Using nanopore reads to recover the complete genomes of ocean viruses without the need for assembly

Over 1,400 high-quality full-length virus genome sequences obtained in single reads from a single MinION run using an assembly-free bioinformatics analysis pipeline

Contact: **publications@nanoporetech.com** More information at: **www.nanoporetech.com** and **publications.nanoporetech.com**

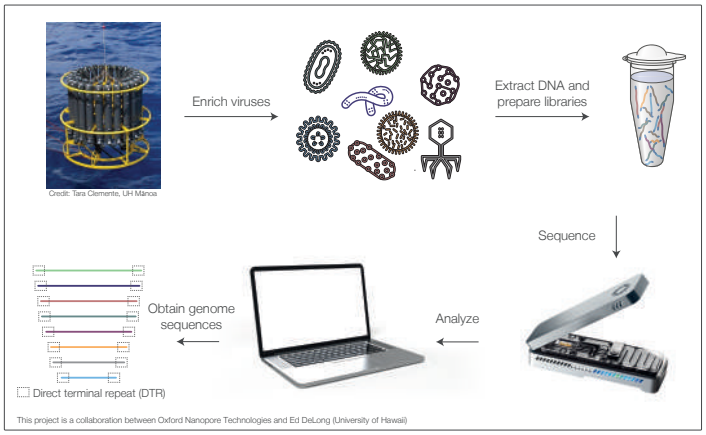

Fig. 1 End-to-end workflow for obtaining complete viral genome sequences without assembly

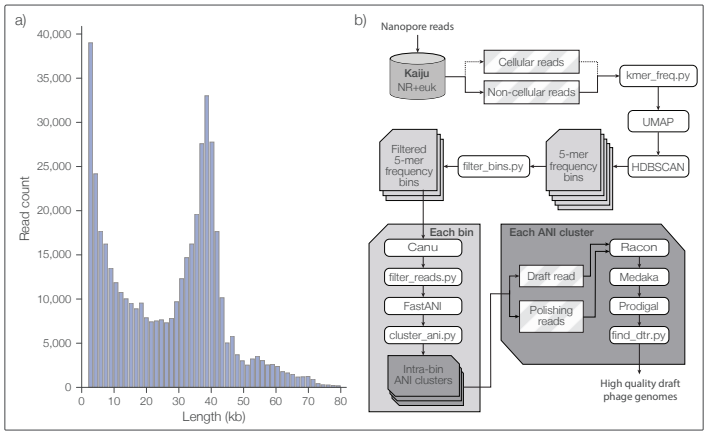

Fig. 2 Deriving viral genome sequences from long reads a) read lengths b) analysis pipeline

Direct recovery of complete viral genome sequences from environmental samples

Viruses are the most abundant biological entities on Earth and play key roles in host ecology, evolution and horizontal gene transfer. The inherent genetic complexity of virus populations poses technical difficulties for recovering complete virus genomes. To address these challenges, we developed an assembly-free, single-molecule nanopore sequencing approach enabling direct recovery of viral genome sequences from environmental samples (Fig. 1). Water was sampled at three different depths from the Pacific Ocean near Hawaii and viral particles were enriched from each fraction. We prepared libraries, sequenced these on a MinION flowcell and analysed the resulting data using a bespoke bioinformatics pipeline (Fig. 2).

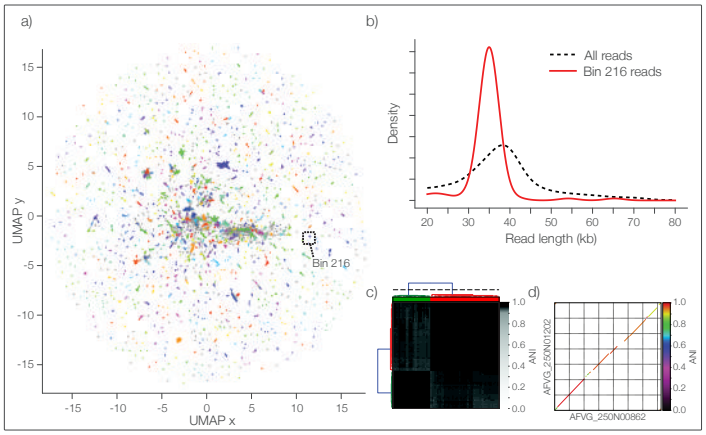

Fig. 3 Binning by k-mer frequencies using *UMAP*

Bioinformatics workflow for assembly-free derivation of full-length genome sequences

The read lengths generated from a single MinION flow cell show peaks corresponding to the expected genome sizes of abundant marine viruses (30–40 kbp, Fig. 2a). We therefore designed an assembly-free bioinformatic pipeline to extract genomes from viral metagenomic sequencing reads (Fig. 2b). We first applied a coarse taxonomic annotation to separate cellular from non-cellular reads. Next, normalized 5-mer frequencies were computed for each read. Dimensionality reduction and clustering tools were then used to embed these 5-mer frequency vectors in two dimensions and call read bins. Additional clustering based on pairwise average nucleotide identity (ANI) within each bin provided a draft read for polishing using the remaining reads.

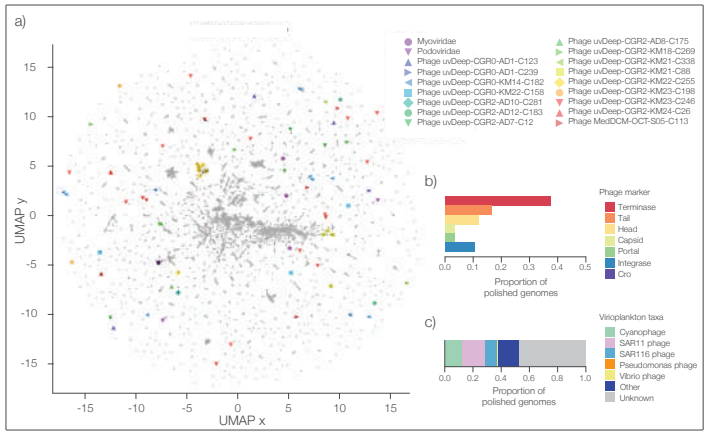

Fig. 4 Comparison of the ocean phage genome sequences to known reference genomes

Reference-free read binning resolves micro-diversity in phage genomes

The dimensionality-reduction tool *UMAP* was used to create a two-dimensional embedding of 5-mer frequency features for each read in the 250 m seawater sample and read bins were called (Fig. 3a). Bin 216 is representative of many other bins in that the read length distribution revealed enrichment for reads of a specific length, suggesting that these reads fully span a virus genome (Fig. 3b). The genome-scale reads within Bin 216 were further clustered by pairwise average nucleotide identity (ANI) values to reveal strain-level differences in virus reads (Fig. 3c). Polished draft genomes from each ANI cluster share large regions of high sequence identity (>95%), although several regions contain significantly diverged sequences (Fig. 3d).

Most phage genome sequences are absent from existing databases and may be novel

Taxonomic annotation of read bins in the two-dimensional embedding of 5-mer frequencies shows that the sequences in some of the bins are matches (i.e. >50% of the binned reads have this annotation) to known marine phages or other viral genomes. However, the large majority of unlabelled bins suggests a level of phage genome diversity that is not well represented in existing sequence databases (Fig. 4a). Fig. 4b shows the proportion of polished genomes with PFAM annotations (bit score > 30) to common virus and prophage marker genes. The proportions of polished genomes with one or more protein matches to common virioplankton taxa in the RefSeq84 database is shown in Fig. 4c.