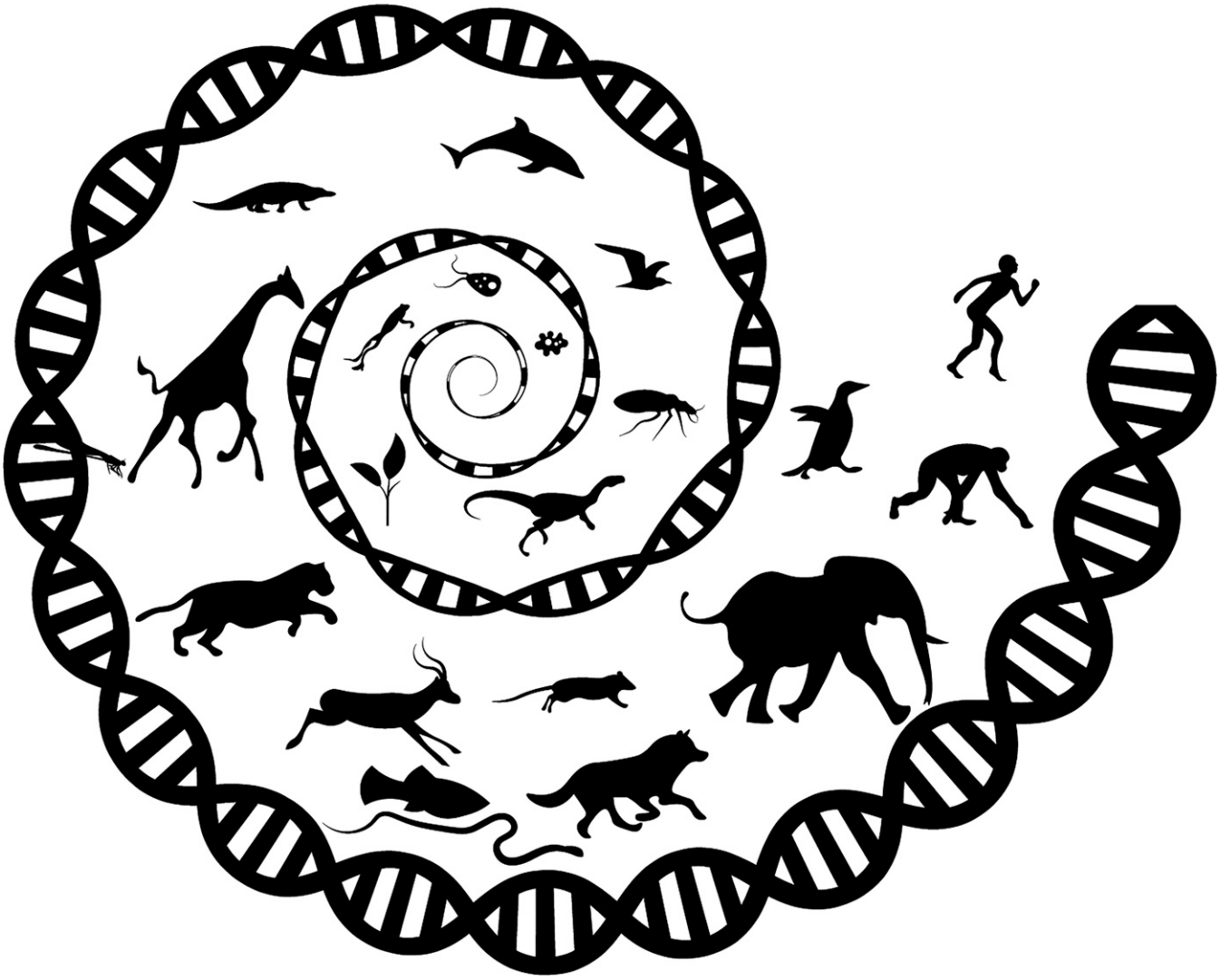
PNAS INNER WORKINGS

Machine learning in evolutionary studies comes of age

Viviane Callier, *Science Writer*



A variety of new machine learning methods are shedding light on longstanding questions about the forces that shape genomes over time. Image credit: Shutterstock/alionaprof.

Whether flowers, mosquitoes, or humans, all organisms tend to interact and reproduce with others of their ilk who are close by. In principle, that means that their genetics can reveal not only their ancestry but also their geography. Taking advantage of this simple insight, evolutionary geneticists recently showed that it was possible to create models, based on scanning thousands of genomes from mosquitoes to elephants to humans (Fig. 1), that match individual genomes to spatial locations and therefore predict where a given individual animal was born (1, 2). The approach has practical import: It could be applied in the context of ecology and conservation by, for example, tracing the origins of elephant tusks and rare woods.

