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

Supplementary methods

Selection of natural river sample for whole-genome enrichment experiment

Twenty-four water samples were collected from Morogoro river (06° 49' 36'' S, 37° 40' 09'' E) in Tanzania from October 2012 to June 2013 using sterile one liter plastic bottles (Thermo Scientific, Nalgene, USA) and the guidance of standard methods for sampling environmental waters [14]. Morogoro river is the main source of drinking water for the populations residing in Morogoro City (315.000 people) and Dar es Salaam City (4.400.000 people) in Tanzania. Cholera is endemic in the region since 1976 with cases reported each year [15]. All collected samples were screened for *V. cholerae* using both culture and molecular techniques. Briefly water (250ml) was filtered through nucleopore-filters (45 mm diameter, 0.2 µm pore size, Track-etch, Whatman Corp) using vacuum filtration and plated on appropriate culture media. For molecular analysis biomass collected on the nuclepore filter was scraped off the filter with a sterile plastic spatula in a sterile petri dish, resuspended three times in 60 µl sterile water and transferred to FTA Card classic (Sigma-Aldrich Corp). FTA card technology is particularly suitable for sampling in remote areas such as Morogoro region as it is quite economic, DNA of the bacterial biomass is stored on the card material at ambient temperature for many years and does not need any freezing or frozen transportation. DNA was then extracted from the FTA classic card back in the laboratory using Rapid Water DNA isolation kit, (MoBio Laboratories, Solana Beach, CA, USA) following the manufacture instructions. Purified DNA was stored at -20°C until analysis.

Total bacterial number and *V. cholerae* concentration were assessed by Real-Time PCR using a capillary-based LightCycler instrument (Roche Diagnostics, Mannheim, Germany) and a standard curve method for quantification. For total bacteria, a LightCycler- FastStart DNA Master SYBR Green I kit (Roche Diagnostics, Milan, Italy) optimized for use with glass capillaries were used following conditions described in Vezzulli *et al.* [16]. The oligonucleotide primers used in the PCR reaction were 967f-5-CAACGCGAAGAACCTTACC-3 and 1046r-5

CGACAGCCATGCANCACCT-3 [17], specific for the domain Bacteria, amplifying positions 567–680 and 965–1063 (V6 hyper variable region) of the *Escherichia coli* numbering of the 16S rRNA. For *V. cholerae* enumeration a Light Cycler TaqMan Master Mix chemistry were used with species-specific primers VcgbpAF-5-CCGCAGCTTCCTTCTACAAC-3, VcgbpAR-5-GGCTTTGGTTAGCGTCTCAG-3 and VcgbpApr-5-FAM-AACCCAGCAGGTCAAATCATTCCAAGTA-BBQ probe following conditions described in Vezzulli *et al* [3]. To address the problem of PCR inhibition all qPCR reactions were performed both on whole and diluted (1:10) DNA samples.

V. cholerae was not cultured in any of the collected water samples. For further analyses and WGE experiment one river sample (hereinafter referred to as “RS sample”) was selected that resulted negative to culturing but weakly positive (<50 genome unit/L) to the *V. cholerae* species-specific qPCR assay.

PCR detection of virulence factor genes

Detection of major *V. cholerae* virulence encoding genes in RS sample were performed with the LightCycler instrument and the LightCycler- FastStart DNA Master SYBR Green I kit (Roche Diagnostics, Milan, Italy) using conditions described in Vezzulli *et al.* [16] and annealing temperature optimised for each set of primers. Primers used were the following:

gene	Primer sequence	reference
ctxA	ctx2-5-cgggcagattctagacctcctg-3 ctx3-5-cgatgatcttggagcattcccac-3	[18]
tcpA	tcpAF-5-cacgataagaaaaccggtcaagag-3 tcpAR/Class-5-ttaccaaatgcaacgccgaatg-3 tcpA R/El Tor-5-cgaaagcaccttctttcacacgttg-3	[19]
rfbN	O1f21-5-tggtttcactgaacagatggg-3 O1r22-5-aggtcatctgtaagtacaacattc-3	[20]
wbfR	O139f2-5-aagcctctttattacgggtgg-3 O139R2-5-gtcaaacccgatcgtaaaggt-3	[20]

16S rRNA gene-based profiling of the bacterial community

A 16SrDNA PCR amplicon library was generated from genomic DNA extracted from RS sample using modified forward primer COM 1F (5-cagcagccgcggtaatac-3) and reverse primer COM2R (5-ccgtcaattcctttgagttt-3) [21] to amplify the V4 and V5 region of the 16S rRNA gene of bacteria. All primers were custom designed to include the 16SrRNA complementary regions plus the complementary sequences to the Illumina specific adapters and flow cell binding sites. All primers were synthesized and HPSF or HPLC purified by Eurofins MWG Operon, Ebersberg, Germany. Three PCR assays were performed. A 1st target enrichment PCR assay with the 16S conserved primers. A 2nd PCR assay, with customised primers and included adapters' complementary regions. A final 3rd PCR, a non-target specific assay, integrated complementary sequences to flow cell binding sites. The obtained library was sequenced on a MiSeq Illumina™ platform (2 × 250 reads).

