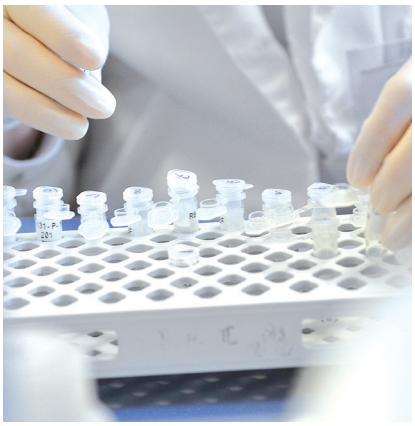
CASE STUDY

Sequencing and Supercomputers

Intel® Xeon® Processor E5 Family
Big Data Analytics
Healthcare & Life Sciences



Spain's National Center of Genomic Analysis (CNAG) is involved in large-scale sequencing projects in areas as diverse as cancer genetics, rare diseases, host-pathogen interactions, preservation of endangered species, evolutionary studies, and the improvement of agriculturally useful species. Its mission is to deliver research and results that help make citizens' lives better. CNAG has worked with Intel and Atos to build its latest analytics platform, which helps drive new insights faster, with wide-ranging applications.

Challenges

- **Billions of bases.** CNAG sequences over 800 billion genomic bases every day and needs to conduct quick, accurate analysis to process large volumes of data as efficiently as possible
- **Needles in haystacks.** Identifying the small number of variations that drive breakthrough insights can be time-consuming and complex

Solutions

- **Strong combination.** Atos Big Data & Security service line developed a tailor-made compute cluster, powered by the Intel® Xeon® processor E5 family, to conduct in-depth high-performance data analytics (HPDA) on genome sequences

Impact

- **Healthy results.** CNAG plans to offer more in-depth and comprehensive genomic analyses to healthcare organizations to improve treatments
- **Leading the way.** The new platform maintains CNAG's position as a leading genomic research facility, enabling it to achieve higher output while driving cost savings

CNAG conducts large-scale DNA sequencing and analysis, ensuring Spain's competitiveness in the strategic field of genomics

"We're certainly handling big data, but what we're really after is big information: the ability to identify the valuable insight from the sequences in front of us. To do this well, we need good data, good analytics and good tools. We quality control these elements carefully."

Ivo Gut,
Director,
National Center of Genomic Analysis

Covering a lot of bases

CNAG works with scientists from universities, hospitals, research centers and biotechnological and pharmaceutical companies across Europe. It is a key player in many international research initiatives, including the International Cancer Genome Consortium (ICGC), the International Rare Disease Research Consortium (IRDiRC) and the International Human Epigenome Consortium (IHEC). CNAG supports around 120 researchers, who conduct around 300 projects each year. This results in it sequencing around 800 Gigabases per day, the equivalent of sequencing eight full human genomes at 30-fold coverage (the coverage required for reliable analysis), making it one of the largest capacity sequencing facilities in Europe.

Genome sequencing is a complex and demanding task. "Each genome is made up of over 3 billion bases. It's not simply a case of starting at the first base in the genome and then identifying them all in order until you have the complete string," says Ivo Gut, director of CNAG. "All genomes are 99.9 percent identical, but it's the variations in that remaining 0.1 percent that we're looking for. To find them, we need to break each genome down into strings of a hundred or so bases, sequence the short strings and rebuild it. It's like doing a jigsaw puzzle with billions of pieces."

Once found, these variations in the genome bases can be used to identify certain characteristics of an organism – anything from eye color, to disease susceptibility, to tolerance to medication. These genetic characteristics, known as phenotypes, can then be applied in developing new medications and clinical practices, growing more disease-resistant crops, or even preserving endangered species. For example, CNAG has supported the genome sequencing of the Iberian lynx, a highly endangered big cat now found only in southern Spain, which is particularly susceptible to infection. With only around 200 left in the wild, it is critical that scientists are able to identify the genetic traits that make these animals so prone to illness, in order to better protect them. The center was also the first organization in the world to sequence the olive genome, using a 1,000-year-old tree near Madrid, and has participated in the genome sequencing of many other agricultural crops including tomatoes, peaches, melons, almonds and fungi.

New fields of study

Building on the success of its first five years in operation, CNAG has now opened two new departments, specializing in biomedical genomics and population genomics. The Biomedical Genomics Department takes data generated within CNAG and combines it with other data sources to gain new insights. For example, it has investigated a particular mutation related to cystic fibrosis which impacts the severity with which the disease manifests in patients. "With access to a larger data set, we can study these differences in more detail than has been previously possible, and spot modifiers in the milder cases that could

be targeted in the more severe cases to help deliver better treatment to people with the condition," says Gut.

