The Economist

Sequencing the world

How to map the DNA of all known plants and animal species on Earth

IN NOVEMBER 2015, 23 of biology's bigwigs met up at the Smithsonian Institution, in Washington, DC, to plot a grandiose scheme. It had been 12 years since the publication of the complete genetic sequence of *Homo sapiens*. Other organisms' genomes had been deciphered in the intervening period but the projects doing so had a piecemeal feel to them. Some were predictable one-offs, such as chickens, honey bees and rice. Some were more ambitious, such as attempts to sample vertebrate, insect and arachnid biodiversity by looking at representatives of several thousand genera within these groups, but were advancing only slowly. What was needed, the committee concluded, was a project with the scale and sweep of the original Human Genome Project. Its goal, they decided, should be to gather DNA

sequences from specimens of all complex life on Earth. They decided to call it the Earth BioGenome Project (EBP).

At around the same time as this meeting, a Peruvian entrepreneur living in São Paulo, Brazil, was formulating an audacious plan of his own. Juan Carlos Castilla Rubio wanted to shift the economy of the Amazon basin away from industries such as mining, logging and ranching, and towards one based on exploiting the region's living organisms and the biological information they embody. At least twice in the past—with the businesses of rubber-tree plantations, and of blood-pressure drugs called ACE inhibitors, which are derived from snake venom—Amazonian organisms have helped create industries worth billions of dollars. Today's explosion of biological knowledge, Mr Castilla felt, portended many more such opportunities.

For the shift he had in mind to happen, though, he reasoned that both those who live in the Amazon basin and those who govern it would have to share in the profits of this putative new economy. And one part of ensuring this happened would be to devise a way to stop a repetition of what occurred with rubber and ACE inhibitors—namely, their appropriation by foreign firms, without royalties or tax revenues accruing to the locals.

Such thinking is not unique to Mr Castilla. An international agreement called the Nagoya protocol already gives legal rights to the country of origin of exploited biological material. What is unique, or at least unusual, about Mr Castilla's approach, though, is that he also understands how regulations intended to enforce such rights can get in the way of the research needed to turn knowledge into profit. To that end he has been putting his mind to the question of how to create an open library of the Amazon's biological data (particularly DNA sequences) in a way that can also track who does what with those data, and automatically distribute part of any commercial value that results from such activities to the country of origin. He calls his idea the Amazon Bank of Codes.

Now, under the auspices of the World Economic Forum's annual meeting at Davos, a Swiss ski resort, these two ideas have come together. On January 23rd it was announced that the EBP will help collect the data to be stored in the code bank. The forum, for its part, will drum up support for the venture among the world's panjandrums—and with luck some dosh as well.

Branching out

The EBP's stated goal is to sequence, within a decade, the genomes of all 1.5m known species of eukaryotes. These are organisms that have proper nuclei in their cells—namely plants, animals, fungi and a range of single-celled organisms called protists. (It will leave it to others to sequence bacteria and archaea, the groups of organisms without proper nuclei.) The plan is to use the first three years to decipher, in detail, the DNA of a member of each eukaryotic family. Families are the taxonomic group above the genus level (foxes, for example, belong to the genus *Vulpes* in the family Canidae) and the eukaryotes comprise roughly 9,300 of them. The subsequent three years would be devoted to creating rougher sequences of one species from each of the 150,000 or so eukaryotic genera. The remaining species would be sequenced, in less detail still, over the final four years of the project.

That is an ambitious timetable. The first part would require deciphering more than eight genomes a day; the second almost 140; the third, about 1,000. For comparison, the number of eukaryotic genomes sequenced so far is about 2,500. It is not, though, the amount of sequencing involved that is the daunting part of the task. That is simply a question of buying enough sequencing machines and hiring enough technicians to run them. Rather, what is likely to slow things down is the gathering of the samples to be sequenced.

