Sensitivity analysis of the method, taxonomic sampling, and genomic partition in the phylogenetic reconstruction of the dengue virus (DENV)

Natalia González Piñeres

Trabajo de Grado para Optar el título de Bióloga

Director
Daniel Rafael Miranda Esquivel
Doctor en Ciencias Naturales

Codirectora
Cinthy Lorena Jiménez Silva
Magíster en Ciencias Básicas Biomédicas

Universidad Industrial de Santander
Facultad de Ciencias
Escuela de Biología
Bucaramanga
2019

To my parents, Isabel Piñeres and Marcos González, for their unconditional love and support.

To my brother, Sergio González, for giving me his precious smile and amazing company in difficult times.

To my best friends, Fred and Vivi, for being an important part of this adventure of becoming a biologist.

Acknowledgments

We are beholden to the División de Investigación y Extensión, Facultad de Ciencias, Universidad Industrial de Santander, for their financial support. The funders did not participate in the research proposal, database assembly and analysis, decision to publish, or preparation of the manuscript.

Table of Contents

Tables List

Figures List

Appendix List

(Appendices G-I can be found on the CD, and also at the database of UIS library)

Resumen

TITULO: ANÁLISIS DE SENSIBILIDAD DEL MÉTODO, MUESTREO TAXONÓMICO Y PARTICIÓN GENÓMICA EN LA RECONSTRUCCIÓN FLOGENÉTICA DEL VIRUS DE DENGUE (DENV) *

AUTOR: NATALIA GONZÁLEZ PIÑERES**

PALABRAS CLAVES: FILOGENIA DENV, CONGRUENCIA TAXONÓMICA, ANÁLISIS DE SENSIBILIDAD, MUESTREO TAXONÓMICO, PARTICIONES GENÓMICAS,

DESCRIPCIÓN:

La reconstrucción filogenética del virus del dengue (DENV) se ha abordado desde múltiples perspectivas que difieren en el método de reconstrucción, el muestreo taxonómico y los datos genómicos. En la mayoría de estos estudios, no se han presentado los motivos para seleccionar dichos enfoques. Por lo tanto, el objetivo del presente estudio fue determinar los efectos de estas variables en la reconstrucción filogenética del DENV. Realizamos los análisis filogenéticos utilizando tres métodos de reconstrucción de árboles: parsimonia, distancia y Maximum Likelihood. Utilizamos 410 secuencias del marco abierto de lectura (ORF) de los cuatro serotipos: DENV1-4. Para estimar el efecto del muestreo taxonómico, obtuvimos submuestreos del 10%, 36% y 75% del total de taxones. Para cada conjunto de datos, utilizamos 22 particiones genómicas del ORF. Comparamos las topologías en términos de recuperación de genotipos y serotipos como grupos monofiléticos, nodos comunes, similitud topológica, congruencia taxonómica y soporte de nodos. Encontramos que la partición genómica sobrepesó las otras variables con respecto al número de nodos comunes, congruencia taxonómica y soporte nodal. Todos los métodos recuperaron la monofilia de los serotipos, independientemente del muestreo taxonómico o la partición genómica; pero para las particiones genómicas más cortas, la recuperación de genotipos disminuyó a medida que aumentó el número de taxones. Además, los resultados mostraron que parsimonia y Maximum Likelihood obtuvieron resultados casi idénticos en términos de congruencia con una topología de evidencia total. Con base en estos resultados, discutimos las implicaciones de cada variable y una partición genómica que podrían mejorar la eficiencia en estudios posteriores.

---

*Trabajo de Grado

**Facultad de Ciencias. Escuela de Biología. Director: Daniel R. Miranda-Esquivel, PhD. En Ciencias Naturales.

Abstract

TITLE: SENSITIVITY ANALYSIS OF THE METHOD, TAXONOMIC SAMPLING, AND GENOMIC PARTITION IN THE PHYLOGENETIC RECONSTRUCTION OF THE DENGUE VIRUS (DENV)*

AUTHOR: NATALIA GONZÁLEZ PIÑERES**

KEY WORDS: DENV PHYLOGENY, TAXONOMIC CONGRUENCE, SENSITIVITY ANALYSIS, TAXON SAMPLING, GENOME PARTITIONS, MONOPHYLY RECOVERY

DESCRIPTION:

The phylogenetic reconstruction of the dengue virus (DENV) has been approached from multiple perspectives that differ in the method of reconstruction, taxonomic sampling, and the genomic data. In most of these studies, the reasons for selecting these approaches have not been reported. Thence, the aim of the present study was to determine the effects of these variables on the phylogenetic reconstruction of the DENV. We performed the phylogenetic analyses using three methods of tree reconstruction: parsimony, distance, and Maximum Likelihood. We used 410 ORF (Open Reading Frame) sequences of all the four serotypes: DENV1-4. To estimate the effect of the taxonomic sampling, we obtained subsamples of the 10%, 36%, and 75% of the total taxa. For every dataset, we used 22 genomic partitions from the ORF. We compared the topologies in terms of recuperation of genotypes and serotypes as monophyletic groups, common nodes, topological similarity, taxonomic congruence, and node support. We found that the genomic partition overpoised the other variables regarding the number of common nodes, taxonomic congruence and nodal support. All the methods recuperated the monophyly of the serotypes, regardless the taxonomic sampling or genomic partition; but for the shortest genomic partitions, the recuperation of genotypes decreased as the number of taxa increased. Moreover, the results showed that parsimony and Maximum Likelihood obtained nearly identical results in terms of congruence to a complete evidence topology. Based on these results we discuss the implications of each variable and one genomic partition that could improve the efficiency in further studies.

---

*Bachelor Thesis

**Facultad de Ciencias. Escuela de Biología. Director: Daniel R. Miranda-Esquivel, PhD. En Ciencias Naturales.

Introduction

The dengue virus (DENV) has been a widely studied model in phylogeny due to its population dynamics (Allicock *et al.*, 2012; Hapuarachchi *et al.*, 2016; Lequime *et al.*, 2016), high evolutionary rate (Jenkins *et al.*, 2002; Twiddy *et al.*, 2003; Costa *et al.*, 2012), genotype diversity (Rico-Hesse, 1990; Holmes, 2003; Lee *et al.*, 2012), responsibility as the cause of one of the reemerging infections with the highest incidence globally (WHO, 2016); and therefore to the fact that it is considered a major public health problem (Gubler, 2002; Bhatt *et al.*, 2013). DENV is grouped into four antigenically differentiated serotypes, designated as DENV-1, DENV-2, DENV-3 and DENV-4 (Wang *et al.*, 2000; Ross, 2010). Within each serotype, there are multiple groups of lineages or genotypes (Rico-Hesse, 1990; Holmes, 2004), which have been considered as monophyletic groups (e.g. Laille and Roche, 2004; Alfonso *et al.*, 2012; Villabona-Arenas and Zanotto, 2013).

In order to obtain information on the evolution and genetic diversity of this virus, different authors have used phylogenetic reconstructions (Araujo *et al.*, 2009; Villabona-Arenas and de Andrade Zanotto, 2011; Chen and Vasilakis, 2011), that widely varied in: the method of reconstruction (parsimony (PA), distance (DI), Maximum likelihood (ML), and Bayesian analysis); the partition of the molecular data, such as the *E* gene (Twiddy *et al.*, 2002; Chen and Vasilakis, 2011; Villabona-Arenas *et al.*, 2016), open reading frame (Aviles *et al.*, 2003; Schreiber *et al.*, 2009; Hapuarachchi *et al.*, 2016), complete genome (Tolou *et al.*, 2001; Caceres *et al.*, 2008; Azhar *et al.*, 2015), *E* and *NS1* gene junction (Rico-Hesse, 1990; Pires Neto *et al.*, 2005; Nur Liyana *et al.*, 2016), or an *E* gene domain (Chungue *et al.*, 1995; Usme-Ciro *et al.*,

2008; Ciccozzi *et al.*, 2014); and the taxonomic sampling or number of tips used, ranging from 8 (Tolou *et al.*, 2001) to 1619 (Ernst *et al.*, 2015).

Given the variety of methodological approaches, it is expected that there will be effects on the reconstructed phylogenies and inferences derived from them (Goldberg, 2003; Planet, 2006; Lam *et al.*, 2010). In fact, a long discussion has been held about the behaviour of certain methods of topological reconstruction (Miyamoto and Fitch, 1995; Leach and Reeder, 2002; Hall, 2005), different genomic partitions (Rokas *et al.*, 2003; Kolaczkowski and Thornton, 2004; Gadagkar *et al.*, 2005), and taxonomic sampling (Hillis, 1998; Rosenberg and Kumar, 2001; Heath *et al.*, 2008), in the nodes obtained and their support (Simmons and Geisler, 2002; Castoe *et al.*, 2004; Kutty *et al.*, 2007).

Nonetheless, in most phylogenetic studies published at present, the reasons for selecting or preferring certain approaches remain unclear. Also, there are no studies about the overall effect of the conditions under which these reconstructions are performed, in the topologies obtained. Consequently, the aim of the present study was to estimate the effect of the method, taxonomic sampling and genomic partition on the phylogenetic reconstruction, using the dengue virus as a model.

1. Objectives

1.1 General Objective

To estimate the effect of the method, taxonomic sampling and genomic partition on the phylogenetic reconstruction of the dengue virus (DENV).

2. Materials and Methods

2.1 Sequence data

We downloaded all available sequences of the ORF for the four serotypes (DENV1-4), until February 20, 2018, from the NCBI Virus Variation database (Brister *et al.*, 2014, 2015). Given the fact that interhost diversity in dengue has determined different evolutionary processes (Parameswaran *et al.*, 2012) we did not include the sequences whose host was not human or not specified. We removed identical sequences, clones and chimeras using the sequence-split algorithm in the usearch program v.8.1.1861 (Edgar, 2010), with a 99.0% identity cutoff. We use this cutoff value for two reasons: first, to ensure the removal of all the identical sequences that could affect the topological resolution (e.g. Lutzoni *et al.*, 2000; DeFilippis and Moore, 2000; Degnan and Rosenberg, 2009); and second, to guarantee that the total taxonomic sampling did not exceed 400 terminals. The latter, in order to minimize on computational times.

In order to identify recombination in the data set, we applied the statistical test of recombination detection, implemented in the program SplitsTree v.4.2 (Bruen *et al.*, 2006). Subsequently, we searched for potential recombination events between the input sequences by using different methods (RDP (Martin and Rybicki, 2000), GENECONV (Padidam *et al.*, 1999), Bootscan (Salminen *et al.*, 1995), Maxchi (Smith, 1992), Chi-maera (Posada and Crandall, 2001), SiSscan (Gibbs *et al.*, 2000), PhylPro (Weiller, 1998), LARD (Holmes *et al.*, 1999), 3Seq (Boni *et al.*, 2007)), included in the pro-gram of analysis of recombinant patterns of viral genomes RDP4 (Martin *et al.*, 2015). We excluded the recombinant sequences identified in this

search, in order to guarantee the non-sub or underestimation of the branch lengths by the mosaic behavior of the recombinant sequences in the phylogeny (Schierup and Hein, 2000a,b; Posada, 2001).

2.2 Genotyping

We aligned the resulting sequences with the multiple sequence alignment algorithm implemented in MUSCLE v.3.8.31 (Edgar, 2004a), using the default parameters, which were established to progressively re ne the alignment, thus ensuring accuracy, convergence of parameters, and efficiency in computation time (Edgar, 2004a,b).

We followed the classification scheme proposed by the authors referenced in Table 2. We assigned the genotypes to each sequence according to the result of the phylogenetic analysis by Maximum Likelihood, using the ORF, based on the rationale proposed by (Klungthong *et al.*, 2008; Schreiber *et al.*, 2009). We further reviewed and compared this allocation taking into account the current taxonomic classification (Appendix H.); and the search results using the algorithm of free alignment sub-typing based on return time distribution RTD (Kolekar *et al.*, 2012), implemented in the Dengue Subtyper server, accessible at http://196.1.114.46:1800/dengue/RTD.html. We conducted this last search with the non-aligned sequences. We designated all the sequences to a particular cluster, determined by one of the criteria if the other two did not show a result, or by two of the criteria agreeing against one of the three, either the ORF phylogeny, the bibliographic revision or the RTD search.

2.3 Partition Delimitation

Using the total aligned sequences of the Open Reading Frame (ORF) , we obtained 22 partitions (P01-P22), corresponding to: the different genes of the DENV genome; combinations of genes, which have been widely used in the phylogenetic reconstruction of DENV; the domains of the *E* and *NS5* genes; structural genes and non-structural genes (Table 1). We organised these partitions taking the number of nucleotides into the account, so P01 would be the smallest partition and P22 the largest one. The purpose of partitioning genes such as *E* and *NS5* was to assess the similarities between the topologies they yield as single domains. We did not take untranslated regions into account due to the lack of completely sampled sequences, and the results of preliminary analyses we performed using these regions (data included in Supporting Files), which showed they yield low topological resolution (< 50% fully resolved nodes).

After defining the genome partitions, we divided the dataset into two sets: total terminals, made up of the terminals of all serotypes; and subsampled terminals. For the latter, we acknowledged that there is an inverse square root relationship between confidence intervals and sample sizes. Hence, in order to cut the margin of error in half, one would need to approximately quadruple the sample size. Thereby, we obtained sub-sets equivalent to 10%, 36%, and 75% of the total terminals, maintaining the proportions of frequency of each serotype in this total sample (DENV1: 38%, DENV2: 28%, DENV3: 19%, DENV4: 15%), thus guaranteeing that the least frequent clade contained at least 2 representatives per genotype.

We chose this number of representatives for two reasons. First, some genotypes such as the Sylvatic or Divergent for DENV1 only had two representatives after the data cull. Second, each node in a tree depicts the last common ancestor of the two lineages that descend from it, therefore two representatives are the minimum acceptable number to evaluate the monophyly of a group of descendants; in this case, serotypes and genotypes. From each of the genome partitions of the total sample (100% of the tips), we did each subsampling (10%, 36%, 75%) using 30 replicates with fixed and specific seeds for each replicate.

2.4 Phylogenetic analyses

For each genome partition of the total terminals and their subsamples, we chose the nucleotide substitution model that best fits the data following the Akaike Information Criterion (AIC), using the Phymltest function of the ape package v.5.3 (Paradis *et al.*, 2004; Popescu *et al.*, 2012), which calls PhyML v.3.0 (Guindon *et al.*, 2010) from the R platform v.3.5.3 (R Core Team, 2019). We used three methods of phylogenetic re-construction: Parsimony (Camin and Sokal, 1965; Farris, 1970; Fitch, 1971), Maximum Likelihood (Cavalli-Sforza and Edwards, 1967; Felsenstein, 1981), and Distance based (Sokal and Sneath, 1963; Saitou and Nei, 1987; Gascuel, 1997).

We used the TNT v.1.5 (Goloboff*et al.*, 2008) program for parsimony reconstruction. This reconstruction was done under equal weights. We set the depth of the searches in dependency of the number of tips presented in each of the sub-matrices. For the subsets of the 10% of the terminals, we performed 30 replicates of random-addition-sequence e orts (RAS) and 5 ratchet

rounds. For the subsets with the 36% and 75% of the terminals, we used new technology searches (xmult) with 100 replicates and 30 ratchet iterations. For the matrices with the complete taxon sampling, we used also new techs (xmult), but adding sectorial searches, 40 iterations of ratchet, 40 rounds of tree-fusing and 40 rounds of tree-drifting. The support values for the nodes were calculated using 100 pseudo replicates of non-parametric Bootstrap.

We did the reconstruction by distance-based method by calculating the distance matrix from the TN93+G+I nucleotide substitution model, using the function dist.dna of the package ape v.5.3 (Paradis *et al.*, 2004; Popescu *et al.*, 2012). We estimated the topology from the previously generated genetic matrix with the BIONJ algorithm V.5.1 (Saitou and Nei, 1987; Gascuel, 1997), using the function bionj of the package ape v.5.3 (Paradis *et al.*, 2004; Popescu *et al.*, 2012). We calculated the support of the nodes by non-parametric Bootstrap of 100 replicates, using the boot.phylo function of the package ape v.5.3 (Paradis *et al.*, 2004; Popescu *et al.*, 2012). We used the functions within the R v.3.5.3 (R Core Team, 2019) platform.

We did the reconstruction by Maximum Likelihood using a reference tree obtained from the selected substitution model, using 30 replicates in RaxML v.8.2.11 (Sta-matakis, 2014). We used the GTR CAT approximation (Stamatakis, 2006) for all the topologies, as it was more practical for two reasons: first, the coding and automation processes, given the amount of trees we wanted to estimate; and second, because it was the nucleotide substitution model selected for most cases.

We calculated the support of the clades by nonparametric Bootstrap of 100 replicates. We inferred all the topologies using the same root, which was conformed by the reference sequences

of the West Nile Virus (GenBank accession number: (NC 009942); Japanese Encephalitis (NC 001437); and Usutu (NC 006551).

2.5 Comparisons between topologies

2.5.1 Node recovery

We evaluated the taxonomic congruence using the common node count metric (Nel-son, 1979; Goloboff, 1997; Ramírez, 2003), which we calculated in terms of the percent-age of shared nodes between the current topology (each of the topologies reconstructed from each genomic partition) and the reference topology (reconstructed from the ORF). Topological similarity

To determine the differences or similarities in terms of the branching pattern, we calculated the Robinson-Foulds distance (Robinson and Foulds, 1981) between the topologies reconstructed from each partition, using the RFdist function of the Phangorn package v.2.4.0 (Schliep, 2011). We divided this distance into RFMax for each case and then multiplied by 100.

In order to summarize and visualize how the genome partitions are grouped according to the similarity between the topologies they generated, we obtained a partition dendrogram for each replic. This, by using the topological distances matrix calculated by the Neighbor Joining NJ algorithm for all the trees of a replic (Saitou and Nei, 1987).

For the case of subsamples, we built a majority-rule consensus dendrogram of the 60% using the consensus function of the ape package v.5.3 (Paradis *et al.*, 2004; Popescu *et al.*, 2012). We chose this percentage taking into account the rationale presented by (Barrett *et al.*, 1991; Gascuel and Berry, 1996; Bryant *et al.*, 2003; Holder *et al.*, 2008, e.g); and previous analyses that showed more resolved summaries (with almost no collapsed branches).

2.5.2 Monophyly recovery. We calculated the monophyly recovery as the percentage of replicates (out of 30 for the subsamples) that recuperated the monophyly of each of the clades of interest; that is, serotypes and/or genotypes as assigned in the typification process of this study.

To test whether there was a linear correlation between the recuperation values (com-mon nodes, monophyly), topological distances and the size of the genomic partition, we performed a Pearson correlation analysis Pearson (1896).

2.5.3 Node support. To evaluate whether each node of the topology, common nodes, and nodes delimiting the monophyly of the clades of interest were representing all of the data, we compared the support values. To discriminate the support, we established three categories: low (50-75%), moderate (76-94%), and high (95-100%), taking into account previous studies comparing this measurement (Erixon *et al.*, 2003; Simmons *et al.*, 2004). To estimate the proportion of nodes with high support for each partition, we calculated the per-centage of topology nodes for each category. All the procedures described before are summarized in Appendix A.

3. Results

3.1 Sequence data

We retrieved 4471 sequences of complete genomes from 57 countries, encompassing a temporal range of 72 years (1944-2016) Appendix G. After excluding the sequences that did not fulfill the expected length; that were identical, clones, chimeras, or potential recombinants, we obtained 410 sequences (9.2% of the total, unfiltered sample) of the complete genomes from 48 countries (67% of the total, unfiltered sample), and a temporal range of 42 years (1944-2016), covering 73% of the total sample. The proportions of the serotypes in this sampling were kept similarly as the original ones (Appendix H.). The total sample includes 155 sequences of DENV1 (37.8%), 116 sequences of DENV2 (28.3%), 79 sequences of DENV3 (19.3%), and 60 sequences of DENV4 (14.6%).

3.2 Genotyping

The final dataset contained at least 2 representants per genotype (Table 3). We obtained sequences from all the lineages previously proposed for each serotype, but the American genotype for DENV2. We excluded the representative sequences for this genotype, either by the identity cutoff in the first letter or as a result of being marked as potential recombinants by the analysis of recombinant patterns of viral genomes. The latter was supported by at least two of the nine recombination detection methods in RDP4 (Appendix I).

3.3 Comparisons between topologies

3.3.1 Monophyly Recovery. As shown by Tables 4-9, the monophyly of the four serotypes was recuperated by all the methods overall (mean: 93.08, ci: 3.77), regardless the taxonomic

sampling (Figure 1). The mean recuperation value for all the serotypes was 100% for the topologies reconstructed by distance methods; 97.84% for the topologies reconstructed by parsimony (sd: 7.14, ci: 2.92); and 81.42% for the topologies reconstructed by Maximum Likelihood (sd: 11.28, ci: 4.61).

In spite of the fact that the taxonomic sampling did not affect the recuperation of serotypes as monophyletic groups, it did affect the recuperation of genotypes. The average proportion of recovered genotypes decreased as the number of tips increased. In the trees reconstructed with parsimony, this proportion ranged from 75.06 % to 58.58%, represented by a recuperation of 18 and 13 out of 19, correspondingly. In the topologies reconstructed with distance, the average recuperation ranged from 74.37% to 54.92% (18 and 13 out of 19) and
in the Maximum Likelihood topologies from 67.57% to 52.86% (18 and 14 out of 19). These values varied for the 36% and 75% of tips, albeit, the number of genotypes recovered stayed the same for all the methods.

Under all the methods of reconstruction, for the full taxonomic sampling, none of the topologies of any of the genome partitions recovered the monophyly of the genotypes: divergent of DENV1; cosmopolitan of DENV2; I and V of DENV3; I and II of DENV4 (Appendix B-C).

Although all the methods inferred similar classification schemes, each of them showed particularities (Appendix B-D). For instance, the topologies re-constructed by parsimony, using the genomic partitions: (C) and (NS5-DII), did not recover the monophyly of the

serotype DENV2. Interestingly, the genome partition (C) always recovered the monophyly of the sylvatic genotype of DENV1, despite the taxonomic sampling (Appendix B-E).

For the distance trees, the monophyly of the sylvatic genotype of DENV1 was recovered by the genomic partitions: (*C*), (*NS2B*), (*C-M*), (*E*-DII), (*NS4B*), (*E*), (*prM*/*M-E*), Structural genes, (*NS5*), and (*E-NS5*) (Appendix C). For Maximum Likelihood trees, the only serotype that was recovered for all the cases was DENV3; the monophyly of the sylvatic genotype of DENV1 was recovered only by the genomic partition (*E*-DIII); and the Cosmopolitan genotype of DENV2 by the partition (*NS4A*).

Considering all the genotypes within each serotype, we found that all the topologies in all the methods, obtained similar average values of monophyly recovery (Appendix B). The highest values corresponded to the genotypes within DENV2; followed by DENV4; DENV3; and DENV1 was the serotype with the lowest values of recuperation of genotype monophyly (Appendix E).

