

# Discovering recent selection forces shaping the evolution of dengue viruses based on polymorphism data across geographic scales

Nien-Kung Li,<sup>1</sup> Jukka Corander,<sup>2,3,4</sup> Yonatan H. Grad,<sup>5,6,†</sup> and Hsiao-Han Chang<sup>†,\*</sup>

<sup>1</sup>Department of Life Science & Institute of Bioinformatics and Structural Biology, National Tsing Hua University, 101, Section 2, Kuang-Fu Road, Hsinchu 300044, Taiwan, <sup>2</sup>Helsinki Institute for Information Technology, Department of Mathematics and Statistics, University of Helsinki, Yliopistonkatu 3, Helsinki 00014, Finland, <sup>3</sup>Department of Biostatistics, University of Oslo, Domus Medica Gaustad Sognsvannsveien 9, Oslo 0372, Norway, <sup>4</sup>Parasites and Microbes, The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK, <sup>5</sup>Department of Immunology and Infectious Diseases and Center for Communicable Disease Dynamics, Harvard T. H. Chan School of Public Health, 677 Huntington Ave, Boston, Massachusetts 02115, USA and <sup>6</sup>Division of Infectious Diseases, Brigham and Women's Hospital, Harvard Medical School, 75 Francis St, Boston, Massachusetts 02115, USA

<sup>†</sup><https://orcid.org/0000-0001-5646-1314>

<sup>†</sup><https://orcid.org/0000-0001-8016-1530>

\*Corresponding author: E-mail: [hhchang@life.nthu.edu.tw](mailto:hhchang@life.nthu.edu.tw)

## Abstract

Incomplete selection makes it challenging to infer selection on genes at short time scales, especially for microorganisms, due to stronger linkage between loci. However, in many cases, the selective force changes with environment, time, or other factors, and it is of great interest to understand selective forces at this level to answer relevant biological questions. We developed a new method that uses the change in  $d_N/d_S$ , instead of the absolute value of  $d_N/d_S$ , to infer the dominating selective force based on sequence data across geographical scales. If a gene was under positive selection,  $d_N/d_S$  was expected to increase through time, whereas if a gene was under negative selection,  $d_N/d_S$  was expected to decrease through time. Assuming that the migration rate decreased and the divergence time between samples increased from between-continent, within-continent different-country, to within-country level,  $d_N/d_S$  of a gene dominated by positive selection was expected to increase with increasing geographical scales, and the opposite trend was expected in the case of negative selection. Motivated by the McDonald–Kreitman (MK) test, we developed a pairwise MK test to assess the statistical significance of detected trends in  $d_N/d_S$ . Application of the method to a global sample of dengue virus genomes identified multiple significant signatures of selection in both the structural and non-structural proteins. Because this method does not require allele frequency estimates and uses synonymous mutations for comparison, it is less prone to sampling error, providing a way to infer selection forces within species using publicly available genomic data from locations over broad geographical scales.

**Key words:**  $d_N/d_S$ ; pMK test; natural selection; dengue virus; geographic scales.

## Introduction

With the decrease in sequencing costs, genome sequences of representatives of a species across geographic and temporal scales are increasingly available (Gratton et al. 2017), providing opportunities to understand the species' recent demographic history and evolution. The ratio of non-synonymous to synonymous substitution rates ( $d_N/d_S$ ), which compares the relative abundance of amino acid altering and preserving mutations, is commonly used to infer selection between species. Because  $d_N/d_S$  is applicable even when recombination is absent (Hedge, Wilson, and Ouellette 2016), it is useful for haploid organisms, such as bacteria and viruses, in which tests based on linkage disequilibrium cannot be applied (Shapiro et al. 2009). Traditionally, a  $d_N/d_S$  larger than one is considered a signal of positive selection, a  $d_N/d_S$  smaller than one is considered a signal of negative selection, and  $d_N/d_S$  is expected to be one under a neutral condition. However, at the

within-species level, since the divergence time among samples is shorter, selection may not have had sufficient time to act within the sampling timeframe, leading to incomplete selection and possible misinterpretation (Rocha et al. 2006; Kryazhimskiy, Plotkin, and Gojobori 2008; Mugal, Wolf, and Kaj 2014). For example, incomplete purifying selection may lead to elevated  $d_N/d_S$  (Hasegawa, Cao, and Yang 1998; Pybus et al. 2007; Peterson and Masel 2009; Park et al. 2015).

Since within-species  $d_N/d_S$  should be interpreted with caution, several studies formally incorporated polymorphisms in their models when estimating  $d_N/d_S$  (Wilson et al. 2011; Mugal et al. 2020; Wilson and Consortium 2020), with the assumption that selection did not change over space and time. Bhatt et al. considered that high-frequency non-synonymous polymorphic sites are possibly adaptive and utilized segregating mutations with different frequencies to better infer the rate of adaptation

(Bhatt, Katzourakis, and Pybus 2010; Bhatt, Holmes, and Pybus 2011). However, since this method requires estimates of allele frequencies, it is sensitive to sampling bias and, therefore, less suitable for analyzing samples in public databases, which often represent datasets collected and sequenced by multiple research groups. Moreover, these models also assume a constant force of selection, an assumption that may be violated in the case of local adaptation.

