The basic text of all life is written in a simple alphabet. Adenine, thymine, guanine, and cytosine form the units that allow DNA to write the books of Earth’s elaborately complex lifeforms. Curiously though, despite doing a seemingly equivalent job at conveying information, the prevalence of certain base pairs (G-C or A-T) in the genome is mysteriously variable, across both phyla and environment. Currently, it’s unclear why this should be.

That makes nucleotide content a fascinating area ripe for study, says Ruth Hershberg, an assistant professor at the Technion-Israel Institute of Technology in Haifa. It’s a highly variable, easy-to-measure trait. One that, she says, “affects every nucleotide in the gene. We really don’t know why it’s so variable, or what determines it, so that makes it interesting.”

It might also be important too, she says. Mutations tend to move in a G-C → A-T direction, meaning G-C rich genomes could potentially have higher mutation rates. Nucleotide content might also affect codon usage and regulatory genes, she says.

Hershberg and colleagues, including Erin Reichenberger, Gail Rosen, and Uri Hershberg (her brother) have recently begun to unpick why nucleotide content varies so widely, publishing their findings online April 9 in Genome Biology Evolution (Reichenberger et al. 2015).

Surprisingly, they found that both phylogeny and environment play a role in shaping nucleotide composition. How natural selection operates on this trait, however, remains unclear.

To do this study, they used already available genomic data (183 metagenomic samples, obtained by different labs, using different sequencing technologies) to examine how nucleotide composition varied across phyla and environments (e.g., soil, water, and the human gut). Metagenomic studies pool data from an environment, rather than starting with pure DNA from a single species.

“I don’t know where that DNA came from,” Hershberg says, “but I get a pool of small sequences that came from bacteria that is in this environment.” That allows a researcher to compare the sequenced chunks to known genomes, giving a directory of which phyla are present in the environment.

Calculating the nucleotide content of each piece of DNA allows the researcher to know, by phyla, the mean and variation of nucleotide content present within that sample. Do that for several environments and…

“What we were able to show was that, on one hand, across different environments, different phyla have characteristic nucleotide content,” says Hershberg. If you happen to be an Actinobacterium, for example, you will be relatively G-C rich (regardless of your environment) whereas your Fermicute neighbor will be more A-T rich. This, perhaps, is no more than what would be expected.

Phyla also differ in the extent to which there is variation in their nucleotide content—some having much, some having very little. Hershberg and her collaborators wanted to know, then, is the variation in average nucleotide content in different environments due to their holding different phyla, or is something else at work?

To address this, they looked at the correlations among different phyla within a particular environment. They found that they do, indeed, correlate, more than you would expect by chance. “This tells us whatever is increasing the G-C content of a certain phylum in a certain environment,” Hershberg says, “it is also increasing the G-C content for different phyla.”

That’s true not just for very different environments, but—to the researcher’s surprise—also within similar environments. This was true within different human guts, for instance. On average, the nucleotide content of my gut flora would be different from the nucleotide content of your gut flora.

“What is in the gut is determining this,” Hershberg says, “but we don’t know what that thing is.”

Their findings, while intriguing, are not necessarily to be considered ironclad. Laurent Duret, an evolutionary genomicsist at the French National Centre for Scientific Research (CNRS) in Lyon, expressed some mild misgivings. “The approach is sound,” says Duret, who was not involved in the work, “the authors clearly show that the base composition of
bacteria from different phyla covaries among metagenome samples.” That said, he suspects the effect might be due to sequencing artifacts, as the preparation of sequencing libraries can introduce biases.

It’s been shown, he says, that sequencing depth varies with the GC-content of DNA fragments (Benjamini and Speed 2012). In a metagenomic experiment, this artifact should lead to overestimating the relative abundance of GC-rich genomes, or underestimating the relative abundance of GC-poor genomes, Duret says. Until that point is investigated, he’ll remain skeptical, he says.

Hershberg discounts this kind of artifact as a major influencer in her results. “We have samples which we know have all been sequenced by the exact same technology in the same lab. We observe significant correlations even when looking only within such samples as well,” she says.

The team is now evaluating environmental parameters: Salinity, pH, and temperature, for example, to see what are the forces driving this apparent byproduct of natural selection. That, says Uri Hershberg, professor at the Drexel University College of Medicine in Philadelphia, is exciting because it addresses a foundational force in the evolution of life.

“In general we are very bad in biology thinking about meta levels of change,” he says. “Here we have something that is still a very big mystery, to study how it is changing—that’s exciting.”

**Literature Cited**

Reichenberger ER, Rosen G, Hershberg U, Hershberg R. Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. Genome Biol Evol. 7(5):1380–1389.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 40(10):e72.

Highlights editor: George Jianzhi Zhang