geneious prime

Application Note

Identification of Zika virus with Geneious Prime

Identification of Zika virus with Geneious Prime

Introduction

A whole pipeline for the analysis of metagenomic Next- Generation Sequencing (NGS) data can be carried out with Geneious Prime[1]. Analysis of complex microbiomes or detection of low-concentration components in mixed genomes can be performed. The latter is useful in the detection of pathogens, such as Zika virus (ZIKV) from plasma, as we describe here.

ZIKV is a positive-strand RNA virus belonging to the Flaviviridae family and is mainly spread by *Aedes mosquitoes*[2], but also transmitted sexually or by blood transfusion[3] (**Fig. 1**).

ZIKV was firstly identified in 1947 in a sentinel rhesus monkey in the Zika valley (Uganda[4]) and subsequently isolated from mosquitoes (1948[4]) and humans (1952[5]). However, it was relatively unknown until it reemerged in 2007, causing an outbreak in Micronesia[6], followed by epidemics in Oceania in 2013-2014[7] and the most recent major outbreak across the Americas[8].

It has been recently identified as a public health emergency by the WHO[9] after the 2015 epidemic, primarily affecting Brazil, due to the rapid spread and threat to the human population.

Zika infection in adults can pass undetected, facilitating the disease diffusion in the population. However, it has been associated with severe brain anomalies and microcephaly in prenatal infants of affected mothers, and consequently, intellectual disability, poor motor function, poor speech, abnormal facial features, seizures, and dwarfism[10]. A sensitive and accurate detection of the ZIKV could help contain the disease, so many efforts are focussed in that direction. A publicly available workflow for analyzing high-depth ZIKV sequencing data is already available for Geneious Prime, developed in collaboration with the O'Connor lab[11].

Here we describe a different pipeline in Geneious Prime, which allows to accurately identify and analyse low-concentration of ZIKV from high-throughput metagenomic data proceeding from a patient with suspected infection.

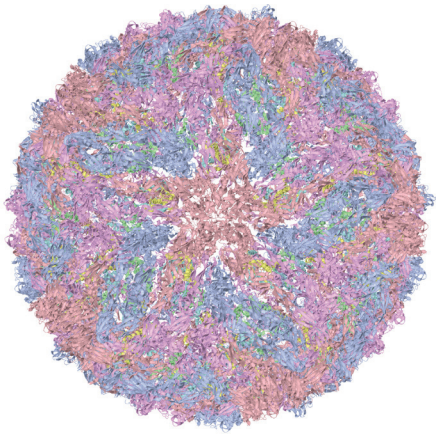

Figure 1. ZIKV structure. Full capsid reconstruction of the ZIKV at its mature and infectious state (PDB 5IRE and EMD-8116), from the 3.8 angstrom resolution cryo Electron Microscopy structure[12] downloaded from the VIPERdb (Entry key 929), imported and visualised in 3D in Geneious, and exported as image.

Data

Single-end Illumina MiSeq data (634,815 reads) were downloaded from the NCBI SRA database, accession number SRR3502889, in FASTQ format. This dataset comprises the RNA metagenome extracted and reverse transcribed from serum of one patient from Bahia, Brazil, suspected of ZIKV infection. These data were submitted from the University of California, San Francisco and Abbott Laboratories Inc., as part of a large project aiming at tracking the evolution and spread of the virus (BioProject PRJNA329513).

Analysis with Geneious Prime

Pre-processing

The FASTQ file was imported into Geneious R10 and reads were processed with 'Annotate & Predict' 'Trim using BBDuk'[13], using the default settings with the options checked to 'Trim Adapters', 'Trim Low Quality (Minimum Quality of 25)', 'Discard short reads (Minimum Length 40 bp)', and 'Keep original order'. The stringent setting of minimum read length of 40 bp was applied to only recover sequences long enough for reliable taxonomic identification. Duplicates were removed after trimming (591,248 reads) using 'Sequence' -> 'Remove duplicate reads'[14], obtaining a total of 398,031 reads post-processing.

