Reading the genome like a history book
Mathematical modeling sheds light on the evolution of human genetic variation

By Kelley Harris

One way to study a genome is to read it like an instruction manual. It contains genes that are easily decoded into the protein building blocks of cells, as well as much more cryptic regulatory codes that dictate when and where each protein should be produced.

Genomes are much more than instruction manuals, however; they are living, evolving documents that are constantly being passed imperfectly from generation to generation. As a graduate student, I learned to look at genetic diversity the way geologists look at sediment: as a historical record of ancient catastrophes and slow but unstoppable drift.

The key to interpreting historical records encoded within the genome turns out to be a body of theory that predates the knowledge that genes come in the shape of a double helix. Starting early in the 20th century, population geneticists like Sewall Wright, Ronald Fisher, and Motoo Kimura explained evolutionary dynamics using equations that describe how ink diffuses through water (1, 2). Kingman later saw the wisdom of turning time backwards and calculating the distribution of possible founding ancestors of a population that exists in the present (3).

Whether they existed in the past or the present, the populations studied by Kingman and his predecessors were fundamentally abstract. Fast forward to the year 2010, however, and a population could suddenly mean a compressed data file of a thousand human genomes. I spent my Ph.D. years working with Rasmus Nielsen and Yun Song to learn what I could about human evolution from the DNA sequences that had been suddenly immortalized by the 1000 Genomes Project (4).

Even in an alignment of 1000 human genomes, most sites are extremely boring, with every sampled chromosome carrying the same allele. However, perhaps 1 out of 100 sites is polymorphic, meaning that some chromosomes carry a different allele because of an ancient mutation in some ancestral chromosome.

Rare variants present in only 1 or 2 out of 2000 chromosomes are likely to be much younger than common variants, meaning that a sample of modern-day genomes contains information about a broad time swath of history. The frequency distribution and spatial arrangement of variable sites within the genome is extremely sensitive to demographic events like population booms and busts and migration between isolated communities. Kingman’s evolutionary theory turns out to be extremely useful for calculating the relationship between present-day diversity and these features of population history.

For the first half of my thesis, I derived a set of equations that describe how past population size changes and migration events are expected to affect the distribution of distances between polymorphic sites (5, 6, 7). Fitting these models to human data, I was able to estimate the severity and duration of the out-of-Africa bottleneck: the period of inbreeding that humans experienced when they first ventured into Eurasia (5). According to our findings, the migrant population that left Africa was only about 10% the size of the original human population in Africa, and it stayed small for a long time (on the order of 40,000 years).

I also studied the distribution of polar bear and grizzly bear genetic differentiation and found that polar bears appear to be a surprisingly young species, just half a million years old (8). I also found that many common polar bear alleles were present at low frequency in a grizzly genome panel, as if they entered the grizzly species through hybridization, but common grizzly alleles did not appear to have spread to polar bears, suggesting that polar bears have successfully integrated into grizzly populations for some time, but grizzly bears have not achieved integration into polar bear communities.

It is remarkable how well Wright and Fisher were able to predict the structure of genetic variation at a time when almost nothing was known about genome architecture. At the same time, it isn’t hard to find features of real genomes that defy the predictions of classical models.

Population geneticists are constantly writing new computer programs to simulate the course of human evolution, and these programs produce simulated genomes that resemble real human genomes in many ways. The resemblance is still not perfect, however, and when we find that simulated data lack certain features of real data, this can provide clues about how our understanding of the evolutionary process remains incomplete.

One discrepancy between real and simulated genomes is that real human genomes contain many more dense clusters of mutations. I decided to study these clusters in detail and found that they could not be explained either by imperfections in DNA sequencing technology or by demographic events that we had previously failed to model. I also found that some mutation clusters in human data show less resemblance to ordinary human mutations than to certain mutation clusters that occur in yeast under conditions of experimentally induced replication stress (9).

Given this resemblance, I proposed that the presence of certain mutation clusters in human genomes could be explained by the action of error-prone DNA polymerase
Global variance in the accumulation of TCC→TTC mutations

Among the populations catalogued in the 1000 Genomes project, TCC→TTC mutations are most prevalent in the Tuscans of Southern Italy, where they make up about 2% of all polymorphisms. This is a significant increase compared to East Asia, where this mutation type makes up only 1.5% of all polymorphisms.

$\zeta$ in the human germline, which may be required to restart stalled DNA replication and can leave behind signature substitutions—for example, often changing “GC” motifs into “AA” motifs (10). Another research group quickly built upon my findings to show that certain regions of the genome appear to be hotspots of Pol-$\zeta$-related instability that can unfortunately cause severe genetic diseases (11).

After calculating that at least 2% of human diversity appears to originate from clustered mutational processes, I started wondering how many distinct mutational processes are working behind the scenes to create our genetic variation. To this end, I developed a method to test whether any mutational processes might be more active in certain human ethnic groups than in others. I looked at young, low-frequency variants that appear to be found exclusively on a single continent—for example, present in Africa but absent from Europe and Asia—and asked how they are distributed among all possible three-letter DNA motifs.

As shown in the figure, the results were unexpectedly striking: The particular DNA motif “TCC” appears to have mutated much more often within the European population than in either Africans or East Asians (12). Almost all of these extra mutations are from C to T, changing TCC into the motif TTC. Although the cause of these mutations remains a mystery for now, it is very clear that some mutagenic force became overactive in the European population during the past 20,000 years or so since Europeans and East Asians started differentiating into separate populations.

This contradicts the popular “molecular clock” model (13), which posits that mutation rates evolve very slowly over perhaps tens of millions of years. Rather, it suggests that DNA replication fidelity is a lot like other biological traits, sometimes evolving by leaps and bounds for reasons that usually elude us.

REFERENCES
8. S. Liu et al., Cell 157, 785 (2014).

10.1126/science.aar2003
Reading the genome like a history book
Kelley Harris

Science 358 (6368), 1265.
DOI: 10.1126/science.aar2003