

EVOLUTIONARY GENETICS

Effect of spatial sampling on phylogeography*Proc. Natl Acad. Sci. USA* **118**, e2105273118 (2021)

Phylogeography, the analysis of genetic evolution over a range of geographical locations, provides insights into evolutionary dynamics across space and time. An ongoing issue in the field is that the conclusions drawn using genetic data sampled across different spatial locations are riddled with biases related to the practical limitations of spatial sampling per se. For example, a key assumption of spatial sampling is that sampling intensity mirrors population density. However, sampling intensity is influenced by various factors, such as socioeconomic factors (for instance, wealthier regions may have better resources for data sampling than more disadvantaged areas) and ease of sampling in certain areas (for instance, it may be more difficult to sample from an underdeveloped area than from a developed one due to the lack of access to transportation in the former), among other factors. Ultimately, these assumptions can lead to biased estimation of key biological parameters.

To better study and address the impact of these biases in phylogeography, Stéphane Guindon and Nicola De Maio recently proposed a mathematical framework that includes two different sampling techniques: the detection sampling scheme considers that the data collection process is proportional to the population density, while, in the survey scheme, the sampled locations do not convey information about the evolutionary process. Both sampling schemes are implemented using a Bayesian

parameter inference that relies on the probability of the sequence alignment given the phylogenetic tree. Under the detection scheme, sampling is taken to be conditioned on the outcome of the evolutionary process, whereas in the sampling scheme, the outcome of the evolution is unknown.

The authors applied both sampling techniques to the West Nile virus dataset, selecting smaller samples from this dataset to mimic the two sampling schemes. They found that, in cases where data points are spatially clustered, a more accurate inference is obtained using the survey scheme. This suggests that at the start of an epidemic, when sampling is done from a higher fraction of infected individuals (meaning, the sampling is likely uniform), the detection scheme would be more suitable. However, at a later stage of the same epidemic, the virus inhabits a larger group of people, but due to lack of resources, not all individuals are sampled and the data is often clustered: in this case, the survey scheme would be a better fit. Because phylogeography has been recently used for elucidating the evolution of SARS-CoV-2, the proposed framework can provide a clearer picture of the evolutionary dynamics and of the spread of the disease, leading to more nuanced intervention strategies in the future.

Ananya Rastogi

Published online: 21 January 2022
<https://doi.org/10.1038/s43588-022-00194-4>