scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A holistic genome dataset of bacteria, archaea and viruses of the Pearl River estuary

Bu Xu[1,2], Fuyan Li[3], Lanlan Cai[4,5], Rui Zhang[5,6], Lu Fan[2,7] ✉ & Chuanlun Zhang[2,7]

Estuaries are one of the most important coastal ecosystems. While microbiomes and viromes have been separately investigated in some estuaries, few studies holistically deciphered the genomes and connections of viruses and their microbial hosts along an estuarine salinity gradient. Here we applied deep metagenomic sequencing on microbial and viral communities in surface waters of the Pearl River estuary, one of China's largest estuaries with strong anthropogenic impacts. Overall, 1,205 non-redundant prokaryotic genomes with ≥50% completeness and ≤10% contamination, and 78,502 non-redundant viral-like genomes were generated from samples of three size fractions and five salinity levels. Phylogenomic analysis and taxonomy classification show that majority of these estuarine prokaryotic and viral genomes are novel at species level according to public databases. Potential connections between the microbial and viral populations were further investigated by host-virus matching. These combined microbial and viral genomes provide an important complement of global marine genome datasets and should greatly facilitate our understanding of microbe-virus interactions, evolution and their implications in estuarine ecosystems.

Background & Summary

Estuaries are transitional environments between ocean and river. Complex and dynamic estuarine ecosystems are distinguishable from oceanic environments by significant variety of physical, chemical and geomorphologic conditions[1–4]. These factors have structured a highly unique estuarine microbial and viral community[5–7]. In addition, most estuarine ecosystems are impacted by strong anthropogenic stresses[1]. Viruses play essential roles in marine ecosystems by mortality[8,9] and reprogramming the metabolic processes of hosts[10]. There is a great interest to investigate the genomic characteristics, evolutionary mechanisms, community composition and interactions of microorganisms and viruses in coastal environments[11,12]. While the abundance, distribution and function of prokaryotes or viruses in estuaries have been reported by using meta-omics approaches[13–18], few studies have investigated bacteria, archaea and viruses simultaneously and none has delineated the potential connections between the microbiome and the virome. Therefore, a holistic estuarine genome dataset recovering both microbiome and virome will allow the analysis of microbe-virus interactions in this unique ecosystem.

The Pearl River is the second largest river in China with an average annual discharge flux of about 3.5×10^{11} m^3 fresh water and 8.87×10^7 tons suspended sediment[19]. Locating in the most densely industrialized and urbanized region in China, the Pearl River is heavily impacted by human activities including agricultural irrigation, industrial and domestic emissions and aquaculture[20,21]. While some ecological and genomic studies on the bacterial or viral communities at the Pearl River estuary (PRE) have been performed[13,15,17], none of them has produced a combined dataset including both the microbial hosts and the viruses. Such a dataset is therefore

[1]School of Environment, Harbin Institute of Technology, Harbin, China. [2]Shenzhen Key Laboratory of Marine Archaea Geo-Omics, Department of Ocean Science and Engineering, Southern University of Science and Technology (SUSTech), Shenzhen, China. [3]Daniel K. Inouye Center for Microbial Oceanography: Research and Education (C-MORE), University of Hawaii, Honolulu, Hawaii, USA. [4]Department of Ocean Science, The Hong Kong University of Science and Technology, Hong Kong, China. [5]State Key Laboratory of Marine Environmental Science, Fujian Key Laboratory of Marine Carbon Sequestration, College of Ocean and Earth Sciences, Xiamen University, Xiamen, Fujian, China. [6]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, China. [7]Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou, China. ✉e-mail: fanl@sustech.edu.cn

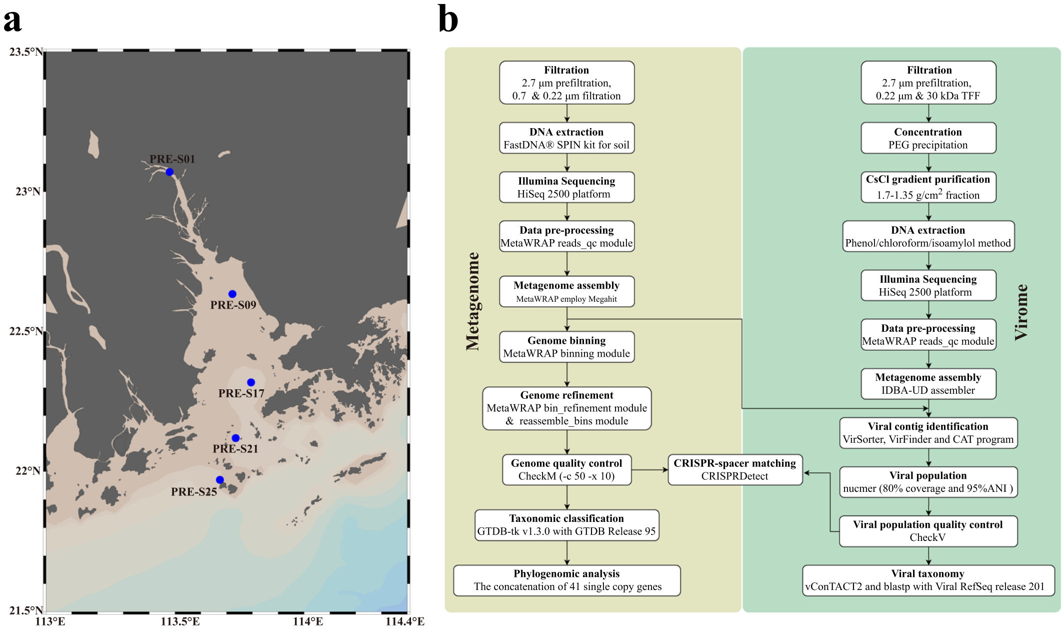

Fig. 1 Sampling sites in the PRE and methods used for this study. (**a**) yellow dots represent the sampling sites. (**b**) the study workflow in processing PRE metagenome sequences.