3.3.2 Common Nodes. Using the complete coding region (ORF) as the reference topology, for the 10% of the taxonomic sampling (53 tips), all the 22 genome partitions recovered at least 50% of the nodes of the ORF topology (Table 10; Appendix C-D). This, in spite of the inference method. In average, the proportion of common nodes from the reference topology was 81.28 % for the topologies inferred by parsimony (sd: 7.76; ci: 0.60); 78.43 % for the

ones inferred by distance (sd: 8.60; ci: 0.67); and 64.71% for the ones inferred by Maximum Likelihood (sd: 5.49; ci: 2.29).

As the number of taxa increased, the proportions of common nodes were the highest (above 65%) and most similar for parsimony and Maximum Likelihood; as we can see their curves in the plot almost overlap (Figure 2 and 3). These two methods were not affected by the taxonomic sampling when using the ten largest genomic partitions in the tree inference: *NS1*; *E*; *NS3*; *NS5*-DII; *prM*/*M-E*; SG; *NS5*; *E-NS5*; NSG). Nonetheless, for the topologies reconstructed with distance methods, only the two largest partitions (*E-NS5*; NSG), recovered more than 65 % of the nodes from the ORF topology in all the taxonomic samples (subsampled tips and total tips).

The proportions of common nodes, when using the total taxa, were also the highest for the topologies reconstructed by parsimony and Maximum Likelihood. However, for all the methods, the proportion of common nodes decreased as the number of taxa increased (Appendix C-D).

Overall, for the smallest sampling size we used (10% of total taxa: 53 tips), all the genomic partitions yielded proportions of above 60% of common nodes with the total evidence topology; except (*E*-DIII) and (*C-M*) in Maximum Likelihood. However, as the number of taxa increases, only the largest partitions (*NS1*; *E*; *NS3*; *NS5*-DII; *prM*/*M-E*; SG; *NS5*; *E-NS5*; NSG) could recover similar topologies, with at least 60% of common nodes, regardless of the reconstruction method. It is important to remark that the highest proportions

of common nodes were obtained by the trees reconstructed with parsimony and Maximum Likelihood. Using the last method, in every taxonomic sampling, from the genomic partition NS3 on, the proportion of common nodes was above 70%.

3.3.3 Topological similarity. Using the Robinson-Foulds distance, we found that the topologies reconstructed by parsimony methods are very similar, with less of 20% of distance, despite the number of tips (Table 11). This method showed to be the least sensitive to the taxonomic sampling, remarkably so, in spite of the genomic partition. Maximum Likelihood, on the other hand, showed very similar values (min: 23.73, max: 25.01, mean: 24.29). The smallest distances, and also in the same range as the ones obtained by parsimony were shown by the topologies reconstructed by Maximum Likelihood, but only with the largest genomic partitions (*E*; *NS3*; *NS5*-DII; *prM*/*M*-*E*; SG; *NS5*; *E*-*NS5*; NSG; ORF).

The highest topological distance values were displayed by the trees inferred by the distance method (Figure 4). Although this method was less sensitive when using the largest genomic partitions, the RF values were greater than 15 (min: 17.43, max: 39.56, mean: 28.24). In terms of how partitions yielded similar nodes, we found that the combination of non-structural genes (NSG), (*NS5*), (*NS5*-DII), and (*E*-*NS5*) generated the most similar topologies to the one inferred by total evidence (Appendix G).

As indirectly shown by the common nodes count, the topologies reconstructed by parsimony and Maximum Likelihood are very similar (Table 11). These topologies showed the smallest RF distance values (Figure 5). In average, the distances between these two

methods, in every taxonomic sampling, were less than 30% (min: 24.24, max: 26.32, mean: 25.25). The most dissimilar methods according to the topologies they yielded were distance and Maximum Likelihood (min: 32.45, max: 57.39, mean: 47.42).

3.3.4 Node Support. In general, the larger the genomic partition, the higher the proportion of nodes with high Bootstrap support (between 95% and 100%) (Figure 6). The trees reconstructed using the nine largest partitions yielded more than 30% of their nodes with high support, and more than 20% with moderate support (Appendix D-E). The trees inferred by parsimony showed higher support values than Maximum Likelihood and distance (Table 12).

The taxonomic sampling affected the support values for the genomic partitions that were smaller than 1000 nucleotides (from *E-NS5* to *NS5*-DI). For instance, as more tips were added into the phylogeny, the support values decreased in the high category for the nodes, reconstructed from these genomic partitions. The support values did not vary as much when using the genomic partitions whose size surpassed 1000 nucleotides (from *E-NS5* to NSG). As an example, considering the results of non-structural genes, when using parsimony and Maximum Likelihood, correspondingly, the average proportion of nodes in the category of high support for was: 76.82% and 67.22% for T10; 68.13% and 66.24% for T36; 66.02% and 64.65% for T75 and 61.99% 66.26% for T100 for parsimony.

The trees reconstructed by distance methods, however, did not show an effect of the length of the genomic partition on the support value. As shown in Figure 5, the only case that was similar to the pattern shown by the two other methods of reconstruction was T10. When

using more than 60 tips (T36, T75, T100), the number of nucleotides in the sequence showed no effect on the relatively low values of node support in the high category.

4. Discussion

To our knowledge this study is the first sensitivity analysis of the overall effects of the taxonomic sampling, genomic subsampling and method of tree reconstruction on the phylogenetic relationships of the dengue virus. Our results show that these variables affected not only the phylogenetic relationships, but also the nodal support.

4.1 Monophyly Recovery

The recuperation of the monophyly of all serotypes depended more on the method of tree reconstruction, than any of the other two variables: sampling size and genomic partition. For instance, distance methods always showed a complete recovery of the serotypes as monophyletic groups, in all subsampling replics and genomic partitions. This is congruent with the initial proposals of serotypification based primarily on the genetic distance (e.g. Rico-Hesse (1990); Pires Neto *et al.* (2005); Ito *et al.* (2007)). The monophyly of genotypes, on the other hand, was affected by the sampling size, being the total tips, the sample size with the smallest number of genotypes recovered. Although a higher taxonomic sampling is often related to a higher phylogenetic accuracy when given a reference classification (e.g. Zwickl *et al.* (2002); Heath *et al.* (2008)), the amount of disrupting taxa can also increase when more tips are included in the tree reconstruction. Additionally, some authors proposed the idea that the proportion of sampled

taxa within a taxonomic group could be more important than the total number of used taxa (Yang and Goldman (1997); Hillis (1998)). In the present study, we took both ideas into consideration in order to reduce any possible bias given the taxonomic sampling.

Regarding the effect of the genomic partition, Klungthong *et al.* (2008) studied the suitability and usefulness of each individual gene region for the molecular genotyping of DENV. They used 56 tips overall and found specific genes that recovered the majority of clades in each serotype. Cuypers *et al.* (2018) argues that only partitions such as *NS1*, *NS3*, and *NS5* provide more accurate topologies with a respectively high nodal support. Nonetheless, our results show that most of the partitions can successfully recover the monophyly of both serotypes and genotypes in DENV.

4.2 Common Nodes

Topologies inferred by parsimony had more nodes in common with the total evidence tree than the trees inferred by the other two methods. Nonetheless, regardless the method, the taxonomic sampling affected the common node count (phylogenetic accuracy) mostly for the first 13 partitions (all below 1000 nucleotides). This implies that as the number of taxa increases, the character sampling plays a more crucial role. Thence, the use of partitions such as the junction of the genes (*E-NS1*), or (*C-prM*/*M*); or genes as (*C*), (*NS4A*), (*NS2A*), and (*NS4B*) is not recommended if: the desired approach is towards the complete evidence, being 60% the minimal expected proportion of common nodes; and the taxonomic sampling comprises more than 100 tips. Otherwise, with a smaller number of tips, any of the genomic partitions presented in this study can be used, except (*E*-DIII) and (*C-prM*/*M*).

4.3 Topological similarity

According to our results, parsimony was the least sensitive method to taxonomic sampling, despite the genomic partition. The aforementioned implies that as the number of taxa is increased, the topological distance between the current topology and the total taxa topology almost decreases in half. This is consistent with previous observations (e.g. Simmons, 2012, 2014). The topologies reconstructed by Maximum Likelihood and distance methods showed that although the topological similarity increased with the number of taxa, these values highly depended on the genomic partition. Consequently, the largest partitions will yield similar topologies regardless the number of taxa.

One crucial questions in this study was whether there was a taxonomic congruence *sensu* Mickevich (1978) or not. As shown by the topological distance and indirectly, by the common nodes count and monophyly recovery; parsimony and Maximum Likelihood propose similar schemes of classification, ergo they are congruent. In fact, they display the same patterns regarding the monophyly recovery and the common nodes recovery, as the number of taxa and nucleotides increases, the recovery is higher, that so even the taxonomic sampling is no longer an issue. These results that were common for over four thousand topologies might indicate that the evolutionary rates of the sequences sites are heterogeneous or change non-identically over time (e.g. Steel and Penny (2000); Kolaczkowski and Thornton (2009)). Moreover, Tuffley and Steel (1997) proposed that these two methods yield equivalent topologies under the assumption of non-common mechanism. Also, DeBry and Abele (1995), demonstrated that with relatively long sequences, maximum parsimony was guaranteed to give the same estimate of the phylogenetic

tree as a Maximum-Likelihood estimator. Their analysis was based on 945 topologies inferred from ribosomal genes of 2000 nucleotides, for 8 tips.

Besides the methods of inference, we found that the ten largest genomic partitions reconstruct very similar topologies. The topological distance tree showed the same clustering for these partitions. Thus, we found two main clusters: one made up of the partitions *E*, structural genes, and the combination of *prM*/*M-E*; and the second made up of the second domain of *NS5*, *NS5*, the combination of *E-NS5*, the non-structural genes and the ORF. From all these partitions, the two that produced the most similar trees to the ones inferred from total evidence were the combination of the genes *E-NS5* and the combination of all non-structural genes.

It is important to clarify that although the largest genomic partitions always recuperated a high amount of nodes and clades, despite the method, the length of the genomic sequences might not be the main reason. The aforementioned, taking into account that there was not a linear trend between this variable and the recuperation values. As an example, the E gene, which is at least 300 nucleotides smaller than the three next larger partitions; always showed higher recuperation values and a closer relationship or similarity to the topologies inferred with the total evidence.

Comparatively, using this combination of *E-NS5* would be more advantageous in terms of not only phylogenetic accuracy, but also computational times. The aforementioned given the fact that this partition has about 4100 nucleotides, which is 6000 less than the ORF.

4.4 Node Support

We found that in spite of the taxonomic sampling, the longer the sequence of the genomic partition, the higher the proportion of nodes with the highest support values. Obtaining higher Bootstrap support values as the length of the sequences increases has been previously reported by e.g. Felsenstein (1985); Rosenberg and Kumar (2001); Rokas *et al.* (2003); Klötzl and Haubold (2016). Notwithstanding the fact that the taxonomic sampling also showed to cause a decrease in the support values as more taxa were included, this did not affect the largest genomic partitions. Consequently, the variable that weighs the most is the size of the genomic sequences. One unexpected result we obtained was parsimony showing higher support values than Maximum Likelihood, which has been reported as a method that tends to overestimate these values (e.g. Simmons and P Norton, 2013).

5. Concluding remarks

As recommended by Hedtke *et al.* (2006), instead of following a set of assumptions, we assessed our variables and measured the variation of the topological reconstruction in the recuperation of monophyly, common nodes, and node support.

In terms of recuperation of the monophyly of the four serotypes, all the methods of reconstruction accomplish it regardless the taxonomic sampling or genomic partition. Nonetheless, Distance methods are 100% efficient to recover these clades. Although the taxonomic sampling has no apparent effect in the recuperation of serotypes, it does affect the

recuperation of genotypes using the three methods of reconstruction, as it decreases as the number of taxa increases.

Regarding the phylogenetic accuracy to the complete evidence topology, both methods Parsimony and Maximum Likelihood obtain the highest proportions of common nodes. Furthermore, they are not affected by the taxonomic sampling when using the ten largest genomic partitions in the tree inference. Nevertheless, the length of the genomic partition is not the key factor to this, because some of the partitions such as the envelope gene consistently display higher recuperation values in contrast to some of the longest partitions. For instance, the combination of the genes *E* and *NS5* yield the most similar topology to the one inferred from the open reading frame. Using this partition would guarantee a recovery of more than the majority of the nodes of the total evidence topology and also, less computational time for the analysis.

Finally, through Parsimony and Maximum Likelihood we obtain congruent schemes of phylogenetic relationships. These nodes present the highest support values, when the topologies are reconstructed from partitions that are longer than 1000 nucleotides. In spite of the method, when we use shorter genomic partitions, the support values decrease when adding more tips into the phylogeny. Thereby, considering the overall effects of these three variables, we propose to take these results into account when doing the phylogenetic reconstruction of DENV and further analyses that take this reconstruction as a starting point.

References

Afreen, N., Deeba, F., Khan, W. H., Haider, S. H., Kazim, S. N., Ishrat, R., Naqvi, I. H., Shareef, M. Y., Broor, S., Ahmed, A. and Parveen, S. (2014) 'Molecular characterization of dengue and chikungunya virus strains circulating in New Delhi, India.', Microbiology and immunology, 58(12), 688-696.

Alfonso, H. L., Amarilla, A. a., Goncalves, P. F., Barros, M. T., Almeida, F. T. D., Silva, T. R., Silva, E. V. D., Nunes, M. T., Vasconcelos, P. F. C., Vieira, D. S., Batista, W. C., Bobadilla, M. L., Vazquez, C., Moran, M., Figueiredo, L. T. and Aquino, V. H. (2012) 'Phylogenetic relationship of dengue virus type 3 isolated in Brazil and Paraguay and global evolutionary divergence dynamics', Virology Journal, 9(1), 124.

Allicock, O. M., Lemey, P., Tatem, A. J., Pybus, O. G., Bennett, S. N., Mueller, B. A., Suchard, M. A., Foster, J. E., Rambaut, A. and Carrington, C. V. F. (2012) 'Phylogeography and population dynamics of dengue viruses in the Americas', *Molecular Biology and Evolution*, *29*(6), 1533-1543.

Anoop, M., Issac, A., Mathew, T., Philip, S., Kareem, N. A., Unnikrishnan, R. and Sreekumar, E. (2010) 'Genetic characterization of dengue virus serotypes causing concurrent infection in an outbreak in Ernakulam, Kerala, South India.', *Indian journal of experimental biology*, *48*(8), 849-857.

Aquino, V. H., Amarilla, A. A., Alfonso, H. L., Batista, W. C. and Figueiredo, L. T. M.

(2009) ‘New Genotype of Dengue Type 3 Virus Circulating in Brazil and Colombia Showed a Close Relationship to Old Asian Viruses’, *PLoS ONE*, *4*(10), e7299.

Araujo, J. M. G., Nogueira, R. M. R., Schatzmayr, H. G., Zanotto, P. M. D. a. and Bello, G. (2009) ‘Phylogeography and evolutionary history of dengue virus type 3.’, Infection, genetics and evolution : *Journal of molecular epidemiology and evolutionary genetics in infectious diseases*, *9*(4), 716-25.

Aviles, G., Meissner, J., Mantovani, R. and St. Jeor, S. (2003) ‘Complete coding sequences of dengue-1 viruses from Paraguay and Argentina’, *Virus Research*, *98*(1), 75- 82.

Azhar, E. I., Hashem, A. M., El-Kafrawy, S. A., Abol-Ela, S., Abd-Alla, A. M. M., Sohrab, S. S., Farraj, S. A., Othman, N. A., Ben-Helaby, H. G., Ashshi, A., Madani, T. A. and Jamjoom, G. (2015) ‘Complete genome sequencing and phylogenetic anal-ysis of dengue type 1 virus isolated from Jeddah, Saudi Arabia’, *Virology Journal*, *12*(1), 1-11.

Barrett, M., Donoghue, M. J. and Sober, E. (1991) ‘Against consensus’, Systematic *Zoology*, *40*(4), 486-493.

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., Drake, J. M., Brownstein, J. S., Hoen, A. G., Sankoh, O., Myers, M. F., George, D. B., Jaenisch, T., Wint, G. R. W., Simmons, C. P., Scott, T. W., Farrar, J. J. and Hay

S. I. (2013) 'The global distribution and burden of dengue', *Nature*, *496*(7446), 504- 507.

Boni, M. F., Posada, D. and Feldman, M. W. (2007) 'An Exact Nonparametric Method for Inferring Mosaic Structure in Sequence Triplets.'.

Brister, J. R., Ako-adjei, D., Bao, Y. and Blinkova, O. (2015) 'NCBI Viral Genomes Resource', *Nucleic Acids Research*, *43*(Database issue), D571-D577.

Brister, J. R., Bao, Y., Zhdanov, S. A., Ostapchuck, Y., Chetvernin, V., Kiryutin, B., Zaslavsky, L., Kimelman, M. and Tatusova, T. A. (2014) 'Virus Variation Resource-recent updates and future directions.', *Nucleic acids research*, *42*(Database issue), D660-5.

Bruen, T. C., Philippe, H. and Bryant, D. (2006) 'A simple and robust statistical test for detecting the presence of recombination', *Genetics*, *172*(4), 2665-2681.

Bryant, D., Janowitz, F. M., Lapointe, F., McMorris, R. F., Mirkin, B. and Roberts, S. (2003) A classification of consensus methods for phylogenetics, in A. M. Society, ed., '*Bioconsensus*, volume 61', Rhode Island, pp. 163-183.

Caceres, C., Yung, V., Araya, P., Tognarelli, J., Villagra, E., Vera, L. and Fernandez, J. (2008) 'Complete nucleotide sequence analysis of a Dengue-1 virus isolated on Easter Island, Chile.', *Archives of virology*, *153*(10), 1967-70.

Camin, J. H. and Sokal, R. R. (1965) 'A method for deducing branching sequences in phylogeny', *Evolution*, *19*(3), 311-326.

Castoe, T. A., Doan, T. M., Parkinson and L., C. (2004) 'Data Partitions and Complex Models in Bayesian Analysis: The Phylogeny of Gymnopthalmid Lizards', *Syst Biol*, *53*(3), 448-469.

Cavalli-Sforza, L. L. and Edwards, A. W. (1967) 'Phylogenetic analysis. Models and estimation procedures', *American journal of human genetics*, *19*(3 Pt 1), 233-257.

Chen, R. and Vasilakis, N. (2011) 'Dengue-quo tu et quo vadis?', *Viruses*, *3*(9), 1562- 608.

Chungue, E., Polynesia, F. and Kouri, P. (1995) 'Molecular epidemiology of dengue-1 and dengue-4. *Viruses*', *2*, 1877-1884.

Ciccozzi, M., Lo Presti, A., Cella, E., Giovanetti, M., Lai, A., El-Sawaf, G., Faggioni, G., Vescio, F., Al Ameri, R., De Santis, R., Helaly, G., Pomponi, A., Metwally, D., Fantini, M., Qadi, H., Zehender, G., Lista, F. and Rezza, G. (2014) 'Phylogeny of dengue and chikungunya viruses in Alhudayda governorate, yemen', *Infection, Genetics and Evolution*, *27*, 395-401.

Costa, R. L., Voloch, C. M. and Schrago, C. G. (2012) 'Comparative evolutionary epidemiology of dengue virus serotypes', Infection, *Genetics and Evolution, 12*(2), 309- 314.

Cuypers, L., Libin, P. J. K., Simmonds, P., Nowe, A., Mu~noz-Jordan, J., Alcantara, L. C. J., Vandamme, A.-M., Santiago, G. A. and Theys, K. (2018) 'Time to Harmonize Dengue Nomenclature and Classification', *Viruses 10*(10),, 569.

Dash, P. K., Sharma, S., Soni, M., Agarwal, A., Sahni, A. K. and Parida, M. (2015) 'Complete genome sequencing and evolutionary phylogeography analysis of Indian isolates of Dengue virus type 1.', *Virus research, 195*, 124-134.

DeBry, R. and Abele, L. (1995) 'The relationship between parsimony and maximum-likelihood analyses: Tree scores and confidence estimates for three real data sets', Molecular Biology and Evolution 12, 291-297.

DeFilippis, V. R. and Moore, W. S. (2000) 'Resolution of phylogenetic relationships among recently evolved species as a function of amount of DNA sequence: an empirical study based on woodpeckers (Aves: Picidae).', *Molecular phylogenetics and evolution, 16*(1), 143-160.

Degnan, J. H. and Rosenberg, N. A. (2009) 'Gene tree discordance, phylogenetic inference and the multispecies coalescent.', *Trends in ecology & evolution, 24*(6), 332-340.

Dettogni, R. S. and Louro, I. D. (2012) 'Phylogenetic characterization of Dengue virus type 2 in Espirito Santo, Brazil.', *Molecular biology reports 39*(1), 71-80.

Edgar, R. C. (2004a) 'MUSCLE: a multiple sequence alignment method with reduced time and space complexity', *BMC bioinformatics 5*, 113.

Edgar, R. C. (2004b) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research 32*(5), 1792-1797.

Edgar, R. C. (2010) 'Search and clustering orders of magnitude faster than BLAST', *Bioinformatics 26*(19), 2460-2461.

Erixon, P., Svennblad, B., Britton, T. and Oxelman, B. (2003) 'Reliability of Bayesian Posterior Probabilities and Bootstrap Frequencies in Phylogenetics', *Systematic Biology 52*(5), 665- 673.

Ernst, T., McCarthy, S., Chidlow, G., Luang-Suarkia, D., Holmes, E. C., Smith, D. W. and Imrie, A. (2015) 'Emergence of a New Lineage of Dengue Virus Type 2 Identified in Travelers Entering Western Australia from Indonesia, 2010-2012', *PLoS Negl Trop Dis 9*(1), e0003442.

Fansiri, T., Pongsiri, A., Klungthong, C., Ponlawat, A., Thaisomboonsuk, B., Jarman, R. G., Scott, T. W. and Lambrechts, L. (2016) 'No evidence for local adaptation of dengue viruses to mosquito vector populations in Thailand', *Evolutionary Applications 9*(4), 608-618.

Farris, J. S. (1970) 'Methods for computing wagner trees', *Systematic Zoology 19*(1), 83- 92.

Felsenstein, J. (1981) 'Evolutionary trees from DNA sequences: A maximum likelihood approach', *Journal of Molecular Evolution 17*(6), 368-376.

Felsenstein, J. (1985) 'CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP.', *Evolution; international journal of organic evolution 39*(4), 783-791.

Fitch, W. M. (1971) 'Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology', *Systematic Biology 20*(4), 406-416.

Gadagkar, S. R., Rosenberg, M. S. and Kumar, S. (2005) 'Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree.', Journal of experimental zoology. Part B, *Molecular and developmental evolution 304*(1), 64- 74.

Gascuel, O. (1997) 'BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data.', *Molecular Biology and Evolution 14*(7), 685-695.

