CRACKING CANCER WITH BIG DATA AND SUPERCOMPUTERS

Cancers are essentially caused by gene mutations, but finding the culprits requires examining millions of base pairs. For more than a decade, the HUMAN GENOME CENTER at THE UNIVERSITY OF TOKYO and its supercomputer SHIROKANE have been integral to international efforts to speed up the search.

When the SHIROKANE supercomputer first came online at the University of Tokyo's Human Genome Center (HGC) in 1997, the assembly and analysis of entire human genomes was still a dream. It would be three years before the multi-billion-dollar project to sequence the human genome would complete its first rough draft.

Today, individual genomes can be analysed with a standard desktop PC, and SHIROKANE's 'great-grandson' is taking on far more ambitious tasks — helping scientists and clinicians around the world explore many thousands of human genome sequences to unearth the causes and vulnerabilities of cancer. "With our current computer architecture and support applications, we can handle millions of cancer genome sequence analysis jobs per month," says Satoru Miyano, director of the HGC (opposite).

Supercomputer tackles cancer

The fourth-generation SHIROKANE system (opposite, bottom) evolved in response to the rapidly growing needs of the oncology research community. In particular, Miyano cites the launch in 2008 of the International Cancer Genome Consortium (ICGC) — a multinational research effort profiling whole-genome and gene expression for dozens of tumour types — as an important driver. At present, the ICGC has 89 cancer projects underway.

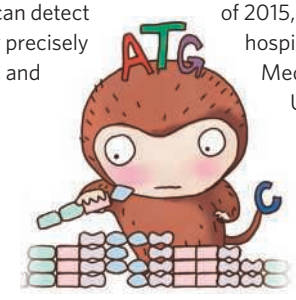
For any given cancer, the cause may lie in any number of different genetic mutations. Some of these are relatively common or easy to spot — for example, a sequence change that disrupts a protein that keeps cell growth in check — while others are subtle and exceedingly rare.

IT EXAMINED THE ENTIRE GENOMES OF BOTH TUMOUR AND HEALTHY TISSUES FROM 2,800 PATIENTS

To identify the latter rare changes, many genomes must be compared in parallel. "To identify mutations that occur in the rarest 1% of cancers, we need to analyse roughly 1,700 tumour samples, and we also need to analyse healthy germ-line genomes for comparison," says Miyano. "That means almost 4,000 samples must be analysed." Accordingly, SHIROKANE is capable of extremely fast and efficient computer processing, and also has the capacity to store and rapidly access large quantities of data during the analysis process.

SHIROKANE makes use of a parallel computing system known as a 'grid engine', which can efficiently manage many projects simultaneously.

The system also benefits from a number of software innovations. In collaboration with Seishi Ogawa's team at Kyoto University, HGC researchers have developed a powerful data analysis pipeline software called Genomon (mascot below) for identifying mutations in tumour genomes, including structural changes that are difficult to pinpoint. Many cancers acquire big insertions, deletions or duplications of DNA sequences that can dramatically change gene expression and cellular function. The most widely used genome sequencing technologies identify these changes by generating many short stretches of sequence data that must be reassembled into a bigger sequence. But these are easily confused by structural arrangements that broadly alter the genomic map. Genomon overcomes this hurdle, according to Miyano. "It can detect mutations very precisely and accurately, and also detects structural variations with an accuracy of more than 90%," he says.

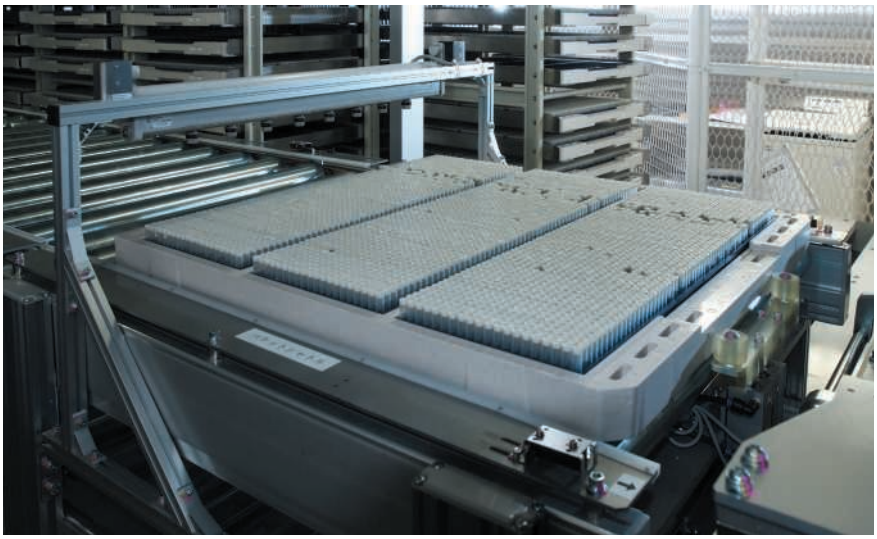


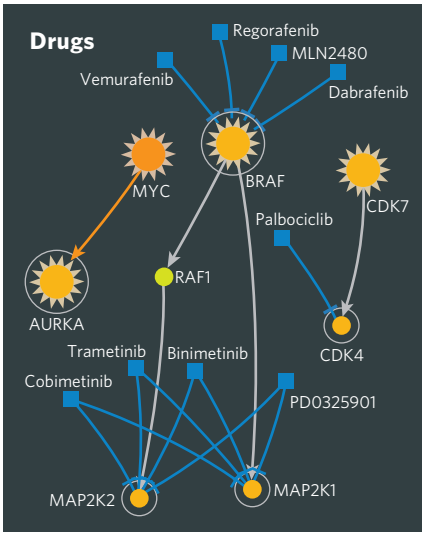
Leaps forward in blood and liver cancers