Station	Size-fraction	Raw read pairs	Read pairs after QC	Contigs (>1 kb)	Prokaryotic MAGs*	Viral contigs (>5 kb)
PRE-S01	0.7–2.7 μm	442,832,413	249,009,883	1,497,910	256	7,539
	0.22–0.7 μm	408,434,124	297,683,700	1,155,059	178	9,479
	<0.22 μm	81,475,964	66,065,092	114,587	Not Applicable	1,373
PRE-S09	0.7–2.7 μm	449,343,573	320,412,019	1,330,764	239	5,027
	0.22–0.7 μm	477,408,408	299,346,960	1,300,578	204	10,069
	<0.22 μm	18,803,045	15,540,200	26,853	Not Applicable	430
PRE-S17	0.7–2.7 μm	461,596,030	322,152,520	920,756	191	3,316
	0.22–0.7 μm	472,049,471	312,545,342	702,047	134	9,404
	<0.22 μm	22,655,869	18,478,415	31,626	Not Applicable	1,043
PRE-S21	0.7–2.7 μm	462,296,138	319,291,227	945,976	182	5,025
	0.22–0.7 μm	475,167,759	306,620,589	929,998	143	16,029
	<0.22 μm	21,784,687	17,794,588	48,363	Not Applicable	1,919
PRE-S25	0.7–2.7 μm	462,183,037	292,078,554	1,160,660	182	6,963
	0.22–0.7 μm	467,795,025	295,708,591	1,001,020	169	12,710
	<0.22 μm	19,058,316	15,808,251	48,012	Not Applicable	2,258

Table 1. Summary of reads, contigs, MAGs and viral contigs of PRE metagenomes. *Completeness >50%, contamination <10%.

urgently demanded to unveil the dynamic and diverse biological processes coupling with physiochemical factors at this estuary.

Here, we sequenced 15 deep-sequencing metagenomes of surface water with three size-fractions collected at five sampling sites along the salinity gradient of the PRE in August 2016 (Fig. 1a). Seawater was filtered through cellulose membranes subsequently. The 0.7–2.7 μm and 0.22–0.7 μm fractions were used to produce particle-attached and free-living prokaryotic metagenomes, respectively. To collect the viral fraction, surface water was prefiltered by using filters of 2.7 μm and 0.22 μm pore-size, subsequently, and then concentrated with 30 kilodalton (kDa) pore-size filters by using tangential-flow filtration. Further concentration and purification were done via polyethylene glycol (PEG) precipitation and cesium chloride (CsCl) step-gradient ultracentrifugation (Fig. 1b). DNA was extracted from the cellular (0.7–2.7 μm and 0.22–0.7 μm) and viral (<0.22 μm) fractions for metagenomic sequencing.

Overall, 13,305,017 contigs were generated by assembling quality checked sequencing reads (Table 1). A total of 1,205 non-redundant metagenome assembled genomes (MAGs) with the estimated completeness ≥50% and contamination ≤10% were reconstructed based on multi-strategy binning according to the MIMAG criteria[22] (Supplementary Table 1). Phylogenomic analysis based on single-copy marker genes showed that these

