DEN-IM: Dengue virus genotyping from amplicon and shotgun metagenomic sequencing

Supplemental Material

Dengue virus reference databases

We have compiled a database of 3858 complete DENV genomes obtained from the NIAID Virus Pathogen Database and Analysis Resource (ViPR) in October 2019 (1) (http://www.viprbrc.org/). The sequences were distributed unevenly throughout the four DENV serotypes, with DENV-1 being the most represented with 1636 sequences (42.72%), followed by DENV-2 with 1067 sequences (27.86%), DENV-3 with 807 sequences (21.07%), and DENV-4 with 320 sequences (8.36%). The selection criteria for the search were as follows: a) complete genome sequence only, b) human or mosquito host, c) collection year (1950-2018). Data available from all countries was included and duplicated sequences were removed and only the sequences with sub-type data were kept. A representative of DENV serotype 1 genotype III was introduced (EF457905, recovered from monkey) as no representatives were available with the search criteria used. This genotype is sylvatic and considered extinct (2,3). Additionally, any sample with IUPAC codes in the sequence provided were excluded.

In order to recover the maximum number of DENV reads from the input HTS data in the first mapping step (Figure 1), we maintained the database with the 3858 complete DENV genomes to retain as much diversity as possible. This database is referred as **DENV mapping database** and is available on GitHub at https://github.com/B-UMMI/DEN-IM/blob/master/ref/DENV_MAPPING_V3.fasta.

For typing purposes, overly similar sequences in the collection were removed from the database by clustering the sequences in each serotype at 98% nucleotide similarity with CD-HIT (4), leaving 161 representative sequences of all described DENV serotypes and genotypes, with 46 DENV-1 sequences (Table S6), 63 DENV-2 (Table S7), 25 DENV-3 (Tables S8) and 27 DENV-4 (Table S9). This database is referred as DENV typing database and is available on GitHub at https://github.com/B-UMMI/DEN-IM/blob/master/ref/DENV_TYPING_V3.fasta. This step is necessary to speed up the classification step for genotyping.

Phylogenetic analysis of typing collection was performed by aligning the full reference genomes with MAFFT (5), in auto mode and with automatic sequence orientation adjustment. A phylogenetic tree was inferred with RAxML (version 8.12.11) (6) using the GTR-Γ substitution model and 500 times bootstrap. Additionally, the same analysis was performed with the envelope protein (E) only, as this region has been used traditionally for sero- and genotyping (7–13), and continues to be the standard in many laboratories for genotyping. The resulting trees are available as supplemental material (Figures S4 to S7) and on Figshare (https://10.6084/m9.figshare.9331826).

The sequence JF459993 from the DENV-1 collection, as of April 2019, was annotated in ViPR as belonging to genotype IV, but in our analysis, it clustered within genotype I clade (Figure S4). The classification of DENV-1 I was also obtained from GenomeDetective Dengue Subtyping Tool (https://www.genomedetective.com/app/typingtool/dengue/), so we proceeded to alter the annotation of this particular sample (Table S6).

43 In order to harmonise dengue nomenclature, the system uses Roman-numeric labels to identify the
44 genotype, with the exception of Serotype 2 (Table S4), which used both Roman-numeric and
45 geographic origin due to the widespread adoption of the latter.