Gascuel, O. and Berry, V. (1996) 'On the Interpretation of Bootstrap Trees: Appropri-ate Threshold of Clade Selection and Induced Gain', *Molecular Biology and Evolution 13*(7), 999.

Gibbs, M. J., Armstrong, J. S. and Gibbs, A. J. (2000) 'Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences.', *Bioinformatics (Oxford, England) 16*(7), 573-582.

Goldberg, T. L. (2003) 'Application of phylogeny reconstruction and character-evolution analysis to inferring patterns of directional microbial transmission', *Preventive Veterinary Medicine 61*(1), 59-70.

Goloboff, P. A. (1997) 'Self-weighted optimization: Tree searches and character state reconstructions under implied transformation costs', *Cladistics 13*(3), 225 - 245.

Goloboff, P. A., Farris, J. S. and Nixon, K. C. (2008) 'TNT, a free program for phylogenetic analysis', *Cladistics 24*(5), 774-786.

Gubler, D. J. (2002) 'Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century', *Trends in Microbiology 10*(2), 100- 103.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O.(2010) 'New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0', *Systematic Biology 59*(3), 307-321.

Hall, B. G. (2005) 'Comparison of the Accuracies of Several Phylogenetic Methods Using Protein and DNA Sequences', *Molecular Biology and Evolution 22*(3), 792-802.

Hapuarachchi, H. C., Koo, C., Kek, R., Xu, H., Lai, Y. L., Liu, L., Kok, S. Y., Shi, Y., Chuen, R. L. T., Lee, K.-S., Maurer-Stroh, S. and Ng, L. C. (2016) 'Intra-epidemic evolutionary dynamics of a Dengue virus type 1 population reveal mutant spectra that correlate with disease transmission', *Scientific Reports 6*(October 2015), 22592.

Heath, T. a., Hedtke, S. M. and Hillis, D. M. (2008) 'Taxon sampling and the accuracy of phylogenetic analyses', *Journal of Systematics and Evolution* 46, 239-257.

Hedtke, S. M., Townsend, T. M., Hillis, D. M. and Collins, T. (2006) 'Resolution of Phylogenetic Conflict in Large Data Sets by Increased Taxon Sampling', *Systematic Biology 55*(3), 522-529.

Hillis, D. M. (1998) 'Taxonomic Sampling, Phylogenetic Accuracy, and Investigator Bias', *Systematic Biology 47*(1), 3-8.

Holder, M. T., Sukumaran, J. and Lewis, P. O. (2008) 'A Justification for Reporting the Majority-Rule Consensus Tree in Bayesian Phylogenetics', *Systematic Biology 57*(5), 814- 821.

Holmes, E. C. (2003) 'Patterns of Intra- and Interhost Nonsynonymous Variation Reveal Strong Purifying Selection in Dengue Virus Patterns of Intra- and Interhost Non-synonymous Variation Reveal Strong Purifying Selection in Dengue Virus', *77*(20), 1-4.

Holmes, E. C. (2004) 'The phylogeography of human viruses.', *Molecular ecology 13*(4), 745-756.

Holmes, E. C. and Twiddy, S. S. (2003) 'The origin, emergence and evolutionary genetics of dengue virus', Infection, *Genetics and Evolution 3*(1), 19-28.

Holmes, E. C., Worobey, M. and Rambaut, A. (1999) 'Phylogenetic evidence for recom- bination in dengue virus.', *Molecular Biology and Evolution 16*(3), 405-409.

Ito, M., Yamada, K.-I., Takasaki, T., Pandey, B., Nerome, R., Tajima, S., Morita, K. and Kurane, I. (2007) 'Phylogenetic analysis of dengue viruses isolated from imported dengue patients: Possible aid for determining the countries where infections occurred', *Journal of Travel Medicine 14(*4), 233-244.

Jenkins, G. M., Rambaut, A., Pybus, O. G. and Holmes, E. C. (2002) 'Rates of molec-ular evolution in RNA viruses: A quantitative phylogenetic analysis', *Journal of Molecular Evolution 54*(2), 156-165.

Klötzl, F. and Haubold, B. (2016) 'Support Values for Genome Phylogenies', *Life (Basel, Switzerland) 6*(1), 11.

Klungthong, C., Putnak, R., Mammen, M. P., Li, T. and Zhang, C. (2008) 'Molecular genotyping of dengue viruses by phylogenetic analysis of the sequences of individual genes.', *Journal of virological methods 154*(1-2), 175-81.

Kolaczkowski, B. and Thornton, J. W. (2004) 'Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous', *Nature 431*(7011), 980- 984.

Kolaczkowski, B. and Thornton, J. W. (2009) 'Long-Branch Attraction Bias and In-consistency in Bayesian Phylogenetics', *PLOS ONE 4*(12), 1-12.

Kolekar, P., Kale, M. and Kulkarni-Kale, U. (2012) 'Alignment-free distance measure based on return time distribution for sequence analysis: Applications to clustering, molecular phylogeny and subtyping', *Molecular Phylogenetics and Evolution 65*(2), 510-522.

Kutty, S. N., Bernasconi, M. V., Sifner, F. and Meier, R. (2007) 'Sensitivity analysis, molecular systematics and natural history evolution of Scathophagidae (Diptera: Cyclorrhapha: Calyptratae)', *Cladistics 23*(1), 64-83.

Laille, M. and Roche, C. (2004) 'Comparisons of Dengue-1 virus envelope glycoprotein gene sequences from French Polynesia', *The American Journal of Tropical Medicine and Hygiene 71*(4), 478-484.

Lam, T. T.-Y., Hon, C.-C. and Tang, J. W. (2010) 'Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections.', *Critical reviews in clinical laboratory sciences 47*(1), 5-49.

Lanciotti, R. S., Lewis, J. G., Gubler, D. J. and Trent, D. W. (1994) 'Molecular evolution and epidemiology of dengue-3 viruses.', *The Journal of general virology 75* (Pt 1), 65-75.

Leache, A. D. and Reeder, T. W. (2002) 'Molecular Systematics of the Eastern Fence Lizard (Sceloporus undulatus): A Comparison of Parsimony, Likelihood, and Bayesian Approaches', *Systematic Biology 51*(1), 44-68.

Lee, K.-S., Lo, S., Tan, S. S.-Y., Chua, R., Tan, L.-K., Xu, H. and Ng, L.-C. (2012) 'Dengue virus surveillance in Singapore reveals high viral diversity through multiple introductions and in situ evolution', *Infection, Genetics and Evolution 12*(1), 77-85.

Lequime, S., Fontaine, A., Ar Gouilh, M., Moltini-Conclois, I. and Lambrechts, L. (2016) 'Genetic Drift, Purifying Selection and Vector Genotype Shape Dengue Virus Intra-host Genetic Diversity in Mosquitoes', *PLoS Genetics 12*(6), e1006111.

Lutzoni, F., Wagner, P., Reeb, V. and Zoller, S. (2000) 'Integrating Ambiguously Aligned Regions of DNA Sequences in Phylogenetic Analyses without Violating Po-sitional Homology', *Systematic Biology 49*(4), 628-651.

Martin, D. P., Murrell, B., Golden, M., Khoosal, A. and Muhire, B. (2015) 'RDP4: Detection and analysis of recombination patterns in virus genomes', *Virus Evolution 1*(1), 1-5.

Martin, D. and Rybicki, E. (2000) 'RDP: detection of recombination amongst aligned sequences.', *Bioinformatics (Oxford, England) 16*(6), 562-563.

Messer, W. B., Gubler, D. J., Harris, E., Sivananthan, K. and de Silva, A. M. (2003) 'Emergence and Global Spread of a Dengue Serotype 3, Subtype III Virus', *Emerging Infectious Disease journal 9*(7), 800.

Miagostovich, M. P., dos Santos, F. B., de Simone, T. S., Costa, E. V., Filippis, a. M. B., Schatzmayr, H. G. and Nogueira, R. M. R. (2002) 'Genetic characterization of dengue virus type 3 isolates in the State of Rio de Janeiro, 2001.', *Brazilian journal of medical and*

biological research = Revista brasileira de pesquisas medicas e biologicas/ Sociedade Brasileira de Biof sica ... [et al.] 35(8), 869-72.

Mickevich, M. F. (1978) 'Taxonomic Congruence', *Systematic Biology 27*(2), 143-158.

Miyamoto, M. M. and Fitch, W. M. (1995) 'Testing species phylogenies and phylogenetic methods with congruence', *Systematic Biology 44*(1), 64-76.

Mondini, A., Bronzoni, R. V. d. M., Nunes, S. H. P., Chiaravalloti Neto, F., Massad, E., Alonso, W. J., Lazzaro, E. S. M., Ferraz, A. A., de Andrade Zanotto, P. M. and Nogueira, M. L. (2009) 'Spatio-Temporal Tracking and Phylodynamics of an Urban Dengue 3 Outbreak in S~ao Paulo, Brazil', *PLOS Neglected Tropical Diseases 3*(5), e448

Nelson, G. (1979) 'Cladistic analysis and synthesis: Principles and definitions, with a historical note on adanson's familles des plantes (1763-1764)', *Systematic Biology 28*(1), 1-21.

Nur Liyana, K., Fauziah, M., Zainah, S., MatRahim, N. A., Salbiah, N., Lau, I., Ong, Y., Khadijah, N., Sharifah Aishah, W., Othman, S. and Thayan, R. (2016) 'Sequence analysis of E / NS1 gene junction of Dengue Type- 1 viruses isolated in Klang Valley 2010 to 2012', *33*(2), 1-11.

Padidam, M., Sawyer, S. and Fauquet, C. M. (1999) 'Possible emergence of new geminiviruses by frequent recombination.', *Virology 265*(2), 218-225.

Paradis, E., Claude, J. and Strimmer, K. (2004) 'APE: Analyses of Phylogenetics and Evolution in R language', *Bioinformatics 20*(2), 289-290.

Parameswaran, P., Charlebois, P., Tellez, Y., Nunez, A., Ryan, E. M., Malboeuf, C. M., Levin, J. Z., Lennon, N. J., Balmaseda, A., Harris, E. and Henn, M. R. (2012) 'Genome-wide patterns of intrahuman dengue virus diversity reveal associations with viral phylogenetic clade and interhost diversity.', *Journal of virology 86*(16), 8546-58.

Pearson, K. (1896) 'Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia', *Philosophical Transactions of the Royal Society of London Series A 187*, 253-318.

Petronio, J. A. G., Vinarao, R. B., Flores, K. M. G. and Destura, R. V. (2014) 'Continued circulation of a single genotype of dengue virus serotype 2 in the Philippines', Asian Pacific *Journal of Tropical Medicine 7*(1), 30-33.

Pires Neto, R. J., Lima, D. M., de Paula, S. O., Lima, C. M., Rocco, I. M. and Fonseca, B. A. L. (2005) 'Molecular epidemiology of type 1 and 2 dengue viruses in Brazil from 1988 to

2001.', Brazilian journal of medical and biological research = *Revista brasileira de pesquisas medicas e biologicas / Sociedade Brasileira de Biofisica ... [et al.] 38*(6), 843-852.

Planet, P. J. (2006) 'Tree disagreement: Measuring and testing incongruence in phylo-genies', *Journal of Biomedical Informatics 39*(1), 86-102.

Popescu, A.-A., Huber, K. T. and Paradis, E. (2012) 'ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R', *Bioinformatics 28*(11), 1536- 1537.

Posada, D. (2001) 'Unveiling the molecular clock in the presence of recombination.'.

Posada, D. and Crandall, K. A. (2001) 'Evaluation of methods for detecting recombination from DNA sequences: Computer simulations.'.

R Core Team (2019) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/

Ramírez, M. J. (2003). The Spider Subfamily Amaurobioidinae (Araneae, Anyphaenidae): a Phylogenetic Revision At the Generic Level', *Bulletin of the American Museum of Natural History*, 277, 1-262

Regato, M., Recarey, R., Moratorio, G., de Mora, D., Garcia-Aguirre, L., Gonzalez, M., Mosquera, C., Alava, A., Fajardo, A., Alvarez, M., D' Andrea, L., Dubra, A., Mart nez, M., Khan, B. and Cristina, J. (2008) 'Phylogenetic analysis of the NS5 gene of dengue viruses isolated in Ecuador', *Virus Research 132*(1-2), 197-200.

Rico-Hesse, R. (1990) 'Molecular evolution and distribution of dengue viruses type 1 and 2 in nature', *Virology 174*(2), 479-493.

Robinson, D. F. and Foulds, L. R. (1981) 'Comparison of phylogenetic trees', *Mathematical Biosciences 53*, 131-147.

Rodpothong, P. and Auewarakul, P. (2012) 'Positive selection sites in the surface genes of dengue virus: phylogenetic analysis of the interserotypic branches of the four serotypes.', *Virus genes 44*(3), 408-414.

Rodriguez-Roche, R., Hinojosa, Y. and Guzman, M. G. (2014) 'First dengue haemorrhagic fever epidemic in the Americas, 1981: insights into the causative agent', *Archives of virology 159*(12), 3239-3247.

Rokas, A., Williams, B. L., King, N. and Carroll, S. B. (2003) 'Genome-scale approaches to resolving incongruence in molecular phylogenies', *Nature 425*(6960), 798-804.

Rosenberg, M. S. and Kumar, S. (2001) 'Incomplete taxon sampling is not a problem for phylogenetic inference.', *Proceedings of the National Academy of Sciences of the United States of America 98*(19), 10751-10756.

Ross, T. M. (2010) 'Dengue Virus', *Clinics in Laboratory Medicine 30*(1), 149-160.

Saitou, N. and Nei, M. (1987) 'The neighbor-joining method: a new method for recon-structing phylogenetic trees.', *Molecular Biology and Evolution 4*(4), 406-425.

Salminen, M. O., Carr, J. K., Burke, D. S. and McCutchan, F. E. (1995) 'Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning.', *AIDS research and human retroviruses 11*(11), 1423-1425.

Schierup, M. H. and Hein, J. (2000a) 'Consequences of recombination on traditional phylogenetic analysis.', *Genetics 156*(2), 879-891.

Schierup, M. H. and Hein, J. (2000b) 'Recombination and the molecular clock.'.

Schliep, K. P. (2011) 'phangorn: phylogenetic analysis in R', *Bioinformatics 27*(4), 592- 593.

Schreiber, M. J., Holmes, E. C., Ong, S. H., Soh, H. S. H., Liu, W., Tanner, L., Aw, P. P. K., Tan, H. C., Ng, L. C., Leo, Y. S., Low, J. G. H., Ong, A., Ooi, E. E., Vasudevan, S. G. and

Hibberd, M. L. (2009) 'Genomic epidemiology of a dengue virus epidemic in urban Singapore.', *Journal of virology 83*(9), 4163-73.

Simmons, M. P. (2012) 'Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data', Molecular *Phylogenetics and Evolution 62*(1), 472-484.

Simmons, M. P. (2014) 'A confounding e ect of missing data on character Conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses.', *Molecular phylogenetics and evolution* 80, 267-280.

Simmons, M. and P Norton, A. (2013) 'Quanti cation and relative severity of in ated branch-support values generated by alternative methods: An empirical example', *Molecular phylogenetics and evolution 67*.

Simmons, M. P., Pickett, K. M. and Miya, M. (2004) 'How Meaningful Are Bayesian Support Values?', *Molecular Biology and Evolution 21*(1), 188-199.

Simmons, N. B. and Geisler, J. H. (2002) 'Sensitivity analysis of different methods of coding taxonomic polymorphism: an example from higher-level bat phylogeny', *Cladistics 18*(6), 571-584.

Smith, J. M. (1992) ‘Analyzing the mosaic structure of genes.’, *Journal of molecular evolution 34*(2), 126-129.

Sokal, R. and Sneath, P. H. (1963) Numerical taxonomy, W. H. Freeman, San Francisco.

Stamatakis, A. (2006) ‘RAxML-VI-HPC: maximum likelihood-based phylogenetic anal-yses with thousands of taxa and mixed models.’, *Bioinformatics (Oxford, England) 22*(21), 2688-90

Stamatakis, A. (2014) ‘RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.’, *Bioinformatics (Oxford, England) 30*(9), 1312-3.

Steel, M. and Penny, D. (2000) ‘Parsimony, likelihood, and the role of models in molecular phylogenetics’, *Molecular Biology and Evolution 17*(6), 839-850.

Tolou, H. J. G., Couissinier-Paris, P., Durand, J.-P., Mercier, V., de Pina, J.-J., de Micco, P., Billoir, F., Charrel, R. N. and de Lamballerie, X. (2001) ‘Evidence for recombination in natural populations of dengue virus type 1 based on the analy-sis of complete genome sequences’, *Journal of General Virology 82*(6), 1283-1290.

Tuffley, C. and Steel, M. (1997) 'Links between maximum likelihood and maximum par- simony under a simple model of site substitution.', *Bulletin of mathematical biology 593*, 581-607.

Twiddy, S. S., Farrar, J. J., Chau, N. V., Wills, B., Gould, E. A., Gritsun, T., Lloyd, G. and Holmes, E. C. (2002) 'Phylogenetic Relationships and Differential Selection Pressures among Genotypes of Dengue-2 Virus 1', *72*, 63-72.

Twiddy, S. S., Holmes, E. C. and Rambaut, a. (2003) 'Inferring the Rate and Time-Scale of Dengue Virus Evolution', *Molecular Biology and Evolution 20*(1), 122-129.

Usme-Ciro, J. a., Mendez, J. a., Tenorio, A., Rey, G. J., Domingo, C. and Gallego-Gomez, J. C. (2008) 'Simultaneous circulation of genotypes I and III of dengue virus 3 in Colombia.', *Virology journal 5*, 101

Villabona-Arenas, C. J. and de Andrade Zanotto, P. M. (2011) 'Evolutionary history of Dengue virus type 4: Insights into genotype phylodynamics', *Infection, Genetics and Evolution 11*(5), 878-885.

Villabona-Arenas, C. J., de Oliveira, J. L., de Capra, C. S., Balarini, K., Loureiro, M., Fonseca, C. R. T. P., Passos, S. D. and de Zanotto, P. M. A. (2014) 'Detection Of Four Dengue

Serotypes Suggests Rise In Hyperendemicity In Urban Centers Of Brazil', *PLoS Neglected Tropical Diseases 8*(2), 3-5.

Villabona-Arenas, C. J., de Oliveira, J. L., de Sousa-Capra, C., Balarini, K., Pereira da Fonseca, C. R. T. and Zanotto, P. M. d. A. (2016) 'Epidemiological dynamics of an urban Dengue 4 outbreak in Sao Paulo, Brazil', *PeerJ 4*, e1892.

Villabona-Arenas, C. J. and Zanotto, P. M. D. A. (2013) 'Worldwide spread of Dengue virus type 1.', *PloS one 8*(5), e62649.

Wang, B., Li, Y., Feng, Y., Zhou, H., Liang, Y., Dai, J., Qin, W., Hu, Y., Wang, Y., Zhang, L., Baloch, Z., Yang, H. and Xia, X. (2015) 'Phylogenetic analysis of dengue virus reveals the high relatedness between imported and local strains during the 2013 dengue outbreak in Yunnan, China: a retrospective analysis.', *BMC infectious diseases 15*, 142.

Wang, E., Ni, H., Xu, R., Barrett, a. D., Watowich, S. J., Gubler, D. J. and Weaver, S. C. (2000) 'Evolutionary relationships of endemic/epidemic and sylvatic dengue viruses.', *Journal of virology 74*(7), 3227-3234

Weiller, G. F. (1998) 'Phylogenetic pro les: a graphical method for detecting genetic re-combinations in homologous sequences.', *Molecular biology and evolution 15*(3), 326- 335.

WHO (2016) 'DENGUE AND SEVERE DENGUE'.

Yang, Z. and Goldman, N. (1997) 'Are big trees indeed easy?', Trends in Ecology & Evolution 12(9), 357.

Zaki, A., Perera, D., Jahan, S. S. and Cardosa, M. J. (2008) 'Phylogeny of dengue viruses circulating in Jeddah, Saudi Arabia: 1994 to 2006.', Tropical medicine & international health : TM & IH 13(4), 584-592.

Zwickl, D. J., Hillis, D. M. and Crandall, K. (2002) 'Increased Taxon Sampling Greatly Reduces Phylogenetic Error', Systematic Biology 51(4), 588-598.

Tables

Table 1.

Genomic partitions used in this study. Cells are shadowed according to the type of partition.

Name	Part.	Type	Size (bp)	Phylogenetic reconstructions using this partition
P01	E-NS1	cg	240	Zaki *et al*. (2008); Aquino *et al*. (2009); Nur Liyana *et al*. (2016)
P02	E-DIII	pd	285	Dettogni and Louro (2012); Afreen *et al*. (2014); Ciccozzi *et al*. (2014)
P03	C	sg	349	Klungthong *et al*. (2008)
P04	NS2B	nsg	392	Klungthong *et al*. (2008)
P05	NS5-D	pd	399	Mondini *et al*. (2009)
P06	E-DI	pd	423	Dettogni and Louro (2012); Afreen *et al*. (2014); Ciccozzi *et al*. (2014)
P07	C-prM/M	cg	427	Anoop *et al*. (2010); Petronio *et al*. (2014); Villabona-Arenas *et al*. (2014)
P08	NS4A	nsg	449	Klungthong *et al*. (2008)
P09	E-DII	pd	474	Dettogni and Louro (2012); Afreen *et al*. (2014); Ciccozzi *et al*. (2014)
P10	prM/M	sg	497	Klungthong *et al*. (2008); Rodpothong and Auewarakul (2012)
P11	NS2A	nsg	670	Klungthong *et al*. (2008)
P12	NS4B	nsg	758	no record of previous use
P13	NS5-DI	pd	786	Mondini *et al*. (2009)
P14	NS1	nsg	1055	Klungthong *et al*. (2008); Aquino *et al*. (2009); Ciccozzi *et al*. (2014)
P15	E	sg	1494	Wang *et al*. (2015); Fansiri *et al*. (2016); Villabona-Arenas *et al*. (2016)
P16	NS3	nsg	1871	Klungthong *et al*. (2008)
P17	NS5-DII	pd	1881	Mondini *et al*. (2009)
P18	prM/M-E	cg	2272	Lanciotti *et al*. (1994); Miagostovich *et al*. (2002); Messer *et al*. (2003)
P19	SG	cg	2340	no record of previous use
P20	NS5	nsg	2709	Holmes and Twiddy (2003); Regato *et al*. (2008); Klungthong *et al*. (2008)
P21	E-NS5	cg	4203	no record of previous use
P22	NSG	cg	7904	no record of previous use
P23	ORF	cg	10253	Rodriguez-Roche *et al*. (2014); Dash *et al*. (2015); Hapuarachchi *et al*. (2016)

Note: Part.: partition; sg: structural gene; nsg: non-structural gene; cg : combined genes; pd: protein domain. The size of partition is expressed as the number of base pairs (bp).

Table 2.