Since  $d_N/d_S$  of a positively selected gene is expected to increase through time and  $d_N/d_S$  of a negatively selected gene is expected to decrease through time (Mugal, Wolf, and Kaj 2014), examining the trend of change in  $d_N/d_S$  can provide information about recent selective forces. The straightforward way to observe the change in  $d_N/d_S$  is to analyze temporal samples. Alternatively, samples across geographic scales may represent different levels of divergence time. It has been shown in several organisms that the genetic differentiation, which decreases with gene flow or migration rate between populations, was higher between continents than within each continent (Romualdi et al. 2002; Bedford et al. 2010; Yukilevich et al. 2010; Miotto et al. 2013; Azarian et al. 2018). Suppose we assume that gene flow decreases with geographic distances (e.g. within-continent gene flow is higher than between-continent gene flow, and within-country gene flow is higher than between-country gene flow) and since divergent time is expected to decrease with gene flow, divergent time between samples increases with geographic distances. In other words, analyzing samples from locations across geographic scales provides an opportunity to obtain  $d_N/d_S$  at different levels of divergence time.

We applied this idea to study recent selective forces acting on four serotypes of dengue viruses, which have been shown to emerge from four sylvatic ancestors independently (Holmes and Twiddy 2003) and differ in their virulence (Fried et al. 2010) and transmissibility (Duong et al. 2015). Sequence data of dengue viruses across geographic scales are publicly available, offering a great opportunity to study the evolution of this important pathogen. The statistical significance was tested using a simple contingency test adapted from the original McDonald–Kreitman (MK) test (McDonald and Kreitman 1991). Across dengue genes, we found that the dominant selective force varied among serotypes and continents, providing evolutionary insights into the phenotypic difference among serotypes.

## Materials and methods

### Genomic data of dengue viruses

We downloaded all the available sequences of four dengue virus serotypes (DENV1–4) from the National Center for Biotechnology Information database on 11 February 2020 (NCBI 1988). After excluding sequences from non-human sources, the sample sizes for serotypes 1, 2, 3, and 4 are 1,150, 865, 707, and 198 from 31, 33, 38, and 22 countries, respectively. Bayesian clustering was performed using the R package *rhierBAPS* (Tonkin-Hill et al. 2018) (*max.depth* = 3).

### $d_N/d_S$ estimation and pairwise McDonald–Kreitman test

The ratio of non-synonymous to synonymous changes ( $d_N/d_S$ ) was calculated between every pair of sequences using the maximum likelihood method implemented in CodeML from PAML (Yang 1997, 2007) (*runmode* = -2, *CodonFreq* = 2), and the average  $d_N/d_S$  was calculated by taking the ratio of the average  $d_N$  to the average  $d_S$ . The number of non-synonymous changes per

non-synonymous site,  $d_N$ , is equal to  $C_N/N$ , where  $C_N$  is the number of non-synonymous changes and  $N$  is the total number of non-synonymous sites. Similarly, the number of synonymous changes per synonymous site,  $d_S$ , is equal to  $C_S/S$ , where  $C_S$  is the number of synonymous changes and  $S$  is the total number of synonymous sites.

To examine the statistical significance of the pattern of  $d_N/d_S$  across geographic scales, we modified the original MK test to what we term a pairwise McDonald–Kreitman (pMK) test. The numbers of synonymous and non-synonymous changes,  $C_S$  and  $C_N$ , were calculated from the numbers of non-synonymous and synonymous sites ( $N$  and  $S$ ), and  $d_N$  and  $d_S$  were estimated from PAML by  $N \times d_N$  and  $S \times d_S$ , respectively.

We compared the non-synonymous and synonymous differences between within-continent between-country and between-continent levels. We first calculated the average pairwise non-synonymous and synonymous differences between all the country pairs and then summed and rounded the differences between continents or within each continent. The two dimensions of the contingency test were (1) within or between continents and (2) non-synonymous or synonymous (Table 1). For within-continent changes, we used either the America or Asia data. Additionally, to test the robustness of the results, we performed the same tests with only values within the interquartile range (pMK\* test). We estimated  $q$ -values using the *qvalue* package (Storey et al. 2020) in R and used  $q < 0.1$  as a standard for statistical significance. To quantify the magnitude of the signal, we also calculated the odds ratio as follows:

$$\frac{A/B}{C/D}.$$

The difference in divergence times,  $S$  ratio, was quantified by the ratio of synonymous changes between and within continents,  $B/D$ .

## Simulations

We used SLiM to simulate positive and negative selection with various values of selection coefficients ( $s = 0.0025, 0.005, 0.01$ , and  $0.02$ ) and migration rates ( $5 \times 10^{-6}, 5 \times 10^{-5}, 5 \times 10^{-4}, 5 \times 10^{-3}$ , and  $5 \times 10^{-2}$ ) (Haller, Messer, and Hernandez 2019). Haploid populations with no recombination event were simulated. The population size was 1,000 per subpopulation, and the mutation rates were  $3 \times 10^{-5}, 1 \times 10^{-5}$ , and  $1 \times 10^{-6}$  for neutral, deleterious, and beneficial mutations, respectively. We considered ten pairs of locations in simulations, which is analogous to five countries per continent. Each simulation scenario was repeated 100 times, and the average was presented.

## Results

We developed and applied the new method based on the change in  $d_N/d_S$  to the dengue virus to understand recent selection in all four serotypes. First, to know if the samples showed multiple levels of divergence time, we examined the population structure of dengue viruses using two approaches: (1) synonymous changes

