AQMM: Enabling Absolute Quantification of Metagenome and Metatranscriptome

Authors: Xiao-Tao Jiang, Ke Yu, Li-Guan Li, Xiaole, Yin, Andong, Li, Tong Zhang*

Affiliations: Environmental Biotechnology Laboratory, Department of Civil Engineering, University of Hong Kong

* Corresponding author Email: zhangt@hku.hk

November 12, 2017

Abstract

Metatranscriptome has become increasingly important along with the application of next generation sequencing in the studies of microbial functional gene activity in environmental samples. However, the quantification of target active gene is hindered by the current relative quantification methods, especially when tracking the sharp environmental change. Great needs are here for an easy-to-perform method to obtain the absolute quantification. By borrowing information from the parallel metagenome, an absolute quantification method for both metagenomic and metatranscriptomic data to per gene/cell/volume/gram level was developed. The effectiveness of AQMM was validated by simulated experiments and was demonstrated with a real experimental design of comparing activated sludge with and without foaming. Our method provides a novel bioinformatic approach to fast and accurately conduct absolute quantification of metagenome and metatranscriptome in environmental samples.

The AQMM can be accessed from https://github.com/biofuture/aqmm.

Keywords: metagenome, metatranscriptome, absolute quantification, differential expression genes

Background

Shotgun metatranscriptomics is a powerful tool in identifying the overall expression of microorganisms in an environment (Alexander et al. 2015, Gifford et al. 2011, Shi et al. 2009, Turner et al. 2013, Yu and Zhang 2012), shedding light on discovering how microbes respond to environmental changes or diseases status (Jorth et al. 2014, Mason et al. 2012) and capturing gene expression patterns for functionally important bacteria in engineering systems (Oyserman et al. 2015, Stark et al. 2014). For these applications, accurate quantification is required to detect the true variations or differential expression genes (DEGs).

Traditionally, the abundance of a transcript in RNA-sequencing (RNA-seq) is thought to be influenced by the library size and inherent dependence on the expression levels of other transcripts as described in a comprehensive review (Rapaport et al. 2013). Following this idea, transcripts in RNA-seq was generally quantified by within-sample normalization. One of the most common quantification methods was RPKM (Mortazavi et al. 2008) (reads per kilobase of exon model per million mapped reads) which considered factors of both the length of gene and library size. Another improved within-sample normalization method was TPM (transcript per million) (Wagner et al. 2012) which only considered the transcript rather than the whole library size and respected the invariance of relative molar RNA concentration (rmc). The TPM was thought to be better fitted in sample comparison due to its unit-free characteristics. The FPKM (substitute the reads with fragments in RPKM) was an adaption of RPKM to pair-end reads. These above methods are all relative quantification (RQ) and suffer from the ‘composition effects’ (the increase of one transcript will decrease other unrelated transcript). To relieve this problem, Robinson and Oshlack proposed a new normalization method “TMM” (trimmed mean of

50 M-values) to detect the DEGs under the hypothesis that most of the genes are not
51 differentially expressed (Robinson and Oshlack 2010), which has been integrated into
52 popular DEGs detection R software edgeR (Robinson et al. 2010). The scaling factor
53 in edgeR for normalization is the TMM value. Another method was to compute the
54 median of the ratio as the scaling factor and it could be conducted by R software
55 DESeq/DESeq2 (Love et al. 2014). It is also based on the assumption that most genes
56 are not DEGs and this method then calculates the scaling factor (median of ratios)
57 associated with this sample to perform further normalization. In the two software, the
58 negative binomial distribution was applied to adjust the distribution of transcript
59 between different conditions to relieve the dispersion effects of deviation from
60 standard passion distribution (Rapaport et al. 2013). Although with these efforts in
61 optimizing the normalization process, these indices were all still RQ based and the
62 relationship could be distorted while performing comparative analysis across samples,
63 especially when borrowing these methods from traditional Eukaryote RNA-seq to
64 current Prokaryote metatranscriptome studies (Conesa et al. 2016). One feasible way
65 to solve the problem was to get the absolute quantification (AQ) of expression level
66 for each transcript. For example, the qRT-PCR has long been applied in RNA-seq or
67 microarray data for AQ (Becker-André and Hahlbrock 1989, Whelan et al. 2003). In
68 addition, there were methods by spiking in exterior/alien RNA in microarray to get the
69 per cell absolute quantification (Kanno et al. 2006) and internal standard approach to
70 estimate per liter expression in marine metatranscriptome (Gifford et al. 2011).
71 However, the experiment to perform spiking internal standard was difficult due to its
72 skill-demanding nature and for metatranscriptome data, factors like the time to add
73 spike-in material, the type and the amount of alien RNA required still needed to be
74 elaborately designed. Hence, it was not as popular as those RQ methods. The

75 quantification methods in the newly developed analyzing pipelines for
76 metatranscriptome like IMP (Narayanasamy et al. 2016), MetaTrans (Martinez et al.
77 2016), COMAN (Ni et al. 2016) and SAMSA (Westreich et al. 2016) were still all
78 based on RQ methods; this would result in accelerated spreading of the inaccurate
79 quantification in many studies.

80 To solve the problem of RQ and get an accurate quantification without performing
81 spike-in experiment, an AQ bioinformatics software package AQMM was developed
82 by combining metagenome and metatranscriptome data to achieve the goal of
83 accurate and comparable quantification. In this study, we firstly introduced the
84 AQMM algorithm flow, and then compared and validated it with RQ methods by
85 simulated metagenome and metatranscriptome data. Moreover, we further applied this
86 algorithm to a real combination of metagenome and metatranscriptome dataset in
87 quantifying genes and transcripts of resistome in six foaming activated sludge (FAS)
88 and non-foaming activated sludge (NFAS) samples.

89 Results

90 Overall view of AQMM algorithm

91 The AQMM **(Fig. 1)** was designed to perform AQ of parallel metagenome and
92 metatranscriptome dataset no matter whether spike-in experiment/internal standard
93 was initially added or not. The major aims were to obtain the AQ of
94 genes/transcripts/taxa in samples and to accurately detect DEGs in metatranscriptome
95 data. The assumptions under the algorithm include: 1) with the known extraction ratio
96 of DNA for a DNA extraction Kit for a type of sample, the total weight of DNA per
97 volume of the sample could be calculated. The weight of the sequenced library of
98 DNA could be estimated with the molecular weight and bases numbers of A, T, C and

99 G in the sample. Then, the ratio of sequenced DNA to total weight of DNA per
100 volume of the sample could be calculated. In addition, by utilizing the universal
101 single-copy phylogenetic marker genes (USCMGs)**,** the number of cells for a
102 metagenome library could be estimated accurately (Nayfach and Pollard 2015). With
103 the above information, cells per volume could be calculated for a metagenome data; 2)
104 Using the same volume of the sample contains the same number of cells for DNA and
105 RNA extraction, the cell number per volume to extract RNA was the same as the
106 parallel DNA sample; 3) With the known ratio of RNA extraction and the rRNA ratio
107 of total RNA, by a similar process, the sequenced RNA weight ratio could be
108 calculated, and then the equivalent cell numbers in a metatranscriptome could be
109 deduced accordingly. With the cell numbers included in the metagenome and
110 metatranscriptome data, the abundance of gene/transcript could be normalized to per
111 cell level. Moreover, as the number of cells per volume is available, per cell
112 quantification could be easily transformed into per volume quantification.