46

Workflow parameters

48 The short-read data is passed as input through the "--*fastq*" parameter, that by default is set to match all
49 files in the "fastq" folder that match the pattern "*_R{1,2}*". Both paired and single-end sequencing
50 data can be passed through with the "--*fastq*" parameter, as defined by the pattern used.
51 In the process to verify the integrity of the short-read raw sequencing data, the integrity of the input
52 files is assessed by attempting to decompress and read the files. An estimation of the depth of coverage
53 is also performed. By default, the input size ("–*genomeSize*") is set to 0.012 Mb and the minimum
54 coverage depth ("–*minCoverage*") is set to 10. If any input file is found to be corrupt, its progression
55 in the workflow is aborted.
56 In   the   FastQC   and   Trimmomatic   module,   FastQC
57 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is run with the parameters "–*extract* –
58 *nogroup* –*format fastq*". FastQC will inform Trimmomatic (14) on how many bases to trim from the
59 3'and 5' ends of the raw reads. By default, Trimmomatic uses the default set of Illumina adapters
60 provided with the workflow but this behaviour can be overwritten with the "–*adapters*" parameter. The
61 additional Trimmomatic parameters "–*trimSlidingWindow*", "–*trimLeading*", "–*trimTrailing*" and "–
62 -*trimMinLength*"can all be set to different values.
63 The removal of low complexity sequences is done with PrinSeq (15) using a custom parameter ("–
64 *pattern*"), which by default is set to the value "A 50%; T 50%; N 50%", removing sequences whose
65 content is at least half composed of a polymeric sequence (A, T or N).
66 To retrieve the reads that map to the DENV reference database, Bowtie2 (16) is run with default
67 parameters with the DENV mapping database as a reference. For paired-end data, the reads and their
68 mates that map to the reference are retrieved with "*samtools view -buh -F 12*" and "*samtools fastq*"
69 commands. In single-end reads, all mapped reads are retrieved with "*samtools view -buh -F 4*" and
70 "*samtools fastq*". The DENV mapping database can be altered with the "–*reference*" parameter, or
71 alternatively, a Bowtie2 index can be provided with the "–*index*" parameter. This allows for the
72 workflow to work with other databases obtained through public and owned DENV genomes. The
73 coverage estimation step is performed on the retrieved DENV reads with the same parameters are the
74 first estimation ("–*genomeSize=0.012*" and "–*minCoverage=10*").
75 In the assembly process, the retrieved DENV reads are firstly assembled with SPAdes Genome
76 Assembler (17) with the options "–*careful* –*only-assembler* –*cov-cutoff*". The coverage cut-off if
77 dictated by the "–*spadesMinCoverage*" and "–*spadesMinKmerCoverage*" parameters, set to 2 by
78 default. If the assembly with SPAdes fails to produce a contig equal or greater than the value defined
79 in the "–*minimumContigSize*" parameter (default of 10000), the data is re-assembled with the
80 MEGAHIT assembler (18) with default parameters. By default, the k-mers to be used in the assembly
81 in both tools ("–*spadesKmers*" and "–*megahitKmers*") are automatically determined depending on the
82 read size. If the maximum read length is equal or greater than 175 nucleotides, the assembly is done
83 with the k-mers "55, 77, 99, 113, 127", otherwise the k-mers "21, 33, 55, 67, 77" are used.
84 To correct the assemblies produced, the Pilon tool (19) is run after mapping the QC'ed reads back to
85 the assembly with Bowtie2 and "*samtools sort*". This process also verifies the coverage and the number
86 of contigs produced in the assembly. The behaviour can be altered with the parameters "–
87 *minAssemblyCoverage*", "–*AMaxContigs*" and "–*genomeSize*", set to "auto", 1000 and 0.01 Mb by
88 default. The first parameter, when set to 'auto', the minimum assembly coverage for each contig
89 required is set to the 1/3 of the assembly mean coverage or to a minimum of 10x. The ratio of contig
90 number per genome MB is calculated based on the genome size estimation for the samples.
91 The contigs larger than the value defined in the "–*size*" parameter (default of 10000 nucleotides) are
92 considered to be complete CDSs and follow the rest to the workflow independently. If no complete
93 CDS is recovered, the QC'ed read data is passed to the mapping to module that does the DENV typing
94 database and consensus generation.

95 The serotyping and genotyping are performed with the Seq_Typing tool (20) with the command
96 "*seq_typing.py assembly*" or "*seq_typing.py reads*", using as reference the provided curated DENV
97 typing database. It is possible to retrieve the genomes of the closest references and include them in the
98 downstream analysis by changing the "*–get_reference*" option to "true". By default, this is not included
99 in the analysis.
100 The CDSs, and the reference sequences if requested, are aligned with the MAFFT tool (5) with the
101 options "*–adjustdirection –auto*". By default, four representative sequences for each DENV serotype (1
102 to 4) from NCBI is also included in the alignment. This option can be turned off by changing the value
103 of "*--includeNCB*I" to "false". If the number of sequences in the alignment is less than 4 these are
104 automatically added.
105 A maximum likelihood phylogenetic tree is obtained with the RaXML tool (6) with the options "*-p
106 12345 -f -a*". Additionally, and by default, the substitution model ("*–substitutionModel*") is set to
107 "GTRGAMMA", the bootstrap is set to 500 ("*–bootstrap*") and the seed to "12345" ("*–seedNumber*").

108

Shotgun Metagenomics Sequencing Data

110 Samples of plasma (n=9) and serum samples (n=13) from confirmed dengue symptomatic patients
111 were collected in Venezuela between 2010-2015 (Table S2) (see Availability of supporting materials).
112 DENV positivity was confirmed by either RT-qPCR (21)(21) or nested RT-PCR (9).

113 As a positive control sample, the supernatant of a viral culture containing DENV-2 strain 16681 was
114 used. The negative control sample consisted of DNA- and RNA-free water (Sigma-Aldrich, St. Louis,
115 MO, USA).

116 A spiked sample was produced consisting of a mixture of four 5 µl of cDNA isolated from clinical
117 samples including all DENV serotypes (DENV-1 to -4). The viral cDNA for these samples was not in
118 equal concentration and the viral copy number in the clinical samples was assessed by RT-PCR (9).
119 The results were as follow: DENV-2 with 1070000 copies/µl, DENV-1 with 117830 copies/µl,
120 DENV-3 with 44300 copies/µl and DENV-4 with 6600 copies/µl.

121 The cDNA libraries were generated using either the NEBNext® RNA First and Second strand
122 modules and the Nextera XT DNA library preparation kit (NXT), or the TruSeq RNA V2 library
123 preparation kit (TS). The libraries were sequenced in MiSeq and NextSeq instruments using 300-
124 cycles v2 paired-end cartridges.

125 The DEN-IM workflow was executed with the raw sequencing data using the default parameters and
126 resources in an HPC cluster with 300 Cores/600 Threads of Processing Power and 3 TB RAM divided
127 through 15 computational nodes, 9 with 254 GB Ram and 6 with 126GB RAM.

