Genome Biology

COMMENT

Rising in the East

Gregory A Petsko*

The publication of the complete genome sequence of the giant panda, *Ailuropoda melanoleuca*, is a watershed moment for genomics, and not just because of the technology used. Before I explain, let me say a few words about that technology, because it is worth commenting on. The sequence, which was published in the 21 January issue of *Nature* (Li *et al*., *Nature* 2010 **463:**311-317, with a nice News and Views piece by Kim Worley and Richard Gibbs on pages 303-304; see also the minireview by Shaun Jackman and Inanç Birol in *Genome Biology* [http://genomebiology.com/2010/11/1/202]), was determined largely at the Beijing Genomics Institute (more on that later), which is not actually in Beijing, but never mind. It's an important genome, in part because the giant panda is a highly endangered species (only a few thousand are known to exist), in part because on the tree of life the panda sits between the human and the dog, but also because it is the first reported mammalian genome sequence to be determined using so-called 'next-generation' sequencing methods.

NGS methods, as they are widely called, use machines that produce very short sequences at very high speed. Compared with more traditional sequencing methodologies, they also cost much less per base pair. Some tests of NGS sequencing have been reported, but none involved the *de novo* assembly of an entire mammalian genome. Only the human genome sequence (2001-2003) and the mouse genome sequence (2002) have been completed with high redundancy and few gaps. Other large genomes, such as those of the dog, rat and monkey are basically drafts (approximately sevenfold coverage).

Genome sequencing is done in stages. After the genome is fragmented, the fragments are sequenced by machines that typically read 1,000 bases at a time. The reads are assembled by merging overlaps at the ends to form continuous sequence fragments (contigs). Traditional mammalian genome sequences contain contigs 100 kilobases long, so that often a complete gene is contained in one, providing reasonable accuracy. Contigs are then ordered into larger semi-continuous stretches, called scaffolds, using a variety of bioinformatics tools. A scaffold will contain a number of contigs separated by gaps. Larger gaps separate the scaffolds from each other. A good draft sequence of a mammalian genome will have perhaps a hundred scaffolds, or even fewer. Some people have wondered whether NGS machines, which typically read less than 100 bases at a time, would ever give comparable accuracy.

The Chinese team has answered that question, with a loud affirmative. The giant panda sequence has 73-fold average redundancy and a median contig length of 40 kilobases. Those are not typographical errors. The high redundancy offsets the assembly error problems that would compromise the quality of the sequence if the coverage were 10-fold or less. However, because of the short fragment read length, there are 3,805 scaffolds. That is not a typo either.

Illumina machines were used for most of the sequence, and the total cost of the sequencing itself has been estimated at less than US$1 million - at least 10 times less than that of a comparable genome done by, say, Sanger sequencing machines. While we are still a way off from the $1,000 human genome sequence, the $100,000 human genome sequence is essentially here.

To me, however, the real import of this paper lies in its geographic origin. The Beijing Genomics Institute (BGI) has its sequencing facility in Shenzhen, near the border with Hong Kong. It is a new but unremarkable building whose 11 floors of relatively plain decor belie the state-of-the-art science going on. It is the brainchild of Yang Huanming, a US-trained scientist who founded BGI in Beijing in 1999 as a private, non-profit research organization. Yang quickly got his fledgling institute involved in China's contribution to the Human Genome Project. Three years later, they made the cover of *Science* by winning the race for the sequence of the rice genome. Using Sanger sequencing machines, they completed that project in just 74 days. The giant panda sequence took 6 months.

In 2007, the BGI made two momentous decisions. They made a huge investment in NGS technology, focusing on the Illumina Solexa machine, and moved their headquarters to Shenzhen. The director is now a home-grown

*Correspondence: petsko@brandeis.edu
Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02454-9110, USA

genome biologist, Jun Wang, who is only 33 years old. He is the last of the 123 authors of the giant panda genome sequence paper.

The goal of the BGI-Shenzhen is to sequence informative genomes from all branches of the tree of life. In 2008, they completed the sequence of the genome from a Han Chinese individual, only the third published complete personal human sequence. Their intention is to sequence at least 100 more individuals within a few years, to explore the enormous ethnic variation in the Chinese population.