**Table 1.** The components of the pMK test.

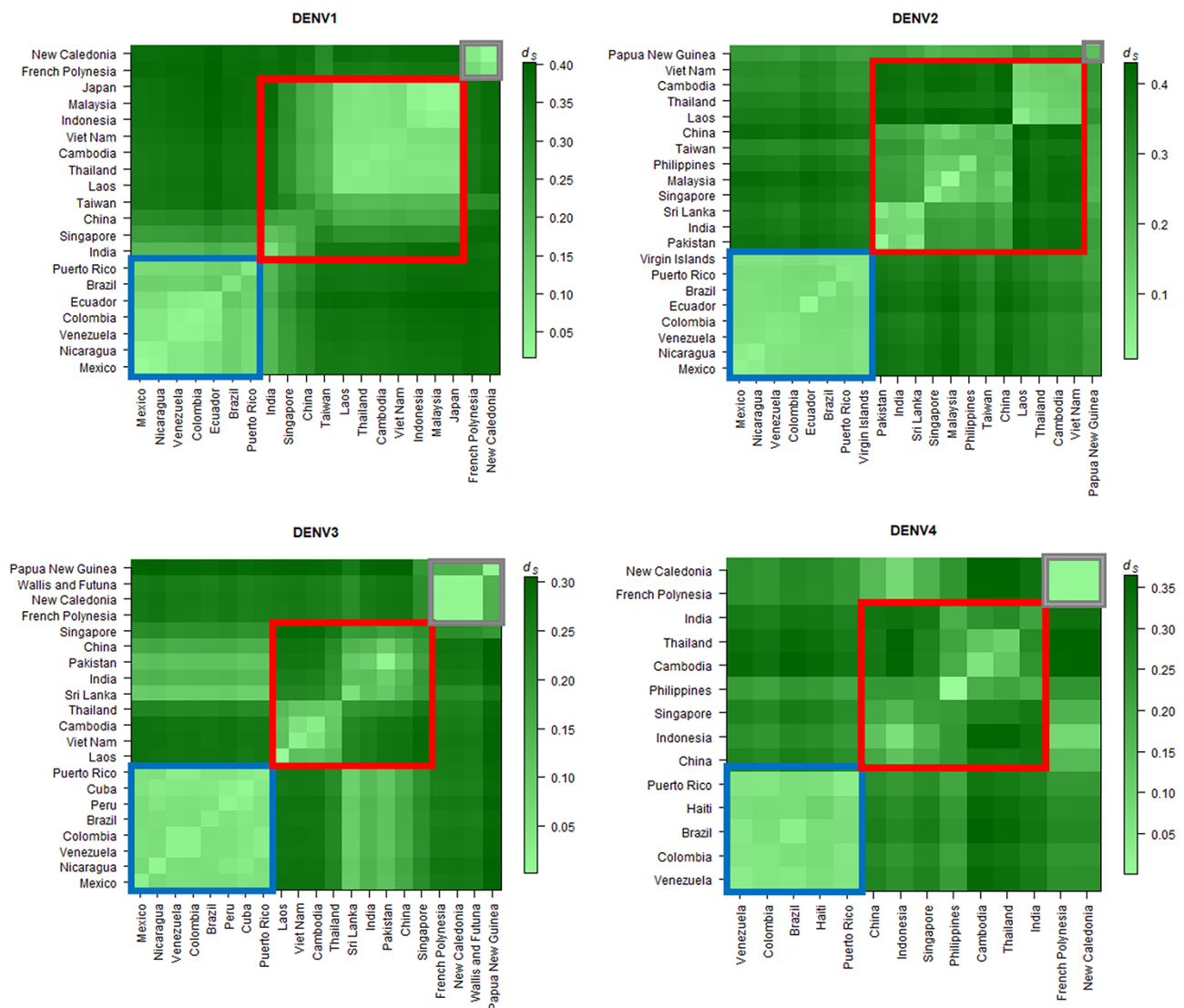
	Between continents	Within continent
Non-synonymous	A	C
Synonymous	B	D

and (2) a Bayesian clustering method. Assuming that synonymous changes are neutral,  $d_s$  is expected to increase with the divergence time between samples (Nei and Kumar 2000) and therefore was used to reveal relative divergence time across geographic scales. The Bayesian clustering method, *rhierBAPS*, assigns samples that are genetically more similar to the same groups. Results from both approaches suggest a clear population structure shaped by continental differences (Fig. 1 and Supplementary Fig. S1). Varying levels of substructure in Asia were found among serotypes by pairwise  $d_s$ , and the divergence among countries in the Americas was lower than that among countries in Asia for all the serotypes (Fig. 1). In addition, groups identified using a Bayesian clustering method (called as 'BAPS groups') tended to be composed of samples from the same continents (Supplementary Fig. S1), revealing substantial genetic differentiation between continents.