128

Amplicon Sequencing Data

130 The accession numbers for the 106 DENV-3 paired-end amplicon sequencing paired-end short-read
131 datasets are available under BioProject PRJNA394021. The accession numbers for the 78 DENV-1
132 amplicon sequencing single-end short-read datasets are available under BioProject PRJNA321963. The
133 Run Accession IDs for both sets were obtained with NCBI's RunSelector and the raw data was
134 downloaded with the GetSeqENA tool (https://github.com/B-UMMI/getSeqENA).

135 The DEN-IM workflow was executed with the raw sequencing data with default parameters and
136 resources in the same HPC cluster as the shotgun metagenomics dataset.

137

Non-DENV Arbovirus Data

139 The accession numbers for the 132 samples, belonging to zika virus (ZKV), chikungunya virus
140 (CHIKV) and yellow fever virus (YFV) amplicon and metagenomic datasets are available as
141 supplemental material (Table S4). As with the amplicon sequencing dataset, the list of Run Accession

142 IDs was obtained with NCBI's RunSelector and the raw data was downloaded with the GetSeqENA
143 tool (https://github.com/B-UMMI/getSeqENA).

144 The DEN-IM workflow was executed with default parameters and resources in the same HPC cluster
145 as the amplicon and shotgun metagenomics datasets.

146

Supplemental tables

148 **Table S1 –** Collection date, serotype confirmation and run accession identifier for the metagenomic
149 sequencing dataset.

150 **Table S2 -** Run accession ID, BioProject SRA Study ID, source and organism present for each sample
151 of the negative control dataset (ZKV – zika virus, CHIKV – chikungunya virus, YFV – yellow fever
152 virus).

153 **Table S3 –** Number of raw base pairs, overall alignment rate against the DENV mapping database,
154 estimated coverage depths and serotype and genotype for 25 shotgun metagenomics sequencing
155 samples.

156 **Table S4 -** Number of raw base pairs, overall alignment rate, in percentage, for the mapping against the
157 DENV database, number of ORFs recovered, and respective serotype and genotype for 106 paired-end
158 amplicon sequencing samples.

159 **Table S5 -** Taxonomic profiling results for the amplicon sequencing samples with less than 70% DENV
160 DNA.

161 **Table S6 -** Number of raw base pairs, overall alignment rate, in percentage, for the mapping against the
162 DENV database, number of ORFs recovered, and respective serotype and genotype for 78 single-end
163 amplicon sequencing samples.

164 **Table S7 -** Representative sequences of serotype 1 diversity in the Dengue Virus Typing Database.

165 **Table S8 -** Representative sequences of serotype 2 diversity in the Dengue Virus Typing Database.

166 **Table S9 -** Representative sequences of serotype 3 diversity in the Dengue Virus Typing Database.

167 **Table S10 -** Representative sequences of serotype 4 diversity in the Dengue Virus Typing Database.
168

169

Supplemental Figures

a)

b)

171
172 **Figure S1 -** DEN-IM report tables. **a)** DEN-IM's quality control report containing information of the number of base-pairs
173 and the number of reads for the analysed samples, the estimated coverage depth before and after mapping, and the
174 percentage of reads in the input data that were trimmed. **b)** DEN-IM's typing report for 24 CDSs recovered from the
175 metagenomic dataset. The ID contains the CDS contig name, the typing result for serotype-genotype, the values for identity
176 and coverage, and the GenBank ID of the closest reference in the Typing Database containing 161 complete DENV
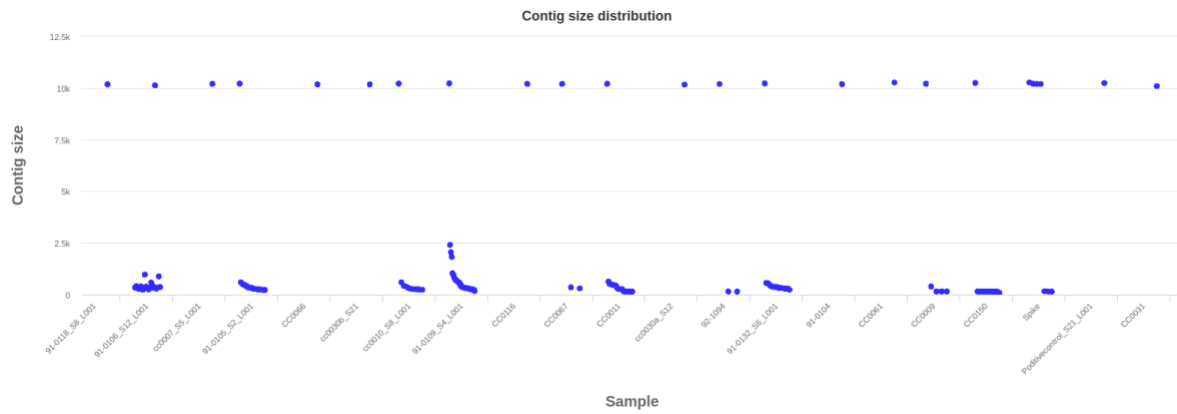177 genomes.

