

# Making Personalized Genomic Medicine a Reality: Understanding the Mechanisms of *Cancer* through Simulation Analysis of Genetic Information

The highest performing supercomputer in the nation to serve the field of life sciences



## The Task

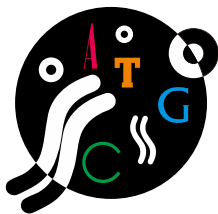
- To achieve personalized medical care through whole-genome analysis
- To ensure frequent access to large amounts of files

## The Solution

- Supercomputers equipped with the Intel® Xeon® processor E5 family
- Intel® Enterprise Edition for Lustre\*: a high-speed distributed file system

## Initial Results

- *Cancer Systems* research activities on the rise, combining Systems Biology and Cancer Pathology
- Extensive Genome Sequencing Research that combines next-generation sequencing with supercomputers



## Human Genome Center

Human Genome Center, The Institute of Medical Science, The University of Tokyo  
4-6-1 Shirokanedai, Minato-ku, Tokyo

**Established:** 1991

### Objectives:

Established as a research center for medical and biological research.

Conducting cutting-edge basic research in the Laboratory of Functional Analysis *in silico*, Genome Database, Genome Technology, etc.

Providing research materials for Japan's researchers as well as technological guidance and education, while working on database construction.

[www.hgc.jp/](http://www.hgc.jp/)

## Genomic information is the blueprint of life – its analysis aims to achieve personal medical care that suits the individual patient

Cancer affects half of the Japanese population, and it causes death in 1 of 3 people. To overcome the illness that is impacting the lives of people so harshly, scientists are now trying new ways to look at the issue. The new strategy lies in Big Data, using supercomputers to identify the genes lurking behind cancer. At the Human Genome Center, The Institute of Medical Science, The University of Tokyo (HGC, IMSUT), the Laboratory of DNA Information Analysis that carries out *cancer systems* research, combining

systems biology and cancer pathology, obtains large amounts of genetic data (genome information) from patients to carry out simulation analysis, using supercomputers equipped with Intel® Xeon® processor E5 family, in an effort to thoroughly understand the mechanisms of cancer.

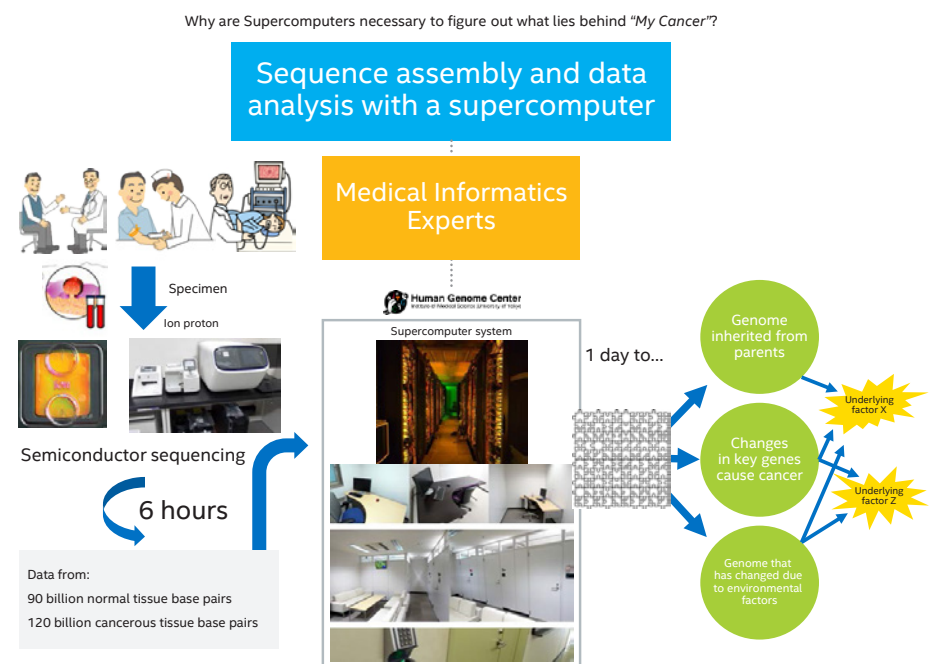
The genome is the complete set of genetic information that is passed on from parents to children; it is the blueprint for the life of an organism. It is made up of a substance called DNA. Its basic building blocks are the four types of nucleotide bases, referred to with the letters A, G, T and C. The DNA of the

human genome is made up of 3 billion of these base pairs – what their sequence is and how the genome works was clarified through the *Human Genome Project*, an international collaborative relationship that started in 1990. In recent years, we see a development of personalized medical care that analyzes the human genome and offers treatments suited to the individual patient. “Considering the aging population of Japan, it is an important task to create a society where people live longer and healthier, supporting each other. Making individualized genomic medical care and prevention a reality will lead to the answer to this issue,” says Prof. Satoru Miyano at HGC.

