

National Center for Genome Resources expands gene sequencing and research operations using Kognitio WX₂ analytical database

WX₂ is allowing us to perform research that just wasn't possible before. Kognitio's technology, willingness to make our installation a success, and understanding of the genomic space, have all helped to make this implementation of immense benefit to us. As genomic sequencing continues to expand, we are confident that WX₂ will fit perfectly into our plans.

Dr. Ernest Retzel, Program Leader, NCGR

The Company

Based in Santa Fe, New Mexico, The National Center for Genome Resources (NCGR) is a non-profit research institute dedicated to improving human health and nutrition through genome sequencing and analysis of the genomic data produced. As a bioinformatics, software, technology and service provider for genome sequencing projects, NCGR offers advanced DNA sequencing services and bioinformatics and software solutions to academic and commercial researchers. NCGR is one of the world's 10 largest gene sequencing centers, and has been in operation since 1994.

NCGR's Genome Center provides large-scale sequencing and analysis services to collaborators worldwide, both as not-for-profit work and also as a service to corporations in agriculture, healthcare and other industries. These corporate activities are vital to maintain an appropriate level of funding for the Center's research, and so need to be completed quickly and to the customer's satisfaction. Current projects the Center is working on include investigations into genetic causes of schizophrenia, the genetic changes that cause cancer, the nature of virulent plant pathogens and the difference between reactions to pneumonia. The Center also provides a web-based query service to customers via its Alpheus software, allowing scientists, bioinformatics professionals and interested parties to investigate data that NCGR has generated, recorded and stored.

The Challenge

It is not simply genome sequencing that provides NCGR with its reputation, status and income: the investigation of a wide range of genomic data it undertakes and the insights generated by the NCGR staff are vital assets. As a result, NCGR does not just store the data it generates before passing it on to third parties. NCGR's research arm provides indispensable benefits by performing bioinformatics on the data created by sequencing. NCGR can provide clients, whether businesses or multinational academic projects, with detailed analysis of sequencing results.

For example, when investigating crop pathogens, NCGR can study the genomic data of both the crop and the pathogen itself to see what genetic data links those plants that are infected or immune and how that corresponds to the data from the pathogen itself. Agricultural corporations might want to discover which genes are specifically linked with plant breeding and growth, to improve crop yields. When investigating diseases such as pneumonia, researchers can study how the virus and the host respond genetically to one another and so predict the behavior of diseases and possible cures.

However, offering this level of genome sequencing and bioinformatics calls for a massive amount of genomic data to be produced, stored and studied. Furthermore, when investigating the data, it is not enough to simply query a subset and extrapolate from those numbers the likely results for the entire set. To ensure accuracy, vital both for corporate customers and for NCGR's not-for-profit work, every single item of data has to be recorded and queried. Without this, results could be inaccurate, either rendering an experiment worthless, or even worse, leading researchers to incorrect conclusions. Querying massive amounts of genomic data to this level of thoroughness could take days, a timescale that was unacceptable to corporate customers who demanded fast results, and far from satisfactory for other work that could be helping to save lives.

To that end, NCGR researchers were aware that their incumbent technology environment could not keep pace with the demands of the organization. Not only were queries taking up to three days to produce results: balancing the needs of corporate customers and NCGR's core work was becoming increasingly difficult, as a simple lack of space on NCGR's database meant there wasn't enough room to process all incoming queries. The growing backlog this produced meant that it was also impossible for NCGR to grow further: taking on more experiments, or performing existing ones in more detail, meant that the overload on the system would have reached a critical point.

The Challenge

With the volumes of data being produced from its production operations in genome sequencing increasing at a near-exponential rate and with clients requiring rapid turnaround on their projects, NCGR realized that it would soon reach the limits of its current technology

The Solution

NCGR deployed Kognitio's WX₂ purpose-built analytical database to load and query multiple terabytes of data, ensuring that its bioinformatics and analytical operations could keep pace with current and future increases in the amounts of data produced by its gene sequencing operations

The ROI

By using Kognitio WX₂, NCGR has been able to accelerate its bioinformatics operations to keep pace with demand, which complements their world-class genome sequencing center and processes. WX₂ has also allowed NCGR to expand the scope of its investigations and experiments, by allowing entire sets of genomic data to be queried in extremely short times.



“WX₂ gives us the ability to investigate 100% of all the data available without limiting our queries which doesn't just help accuracy: it also reveals new avenues for investigation that we might otherwise miss.”

Dr. Ernest Retzel, Program Leader, NCGR

“Essentially, our bioinformatics operation was lagging behind; it could not progress because its sequencing operation could not match the level of output needed,” said Dr Ernest Retzel, Program Leader, NCGR. “Our customers range from large corporations looking to improve their yields, to charitable bodies investigating schizophrenia, to university departments investigating different strains of tuberculosis. Ideally, we would want to serve everyone, but we already had a backlog of data and faced delaying several research projects if we were unable to improve the throughput of our queries; without an upgraded system, our work would simply grind to a halt.”