113 **Comparing and validating AQMM with RQ methods using simulated**
114 **metatranscriptome demo**

115 To reveal the problem of RQ methods like RPKM, edgeR and DESeq2 and to assess
116 the effectiveness of AQMM, simulated metatranscriptomic datasets comprised of
117 known community structure and expression levels were generated **(Fig. 2; Details in**
118 **methods).** The simulated data was with known ground truth absolute expression for
119 each gene. For simplicity, to focus on the quantification of metatranscriptome in
120 identifying DEGs, we assume the DNA content are not changed like what happens in
121 a reactor with a stable biomass concentration, however the gene expression under
122 condition A and B are significantly changed with fold of 2 or 16 in part of the bacteria
123 like what happens under sharp environmental change. In order to focus only on the of

124 influence normalization methods, in generation of the simulated metatranscriptome,
125 the base qualities of were all set with 50 and to eliminate the influence of mapping
126 process, the mapping criteria of bowtie2 was set to exactly match without gap and
127 mismatch allowed (bowtie2 parameters, -N 0 -L 31, --rdg 100,150 --rfg 100,150
128 --gbar 100,150). The result of DEGs detection was in **Table 1**. We can observe that
129 compared with ground truth, the RQ methods detect quite a large portion of false
130 positive higher gene expression under condition A. On the contrary, the AQMM
131 method which aims to obtain the AQ has limited errors detection even with a given
132 variance in RNA extraction efficiency **(Table 1)**. Noticeably, in real combination of
133 metagenome and metatranscriptome, the metagenome could also be totally different,
134 and in this case, the AQMM is still applicable.

135 **Case study: AQ of activate resistome in FAS and NFAS**

136 The AQMM was applied in the six metagenome and metatranscriptome dataset of
137 FAS and NFAS, the AQ of the sequenced cells generated by the pipeline were shown
138 in **Table 2**. In detail, the metagenome contained 8 to 11.8 GBs data and
139 metatranscriptome with a depth between 13 and 16 GBs for each sample. The "per
140 cell/volume" quantifying values were the fundamental of normalizing to cells or
141 volume in order to perform comparison among different studies. The cell number per
142 milliliter in literature was at 3.3E+09 using flow Cytometer to quantify(Foladori et al.
143 2010) and was from 2.1E+09 to 5.5E+09 using CFU and flow Cytometer (Manti et al.
144 2008) level for AS which was a bit lower than the obtained number in this study at the
145 magnitude of E+10 cells per milliliter**.** Overall number of mRNA molecules per cell
146 are 387.98 ± 102.86 and 235.21 ± 30.59 averagely for FAS and NFAS, respectively
147 **(Table S1)**, which is consistent with previous observation of coastal bacterioplankton
148 by 142-238 mRNA molecules per Cell (Gifford et al. 2011, Moran et al. 2013).

149

150 As WWTPs become the hot-spot of antibiotic resistant genes (ARGs) to the receiving
151 environment. Hence, the expressions of ARGs in the AS were in great concerns and
152 further profiled. Overall, the abundance of ARGs per cell in FAS and NFAS were
153 0.0517 ± 0.0034 and 0.0483 ± 0.0041; and the transcript of ARGs per cell were
154 0.0140 ± 0.0039 and 0.0059 ± 0.0009, respectively **(Table S2 & S3)**. The overall
155 transcription of ARGs was significantly higher in FAS compared with NFAS. At DNA
156 level, only tetracycline resistance gene was higher in FAS and beta-lactam was higher
157 in NFAS, other types were not significantly different. However, at transcript level, all
158 the types were all significantly higher in FAS. Among the nine transcribed ARGs
159 types, beta-lactam and sulfonamide resistance genes were the most abundant
160 expressed ARG types in both FAS and NFAS. Per volume ARGs abundance and
161 expression at type level were shown in **Fig. 3**. The overall ARGs abundance per
162 milliliter AS in FAS and NFAS were 2.51E+09 ± 2.44E+08 and 2.66E+09 ±
163 5.63E+08; and the transcript of ARGs per milliliter were 9.83E+09 ± 3.82E+08 and
164 4.49E+09 ± 5.10E+08, respectively. With the AQ results, the transcripts per copy gene
165 (TPCG), which represents of the transcribe rate could be further derived. The
166 unclassified, quinolone, multidrug and beta-lactam were more active in FAS
167 compared with NFAS in terms of TPCG, **(Table S4)**. For the detected ARGs, the host
168 taxonomy was assigned by LCA algorithms using all the genes annotation in the same
169 Contig. Thirteen orders were detected to carry ARGs and eleven of them were
170 transcribed (**Fig. 3)**. The most ARGs transcribed order was Enterobacteriales. The
171 active ARGs in bacteria enclosed in foams of FAS posed potential threats for the
172 public as ARGs carrying bacteria could spread into the air from the foams bubbles.

173

174 The co-expression of ARGs and MRGs was also studied to check whether there were
175 co-expression effects at the RNA level. Using this dataset, we observed co-expression
176 within ARGs, within MRGs, and between ARGs and MRGs (**Fig. 4)**. Numerous types
177 of MRGs were detected in the metagenome and metatranscriptome. The most
178 abundant MRG was Cu resistant genes and for the ARGs, beta-lactam, tetracycline
179 and aminoglycoside were the most expressed types. The highest number of
180 co-expression within MRGs was Cr and Fe; while within ARGs was beta-lactam and
181 tetracycline. The most MRG and ARGs co-expression was Cr, which co-expression
182 with nine types of ARGs. This was the first transcript level evidence of the
183 co-expression of ARGs and MRGs in AS.