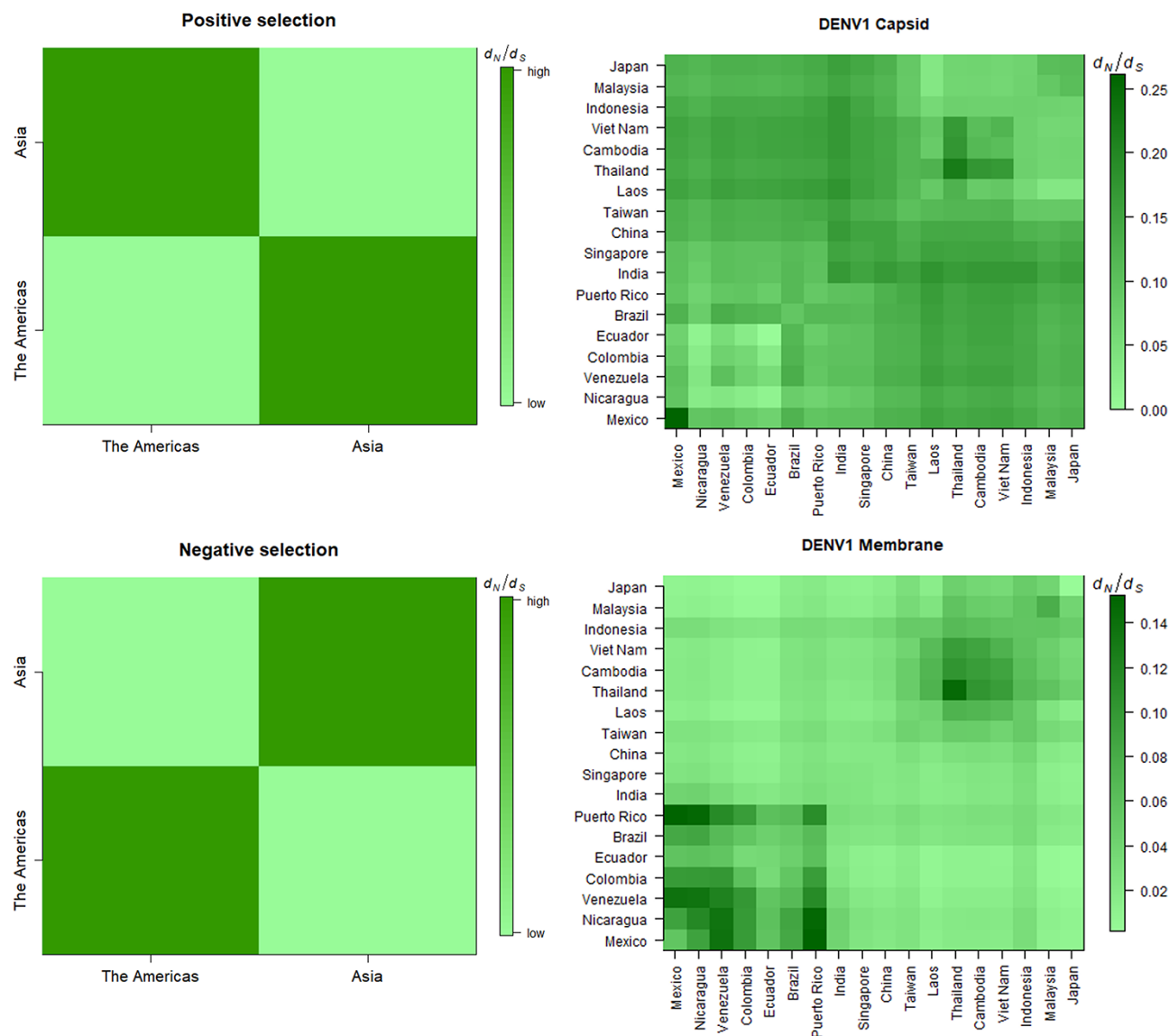
Since both synonymous changes and Bayesian clustering suggest that gene flow between continents was significantly lower

than that within continents, we examined the changes in average  $d_N/d_S$  and inferred the dominant selective force using samples covering broad geographic scales and representing a wide range of divergence times. Since DENV2 and DENV3 showed a population substructure within Asia, we separated Asian countries into two subgroups for the following analysis (Fig. 1). To obtain the average  $d_N/d_S$  for each pair of countries, we calculated the averages of  $d_N$  and  $d_S$  over all pairs of samples and then took the ratio. For genes dominated by positive selection, the average  $d_N/d_S$  should increase with the increasing geographical distance and the opposite was expected for genes dominated by negative selection (Fig. 2). It is also possible that the selective force varied between continents (Supplementary Fig. S2)—for example, it could appear only in one continent or differ between two continents.

To further determine the statistical significance of changing  $d_N/d_S$ , we developed a pMK test (Supplementary Fig. S3). The first dimension of the contingency table (Table 1) is synonymous



**Figure 1.** Population substructure of dengue viruses. The population substructure of dengue viruses was characterized by the average synonymous substitution rates ( $d_s$ ) between each pair of countries. Between-continent differentiation was higher than within-continent differentiation. Countries in Asia, the Americas, and Oceania are indicated by red, blue, and grey borders, respectively. For DENV2 and DENV3, Asian countries were separated into two subgroups (DENV2 group 1 includes Laos, Thailand, Cambodia, and Vietnam; DENV2 group 2 includes Pakistan, India, Sri Lanka, Singapore, Malaysia, Philippines, Taiwan, and China; DENV3 group 1 includes Laos, Vietnam, Cambodia, Thailand, and Singapore; DENV3 group 2 includes Sri Lanka, India, Pakistan, and China).

