breakthrough thinking GLS GenoLogics

Enabling the **USC Epigenome Center** to keep up with Constantly Changing Next Generation Sequencing Technology and Rising Throughput

The USC Epigenome Center occupies an entire floor of the newly constructed Harlyne Norris Research Tower of the USC Norris Comprehensive Cancer Center.

"We required a centralized lab and data management system that would allow our facility to seamlessly track samples across many projects and automate data capture for integrated analysis."

Dr. Peter. W. Laird, Principal Investigator and Director USC Epigenome Center

Case Study Summary: GenoLogics is enabling the USC Epigenome Center to accelerate their research, improve communications with customers and collaborators, while their lab scales with quickly changing genomics technologies and throughput.

A leader in cutting-edge epigenomics research, and a member of The Cancer Genome Atlas, the USC Epigenome Center

The University of Southern California established the first dedicated Epigenome Center in the nation in July 2007, to conduct high throughput epigenomics and leverage USC's long standing expertise in the field of epigenetics. The USC Epigenome Center conducts genome-wide epigenetic analyses and technology development for their cutting-edge epigenomic and population-based genomic research.

In addition to the Center's own epigenetics research, they also provide core services to the USC community using Illumina BeadArray assays, TaqMan quantitative PCR, and most recently Illumina next generation DNA sequencing. Regarding throughput, the sequencing center has three Illumina GAIIx sequencers and currently conducts between three and five sequencing runs and performs upwards of 2,000 Infinium BeadArray assays per week.

The USC Epigenome Center was recently awarded funding to participate in The Cancer Genome Atlas (TCGA) consortium to collect epigenomic data from all major human cancers over the next five years, in collaboration with Dr. Steve Baylin at Johns Hopkins University. The center will be the sole epigenetic data production facility for TCGA, which is a long term, comprehensive initiative funded by the NCI and the National Human Genome Research Institute to generate an understanding of the molecular basis of cancer. During the pilot phase of the project the Center used Illumina's Infinium Methylation BeadArray assay to profile glioblastomas (1), and will continue to use this

technology in the upcoming phase. However, the center has recently optimized whole-genome bisulfite sequencing for human tumor samples using GAIIx technology, and will incorporate this much higher resolution technique over the next five years as sequencing costs drop.

In search of a robust solution to manage an increasing amount of NGS data, collaborator communications and increase operational efficiency

Like many other next generation sequencing labs, the Epigenome Center is scrambling to keep up with continually evolving protocols and updates to Illumina's analysis pipeline software, which have lead to a 15-fold increase in generated sequencing data since the Center ventured into next generation sequencing. The writing was on the wall and the Bioinformatics team realized early on, after receiving its first Genome Analyzer, that they would require a centralized data and lab management informatics system.

The system they needed had to be capable of handling the following:

- The Center has to manage samples and data from a number of different labs and a diverse group of participating scientists: external collaborating investigators who submit samples to the lab, internal investigators within the Epigenome Center and the Center's lab technicians and data analysts.
- The Center is receiving an increasing number of requests to provide information on which data analyses were performed, how the analyses were performed and what methods were used to prepare samples prior to analysis. As a core service center, they needed sophisticated tools to communicate this information to customers easily.

The Build vs. Buy Decision

When did the USC Epigenome Center select the Geneus system and what criteria did you use, Dr. Berman?

"We did our own analysis. We looked at other products on the market, and what it would take to implement such a system. Part of it just came down to bioinformatics resources. As you know the whole field is just… so short on bioinformaticians - it's hard to find good ones and it's expensive. To be able to have ours just spending their time working on the science aspects of the analysis, and not having to spend all their time working on sample tracking when we could buy a product to do this, was a major factor.

Some of the products out there, some of the widely-used third party products, are really focused on the analysis side and don't concentrate so much on the sample processing side where GenoLogics is particularly strong. Especially in terms of its flexibility and the level to which you can customize it for Illumina platforms, this was a big factor. The decision was made easier as the Geneus software offered very tight out-of-the-box integrations to the Illumina Genome Analyzer and other Illumina BeadArray platforms in our lab.