Bioinformatic analysis of NGS data was performed using the Microbial Genomics module (version 1.3) work-flow of the CLC Genomics workbench (version 9.5.1). After quality trimming based on quality scores (quality nucleotide limit 0.05), trim of ambiguous nucleotides (n=2) and length trimming, reads were clustered at 97% level of similarity into Operational Taxonomical Units (OTUs). Chimera detection and removal was performed by the kmer searches pipeline of the Microbial Genomics module (version 1.3). Ribosomal RNA gene reads were classified against the non-redundant version of the SILVA SSU reference taxonomy (release 119; http://www.arb-silva.de).

Synthetic metagenome (SM)

A synthetic metagenome was prepared, as a positive control, by mixing equal amounts of genomic DNA (100ng/ul) in phosphate-buffered saline (PBS) of the following bacterial strains: *Vibrio cholerae* ATCC 39315, *Escherichia coli* MG1655, *Salmonella enterica* ATCC 19430, *Pseudomonas fluorescens* VET67*, Klebsiella pneumoniae* E12 , *Serratia marcescens* E19, *Providencia stuartii NE35*, *Citrobacter freundii NE111.* All strains were grown overnight in Luria Bertani (LB) broth (Scharlau) at 37°C. DNA was extracted from each single strain using the High Pure PCR Template Preparation Kit (Roche Diagnostics, Mannheim, Germany) according to the manufacturer's instructions. The amount of extracted DNA was quantified using the Quantifluor double-stranded DNA quantification kit (Promega Italia, Milan, Italy).

Whole Genome Enrichment and Sequencing

A Whole Genome Enrichment (WGE) protocol was developed and applied for target enrichment of the *V. cholerae* metagenome [4] in natural water samples. The protocol was based on the use of biotinylated RNA baits (on average >100-mer) for target capturing of *V. cholerae* DNA via hybridization. Baits were produced using the MYcroarray WGE proprietary technology (MYcroarray, Ann Arbor, MI, USA) and made out from genomic DNA extracted from different *V. cholerae* strain 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). 500ng (1.1x10^8 genome unit) of total baits were used for a capture and were capable of enriching single-copy nuclear loci i.e. >99.5% of the capture target region. Based on this assumption the detection limit of the assay downstream of water filtration and DNA extraction was theoretically estimated to be close to 1 genome unit per reaction. Environmental DNA extracted from RS and SM samples were sized on an Agilent Bioanalyzer and enzymatically fragmented using the NEBNext dsDNA Fragmentase (New England Biolabs Inc, MA, USA) protocol to an average size of about 250bp. The fragmented DNA was used for the production of an indexed library for next-generation sequencing on the Illumina platform (Illumina, Inc) using the NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs Inc, MA, USA). About 200 ng of the produced library was used for *V. cholerae* DNA target capturing using the MYbaits protocol (MYcroarray, Ann Arbor, MI, USA) following the manufacture instructions. Briefly the genomic DNA library was heat-denatured and hybridized to the RNA baits in stringent conditions. After hybridization, the biotinylated baits hybridized to captured material were pulled out of the solution with streptavidin-coated magnetic beads and the captured genomic DNA was released by chemical degradation of the RNA baits. Post-capture PCR amplification was finally carried out. RS and SM samples libraries were pooled and sequenced on a MiSeq Illumina™ platform (2 × 250 reads). Sequence reads data were archived at NCBI Sequence Read Archive (SRA) with accession number: SRP078027.

