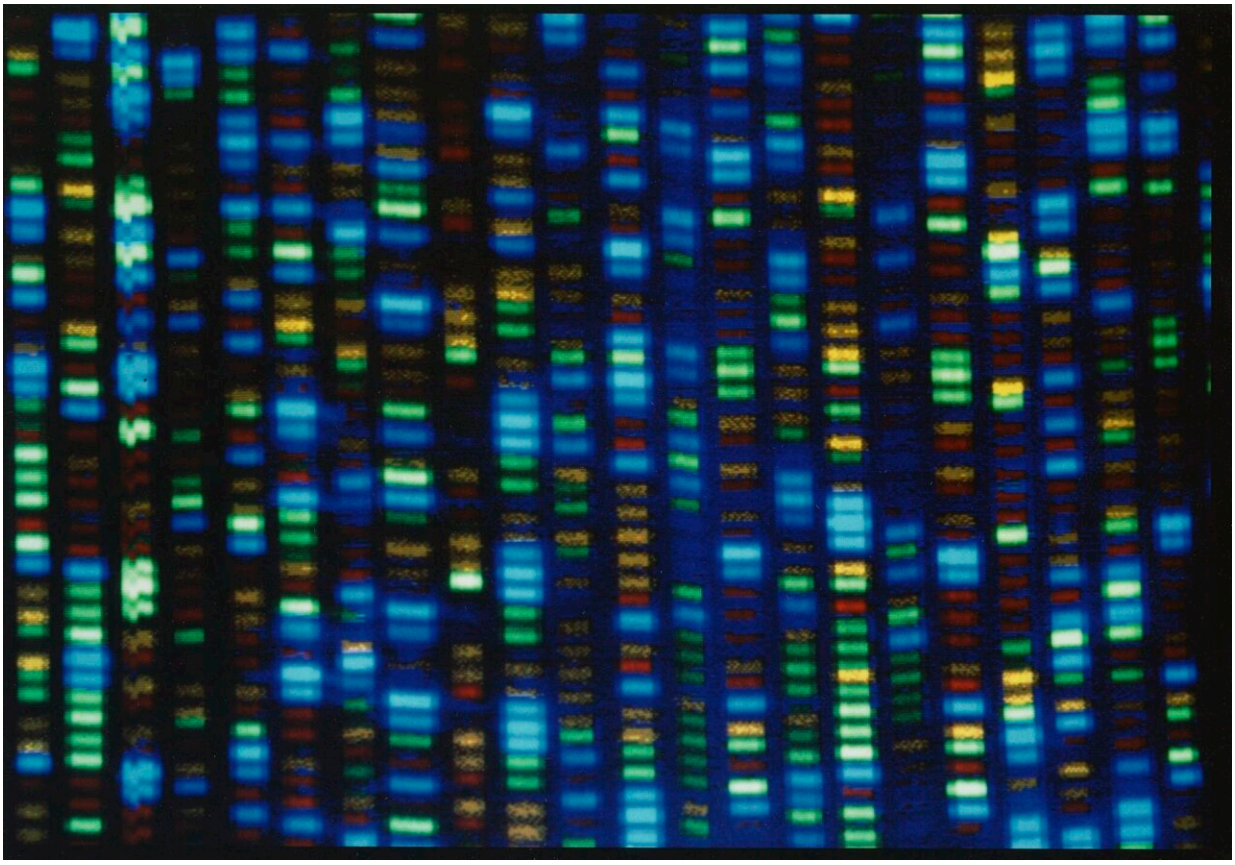
Big data's hidden cost: The carbon footprint of computational science

January 20 2023, by Craig Brierley



Output from a DNA sequencer. Credit: National Human Genome Research Institute

As the climate emergency and cost-of-living crisis focus our minds on how to reduce energy, a group of scientists have highlighted the hidden

environmental cost behind some of our major breakthroughs.

High performance computing has transformed how research works and our ability to make previously unthinkable discoveries. We're able to model our future climate with unprecedented accuracy. We're able to predict what a protein looks like from its genetic code. We even know what a black hole 55 million light-years away looks like.

But while few people would argue against such progress, it comes with a cost.

In 15 years of writing about medical research, I have found myself writing countless stories about genome-wide association studies, where researchers compare the DNA of potentially hundreds of thousands of people—patients and healthy 'controls'—to look for genetic variants that increase our risk of developing a particular disease. Never once did I find myself considering the environmental impact of such studies.

It turns out that it can be quite staggering.