184 **Discussion**

185 Metatranscriptome enabled the study of whole metabolic pathways expression of the
186 system and many studies had already taken this advantage for different environments,
187 such as in marine (Mason et al. 2012), rhizosphere of the plant (Turner et al. 2013),
188 human oral disease (Jorth et al. 2014). Each study has specific method to integrate the
189 metagenome and metatranscriptome information to understand the microbes and their
190 activities in the system. The quantification of metatranscriptome was generally RQ
191 based methods. The RQ methods are problematic as they may not be able to reflect
192 the actual expression level of a population in the whole community. Due to the
193 relative characteristics, the RQ methods are always suffer from the so-called
194 composition effects, which indicates that the upgrade of one gene should definitely
195 make other genes downgrade. Additionally, the RQ methods are just a relative portion
196 rather than a value with biological implications. On the contrary, the AQ could be
197 more biological meaningful at per cell/volume unit. Hence, it was necessary to
198 conduct AQ to compare different samples. In this study, we proposed an AQ method

199 and developed a set of algorithms to conveniently calculate the absolute number of
200 sequenced cells for each RNA library by borrowing cell numbers from a
201 corresponding data set of DNA library of the same sample.

202 Noticeably, there were several hypotheses for the application of the proposed method.
203 Firstly, the sample used to extract DNA and RNA should contain the same cell
204 numbers per volume which could be easily met with sufficient mixing of samples.
205 Secondly, the DNA and RNA extraction efficiency should be estimated, as well as the
206 rRNA ratio in total RNA. This was likely difficult to achieve. However, for an
207 environmental sample, generally literature based data could be used for the extraction
208 kit, for example, to FastDNA SPIN Kit for Soil, the extract efficiency was estimated
209 as 28.4% (Mumy and Findlay 2004). Most importantly, as the parallel samples were
210 extracted under the same condition, the difference between samples was minimized
211 **(DNA extraction data, unpublished)**. This AQMM method is capable of performing
212 absolute quantification of both metagenome and metatranscriptome without the
213 requirement to do complex spike-in experiments. Importantly, AQMM avoids the RQ
214 problems of composition effects and able to detect accurate DEGs. Hence, the
215 proposed AQMM is a method in between experimental spike-in based AQ methods
216 and those improved RQ methods of TMM based edgeR.

217 With AQ, a number of indices with various biological meaning were proposed in this
218 study (Methods), for example, the transcript per copy gene (TPCG) index is a
219 reflection of the transcribe rate of the gene, which could never be delivered by RQ
220 methods. It was demonstrated with simulating RNA-seq that the organism abundance
221 (community structure) was important at normalizing metatranscritptome data in
222 identifying DEGs (Klingenberg and Meinicke 2017). The gene per cell (GPC) and
223 transcript per cell (TPC) in AQMM are global level normalization indices and the

224 scaling factor is the total number of cells in the DNA or RNA library. This global
225 scaling factor could be easily transformed into taxa specific scaling factors with the
226 relative quantification of different taxa with indices of transcript of taxon A per cell
227 (TTPC). Hence, the normalization in AQMM is well fit for the factor of microbial
228 abundance in metatranscriptome data.

229 AS is important biological wastewater treatment process and this system is considered
230 as a hot spot for ARG dissemination into the receiving water. The foaming of AS
231 would result in spreading of foams with AS bacteria into the surrounding environment.
232 Understanding the active resistome and the host bacteria in foaming AS enables
233 engineers understanding the risk of sludge foaming incurred to the surrounding
234 environment. We observed a wide profile of active ARG types in the FAS, the
235 identification of opportunity pathogen bacteria *Pseudomonas* carrying active ARGs
236 alerts us the risk of spreading ARGs-carrying bacteria. Additionally, per cell mRNA
237 molecules is an important indication of the activity of the cell, generally natural
238 bacterial communities was observed to hold a lower inventory of transcripts (Moran et
239 al. 2013); and the absolute quantification obtained with AQMM was well-fitted with
240 previous observation.

241 Conclusions

242 In this study, we filled the gap of lacking a bioinformatic algorithm to perform AQ of
243 metatranscriptomic data. The developed AQMM was demonstrated to gain enhanced
244 performance at identifying DEGs compared with those RQ methods benchmarked
245 with simulated metagenomic and meatranscriptomic data. Additionally, with the
246 AQMM, the active resistome in foaming and normal activated sludge were quantified
247 to per cell/volume level and even down to the transcription per copy gene. The active

248 ARG host were quantified and the co-expression of MRGs and ARGs was revealed

249 for the first time in AS.

250 Materials and methods

251 Absolute quantification of gene abundance and transcript expression

252 We developed a package of scripts AQMM (absolute quantification of metagenome

253 /metatranscriptome) to perform comparative analysis.

254 The formula for cells per mL:

255 $$\boldsymbol{C} = N_c \,/\, \frac{L_{size} * 10^9 * \frac{(R_A * 313.2 + R_T * 304.2 + R_C * 289.18 + R_G * 329.21)}{6.022 * 10^{23}}}{X/\alpha} \qquad (1)$$

256 C is value of cell numbers per mL AS

257 N_c is the estimated cell numbers for the sequenced DNA library with USCMGs

258 L_{size} is the sequencing depth

259 R_A, R_T, R_C and R_G are ratios of A, T, C and G

260 X is the overall extracted weight (ng) of DNA for 1 mL AS

261 α is DNA extraction efficiency, for FAST DNA Kit for Soil, α is estimated as 28.2%

262 (Mumy and Findlay 2004).

263 The sequenced cells for RNA sequencing, for a RNA-seq with library size of L_{size}

264 after removing all ribosomal RNA, the equivalent sequenced cells for this sample is

265 $$\boldsymbol{E_c} = C * \frac{L_{size} * 10^9 * \frac{R_A * 329.2 + R_U * 306.2 + R_C * 305.2 + R_G * 345.2}{6.022 * 10^{23}}}{Y * \gamma/\beta} \qquad (2)$$

266 E_c is the estimated number of cells sequenced for this RNA library

267 C is value of cell numbers per mL AS

268 L_{size} is the sequencing depth

269 R_A, R_U, R_C and R_G are ratios of A, U, C and G, the value they multiplied are molecular

270 weight

271 Y is the overall extracted weight (ng) of RNA for 1 mL AS

272 β is RNA extraction efficiency, the estimated β is about 7.5% as used in this study.
273 This value was deduced from AS empirical data of proportion of RNA biomass by
274 engineering perspective and the extracted RNA biomass.
275 γ is non-ribosomal RNA ratio, for AS the estimated γ is about 0.03.