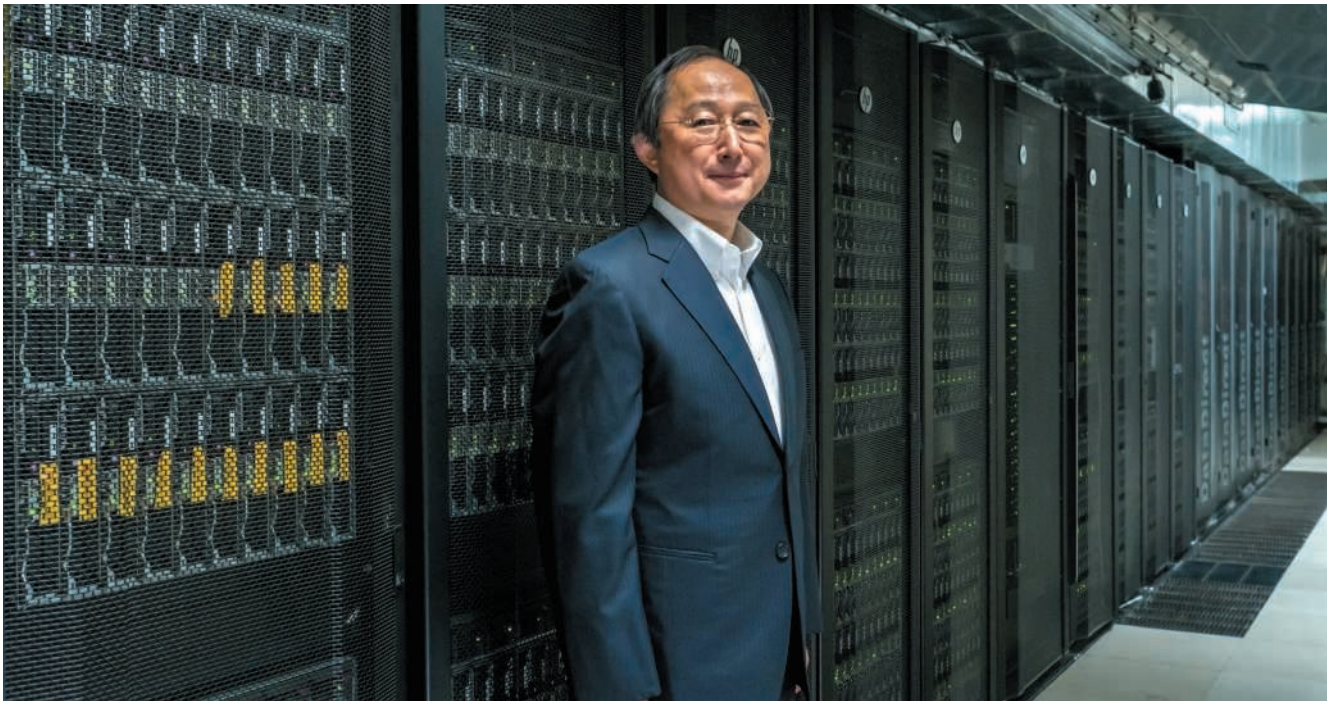
Cancer researchers have already racked up some high-profile successes with SHIROKANE. For example, the HGC team made 1 petabyte of storage (a million gigabytes) available on the supercomputer for an ICGC-led survey of numerous cancer types, which examined the entire genomes of the tumour tissues and healthy tissues from 2,800 patients, data that was made available to many researchers around the world.

The Japan team has also used SHIROKANE to drill down into cancers that were until recently poorly understood. In 2016, researchers analysed the whole genome sequences of 300 liver cancer patients and comprehensively identified recurrent mutations, structural variations and virus integrations. They found mutations related to liver carcinogenesis recurrently occurred in non-coding regions, underlining the drug discovery potential of this type of analysis.

In addition, since the summer of 2015, the HGC and the hospital of The Institute of Medical Science at The University of Tokyo (IMSUT) have been working closely with artificial intelligence systems. After six months of 'training' these systems, it is







now enabling clinicians working in haematology and oncology at IMSUT's hospital to draw reliably on insights from genome sequencing and analysis. "In just ten minutes, we can get an analysis of mutational data, with some recommendations that associate gene mutations with approved drugs or drugs that are now in clinical trials (pictured, top right)," says Miyano. "The final recommendation is made by clinicians, of course, but the progress is very good and the work of our research hospital these days will not move forward without input from artificial intelligence systems."

Genome analysis is rapidly becoming common medical practice, and Miyano and his colleagues have been working with the Japanese government to help expand national investment in this space. "Funding and implementation of cancer genomic medicine will rapidly increase from next year," he says. On the other hand, this will also mean a huge influx of additional data to process, increasing the number of samples stored in a biobank (top left) at The University of Tokyo and creating an urgent demand for more computational storage in the near future. Miyano anticipates that SHIROKANE's capacity will be inadequate within five years.

Plans are already underway for the next SHIROKANE upgrade, but Miyano hopes for even broader investment in supercomputing for genomics. "Japan should have several systems," he says. It is, not only essential to Japan's position as a research superpower, he says, but also to its ability to expand its presence in global multilateral efforts like the ICGC, which have become cornerstones of modern biological research. "Cancer genomics cannot be done by a single country," says Miyano. "Many countries must collaborate." ■

THE UNIVERSITY OF TOKYO

Web: hgc.jp/english
Email: info@hgc.jp