It's one of a variety of new machine learning methods that are finding applications in the field, shedding light on longstanding questions about the forces that shape genomes, such as selection and genetic drift. These approaches have already demonstrated the potential to deal with the messiness of real data in ways that formal population genetic theory cannot. In the early 2000s, researchers who were "thinking a little bit ahead," says evolutionary geneticist Andrew

Published April 19, 2022.

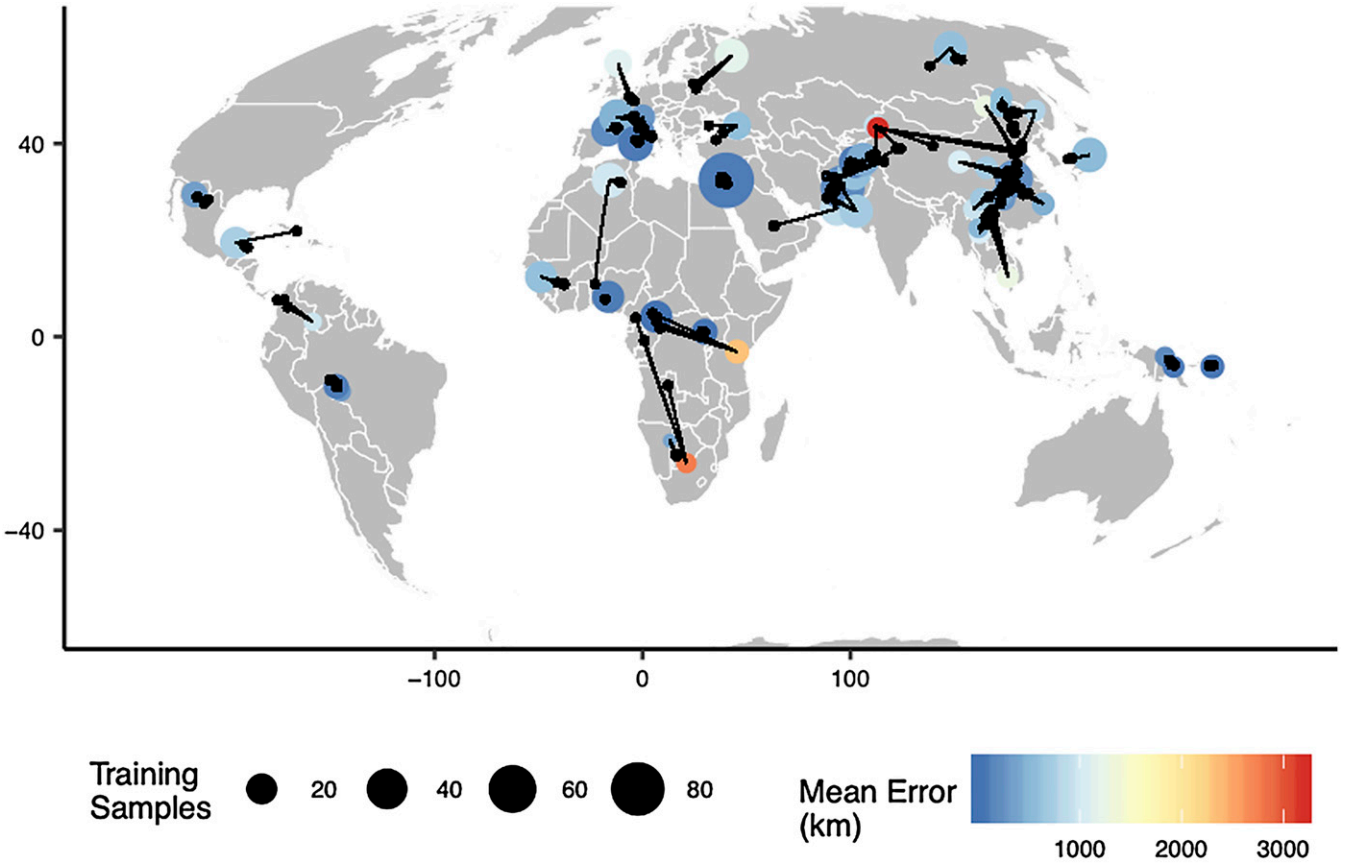


Fig. 1. In one example of how deep learning can be applied to evolutionary genetics, researchers used the known locations of individuals in the Human Genome Diversity Project to train a deep learning algorithm that matches genomes to geographic location. Then they used the algorithm to predict the geographic origin of test genomes. The colored circles represent the magnitude of error in the predictions. Reprinted from ref. 2, which is licensed under CC BY 4.0.