As a bleeding-edge epigenetics center, we didn't think we could ever make do with commercial analysis tools, we want to be leaders in epigenomic data analysis here.

So, we felt the best decision for our center would be to purchase something that was really very strong on the lab side, but also provided us with ability to connect to our own custom analysis tools, which are written in a variety of programs including PERL, Java, R, etc. That's basically how we made our decision. Also, we needed to be up and running quickly and figured that even if we allocated resources to developing such a product ourselves, it would take somewhere in the two year timeframe.

The other issue was that we would probably have to allocate programmer resources whenever we needed to change the system and improve it, so we'd have to do continuous maintenance, whereas we can get that level of support from GenoLogics, and they are very good about working very closely with us to meet our needs."

- Another requirement for this system was robust quality assurance capabilities. The system would need to be able to track all facets of their experiments, including capturing information for increasingly sophisticated experimental and data quality metrics.
- The moving target of their DNA technologies required a system built to handle rapidly evolving lab workflows and analysis workflows, and ideally one with rich built-in support for the specific Illumina technologies in their lab.
- The USC Epigenome Center operates with a philosophy of tight integration between the lab side and data analysis side, so they wanted a system which could provide a tight level of integration between sample tracking and their custom built bioinformatic data analysis pipelines.

After defining the top level requirements for a lab and data management system the real question that rose to the top became - whether to use their own software development team to build it themselves, or whether to buy a third party solution.

Dr. Benjamin Berman, Senior Research Associate at the USC Epigenome Center, leads the bioinformatics and data management effort for next generation sequencing at the Center.

In conversation with Dr. Berman, he stated, *"As a bleeding-edge epigenetics center, we didn't think we could ever make do with commercial analysis tools, we want to be leaders in epigenomic data analysis here. The decision we made was to free up our in-house software developers for specialized analysis and data processing problems, and to purchase the lab information management software system. We chose the Geneus software solution as a great solution with support for a rich level of sample tracking, lab inventory tracking, automatic report generation, QC/QA capabilities, collaborator data access, and most importantly for its ability to adapt to our changing lab environment."*

"The decision we made was to free up our in-house software developers for specialized analysis and data processing problems, and to purchase the lab information management software system."

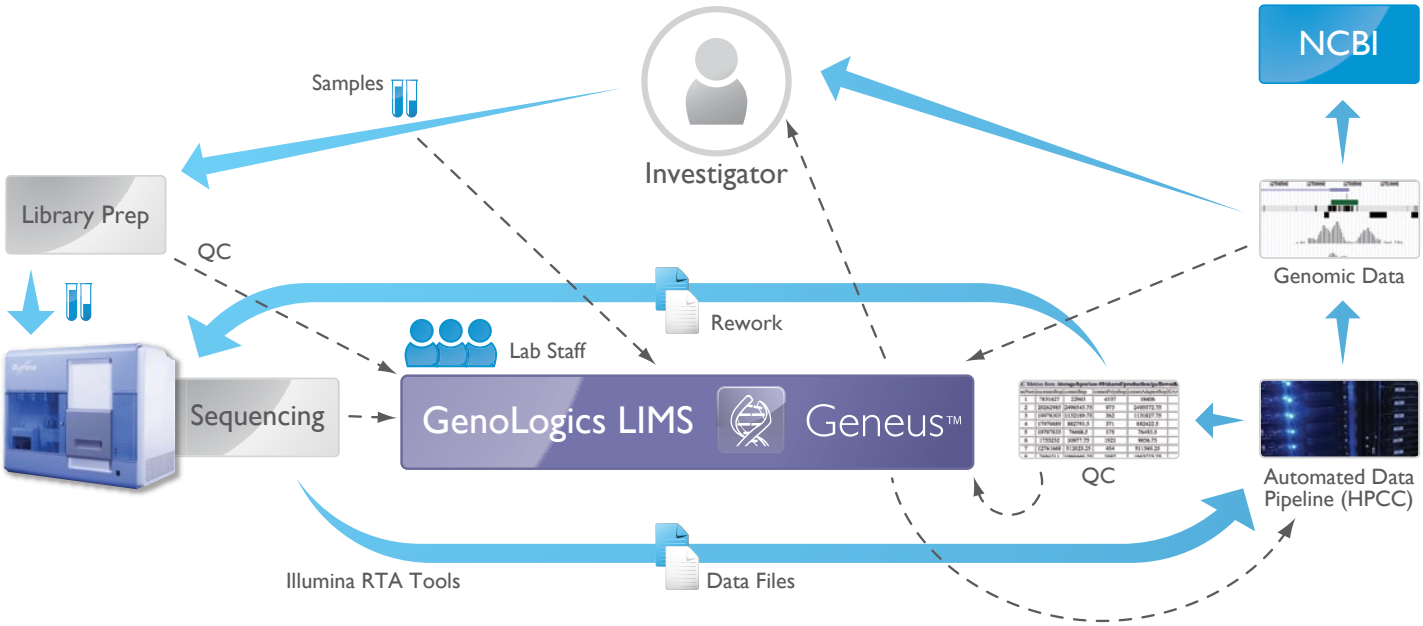
Dr. Benjamin Berman, Senior Research Associate, Bioinformatics team, USC Epigenome Center