Mapping to the reference

The processed reads were then aligned to a Brazilian ZIKV[15] (ZikaSPH2015/Brazil/2015, Accession KU321639) using Geneious Prime mapper. The default 'Medium Sensitivity/Fast' sensitivity setting were modified by selecting the 'Custom Sensitivity' and activating the option 'Minimum overlap Identity' 88%. This identity filter was activated to reduce the probability of aligning sequences from phylogenetically

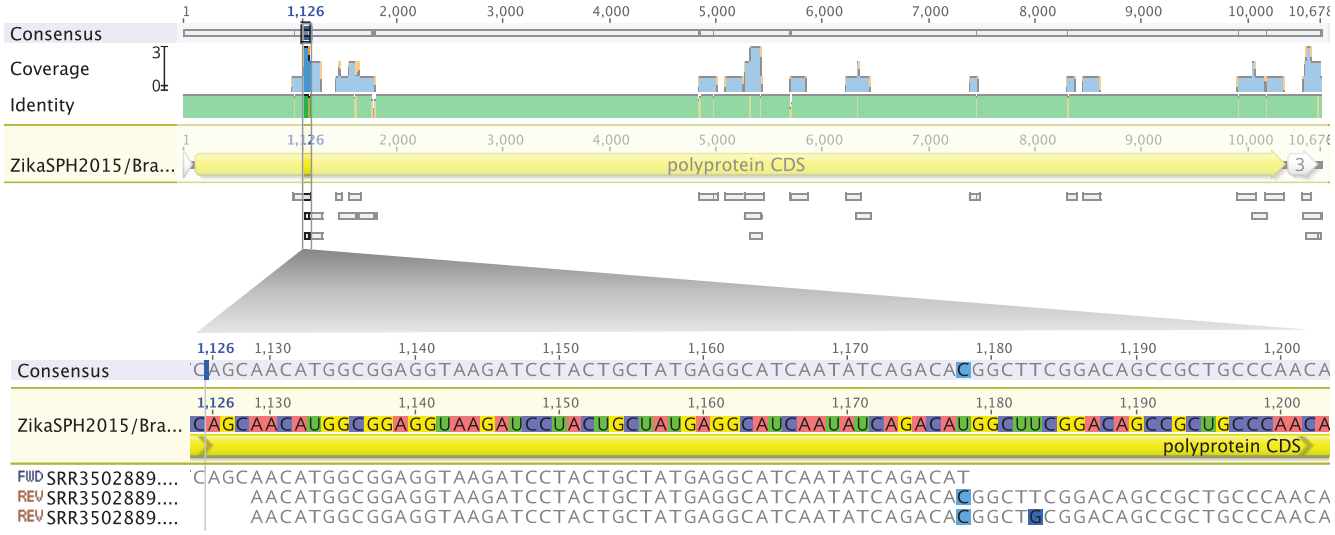

Figure 2. Mapping to reference results in the Contig view. Alignment visualisation in the interactive viewer available in Geneious, zooming in one of the aligned regions is shown. Disagreements to the reference are highlighted, following the quality colouring scheme. Heterozygous positions can be due to sequencing errors (probably the blue G here, with low base quality), de novo mutations, or presence of more than one ZIKV.

distant species. This value was rounded down from the minimum pairwise identity between ZIKV genomes (87.8%), from a representative data set[15] of both pre-epidemic and epidemic strains. Pairwise identity was calculated in Geneious Prime after aligning the selected dataset and exploring the Statistics tab within the Alignment View (see section 'Phylogenetic analysis within the Zika virus variability').

Twenty-four reads aligned to the reference sequence, as shown in the Contig View (**Fig. 2**). The statistic tab provided the following information: pairwise identity of 98.4%, mean coverage of 0.3, covering the 2,384 bp of the reference sequence at least once (22.3% of the total 10,678 bp).

A consensus sequence was constructed based on the highest quality threshold (as set in the Display tab in the Contig view) and extracted using 'Tools' -> 'Generate consensus sequence', calling "?" (base or gap) in absence of coverage.

BLAST search

BLAST was used to confirm the specificity of the reads aligned to the Zika reference. A batch BLAST search of the 24 nucleotide sequences was performed using the Megablast algorithm against the non-redundant (nr) nucleotide database of GenBank, EMBL, DDBJ, PDB and RefSeq. The matching region of the best hit per read was retrieved.

The BLAST search provided matches with ZIKV for all sequences with pairwise identity spanning from 100% (in 66.6% of cases) to 98.8% with E values lower than e-13 (Table 1).