276 Based on the two AQ numbers of cells for each sample, the gene or transcript
277 abundance matrix could be further normalized into the following indices.
278 **GPC** (Gene per Cell): an indication of the overall abundance of the gene in system.
279 $$\mathrm{GPC} = \frac{N_{read} * L_{read}/L_{gene}}{N_c} \quad (3)$$
280 **TPC** (Transcript per Cell): an indication of overall activity of the gene in system.
281 $$\mathrm{TPC} = \frac{N_{read} * L_{read}/L_{gene}}{E_c} \quad (4)$$
282 **TPCG** (Transcript per copy gene): an indication of the absolute activity of one copy
283 gene in the system, equivalent to transcribe rate for each gene.
284 $$\mathrm{TPCG} = {}^{\mathrm{TPC}}/_{GPC} \quad (5)$$
285 **GTPC** (Gene of taxon A per Cell): an indication of the overall abundance of the taxon
286 in system averagely.
287 $$\mathrm{GTPC} = \sum_{i=1}^{n} GPC_i \quad (6)$$
288 **TTPC** (Transcript of taxon A per Cell): an indication of overall activity of the taxon
289 in system averagely.
290 $$\mathrm{TTPC} = \sum_{i=1}^{n} TPC_i \quad (7)$$
291 **ATCT** (Averagely transcript per copy gene of taxon A): indication of the averagely
292 absolute activity per copy expressed gene in taxon A
293 $$\mathrm{ATCT} = \frac{1}{n}\sum_{i=1}^{n} TPCG_i \quad (8)$$
294 N_c is the estimated cell numbers for the sequenced DNA library,
295 N_{read} is the number of reads or transcript mapping to the target gene

296 L_{read} is the length of reads

297 L_{gene} is the length of the target gene

298 n is the number of genes affiliated to taxa A.

299 When the number of cells per mL was obtained, using the GPC, genes per mL could

300 be calculated.

301 **Simulating metatranscriptome data**

302 To validate our method and comparing with those RQ methods in identifying the

303 DEGs, simulated data was generated by workflow illustrated in **Fig. 2**. For simplicity,

304 the DNA was set unchanged to mimic the activated sludge community composition

305 with 16 strains from different phylogeny. The metatranscriptome data sets were

306 generated for two conditions A and B, each with three biological duplications; for the

307 condition A and B, there were part of the strains with folds of significantly changed

308 expression (**Table S5**). To only focus on the quantification method, all the system

309 errors caused by other factors like base qualities, cDNA synthesis, assembly, mapping

310 parameters were not considered.

311 **Sampling**

312 AS samples were collected in Shatin wastewater treatment plant at three locations

313 along the flow direction while serious foaming happened at 2016-04-08 and nearly no

314 foaming happened at 2016-04-25. Samples were collected on site by storing in liquid

315 nitrogen immediately and then transported to the laboratory for RNA extraction. The

316 DNA samples were mixed with 1:1 100% ethanol and AS and then stored at -20 °C

317 fridge. Totally six samples were collected for both DNA and RNA samples alongside

318 the segment aeration tank in three locations as depicted in **Fig. 5**.

319 **Whole DNA, total RNA extraction, removal of ribosomal RNA, cDNA synthesis**
320 **and next generation sequencing**

321 FAST DNA Kit was used to extract total DNA from 1 mL mixed AS samples. RNeasy
322 Mini was used to extract the total RNA from 0.5 mL AS stored in liquid nitrogen. The
323 extracted RNA was then processed by DNase I to eliminate the DNA in the RNA
324 samples. Then both Illumina Ribo-Zero rRNA removal KIT (Bacteria) and Ribo-Zero
325 rRNA removal KIT (Human/Mouse/Rat) was applied for each sample to remove
326 rRNA from Prokaryote and Eukaryote respectively in order to get the total clean
327 non-ribosomal RNA. Generally, metatranscriptome rRNA depletion was only used the
328 Ribo-Zero for Bacteria, in this study, the addition of Eukaryote rRNA removal was
329 due to a fact that by only using the Ribo-Zero Bacteria rRNA removal Kit for AS,
330 there was still over half of RNA were rRNA from Eukaryote (our previous experiment,
331 data unpublished). To get more non-rRNA, the Ribo-Zero rRNA Kit to remove
332 Eukaryote was also used. RNA then was fragmented into 170 bps library and was
333 reverse-transcribed to construct cDNA library for sequencing. The quality of DNA
334 and RNA were assessed with Agilent 2100 Bioanalyzer (Agilent Technologies, Palo
335 Alto, CA, USA). All the samples was sent to sequence, considering the complexity of
336 AS and the aims of this study to detect the expression of low abundance gene, we
337 gave each sample a very deep sequencing depth which doubled the sequencing depth
338 in previous studies. All the samples were sequenced with Hiseq 4000 in
339 BGI-ShenZhen. DNA samples with PE-150 with library size of 300 bps. And RNA
340 with PE101 of library size 170 bps.

341 **Bioinformatics analysis**

342 Quality filtering was firstly performed on DNA and RNA reads to keep only high
343 quality reads using trimmomatic v1.04 (Bolger et al. 2014). DNA datasets were

344 pooled together and assembled by CLC Genomics Workbench 6.5.3 (CLC Bio,
345 Aarhus, Denmark, https://www.qiagenbioinformatics.com/) with default parameters.
346 Finally, 1,430,611 contigs with length over 100 bps (N50, 2,416 bps; 2,457,704,443
347 bps length in total) were obtained and 74.5% of reads could be mapped back to these
348 Contigs. All these contigs were sent to predict genes with Prodigal (version 1.5)
349 (Hyatt et al. 2010) using `-meta` parameter and finally 3,234,330 genes were obtained.
350 By removing exactly the same genes using USEARCH (version 8.0.1623) (Edgar
351 2010) unique command (parameters -fastx_uniques), 3,234,246 million genes were
352 kept; this set was defined as 'unique gene set'. Reads were mapped back to the contig
353 set and 'unique gene set' to obtain reads coverage matrixes for contigs and genes. The
354 matrix of genes was finally normalized to cell numbers. For metatranscriptome
355 samples, after quality filtering, the SortMeRNAv1.9 was used to remove all the
356 possible ribosomal RNA by aligning to six databases of bacteria, archaea and
357 eukaryotic small and large subunits (Kopylova et al. 2012). RNA reads for each
358 sample were then mapped back to the `unique gene set` to get the transcript coverage
359 for each gene with CLC genomic workbench 6.5.3 using parameters of gap penalty 2,
360 gap extension 3, length fraction 0.8 and similarity at least 0.9.