He added, *"As a bonus, the decision was made easier as this particular software offered very tight built-in integration to the Illumina Genome Analyzer and Illumina BeadArray platforms in our lab."*

An in-depth look at the custom Genome Analyzer data analysis workflow at the Epigenome Center

Now for an in-depth look at how the Epigenome Center is deploying the Geneus system in their lab to handle their data management and increase technician access to sophisticated data analyses through the use of flexible application programming interfaces (API's) in Geneus. The workflow starts with an investigator (let's assume the Center is providing their services to an external investigator) who first interacts with Geneus using a built-in web-based collaboration tool called LabLink to upload an Excel-based sample data sheet(s) and initiate a new project. The lab's system administrators can define the template for this and can include any custom fields for information they want to track in addition to standard fields. One example might be whether a DNA sample has been pre-fragmented, and if so, what technique and conditions were used. All sample information is then transferred into the Geneus lab and data management system.

The Geneus LIMS centralizes and manages all the pertinent information that flows through the USC lab. This diagram shows Geneus managing and tracking samples, the QC process, library prep, the sequencing run, and kicking off their automated data analysis pipeline.



The information from the PI then enters the lab environment. All steps involved in library preparation from the submitted samples are tracked, including quality control information along with quality control files generated during library preparation. The lab then uses an interface in Geneus, customized by Dr. Berman's team, to select and initiate analyses for next generation sequencing experiments. This effectively allows any bioinformatics researcher within the Center to initiate a range of data analysis workflows from a simplified Graphical User Interface (GUI) in Geneus without requiring programming skills. Initiating these workflows within the complete experimental context has the benefit of ensuring that the correct analyses are carried out.

Dr. Berman discusses below some of the specifics of their custom analysis workflow integration to Geneus.

> *"It starts within Geneus… for example, a flow cell "alignment and genomic coverage" analysis that the lab staff or project manager would run immediately after prepping the sequencing run is initiated from a custom view within the system. If the results are found to be poor, we can re-run the analysis with different parameters from within Geneus and each of these analyses will remain persistent and traceable within the Geneus. The results are tied via a unique identifier back to the analysis that was run and [parameters] that were used. The flexible infrastructure of the system allows us to easily create as many analysis interfaces as we need, each using different input parameters. We can also run analyses that combine sequencing lanes on different flowcells, which is an essential feature."*

Dr. Berman's data analysis wizards make use of a functionality in Geneus called Automated Informatics (AI) that enables execution of any Linux executable and it provides hooks for the executed program to call back to Geneus to request additional information needed for the analyses. Once the user clicks "start", AI initiates the workflow by calling a Linux PERL script written by the Epigenome Center, and this scripts queries Geneus' database using an abstract object-level REST API (Application Programming Interface). REST is an

Internet client server communication standard in which clients request information from a server using the web protocol HTTP and some standard encoding language. In Geneus' implementation of REST API's, results are returned as structured data in an XML format in response to queries from USC's data analysis pipeline.