Figure S2 - Contig size distribution for the shotgun metagenomics sequencing dataset. Each dot depicts an assembled DENV contig. Above the 10Kb are full CDS of DENV.
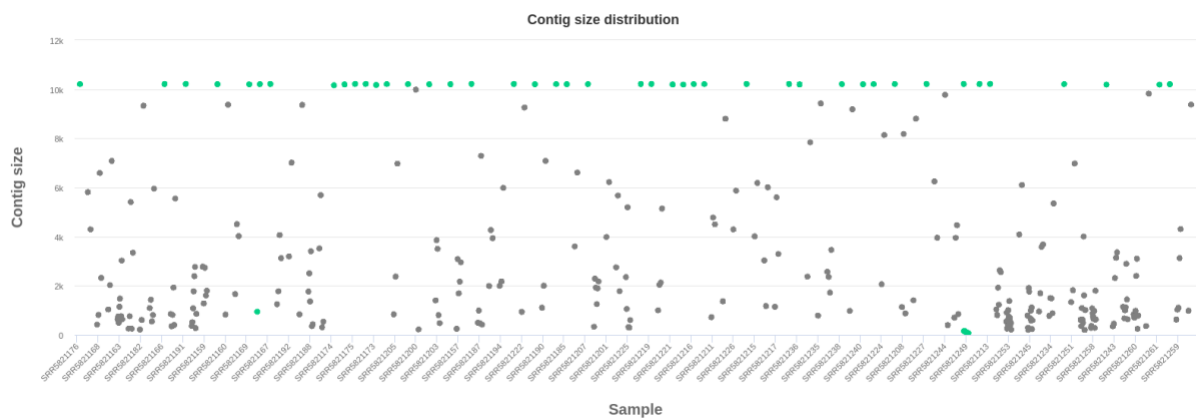

Figure S3 - Contig size distribution of the amplicon sequencing dataset with 106 paired-end samples. Each dot depicts an assembled DENV contig. Above the 10Kb are full CDS of DENV. Contigs belonging from samples that assembled a complete DENV CDS are highlighted in green, whereas the remaining are coloured in grey.
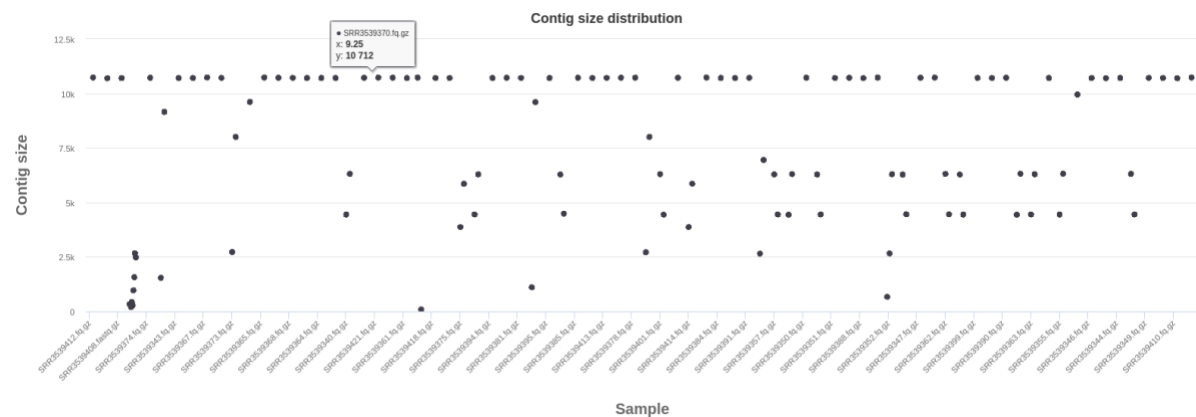

