Manuscript

Click here to view linked References

Click here to download Manuscript 1. Vezzulli et al REVISED (Main text).doc

Whole-genome enrichment provides deep insights into *Vibrio cholerae* metagenome from an African river

Vezzulli L[1,4]*, Grande C[1,4], Tassistro G[1], Brettar I[2], Höfle MG[2], Pereira RPA[2], Mushi D[2], Pallavicini A[3], Vassallo P[1], Pruzzo C[1]

[1] Department of Earth, Environmental and Life Sciences (DISTAV), University of Genoa, Genoa, Italy
[2] Department of Vaccinology and Applied Microbiology, Helmholtz Centre for Infection Research, Braunschweig, Germany
[3]Department of Life Sciences, University of Trieste, Trieste, Italy
[4]These authors contributed equally to this work

*corresponding author:
Luigi Vezzulli
Department of Earth, Environmental and Life Sciences (DISTAV), University of Genoa, Genoa, Italy
luigi.vezzulli@unige.it

Running title: *Vibrio cholerae* metagenome in African river

Abstract

The detection and typing of *Vibrio cholerae* in natural aquatic environments encounter major methodological challenges related to the fact that the bacterium is often present in environmental matrices at very low abundance in nonculturable state. This study applied, for the first time to our knowledge, a whole-genome enrichment (WGE) and next generation sequencing (NGS) approach for direct genotyping and metagenomic analysis of low abundant *V. cholerae* DNA (<50 genome unit/L) from natural water collected in the Morogoro river (Tanzania). The protocol is based on the use of biotinylated RNA baits for target enrichment of *V. cholerae* metagenomic DNA via hybridization.

An enriched *V. cholerae* metagenome library was generated and sequenced on a Illumina MiSeq platform. Up to 1.8X10^7 bp (4.5x mean read depth) were found to map against *V. cholerae* reference genome sequences representing an increase of about 2500 times in target DNA coverage compared to theoretical calculations of performance for shotgun metagenomics. Analysis of metagenomic data revealed the presence of several *V. cholerae* virulence and virulence associated genes in river water including major virulence regions (*e.g.* CTX prophage and *Vibrio* pathogenicity island-1) and genetic markers of epidemic strains (*e.g.* O1-antigen biosynthesis gene cluster) that were not detectable by standard culture and molecular techniques. Overall, besides providing a powerful tool for direct genotyping of *V. cholerae* in complex environmental matrices this study provides a "proof of concept" on the methodological gap that might currently preclude a more

comprehensive understanding of toxigenic *V. cholerae* emergence from natural aquatic environments.

Main text

Vibrio cholerae, the causative agent of epidemic cholera, is naturally found in the aquatic environment that, according to the “cholera paradigm”, is believed to play an important role in cholera epidemiology [1]. However, detection and typing of the bacterium from environmental sources is not straightforward due mostly to existing methodological limitations. *V. cholerae* is often present in the aquatic environment in a viable but not culturable physiological (VBNC) state thus being not longer detectable by conventional (culture-based) microbiological methods [2]. In addition *V. cholerae* cells might be present in environmental matrices at very low abundance within complex microbial communities and this may also hamper their detection by PCR or shotgun metagenomic techniques [3]. All of these issues strongly limit our capability to track epidemic outbreaks (e.g. tracking the source of disease outbreaks) and may pose the fundamental question on whether the role of the environment as a reservoir of toxigenic *V. cholerae* strains or their genes is by far underestimated.

To address this important methodological challenge, this study applied, for the first time to our knowledge, a whole-genome enrichment (WGE) and next generation sequencing (NGS) approach for direct genotyping and metagenomic analysis of low abundant *V. cholerae* DNA in complex environmental samples. The protocol is based on the use of biotinylated RNA baits for target enrichment of *V. cholerae* metagenomic DNA via hybridization [4] (supplementary methods). Baits were produced from genomic DNA extracted from different *V. cholerae* strains representative of the main pathotypes (*V. cholerae* N16961 [serogroup O1, biotype El Tor], *V. cholerae* O395 [serogroup O1, biotype classical], *V. cholerae* MO10 [serogroup O139] and *V. cholerae* TMA21 [serogroup non O1/O139]) using MYcroarray WGE proprietary technology (MYcroarray, Ann Arbor, MI, USA). WGE was applied on a synthetic metagenome (SM) sample composed of equal amount of genomic DNA from *V. cholerae* N16961 and other phylogenetically affiliated bacterial strains and a natural water sample (RS) collected in the Morogoro river (Tanzania) (see supplementary methods for detailed composition of SM sample and rationale adopted in selection of RS sample).