The BGI has about 30 Illumina Genome Analyzers, and can produce tens of Gigabases of sequence per day. The institute is exploring the use of other technologies, such as the SOLiD system developed by Life Technologies. It has a supercomputer center comprising 500 Linux nodes to do the assembly and analysis, and it needs it: the sequencing generates 10 terabytes of raw data every 24 hours. The computer center alone has an annual budget of about $9 million; the annual budget of the institute is $30 million.

I know what you're thinking: "I could do the same thing here if I had that kind of support from my government." The only problem with that is that you're mistaken. The BGI is a totally private organization, and doesn't derive a single cent of its budget from direct appropriations. It exists entirely on competitive contracts and grants, income from some spin-off companies, plus some private donations.

And this is just one institute of many in the exploding Chinese scientific landscape. I could instead have told you about the National Institute of Biological Sciences in Beijing (China's version of the legendary MRC Laboratory of Molecular Biology in Cambridge, UK), where scientists have successfully produced fertile mice from induced pluripotent stem cells. Or the 10 different Institutes of the Shanghai Institutes for Biological Sciences of the Chinese Academy of Sciences, China's version of the intramural research program of the National Institutes of Health.

But instead, let me tell you about the Kungming Institute of Botany, which is located in the capital of Yunnan Province, close to Tibet. In addition to doing first-rate botanical work, this institute contains the State Key Laboratory of Phytochemistry and Plant Resources of West China, which focuses on the search for bioactive molecules from natural sources. In this unique research facility, teams of chemists screen the vast biodiversity of the region and local ethnobotanical knowledge to discover compounds that can be developed into new drugs for unmet therapeutic needs and agrochemicals that do not harm the environment, and then synthesize them and make analogs of them. I toured the institute with an American synthetic organic chemist, and every other poster he would grab my arm, point to something, and say, "I've never seen anything like that [molecule or reaction] before!" In other words, the Kungming Institute of Botany, an institute you've never heard of in place a thousand miles off the beaten track, is one of the great centers of natural product chemistry in the world.

At a time when the United States is talking about three years of level government spending and an anti-intellectual movement I once thought was fading looks to be stronger than ever (more on that next month), China is beginning to tap the vast resource of its enormous population. Chinese culture has a strong work ethic, the government is pouring money into science, higher education is trying to emulate that of the United States, and living conditions have improved to the point that many foreign-trained Chinese scientists are going back home instead of remaining abroad permanently. Their research system, which is less hierarchical than that of Japan or Korea, is much better than either of those two countries in allowing young scientists, women as well as men, to be independent and advance. I could say something as well about the more gradual, but nonetheless impressive, rise of science in India, or its rapid rise in Singapore. The Far East, once a scientific backwater, is becoming a powerhouse.

In 1854 the American Indian Chief Seattle, considering whether to sign an unfavorable treaty, uttered these words:

But why should I mourn at the untimely fate of my people? Tribe follows tribe, and nation follows nation, like the waves of the sea. It is the order of nature, and regret is useless. Your time of decay may be distant, but it will surely come, for even the White Man whose God walked and talked with him as friend to friend, cannot be exempt from the common destiny.

I have always believed that not only was he right, but that sometime during my lifetime would be the time where future historians would draw their imaginary line and say, here marks the beginning of the fall of Western civilization and the rise of the East. I don't actually know if that's true, of course, but this much seems certain: Western scientific hegemony is fading fast. If you doubt it, just look at how many of the interesting and important papers in the leading journals are starting to come out of China, Korea and Singapore, and still come out of Japan. You could start with the 21 January issue of *Nature*. You can't miss it - it has a pair of giant pandas on the cover.

I feel sorry for those scientists who published other papers in that issue. They probably spent a fair amount of time and effort making illustrations they hoped would be selected for the cover. They never had a chance.

Published: 29 January 2010

doi:10.1186/gb-2010-11-1-102
Cite this article as: Petsko GA: Rising in the East. *Genome Biology* 2010, 11:102.