When speaking of genomic medicine, the case of Hollywood actress Angelina Jolie is still fresh in everyone’s memory. In May 2013, she announced that she had undergone a double mastectomy and reconstructive surgery as a breast cancer prevention measure. The deci-

sion was made because, as a result of genomic testing, she had learned that she carries a mutated gene (BRCA1) that can lead to an increased incidence of breast cancer and ovarian cancer. The chance of her developing breast cancer was determined to be 87%.

This shows how risks can be assessed by examining the genes that are highly relevant to cancer, and early treatment becomes a possibility. Even if cancer has already developed, side effects of anti-cancer medications can be predicted, and the most effective cancer-fighting treatments can be administered. In the future that we are striving to achieve, appropriate treatment methods can be selected based on anyone’s own DNA information while avoiding side effects, so we could say, “For your abnormal genes, anti-cancer medications X and Y would work, but since X has a high chance of causing side effects in your case, let’s use Y.”





## The Intel® Xeon® processor family speeds up supercomputers, creating a good balance between floating-point and integer operations

### Using mathematics and super-computers, systems biology and cancer pathology come together in *Cancer Systems* research to pioneer innovative cancer care

To begin with, cancer is an illness where abnormalities of several genes within the cell's genome cause normal cells to turn into malignant cells (cancer cells). The cancer cells develop this way, then start to disturb the function of the organ by forming a *tumor*, multiplying according to instructions coming from within themselves. Next, the cell division continues indefinitely, ignoring instructions from the outside to stop multiplying, and the cancer invades healthy cells and adjacent tissue, spreading and thus expanding.

Again, it is the *gene network* that lies at the heart of this abnormal cancer system. The phenomenon of life is based on gene networks; they are everywhere, and have a significant impact on cancer. Mutations that occur in cancer cells are the result of changes that happen in the gene networks, caused by genetic factors that are inherited from one's parents (DNA), genetic changes that accumulate in

one's cells as the person ages (cancer genome), and environmental factors, etc., leading to mutations in genome modifications (epigenome). "The key to cancer research," says Prof. Miyano, "is to pinpoint the abnormalities in the system that will determine such things as the malignancy of the cancer, responsiveness to treatment, ease of side effect development, etc. However, cancer cells are complex—continuously changing their shape, they infiltrate their surroundings and keep repeating their transformations. Additionally, cancer stem cells, normal cells and immune inflammatory cells 'mate,' developing resistance to anti-cancer medications, and this makes things especially difficult," explains the professor.

Traditionally, the focus of cancer research has been *oncology*, *cancer pathology*, and *cancer biology*, but at the HGC, a completely new *cancer systems* research is envisioned: *systems biology*, *bioinformatics* and *statistical*

"With the improvement of sequencers that collect DNA information, it became possible to acquire whole-genome information at a low cost and within a short timeframe. In order for personalized genome-based medical care to become a reality in the future, it is essential to have supercomputers equipped with the high-performance Intel® Xeon® processor E5 family."

*Human Genome Center,  
The Institute of Medical Science,  
The University of Tokyo*