Bacterial concentrations in RS sample was of 2x10^{10} genome unit/L and 16SrDNA profiling analysis of the bacterial community estimated that the sample contained 205 OTUs of which 185 were classified with SILVA reference sequences. The bacterial community was dominated by the class of *Gammaproteobacteria* that accounted for nearly 80% of the overall community structure (Fig. 1). Among this class the most dominant bacterial genus was *Stenotrophomonas* within the order *Xanthomonadales* and *Aeromonas* within the order *Aeromonadales* which represented 33%

and 26% of the overall bacterial community, respectively (Fig. 1). Bacteria belonging to the *Vibrio* genus represented less than <0.02 % of the bacterial community whilst *V. cholerae* concentration was estimated to be less than 50 genome unit/L by applying a species-specific PCR protocol [3]. In addition, the sample tested negative for standard culturing (using *V. cholerae* selective media) and PCR specifically targeting *V. cholerae* O1/O139 antigen markers and the main virulence factors e.g. genes encoding for the cholera toxin (*ctxA*) and the toxin coregulated pilus (*tcpA*) (supplementary methods).

Enriched *V. cholerae* genome libraries were generated for both samples and sequenced on a Illumina MiSeq platform. In order to evaluate method performance metagenomic reads (average read length= 251bp) were mapped to *V. cholerae* reference genome sequences using the mapping tool of the CLC Genomics Workbench software (version 9.5.1) (supplementary methods). The success of the enrichment was evident for the SM sample where a total of $1.70\text{x}10^9$ bp out of $1.74\text{x}10^9$ bp mapped against *V. cholerae* N16961 reference genome sequence whilst, on average, $4.2\text{x}10^7$ bp mapped against control genome sequences from other species in the community (Fig. S1). This means that more than 97% of mapped reads belong to *V. cholerae*. In addition coverage was highly uniform across the two *V. cholerae* chromosomes suggesting that the stringency of hybridization was optimal to capture the majority of *V. cholerae* genome content but not of those of phylogenetically related species (Fig. S1).

In the RS sample up to $1.8\text{X}10^7$bp (4.5x mean read depth) out of $2.9\text{x}10^9$bp were found to map against reference *V. cholerae* genome sequences (Table S1, Fig. 2). This represents an increase of about 2500 times in target DNA coverage compared to theoretical calculations of performance for shotgun metagenomics (Table S2). Interestingly the highest number of reads were allocated to *V. cholerae* O1 genome sequences including *V. cholerae* 4784 isolated from Tanzania (Table S1). Albeit assembly for large majority of reads was not possible (*e.g.* mainly due to complex nature of environmental DNA and low concentrations of *V. cholerae* DNA in the sample), the coverage obtained allows for identification of specific genes and genetic regions within the *V. cholerae* pangenome. Amongst relevant findings, reads mapping against *V. cholerae* genome specific regions encoding for somatic antigens O1 and O139 (supplementary methods) revealed that the O1-antigen biosynthesis gene cluster (*wbe*) was present in the metagenome whilst the O139 gene cluster was lacking (Fig. 2). In addition, mapping reads against virulence factor (VFDB, http://www.mgc.ac.cn/VFs) [5] and antibiotic resistance genes database (ARDB, https://ardb.cbcb.umd.edu) [6] showed the presence of *V. cholerae* T6SS-encoding gene cluster and the MARTX region encoding for the RTX toxin gene (Table 1). A ca 3000bp consensus sequence showing >99% nucleotide identity with SXT-related integrating conjugative elements (ICEs) of *V. cholerae* was found containing genes encoding for a transposase and *strA-strB* streptomycin resistance proteins. Interestingly, the SXT element is commonly found in *V cholerae* O1 and O139 isolates from Africa but is not present in *V. cholerae* N16961 [7]. Specific sequences for genes encoding the two major virulence factors e.g. the CTX prophage containing cholera toxin genes (*ctxAB*) and the vibrio pathogenicity island-1 (VPI-1) containing genes required for toxin-coregulated-pilus (TCP) biogenesis were also found (Fig. 2).

