. (G) Events exhibiting Cy5 and A488 fluorescence signal above the cut-off threshold for the negative controls were collected in sorting gate 4; (H) purity of the sorted sample after an initial round of sorting.

Appendix A-13: Taxonomic classification of *Haliangium* using the NCBI and SILVA taxonomy

	NCBI taxonomy	SILVA taxonomy
Phylum	Proteobacteria	
Subphylum	Delta/epsilon subdivision	-
Class	Deltaproteobacteria	
Order	Myxococcales	
Suborder	Nannocystineae	-
Family	Kofleriaceae	Haliangiaceae
Genus	Haliangium	

Appendix A-14: Genes that are unique to Haliangium cluster0 but not to its closest neighbour, Haliangium ochraceum DSM 14365

Genome	Category	Subcategory	Subsystem	Role
HC	Clustering-based subsystems	No subcategory	Sigma-B stress response cluster 1	Anti-sigma B factor RsbT
HC	Clustering-based subsystems	No subcategory	Sigma-B stress response cluster 1	Phosphoserine phosphatase RsbX (EC 3.1.3.3)
HC	Clustering-based subsystems	No subcategory	Sigma-B stress response cluster 1	RsbR, positive regulator of sigma-B
HC	Clustering-based subsystems	No subcategory	Sigma-B stress response cluster 1	RsbS, negative regulator of sigma-B
HC	Cofactors, Vitamins, Prosthetic Groups, Pigments	Folate and pterines	Folate Biosynthesis	Thymidylate synthase thyX (EC 2.1.1.-)
HC	DNA Metabolism	DNA repair	Uracil-DNA glycosylase	Uracil-DNA glycosylase, family 5
HC	Nucleosides and Nucleotides	Detoxification	Nudix proteins (nucleoside triphosphate hydrolases)	NADH pyrophosphatase (EC 3.6.1.22)

HC	Protein	Protein	tRNA	Glycyl-tRNA
	Metabolism	biosynthesis	aminoacylation, Gly	synthetase (EC 6.1.1.14)
HC	Respiration	no subcategory	Biogenesis of c-type cytochromes	Cytochrome c heme lyase subunit CcmH
HC	Respiration	no subcategory	Biogenesis of c-type cytochromes	Periplasmic thiol:disulfide interchange protein DsbA
HC	Virulence, Disease and Defence	Invasion and intracellular resistance	Listeria surface proteins: Internalin-like proteins	Internalin-like protein (LPXTG motif) Lmo0333 homolog
HC	Virulence, Disease and Defence	Resistance to antibiotics and toxic compounds	Resistance to chromium compounds	Chromate transport protein ChrA

HC: *Haliangium clusterio*

Appendix A-15: Box-and-whisker graph depicting the distribution of genome completeness from 201 single-cell amplified genomes (Rinke *et al.*, 2013) using the concept of essential-single copy genes. Minimum, 25th percentile, median, 75th percentile and maximum genome completeness corresponded to a value of 4%, 20%, 36%, 57.50% and 100% respectively.

Appendix A-16: Box-and-whisker graph depicting the distribution of the number of contigs generated per genome from the 201 single-cell amplified genomes (Rinke *et al.*, 2013). Minimum, 25th percentile, median, 75th percentile and maximum number of contigs corresponded to a value of 7, 30, 48, 84 and 190 respectively.

Appendix A-17: A visualisation of the differential coverage binning plot of contigs from UPWRP_1. Differential coverage binning performed using sorted samples with lower number of collected events: (A) 5 events and (B) 10 events in the X-axis resulted in poor clustering of contigs. Each scaffold is represented by a circle that is scaled by their length, and coloured based on the taxonomic classification of essential single-copy genes. Only scaffolds of >3000 bp were displayed.

Appendix A-18: Comparison of genome coverage across multiple samples that contained different number of events. Genome coverage was quantified by measuring the total number of unique essential single-copy genes. Sorted samples containing: (A) 5; (B) 10 and (C) 1000 collected events. Number of events is demarcated on the left side of each figure. Identity of the essential single-copy genes is listed on the X-axis of each figure. Number of copies of essential single-copy genes is demarcated with a colour-code on the right side of each figure.