Accession	Grade	% Pairwise Identity	E Value	Query coverage	Sequence Length	Organism	Hit start	Hit end
KX856011	100.00%	100.00%	2.37E-77	100.00%	161	Zika virus	10,663	10,503
KX856011	100.00%	100.00%	3.55E-54	100.00%	119	Zika virus	10,016	10,134
KX856011	100.00%	100.00%	3.98E-37	100.00%	88	Zika virus	7,359	7,446
KX856011	100.00%	100.00%	2.37E-77	100.00%	161	Zika virus	8,581	8,421
KX856011	100.00%	100.00%	5.74E-41	100.00%	95	Zika virus	1,538	1,632
KX827309	100.00%	100.00%	7.90E-56	100.00%	122	Zika virus	6,145	6,266
KX811222	100.00%	100.00%	2.37E-77	100.00%	161	Zika virus	5,423	5,263
KX811222	100.00%	100.00%	8.45E-77	100.00%	160	Zika virus	1,014	1,173
KX811222	100.00%	100.00%	2.37E-77	100.00%	161	Zika virus	5,084	5,244
KX101066	100.00%	100.00%	2.37E-77	100.00%	161	Zika virus	10,219	10,059
KX856011	100.00%	100.00%	9.68E-27	100.00%	69	Zika virus	10,489	10,557
KX856011	99.70%	99.40%	1.10E-75	100.00%	161	Zika virus	1,274	1,114
KU820897	99.70%	99.40%	1.10E-75	100.00%	161	Zika virus	9,892	10,052
KX856011	99.60%	99.20%	1.81E-57	100.00%	128	Zika virus	10,650	10,523
KX842449	99.60%	99.10%	3.33E-49	100.00%	113	Zika virus	6,296	6,408
KX856011	99.50%	99.00%	1.70E-41	100.00%	99	Zika virus	5,300	5,398
KX856011	99.40%	98.80%	1.02E-32	100.00%	83	Zika virus	8,275	8,357
KX856011	99.40%	98.80%	5.12E-74	100.00%	161	Zika virus	1,274	1,114
KX811222	98.40%	100.00%	1.42E-74	96.89%	156	Zika virus	5,254	5,409
KX520666	98.40%	99.40%	1.84E-73	97.52%	157	Zika virus	4,775	4,931
KX856011	98.10%	100.00%	5.12E-74	96.27%	155	Zika virus	1,441	1,595
KX856011	93.20%	100.00%	4.02E-65	86.34%	139	Zika virus	5,833	5,695
KX856011	92.90%	100.00%	1.44E-64	85.71%	138	Zika virus	1,624	1,761
KX856011	75.00%	100.00%	2.82E-13	100.00%	44	Zika virus	1,454	1,411

Table 1. BLAST results. First hit retrieved for each of the 24 reads aligned to the reference sequence. The Grade is a specific summary statistic provided in Geneious Prime calculated from the pairwise identity, the E value and the query coverage, weighting 0.25, 0.5 and 0.25 respectively. This allows to sort hits such that the longest, highest identity hits are at the top.

Phylogenetic analysis within the Zika virus variability

A representative set[15] of 9 pre-epidemic and 16 epidemic strains of ZIKV, comprising 5 full genomes and 20 full polyproteins (covering 95% of the full-genome), were downloaded from the NCBI Nucleotide database directly from Geneious Prime, including the closest outgroup, the *Spondweni virus*, for phylogenetic reconstruction. These sequences were aligned using 'Align/Assemble' -> 'Multiple Alignment' using MUSCLE[16] with default parameters. The consensus sequence of assembled reads from the patient was then aligned to this dataset using the 'Consensus Align' function under 'Align/Assemble' -> 'Multiple alignment', using again MUSCLE[16] with default parameters.

The resulting alignment was then masked for all sites

containing ambiguities (including all missing data) using ‘Tools’ -> ‘Mask Alignment’ and selecting ‘Site containing:’ Ambiguities.

The alignment with the masked annotations was subsequently used as input for building a phylogeny based on the partial consensus sequence recovered. We used the ‘Tree’ button selecting a maximum likelihood (ML) approach using the plugin PhyML[17]. Masked sites were excluded, the Jukes Cantor mutation model was chosen following[15], and 100 bootstraps were applied. The resulting tree was edited using wwthe Tree view available in Geneious Prime and exported as image, see **Fig. 3**.

