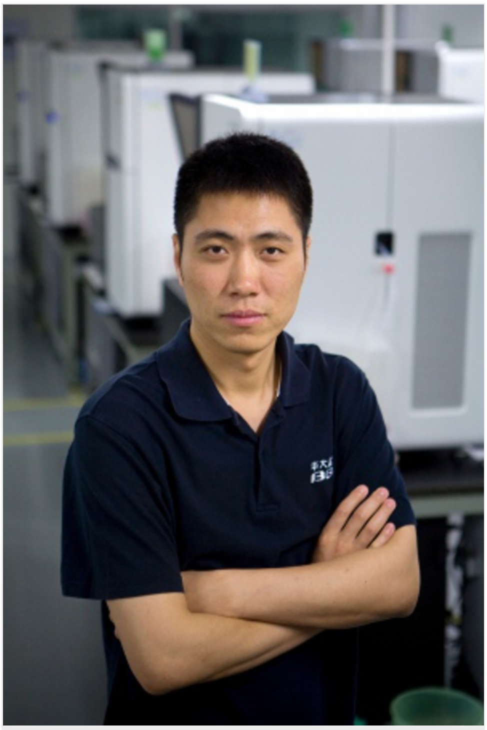
Exclusive: Genomics pioneer Jun Wang on his new AI venture

Visionary leader of China's BGI tells *Nature* why he is stepping down to build a health-monitoring system based on a million genomes.

David Cyranoski

28 July 2015

Jun Wang is one of China's most famous scientists. Since joining the genome-sequencing powerhouse BGI when it started up 16 years ago, he has participated in some of its biggest accomplishments. These include sequencing the first genome of an Asian person[1], the giant panda[2] and the human gut microbiome[3], as well as contributions to the Human Genome Project. Wang has led BGI since 2007 (when it stopped using the name Beijing Genomics Institute and moved its headquarters to Shenzhen). But on 17 July, the institute announced that he will give up that position to pursue research into artificial intelligence (AI).


BGI

Jun Wang.

In an exclusive interview, Wang tells *Nature* why he now plans to devote himself to a new "lifetime project" of creating an AI health-monitoring system that would identify relationships between individual human genomic data, physiological traits (phenotypes) and lifestyle choices in order to provide advice on healthier living and to predict, and prevent, disease. This interview has been edited and condensed for clarity.

What is the concept behind your AI project?
Basically, I am just trying to feed an AI system with masses of data. Then that system could learn to understand human health and human life better than we do. The AI will try to draw a formula for life. Life is digital, like a computer program — if you want to understand the results of the programming, how the genes lead to phenotypes, it is sufficiently complicated for you to need an AI system to figure out the rules.

> **"I am a risk-taker. I am betting all of my credibility on this."**

The AI system will basically consist of two components. The first is the big supercomputing platforms. We already have access to those through cloud computing and supercomputing centres. These will run or devise algorithms that look for relationships between genes, lifestyle and environmental factors, and predict phenotypes. The other thing is big data. We want to have data from one million individuals. And we want the data to be alive, in the sense that they can update their phenotype information at any time point. Other big computing companies, such as Google, could eventually do this, but we want to do it first. And we have the experience with the big data.

You need a million genomes for this?
To really understand complex traits that are determined by a lot of genes, such as height, you will need a million samples. We have 100,000

now, and it is just not enough. But I don't want to end at one million. Ten million, 100 million will be next. And it is not just the genome. We will get data at many levels — genomics, proteomics, metabolomics and lipidomics. As well as the other 'omics' data, we will include your life information, how much you exercise, environmental data. All this will be part of it. We will have one terabyte per person, so one exabyte for a million people.

Everything from the genotype to the phenotype can be digitized. To make AI work, we have to make it digital. I am more concerned with digital life in general than just genomics. It's not a million-genome project, it's a million-digitized-lives project.

What is the end goal?
The end goal is to develop an ecosystem. It will be a virtual village. When people 'stay' in the virtual village, it will advise them on how to live more healthily and for longer, including hints about what they should eat and what exercise they should be taking. It could tell them about warning signs before they get depressed; when they are stressed, it could tell them how to release the stress. All the advice will be based on various factors including genetic make-up and lifestyle**.** Individuals, doctors, researchers, pharma companies will all be part of it.

Even if the AI system can't find the answers, maybe the pharma companies will. The data will still be valuable.

How will you finance the project?
I am thinking that I may try to raise 10 billion yuan (US$1.6 billion) to make the first prototype [based on a million samples]. I don't really think about how to get the money. If you do the right thing — if you have a virtual village where people live longer and are healthier — the money will follow. I will test the water on several business models. Each one will have ways of trying to make people stick to the platform. I don't know which will win. Maybe we will have to recruit 10 million people to get a 1-million-person complete dataset. But who cares. Let's just do it.

Is this project the only reason that you stepped down from leading BGI? There are rumours that you were forced out because you failed to bring the company to an IPO [initial public offering].
People speculated a lot on the reasons, and it was quite annoying. The plans for an IPO are on the right track. One of the main reasons I resigned as chief executive is that the current BGI model is increasingly a diagnostics or research-service model, and it is quite stable now. I want to do more than that. But BGI will still be a very important part of the new initiative.

What do you say to the concern that this sounds too ambitious?
I don't know. I just do it. I've been through that a lot. People have told me I had crazy ideas before — the rice genome project or the short-read assembly of the panda genome [an influential demonstration of next-generation sequencing that used small fragments, or 'short reads', of DNA]. But you know, it turned out pretty good.

This is my lifetime project. Before I retire, I want to make this happen. I'm 39, so I hope to make the whole thing happen in the next 20 years. I'm a little bit nervous, but also excited, because I know I'm doing the right thing. I am a risk-taker. I am betting all of my credibility on this.

Nature | doi:10.1038/nature.2015.18091

References

1. Qin, J. *et al*. *Nature* **464**, 59–65 (2010).
2. Li, R. *et al*. *Nature* **463**, 311–317 (2010).
3. Wang, J. *et al*. *Nature* **456**, 60–65 (2008).

Nature ISSN 0028-0836 EISSN 1476-4687

SPRINGER NATURE

© 2019 Macmillan Publishers Limited, part of Springer Nature. All Rights Reserved.
partner of AGORA, HINARI, OARE, INASP, CrossRef and COUNTER