Taken together, these findings support the presence of a toxigenic *V cholerae* O1 metagenomic DNA in river water that was not detectable by standard culture and molecular techniques. This is consistent to the fact that *V. cholerae* O1 strains are responsible for most cholera outbreaks in Tanzania [8-9]. The nature of such DNA might be cellular but also extracellular (eDNA) and/or having a viral origin (*e.g.* VPIφ and CTXφ phages) thus warranting further investigation. For instance, eDNA may represent a reservoir of virulence genes as it can survive for long period of time when bound to environmental compounds such as clay minerals, larger organic molecules and other charged particles [10]. eDNA is also involved in horizontal gene transfer [11-12] and may contribute to emergence of virulent *V. cholerae* strains from the aquatic environment [13].

Overall, this study successfully applied a new cutting edge high resolution technique for direct genotyping and metagenomic analysis of low abundant *V. cholerae* DNA in environmental samples. Due to the high costs and technical difficulties this protocol is not intended to be used in routine microbiological control practices but it is instead designed for research purpose and/or in-depth outbreak investigation studies. Evidence for a “hidden” metagenomic DNA, including virulence genes and genetic markers of epidemic strains no detectable by commonly employed culture and molecular techniques provides a “proof of concept” on the methodological gap that might currently preclude a more comprehensive understanding of toxigenic *V. cholerae* emergence from aquatic environments. Filling such a gap, may lead to a breakthrough in addressing the “cholera paradigm” in cholera affected areas.

Acknowledgements

We are particularly indebted to Dr. Alison Devault (MYcroarray, USA) for helpful assistance and advice in WGE analysis. This work was supported by the European FP7 project “Protecting the health of Europeans by improving methods for the detection of pathogens in drinking water and water used in food preparation” –AQUAVALENS (grant number 311846). The HORIZON2020 project “Preventing and mitigating farmed bivalve disease- VIVALDI (grant number 678589)” is also acknowledged.

References

1. Colwell RR (1996) Global climate and infectious disease: the cholera paradigm. Science

274: 2031–2025

2. Xu HS, Roberts N, Singleton FL, Attwell RW, Grimes DJ, Colwell, RR (1982) Survival and viability of nonculturable *Escherichia coli* and *Vibrio cholerae* in the estuarine and marine environment. Microb Ecol 8: 313–323

3. Vezzulli L, Stauder M, Grande C, Pezzati E, Verheye HM, Owens NJP et al. (2015) gbpA as a novel qPCR target for the species-specific detection of *Vibrio cholerae* O1, O139, non-O1/non-O139 in environmental, stool, and historical Continuous Plankton Recorder Samples. Plos One 10: e0123983 doi:10.1371/journal.pone.0123983

4. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27(2):182-189

5. Chen LH, Xiong ZH, Sun LL, Yang J, Jin Q (2012) VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res 40: D641-D645

6. Liu B, Pop M (2009) ARDB-Antibiotic Resistance Genes Database. Nucleic Acids Res 37: D443-7

7. Burrus V, Quezada-Calvillo R, Marrero J, Waldor MK (2006) SXT-Related Integrating Conjugative Element in New World *Vibrio cholerae*. Appl Environ Microbiol 72(4): 3054–3057

8. Acosta CJ, Galindo CM, Kimario J, Senkoro K, Urassa H, Casals C et al (2001) Cholera outbreak in Southern Tanzania: risk factors and patterns of transmission. Emerg Infect Dis 7:583-7

9. Naha A, Chowdhury G, Ghosh-Banerjee J, Senoh M, Takahashi T, Ley B et al (2013) Molecular characterization of high-level-cholera-toxin-producing El Tor variant *Vibrio cholerae* strains in the Zanzibar Archipelago of Tanzania. J Clin Microbiol 51:1040-1045

10. Crecchio C, Stotzky G. (1998) Binding of DNA on humic acids: effect on transformation of Bacillus subtilis and resistance to DNase. Soil Biol Biochem 30:1061–1067