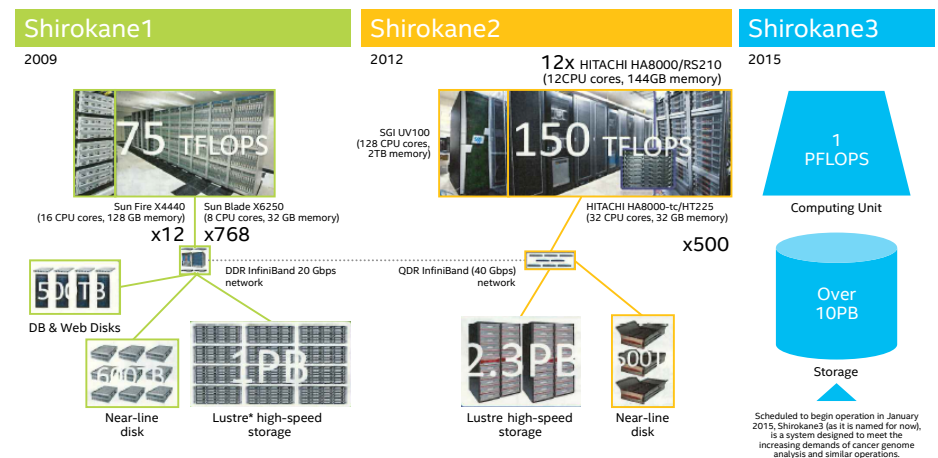
**Professor Satoru Miyano  
Ph.D.**

*genetics* are being combined, using mathematics and supercomputers. Since 2010, the HGC has been doing research focused on two things: *fully embracing a systems biology approach in explaining cancer pathology, and based on life systems data, pioneering and clinically implementing innovative cancer treatment options*. Currently, they are working on systematically deciphering cancer, analyzing extensive gene networks, using the *Supercomputers at the HGC* and the shared-use supercomputer named 'K computer,' which is installed at RIKEN Advanced Institute for Computational Science.

To be able to analyze the human genome with supercomputers, a *sequencer* is a must-have: a machine that reads the DNA data consisting of three billion base pairs. Previous sequencers were extremely expensive to purchase, and on top of that, expensive reagents were needed in proportion to the sequenced amount, so acquiring DNA information from the human genome cost considerable money and time. However, in recent years the development of cutting-edge sequencers (such as the semiconductor sequencer) is moving forward, so the per-person cost of acquiring DNA information is decreasing:

it now costs around 100,000 Japanese yen for a few hours, so it is becoming possible to acquire one's full DNA information. Sequencing technologies of the future, combining nanopores and CMOS chips, are expected to bring these numbers down to 100,000 Japanese yen for 1 hour, and eventually to the range of 10,000 Japanese yen for 1 hour.

Currently at the HGC, scientists are diagnosing the risks associated with cancer, using these next-generation sequencers to read the whole-genome information of cancer patients, then supercomputers analyze the *genome inherited from the parents, changes in key genes that cause cancer, and the genome that has changed due to environmental factors*. "Since clinical transcription/interpretation of whole-genome information has become a reality," says Prof. Miyano, "the objective of sequencing has shifted from the scientific research level to the clinical level of practical medical care, so from here on we should be moving toward medical care that is based on the personal genome. In the past, if we had a colorectal cancer patient, for example, only the mutations in genes specific to the large intestine were analyzed. Other genes—for example, those behind lung



cancer—were ignored. However, once we can work with whole-genome information, the genes causing the cancer could be analyzed in their complexity, and we could select treatment options that provide total care.”

**While the Intel® Xeon® processor contributes to the high-speed computations of genome information, Lustre\*, the high I/O performance distributed file system, supports access to large amounts of data**

Cancer systems research at Prof. Miyano's laboratory proceeds with the help of an important system platform: Supercomputer SHIROKANE. Currently, as of March, 2014, the computing capacity of SHIROKANE is 225 TFlops (peak performance), with 22,000 total cores, and as a specialty supercomputer for the life sciences, it boasts the highest computing performance in the nation. SHIROKANE consists of two supercomputers, Shirokane1 (75 TFlops) and Shirokane2 (150 TFlops), and both are configured with two types of distributed memory servers. One of Shirokane1's distributed memory servers is equipped with an Intel Xeon E5 family processor. When asked about the benefits of using the Intel Xeon processor E5 family, Prof. Miyano said, “We feel the advantage is that the performance balance is really good between integer computing that works with the information acquired from the sequencer and floating-point operations that analyze gene networks; computations are done with high speed in both cases.”

As a common storage solution for saving all files, a distributed file system,

Lustre\*, is used. With a total of 3PB disk space provided, that is, 1 PB for Shirokane1 and 2 PB for Shirokane2, access is available to large amounts of data, anywhere from a few hundred gigabytes to several tens of terabytes.

And who are the ones using SHIROKANE? They include all researchers associated with genome research; it is available for outside academic institutions as well as the private sector. Currently, there are approximately 600 user IDs in use, about half of which belong to members of the HGC, and the other half to outside academic institutions and private sector users combined. Various genome research tools are also ready for use: for example, with *SiGN-BN* and *SiGN-SSM*, static and dynamic gene networks can be estimated from gene expression data; *Cell Illustrator Online* combines the functions of *in vivo* pathway drawing and simulation.