361 Taxonomy composition of the metagenome was generated with MEGAN6 (Huson et
362 al. 2015). In detail, all genes were aligned to NCBI NR database (version 201603)
363 with diamondv1.09 (Buchfink et al. 2015) to find out the homology proteins. To each
364 gene, the local common ancestors (LCA) were applied using the taxonomy
365 information of the hit NR protein in NCBI taxonomy database (Acland et al. 2014)
366 and then this gene was annotated with the common ancestor taxonomy. We further
367 processed the NCBI taxonomy annotation results to remove those subdivisions and
368 subgroups to format the annotation to 7 levels from kingdom to species. Among total

369 3,234,246 unique genes predicted, 2,348,907 could be aligned to NR database. The
370 remaining 885,339 (27.3%) genes could not be annotated with the NR database. The
371 abundance of each taxon was a sum of all the annotated genes under that taxon in
372 every sample. Antibiotic resistant genes (ARGs) were annotated with SARG database
373 which contained a type-subtype structure annotation (Yang et al. 2016). Metal
374 resistance genes (MRGs) were detected by aligning the "unique gene set" to the MRG
375 database (Li et al. 2017). Absolute abundance and transcript was determined by
376 AQMM.

377 **Declarations**

378 *Data availability*

379 The metagenome and metatranscriptome raw data were deposited in NCBI SRA under
380 accession number XXX.

381 *Analyzing document*

382 The analyzing document for the whole data analysis and simulation process could be
383 accessed from
384 https://github.com/biofuture/aqmm/blob/master/Analysing_document.txt

385 **Conflict of interest**

386 The authors declare no conflict of interest

387 **Acknowledgements**

388 The authors would like to thank GRF of Hong Kong for financial support
389 (172099/14E). Xiaotao Jiang would like to thank The University of Hong Kong for
390 the postgraduate scholarship, and Dr. Ke Yu, Dr. Li-Guan, and Dr. Andong, Li, would
391 like to thank The University of Hong Kong for postdoc fellowship.

392 ## Contributions

393 T. Zhang and X.-T. Jiang design the study of quantification. X.-T. Jiang developed the
394 software and performed the wet-lab and simulation experiments. X.-T. Jiang
395 performed the bioinformatics analyses. X.-T. Jiang, A.D. Li and K. Y. did the DNA
396 and RNA extraction experiment. L.-G. Li did the MRG analyses. T. Zhang and X.-T.
397 Jiang wrote the manuscript. T. Zhang, X.-T. Jiang, A.D. Li, L.G. Li and X.L. Yin
398 revised the manuscript.

399 ## Reference

400 Acland, A., Agarwala, R., Barrett, T., Beck, J., Benson, D.A., Bollin, C., Bolton, E., Bryant,
401 S.H., Canese, K. and Church, D.M. (2014) Database resources of the national center for
402 biotechnology information. Nucleic acids research 42(Database issue), D7.
403 Alexander, H., Jenkins, B.D., Rynearson, T.A. and Dyhrman, S.T. (2015) Metatranscriptome
404 analyses indicate resource partitioning between diatoms in the field. Proceedings of the
405 National Academy of Sciences 112(17), E2182-E2190.
406 Becker-André, M. and Hahlbrock, K. (1989) Absolute mRNA quantification using the
407 polymerase chain reaction (PCR). A novel approach by a P CR aided t ranscipt t itration assay
408 (PATTY). Nucleic acids research 17(22), 9437-9446.
409 Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina
410 sequence data. Bioinformatics, btu170.
411 Buchfink, B., Xie, C. and Huson, D.H. (2015) Fast and sensitive protein alignment using
412 DIAMOND. Nature Methods 12(1), 59-60.
413 Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A.,
414 Szcześniak, M.W., Gaffney, D.J., Elo, L.L. and Zhang, X. (2016) A survey of best practices
415 for RNA-seq data analysis. Genome biology 17(1), 13.
416 Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST.
417 Bioinformatics 26(19), 2460-2461.
418 Foladori, P., Bruni, L., Tamburini, S. and Ziglio, G. (2010) Direct quantification of bacterial
419 biomass in influent, effluent and activated sludge of wastewater treatment plants by using
420 flow cytometry. Water Research 44(13), 3807-3818.
421 Gifford, S.M., Sharma, S., Rinta-Kanto, J.M. and Moran, M.A. (2011) Quantitative analysis
422 of a deeply sequenced marine microbial metatranscriptome. The ISME journal 5(3), 461-472.
423 Huson, D., Beier, S., Buchfink, B., Flade, I., Górska, A., El-Hadidi, M., Mitra, S.,
424 Ruscheweyh, H.-J. and Tappu, R. (2015) MEGAN6-Microbiome analysis involving hundreds

425 of samples and billions of reads, preparation.
426 Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J. (2010)
427 Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC
428 bioinformatics 11(1), 119.
429 Jorth, P., Turner, K.H., Gumus, P., Nizam, N., Buduneli, N. and Whiteley, M. (2014)
430 Metatranscriptomics of the human oral microbiome during health and disease. MBio 5(2),
431 e01012-01014.
432 Kanno, J., Aisaki, K.-i., Igarashi, K., Nakatsu, N., Ono, A., Kodama, Y. and Nagao, T. (2006)
433 " Per cell" normalization method for mRNA measurement by quantitative PCR and
434 microarrays. BMC genomics 7(1), 64.
435 Klingenberg, H. and Meinicke, P. (2017) How To Normalize Metatranscriptomic Count Data
436 For Differential Expression Analysis. bioRxiv, 134650.
437 Kopylova, E., Noé, L. and Touzet, H. (2012) SortMeRNA: fast and accurate filtering of
438 ribosomal RNAs in metatranscriptomic data. Bioinformatics 28(24), 3211-3217.
439 Li, L.G., Xia, Y. and Zhang, T. (2017) Co-occurrence of antibiotic and metal resistance genes
440 revealed in complete genome collection. Isme Journal 11(3), 651-662.
441 Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and
442 dispersion for RNA-seq data with DESeq2. Genome biology 15(12), 550.
443 Manti, A., Boi, P., Falcioni, T., Canonico, B., Ventura, A., Sisti, D., Pianetti, A., Balsamo, M.
444 and Papa, S. (2008) Bacterial cell monitoring in wastewater treatment plants by flow
445 cytometry. Water Environment Research 80(4), 346-354.
446 Martinez, X., Pozuelo, M., Pascal, V., Campos, D., Gut, I., Gut, M., Azpiroz, F., Guarner, F.
447 and Manichanh, C. (2016) MetaTrans: an open-source pipeline for metatranscriptomics.
448 Scientific reports 6, 26447.
449 Mason, O.U., Hazen, T.C., Borglin, S., Chain, P.S., Dubinsky, E.A., Fortney, J.L., Han, J.,
450 Holman, H.-Y.N., Hultman, J. and Lamendella, R. (2012) Metagenome, metatranscriptome
451 and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. The
452 ISME journal 6(9), 1715-1727.
453 Moran, M.A., Satinsky, B., Gifford, S.M., Luo, H., Rivers, A., Chan, L.-K., Meng, J., Durham,
454 B.P., Shen, C. and Varaljay, V.A. (2013) Sizing up metatranscriptomics. The ISME journal
455 7(2), 237.
456 Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and
457 quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5(7), 621-628.
458 Mumy, K.L. and Findlay, R.H. (2004) Convenient determination of DNA extraction
459 efficiency using an external DNA recovery standard and quantitative-competitive PCR.
460 Journal of Microbiological Methods 57(2), 259-268.
461 Narayanasamy, S., Jarosz, Y., Muller, E.E., Heintz-Buschart, A., Herold, M., Kaysen, A.,
462 Laczny, C.C., Pinel, N., May, P. and Wilmes, P. (2016) IMP: a pipeline for reproducible