Classification schemes for the assignation of DENV genotypes

Serotype	Genotype	Reference
DENV1	**I:** Hawai; **II**: Tailandia; **III**: Malasia; **IV**: Australia/Pacífico Sur; **V**: América/África	Goncalvez *et al.*, 2002; Weaver & Vasilakis, 2009; Chen & Vasilakis, 2011
DENV2	**I**: Asiático II; **II:** Asiático I; **III:** Americano/Asiático; **IV:** Cosmopolita; **V:** Americano; **VI:** Selvático	Rico-Hesse *et al.*, 1997; Twiddy *et al.*, 2002; Weaver & Vasilakis, 2009; Chen & Vasilakis, 2011
DENV3	**I:** Indonesia; **II:** Tailandia; **III:** India; **IV:** Puerto Rico; **V:** Filipinas	Castro & Aquino, 2010; Chen & Vasilakis, 2011; Yasamita *et al.*, 2013
DENV4	**I:** Filipinas; **IIA:** China; **IIB:** Indonesia, Tahití, Nueva Caledonia; **III:** Tailandia; **IV:** Selvático	Weaver & Vasilakis, 2009; Chen & Vasilakis, 2011; Villabona-Arenas & Zanotto, 2011; Yasamita *et al.*, 2013

Table 3.

Taxonomic representation of DENV virus clades

Serotype	Genotype	Number of tips per sampling size (percentage)			
		100	**75**	**36**	**10**
DENV1	Divergent	**2**	**2**	**2**	**2**
	I	71	33	10	3
	II	22	22	10	3
	III	58	33	10	3
	Sylvatic	**2**	**2**	**2**	**2**
	Total	155	92	34	13
DENV2	Asian-American	36	33	7	3
	Asian-I	52	33	7	3
	Asian-II	**3**	**3**	**3**	**3**
	Cosmopolitan	7	7	7	3
	Divergent	**2**	**2**	**2**	**2**
	Sylvatic	16	15	7	3
	Total	116	93	33	17
DENV3	I	5	5	5	3
	II	4	4	4	3
	III	57	45	23	3
	V	13	13	13	3
	Total	79	67	45	12
DENV4	I	4	4	4	3
	II	51	45	25	3
	III	**3**	**3**	**3**	**3**
	Sylvatic	**2**	**2**	**2**	**2**
	Total	60	54	34	11
Total		410	306	146	53

Table 4.

Recuperation of the monophyly of all serotypes.

Monophyly of all serotypes					
Method	**Sampling**	**mean**	**sd**	**se**	**ci**
PA	T10	97.83	7.20	1.50	2.94
	T36	97.86	7.08	1.48	2.89
	T75	97.86	7.08	1.48	2.89
	T100	97.83	7.20	1.50	2.94
DI	T10	100.00	-	-	-
	T36	100.00	-	-	-
	T75	100.00	-	-	-
	T100	100.00	-	-	-
ML	T10	80.65	10.04	2.09	4.10
	T36	80.07	9.51	1.98	3.89
	T75	80.18	10.98	2.29	4.49
	T100	84.78	14.58	3.04	5.96

Note: Each case is shown with its inference method and taxonomic sampling.

Method PA: Parsimony; DI: Distance; ML: Maximum Likelihood. Sampling T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips.

Table 5.

Recuperation of DENV1 genotypes

Genotypes DENV1					
Method	**Sampling**	**mean**	**sd**	**se**	**ci**
PA	T10	57.39	12.92	2.69	5.28
	T36	59.19	4.20	0.88	1.72
	T75	55.77	7.30	1.52	2.99
	T100	51.30	11.80	2.46	4.82
DI	T10	63.59	7.07	1.47	2.89
	T36	61.71	11.42	2.38	4.67
	T75	60.09	10.13	2.11	4.14
	T100	60.87	16.49	3.44	6.74
ML	T10	52.55	11.18	2.33	4.57
	T36	49.83	14.65	3.05	5.99
	T75	46.32	15.49	3.23	6.33
	T100	44.35	15.90	3.32	6.50

Note: Each case is shown with its inference method and taxonomic sampling.

Method PA: Parsimony; DI: Distance; ML: Maximum Likelihood. Sampling T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips.

Table 6.

Recuperation of DENV2 genotypes

Genotypes DENV2					
Method	**Sampling**	**mean**	**sd**	**se**	**ci**
PA	T10	88.09	19.34	4.03	7.90
	T36	82.66	2.59	0.54	1.06
	T75	82.03	4.12	0.86	1.68
	T100	81.16	5.74	1.20	2.35
DI	T10	89.32	4.36	0.91	1.78
	T36	77.71	8.88	1.85	3.63
	T75	74.98	11.26	2.35	4.60
	T100	73.19	14.86	3.10	6.07
ML	T10	86.50	7.89	1.64	3.22
	T36	76.38	7.94	1.65	3.24
	T75	74.25	9.99	2.08	4.08
	T100	75.36	11.09	2.31	4.53

Note: Each case is shown with its inference method and taxonomic sampling.

Method PA: Parsimony; DI: Distance; ML: Maximum Likelihood. Sampling T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips.

Table 7.

Recuperation of DENV3 genotypes

Genotypes DENV3					
Method	**Sampling**	**mean**	**sd**	**se**	**ci**
PA	T10	48.04	10.61	2.21	4.34
	T36	42.72	10.63	2.22	4.34
	T75	43.15	11.11	2.32	4.54
	T100	42.39	11.76	2.45	4.81
DI	T10	50.29	3.36	0.70	1.37
	T36	41.16	10.38	2.16	4.24
	T75	26.63	7.95	1.66	3.25
	T100	26.09	9.16	1.91	3.75
ML	T10	42.93	8.78	1.83	3.59
	T36	38.95	11.29	2.35	4.61
	T75	37.72	12.34	2.57	5.04
	T100	38.04	14.83	3.09	6.06

Note: Each case is shown with its inference method and taxonomic sampling.

Method PA: Parsimony; DI: Distance; ML: Maximum Likelihood. Sampling T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips.

Table 8.

Recuperation of DENV4 genotypes

Genotypes DENV4					
Method	**Sampling**	**mean**	**sd**	**se**	**ci**
PA	T10	88.80	21.37	4.46	8.73
	T36	66.99	5.40	1.13	2.21
	T75	53.59	1.32	0.27	0.54
	T100	50.00	0.00	0.00	0.00
DI	T10	89.49	10.97	2.29	4.48
	T36	64.93	8.61	1.80	3.52
	T75	52.46	3.09	0.64	1.26
	T100	48.91	5.21	1.09	2.13
ML	T10	82.57	16.03	3.34	6.55
	T36	59.96	10.86	2.27	4.44
	T75	48.84	8.42	1.76	3.44
	T100	44.57	12.96	2.70	5.30

Note: Each case is shown with its inference method and taxonomic sampling.

Method PA: Parsimony; DI: Distance; ML: Maximum Likelihood. Sampling T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips.

Table 9.

Recuperation of all genotypes and serotypes

All serotypes and genotypes					
Method	**Sampling**	**mean**	**sd**	**se**	**ci**
PA	T10	79.02	2.27	0.47	0.93
	T36	70.53	2.90	0.60	1.18
	T75	67.37	3.48	0.72	1.42
	T100	65.41	4.99	1.04	2.04
DI	T10	79.02	2.27	0.47	0.93
	T36	70.53	2.90	0.60	1.18
	T75	67.37	3.48	0.72	1.42
	T100	65.41	4.99	1.04	2.04
ML	T10	79.02	2.27	0.47	0.93
	T36	70.53	2.90	0.60	1.18
	T75	67.37	3.48	0.72	1.42
	T100	65.41	4.99	1.04	2.04

Note: Each case is shown with its inference method and taxonomic sampling.

Method PA: Parsimony; DI: Distance; ML: Maximum Likelihood. Sampling T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips.

Table 10.

Average common nodes count with the ORF

Method	Sampling	mean	sd	se	ci
PA	T10	81.28	7.76	0.31	0.60
	T36	69.98	12.94	0.51	1.00
	T75	61.62	15.75	0.62	1.22
	T100	57.50	17.57	3.75	7.34
DI	T10	78.43	8.60	0.34	0.67
	T36	50.51	3.17	0.67	1.32
	T75	34.73	1.45	0.31	0.60
	T100	45.14	15.53	3.31	6.49
ML	T10	64.71	5.49	1.17	2.29
	T36	48.80	2.98	0.63	1.24
	T75	37.89	2.18	0.47	0.91
	T100	56.98	18.23	3.89	7.62

Note: Each case is shown with its inference method and taxonomic sampling.

Method PA: Parsimony; DI: Distance; ML: Maximum Likelihood. Sampling T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips.

Table 11.

Average RF distance for each case.

Against the total tips topology					
Method	**Sampling**	**mean**	**sd**	**error**	**ci**
PA	T10	10.71	2.67	0.49	0.95
	T36	7.98	1.52	0.28	0.54
	T75	4.54	0.89	0.16	0.32
DI	T10	17.43	4.49	0.82	1.61
	T36	39.56	3.06	0.56	1.09
	T75	27.74	3.38	0.62	1.21
ML	T10	25.01	5.51	1.01	1.97
	T36	24.14	4.34	0.79	1.55
	T75	23.73	3.03	0.55	1.08
Between methods					
Methods	**Sampling**	**mean**	**sd**	**error**	**ci**
PAML	T10	25.11	4.83	0.88	1.727
	T36	24.24	2.98	0.54	1.065
	T75	25.41	1.79	0.33	0.640
	T100	26.32	-	-	-
PADI	T10	22.52	4.17	0.76	1.491
	T36	33.70	2.67	0.49	0.957
	T75	41.60	1.57	0.29	0.563
	T100	42.58	-	-	-
MLDI	T10	32.45	5.17	0.94	1.850
	T36	45.22	3.59	0.65	1.283
	T75	54.61	2.10	0.38	0.751
	T100	57.39	-	-	-

Note: Each case is shown with its inference method and taxonomic sampling.

Method PA: Parsimony; DI: Distance; ML: Maximum Likelihood. Sampling T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips.

Table 12.

Average Bootstrap node support in the topologies.

Method	Sampling	Level	mean	sd	error	ci
PA	T10	Low	12.13	3.58	0.65	1.28
	T10	Moderate	13.36	3.75	0.68	1.34
	T10	**High**	**47.25**	**3.34**	**0.61**	**1.20**
	T36	Low	15.60	2.70	0.49	0.96
	T36	Moderate	15.72	2.68	0.49	0.96
	T36	**High**	**33.26**	**2.22**	**0.41**	**0.80**
	T75	Low	17.79	1.65	0.30	0.59
	T75	Moderate	16.30	1.51	0.27	0.54
	T75	**High**	**27.16**	**1.14**	**0.21**	**0.41**
	T100	Low	19.09	0.00	0.00	-
	T100	Moderate	15.98	0.00	0.00	-
	T100	**High**	**25.16**	**0.00**	**0.00**	**-**
DI	T10	Low	15.31	5.01	3.19	2.05
	T10	Moderate	14.58	4.55	3.04	1.86
	T10	**High**	**45.79**	**16.01**	**9.55**	**6.54**
	T36	Low	20.39	3.65	4.25	1.49
	T36	Moderate	16.41	1.89	3.42	0.77
	T36	**High**	**24.94**	**12.20**	**5.20**	**4.99**
	T75	Low	18.81	1.90	3.92	0.78
	T75	Moderate	15.00	2.64	3.13	1.08
	T75	**High**	**23.60**	**6.67**	**4.92**	**2.73**
	T100	Low	18.88	2.31	3.94	0.94
	T100	Moderate	14.03	3.11	2.93	1.27
	T100	**High**	**23.62**	**7.07**	**4.93**	**2.89**
ML	T10	Low	19.87	5.16	0.94	1.85
	T10	Moderate	22.32	4.68	0.86	1.68
	T10	**High**	**34.48**	**3.88**	**0.71**	**1.39**
	T36	Low	21.74	3.26	0.60	1.17
	T36	Moderate	20.07	2.92	0.53	1.05
	T36	**High**	**26.03**	**2.40**	**0.44**	**0.86**
	T75	Low	20.76	1.92	0.35	0.69
	T75	Moderate	17.78	1.66	0.30	0.60
	T75	**High**	**23.38**	**1.27**	**0.23**	**0.46**
	T100	Low	19.99	0.00	0.00	-
	T100	Moderate	16.52	0.00	0.00	-

T100	**High**	**21.76**	**0.00**	**0.00**	**-**

Cells are shadowed to emphasize on the high support level. Method PA: Parsimony; DI: Distance; ML: Maximum Likelihood. Sampling T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips. Level Low (50-75%); Moderate (76-94%); High (95-100%).

Figures

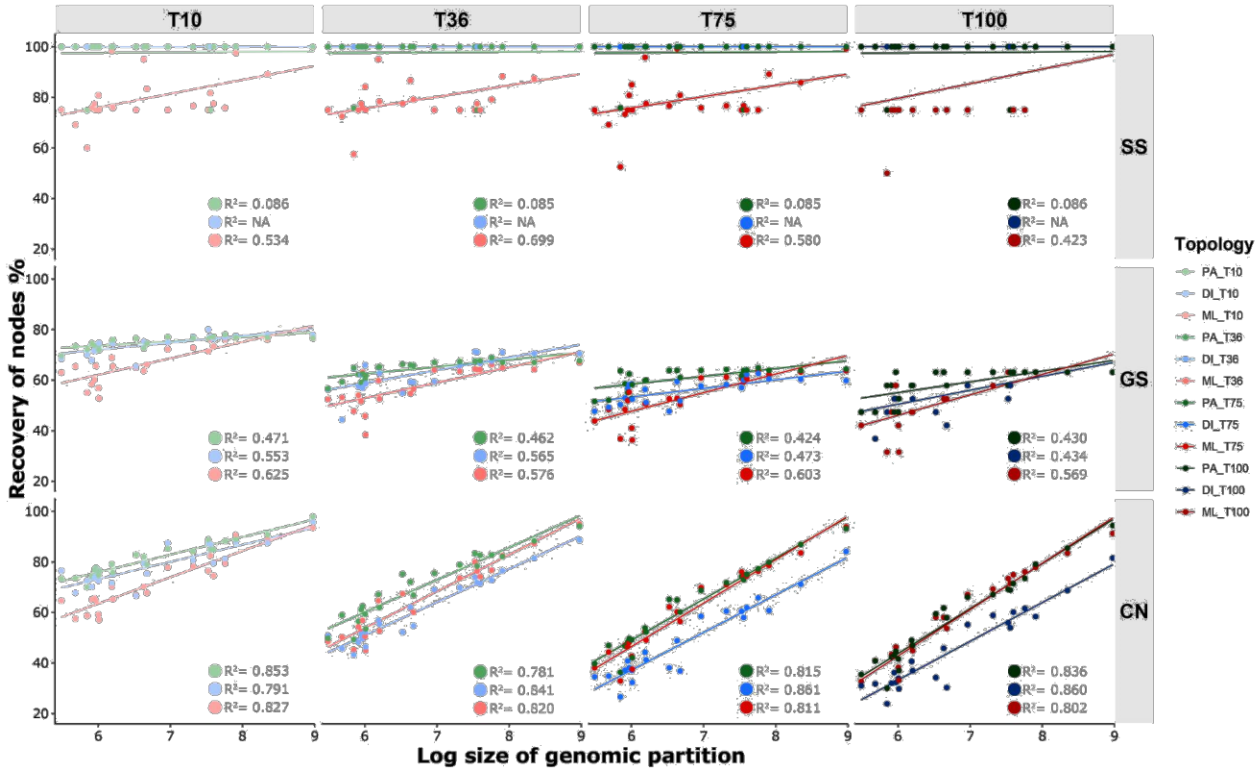

*Figure 1. Recuperation of monophyletic clades in DENV*S. All serotypes; GS. All genotypes; CN. Common nodes (having the ORF as the reference topology). Each case is depicted with its inference method (PA: Parsimony; DI: Distance; ML: Maximum Likelihood), and taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips). Linear regression coefficients (R^2) are reported for each case.

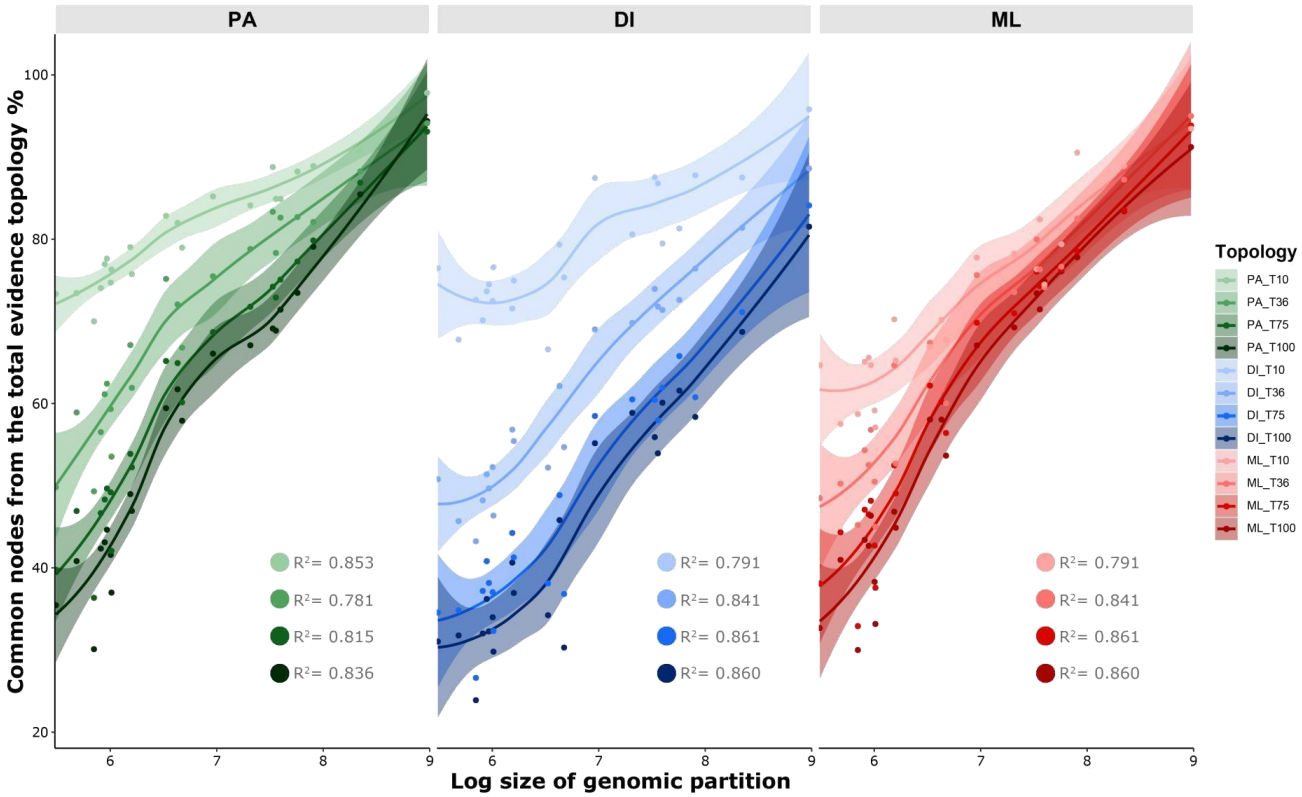

Figure 2. Effect of the taxonomic sampling on common nodes. Each case is depicted with its inference method (PA: Parsimony; DI: Distance; ML: Maximum Likelihood), and taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips). Linear regression coefficients (R^2) are reported for each case.

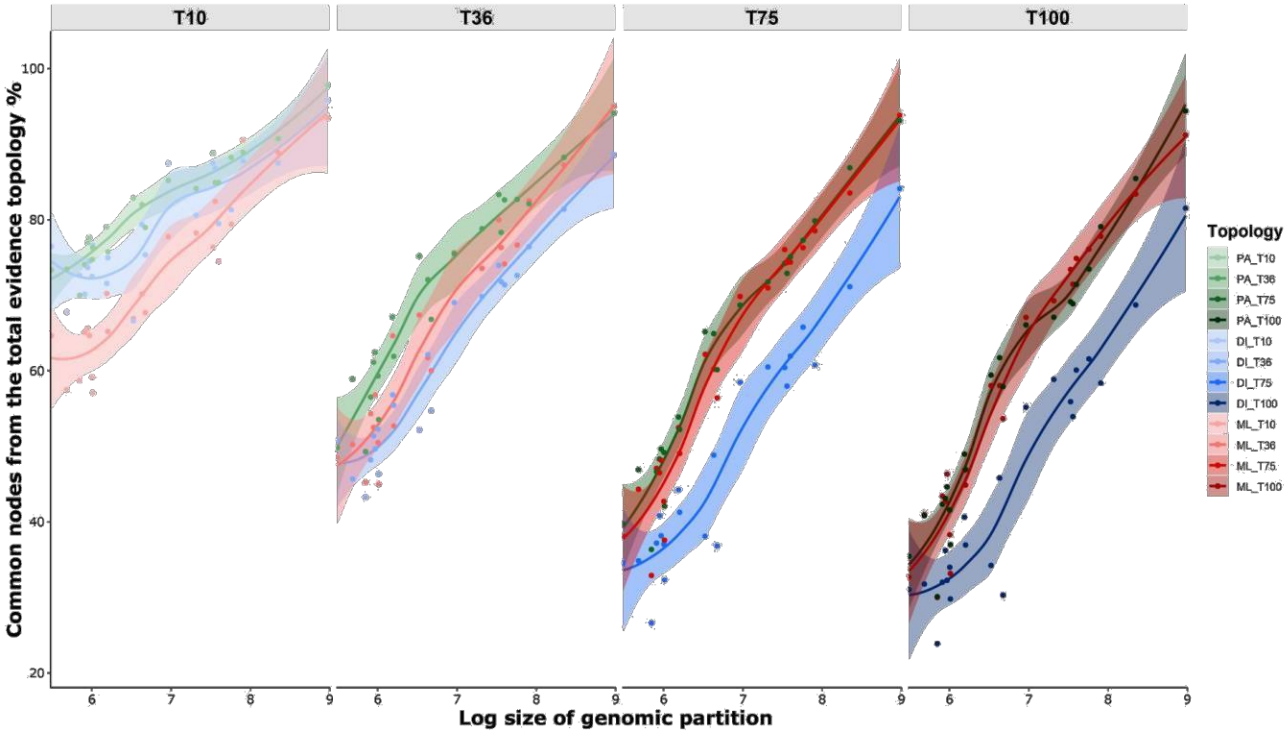

Figure 3. Effect of the tree inference method on common nodes. Each case is depicted with its inference method (PA: Parsimony; DI: Distance; ML: Maximum Likelihood), and taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips).