About the operational conditions of SHIROKANE, Dr. Ayumu Saito, technical specialist at the Laboratory of DNA Information Analysis, HGC, IMSUT, said the following:

“We are always operating at above 80% capacity; the number of executed jobs reaches an average of about 2 million monthly. This number is about twenty times that of the University of Tokyo Information Technology Center, a facility used jointly by eight national universities. The huge amount of input and output files per job is one thing that makes the supercomputers of the genomic medicine laboratories unique, but also, it's not uncommon within a single job set that thousands, even tens of thousands, of files are created and

“The objective of sequencing has shifted from the scientific research level to the clinical level of practical medical care, so from here on we should be moving toward medical care that is based on the personal genome.”

**Professor Satoru Miyano  
Ph.D.**

"Since the importance of storage systems will be growing more than ever before, we are looking forward to high management capacity with the Intel® Enterprise Edition for Lustre\* in the next generation of supercomputers."

Laboratory of DNA  
Information Analysis  
Human Genome Center,  
The Institute of Medical Science,  
The University of Tokyo

**Dr. Ayumu Saito**  
**Technical Specialist**

referred to, and as many as several tens of thousands of file accesses are happening every second. With adequate computing resources and a job scheduling system, SHIROKANE guarantees the ability to smoothly compute enormous jobs."

Regarding the benefits of using Lustre, Prof. Miyano added the following, talking about Lustre's high I/O performance for accessing large amounts of files.

"Since Lustre has such a high I/O performance, we don't have to worry about great drops in its performance, even if large amounts of files, even several tens of thousands, come together in one job set — simultaneous, parallel sequence analysis is possible. When we first introduced Lustre in 2009, we kept running into obstacles, but by now we have reached a stable state, and as a result, Lustre has been able to help us greatly with high-speed access to large numbers of files."

### **Ahead of the world: Discovering the genes responsible for MDS (Myelodysplastic Syndrome)**

The HGC's supercomputer SHIROKANE has contributed to numerous findings marking the history of cancer research. One of these findings is the discovery of the genes behind Myelodysplastic Syndrome (MDS). MDS is a form of *blood cancer* that prevents the body from making blood in the marrow, and this can lead to such things as acute myeloid leukemia. It often occurs in elderly patients, affecting around 5,000 people yearly within Japan. Its cause was previously unknown.

At the HGC, in collaboration with Prof. Seishi Ogawa, who worked at the Tokyo University Hospital at the time, research was launched using large-scale, next-generation sequencing and supercomputers to pinpoint the genes causing MDS. They were identified in merely a year (between July 2010 and September 2011), and the results were published in the journal *Nature*.

Specifically, exon (information for producing protein, separated from the total DNA information) was extracted from 29 cancer specimens, then, using the 75 TFlops (6000 core) capacity of *Shirokane1*, they were searching for the mutant genes. Eventually, they determined 268 locations where changes could be observed, and pointed out that the cause of MDS lies in abnormalities of the genes involved in the maturation pathway of the RNA (RNA splicing) that is copied from the DNA. This finding was validated using an additional 582 specimens.

"While research institutions around the world were working on trying to understand MDS, our team was successful in getting to the bottom of the unresolved problem ahead of everyone else. There is no question that supercomputers have contributed to this achievement." (Prof. Miyano)

Additionally, SHIROKANE is used in research on lung cancer, the number one cause of death of all cancers in Japan. At the HGC, in collaboration with Professor Takashi Takahashi of Nagoya University Graduate School of Medicine, the genes that are responsible for the recurrence of lung cancer are being analyzed using pulmonary adenocarcinoma gene