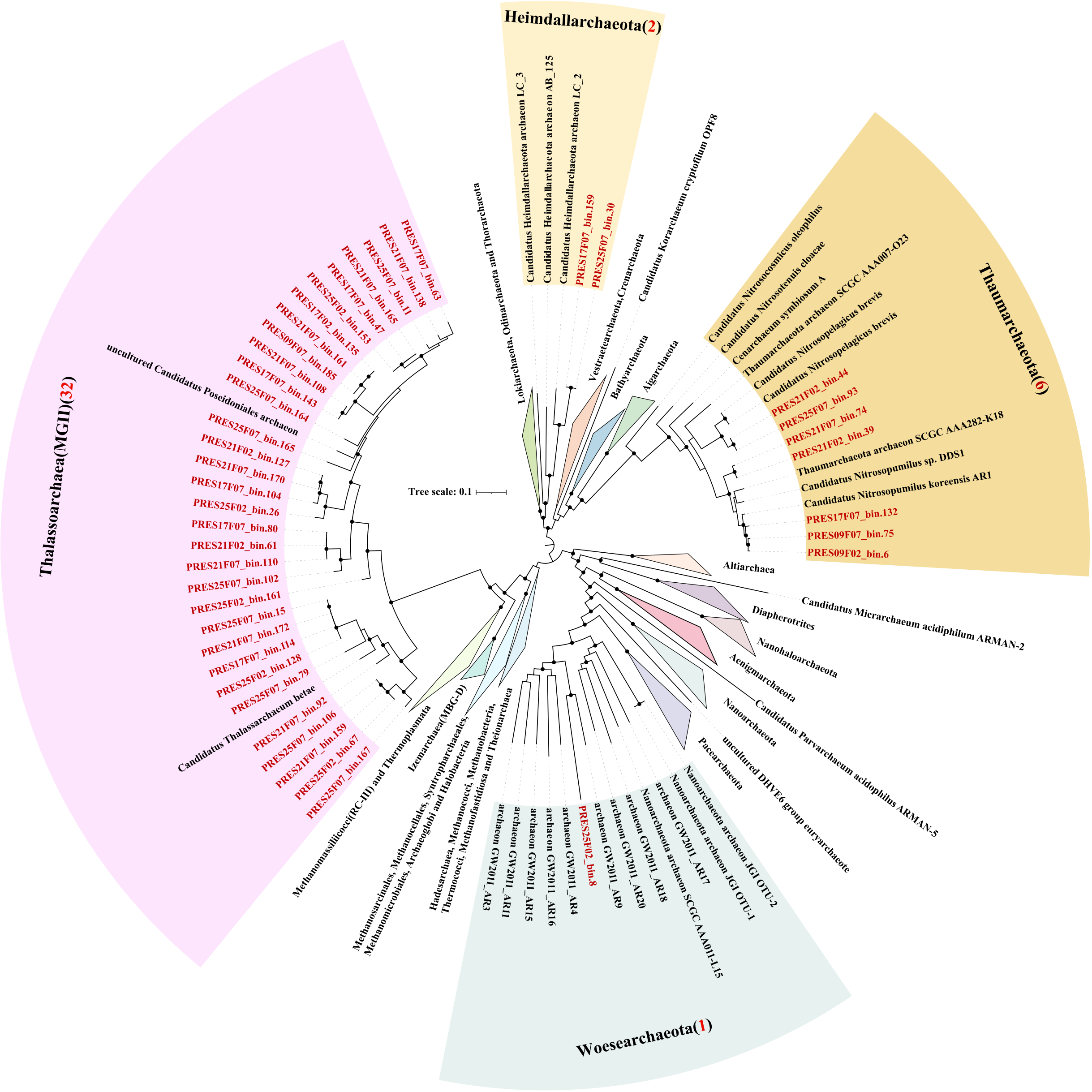
Heimdallarchaeota(2)
Candidatus Heimdallarchaeota archaeon LC_3
Candidatus Heimdallarchaeota archaeon AB_125
Candidatus Heimdallarchaeota archaeon LC_2
PRES17F07_bin.159
PRES25F07_bin.30
Lokiarchaeota, Odinarchaeota and Thorarchaeota
Vestraetearchaeota,Crenarchaeota
Candidatus Korarchaeum cryptofilum OPF8
Bathyarchaeota
Aigarchaeota
Thaumarchaeota(6)
Candidatus Nitrosocosmicus oleophilus
Candidatus Nitrosotenuis cloacae
Cenarchaeum symbiosum A
Thaumarchaeota archaeon SCGC AAA007-O23
Candidatus Nitrosopelagicus brevis
Candidatus Nitrosopelagicus brevis
PRES21F02_bin.44
PRES25F07_bin.93
PRES21F07_bin.74
PRES21F02_bin.39
Thaumarchaeota archaeon SCGC AAA282-K18
Candidatus Nitrosopumilus sp. DDS1
Candidatus Nitrosopumilus koreensis AR1
PRES17F07_bin.132
PRES09F07_bin.75
PRES09F02_bin.6
Altiarchaea
Candidatus Micrarchaeum acidiphilum ARMAN-2
Diapherotrites
Nanohaloarchaeota
Aenigmarchaeota
Candidatus Parvarchaeum acidophilus ARMAN-5
Nanoarchaeota
uncultured DHVE6 group euryarchaeote
Pacearchaeota
Nanoarchaeota archaeon JGI OTU-2
Nanoarchaeota archaeon JGI OTU-1
archaeon GW2011_AR17
Nanoarchaeota archaeon SCGC AAA011-L15
Nanoarchaeota archaeon GW2011_AR18
archaeon GW2011_AR20
archaeon GW2011_AR9
PRES25F02_bin.8
archaeon GW2011_AR4
archaeon GW2011_AR16
archaeon GW2011_AR15
archaeon GW2011_AR11
archaeon GW2011_AR3
Woesearchaeota(1)
Hadesarchaea, Methanococci, Methanobacteria,
Thermococci, Methanofastidiosa and Theionarchaea
Methanosarcinales, Methanocellales, Syntropharchaeales,
Methanomicrobiales, Archaeoglobi and Halobacteria
Izemarchaea(MBG-D)
Methanomassiliicocci(RC-III) and Thermoplasmata
Thalassoarchaea(MGII)(32)
PRES25F07_bin.167
PRES25F02_bin.67
PRES21F07_bin.159
PRES25F07_bin.106
PRES21F07_bin.92
Candidatus Thalassarchaeum betae
PRES25F07_bin.79
PRES25F02_bin.128
PRES17F07_bin.114
PRES21F07_bin.172
PRES25F07_bin.15
PRES25F02_bin.161
PRES25F07_bin.102
PRES21F07_bin.110
PRES21F02_bin.61
PRES17F07_bin.80
PRES25F02_bin.26
PRES17F07_bin.104
PRES21F07_bin.170
PRES21F02_bin.127
PRES25F07_bin.165
uncultured Candidatus Poseidoniales archaeon
PRES25F07_bin.164
PRES17F07_bin.143
PRES21F07_bin.108
PRES09F07_bin.185
PRES21F07_bin.161
PRES17F02_bin.135
PRES25F02_bin.153
PRES17F07_bin.47
PRES21F07_bin.165
PRES25F07_bin.11
PRES21F07_bin.138
PRES17F07_bin.63
Tree scale: 0.1

