GENOME WATCH

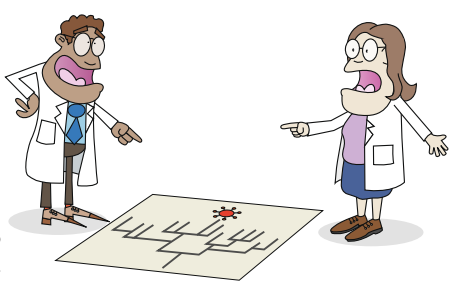
Recombination should not be an afterthought

Russell Y. Neches, Matthew D. McGee and Nikos C. Kyrpides

This month's Genome Watch highlights how the search for the origins of SARS-CoV-2 emphasizes the need for integrated phylogenetic methods.

In one view of evolution, species split into daughter species, adapt and change over time. These events can be depicted as a tree, in which each branch represents a species. In the other, individuals swap genetic material to produce genetically distinct offspring through mechanisms such as hybridization or lateral gene transfer. These mechanisms blur species boundaries or erase them altogether. These events can be depicted as a network. This is a technical duality, not a fundamental one. Evolution is one process with many mechanisms, but most software implement only a subset of these. Simplifications are necessary, but unfortunately it is not always evident which aspects of evolution drive outcomes. This is especially true for novel systems, such as COVID-19.

Writing in September 2020, the scientific consensus is that SARS-CoV-2 did not gain the ability to infect humans as the result of recombination with another virus. The road to this conclusion has been a difficult one, but the struggle holds some important lessons. In January 2020, Pradhan et al.[1] suggested in a much discussed and now withdrawn preprint that short sequence matches between SARS-CoV-2 and HIV comprised an "uncanny similarity" that might be evidence of recombination between the two wildly different viruses. It was a textbook example of reading too much into BLAST results. In another preprint in April 2020, Wang et al.[2] observed that although SARS-CoV-2 is most closely related to bat coronaviruses, it has a segment exhibiting high amino acid identity to pangolin coronavirus. Their results for divergence times were consistent with a recombination event, and they led with that hypothesis in careful language. These results can also be explained by other processes whose likelihoods were not evaluated. In July 2020, Li and colleagues[3] investigated these possibilities and arrived at similar conclusions. Unfortunately, they used a method that predates the birth and maturation of Bayesian phylogenetics, and their results have not been replicated. In July, Boni et al.[4] found that all loci of SARS-CoV-2 diverged within bat coronavirus lineages. They were able to reach what seems to be the emerging scientific consensus for three reasons. First, instead of focusing on the lineages of immediate interest, they performed a broad phylogenetic analysis of 68 ancestral lineages. Second, they used these trees to test hypotheses about the data. Third, they explicitly searched for recombination breakpoints and directly modelled distinct evolutionary processes, including recombination, that could explain their data. This allowed them to offer their conclusions about recombination on a foundation of statistical rigor.



Credit: Philip Patenall/ Springer Nature Limited

It is well known that viruses can have non-tree-like evolution. The segmented genome structure of the influenza viruses reassorts with seasonal frequency. Western equine encephalitis virus is a recombinant hybrid of two other viruses of the same genus. And molecular biologists have used viral recombination in the laboratory since 1952. As recombination behaviour is a major risk factor for developing vaccines and treatments for SARS-CoV-2, why were early efforts to model it incomplete? Phylogenetic trees are necessary to understand viral outbreak patterns and their deeper evolutionary history. Early molecular phylogenetics matured on a diet of mitochondrial data, whose inheritance pattern limits recombination. This led to the perception that evolution can be accurately modelled by simply accounting for differences in substitution rates but assuming a common topology, as in a mitochondrial genome. A lack of recombination within each investigated locus is usually assumed. Fortunately, ancestral recombination graphs can explicitly model recombination events alongside evolutionary divergence between sequences. Once regarded as computationally intractable, contemporary statistical frameworks[5] are now up to the task (for small genomes). Integrated phylogenomic frameworks can directly answer fundamental questions that merit investigation for any biological system: did recombination occur? How recently? Which lineages and loci were involved? With these powerful tools, biology's split view of evolution can be made whole.

Russell Y. Neches[1]✉, *Matthew D. McGee*[2]✉ *and Nikos C. Kyrpides*[1]✉
[1]*DOE Joint Genome Institute, Berkeley, CA, USA.*
[2]*School of Biological Sciences, Monash University, Melbourne, Victoria, Australia.*
✉*e-mail: JGI-Microbe@lbl.gov*

https://doi.org/10.1038/s41579-020-00451-1

1. Pradhan, P. et al. Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag. Preprint at *BioRxiv* https://doi.org/10.1101/2020.01.30.927871 (2020).
2. Wang, H., Pipes, L. & Nielsen, R. Synonymous mutations and the molecular evolution of SARS-Cov-2 origins. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.20.052019 (2020).
3. Li, X. et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci. Adv.* eabb9153 (2020).
4. Boni, M. F. et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* https://doi.org/10.1038/s41564-020-0771-4 (2020).
5. Vaughan, T. G. Inferring ancestral recombination graphs from bacterial genomic data. *Genetics* **205**, 857–870 (2017).