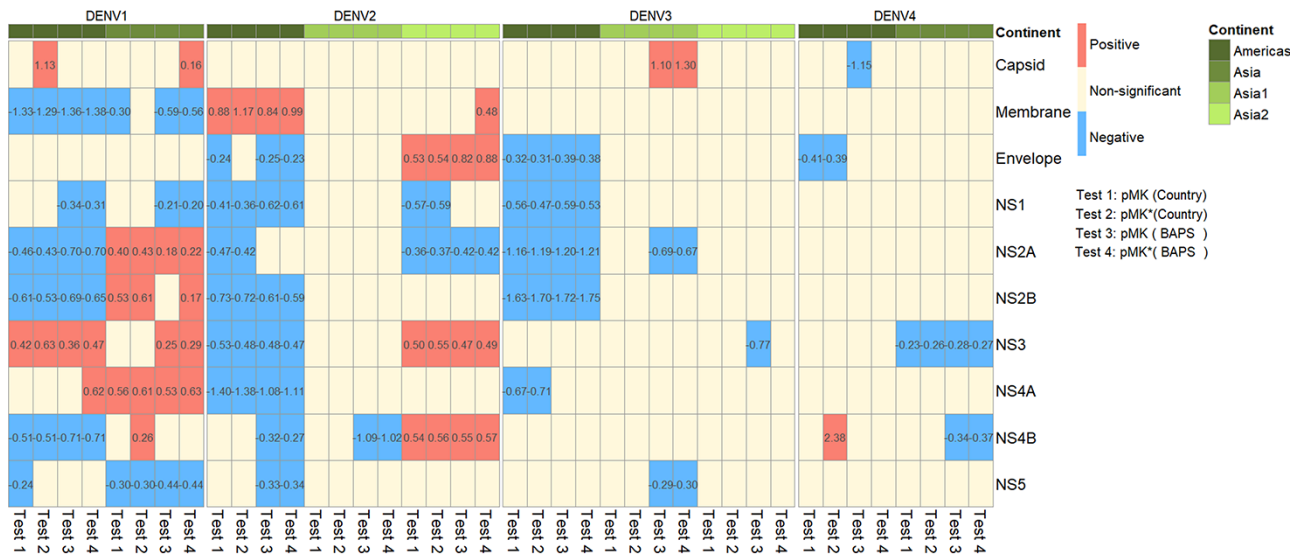


**Figure 2.** The patterns of  $d_N/d_S$  under positive and negative selection. For genes dominated by positive selection, we expected larger between-continent  $d_N/d_S$ ; for genes dominated by negative selection, we expected larger within-continent  $d_N/d_S$ . For each of the two scenarios, an empirical example is shown on the right.

or non-synonymous, and the second dimension is within or between continents. Compared to synonymous changes, more non-synonymous changes between continents than within continents were expected in a gene dominated by positive selection and the opposite applied to negative selection. With the contingency table (Table 1), we calculated the log odds ratio  $\left(\frac{A/B}{C/D}\right)$  to quantify the intensity of the signal and the S ratio ( $B/D$ ) to quantify the difference in the divergence time. The median log S ratio was 1.46 (DENV1), 1.59 (DENV2), 1.55 (DENV3), and 1.45 (DENV4) for the Americas; 0.46 (DENV1) and 0.01 (DENV4) for Asia; 1.15 (DENV2) and 0.45 (DENV3) for Asia group 1; and 0.65 (DENV2) and 0.57 (DENV3) for Asia group 2, suggesting a higher difference in divergence times when the pMK test was applied to the Americas. Because geographic locations of imported cases did not reflect where the infection occurred and could potentially influence our analysis, we determined ‘continents’ in two ways: first, by each

sample’s geographic location and second by the geographic origin of the majority of samples within a group identified by *rhierBAPS* (Fig. 3 and Supplementary Fig. S4).

We found that selective forces varied between serotypes (in the capsid, membrane, NS3, and NS4A genes) and continents (envelope, NS2A, NS2B, and NS3) (Fig. 3). The number of genes showing evidence of positive selection was higher in Asia than in the Americas. The results from the tests based on country grouping and BAPS grouping were consistent, with the latter providing more significant results (59 [country] vs 73 [BAPS]). To examine the impact of outliers, we also performed the test with only values within the interquartile range (pMK\* test). The results generally remained similar, with some differences in statistical significance (5 [country] or 7 [BAPS] out of 100 pairs of tests) but no difference in the type of selection. The intensity of the signal ranged from 0.16 to 2.38 for positive selection and from -1.75 to -0.20 for negative selection, with the capsid of DENV1 and DENV3 and the



**Figure 3.** The summarized results of the pMK test. The figure summarizes the results of the pMK test for ten genes across all four serotypes and two continents (the Americas, Asia, Asia group 1, and Asia group 2). The color of each box indicates positive selection (red), negative selection (blue), or non-significant results (ivory); the log odds ratios of significant results are shown in the box. In Test 1 and Test 2, the continent of a sample was determined by its geographic location (labeled by 'country'). In Test 3 and Test 4, the continent of a sample was determined by the geographic origin of the majority of samples within a BAPS group (labeled by 'BAPS'). In Test 1 and Test 3, all the values were used (pMK test); in Test 2 and Test 4, only values within the interquartile range were used (pMK\* test). Selective forces varied between serotypes (in the capsid, membrane, NS3, and NS4A genes) and continents (envelope, NS2A, NS2B, and NS3). The results from Test 1 to Test 4 were consistent, while the number of significant results differed.

membrane of DENV2 showing the greatest signal of positive selection and NS2B of DENV3 showing the strongest signal of negative selection (Fig. 3).

We further characterized the selective force for each domain of the two genes with known structures (Modis et al. 2004, 2005; Nayak et al. 2009; Hertz et al. 2017), the envelope protein, and NS1 (Supplementary Fig. S4 and Table S1). While NS1 was found to be under purifying selection if the whole gene was analyzed, evidence of positive selection was found in Domain 2, suggesting varying selection forces among different functional domains. A similar pattern of varying selective force across domains was also found in the envelope protein. For example, for Serotype 3 in the Americas, the whole gene and Domain 2 showed signals of negative selection, while Domain 1 was shown to be under positive selection.

Finally, we performed simulations to explore how the variation of migration rates between populations and selection intensity influenced the power of the pMK test and the magnitude of the selection signal (Fig. 4 and Supplementary Fig. S5). As expected, the power of the pMK test and the intensity of the signal increased with the strength of selection. Moreover, the larger the difference in migration rates between populations, or say, the larger the difference in divergence times, the higher the power of the pMK test and the magnitude of the selection signal.

When the log S ratio, which represents the difference in divergence times, was greater than 1 (Supplementary Fig. S6), positive selection with a selection coefficient greater than 0.0025 and negative selection with a selection coefficient smaller than -0.01 can generally be detected (Supplementary Fig. S6). Since the magnitude of the signal was influenced by both the selection coefficient and the difference in migration rates (Fig. 4C and Supplementary Fig. S5C), the log odds ratio cannot be used to infer the strength of selection directly; however, under the same set of migration rates, the log odds ratio can be used to compare the relative strength of selection between genes.