Fig. 2 Phylogenomic analysis of archaeal MAGs. The maximum likelihood tree was reconstructed based on the concatenation of 41 single copy marker genes spanning a set of 41 MAGs (in red) obtained in this study and a set of 163 reference genomes (in black). The number of MAGs discovered in this study in each phylum is indicated in the parenthesis after the phylum name. The bootstrap values >0.9 are shown as dots on nodes. The tree is unrooted. Source data are provided as a Source Data file.

MAGs belonged to 32 bacterial and four archaeal phyla according to the Genome Taxonomy Database (GTDB) taxonomy[23] (Fig. 2, 3). We found that 24.8% and 86.8% of total MAGs did not have close relatives at genus and species level based on 95% average nucleotide identity (ANI). A total of 78,502 non-redundant viral contigs were predicted from the cellular microbiomes (0.2–2.7 μm) and viromes (<0.2 μm). They were then clustered into 56,289 viral populations[24–26]. Taxonomic classification of viral populations was performed based on closest relative affiliation[24] (Supplementary Table 2). Only 15.3% populations could be assigned according to the RefSeqVirus database leaving the rest majority unclassified. A total of 15 viral families were identified including ssDNA, dsDNA and ssRNA viruses and the primary group belongs to order *Caudovirales* (Table 2). Virus-host pair prediction was performed based on clustered regularly interspaced short palindromic repeats (CRISPR) -spacer matching and 11 virus-host pairs were identified (Fig. 4). Among them, an *Acinetobacter junii* and a Rickettsiales bacterium were found being infected by more than one type of virus.

All of the primary contigs, non-redundant MAGs and viral-like contigs have been deposited in the National Center for Biotechnology Information (NCBI) BioProject database and the *figshare* website. The microbial and viral genomes provided here suggest great biological diversity in the PRE ecosystems. This combined dataset

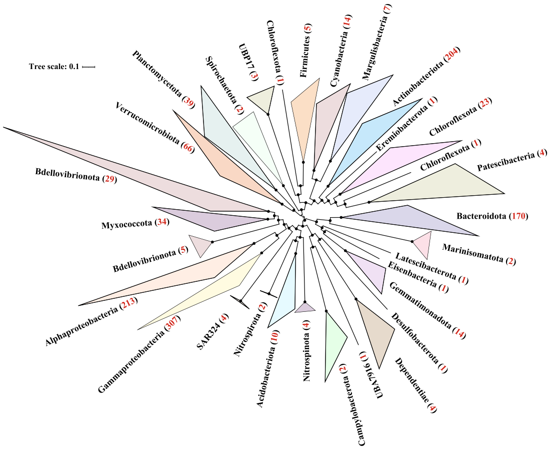

Fig. 3 Phylogenomic analysis of bacterial MAGs. The maximum likelihood tree was reconstructed based on the concatenation of 41 single copy markers. The number of MAGs discovered in this study in each phylum is indicated in the parenthesis after the phylum name. Number of MAGs from the PRE metagenomes in each phylum or class are indicated in between parenthesis in red. The bootstrap values >0.9 are shown as dots on nodes. The tree is unrooted. Source data are provided as a Source Data file.

Major Taxa	PRES01V*	PRES01FL**	PRES01PA***	PRES09V	PRES09FL	PRES09PA	PRES17V	PRES17FL	PRES17PA	PRES21V	PRES21FL	PRES21PA	PRES25V	PRES25FL	PRES25PA
Unassigned	1,301	8,223	6,572	398	8,620	4,464	967	7,582	2,821	1,765	13,416	4,280	2,086	10,195	5,715
Myoviridae	9	933	712	4	1,207	459	17	1,512	314	76	1,926	520	76	1,840	833
Siphoviridae	29	159	105	18	117	55	19	106	93	23	167	75	20	151	106
Podoviridae	24	73	56	10	45	14	22	74	20	50	233	46	63	170	63
Autographiviridae	1	7	6	0	11	2	18	53	22	5	134	37	11	198	168
Phycodnaviridae	0	36	32	0	37	12	0	50	32	0	97	45	2	111	54
Demerecviridae	0	31	31	0	25	17	0	22	6	0	37	12	0	32	12
Mimiviridae	0	12	23	0	5	2	0	1	1	0	12	7	0	5	2
Iridoviridae	0	1	0	0	0	1	0	1	0	0	3	1	0	3	5
Herelleviridae	1	2	1	0	2	1	0	3	0	0	2	1	0	2	0
Microviridae	0	0	0	0	0	0	0	0	6	0	0	1	0	2	2
Lavidaviridae	8	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Inoviridae	0	1	1	0	0	0	0	0	1	0	1	0	0	1	1
Poxviridae	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Metaviridae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Marseilleviridae	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2. Nonredundant contigs of abundant viral populations in samples. *V, viral fraction (siz <0.22 μm); **FL, free-living cellular fraction (size 0.22–0.7 μm); ***PA, particulate-associated cellular fraction (size 0.7–2.7 μm).

allows for systematic study on microbial-virial interactions including the regulatory mechanisms of viruses in manipulating estuarine biogeochemistry under anthropogenic impacts.