463 reference-independent integrated metagenomic and metatranscriptomic analyses. Genome
464 biology 17(1), 260.
465 Nayfach, S. and Pollard, K.S. (2015) Average genome size estimation improves comparative
466 metagenomics and sheds light on the functional ecology of the human microbiome. Genome
467 biology 16(1), 51.
468 Ni, Y., Li, J. and Panagiotou, G. (2016) COMAN: a web server for comprehensive
469 metatranscriptomics analysis. BMC genomics 17(1), 622.
470 Oyserman, B.O., Noguera, D.R., del Rio, T.G., Tringe, S.G. and McMahon, K.D. (2015)
471 Metatranscriptomic insights on gene expression and regulatory controls in Candidatus
472 Accumulibacter phosphatis. The ISME journal.
473 Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D.
474 and Betel, D. (2013) Comprehensive evaluation of differential gene expression analysis
475 methods for RNA-seq data. Genome biology 14(9), 3158.
476 Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for
477 differential expression analysis of digital gene expression data. Bioinformatics 26(1),
478 139-140.
479 Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential
480 expression analysis of RNA-seq data. Genome biology 11(3), R25.
481 Shi, Y.M., Tyson, G.W. and DeLong, E.F. (2009) Metatranscriptomics reveals unique
482 microbial small RNAs in the ocean's water column. nature 459(7244), 266-U154.
483 Stark, L., Giersch, T. and Wünschiers, R. (2014) Efficiency of RNA extraction from selected
484 bacteria in the context of biogas production and metatranscriptomics. Anaerobe 29, 85-90.
485 Turner, T.R., Ramakrishnan, K., Walshaw, J., Heavens, D., Alston, M., Swarbreck, D.,
486 Osbourn, A., Grant, A. and Poole, P.S. (2013) Comparative metatranscriptomics reveals
487 kingdom level changes in the rhizosphere microbiome of plants. The ISME journal 7(12),
488 2248-2258.
489 Wagner, G.P., Kin, K. and Lynch, V.J. (2012) Measurement of mRNA abundance using
490 RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences 131(4),
491 281-285.
492 Westreich, S.T., Korf, I., Mills, D.A. and Lemay, D.G. (2016) SAMSA: a comprehensive
493 metatranscriptome analysis pipeline. BMC bioinformatics 17(1), 399.
494 Whelan, J.A., Russell, N.B. and Whelan, M.A. (2003) A method for the absolute
495 quantification of cDNA using real-time PCR. Journal of immunological methods 278(1),
496 261-269.
497 Yang, Y., Jiang, X.T., Chai, B.L., Ma, L.P., Li, B., Zhang, A.N., Cole, J.R., Tiedje, J.M. and
498 Zhang, T. (2016) ARGs-OAP: online analysis pipeline for antibiotic resistance genes
499 detection from metagenomic data using an integrated structured ARG-database.
500 Bioinformatics 32(15), 2346-2351.

Yu, K. and Zhang, T. (2012) Metagenomic and metatranscriptomic analysis of microbial community structure and gene expression of activated sludge. Plos One 7(5), e38183.