Kern at the University of Oregon in Eugene, "saw that there was going to be this collision happening, where we have this giant corpus of mathematical theory, and all of a sudden reality was going to intrude on our models in the form of all the genomic data that we're about to collect." What the field needed, he adds, were "more ways to bridge that gap between theory and biology." Evolutionary biologists are starting to do exactly that.

Testing Theory with Simulation

When it comes to evolution, researchers can typically observe the result of evolutionary processes but not the inputs or processes themselves. Those results include the fossil record, the diversity of living species, and the amount of genetic variation in the genomes of a population. But ideally, researchers could infer details about the processes that generated these patterns. Machine learning algorithms can serve as the map between input and output.

Machine learning algorithms are typically trained on datasets where inputs and outputs are both known. For instance, training an algorithm to recognize images that include cats might require presenting pictures, some of which have cats. With enough feedback on correct and incorrect guesses, the algorithm would eventually become adept at recognizing cats. In population genetics, researchers would like to have the equivalent of cat pictures to train machine learning algorithms to recognize the signatures of selection, drift, and migration in the genomes of real populations.

But there's a problem. Typically, the researchers don't have labeled data (the equivalent of "cat pictures") to train the algorithms. "In evolution, we never really have ground truth data," explains Jeff Spence, a geneticist and postdoctoral fellow at Stanford University, CA. A given evolutionary process, he notes, happened only once, in the past.

So researchers started to create so-called "synthetic genomes." Using powerful computers, they can create a fictional population of genomes, simulate the processes of selection or drift, and observe the resulting pattern of genetic diversity after many generations. By doing so, they have in hand both the process and the resulting pattern of genetic diversity—in other words, they have the tools necessary to train the machine learning algorithm.

As with the cat pictures, the algorithm needs a lot of training data to hone its accuracy. But once an algorithm has been trained on synthetic data, it can, in principle, be used to make an inference about past selection, drift, or migration from the observed genetic diversity in a real population.

Attempts at explicating how modern populations took shape over time bring the power of the machine learning approach into full view. South American populations, for example, have three ancestral populations: the Native American, the European colonists, and the enslaved people brought from Sub-Saharan Africa. Most previous

models of admixture assumed random mixing. In fact, populations tend to be somewhat stratified, meaning that there is some assortative mating—individuals from the same group are more likely to mate with each other. There can also be sex bias (e.g., males from one group are more likely to mate with females of another group, or vice versa). Models that assume random admixture will predict that mixing happens faster than it actually does and therefore may do a poor job estimating the timing of admixture events.

To understand admixture in South American populations, evolutionary geneticist Alex Mas-Sandoval, a postdoc at Imperial College London, UK, simulated various admixture scenarios by varying the degree of assortative mating and sex-bias, thus producing populations of simulated genomes. He then trained a machine learning algorithm to match the populations of the simulated genomes to the parameters that generated them. The final step entailed using the trained algorithm to infer the amount of stratification and sex-bias in real data from admixed South American populations. Among his findings: Males of European ancestry were more likely to mate with females with lower proportions of European ancestry, and males with lower proportions of Native American ancestry were more likely to mate with females with higher proportion of Native American ancestry.

The method itself, Mas-Sandoval says, is perhaps more important than these results. "Until now, most of the studies analyzing admixture in the Americas were assuming random mating, which is not obviously the case, because all the populations of the cities of the Americas are stratified," he says.

Careful Assumptions

Although very powerful, these simulation-based machine learning approaches do share important caveats with other computational methods for evolutionary inference—not to mention traditional population genetic approaches. All these methods hinge on investigators' assumptions about the evolutionary process, including factors such as population size, the frequency of mutations, and the strength of selection. The approaches "rely on prior beliefs about models and their parameterizations," explains Kern's former postdoc Dan Schrider, who is now an assistant professor at University of North Carolina, Chapel Hill.

To make a reliable inference with machine learning methods, the training data have to be very similar to the real data about which researchers want to make an inference. "It's very difficult to create realistic simulated data," explained Sara Mathieson, a computer scientist at Haverford College, PA, in a talk at the Society for the Study of Evolution virtual meeting last June. "And if the data used to train the model [are] not a good fit for the real data, then we can't really be confident of the [inference] results on the real data."