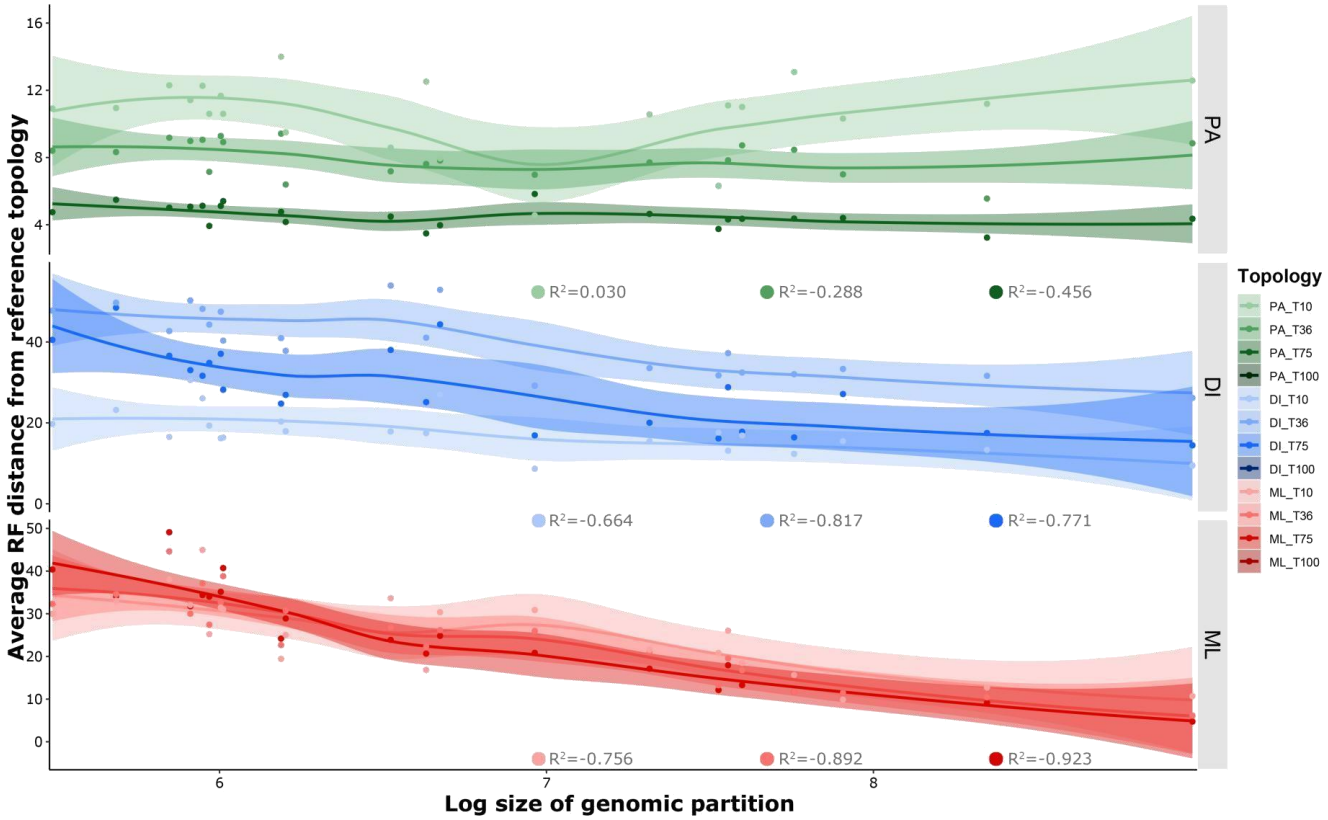

Figure 4. Effect of the taxonomic sampling on the topological distance. Each case is depicted with its inference method (PA: Parsimony; DI: Distance; ML: Maximum Likelihood), and taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips). Linear regression coefficients (R^2) are reported for each case.

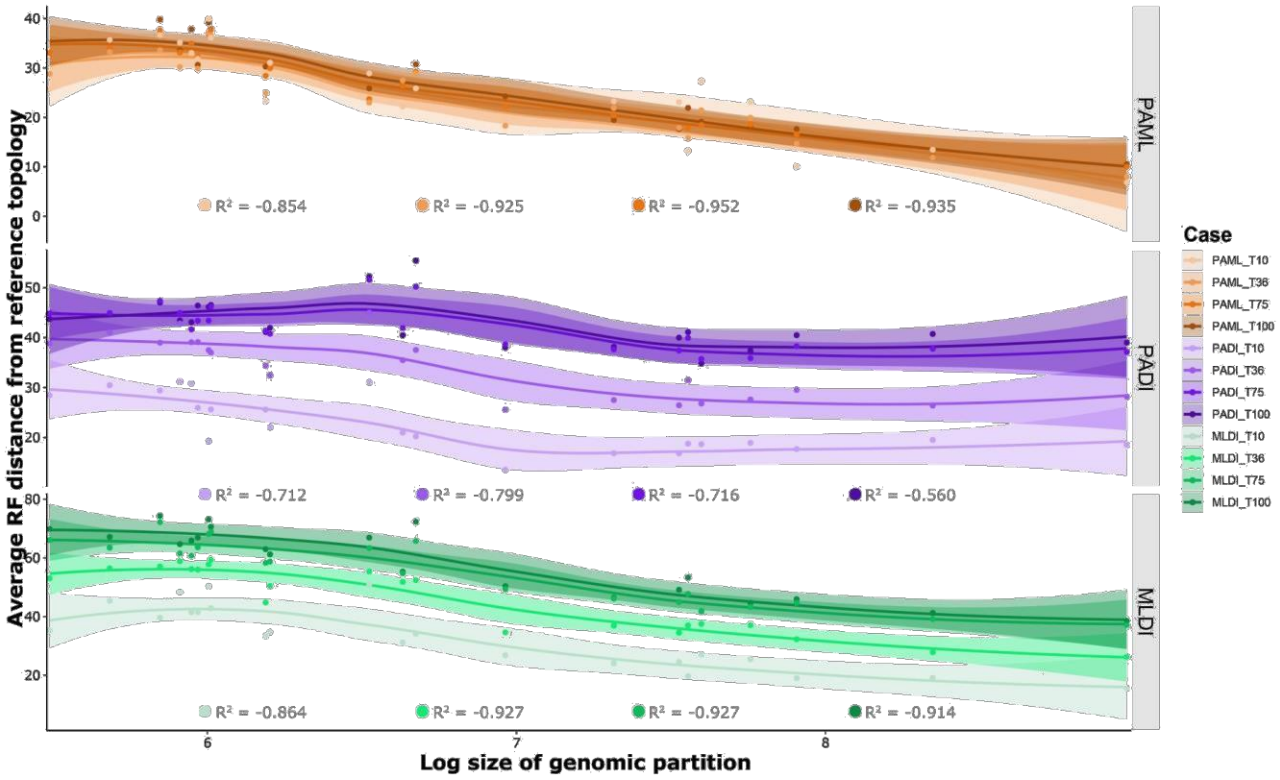

Figure 5. Topological similarity between methods. Each case is depicted with its inference method (PA: Parsimony; DI: Distance; ML: Maximum Likelihood), and taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips). Linear regression coefficients (R^2) are reported for each case.

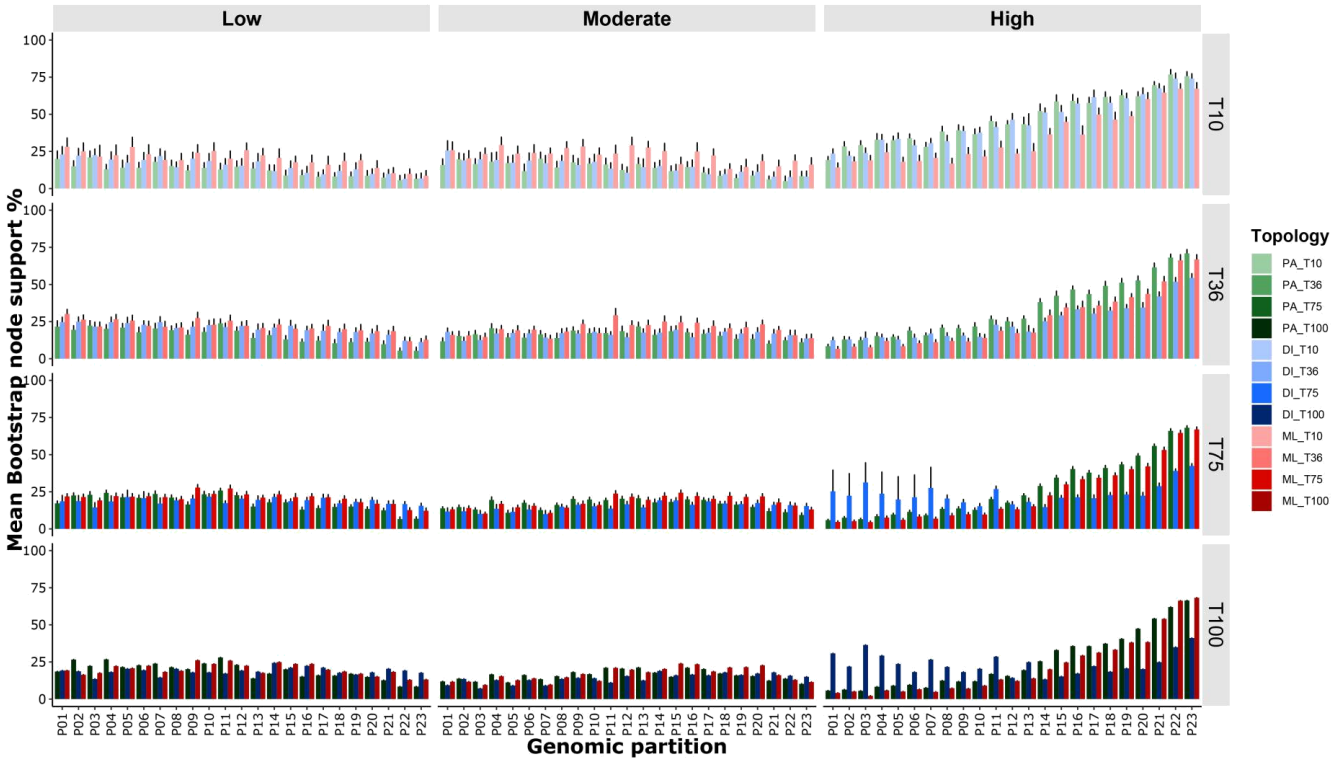

Figure 6. Effect of the variables on the support. The Bootstrap support values are shown in three categories: low (50-75%), moderate (76-94%), and high (95-100%); for each of the genomic partitions (P01 : *E-NS1*; P02: *E*-DIII; P03: *C*; P04: *NS2B*; P05: *NS5*-D; P06: *E*-DI; P07: *C- prM/M*; P08: *NS4A*; P09: *E*-DII; P10: *prM/M*; P11: *NS2A*; P12: *NS4B*; P13: *NS5*-DI; P14: *NS1*; P15: *E*; P16: *NS3*; P17: *NS5*-DII; P18: *prM/M-E*; P19: SG; P20: *NS5*; P21: *E-NS5*; P22: NSG, P23: ORF). Each case is depicted with its inference method (PA: Parsimony; DI: Distance; ML: Maximum Likelihood), and taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tip

Appendix

Appendix A. Summary of the procedures

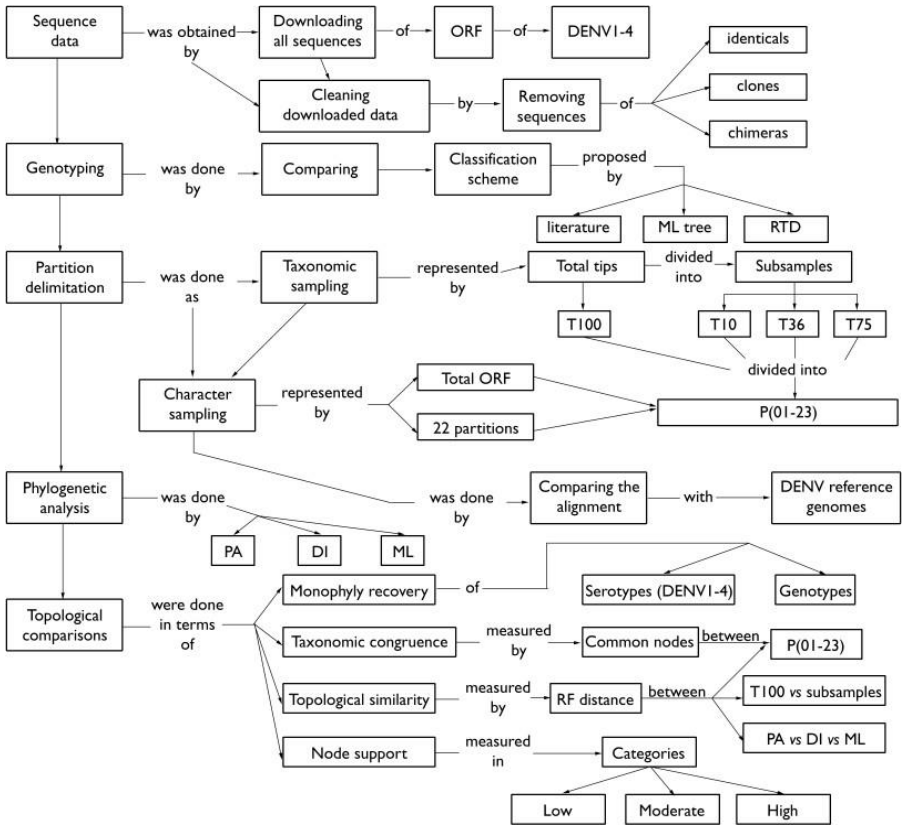

Figure 1. Conceptual map showing the basic steps of the procedures described here.

Appendix B. Recuperation of monophyly by partitions

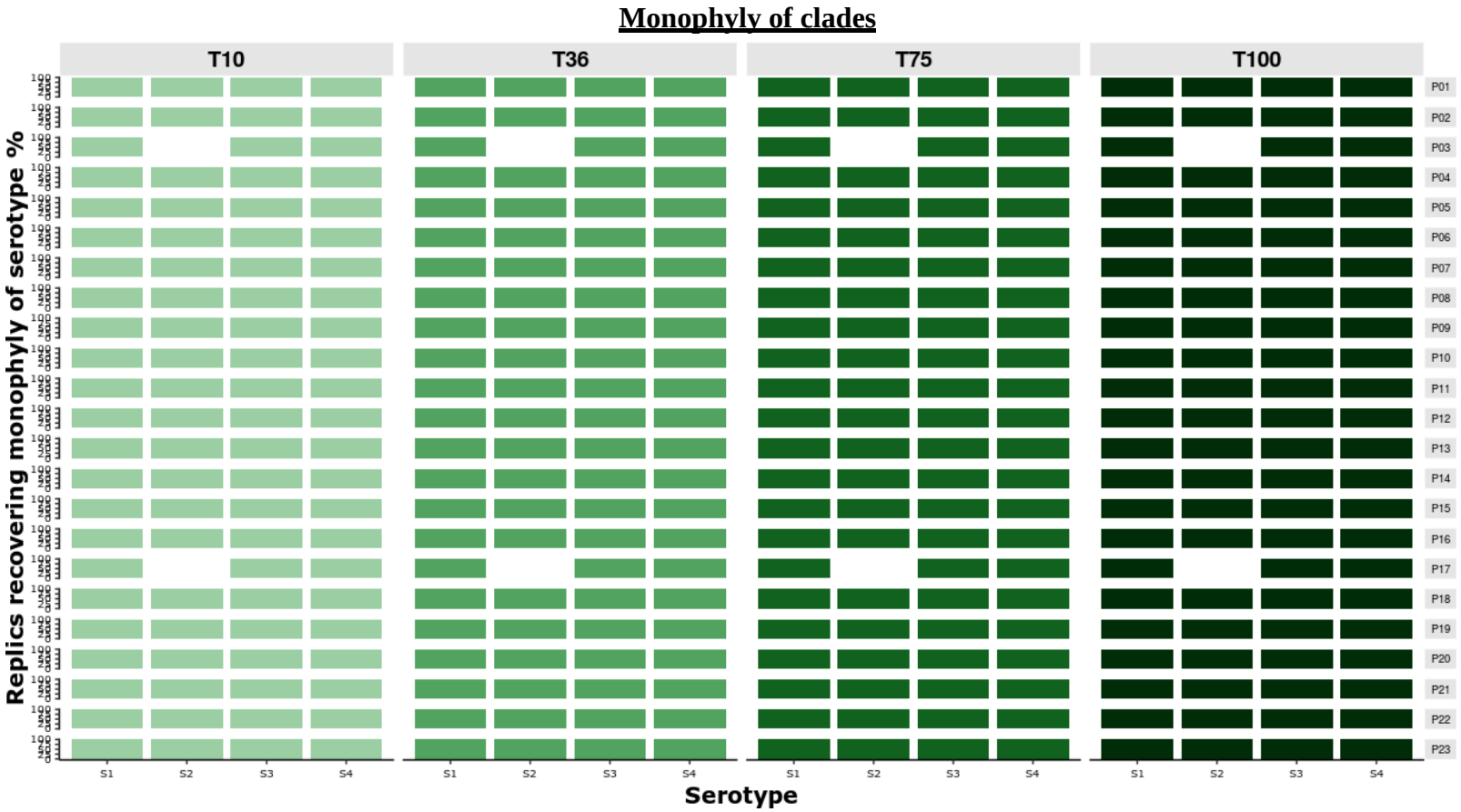

Figure 2. Effect of the taxonomic sampling on the recovery of serotypes as monophyletic groups using the parsimony method. Percentage of topologies (out of 30 replics) recovering the monophyly of the clades (S1: DENV1; S2: DENV2; S3: DENV3; S4: DENV4), for each taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips); and genomic partition (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF).

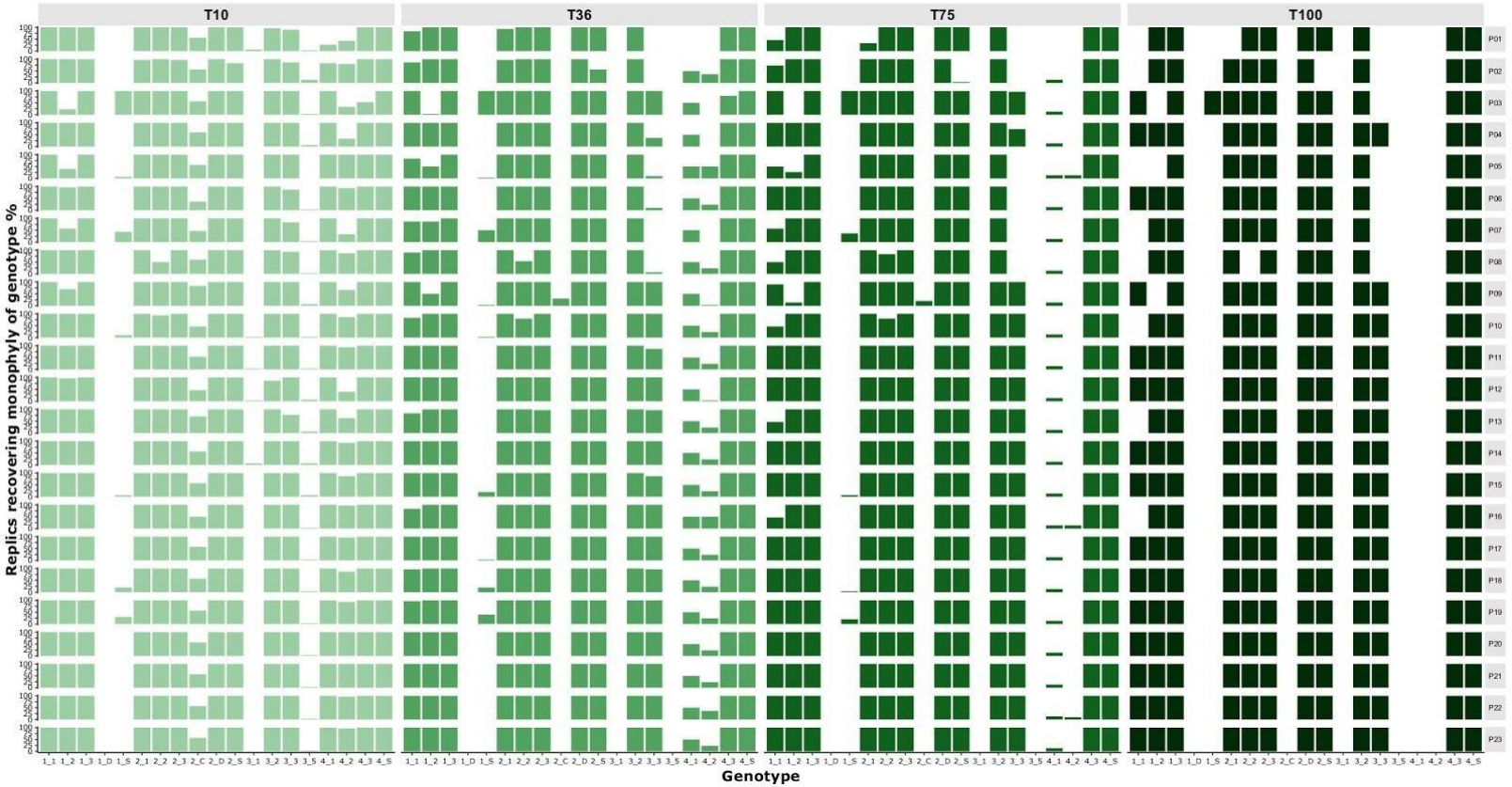

Figure 3. Effect of the taxonomic sampling on the recovery of genotypes as monophyletic groups using the parsimony method. Percentage of topologies (out of 30 replics) recovering the monophyly of the clades (1_1: s1g1I; 1_2: s1g1II; 1_3: s1g1III; 1_S: s1g1SI; 1_D: s1g1DI; 2_S: s2gs2; 2_C: s2gC; 2_1: s2gA1; 2_2: s2gA2; 2_3: s2gAA; 2_D: s2g2DI; 3_1: s3gI; 3_2: s3gII; 3_3: s3gIII; 3_5: s3gV; 4_1: s4gI; 4_2: s4gII; 4_S: s4gs; 4_3: s4gIII), for each taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips); and genomic partition (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF).