Figure S4 - Contig size distribution of the amplicon sequencing dataset with 78 single-end samples. Each dot depicts an assembled DENV contig. Above the 10Kb are full CDS of DENV.
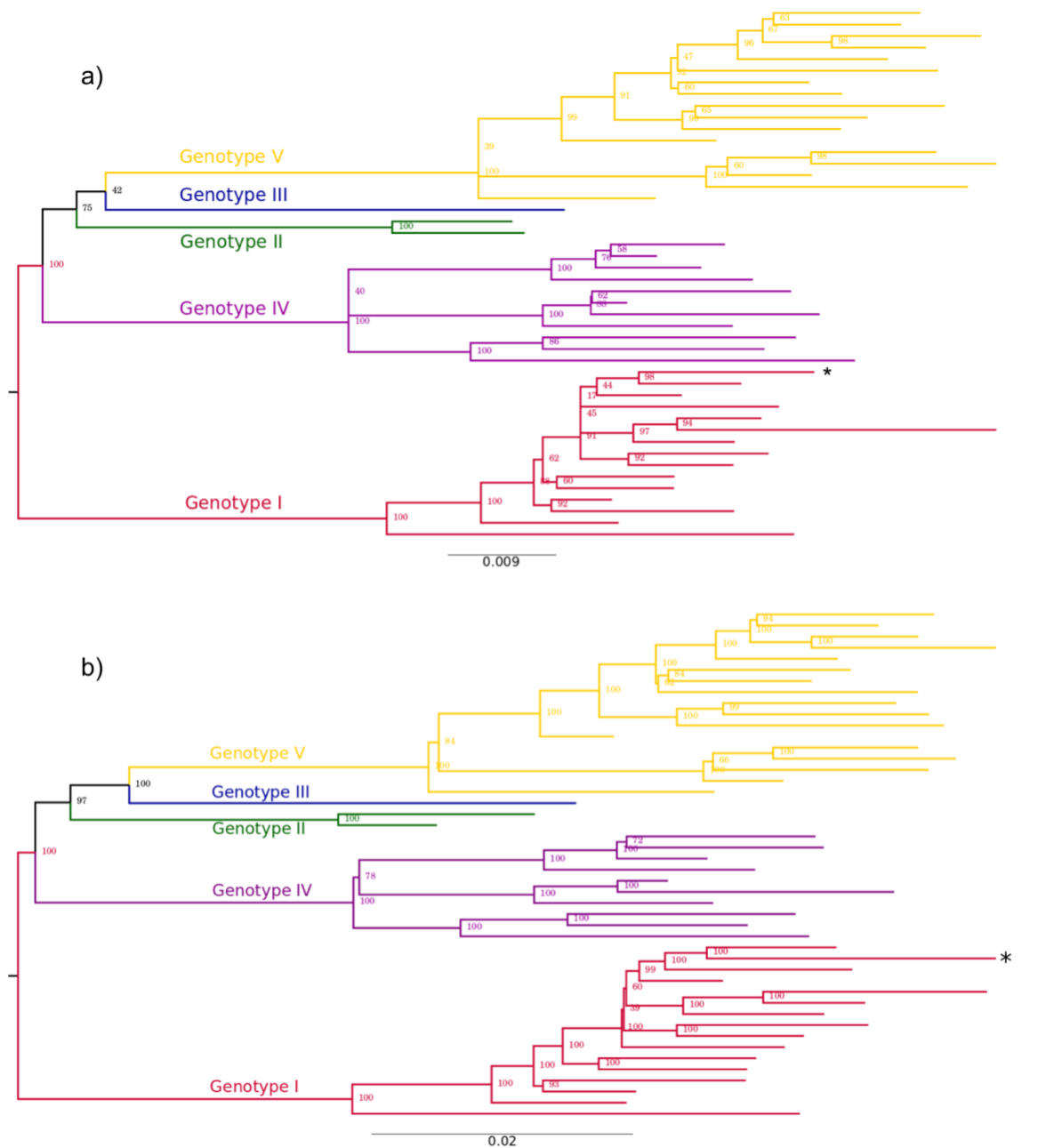
Figure S5 - Maximum Likelihood inference of the multiple sequence alignment of the 46 DENV-1 complete genomes in the typing dataset, with **a)** envelope region and **b)** whole genome sequence. 1635 complete DENV-1 genomes were clustered at 98\% nucleotide identity and the representative genomes were aligned with MAFFT. A maximum likelihood tree was inferred with RAxML. The tree is coloured according to genotype (red: genotype I; green: genotype II; blue: genotype III; purple: genotype IV). The sample JF459993, marked