For the sequencing, Harris Lewin, a genomicist at the University of California, Davis, who was one of the EBP's founding spirits, estimates that extracting decent-quality genetic data from a previously unexamined species will require between $40,000 and $60,000 for labour, reagents and amortised machine costs. The high-grade family-level part of the project will thus clock in at about $500m.

Big sequencing centres like BGI in China, the Rockefeller University's Genomic Resource Centre in America, and the Sanger Institute in Britain, as well as a host of smaller operations, are all eager for their share of this pot. For the later, cruder, stages of the project Complete Genomics, a Californian startup bought by BGI, thinks it can bring the cost of a rough-and-ready sequence down to $100. A hand-held sequencer made by Oxford Nanopore, a British company, may be able to match that and also make the technology portable.

The truly daunting part of the project is the task of assembling the necessary specimens. Some of them, perhaps 500,000 species, may come from botanical gardens, zoos or places like the Smithsonian (the herbarium of which boasts 5m items, representing around 300,000 species). The rest must be collected from the field. Dr Lewin hopes the project will spur innovation in collection and processing. This could involve technology both high (autonomous drones) and low (enlisting legions of sample-hunting citizen scientists). It does, though, sound like a multi-decade effort.

It is also an effort in danger of running into the Nagoya protocol. Permission will have to be sought from every government whose territory is sampled. That will be a bureaucratic nightmare. Indeed, John Kress of the Smithsonian, another of the EBP's founders, says many previous sequencing ventures have foundered on the rock of such permission. And that is why those running the EBP are so keen to recruit Mr Castilla and his code bank.

Banking on it

The idea of the code bank is to build a database of biological information using a blockchain. Though blockchains are best known as the technology that underpins bitcoin and other crypto-currencies, they have other uses. In particular, they can be employed to create "smart contracts" that monitor and execute themselves. To obtain access to Mr Castilla's code bank would mean entering into such a contract, which would track how the knowledge thus tapped was subsequently used. If such use was commercial, a payment would be transferred automatically to the designated owners of the downloaded data. Mr Castilla hopes for a proof-of-principle demonstration of his platform to be ready within a few months.

In theory, smart contracts of this sort would give governments wary of biopiracy peace of mind, while also encouraging people to experiment with the data. And genomic data are, in Mr Castilla's vision, just the start. He sees the Amazon Bank of Codes eventually encompassing all manner of biological compounds—snake venoms of the sort used to create ACE inhibitors, for example—or even behavioural characteristics like the congestion-free movement of army-ant colonies, which has inspired algorithms for co-ordinating fleets of self-driving cars. His eventual goal is to venture beyond the Amazon itself, and combine his planned repository with similar ones in other parts of the world, creating an Earth Bank of Codes.

Plenty needs to go right for this endeavour to succeed, concedes Dominic Waughray, who oversees public-private partnerships at the World Economic Forum. Those working on different species must agree common genome-quality standards. People need to be enticed to study hitherto neglected organisms. Countries which share biological resources (the Amazon basin, for example, is split between nine states) should ideally co-operate on common repositories. And governments must resist lobbying from vested interests in the extractive industries, keen to preserve access to land, minerals or timber, which Mr Castilla's scheme aims ultimately to curtail.

As to the money, that is the reason for the announcement at Davos. By splashing the tie-up between the EBP and the code bank in front of many of the world's richest people, those behind the two enterprises are not so discreetly waving their collecting tins. The EBP has already been promised $100m of the $500m required for its first phase. The code bank, meanwhile, has piqued the interest of the Brazilian and Peruvian governments.

For the participants, the rewards of success would differ. Dr Lewin, Dr Kress and their compadres would, if the EBP succeeds, be able to use the evolutionary connections between genomes to devise a definitive version of the tree of eukaryotic life. That would offer biologists what the periodic table offers chemists, namely a clear framework within which to operate. Mr Castilla, for his part, would have rewritten the rules of international trade by bringing the raw material of biotechnology into an orderly pattern of ownership. If, as many suspect, biology proves to be to future industries what physics and chemistry have been to industries past, that would be a feat of lasting value.