Methods

Sampling, DNA extraction and sequencing. Bacterial, archaeal and viral sample collection and particle size-based fractionation was done by filtration[27]. To obtain the cellular fractions, about 500 L surface water (0.5–1.0 m in depth) was collected at each sampling site in PRE in August 2016 within three days (Fig. 1a, Table 3). The water samples were first filtered through 2.7 μm pore-size glass fiber filters (Shanghai Mosutech, Shanghai, China) to remove large particles and the filtrates were then successively filtered through 0.7 and 0.22 μm pore-size membrane filters (Pellicon cartridge, Millipore Corp., Billerica, MA, USA) to collect particulate associated and free-living microbial cells, respectively. The filters were stored in liquid nitrogen temporarily on board and then transferred to −80 °C freezers when back to laboratory for long-term storage until further processing. To collect

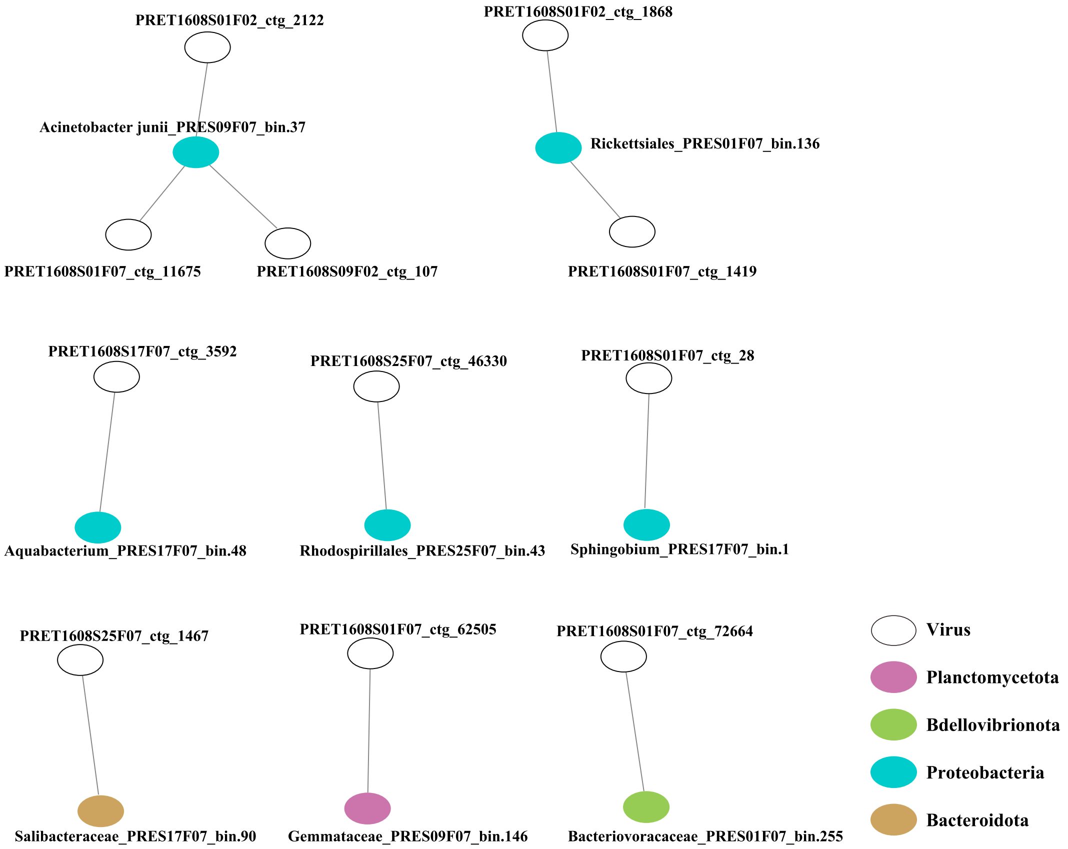

Fig. 4 Network analysis of virus-host pairs. The hollow circles represent the viruses. The solid circles represent the prokaryotic hosts. The colors indicate the phyla of the hosts.