with a star, is currently annotated in ViPR as belonging to genotype IV but, given to the good phylogenetic support, it was re-classified as belonging to the genotype I.
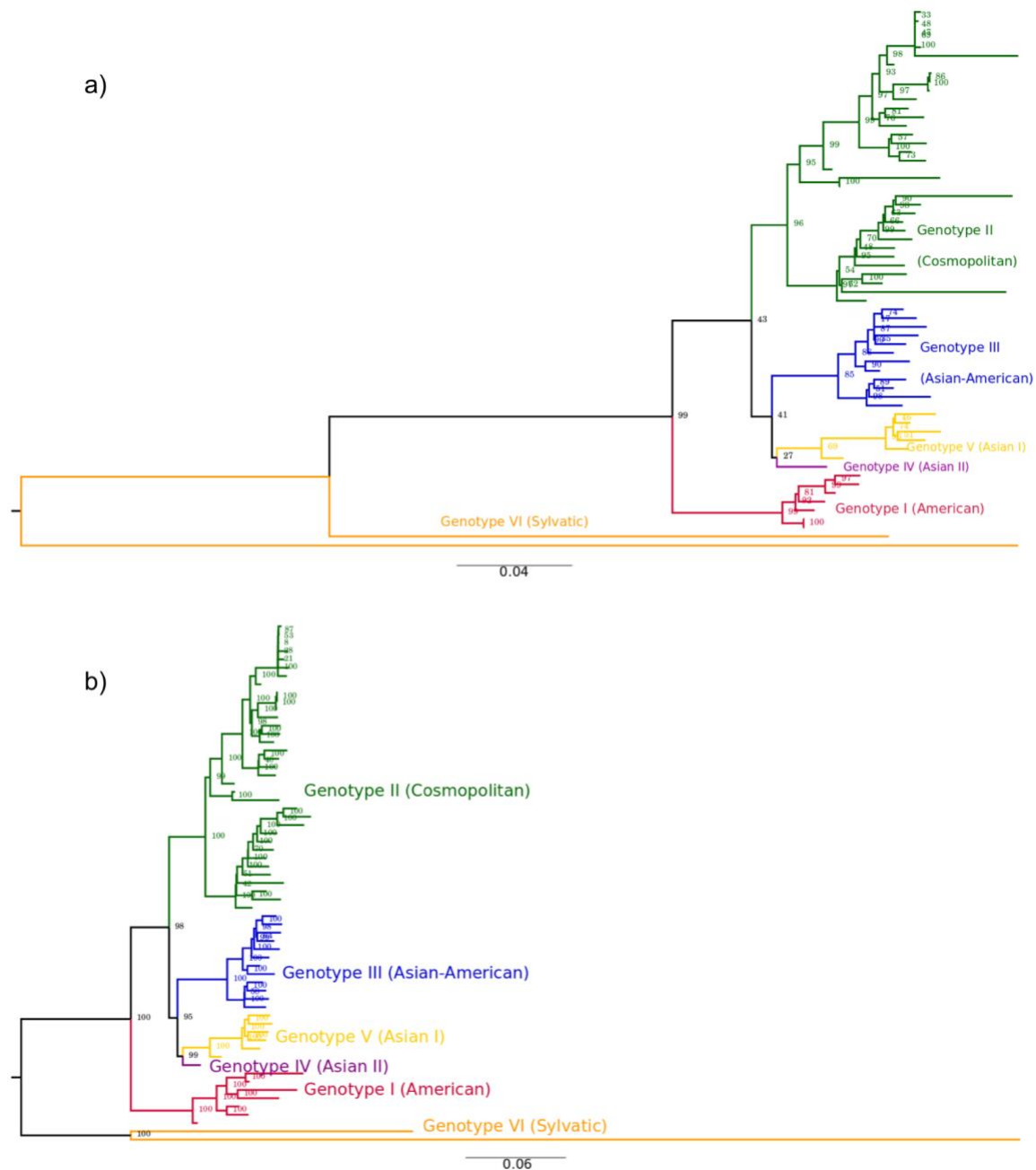
197
198 **Figure S6** - Maximum Likelihood inference of the multiple sequence alignment of the 63 DENV-2 complete genomes in the
199 typing dataset, with **a)** envelope region and **b)** whole genome sequence. 1067 complete DENV-1 genomes were clustered at
200 98\% nucleotide identity and the representative genomes were aligned with MAFFT. A maximum likelihood tree was
201 inferred with RAxML. The tree is coloured according to genotype (red: genotype I; green: genotype II; blue: genotype III;
202 purple: genotype IV).