11. Vlassov VV, Laktionov PP, Rykova EY (2007) Extracellular nucleic acids. Bioessays 29(7):654-67

12. Nielsen KM, Johnsen PJ, Bensasson D, Daffonchio D (2007) Release and persistence of extracellular DNA in the environment Environ. Biosafety Res 6:37–53

13. Blokesch M, Schoolnik GK (2007) Serogroup conversion of *Vibrio cholerae* in aquatic reservoirs. PLoS Pathog 3:733-742

Figure 1

Relative abundances of bacterial classes and genera found in Morogoro river water (Tanzania). Relative abundances were calculated from a total of 16.344 reads classified in OTUs deriving from an input database of 57.978 trimmed reads (average size 409bp).

Figure 2

Metagenomic reads from RS sample assigned to *V. cholerae* N16961 reference genome (accession: AE003852/AE003853) and selected reference genomic regions encoding for major virulence and virulence associated genes: MARTX: *V. cholerae* Rtx toxin gene cluster (accession: AF119150.1); O139: *V. cholerae* O139 MO10 cont1.55 (accession: AAKF03000053.1); O1: *V. cholerae* O1 *wbe* gene cluster (accession: KC152957.1); VPI-1: *V.cholerae tcp* gene cluster, (accession: X64098.1); CTXφ: *Vibrio* phage CTX chromosome I (accession: NC_015209.1). Read mapping was performed under stringent conditions (100% minimum read length matching the reference at >98% nucleotide identity) using the mapping tool of the CLC Genomics Workbench software (version 9.5.1) (supplementary methods).

Table 1.

Metagenomic reads from RS sample assigned to *V. cholerae* virulence genes using the virulence factor database [5]. Read mapping was performed under stringent conditions (100% minimum read length matching the reference at >98% nucleotide identity) using the mapping tool of the CLC Genomics Workbench software (version 9.5.1) (supplementary methods). Specificity of reads matching reference sequences was also assessed by running Blastn software against NR database (version 2.2.28+; http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Function	Virulence factor	Gene (total reads assigned)
Secretion system	*Type 6 Secretion system (T6SS)*	VFG2088 icmF/vasK (195), G00326 vgrG-2 (135), VFG2084 clpV/vasG (119), VFG2085 vasH (81), VFG2080 vasC (80), VFG2093 vipB/tssC (75), VFG2078 vasA (74), VFG2082 vasE (65), VFG2089 vasL (63), VFG2091 vgrG-3 (59), G00325 hcp-2 (58), VFG2083 vasF/tssL (53), VFG2087 vasJ (43), VFG2079 vasB (39), VFG2094 VCA0109 (31), VFG2086 vasI (24), VFG2081 vasD (19), VFG2092 vipA/tssB (16), VFG2090 VCA0122 (15)
Toxin	*Multifunctional autoprocessing RTX toxin (MARTX)*	VFG0983 RtxA (259), R004437 rtxB (60), R004441 rtxD (45), R004434|rtxC (25)
	Cholera toxin (CT)	VFG0107 ctxA (6), VFG0108 ctxB (2)
	Zona occludens toxin (zot)	VFG0109 zot (6)
	Accessory cholera	VFG0110 ace (2)
Adherence	*Toxin-coregulated pilus (TCP)*	VFG0098 tcpT (7), VFG0100 tcpF (7), VFG0092 tcpB (5), VFG0094 tcpC (5), VFG0091 tcpA (4), VFG0096 tcpD (4), VFG0099 tcpE (4), VFG0102 tcpJ (4), VFG0088 tcpI (3), VFG0089 tcpP (3), VFG0093 tcpQ (2), VFG0097 tcpS (2),
	Accessory colonization factor (ACF)	VFG0106 acfD (10), VFG0104 acfB (9), VFG0105 acfC (3)