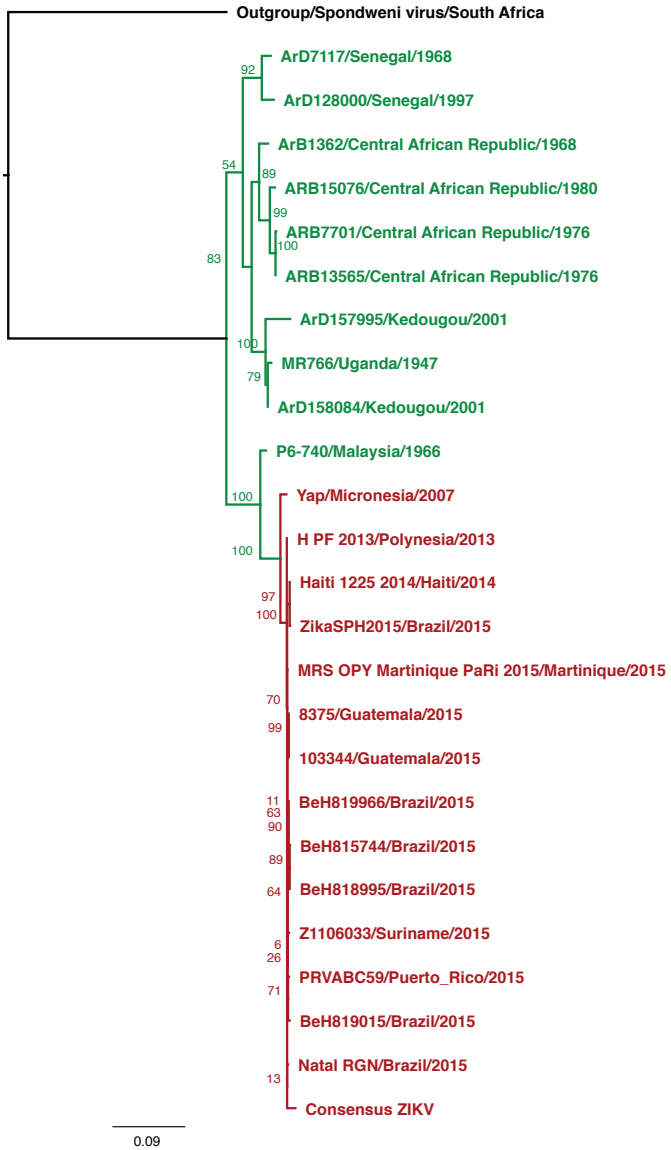

Figure 3. Maximum Likelihood tree. ML phylogenetic tree contextualising the tested sample (in bold, Consensus ZIKV) within the variability of pre-epidemic (green) and epidemic (red) ZIKV, including the Spondweni virus (black) as outgroup (dataset described in [15]). The phylogenetic reconstruction was performed in Geneious using the plugin PhyML [17]. ZIKV sequences are represented as strain/country/year; bootstrap support values above 80 are reported at nodes.

Summary

Geneious Prime provides an integrated user-friendly platform for performing NGS metagenomic analysis, such as the approach described here, which enables rapid early identification of ZIKV in suspected patients. We were able to unequivocally identify 24 Zika-derived Illumina reads, from an initial mixture of 634,815 reads, by aligning the dataset to a Zika reference genome using strict criteria, and confirming their suspected source by a BLAST search. Moreover, we could reconstruct the partial consensus sequence of the strain present in the patient tested. This allowed the phylogenetic investigation of the data, which showed the strain clustered within the variability of the epidemic strains of Brazil and the Americas.

References

1. Kearse, M. *et al.* Bioinformatics 28, 1647–1649 (2012).
2. Malone, R. W. *et al.* PLoS Negl. Trop. Dis. 10, e0004530 (2016).
3. Chen, L. H. & Hamer, D. H. Ann. Intern. Med. 164, 613–615 (2016).
4. Dick, G. W. A., Kitchen, S. F. & Haddow, A. J. Trans. R. Soc. Trop. Med. Hyg. 46, 509–520 (1952).
5. MacNamara, F. N. Trans. R. Soc. Trop. Med. Hyg. 48, 139–145 (1954).
6. Duffy, M. R. *et al.* N. Engl. J. Med. 360, 2536–2543 (2009).
7. Cao-Lormeau, V.-M. *et al*. Emerg. Infect. Dis. 20, 1085–1086 (2014).
8. Zanluca, C. *et al.* Mem. Inst. Oswaldo Cruz 110, 569–572 (2015).
9. World Health Organization. Zika strategic response framework and joint operations plan, January-June 2016.
10. Mlakar, J. *et al.* N. Engl. J. Med. 374, 951–958 (2016).
11. Zequencer: https://bitbucket.org/dhoconno/zequencer
12. Sirohi, D. et al. Science 352, 467–470 (2016).
13. Bushnell, B. http://seqanswers.com/forums/showthread.php?t=42776
14. Bushnell, B. https://sourceforge.net/projects/bbmap
15. Zhu, Z. *et al.* Emerg. Microbes Infect. 5, e22 (2016).
16. Edgar, R. C. Nucl. Acids Res. 32, 1792–1797 (2004).
17. Guindon, S. & Gascuel, O. Syst. Biol. 52, 696–704 (2003).

Published September 2017