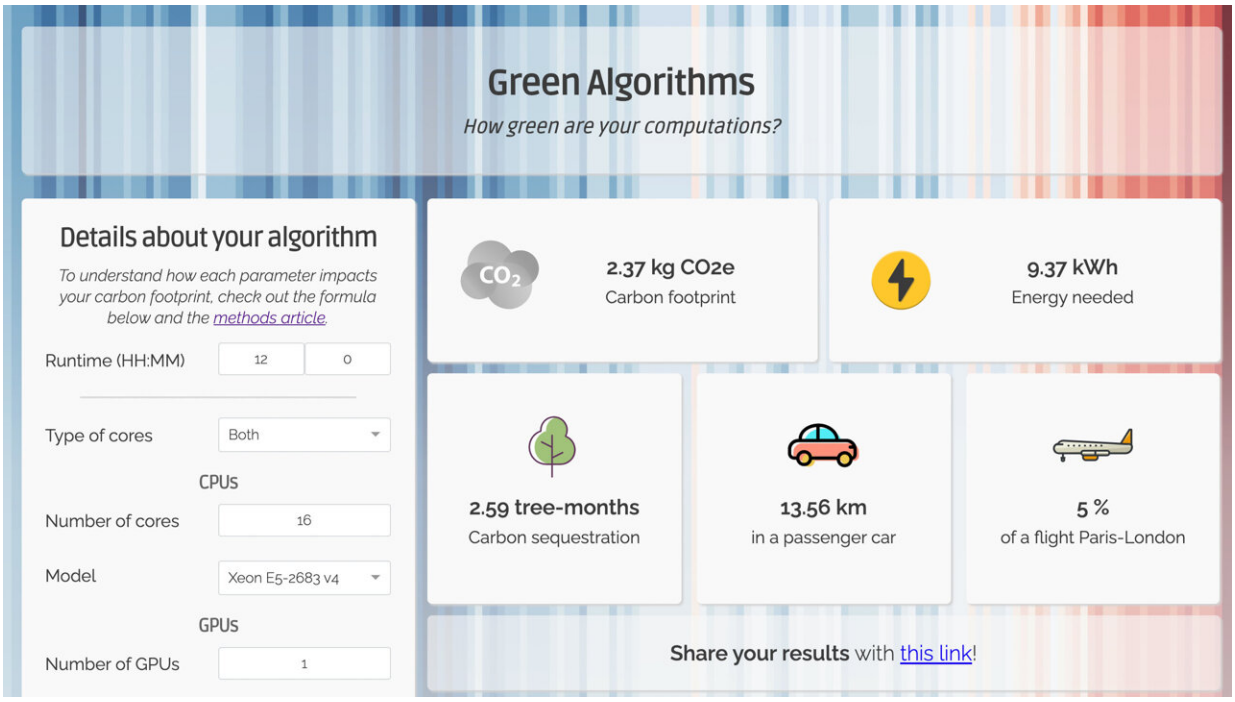
Early this year, a team from Cambridge, together with colleagues at the Baker Institute in Melbourne, Australia, published research showing that a genome-wide association study (GWAS) trawling data from 500,000 participants registered to a biobank database would create a carbon footprint of 17.3kg of CO2e (carbon dioxide equivalent) for each genetic trait being studied.

But in fact, researchers would commonly look at thousands of traits. The same GWAS run for 1,000 traits would generate 17.3 metrics tons of CO2e. That's equivalent to 346 flights between Paris and London. (The researchers point out that upgrading the software used to the latest version would reduce this by three-quarters.)

At the start of 2020, Loic Lannelongue was in the middle of a Ph.D. in health data science at Cambridge's Department of Public Health and Primary Care. He was a computational biologist, using machine learning to predict how proteins interact in the human body. One of his collaborators was Jason Grealey, an academic based at University of Melbourne, Australia. Lannelongue was watching on the news—and hearing first hand from Grealey—about the bushfires tearing through Australia. This made him reflect on the climate emergency and the part we all play.

A few months earlier, Lannelongue had read about a study that equated training artificial intelligence (AI) to the carbon footprint of five cars over their lifetimes. He began to wonder what the impact of his own work was, and together with Grealey decided to work it out, expecting to find an online calculator that they could just plug their numbers into.

"We started thinking it would be a two week project, a nice break from our Ph.D. research," says Lannelongue, "just figuring out what the carbon footprint of what we were doing was to get a number and probably tweeting about it. Except there was nothing out there. We realized that there was a massive gap, that computational scientists weren't really thinking about their carbon footprint yet."