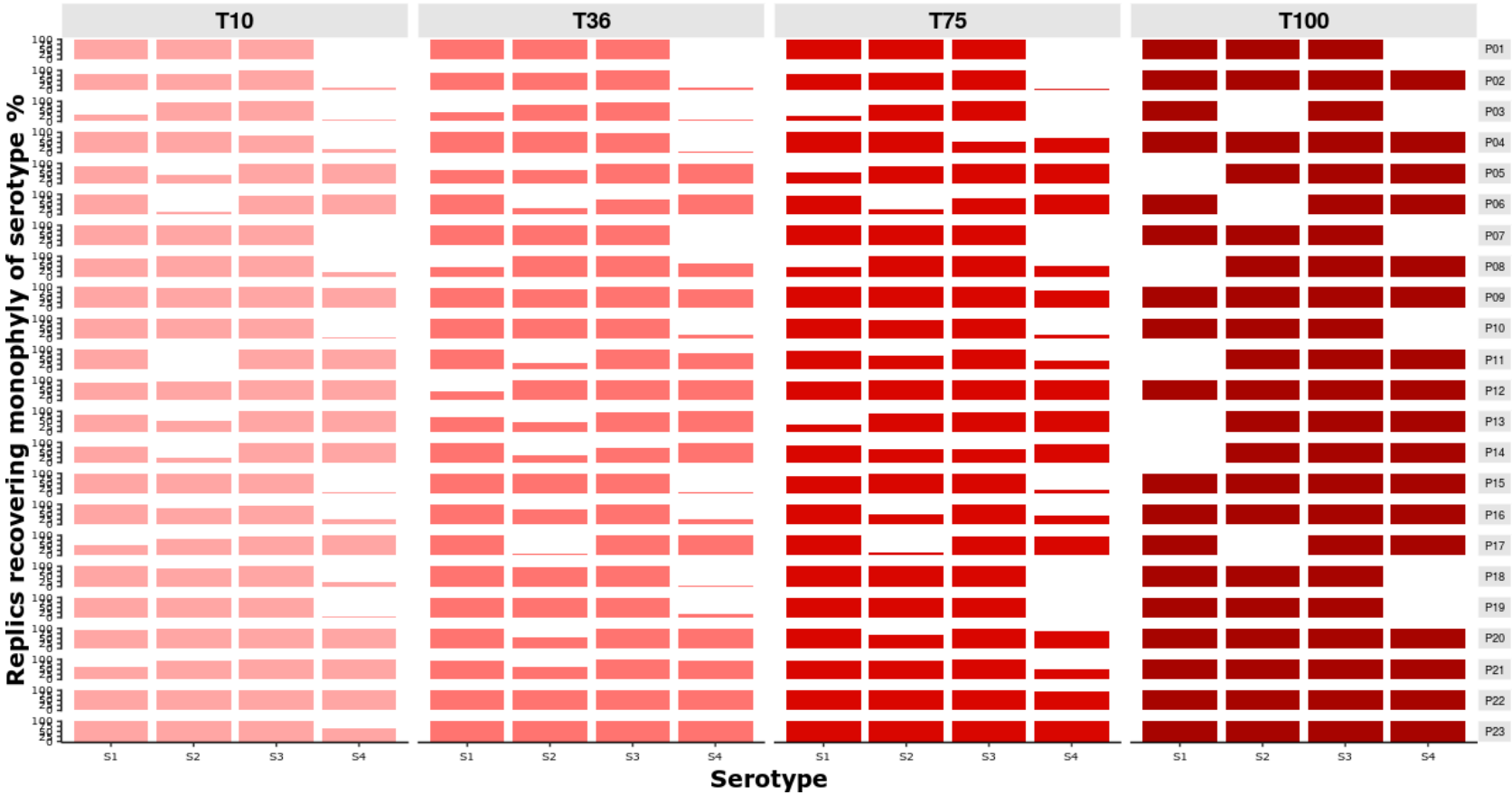

Figure 4. Effect of the taxonomic sampling on the recovery of serotypes as monophyletic groups using Maximum Likelihood. Percentage of topologies (out of 30 replics) recovering the monophyly of the clades (S1: DENV1; S2: DENV2; S3: DENV3; S4: DENV4), for each taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips); and genomic partition (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF).

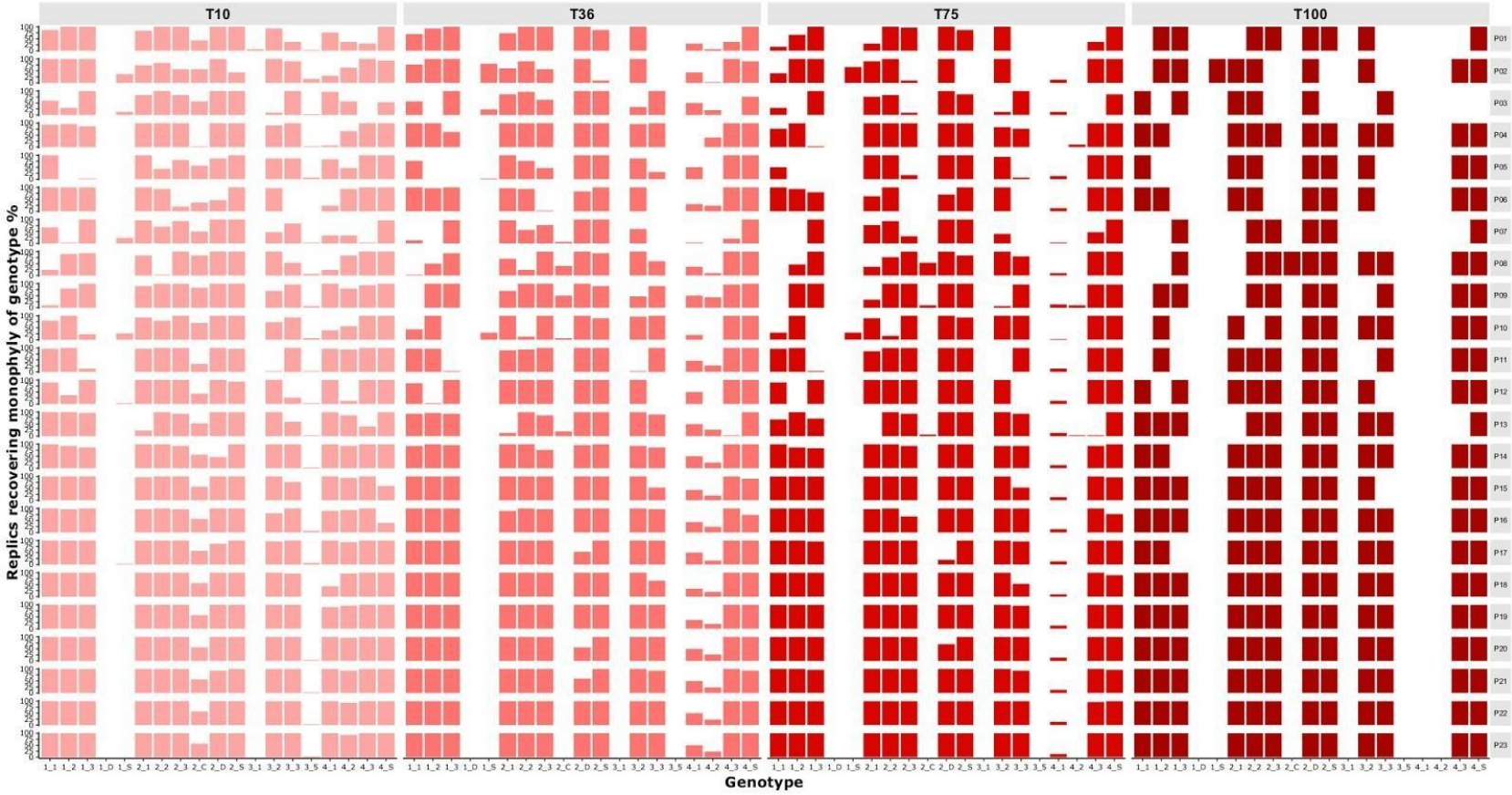

Figure 5. Effect of the taxonomic sampling on the recovery of genotypes as monophyletic groups using Maximum Likelihood. Percentage of topologies (out of 30 replics) recovering the monophyly of the clades (1_1: s1g1I; 1_2: s1g1II; 1_3: s1g1III; 1_S: s1g1SI; 1_D: s1g1DI; 2_S: s2gs2; 2_C: s2gC; 2_1: s2gA1; 2_2: s2gA2; 2_3: s2gAA; 2_D: s2g2DI; 3_1: s3gI; 3_2: s3gII; 3_3: s3gIII; 3_5: s3gV; 4_1: s4gI; 4_2: s4gII; 4_S: s4gs; 4_3: s4gIII), for each taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips); and genomic partition (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF).

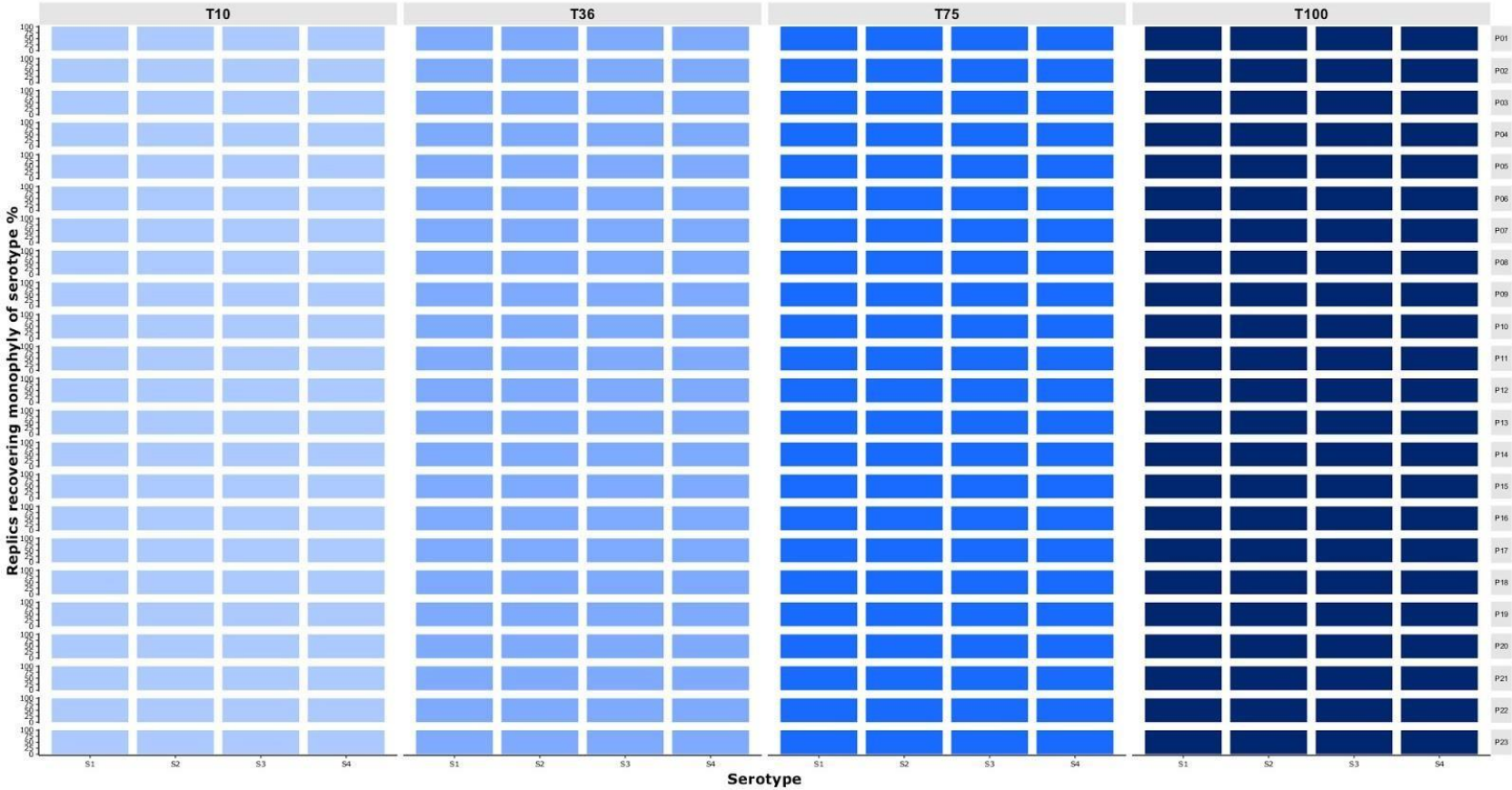

Figure 6. Effect of the taxonomic sampling on the recovery of serotypes as monophyletic groups using distance methods. Percentage of topologies (out of 30 replics) recovering the monophyly of the clades (S1: DENV1; S2: DENV2; S3: DENV3; S4: DENV4), for each taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips); and genomic partition (P01 : E-NS1; P02: E- DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF).

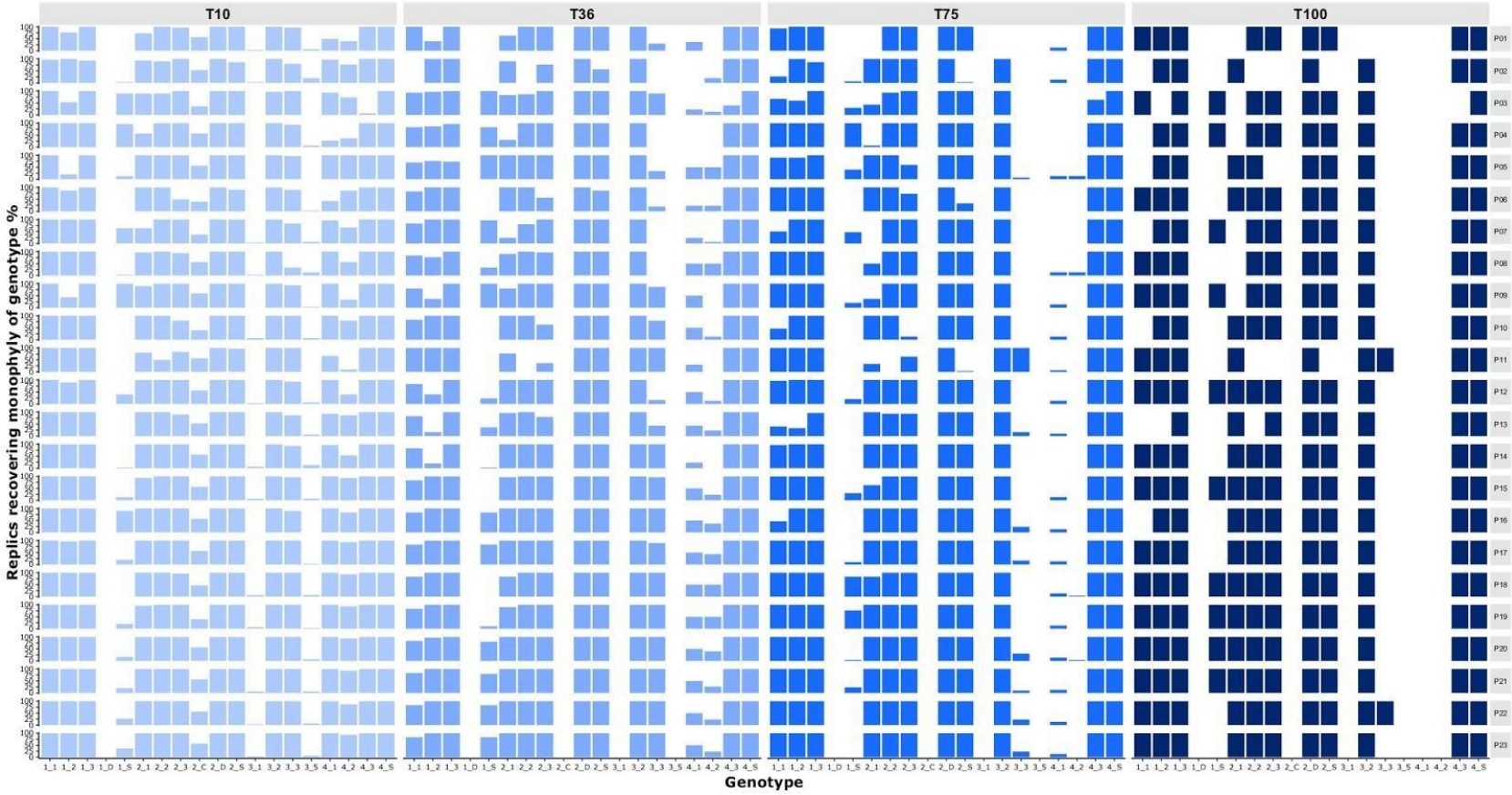

Figure 7. Effect of the taxonomic sampling on the recovery of genotypes as monophyletic groups using distance methods. Percentage of topologies (out of 30 replics) recovering the monophyly of the clades (1_1: s1g1I; 1_2: s1g1II; 1_3: s1g1III; 1_S: s1g1SI; 1_D: s1g1DI; 2_S: s2gs2; 2_C: s2gC; 2_1: s2gA1; 2_2: s2gA2; 2_3: s2gAA; 2_D: s2g2DI; 3_1: s3gI; 3_2: s3gII; 3_3: s3gIII; 3_5: s3gV; 4_1: s4gI; 4_2: s4gII; 4_S: s4gs; 4_3: s4gIII), for each taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips); and genomic partition (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF).

Appendix C. Recuperation of common nodes

Common nodes

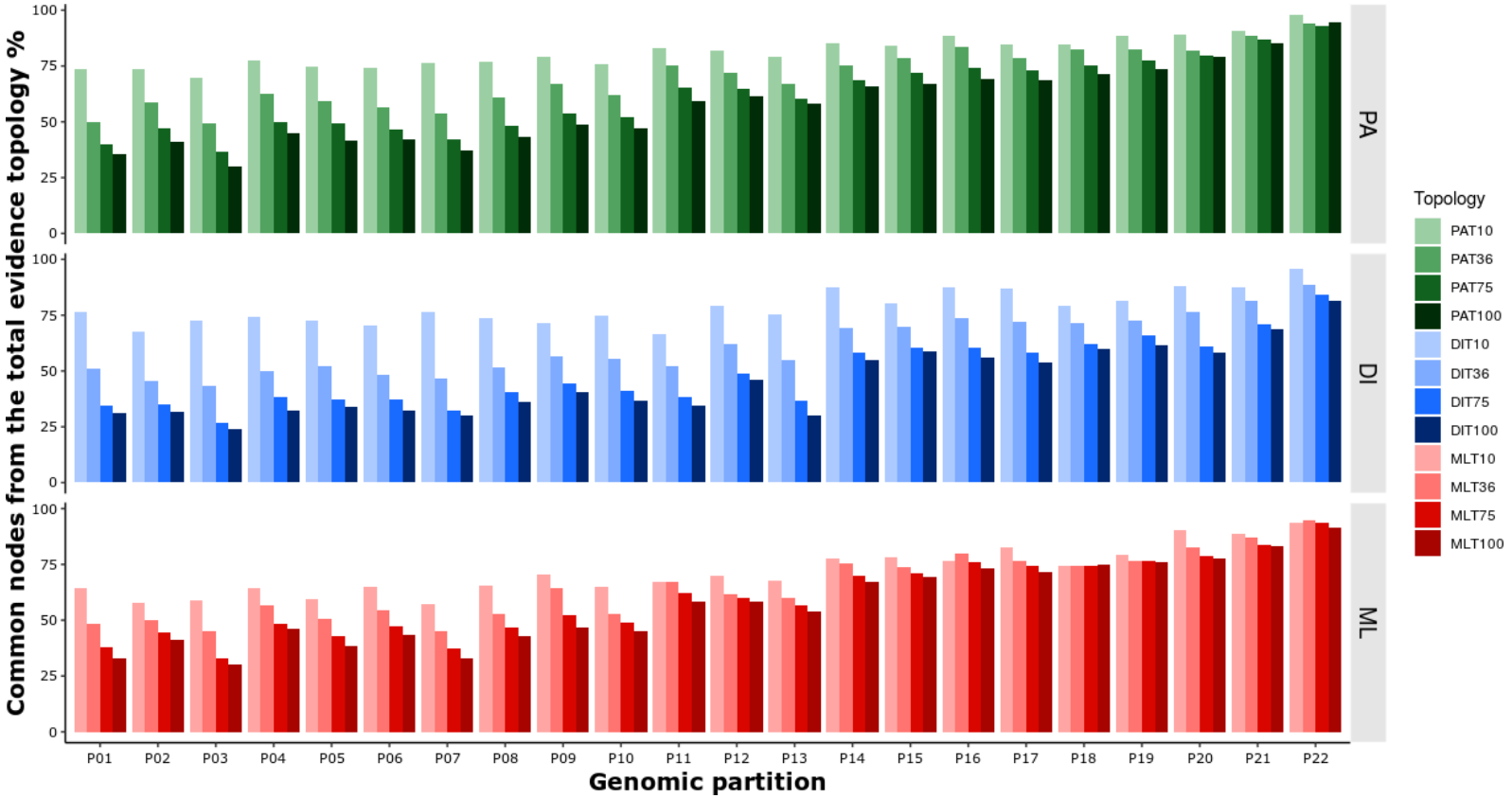

Figure 8. Effect of the genomic partition on node recovery. Percentage of common nodes shared between the topologies inferred from total evidence (P23: ORF) and each genomic partition (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG). Each case is depicted with its inference method (PA: Parsimony; DI: Distance; ML: Maximum Likelihood), and taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips).

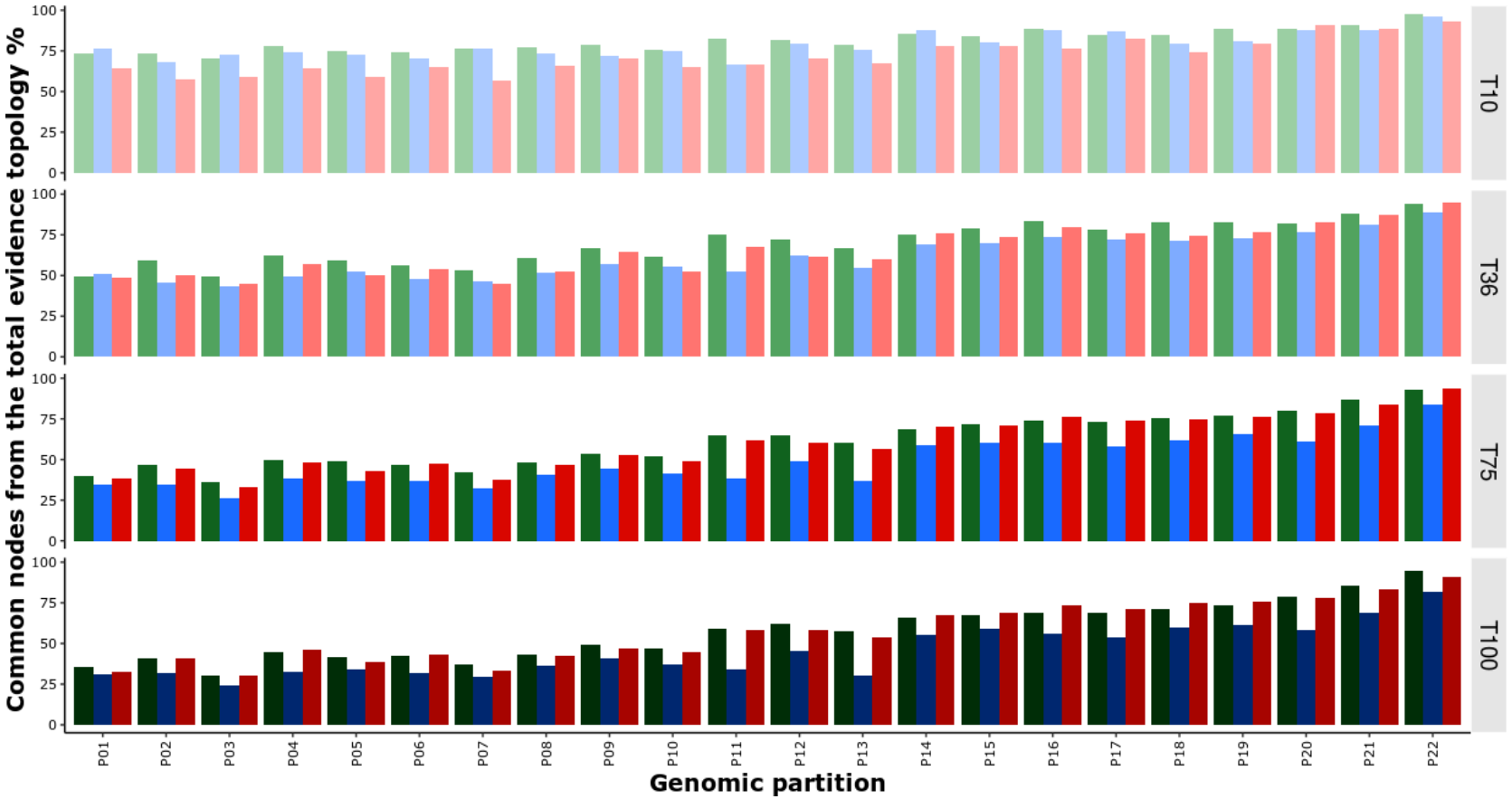

Figure 9. Effect of the taxonomic sampling on node recovery. Percentage of common nodes shared between the topologies inferred from total evidence (P23: ORF) and each genomic partition (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG). Each case is depicted with its inference method (PA: Parsimony; DI: Distance; ML: Maximum Likelihood), and taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips). Methods by colors and their shades: PA: green; DI: blue; ML: red.