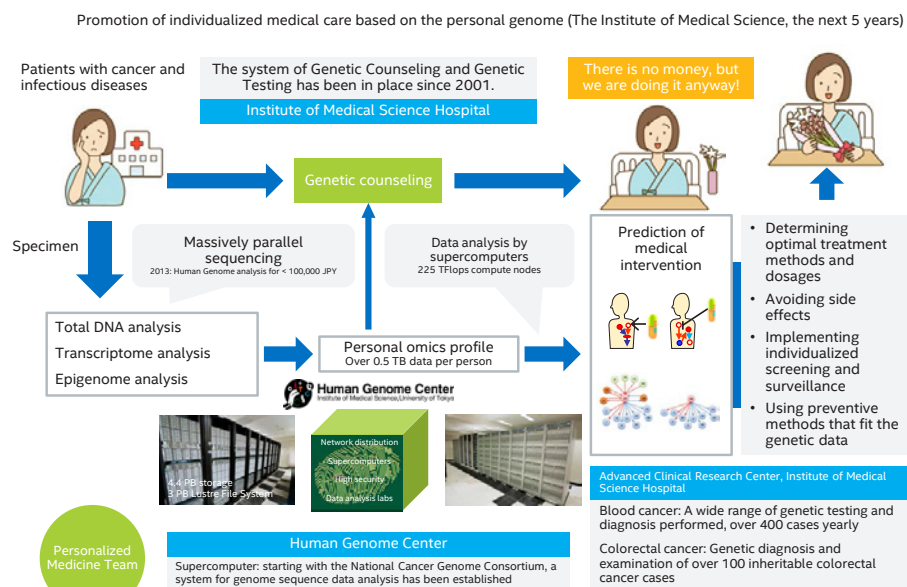


specimens from 124 patients. Simulations of genetic relevance and recurrence after 5 years helped identify 14 small gene networks and the hub of genes that affect whether lung cancer recurs and causes death.

## Enhancing the performance and functions of supercomputers contributes to the promotion of individualized medical care, the future of medicine worldwide

At the HGC, the plan for the future is to continuously enhance the performance and functions of SHIROKANE. The rapid development of sequencers and the decreasing per-individual cost of human genome analysis explains this policy. Once it becomes possible to acquire not only genome information, but also, for example, the epigenome (environmental changes), or transcriptome (mRNA) with next-generation sequencers, there is no question that there is going to be an even higher reliance on high-performance supercomputers and extensive storage.

"If we consider the rapid increase in analyzed data, we can say that by 2015, the computing power should increase to 1 PFlops and the storage capacity to 128 PB. By 2020, ideally the performance should improve to 1 EXFlops, which equals 1,000 times 1 PFlops, and storage to 2 EXB. If we start to seriously apply genome analysis in clinical pathology, that should lead to acquiring genomes from 10 million, 50 million people. For that, it will be necessary to introduce inexpensive disks, and we will have to figure out policies such as saving the data distributed in the country. Additionally,



as the amount of analyzed data grows explosively, it will become urgent to train experts in data analysis and information systems," said Prof. Miyano.

As storage capacity increases, it is also important to ensure the ease and availability of management systems. "Since the importance of storage systems will be growing more than ever before," says Dr. Saito, "we are looking forward to high management capacity with the Intel® Enterprise Edition for Lustre in the next generation of supercomputers."

Clinical sequencing based on whole-genome analysis that started in America is now expanding to the entire world; in Japan, the University of Tokyo Institute of Medical Science Hospital, the HGC and the IMSUT Advanced Clinical Research Center (ACRC) are working on collaborative projects. The HGC will be contributing to the

promotion of personal medical care by always having a system for genome sequence analysis ready, through continued enhancement of the performance and functions of the Supercomputers.

For more information about the Intel® Xeon® processor E5 family, please go to the following websites:

[www.intel.co.jp/xeonE5/](http://www.intel.co.jp/xeonE5/)

[www.intel.com/lustre](http://www.intel.com/lustre)

[www.intel.com/healthcare/bigdata](http://www.intel.com/healthcare/bigdata)



**Prof. Satoru Miyano**  
**Ph.D.**

Human Genome  
Center, The Institute  
of Medical Science,  
The University  
of Tokyo



**Dr. Ayumu Saito**  
**Technical Specialist**

Laboratory of DNA  
Information Analysis  
Human Genome  
Center, The Institute  
of Medical Science,  
The University  
of Tokyo



This document and the information given are for the convenience of Intel's customer base and are provided "AS IS" WITH NO WARRANTIES WHATSOEVER, EXPRESS OR IMPLIED, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NON-INFRINGEMENT OF INTELLECTUAL PROPERTY RIGHTS. Receipt or possession of this document does not grant any license to any of the intellectual property described, displayed, or contained herein. Intel® products are not intended for use in medical, lifesaving, life-sustaining, critical control, or safety systems, or in nuclear facility applications.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to [www.intel.com/performance](http://www.intel.com/performance).

Intel does not control or audit the design or implementation of third-party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase.

Copyright © 2014, Intel Corporation. All rights reserved. Intel, the Intel logo, and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.