203
204 **Figure S7 -** Maximum Likelihood inference of the multiple sequence alignment of the 25 DENV-3 complete genomes in the
205 typing dataset, with **a)** envelope region and **b)** whole genome sequence. 807 complete DENV-3 genomes were clustered at
206 98\% nucleotide identity and the representative genomes were aligned with MAFFT. A maximum likelihood tree was inferred
207 with RAxML. The tree is coloured according to genotype (red: genotype I; green: genotype II; blue: genotype III; purple:
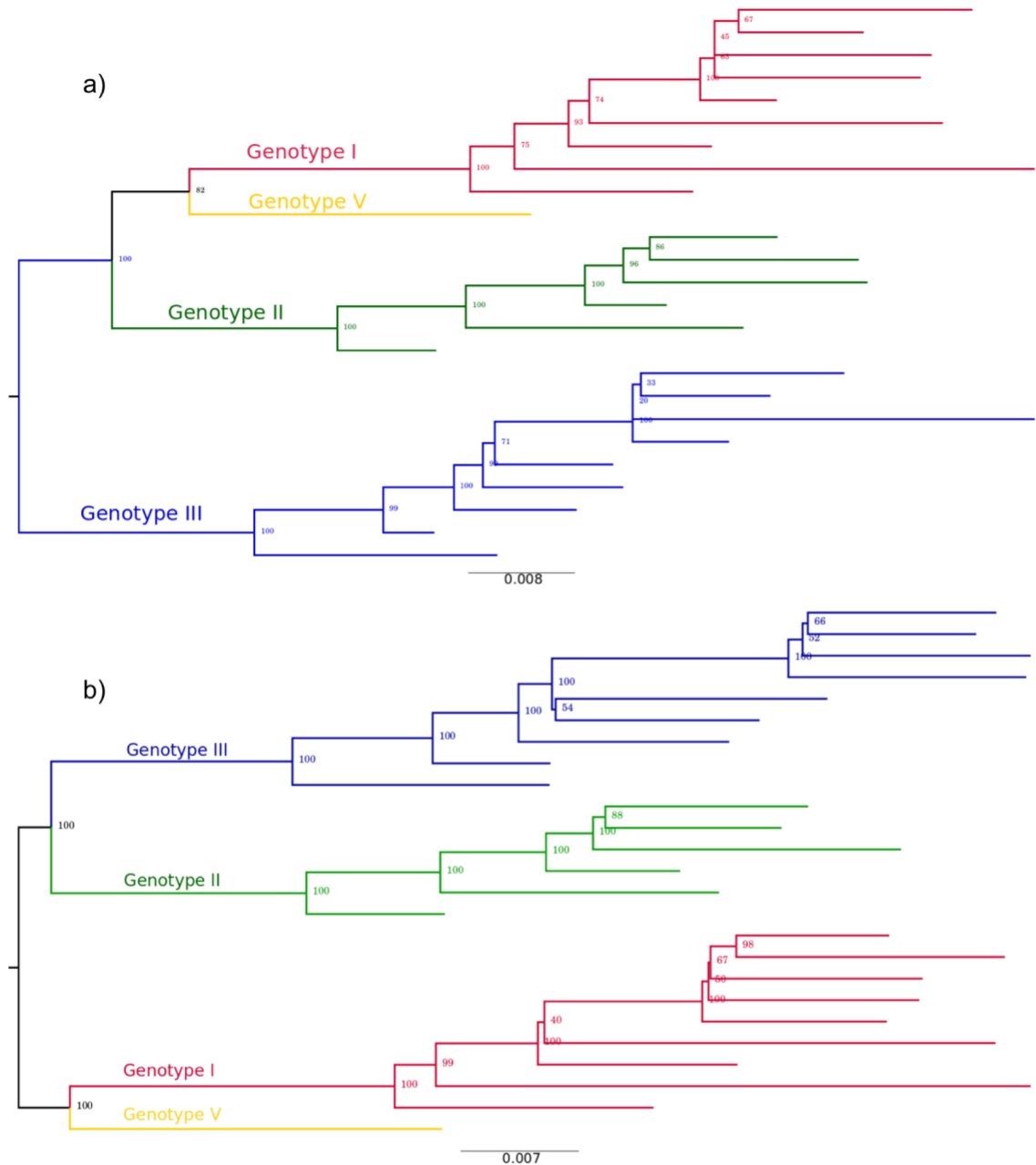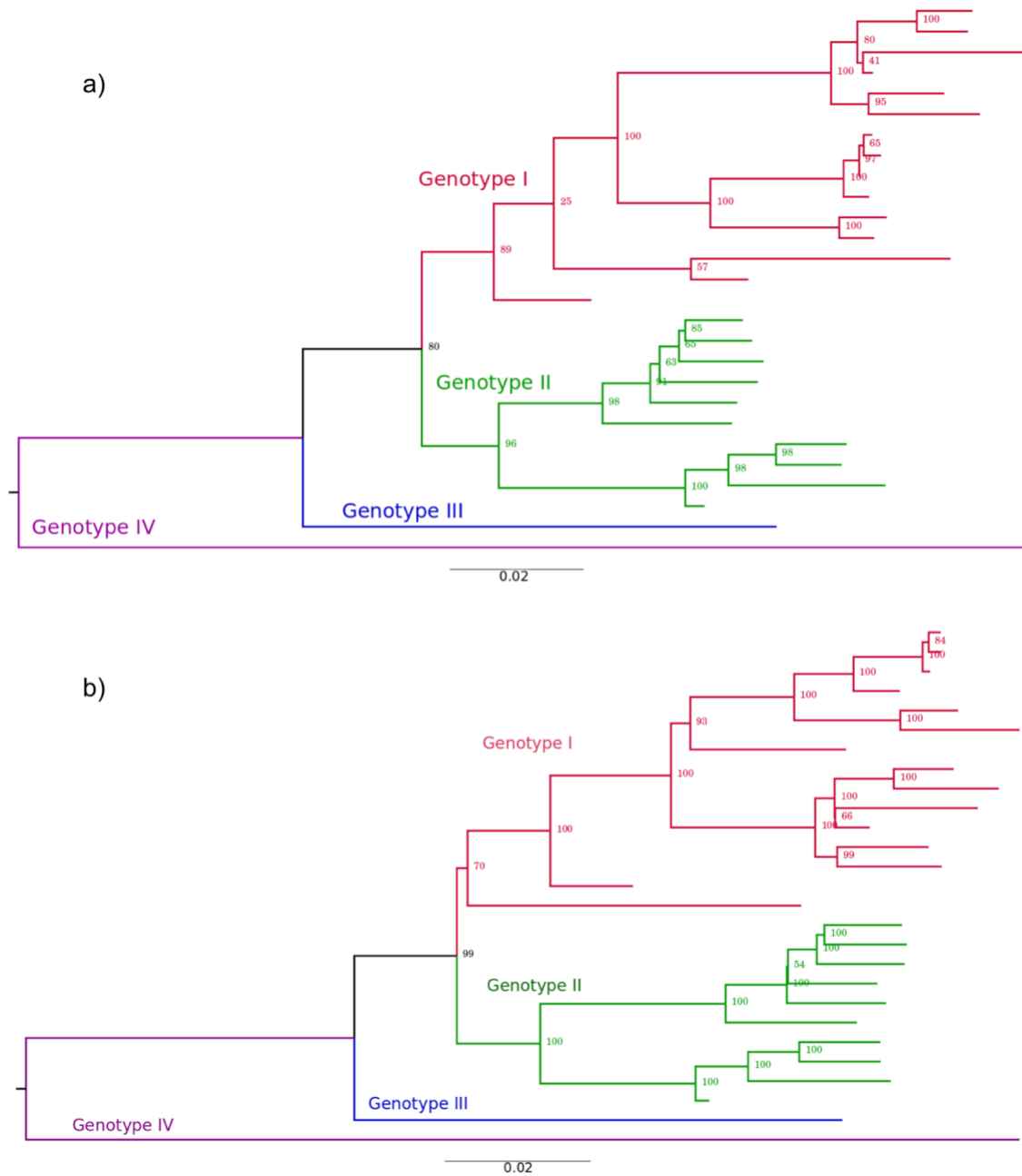208 genotype IV).