To solve this problem, Mathieson has been using something called generative adversarial networks (GANs)—the same types of algorithms that are used to make deep fake images of human faces or voices. The GAN algorithms work by pitting a data generator (the population simulator) against a discriminator. In an iterative process, the generator simulates genomes, and the discriminator tries to tell the fake from the real. As the generator and discriminator try to outdo each other, the generator starts to produce simulated genomes that are increasingly similar to the real population of genomes. Eventually, explains Mathieson, the algorithm homes in on the specific processes and parameter values—such as the degree of mixing, migration, or selection—that produced the real genomes.

Flora Jay, a geneticist at Paris-Saclay University, France, and her colleagues have also been using GANs to create simulated genomes, but with a different aim. Simulated genomes can capture the useful statistical features of the genomes of special populations without risking any loss of privacy—a growing concern with real human data. Indeed, while some populations are well represented in public genome databases, other populations are not; this can lead to biases in analyses. To illustrate this issue, Jay and her colleagues created simulated genomes from samples in an Estonian biobank dataset that was not publicly available (3). The genomes they created with the GAN method successfully captured the statistical features of the Estonian biobank population—which the team confirmed by collaborating with researchers from Tartu University, Estonia, who have special permission to access the actual biobank data. She argues that these simulated genomes could help augment public genome databases without risking the loss of privacy of individuals in that biobank.

Detecting Subtle Signals

One of the strengths of machine learning approaches is that they can pick up signals of selection that are too subtle for traditional methods to detect. For example, Schrider and Kern used machine learning to distinguish between two different varieties of selection in human genomes: soft and hard selective sweeps. Hard selective sweeps occur when a new mutation arises and confers an advantage, so it spreads rapidly through the population. Soft selective sweeps occur when a change in the environment suddenly makes a preexisting mutation advantageous, enabling that preexisting mutation to spread throughout the population. Soft selective sweeps often mean adaptation happens more quickly, because the population doesn't have to wait for a mutation to arise.

Schrider and Kern simulated hard and soft selective sweeps and then trained a machine learning algorithm to distinguish simulated genomes that resulted. They then applied the trained algorithm to examine hard and soft sweeps in real human genomes. Because human population size was quite small historically, "there's not a lot of genetic diversity, and under those conditions, you expect hard sweeps to be more common," Schrider notes. "But that's not what we found. We found that while there were a number of strong clear signals of hard sweeps, there were also many more signatures of soft selective sweeps," he says (4–6).

"The classic population genetic approaches for detecting selection can be quite underpowered," explains David Enard, a population geneticist at the University of Arizona, Tucson, who studies how the selection pressure from ancient

epidemics, occurring 20,000 to 50,000 years ago, shaped human genomes (7, 8). Machine learning approaches allow researchers to detect events farther back in time, picking up subtle signals that wouldn't be captured by conventional methods, he says. Enard is now using these approaches to reanalyze data studied via conventional methods, trying to detect more subtle signals of selection from ancient viruses on human genomes, signals that he may have missed.

Ultimately, these new approaches should help inform longstanding debates about the various forces, such as drift or selection, that have shaped real human genomes. "There's still a lot that we need to do as a field to bring all of these evolutionary forces together," says Brown University (Providence, RI) population geneticist Sohini Ramachandran, "to understand what has truly shaped our genomes."

1. C. J. Battey, P. L. Ralph, A. D. Kern, Space is the place: Effects of continuous spatial structure on analysis of population genetic data. *Genetics* **215**, 193–214 (2020).
2. C. J. Battey, P. L. Ralph, A. D. Kern, Predicting geographic location from genetic variation with deep neural networks. *eLife* **9**, e54507. (2020).
3. B. Yelmen *et al.*, Creating artificial human genomes using generative neural networks. *PLoS Genet.* **17**, e1009303. (2021).
4. D. R. Schrider, A. D. Kern, S/HIC: Robust identification of soft and hard sweeps using machine learning. *PLoS Genet.* **12**, e1005928 (2016).
5. D. R. Schrider, F. K. Mendes, M. W. Hahn, A. D. Kern, Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics* **200**, 267–284 (2015).
6. D. R. Schrider, A. D. Kern, Soft sweeps are the dominant mode of adaptation in the human genome. *Mol. Biol. Evol.* **34**, 1863–1877 (2017).
7. D. Enard, D. A. Petrov, Evidence that RNA viruses drove adaptive introgression between Neanderthals and modern humans. *Cell* **175**, 360–371.e13 (2018).
8. Y. Souilmi *et al.*, An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in East Asia. *Curr. Biol.* **31**, 3504–3514.e9 (2021).