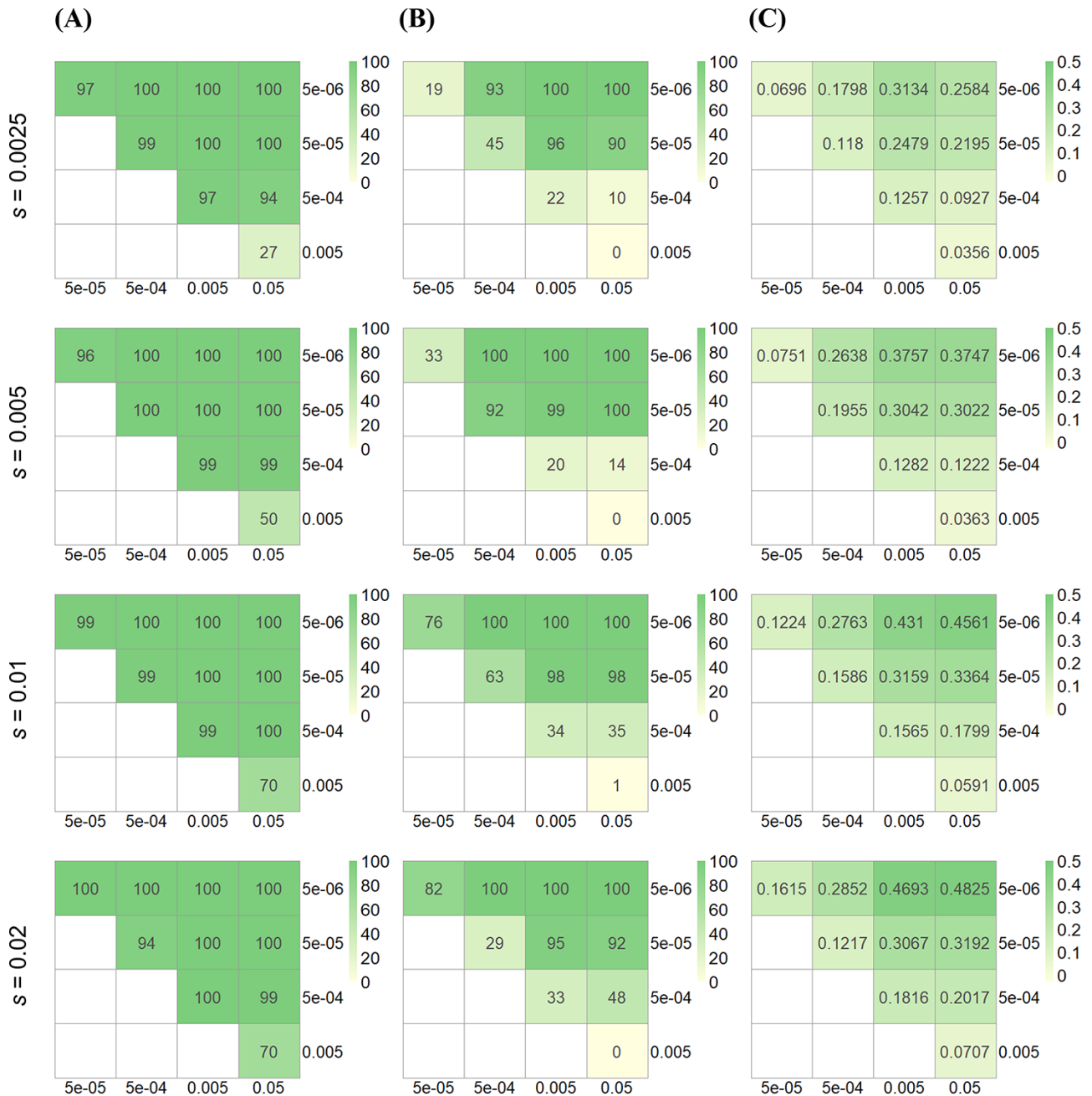
## Discussion and conclusion

We developed a new approach based on the change in  $d_N/d_S$  to infer the dominating selective force using within-species polymorphism data. Because selective constraints likely influence most sites in most genes, if considering a gene as a whole,  $d_N/d_S$  is rarely greater than 1. However, some particular sites in a gene may be under positive selection, and using the threshold of 1 to identify positive selection can potentially overlook interesting and important biological observations. Since our method does not require  $d_N/d_S$  to be greater than 1 for positive selection, it better captures these signals. In fact, none of the genes showed average  $d_N/d_S$  greater than 1 in our analysis. Moreover, using the average pairwise  $d_N/d_S$  across geographic scales in the pMK test, our method considers all the polymorphic differences, not just the fixed differences between locations or species that is used in the MK test (e.g. Parsch, Zhang, and Baines 2009; Fay 2011), increasing the chance of capturing recent selective events. However, it is important to note that, even if a gene is inferred to be 'dominated' by positive selection, some sites (and likely the majority) may still be under purifying selection.

Our analysis did not assume a single value of  $d_N/d_S$  across all geographic locations, nor did we assume the same selective force throughout the gene. However, similar to other studies using synonymous sites as a control (Spielman and Wilke 2015), we did assume that synonymous mutations are neutral. Thus, if synonymous sites are under selection, it can lead to false-positive or false-negative results. Potentially, comparing the change in  $d_S$  between genes across geographic scales can provide insights into selection on synonymous sites and help interpret the overall results of  $d_N/d_S$ . Moreover, we also assumed that, compared to the difference in divergence time between samples from different continents, the difference in sampling time is relatively small.

Through this new approach, we identified dominant selective forces acting on the evolution of four dengue serotypes. We found signatures of positive selection in the envelope proteins,





**Figure 4.** Simulations showed that the power of the pMK test varied with migration rates and selection coefficients (positive selection). The power of the pMK test across different levels of divergence time was examined through simulating various migration rates. Higher migration rates represent lower divergence times. (A) The number in each box indicates the number of replicates showing the expected pattern of positive selection (i.e.  $A/B > C/D$ ) out of 100 replicates. The majority of replicates showed expected patterns, and the consistency was the lowest when migration rates were both high. (B) The number in each box indicates the number of replicates with significant pMK test results (p-value  $< 0.05$ ) out of 100 replicates. The number of significant replicates increased with the difference in migration rates and the selection coefficient. (C) The number in each box represents the average log odds ratio (i.e.  $\log\left(\frac{A/B}{C/D}\right)$ ). The average log odds ratio increased when the selection coefficient and the difference in migration rates increased.