Figures and legends

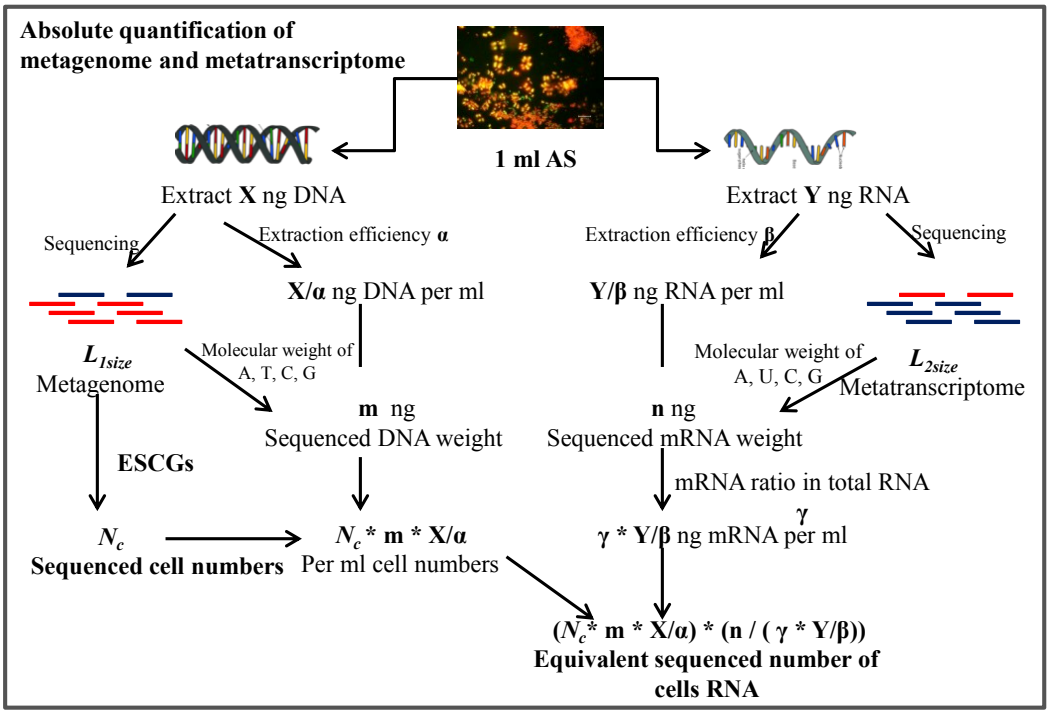
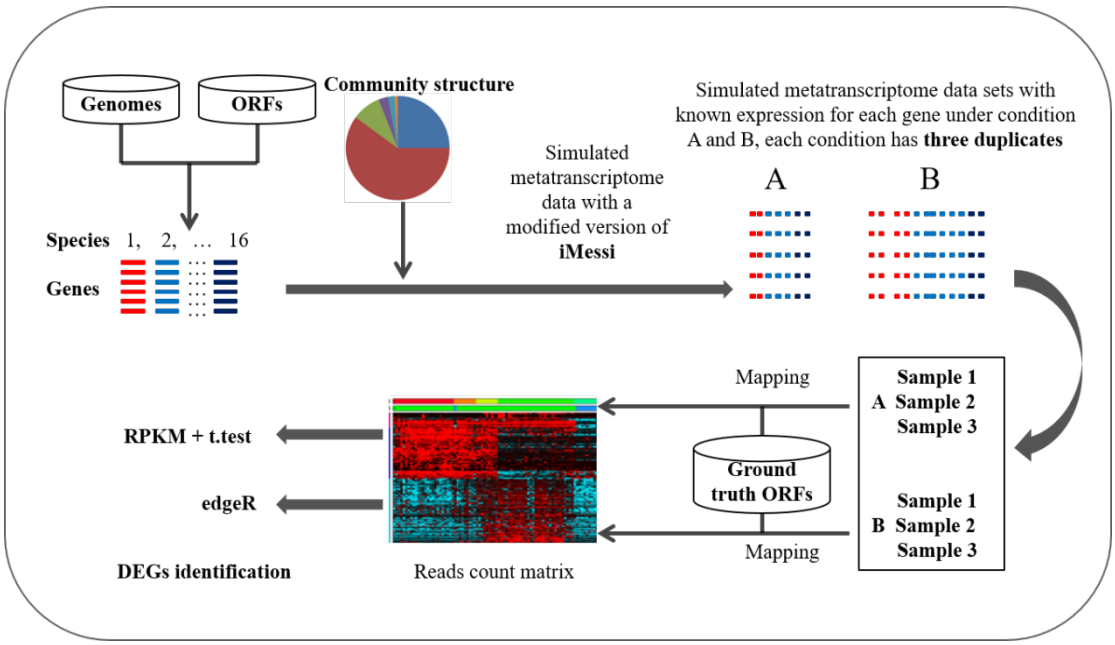
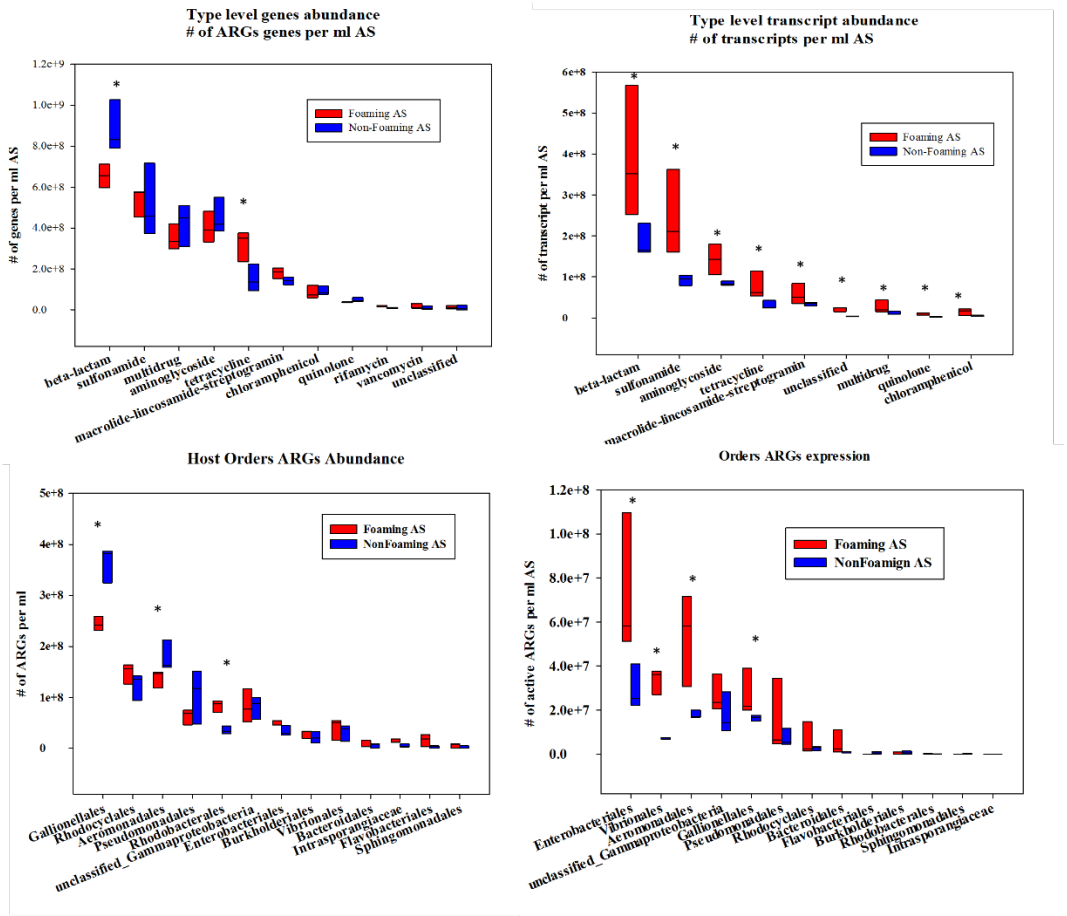


Fig. 1: Schematic flow diagram for absolute quantification of metagenome and metatranscriptome to cell/volume level.



509 **Fig. 2** Flowchart of the simulation datasets generation and analyzing process to get

510 the differential expression genes.