Station	PRE-S01	PRE-S09	PRE-S17	PRE-S21	PRE-S25
Latitude (°N)	23.0717	22.634742	22.319517	22.120133	21.9717
Longitude (°E)	113.479733	113.722569	113.795633	113.735033	113.67375
Sampling time	2016.08.23 12:50	2016.08.22 12:28	2016.08.21 15:58	2016.08.21 11:21	2016.08.20 12:00
Temperature (°C)	30.8	31.5	28.6	27.3	27.4
Salinity (PSU)	0.12	1.17	11.24	22.21	28.05
pH	6.5	6.77	7.42	7.83	8.06
DO (μM)	39.38	123.16	138.75	138.44	147.81
DOC (μM)	182.67	137.5	113.08	99	78.58
TDN (μM)	361.29	197.36	165.71	87.64	39.79
NO^{3-} (μM)	94.69	132.12	82.55	32.17	28.9
NO^{2-} (μM)	23.49	1.628	19.09	11.7	9.5

Table 3. Sampling locations and bulk properties of PRE surface water.

viral particles, 200 L prefiltered seawater was further filtered through 2.7 μm and 0.2 μm pore-size membrane filters. A tangential-flow filtration 30 kDa cartridge was (0.5 m^2 Pellicon cartridge, Millipore Corp., Billerica, MA, USA) applied to increase viral particle concentration till a final liquid volume of 2 L and the liquid was kept at 4 °C till further process[28]. Physiochemical measurements of water and the methods to generate these measurements have been published by He *et al.*[23]. The measurements are also available in Table 3.

DNA was extracted from the 0.2 and 0.7 μm pore-size membrane filters by using the FastDNA® SPIN kit for soil (MP Biomedicals, Solon, OH, USA) following the manufacturers' instructions. For virome samples, a series of enrichment operations were applied to increase the concentration of the virial suspension[28] (Fig. 1b). Firstly, PEG8000 (10% w/v) was dissolved in DNase I (Sigma-Aldrich) treated viral concentrate and incubated at 4 °C overnight to precipitate viral particles. The PEG pellet was resuspended after centrifugation (10, 000 × g for 1 h) and then purified by CsCl density gradient ultracentrifugation (1.7, 1.5, and 1.35 g/mL CsCl layers). After centrifugation, viral like particles was concentrated in 1.5–1.35 g/mL CsCl layers according to the physical properties of various virions. After collection and purification, a phenol-chloroform extraction following the ethanol precipitation method was applied to extract viral genomic DNA[14,28].

The extracted prokaryotic and viral DNA were fragmented by sonication to a size of 350 bp. The DNA fragments were then end-polished, A-tailed, and ligated with the full-length adaptor to construct TruSeq metagenome libraries. Libraries were analyzed for size distribution using the Agilent2100 Bioanalyzer (Agilent, USA) and quantified using real-time PCR. They were then sequenced on an Illumina HiSeq 2500 platform at Novogene Bioinformatics Technology Co., Ltd. (Beijing, China) to generate 150 bp paired-end reads. The FASTQ files containing raw reads are available on NCBI. The overall study workflow is show in Fig. 1b.

Sequence quality check and assembly.

The reads_qc module of MetaWRAP (v1.2.1)[29] was applied for adaptor trimming and contamination removal for the raw sequencing reads to generate high-quality clean reads by calling Cutadapt[30] and FastQC[31] with the default parameters. Clean reads of the cellular fractions were assembled into contigs by using MetaWRAP employing megahit with k-mer values list of 21, 29, 39, 59, 79, 99, 119 and 141[29]. The IDBA-UD software (v1.1.3) was applied to assemble the viral metagenomes with default parameters[32]. Contigs of length longer than 1 kb were used for further analysis as suggested by the MIMAG and the MIUViG standards[22,26] (Table 1).

MAG generation, refinement, quality check and taxonomic annotation.

For each prokaryotic metagenome, MAGs were recovered by using the binning module and bin_refinement module of MetaWRAP[29]. First, the binning module of MetaWRAP employing METABAT[33] and CONCOCT[34] was applied to recover the original genome MAGs sets based on tetranucleotide frequencies and read coverage. These MAGs sets were pooled and dRep (v2.6.2) was performed to remove redundant MAGs[35]. The bin_refinement module of MetaWRAP was used to refine the MAGs to produce final MAGs. The completeness and contamination of archaeal and bacterial MAGs were estimated by running CheckM (v1.0.11)[36] (Supplementary Table 1). Taxonomic classification of the final MAGs was conducted by using GTDB-tk (v1.3.0, Release 95)[37] (Supplementary Table 1). MAGs are considered of the same species if they have ANI values larger than 95% by compared to a reference genome.

Phylogenomic analysis.

We used 41 single-copy marker proteins to infer the maximum likelihood trees of archaeal and bacterial MAGs[38,39], respectively. Specifically, putative coding DNA sequences for each draft genome were predicted by using Prodigal (v2.6.3; -m -p meta)[40]. Putative single copy genes of each MAGs were identified by using hmmsearch (HMMER v.3.1b2; -E 1E-5)[41] based on Hidden Markov Models (HMMS) described by Sunagawa *et al.*[39]. Amino acid sequences of these genes were aligned, respectively, by using Clustal Omega (v1.2.4)[42] and further automatically trimmed by using trimAL (v1.4.1; -automated1)[43]. The alignments of proteins were concatenated by using ScaFos (v1.2.5) and missing data were filled with gaps[44]. The phylogenomic tree of concatenated alignment was reconstructed by using IQ-TREE (v.2.0.3; -st AA -m LG + PMSF + G -B 1000 --bnni)[45] and visualized in the Interactive Tree of Life (iTOL, v.5.1.1)[46].

