

## Customer Success Story

# Genomic research evolves with HPC

## Customer

Sanger Institute

## Industry

Life Sciences

## Challenges

- Genetic sequencing machines produce 120 terabytes of raw data per week that need to be processed for analysis
- Research on genomes of numerous species generates 100,000s of processes per week, that must be scheduled and completed as efficiently as possible

## Solution

Platform LSF

## Results

- Researchers can make rapid advances in science by quickly comparing similar genomic structures
- Ability to perform massive, regular updates to the genome browser database to incorporate the latest research advances
- Excellent support from Platform allows the Institute to deal with the unique issues that any business faces in running a heterogeneous HPC infrastructure

“The genomics industry as a whole has realized that what we need to do would be ludicrously expensive without cluster computing.”

Tim Cutts  
Platform LSF Administrator,  
Sanger Institute

The Wellcome Trust Sanger Institute is a genome research centre set up in 1992 by the Wellcome Trust and the Medical Research Council in order to further the knowledge of genomes. It plays a substantial role in the sequencing and interpretation of the human genome to underpin research on human biology and disease.

The Sanger Institute realized that the research they were doing would be “ludicrously expensive on single machines,” according to Tim Cutts, Platform LSF Administrator. “Traditional supercomputers were not set up to deal with the sorts of problems we were facing.”

## Twelve Clusters, Hundreds of Nodes

The Institute has a total of twelve clusters, eight of which, including the three largest, run Platform LSF. It is a multi-vendor, heterogeneous Linux environment, with dual and quad core IBM and HP machines, as well as several SGI Altix machines in the various clusters. The architecture varies from 32-bit systems to some 64-bit Opteron and 64-bit Itanium systems. This cluster environment is a perfect fit for Platform LSF which was designed to work with the most complex IT environments. The largest cluster has 710 nodes and is used for general purpose research by the Institute’s researchers. Platform LSF is a High Performance Computing (HPC) management software solution that intelligently schedules parallel and serial workloads.

## Data Inputs Grow by Orders of Magnitude

HPC is important to the Sanger Institute than ever before. The Institute recently purchased 30 new genome sequencing machines, each of which produces two orders of magnitude more data than the previous generation of sequencers. The raw data is sent to a Platform LSF cluster for processing.

The machines run third-party “sequencing pipeline” software. “They generate absolutely stupendous quantities of data. A single one of these instruments has the sequencing capacity of our entire institute three years ago,” says Cutts. “A sequencing reaction run on the machine takes around three days,” he continues.

“Each run generates around two terabytes of data, so that’s four terabytes per machine per week.” Multiply that by 30 machines and the Institute is processing 120 terabytes of raw data a week. This makes it much more vital to have a robust workload management solution available to ensure that the enormous amount of information is processed as efficiently and as quickly as possible. The sequencing is carried out on the 128-node Platform LSF cluster. “This cluster handles the huge quantity of data that come off of the new DNA sequencing machines,” says Cutts.

## Accelerating Genomic Research

The Sanger Institute was a key player in the Human Genome Project, delivering almost one third of the work involved. Since the mid-1990s, Platform LSF has improved workload management, researchers’ efficiency and time-to-results for the Human Genome Project by allowing the Institute to run up to half a million sequence matching jobs a day.

“The Human Genome Project used enormous, scalable compute power to market the data available throughout the project. The Project was as much an exercise in IT and systems needs as in lab science,” says Phil Butcher, Head of IT at the Sanger Institute. According to Butcher the Institute was able to finish sequencing the human genome two years ahead of schedule partly because of the investments in flexible systems and software.

## Finding Causes and Cures for Disease

A smaller 100-node cluster, also powered by Platform LSF, handles requests submitted by people externally to the Ensembl Genome Browser. Ensembl, a joint project with the European Bioinformatics Institute, amasses genome information for 20 species including chimpanzees, mice, cows, dogs and humans. This site allows researchers to compare their own gene sequences with information available from the Sanger Institute. Users can paste their information into the web site and ask it to “find something in the database that looks like it,” says Cutts.

“Let’s say you are interested in muscular dystrophy,” says Cutts. “You would be able to compare a particular gene that you know has been associated with muscular dystrophy with the equivalent gene in 20 other organisms. And then use the information to construct for example a mouse model of the disease.”

Platform LSF is crucial to this process as the job of updating the genetic annotation database is a computationally intensive process. “Since the data is being constantly updated by laboratories all over the world, we recalculate the entire Ensembl database from scratch every two months on the large cluster.”

## Support is the Key

As for many of Platform’s customers, a high standard of support is critical to the Sanger Institute. “The standard of support we have had has been excellent,” he says. “No software is bug free, so we need a company that is actually willing to say, ‘Oh, yeah, okay, we’ll fix that as quickly as possible’.”

The Wellcome Trust Sanger Institute does not endorse commercial products.



Platform Computing is the leader in cluster, grid and cloud management software - serving more than 2,000 of the world’s most demanding organizations since 1992. Our workload and resource management solutions deliver IT responsiveness and lower costs for enterprise and HPC applications. Platform has strategic relationships with Cray, Dell™, HP, IBM®, Intel®, Microsoft®, Red Hat®, Fujitsu and SAS®. Visit [www.platform.com](http://www.platform.com).

### World Headquarters

Platform Computing Corporation  
3760 14th Avenue  
Markham, Ontario  
Canada L3R 3T7  
Tel: +1 905 948 8448  
Fax: +1 905 948 9975  
Toll-free Tel: 1 877 528 3676  
[info@platform.com](mailto:info@platform.com)

### Sales - Headquarters

Toll-free Tel: 1 877 710 4477  
Tel: +1 905 948 8448

### North America

New York: +1 212 888 6270  
San Jose: +1 408 392 4900

### Europe

Bramley: +44 (0) 1256 883756  
London: +44 (0) 20 3206 1470  
Paris: +33 (0) 1 41 10 09 20  
Düsseldorf: +49 2102 61039 0

### Asia-Pacific

Beijing: +86 10 82276000  
Xi’an: +86 029 87607400  
Tokyo: +81(0)3 6302 2901  
Singapore: +65 6307 6590