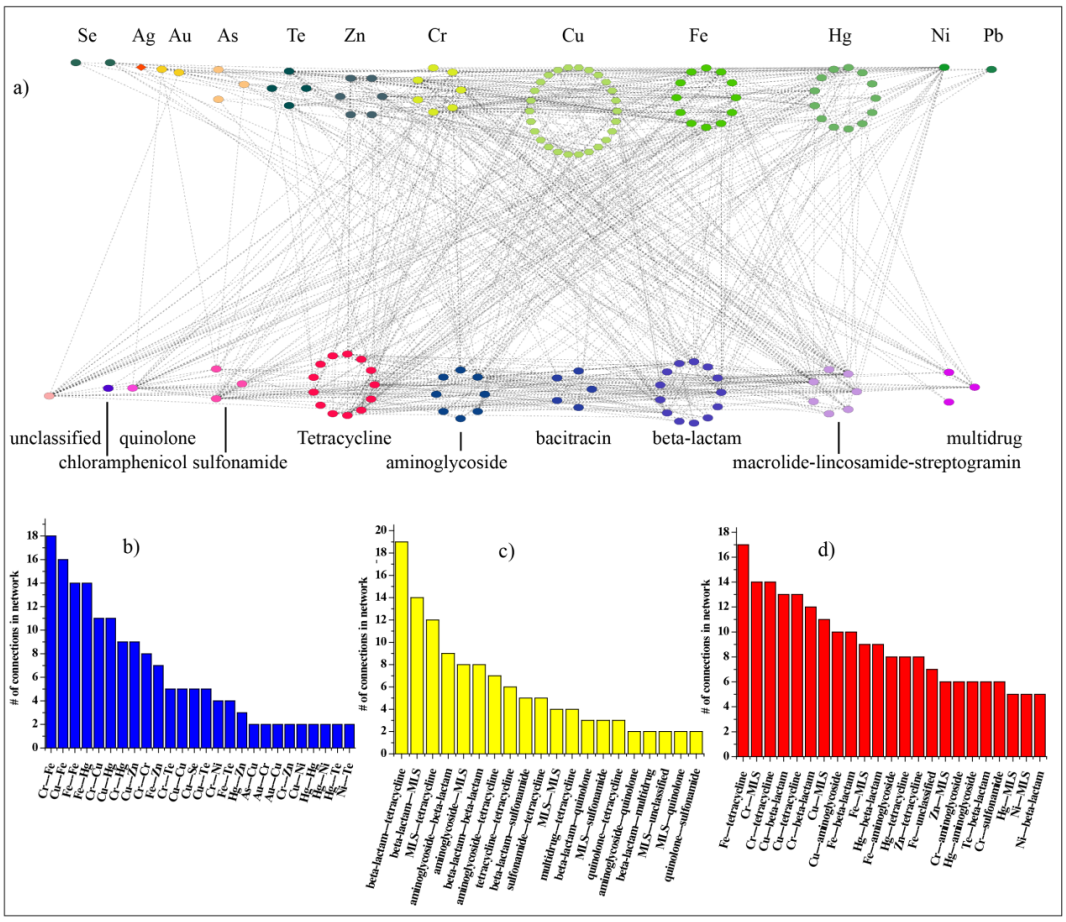
511

512 **Fig. 3:** Absolute quantification of type level ARGs abundance and transcription in

513 FAS and NFAS. ARGs-carry hosts abundance and expression. * represents significant

514 difference (P-value < 0.05).

515



516

517 **Fig. 4:** Co-expression of ARGs and MRGs in Shatin AS, a) was the network of ARGs
518 and MRGs expression; b) was statistical of co-expression within MRGs; c) was
519 statistical of co-expression with ARGs; d) was statistical of co-expression of ARGs
520 and MRGs. Lines in the network represented Spearman association over 0.6, *P*-value
521 0.05 the *P*-value was adjusted with B-H method.

522

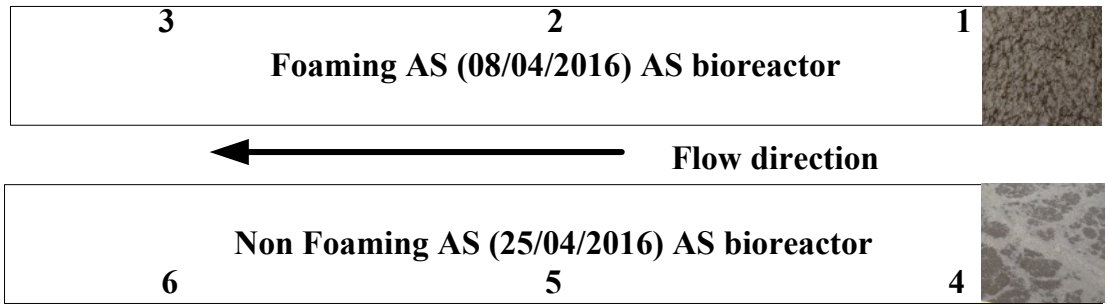


Fig. 5 Samples were collected for foaming activated sludge at 08/04/2016 and non-foaming activated sludge at 25/04/2016 alongside the bioreactor at Shatin wastewater treatment plant.

529 **Table 1** Comparing relative quantification methods with AQMM on detection of

530 DEGs for simulated metatranscriptome data.

	# of genes Higher expression in B	No expression difference	# of genes Higher expression in A
Theoretical Ground Truth	**28524**	**36572**	**0**
RPKM+t-Test (P < 0.05)	16477	11558	**37062**
edgeR	18278	20778	**26040**
AQMM-5%-variation	28744.72 ± 143.53	35807.52 ± 48.08	**543.77 ± 129.72**
AQMM-10%-variation	28740.83 ± 298.43	35801 ± 188.31	**554.17 ± 256.81**
AQMM-20%-variation	28549.48 ± 1007.17	35941.86 ± 919.86	**604.66 ± 654.76**
AQMM-50%-variation	16673.93 ± 9394.27	47694.99 ± 9600.33	727.08 ± 1775.09

531

532

533 **Table 2:** Summary of sequencing outputs and absolute quantification of each sample

534 at cell level with AQMM.

Sample ID	Type	Library size (bps clean data)	Total extracted DNA and RNA (ng/mL)	Estimated sequenced cells *	Estimated cells per mL *
DNA1	Foaming AS	8,567,524,200	49,140	1,541	**6.11E+10**
DNA2	Foaming AS	11,786,228,700	54,600	2,179	**6.98E+10**
DNA3	Foaming AS	10,108,576,800	58,380	1,919	**7.66E+10**
DNA4	Normal AS	8,755,895,700	57,974	1,425	**6.52E+10**
DNA5	Normal AS	9,196,724,100	66,752	1,541	**7.73E+10**

DNA6	Normal AS	11,185,847,400	75,194	1,957	**9.09E+10**
RNA1	Foaming AS	14,894,959,100	12,270	**98,936**	
RNA2	Foaming AS	13,598,855,700	12,710	**99,744**	
RNA3	Foaming AS	15,551,044,400	20,290	**78,449**	
RNA4	Normal AS	15,376,790,700	8,350	**160,343**	
RNA5	Normal AS	16,156,607,900	8,790	**189,776**	
RNA6	Normal AS	13,700,741,100	10,735	**154,925**	

*: Estimated sequenced cells for DNA libraries was using MicrobeCensus and for RNA libraries using AQMM. The assumption for AQMM was that per ml sample used for DNA and RNA extraction contained the same number of cells.

535