Sequence data analysis

Read-pairs were subject to quality control and reads were quality trimmed based on minimum length (75bp), quality scores (quality nucleotide limit 0.05 based on Phred scale) and the presence of ambiguous nucleotides (n=2). Adapter trimming was also performed. Trimmed reads were than mapped against reference genome sequences (*V. cholerae* N16961, Accession: AE003852/AE003853; *V. cholerae* 4784, Accession: CWRV01000000; *V. cholerae* O395, Accession CP000626/ CP000627; *V. cholerae* MO10, Accession AAKF03000000; *V. cholerae* TMA21, Accession ACHY00000000; *Vibrio mimicus* ATCC 33655, Accession: NZ_LOSJ01000001/ NZ_LOSJ01000002; *E. coli* MG1655, Acession: NC_002695; *P. stuartii* MRSN 2154, Acession: NC_017731; *S. marcescens* Db11, Acession: NZ_HG326223; *K. pneumoniae* HS11286, Accession: NC_016845; *C. freundii* CFNIH1, Accession: NZ_CP007557; *P. fluorescens* F113, Accession: NC_016830; *S. enterica* LT2, Accession: NC_003197) using the mapping tool of the CLC Genomics Workbench (version 9.5.1) from Qiagen. A length fraction of 1.0 and similarity fraction of 0.98 were employed in the analysis (e.g. 100% minimum read length matching the reference at >98% nucleotide identity). Masking of 16rDNA encoding genes was applied. Prediction and annotation of virulence genes was performed by mapping reads against the virulence factor [5] and antibiotic resistance genes [6] database and selected genomic regions (*V. cholerae* Rtx toxin gene cluster, Accession: AF119150.1; *V. cholerae* O1 *wbe* gene cluster, Accession: KC152957.1; *V. cholerae* O139 MO10 cont1.55, Accession: AAKF03000053.1; *Vibrio* phage *CTX* chromosome I, Accession: NC_015209.1; *V.cholerae tcp* gene cluster, Accession: X64098.1) using the mapping tool of the CLC Genomics Workbench with same settings as described above. 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).

Sample contamination

To avoid laboratory contamination of treated samples all the analyses including DNA extraction, DNA amplification and NGS library preparations were carried out in a separate laboratory (non aquatic/non microbiological laboratory) using a dedicated set of pipettes, reagents, and consumables.

References

14. APHA, AWWA, WEF. Standard Methods for examination of water and wastewater. 22nd ed. Washington: American Public Health Association; 2012, 1360 pp. ISBN 978-087553-013-0

15. Ali M, Lopez AL, Ae You Y, Eun Kim Y, Sah B, Maskery B et al (2012) The global burden of cholera. Bulletin of the World Health Organization 90:209-218A

16. Vezzulli L, Brettar I, Pezzati E, Reid PC, Colwell RR, Höfle MG et al (2012) Long-term effects of ocean warming on the prokaryotic community: evidence from the vibrios. ISMEJ 6:21-30

17. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR (2006) Microbial diversity in the deep sea and the underexplored ‘rare biosphere’. Proc Natl Acad Sci USA 103: 12115–12120

18. Fields PI, Popovic T, Wachsmuth K, Olsvik O (1992) Use of polymerase chain reaction for detection of toxigenic Vibrio cholerae O1 strains from the Latin American cholera epidemic. J Clin Microbiol 30:2118-2121

19. Singh DV, Matte MH, Matte GR (2001) Molecular Analysis of *Vibrio cholerae* O1, O139, non-O1, and non-O139 strains: clonal relationships between clinical and environmental isolates. Appl Environ Microbiol 67:910–921

20. Hoshino K, Yamasaki S, Mukhopadhyay AK, Chakraborty S, Basu A, Bhattacharya SK et al (1998) Development and evaluation of a multiplex PCR assay for rapid detection of toxigenic *Vibrio cholerae* O1 and O139. FEMS Immunol Med Microbiol 20: 201-207

21. Schwieger, F, Tebbe C (1998) A new approach to utilize PCR-single-strand-conformation-polymorphism for 16S rRNA gene-based microbial community analysis. Appl Environ Microbiol 64:4870–4876

Table S1

Metagenomic reads assigned to *V. cholerae* reference genome sequences. 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).

Strain	**Serogroup**	**Biotype**	**Geographical origin**	**Mapped reads (bp)**	**Coverage**	**Accession**
V. cholerae N16961	O1	El Tor	Bangladesh	1.8x10^7	4.5X	AE003852/AE003853
V. cholerae 4784	O1	El Tor	Tanzania	1.8x10^7	4.5X	CWRV01000000
V. cholerae O395	O1	Classical	India	1.7x10^7	4.2X	CP000626/ CP000627
V. cholerae MO10	O139		India	1.7x10^7	4.2X	AAKF03000000
V. cholerae TMA21	non-O1/O139		Brazil	1.5x10^7	4.1X	ACHY00000000
V. mimicus ATCC 33655				5.0x10^5	0.1X	NZ_LOSJ01000001/2
E. coli MG1655				4.0x10^5	0.1X	NC_002695