The Solution: Kognitio WX₂

After investigating other options, NCGR was approached by Kognitio with the offer of using a WX₂ analytical database to store and investigate its sequencing data. Kognitio went over the background specifications of NCGR's needs and delivered a proof of concept. Throughout this process, NCGR was impressed by Kognitio's commitment to the area of genomics and agreed to a three-month evaluation of the technology. Kognitio provided technical support and best-practice consultancy throughout the set-up and trial period and demonstrated a clear understanding of NCGR's work and goals. Since that initial implementation, NCGR has subsequently increased its use of WX₂ and its license to allow it to store and query unlimited amounts of data.

WX₂ uses Massively Parallel Processing to provide a powerful data warehouse. The solution uses the brute force of a large number of machines running in parallel with one another to quickly integrate and query data without the need to structure it in sets or provide specific rules for queries. By performing queries “in-memory”, rather than relying on accessing data from disk storage, WX₂ is able to quickly investigate a large amount of data near-instantaneously. Thanks to Massively Parallel Processing and in-memory performance, WX₂ can query data as soon as it is entered into a database, with no need to wait for further data management processing which can delay end-user access by hours, if not days.

For NCGR's researchers, this means that genomic data can be stored and queried as soon as it is produced by the sequencing machines. WX₂ is powerful enough that huge sets of data can be investigated in their entirety within a few minutes, rather than the days previously taken. NCGR can now provide a data warehouse that serves every one of its ongoing experiments, from discovering the genes that cause adverse responses to vaccinations, to creating a full database on the genomic data of every type of legume.

The Benefits: NCGR's scope for further experiments is massively increased

The first effect of WX₂'s increased speed and capability has been to eliminate the backlog of genomic data projects that was previously troubling NCGR. With query times reduced from days to minutes, NCGR has been able to guarantee quick results for its corporate customers for investigations into subjects such as drug effectiveness, while still having plenty of time and space spare to perform experiments into subjects such as preventing rare genetic diseases by studying gene mutations. This is despite the amounts of data involved growing by more than 100 times since WX₂'s initial installation at NCGR, thanks to a greatly increased capacity.

As well as the increased speed, NCGR's IT team has reported that it can run more sophisticated and complicated queries on WX₂ than were previously possible. With no need to tailor data entry and queries to fit with the needs of a structured database, NCGR's researchers have far more freedom to devise experiments and queries without feeling constrained by time pressure or the risk of an overly prescriptive set of results. Data can also be investigated in parallel, so as to reveal separate trends and again reduce the time taken by what would previously have been two separate investigations. For example, when investigating crop fertility, rather than simply comparing the genomic data of two samples to detect anomalies and see how the fertile crop differs from others, NCGR can also investigate the positioning of those different genetic markers, and see how they affect their neighbors and which of those different genes is most likely to be the primary cause of greater or reduced fertility. All this can be done in far less time than a simple query would have taken previously.

“18 months ago, we had a single sequencing instrument: now we have eight,” said Dr. Ernest Retzel. “With each machine producing 10 to 20 times as much data as before, it would have been impossible for us to expand our investigations in this way with our old database solution. The ability to investigate 100 percent of all the data available without limiting our queries to cope with a less-than-capable database doesn't just help accuracy: it also reveals new avenues for investigation that we might otherwise miss. Essentially, WX₂ is allowing us to perform research that just wasn't possible before. Kognitio's technology, willingness to make our installation a success, and understanding of the genomic space, have all helped to make this implementation of immense benefit to us. As genomic sequencing continues to expand, we are confident that WX₂ will fit perfectly into our plans.”

The Future

NCGR's use of WX₂ has enabled it to grow as one of the first large-scale bioinformatics implementations in the world. It has also allowed the development and implementation of NCGR's Alpheus software, permitting NCGR's customers to investigate their data remotely using a web-based service. By providing this sequence querying as a service, NCGR can free up its own researchers for other tasks such as disease investigations and fast-response projects and also provide an expanding new service to corporate customers who have the freedom to tailor their own investigations to their desires.

NCGR is set to continue expanding its operations to deal with the ever-increasing amount of genetic information being uncovered: researchers expect data requirements to grow more than tenfold in the next year. In the short term, this will manifest as the ability to perform more experiments, in more detail, on more data, providing more vital information on diseases such as schizophrenia and cancer, as well as more corporate opportunities. In the long term, this will allow NCGR to take on tasks with a much shorter turn-around time; an example might be investigating sudden and virulent disease outbreaks.

“Technology has advanced to the point where research such as the original Human Genome Project, which took hundreds of millions of dollars and several years to complete, could now be replicated in a few months and for a fraction of the original price,” said Dr. Ernest Retzel. “This is thanks to tools such as WX₂, which are at the core of this increase in power and reduction in times and costs. In the future, we will see our research capability soar and our research times plummet still further, and WX₂ will continue to be at the heart of that.”