209
210 **Figure S8 -** Maximum Likelihood inference of the multiple sequence alignment of the 27 DENV-4 complete genomes in the
211 typing dataset, with **a)** envelope region and **b)** whole genome sequence. 320 complete DENV-4 genomes were clustered at
212 98\% nucleotide identity and the representative genomes were aligned with MAFFT. A maximum likelihood tree was inferred
213 with RAxML. The tree is coloured according to genotype (red: genotype I; green: genotype II; blue: genotype III; purple:
214 genotype IV).

215

References

1. Pickett BE, Greer DS, Zhang Y, Stewart L, Zhou L, Sun G, et al. Virus pathogen Database and Analysis Resource (ViPR): A comprehensive bioinformatics Database and Analysis Resource for the Coronavirus research community. Viruses. 2012;4(11):3209–26.

2. Villabona-Arenas CJ, Zanotto PM de A. Worldwide Spread of Dengue Virus Type 1. PLoS One. 2013;8(5):e62649.

3. Vasilakis N, Weaver SC. Chapter 1 The History and Evolution of Human Dengue Emergence. In: Advances in Virus Research [Internet]. Elsevier Inc.; 2008. p. 1–76. Available from: http://dx.doi.org/10.1016/S0065-3527(08)00401-6

4. Li W, Godzik A. Cd-hit : a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics [Internet]. 2006;22(13):1658–9. Available from: https://academic.oup.com/bioinformatics/article/22/13/1658/194225

5. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics [Internet]. 2018;34(March):2490–2. Available from: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty121/4916099

6. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.

7. Rico-Hesse R. Molecular evolution and distribution of dengue viruses type 1 and 2 in nature. Virology [Internet]. 1990 Feb;174(2):479–93. Available from: https://linkinghub.elsevier.com/retrieve/pii/004268229090102W

8. Rico-Hesse R. Microevolution and virulence of dengue viruses. Adv Virus Res [Internet]. 2003;59:315–41. Available from: https://www.ncbi.nlm.nih.gov/pubmed/14696333

9. Lanciotti R, Calisher C, Gubler D, Chang G, Vorndam A. Rapid detection and typing of dengue viruses from clinical samples by using reverse transcriptase-polymerase chain reaction. J Clin Microbiol [Internet]. 1992;30(3):545–51. Available from: http://jcm.asm.org/content/30/3/545.short

10. Trent DW, Lanciotti RS, Gubler DJ. Molecular evolution and phylogeny of dengue-4 viruses. J Gen Virol [Internet]. 1997 Sep 1;78(9):2279–84. Available from: http://jgv.microbiologyresearch.org/content/journal/jgv/10.1099/0022-1317-78-9-2279

11. Klungthong C, Zhang C, Mammen MP, Ubol S, Holmes EC. The molecular epidemiology of dengue virus serotype 4 in Bangkok, Thailand. Virology. 2004;329(1):168–79.

12. Zhang C, Jr MPM, Klungthong C, Monkongdee P, Nimmannitya S, Kalayanarooj S, et al. Clade Replacements in Dengue Virus Serotypes 1 and 3 Are Associated with Changing Serotype Prevalence Clade Replacements in Dengue Virus Serotypes 1 and 3 Are Associated with Changing Serotype Prevalence †. J Virol. 2005;79(24):15123–30.

13. Zhang C, Mammen MP, Chinnawirotpisan P, Klungthong C, Rodpradit P, Nisalak A, et al. Structure and age of genetic diversity of dengue virus type 2 in Thailand. J Gen Virol. 2006;87(4):873–83.

14. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

15. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863–4.

16. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods [Internet]. 2012 Apr 4;9(4):357–9. Available from: http://www.nature.com/articles/nmeth.1923

17. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A

262   New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput
263   Biol [Internet]. 2012 May;19(5):455–77. Available from:
264   http://online.liebertpub.com/doi/abs/10.1089/cmb.2012.0021

265   18.   Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: An ultra-fast single-node solution
266   for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics.
267   2015;31(10):1674–6.

268   19.   Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An
269   integrated tool for comprehensive microbial variant detection and genome assembly
270   improvement. PLoS One. 2014;9(11).

271   20.   Machado MP, Ribeiro-Gonçalves B, Silva M, Ramirez M, Carriço JA. Epidemiological
272   Surveillance and Typing Methods to Track Antibiotic Resistant Strains Using High
273   Throughput Sequencing. Methods Mol Biol [Internet]. 2017;1520:331–56. Available from:
274   http://www.ncbi.nlm.nih.gov/pubmed/27873262

275   21.   Santiago GA, Vergne E, Quiles Y, Cosme J, Vazquez J, Medina JF, et al. Analytical and
276   Clinical Performance of the CDC Real Time RT-PCR Assay for Detection and Typing of
277   Dengue Virus. PLoS Negl Trop Dis. 2013;7(7):36–8.

278