Viral contig identification, dereplication and taxonomic classification.

Following assembly, putative viral contigs were identified from contigs of all the three size fractions with length greater than 1.5 kb by using VirSorter (v1.0.6)[47] and VirFinder (v1.1)[48] as described by Gregory *et al.*[24]. First, contigs identified as 'lytic/prophage categories 1 and 2′ and 'circular' by VirSorter were assigned as viral contigs. The rest contigs of length >5 kb were kept for further classification. Among them, those identified as 'lytic/prophage categories 1,2′ by VirSorter, or as viruses by VirFinder with score >0.9 ($p < 0.05$) were assigned as viral contigs. Those identified as 'lytic/prophage category 3′ by VirSorter and as viruses by VirFinder with score 0.7–0.9 ($p < 0.05$) were also assigned as viral contigs. Those identified as 'lytic/prophage category 3' by VirSorter but not as viruses by VirFinder with score >0.7 ($p < 0.05$), and those identified as viruses by VirFinder with score 0.7–0.9 ($p < 0.05$) but not as 'lytic/prophage categories 1–3' by VirSorter were further analyzed through CAT[49] and only those having 40% genes classified as viruses were kept. In total, 97,003 viral contigs were identified. Redundancy of these contig sequences was removed by using CD-HIT at 99% identity (v4.6.8, −c 0.99 −aS 0.99)[50]. The resulting 78,502 non-redundant viral contigs were further grouped into 56,289 viral populations by using nucmer based on the criterion that virial contigs in the same population share 80% of their genes and have 95% average nucleotide identify as previously described[51,52] (Fig. 1b). CheckV (v0.8.1) was used to determine the completeness and quality of the identified viral populations[53] (Supplementary Table 3). We used VirSorter to identify prophages by the *de novo* predictions of categories 4 and 5[47].

Taxonomic classification of viral populations was performed with a complementary approach by using vConTACT2[54] and blastp[55]. First, the ORFs of each population were derived by using prodigal[40]. Second, the protein sequences of population contigs >10 kb were analyzed by using vConTACT2 with Viral RefSeq release 201 based on genome gene-sharing profiles. Then, family level taxonomy of the remaining population including those that could not be assigned by vConTACT2 were further defined by closest relative affiliation using blastp against the Viral RefSeq database with the following principle: identity ≥30%, bit-score ≥50, and E value ≤0.001. Only the population with more than half of proteins assigned to the same viral family was considered as a viral family (Supplementary Table 2).

Host prediction of viral sequences.

In order to link viral contigs to their putative microbial hosts, CRISPR spacers in MAGs were identified by using CRISPRDetect (v2.5)[56]. Spacer sequences were then matched to viral contigs by using fuzznuc[57]. Host and virus infection networks were reconstructed in Cytoscape (v3.8.0)[58].

Data Records

Raw reads generated in this study have been deposited in the National Center for Biotechnology Information BioProject database with the project ID PRJNA763043[59]. Contigs, MAGs, viral genomes and source data files including the genome trees and associated amino acid alignments have been deposited in the *figshare* website[60]. A full copy of this dataset is also available in the National Omics Data Encyclopedia (https://www.biosino.org/node/) with the project ID OEP001662[61].

Technical Validation

Additional technical validation should be applied by researchers to confirm the accuracy of draft MAGs and VAGs used for specific downstream purposes.

Code availability

All versions of third-party software and scripts used in this study are described and referenced accordingly in the Methods sub-sections for ease of access and reproducibility.

Received: 1 October 2021; Accepted: 12 January 2022;
Published online: 14 February 2022