Meanwhile, the Population Genomics Department is seeking opportunities to analyze the genomes of a given population or group of people. This is important when assessing the potential impact or effectiveness of new clinical treatments, and how this might vary depending on the population to which they are administered. "Say we're looking at a new diabetes medication," says Gut. "We might test it on a group of Finnish diabetics versus a group of Spanish control patients by adding a correcting factor to remove the population effects, and we'll be able to say for certain whether any impact of the drug is influenced by a phenotype within either group for increased or reduced susceptibility to the disease."

CNAG also works on a number of projects funded by the European Commission – for example, collaborating with other institutions around the region to research lung disease, and conducting a study of 20,000 breast cancer patients.

Data, analytics and tools

The work that CNAG does has huge potential across a wide range of disciplines. The main challenge facing Gut and his colleagues in reaching this potential is scale. In every genome of 3.2 billion bases, there are many base variations that are potentially responsible for diseases, making them challenging and time-consuming to spot. "At the moment we're able to identify some of these variations, but the aim is to be able to locate and accurately predict the effect of every one of them in whichever genome we're looking at," says Gut. "This will enable us to make big-picture predictions about, say, how a specific type of cancer is likely to respond to a given combination of medication. Once we get to this stage – and it's not far off – we'll be able to put the findings to use in a clinical environment and have an impact on how many diseases are treated."

With such complex analytics, a powerful computing platform is essential. Analysis of each genomic sequence may take hundreds of CPU hours, which limits the number of projects that can be carried out, and makes repeating an analysis difficult. "This is fine for one-off research projects, but we want to be able to offer sequence analysis on an industrial scale, with the ability to repeat one analysis a number of times if required," says Gut. "We're certainly handling big data now – and it's growing all the time – but what we're really after is big information: the ability to identify the really valuable insight from the sequences in front of us. To do this well, we need good data, good analytics and good tools. We quality control all these elements very carefully."

When CNAG built its sequencing and analytics environment, it issued a tender and identified Atos as its solution provider of choice. Gut explains: "It was not only about increasing the sequencing capacity by procuring new equipment, but designing from scratch the appropriate computer infrastructure, with the assistance of a technological leader with expertise in the field of genomics. Choosing a flexible core infrastructure enabling limitless growth, together with genomics projects, was also essential."

Five years in, it refreshed its supercomputing platform to ensure the scalability and performance it needs to support this continued up-scaling were in place. Specialized sequencers feed the genomic sequence data to the analytics platform, which is run on a cluster of 52 bullx R4222E2* servers powered by Intel Xeon processors E5 family.

Lessons Learned

Supercomputing platforms like the one at CNAG can generate amazing results. However, for them to scale well as data volumes and the demand for insight increase, it's essential to have the flexibility to grow. CNAG worked with Intel and Atos to design and implement a sophisticated analytics platform that can grow seamlessly over time.

"We've found that working with Atos as our single point of contact for the whole solution has made managing it very simple for us, and enabled us to focus on our research rather than technology administration," says Simon Heath, chief technology officer, CNAG. "We've also been very impressed by the latest Intel Xeon processors that we have deployed, which by our estimates have contributed to a ten-fold increase in analytics software performance[1]."

Diving deeper

With the latest platform in place, CNAG is looking forward to providing more granular and varied insights to its end users. For instance, hospitals will be able to use its genomic research in different ways to combat different types of illness. "When treating a rare disease, a physician may never have seen another case of it in their whole career," explains Gut. "But we can provide them with phenotype information that could help them make an identification and diagnosis faster. In many cases, this can lead to quick and effective treatment – such as putting the patient on a course of vitamin supplements. On the other hand, in the oncology unit, diagnosis is not so much of an issue. Here, it's more about identifying the right medication or combination of treatments to have the greatest success against a given type of tumor – and this is where the phenotype data will come in use here."

Furthermore, CNAG will soon reinforce its capacity with a few new sequencers, each of them producing 3.2TBases per week that corresponds to 32 full human genomes at 30x. This new capacity will go hand in hand with the expansion of its supercomputing systems, with which CNAG will expand not only its sequencing capacity, but also the scope of its research. This in turn will help CNAG maintain its position as a center of international reference in genomic research, ensuring Spain's leading position in the strategic field of genomics.

www.cnag.crg.eu

Find the solution that's right for your organization. View **success stories from your peers**, learn more about **server products for business** and check out the **IT Center**, Intel's resource for the IT Industry.

[1] 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 http://www.intel.com/performance.

0615/JNW/RLC/XX/PDF

332686-001EN