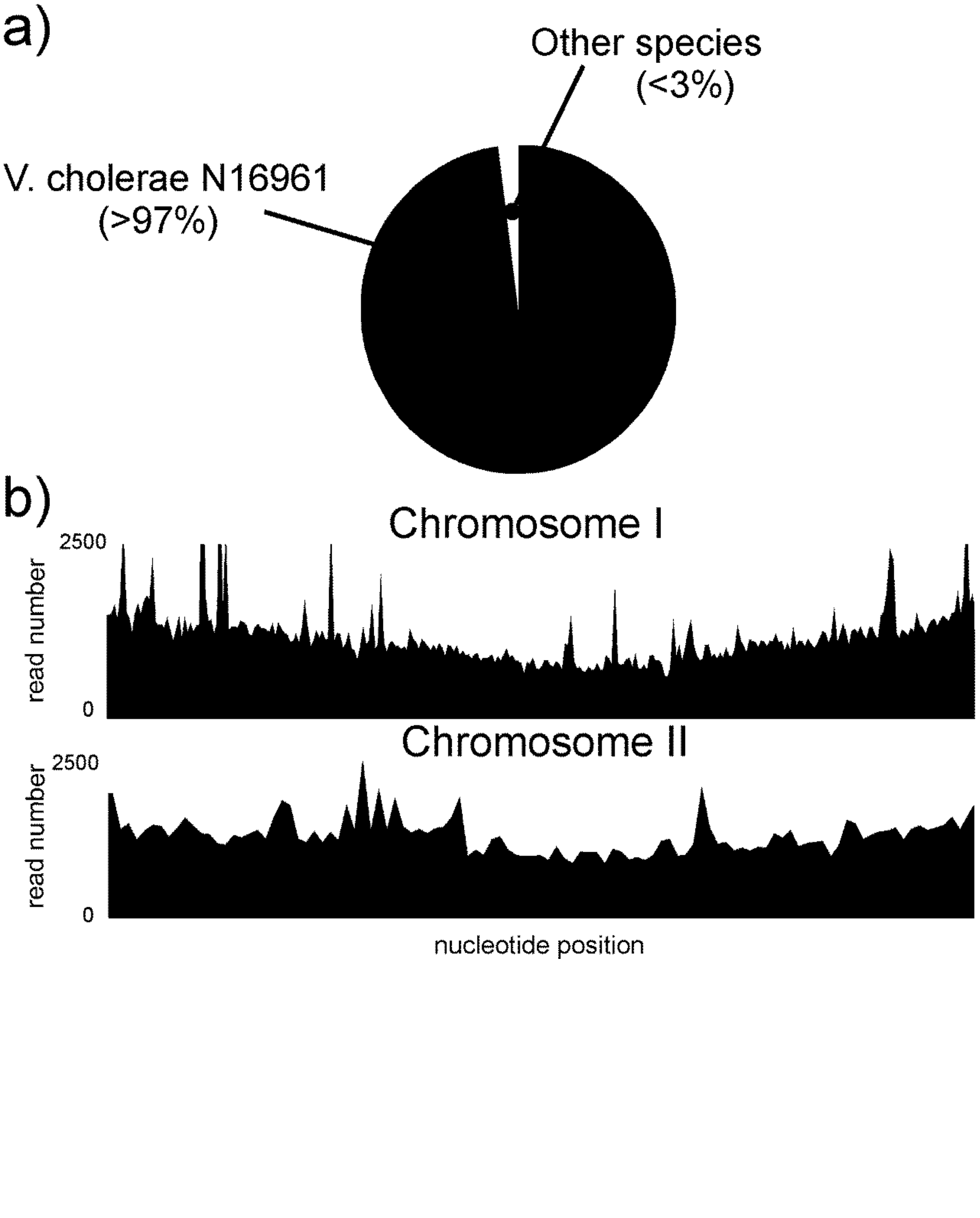
Table S2

Performance of whole genome enrichment (WGE) protocol applied in this study compared with theoretical calculations of performance of a shotgun metagenomic protocol (SMP).

Sample information	
Sample site	Morogoro river (Tanzania)
Sample ID	RS
Sample type	River water
Sample volume (ml)	250
Total extracted DNA (µg)	3.0
Total bacterial concentration (genome unit) (A)	5.0x10^{9}
V. cholerae concentration (genome unit) (B)	50
Present study (WGE)	
Protocol	WGE
Sequencing system	Illumina MiSeq
Number of reads	1.2 x10^{7}
Average read length	251
V cholerae N16961 genome size (number of reads) (C)	1.6 x10^{4}
Reads mapping on *V. cholerae* N16961 genome (D)	7.3 x10^{4}
V. cholerae genome coverage X_1=(D/C)	**4.5X**
Theoretical calculations (SMP)	
Protocol	SMS
Sequencing system	Illumina HiSeq X Series
Max number of reads (E)	6.0 x10^{9}
Average read length	150
Average bacterial genome size (number of reads) (F)	3.3 x10^{4}
Total bacterial concentration (number of reads) (A*F)	1.7 x10^{14}
V. cholerae genome expected coverage X_2=(E/[A*F]*B)	**0.0018X**
WGE:SMP performance ratio (X_1/ X_2)	**2508**

Figure S1

Percentage of metagenomic reads from SM sample assigned to *V. cholerae* N16961 reference genome (accession: AE003852/AE003853) (a) and obtained coverage (b)


a)
Other species
(<3%)
V. cholerae N16961
(>97%)
b)
Chromosome I
read number
Chromosome II
read number
nucleotide position