NS2A, NS2B, and NS4B, which have been reported in previous studies (Twiddy et al. 2002a; Twiddy, Woelk, and Holmes 2002b; Bennett et al. 2003; Lin et al. 2019). Additionally, we found that positive selection also acted on the evolution of capsid, membrane, NS3, and NS4A in at least one region and one serotype. By comparing our results with a previous study based on the site model (Twiddy, Woelk, and Holmes 2002b), we found that our approach was able to capture signals of positive selection for a

gene even when only a small proportion of sites were under positive selection (e.g. envelope of DENV2). Moreover, while NS1 was shown to be under purifying selection in our study and previous studies (Twiddy, Woelk, and Holmes 2002b; Lin et al. 2019), we performed additional analysis for each domain separately and identified domains of NS1 influenced by positive selection. Finally, we found that the dominant selective forces inferred in this study differed between continents in some of the genes, and the reason

behind this remains to be explored. Since dengue viruses became widespread in the Americas later than in Asia, it is possible that (1) selective pressures imposed by human immune responses differed between continents due to the difference in the proportion of the population that had been infected by dengue viruses before or (2) the recent bottleneck decreased the power to detect positive selection in the Americas (Parsch, Zhang, and Baines 2009).

While we applied our method to an organism with a clear population structure, it can also be used for organisms where the population structure is either more continuous or less obvious. If there is a clear pattern of isolation by distance, a correlation test or segmented linear regression between  $d_N/d_S$  and the geographical distance can potentially be used to infer dominant selective forces. A similar concept can be applied to temporal data when sufficient genomic data from multiple time points are available.

In summary, this study presents a new and simple method to detect selection at the within-species level. Since publicly available genomic data from locations over broad geographical scales are more common than temporal data, our idea of using samples across geographic scales to represent a range of divergence time opens up opportunities for more organisms. Our results suggest that dominant selective forces varied among serotypes and continents in dengue viruses and provide insights into the evolution and biology of dengue viruses and candidate regions that may warrant further investigation.

## Data availability

The sequences used in this study and the example code for the pMK test can be downloaded from our GitHub repository ([https://github.com/hhc-lab/dengue\\_selection](https://github.com/hhc-lab/dengue_selection)).

## Supplementary data

Supplementary data are available at *Virus Evolution* online.

## Acknowledgements

The authors thank Yu-Ping Poh, Taj Azarian, Caroline Buckee, and Brian Arnold for helpful discussions. This study was supported by the National Science and Technology Council in Taiwan (NSTC 111-2636-B-007-011). H.-H.C. was supported by the Yushan Scholar Program. J.C. was funded by ERC grant no. 742158. The funders had no role in the preparation of the manuscript.

**Conflict of interest:** None declared.

## References

- Azarian, T. et al. (2018) 'Global Emergence and Population Dynamics of Divergent Serotype 3 CC180 Pneumococci', *PLoS Pathogens*, 14: e1007438.
- Bedford, T. et al. (2010) 'Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2)', *PLoS Pathogens*, 6: e1000918.
- Bennett, S. N. et al. (2003) 'Selection-Driven Evolution of Emergent Dengue Virus', *Molecular Biology and Evolution*, 20: 1650–8.
- Bhatt, S., Holmes, E. C., and Pybus, O. G. (2011) 'The Genomic Rate of Molecular Adaptation of the Human Influenza A Virus', *Molecular Biology and Evolution*, 28: 2443–51.
- Bhatt, S., Katzourakis, A., and Pybus, O. G. (2010) 'Detecting Natural Selection in RNA Virus Populations Using Sequence Summary Statistics', *Infection, Genetics and Evolution*, 10: 421–30.
- Duong, V. et al. (2015) 'Asymptomatic Humans Transmit Dengue Virus to Mosquitoes', *Proceedings of the National Academy of Sciences of the United States of America*, 112: 14688–93.
- Fay, J. C. (2011) 'Weighing the Evidence for Adaptation at the Molecular Level', *Trends in Genetics*, 27: 343–9.
- Fried, J. R. et al. (2010) 'Serotype-Specific Differences in the Risk of Dengue Hemorrhagic Fever: An Analysis of Data Collected in Bangkok, Thailand from 1994 to 2006', *PLoS Neglected Tropical Diseases*, 4: e617.
- Gratton, P. et al. (2017) 'A World of Sequences: Can We Use Georeferenced Nucleotide Databases for a Robust Automated Phylogeography?', *Journal of Biogeography*, 44: 475–86.
- Haller, B. C., Messer, P. W., and Hernandez, R. (2019) 'SLiM 3: Forward Genetic Simulations beyond the Wright-Fisher Model', *Molecular Biology and Evolution*, 36: 632–7.
- Hasegawa, M., Cao, Y., and Yang, Z. (1998) 'Preponderance of Slightly Deleterious Polymorphism in Mitochondrial DNA: Nonsynonymous/Synonymous Rate Ratio Is Much Higher within Species than between Species', *Molecular Biology and Evolution*, 15: 1499–505.
- Hedge, J., Wilson, D. J., and Ouellette, F. (2016) 'Practical Approaches for Detecting Selection in Microbial Genomes', *PLOS Computational Biology*, 12: e1004739.
- Hertz, T. et al. (2017) 'Antibody Epitopes Identified in Critical Regions of Dengue Virus Non-Structural 1 Protein in Mouse Vaccination and Natural Human Infections', *The Journal of Immunology*, 198: 4025–35.
- Holmes, E. C., and Twiddy, S. S. (2003) 'The Origin, Emergence and Evolutionary Genetics of Dengue Virus', *Infection, Genetics and Evolution*, 3: 19–28.
- Kryazhimskiy, S., Plotkin, J. B., and Gojoberi, T. (2008) 'The Population Genetics of  $dN/dS$ ', *PLoS Genetics*, 4: e1000304.
- Lin, J. J. et al. (2019) 'Many Human RNA Viruses Show Extraordinarily Stringent Selective Constraints on Protein Evolution', *Proceedings of the National Academy of Sciences of the United States of America*, 116: 19009–18.
- McDonald, J. H., and Kreitman, M. (1991) 'Adaptive Protein Evolution at the Adh Locus in *Drosophila*', *Nature*, 351: 652–4.
- Miotto, O. et al. (2013) 'Multiple Populations of Artemisinin-Resistant *Plasmodium falciparum* in Cambodia', *Nature Genetics*, 45: 648–55.
- Modis, Y. et al. (2004) 'Structure of the Dengue Virus Envelope Protein after Membrane Fusion', *Nature*, 427: 313–9.
- et al. (2005) 'Variable Surface Epitopes in the Crystal Structure of Dengue Virus Type 3 Envelope Glycoprotein', *Journal of Virology*, 79: 1223–31.
- Mugal, C. F. et al. (2020) 'Polymorphism Data Assist Estimation of the Nonsynonymous over Synonymous Fixation Rate Ratio Omega for Closely Related Species', *Molecular Biology and Evolution*, 37: 260–79.
- Mugal, C. F., Wolf, J. B., and Kaj, I. (2014) 'Why Time Matters: Codon Evolution and the Temporal Dynamics of  $dN/dS$ ', *Molecular Biology and Evolution*, 31: 212–31.
- National Center for Biotechnology Information. (1988) Bethesda, MD, National Library of Medicine (US), National Center for Biotechnology Information. <<https://www.ncbi.nlm.nih.gov/>> accessed 11 Feb 2020.
- Nayak, V. et al. (2009) 'Crystal Structure of Dengue Virus Type 1 Envelope Protein in the Postfusion Conformation and Its Implications for Membrane Fusion', *Journal of Virology*, 83: 4338–44.
- Nei, M., and Kumar, S. (2000) *Molecular Evolution and Phylogenetics*. Oxford; New York: Oxford University Press.
- Park, D. J. et al. (2015) 'Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone', *Cell*, 161: 1516–26.