References

1. Best, J. Anthropogenic stresses on the world's big rivers. *Nat. Geosci.* **12**, 7–21 (2018).
2. Carvalho, T. M. & Fidelis, T. The relevance of governance models for estuary management plans. *Land Use Policy* **34**, 134–145 (2013).
3. Zapata, C., Puente, A., Garcia, A., Garcia-Alba, J. & Espinoza, J. Assessment of ecosystem services of an urbanized tropical estuary with a focus on habitats and scenarios. *PLoS One* **13**, e0203927 (2018).
4. Campbell, B. J. & Kirchman, D. L. Bacterial diversity, community structure and potential growth rates along an estuarine salinity gradient. *ISME J.* **7**, 210–220 (2013).
5. Sun, M. *et al*. Uncultivated viral populations dominate estuarine viromes on the spatiotemporal scale. *mSystems* **6**, e01020–20 (2021).
6. Liu, L., Cai, L. & Zhang, R. Co-existence of freshwater and marine T4-like myoviruses in a typical subtropical estuary. *FEMS Microbiol. Ecol.* **93** (2017).
7. Xie, W. *et al*. Localized high abundance of Marine Group II archaea in the subtropical Pearl River Estuary: implications for their niche adaptation. *Environ. Microbiol.* **20**, 734–754 (2018).
8. Chen, X., Ma, R., Yang, Y., Jiao, N. & Zhang, R. Viral regulation on bacterial community impacted by lysis-lysogeny switch: a microcosm experiment in eutrophic coastal waters. *Front. Microbiol.* **10**, 1763 (2019).
9. Manea, E. *et al*. Viral infections boost prokaryotic biomass production and organic C cycling in hadal trench sediments. *Front. Microbiol.* **10**, 1952 (2019).
10. Warwick-Dugdale, J., Buchholz, H. H., Allen, M. J. & Temperton, B. Host-hijacking and planktonic piracy: how phages command the microbial high seas. *Virol. J.* **16**, 15 (2019).
11. Tyson, G. W. *et al*. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
12. Anantharaman, K. *et al*. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
13. Ni, Z., Huang, X. & Zhang, X. Picoplankton and virioplankton abundance and community structure in Pearl River Estuary and Daya Bay, South China. *J. Environ. Sci. (China)* **32**, 146–154 (2015).
14. Cai, L., Zhang, R., He, Y., Feng, X. & Jiao, N. Metagenomic analysis of virioplankton of the subtropical Jiulong River estuary, China. *Viruses* **8**, 35 (2016).
15. Zhang, C. *et al*. The communities and functional profiles of virioplankton along a salinity gradient in a subtropical estuary. *Sci. Total Environ.* **759**, 143499 (2021).
16. Liu, Y. *et al*. Differences in metabolic potential between particle-associated and free-living bacteria along Pearl River Estuary. *Sci. Total Environ.* **728**, 138856 (2020).
17. Fortunato, C. S. & Crump, B. C. Microbial gene abundance and expression patterns across a river to ocean salinity gradient. *PLoS One* **10**, e0140578 (2015).
18. Natascha, S. & Meinhard, S. Composition and dynamics of particle-associated and free-living bacterial communities in the Weser estuary, Germany. *Aquat. Microb. Ecol.* **30**, 221–237 (2003).
19. Zhang, J. *et al*. The subtropical Zhujiang (Pearl River) Estuary: Nutrient, trace species and their relationship to photosynthesis. *Estuarine Coastal and Shelf Science* **49**, 385–400 (1999).
20. Li, Y. *et al*. Distribution, seasonality, and fluxes of dissolved organic matter in the Pearl River (Zhujiang) estuary, China. *Biogeosciences* **16**, 2751–2770 (2019).
21. Peng, X. *et al*. Persistence, temporal and spatial profiles of ultraviolet absorbents and phenolic personal care products in riverine and estuarine sediment of the Pearl River catchment, China. *J. Hazard Mater.* **323**, 139–146 (2017).
22. Bowers, R. M. *et al*. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
23. He, C. *et al*. Molecular composition and spatial distribution of dissolved organic matter (DOM) in the Pearl River Estuary, China. *Environ. Chem.* **17**, 240–251 (2020).
24. Gregory, A. C. *et al*. Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109–1123 e1114 (2019).
25. Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ.* **5**, e3817 (2017).
26. Roux, S. *et al*. Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).
27. Karsenti, E. *et al*. A holistic approach to marine eco-systems biology. *PLoS Biol.* **9**, e1001177 (2011).
28. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).
29. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
30. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
31. Brown, J., Pirrung, M. & McCue, L. A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* **33**, 3137–3139 (2017).

32. Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
33. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ.* **7**, e7359 (2019).
34. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
35. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
36. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
37. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2019).
38. Martinez-Gutierrez, C. A. & Aylward, F. O. Phylogenetic signal, congruence, and uncertainty across bacteria and archaea. *Mol. Biol. Evol*, msab254 (2021).
39. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).
40. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
41. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–37 (2011).
42. Sievers, F. & Higgins, D. G. in *Multiple Sequence Alignment Methods* (ed. Russell, D. J.). Ch. 6, 105–116 (Humana Press, 2014).
43. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
44. Roure, B., Rodriguez-Ezpeleta, N. & Philippe, H. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol. Biol.* **7**(1), 1–12 (2007).
45. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
46. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
47. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ.* **3**, e985 (2015).
48. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
49. von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 217 (2019).
50. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
51. Brum, J. R. *et al.* Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
52. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
53. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
54. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
55. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
56. Biswas, A., Staals, R. H., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics.* **17**, 356 (2016).
57. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the european molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).
58. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
59. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP337539 (2021).
60. Xu, B. *et al.* Pearl river estuary microbiome and virome. *figshare* https://doi.org/10.6084/m9.figshare.16618255.v7 (2021).
61. *National Omics Data Encyclopedia* https://www.biosino.org/node/project/detail/OEP001662 (2021).

Acknowledgements

This study was supported by National Natural Science Foundation of China (Nos. 91951120, 91851210, 42141003), the State Key R&D Project of China Grant (No. 2018YFA0605800), the Shenzhen Key Laboratory of Marine Archaea Geo-Omics, Southern University of Science and Technology (ZDSYS201802081843490), the Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (No. K19313901), and the Project of Educational Commission of Guangdong Province of China (No. 2020KTSCX123). Computation in this study was supported by the Centre for Computational Science and Engineering at the Southern University of Science and Technology.

Author contributions

R.Z., L.F. and C.Z. conceived this study. B.X., F.L. and L.C. collected the samples and extracted DNA. B.X. analyzed the metagenome data, produced the genomes and conducted all other analyses. B.X., R.Z., L.F. and C.Z. interpreted the results and drafted the manuscript. All authors contributed to the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-022-01153-4.

Correspondence and requests for materials should be addressed to L.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.