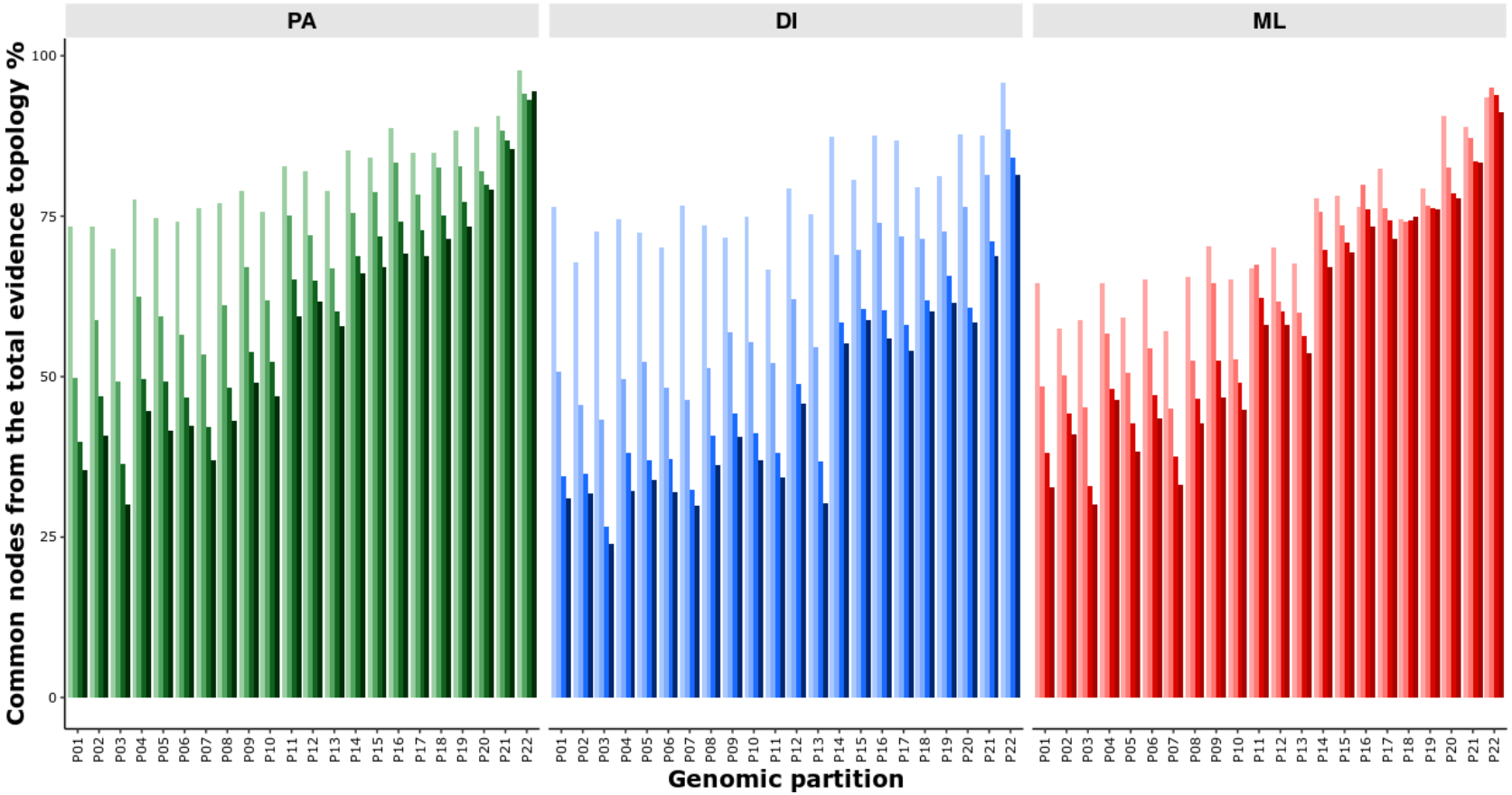

Figure 10. Effect of the method on node recovery. Percentage of common nodes shared between the topologies inferred from total evidence (P23: ORF) and each genomic partition (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG). Each case is depicted with its inference method (PA: Parsimony; DI: Distance; ML: Maximum Likelihood), and taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips).

Appendix D. Topological similarity

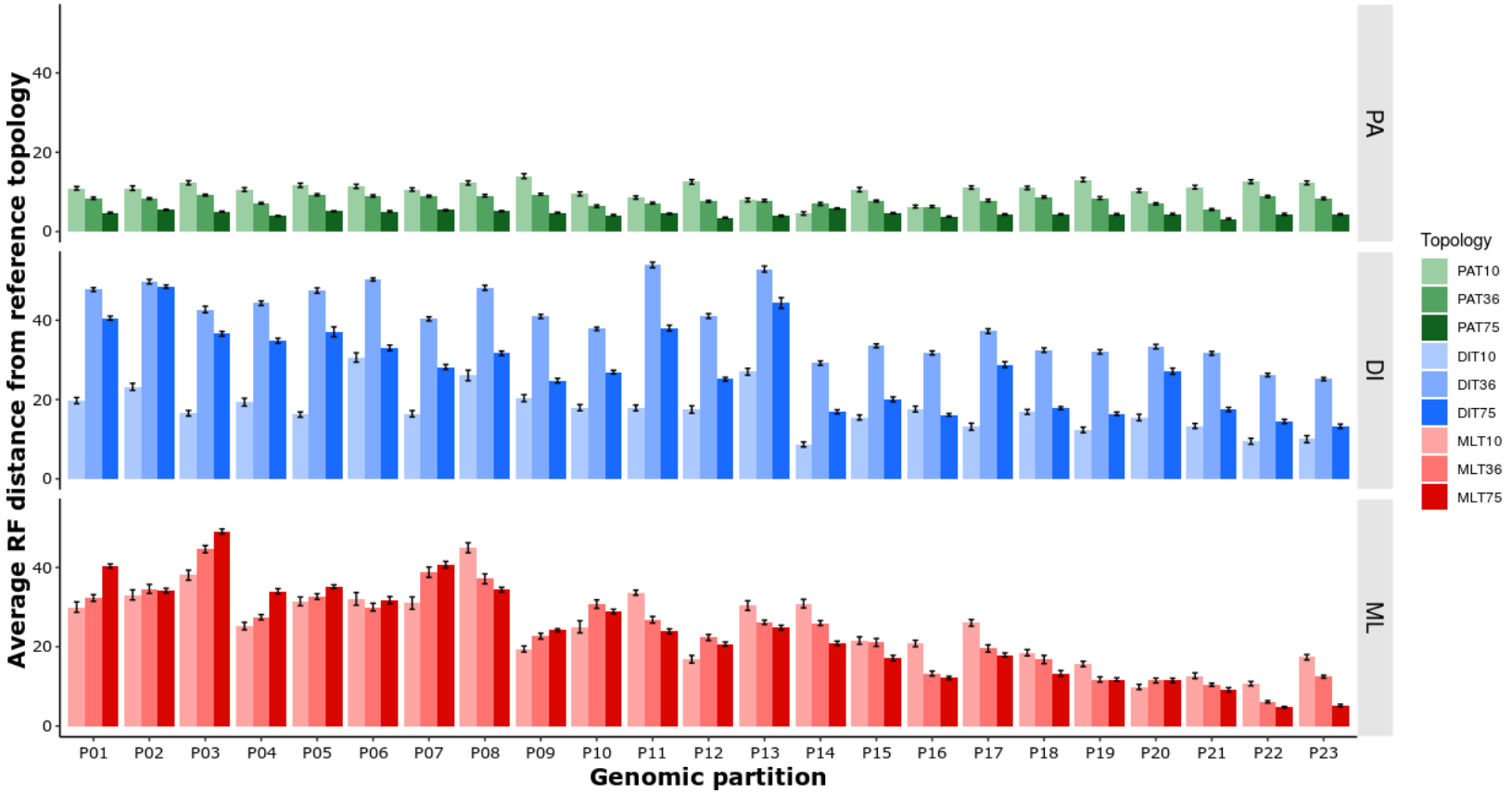

Figure 10a. Effect of the sampling size. Topological distance (RF) between the reference topology (the tree inferred using all the terminals) and the pruned topologies inferred from each taxonomic subsample and taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips). The graphic shows the distances for each genomic partition (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF). Each case is depicted with its inference method (PA: Parsimony; DI: Distance; ML: Maximum Likelihood). Methods by colors and their shades: PA: green; DI: blue; ML: red.

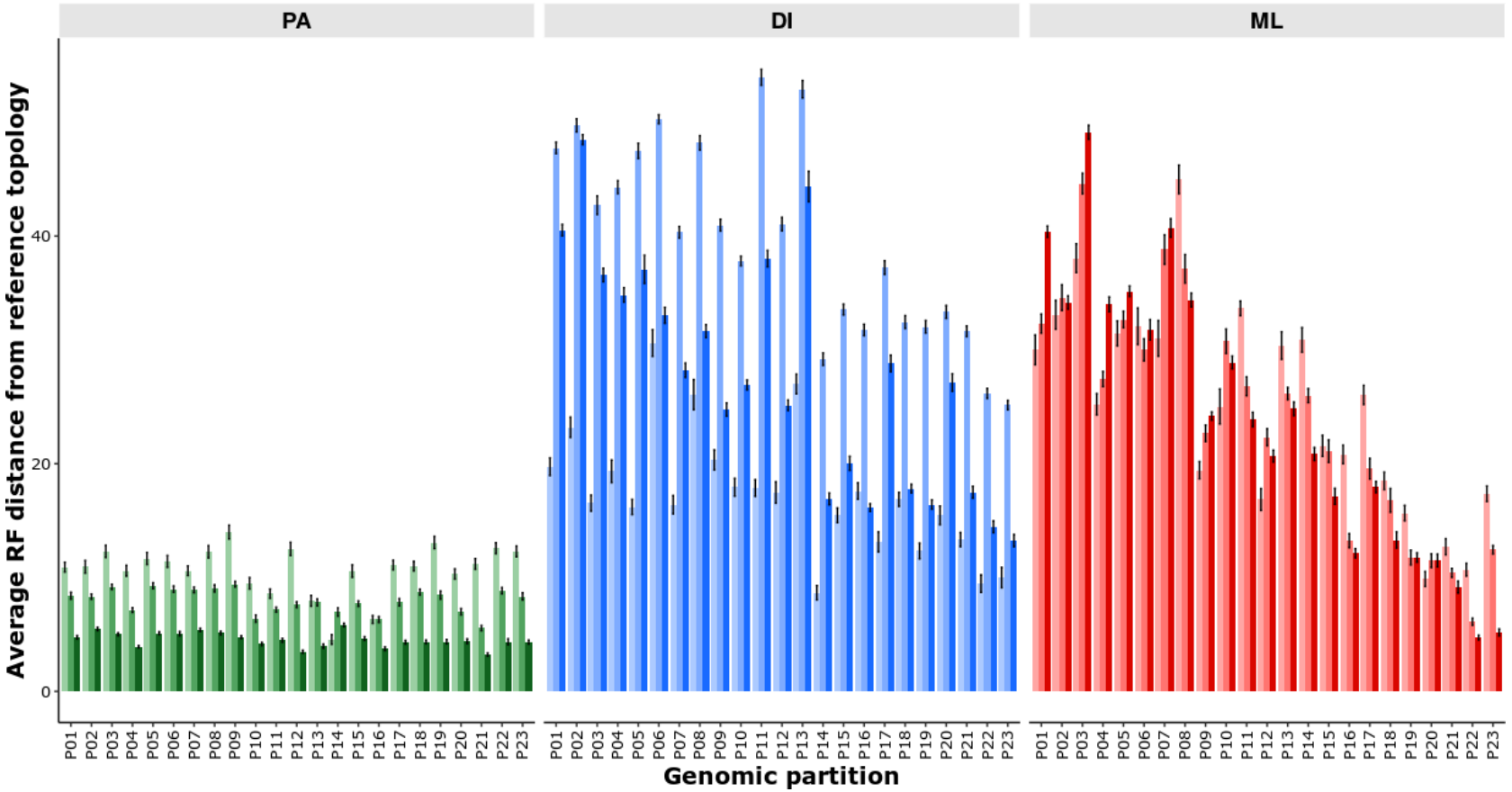

Figure 10b. Effect of the sampling size and method. Topological distance (RF) between the reference topology (the tree inferred using all the terminals) and the pruned topologies inferred from each taxonomic subsample and taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips). The graphic shows the distances for each genomic partition (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF). Each case is depicted with its inference method (PA: Parsimony; DI: Distance; ML: Maximum Likelihood). Methods by colors and their shades: PA: green; DI: blue; ML: red.

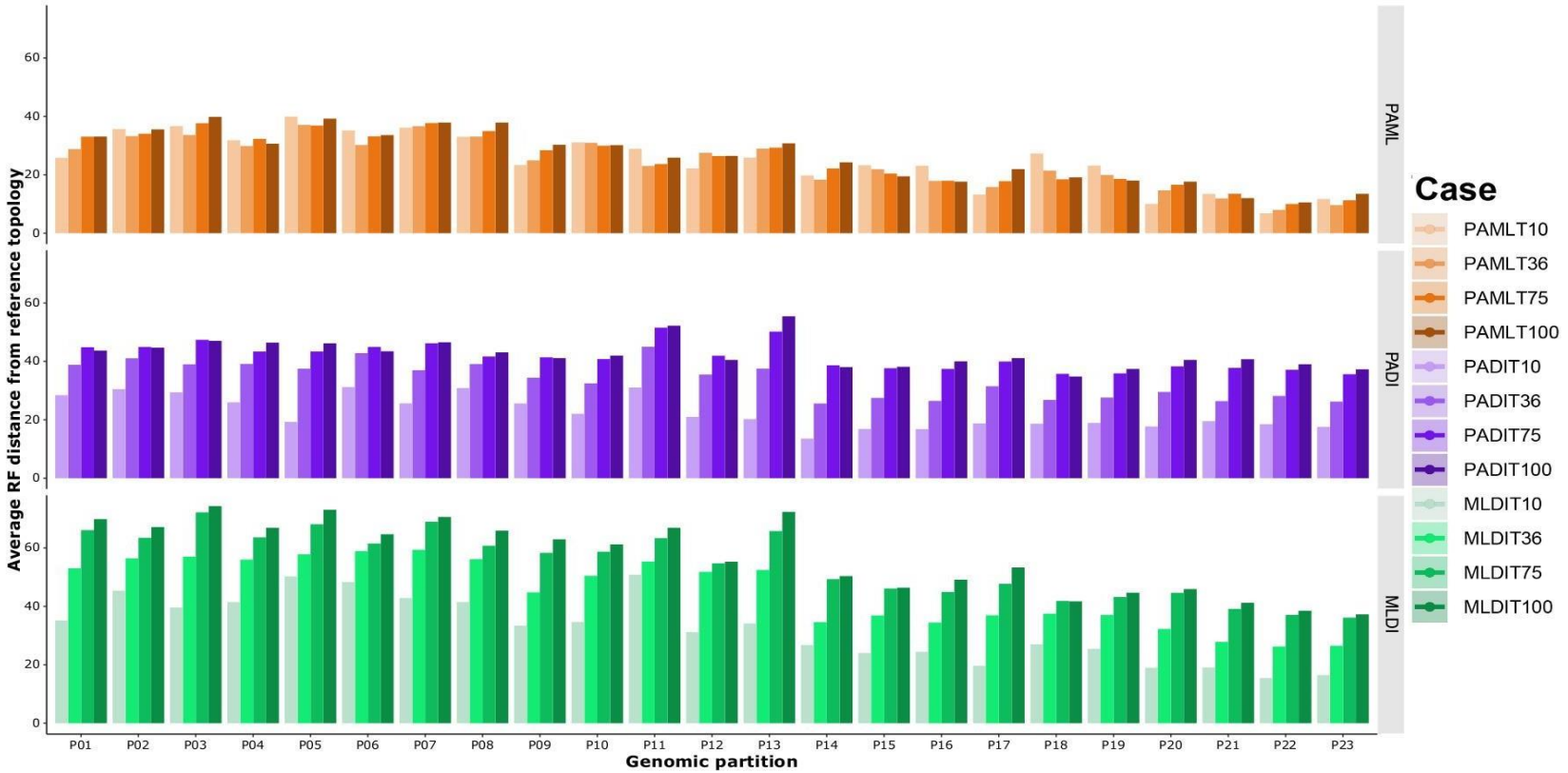

Figure 11a. Topological similarity between methods. The graphic displays the distances for each genomic partition (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF). Each case is depicted with its inference methods (PAML: Parsimony and Maximum Likelihood; PADI: Parsimony and Distance; MLDI: Maximum Likelihood and Distance). It shows that the trees yielded by Parsimony and Maximum Likelihood are more similar between them than with distance methods.

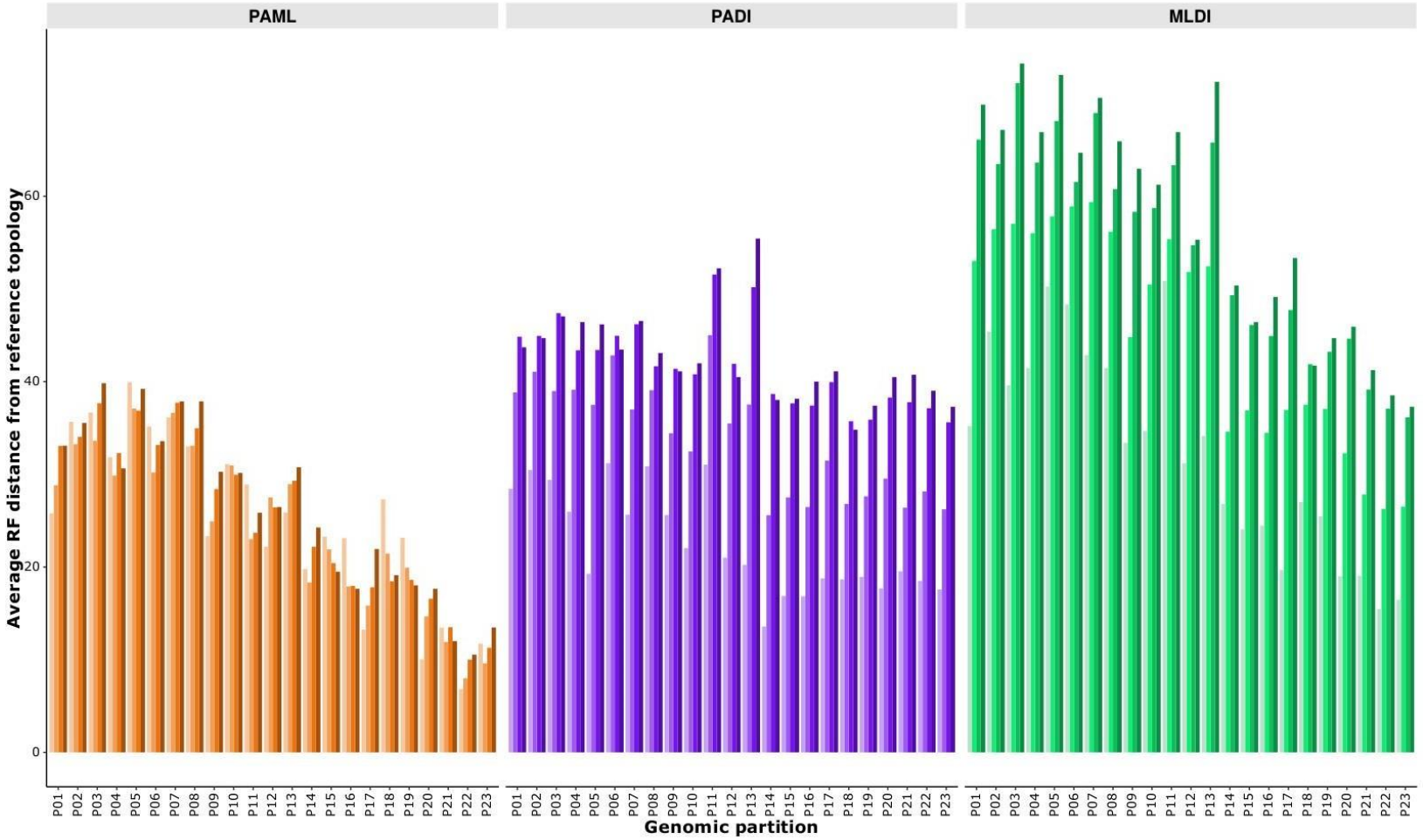

Figure 11b. Topological similarity between methods. The graphic displays the distances for each genomic partition (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF). Each case is depicted with its inference methods (PAML: Parsimony and Maximum Likelihood; PADI: Parsimony and Distance; MLDI: Maximum Likelihood and Distance). It shows that the trees yielded by Parsimony and Maximum Likelihood are more similar between them than with distance methods.