Credit: University of Cambridge

Since then, with the support of his supervisor, Dr. Michael Inouye, Lannelongue has been spending half of his time working on this project, leading to the development of Green Algorithms, a simple online calculator that allows researchers to work out the carbon footprint of their computing work.

This is not the first time the research community has turned the spotlight on its own practices. Some in the community have already been asking questions about the impact of flying across the globe to present their findings at scientific conferences, for example.

Others have raised the issue of plastic and chemical waste and energy requirements from so-called 'wet labs'—that is, laboratories where experimental work takes place. Computer labs also have a significant

impact: equipment needs updating and replacing every few years at a minimum, while even data storage itself requires energy.

And then there is the computing work itself, of which there is a phenomenal amount these days. To give you an idea of its scale, in 2020, the now-concluded US-based XSEDE (the Extreme Science and Engineering Discovery Environment—a virtual system to allow scientists to share computing resources, data and expertise) alone saw researchers use 9 billion compute hours, or 24 million hours per day.

"For powerful calculations, either you need a lot of cores—you basically plug together a lot of computers and they all do the work for you—or you need a lot of memory. Either way, this takes energy."

Part of the problem, he says, is that computing can feel as if it comes at no cost. Research groups often have free access to high performance computing (HPC) facilities at their institution.

"When you first arrive as a Ph.D. student, you're like a kid in a candy store—you basically have unlimited computing power at your fingertips. It's brilliant and it enables great research, so it definitely shouldn't stop, but the problem is you just think it's free."

He gives the example of a process in machine learning called hyperparameter tuning, which involves testing different configurations of your model to work out which works best. "You never know when you've hit the maximum. It just keeps getting better until at some point, you say, 'Well I think I've made it as good as I can'."

"But let's say you're at the end of day and you think, 'Who knows, maybe I could just keep it running overnight. Maybe I'll get that extra half a percent of accuracy. It doesn't cost anything and no one's using the computers'. But actually, there is a cost—there's a carbon cost."

What he wants is not to limit research, but to cut computational waste, "to get people to think: 'Do I really need to do that? Probably not.'"

Lannelongue confesses that when they first launched Green Algorithms, he was skeptical as to whether people would use it. In the first few months, it was used only a few dozen times per month—mostly from users in his own lab, the thinks. But since then it has taken off and they get upwards of 300 users a week from around the world.

Even so, he recognizes that the tool may be "cumbersome" for some people to use, as it requires them to manually input their data. This is why they are working on Green Algorithms 4HPC (which is already available in beta form on GitHub), which uses data logs from the HPC centers to automatically calculate a project's carbon footprint.

"Lots of departments are interested in this as it's a painless way for scientists to implement it. A department can monitor the entire carbon footprint of the work being done there—it's not only individual scientists, but whole groups that can start saying, 'OK, let's monitor our carbon footprint and see what's our impact is month on month'."

He would like to see more transparency from research groups, and this is why his team now routinely calculate their carbon footprint using the Green Algorithms tool and include it at the end of their research papers.

It's easy to assume that as algorithms and the computers that power them become ever more efficient, the carbon footprint of computational science will fall, as it did in the biobank example. But this is not necessarily the case, due to the 'rebound effect."

"If you make a tool ten times more efficient, scientists will use it 100 times more," says Lannelongue.

"I mean, it's brilliant, that's how innovation works. But that's why we need to be able to track more precisely that actually what we do results in lower energy—otherwise, we may do all the hard work and then we realize that energy bills are as high as they've ever been."

Ultimately, he believes, there will have to be an element of personal responsibility when it comes to reducing the carbon footprint of computational science. "People think "I don't need to change how I'm acting, all the data centers will soon be powered by wind and solar." I would love it if it was true—it's just we know it isn't. We need to act now, and then if in the future, we arrive to a point where it doesn't matter anymore, then brilliant, we can resume our guilt-free lives."

And has his work changed how he himself works?

"Sadly, yes," he laughs. He was that proverbial kid in a candy story, running multiple analyses just because he could. Now, though, while he's still continuing with his research and still using machine learning, he is more mindful of the resources he uses. He will stop and ask himself if he really need that extra memory or to run his analysis one more time to be on the safe side. Instead, he will take time to figure out exactly what he needs before beginning the job.

"I know it's for the best," he says, before adding wistfully, "but I liked the innocence of not knowing. That was a nice time."

More information: Green Algorithms: www.green-algorithms.org/

Provided by University of Cambridge