Figure 1

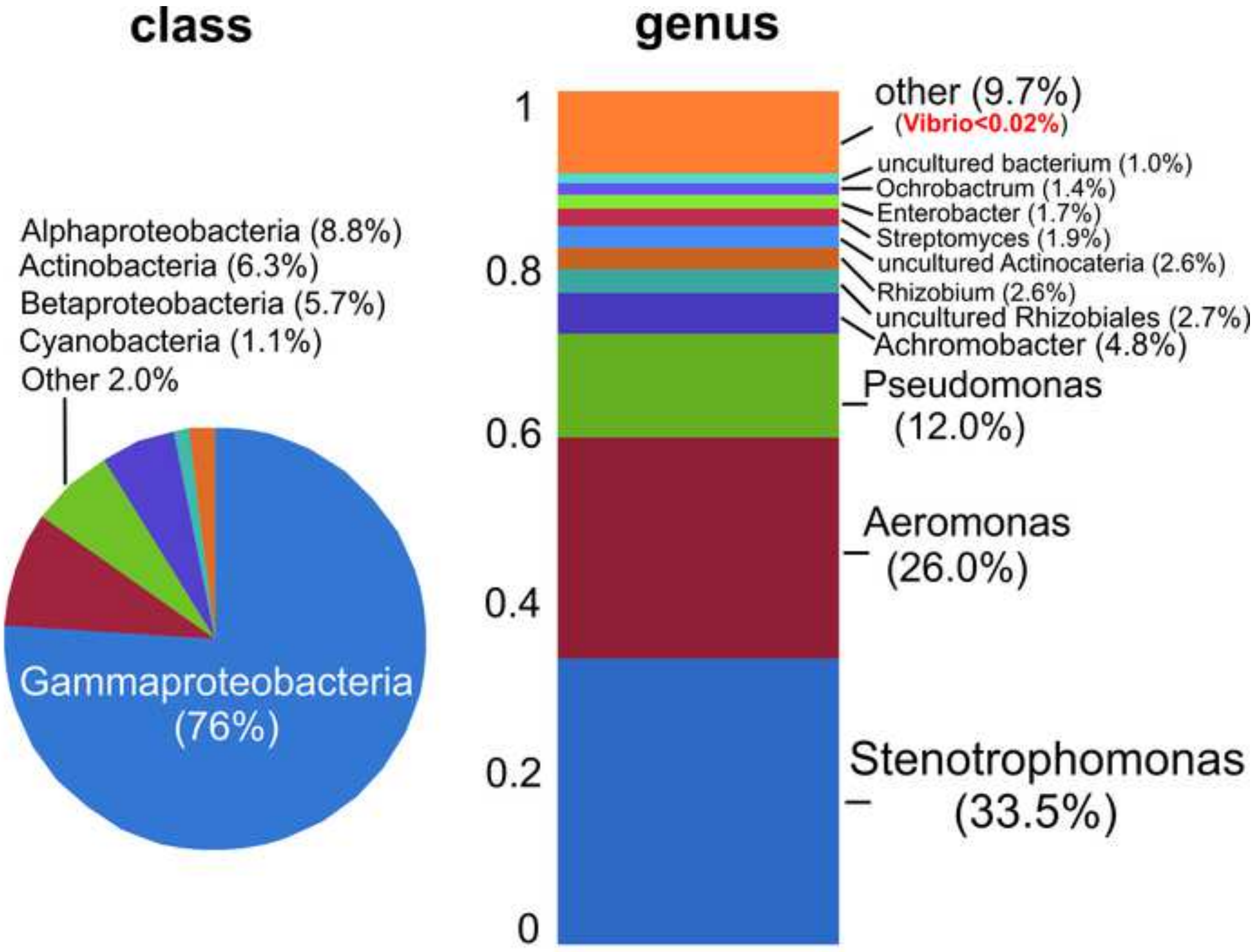
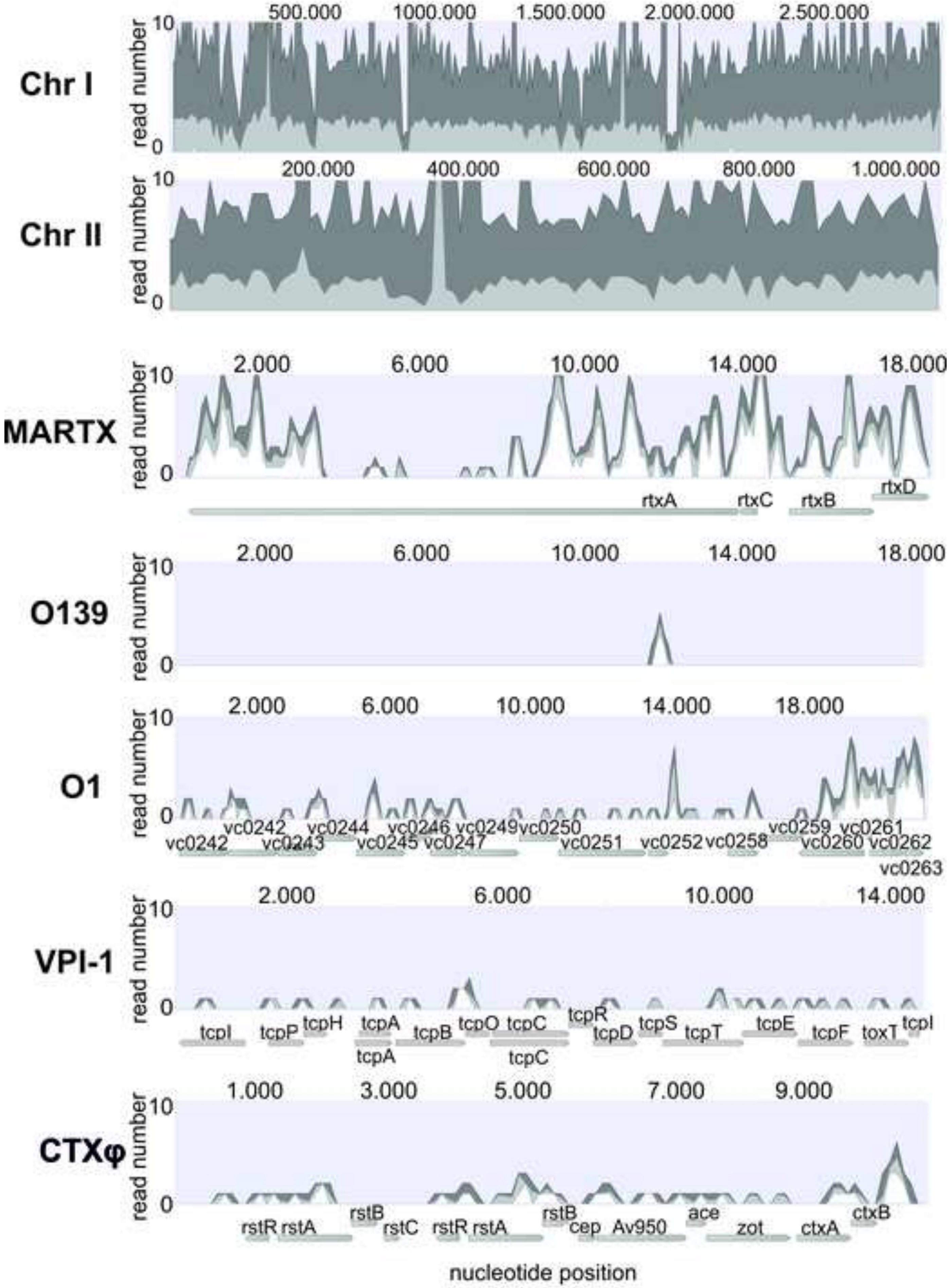

class
Alphaproteobacteria (8.8%)
Actinobacteria (6.3%)
Betaproteobacteria (5.7%)
Cyanobacteria (1.1%)
Other 2.0%
Gammaproteobacteria (76%)
genus
1
0.8
0.6
0.4
0.2
0
other (9.7%)
(Vibrio<0.02%)
uncultured bacterium (1.0%)
Ochrobactrum (1.4%)
Enterobacter (1.7%)
Streptomyces (1.9%)
uncultured Actinocateria (2.6%)
Rhizobium (2.6%)
uncultured Rhizobiales (2.7%)
Achromobacter (4.8%)
Pseudomonas (12.0%)
Aeromonas (26.0%)
Stenotrophomonas (33.5%)

Figure 2

Click here to download Figure 3. Vezzulli et al (Figure 2).tif



Supplementary Material

Click here to access/download

Supplementary Material

4. Vezzulli et al REVISED (Supplementary Material).docx