- Parsch, J., Zhang, Z., and Baines, J. F. (2009) 'The Influence of Demography and Weak Selection on the McDonald-Kreitman Test: An Empirical Study in *Drosophila*', *Molecular Biology and Evolution*, 26: 691–8.
- Peterson, G. I., and Masel, J. (2009) 'Quantitative Prediction of Molecular Clock and  $K_a/K_s$  at Short Timescales', *Molecular Biology and Evolution*, 26: 2595–603.
- Pybus, O. G. et al. (2007) 'Phylogenetic Evidence for Deleterious Mutation Load in RNA Viruses and Its Contribution to Viral Evolution', *Molecular Biology and Evolution*, 24: 845–52.
- Rocha, E. P. et al. (2006) 'Comparisons of dN/dS are Time Dependent for Closely Related Bacterial Genomes', *Journal of Theoretical Biology*, 239: 226–35.
- Romualdi, C. et al. (2002) 'Patterns of Human Diversity, within and among Continents, Inferred from Biallelic DNA Polymorphisms', *Genome Research*, 12: 602–12.
- Shapiro, B. J. et al. (2009) 'Looking for Darwin's Footprints in the Microbial World', *Trends in Microbiology*, 17: 196–204.
- Spielman, S. J., and Wilke, C. O. (2015) 'The Relationship between dN/dS and Scaled Selection Coefficients', *Molecular Biology and Evolution*, 32: 1097–108.
- Storey, J. D. et al. (2020) *Qvalue: Q-Value Estimation for False Discovery Rate Control*. R Package Version 2.20.0. <<http://github.com/jdstorey/qvalue>> accessed 1 Mar 2021.
- Tonkin-Hill, G. et al. (2018) 'Rhierbaps: An R Implementation of the Population Clustering Algorithm Hierbaps', *Wellcome Open Research*, 3: 93.
- Twiddy, S. S. et al. (2002a) 'Phylogenetic Relationships and Differential Selection Pressures among Genotypes of Dengue-2 Virus', *Virology*, 298: 63–72.
- Twiddy, S. S., Woelk, C. H., and Holmes, E. C. (2002b) 'Phylogenetic Evidence for Adaptive Evolution of Dengue Viruses in Nature', *Journal of General Virology*, 83: 1679–89.
- Wilson, D. J., and Consortium, C. R. (2020) 'GenomeMap: Within-Species Genome-Wide dN/dS Estimation from over 10,000 Genomes', *Molecular Biology and Evolution*, 37: 2450–60.
- Wilson, D. J. et al. (2011) 'A Population Genetics-Phylogenetics Approach to Inferring Natural Selection in Coding Sequences', *PLoS Genetics*, 7: e1002395.
- Yang, Z. (1997) 'PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood', *Computer Applications in the Biosciences : CABIOS*, 13: 555–6.
- (2007) 'PAML 4: A Program Package for Phylogenetic Analysis by Maximum Likelihood', *Molecular Biology and Evolution*, 24: 1586–91.
- Yukilevich, R. et al. (2010) 'Patterns and Processes of Genome-Wide Divergence between North American and African *Drosophila melanogaster*', *Genetics*, 186: 219–39.