Appendix E. Nodal Support

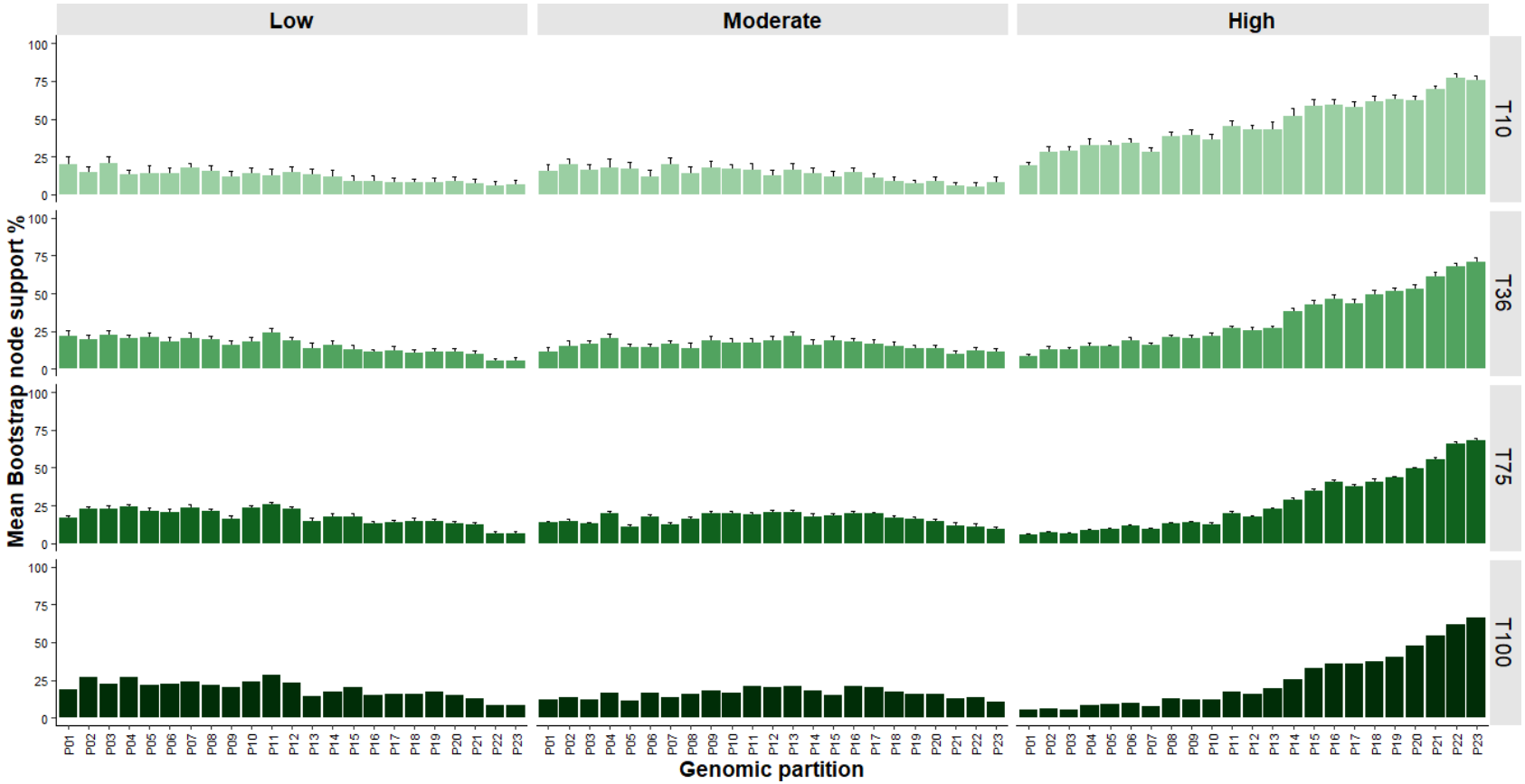

Figure 12. Nodal support in trees reconstructed by Parsimony. The Bootstrap support values are shown in three categories: low (50-75\%), moderate (76-94\%), and high (95-100\%); for each of the genomic partitions (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF). Each case is depicted with its taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips).

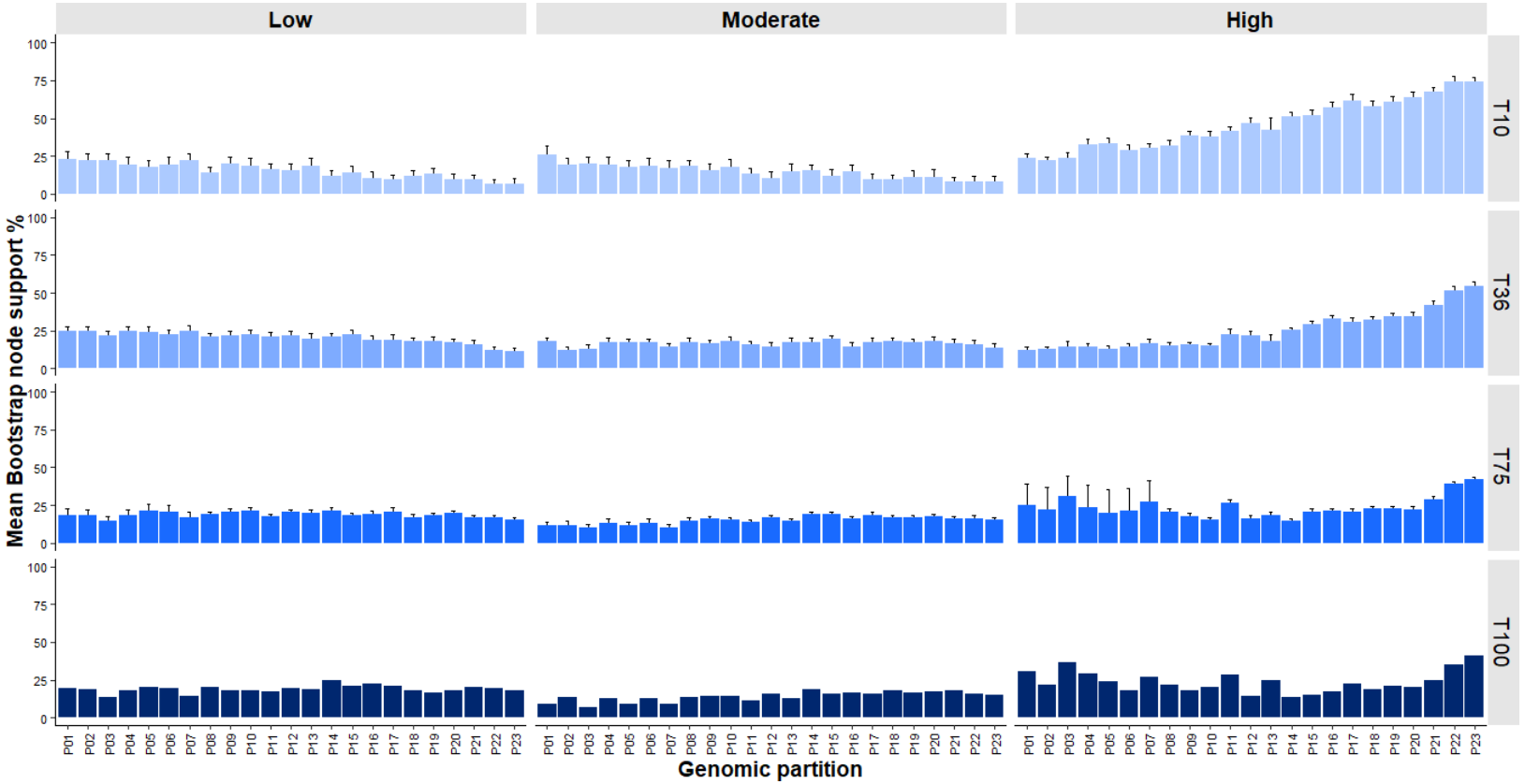

Figure 13. Nodal support in trees reconstructed by distance. The Bootstrap support values are shown in three categories: low (50- 75\%), moderate (76-94\%), and high (95-100\%); for each of the genomic partitions (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF). Each case is depicted with its taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips).

Figure 14. Nodal support in trees reconstructed by Maximum Likelihood. The Bootstrap support values are shown in three categories: low (50-75\%), moderate (76-94\%), and high (95-100\%); for each of the genomic partitions (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF). Each case is depicted with its taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips).

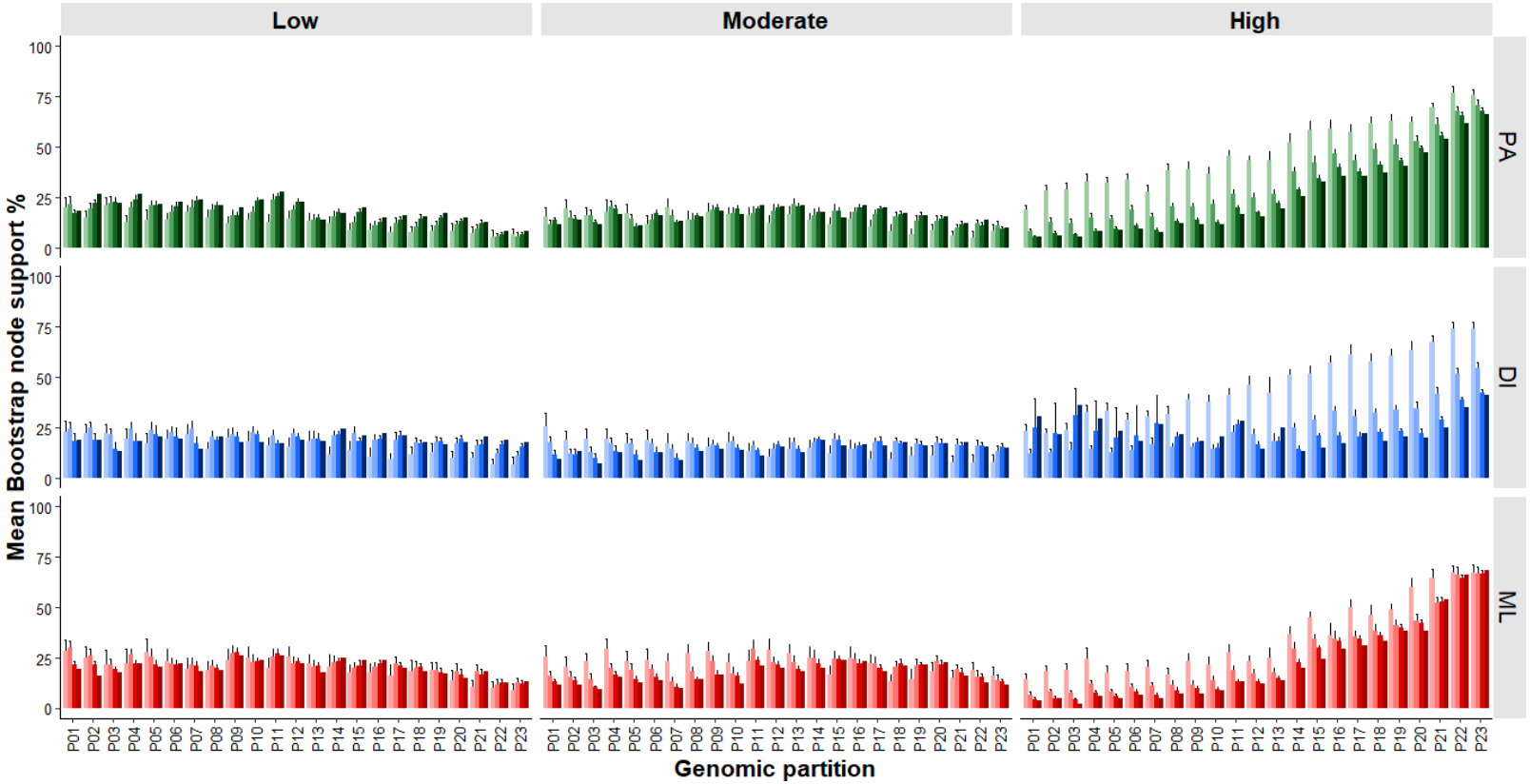

Figure 15. Nodal support in trees reconstructed by Maximum Likelihood. The Bootstrap support values are shown in three categories: low (50-75\%), moderate (76-94\%), and high (95-100\%); for each of the genomic partitions (P01 : E-NS1; P02: E-DIII; P03: C; P04: NS2B; P05: NS5-D; P06: E-DI; P07: C-prM/M; P08: NS4A; P09: E-DII; P10: prM/M; P11: NS2A; P12: NS4B; P13: NS5-DI; P14: NS1; P15: E; P16: NS3; P17: NS5-DII; P18: prM/M-E; P19: SG; P20: NS5; P21: E-NS5; P22: NSG, P23: ORF). Each case is depicted with its inference method (PA: Parsimony; DI: Distance; ML: Maximum Likelihood), and taxonomic sampling (T10: 53 tips; T36: 146 tips; T75: 306 tips; T100: 410 tips).

Appendix F. General comparisons

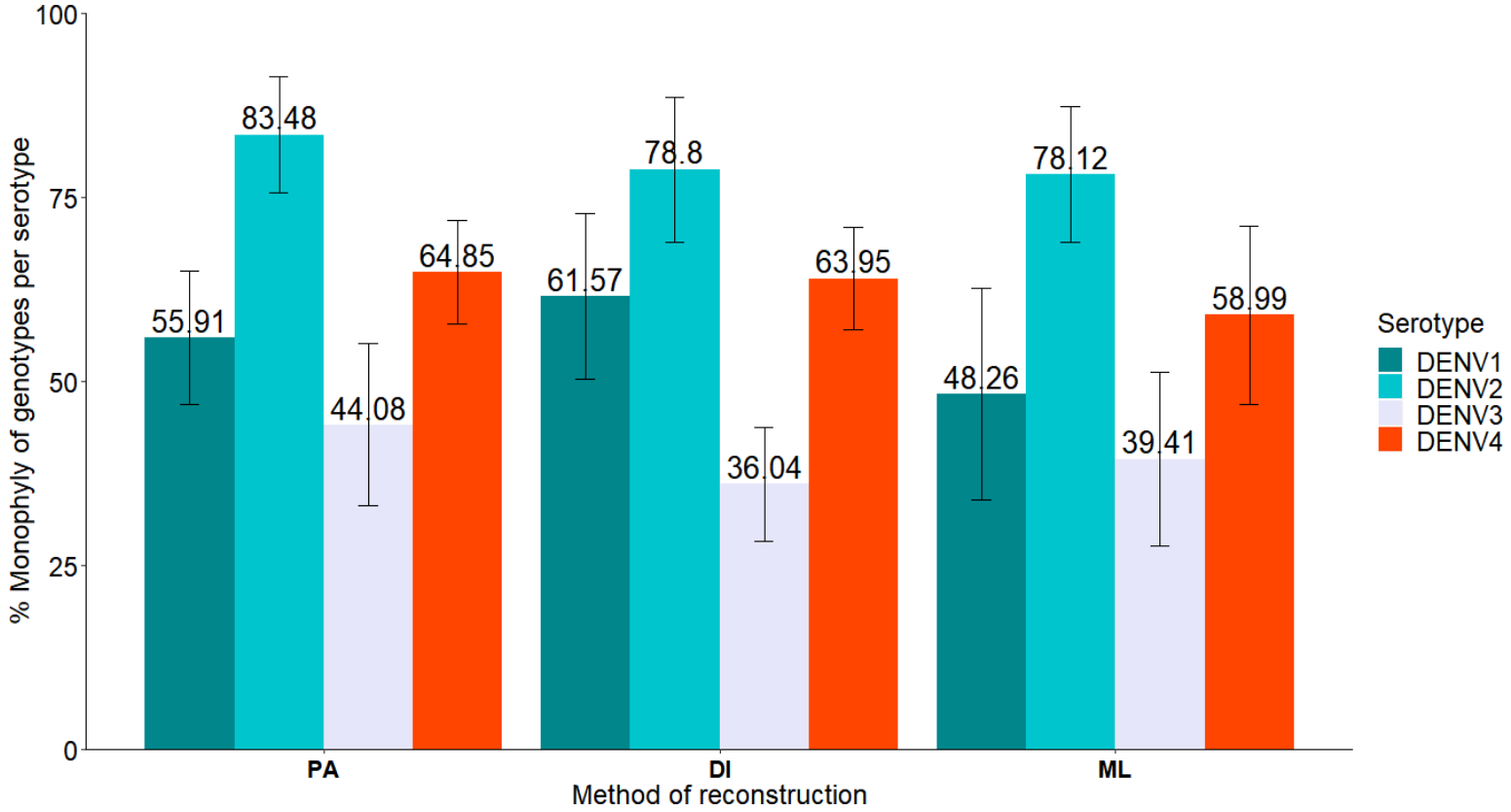

Figure 16. Overall recuperation of monophyly of all the clades per serotype, for all the reconstruction methods, and the sampling sizes altogether.

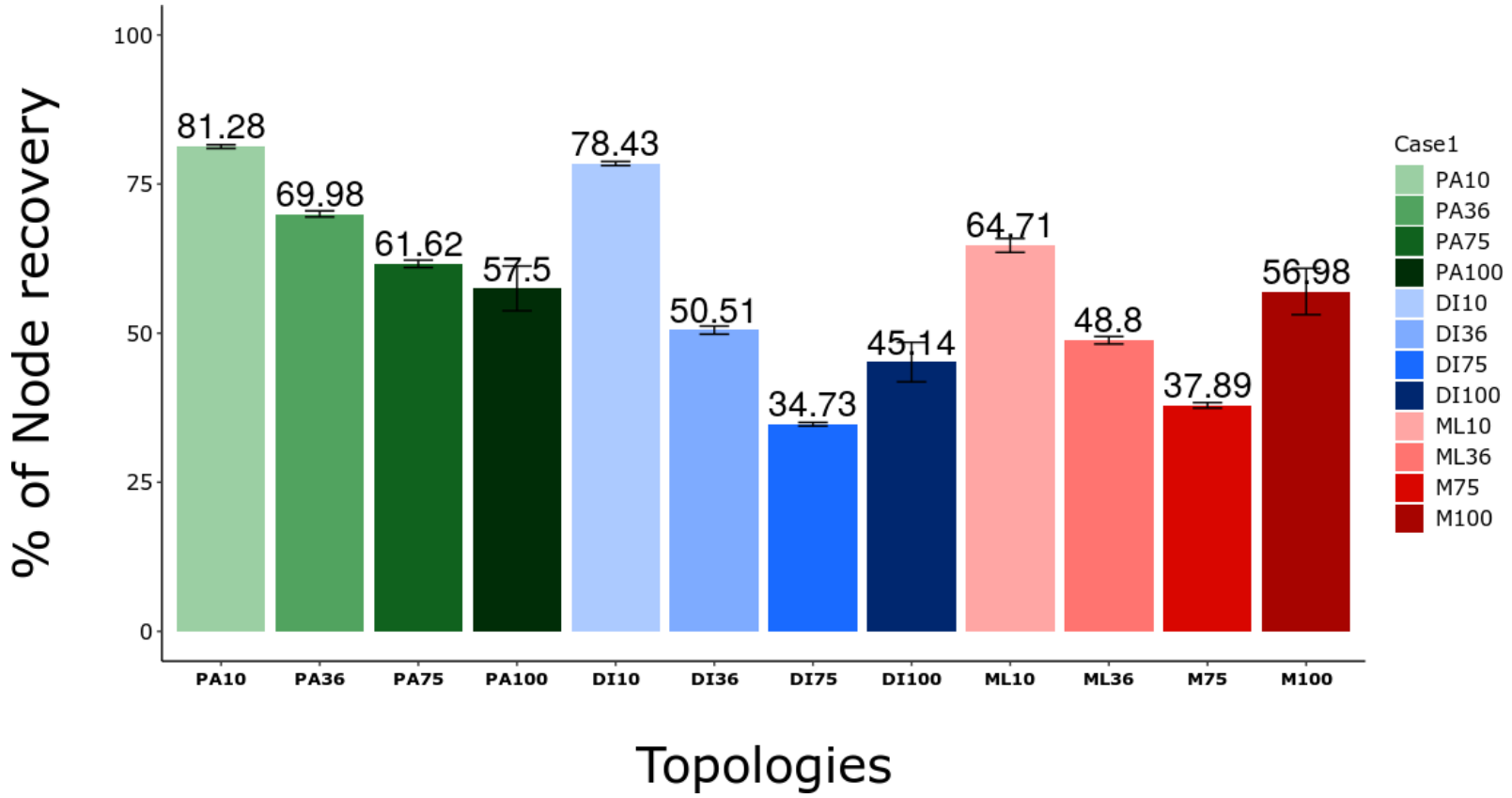

Figure 17. Average node recovery with the reference topology (total evidence).

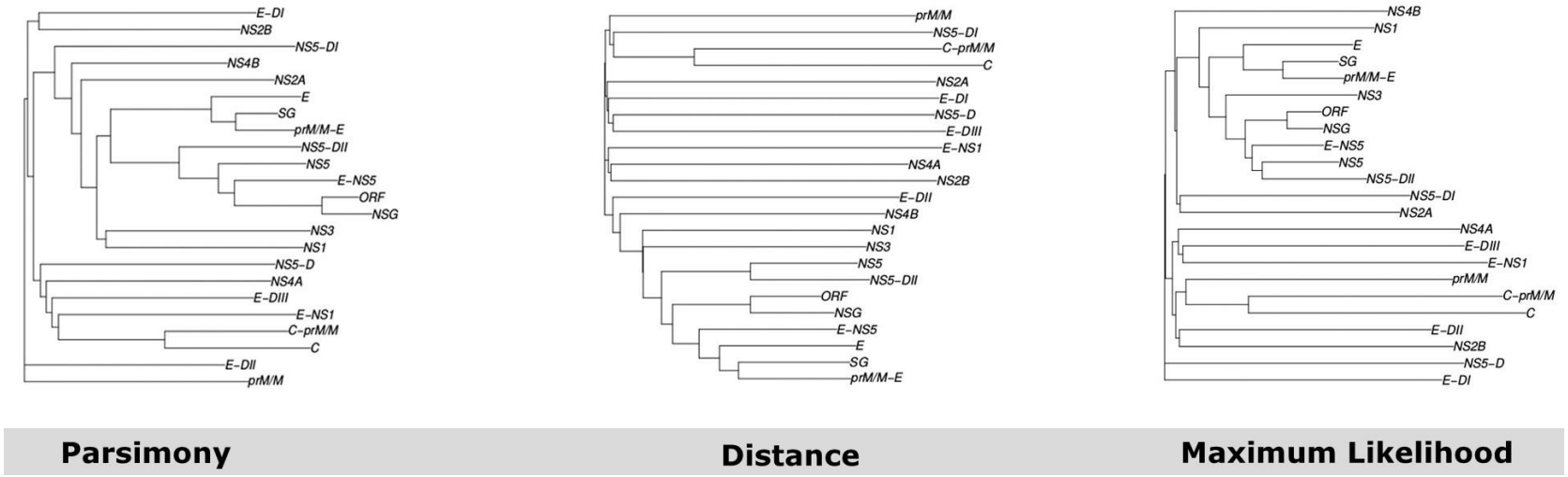

Figure 18. Topological similarity between genomic partitions.