Here are three examples of different kinds of information the USC Epigenome Center's analysis pipeline scripts request from their Geneus server:

- The first type of information the launcher script requests are parameters entered on the data analysis initiation form, such as problematic sequencing cycles or flow cell tiles to be manually excluded from the analysis, type and lane identifier of a control sample on the flow cell, and the basic data analysis workflow(s) to perform.
- The second type of information is sample specific, such as the source organism in order to pick the reference genome for alignments.
- The third type of information requested from Geneus are details about library construction or lab processing steps – for instance which adapter oligos were used (important for sequence filtering) or whether the DNA was modified by bisulfite treatment (necessary to do the correct kind of genome alignment). They also frequently use the estimated DNA fragmentation size in order to infer paired end positions from single-ended sequencing experiments, a key step in ChIP-seq and RNA-seq analysis.

At this stage, the launcher script passes information it has received from the Geneus server to the Epigenome Center's automated bioinformatics pipeline which uses a system called Pegasus to automatically run individual steps on the Center's compute cluster located at the USC High Performance Computing Center. Pegasus is an open source workflow automation system developed by the cluster computing group at the USC Information Sciences Institute. It provides bioinformaticians with a framework to define complex analytical workflows in an abstract way, and then enables their execution. Pegasus workflow execution plans can be deployed using various platforms such as the USC High Performance Computing Center, Amazon EC2, or the National Science Foundation's TeraGrid.

During our interview, Dr. Berman mentioned that work has begun, in collaboration with Dr. Stan Nelson at the sequencing center at UCLA, to create generalized Pegasus workflows and processing steps with the intention of distributing them as open source tools for the rest of the research community.

Dr. Berman wrapped up by mentioning that one of the best things about the Geneus system is that it frees up a lot of time for their software developers. Their developers no longer have to spend time writing sample tracking or lab management GUIs, Geneus does this within its core functionality, so they can focus on writing more tools for generating better quality control reports, data normalization methods, and biological analyses.

Results: Increased Operational Efficiency and Improved Collaborator Communications while Freeing up Valuable Resources and Saving Time

The Epigenome Center is starting to realize a number of operational efficiencies, and time and cost savings, in a number of key areas due to the implementation of the Geneus data and lab management system, from GenoLogics.

The following quotes are from Dr. Benjamin Berman

- **Streamlined communications with collaborators and sample submission for customers - the use of the LabLink web-collaboration tool reduces time spent dealing with customer requests.**

"The really key thing is that we have access to the data at many different levels, from external investigators who

submit samples to us, to investigators within our center, and the lab technicians. Since we're a core center we need pretty sophisticated tools to deal with our customers."

- **Improved core services with centralized data management and reliable sample tracking, saves significant amount of time which was previously spent chasing down samples and managing vast amounts of data.**

"The thing we don't want is for a lab technician or customer to have to email a bioinformatician to get information about samples. This bogs down the bioinformatics staff and slows down the pace of research."

- **System versatility protects LIMS investment over the long term.**

"A key requirement for the system is - it has to handle a really quickly evolving lab workflow and analysis workflow."

- **Significant software development time and cost savings resulting from existing integrations to Illumina platforms, and the flexibility of the LIMS system to enable bioinformaticians to make use of API's for more customized work.**

"The requirements for this system would include built-in support for the Illumina platforms that we use. And we wanted it to provide a really tight level of integration with our custom data analysis pipeline."

- **Greater ability to add scientific value to the data - purchasing a system that manages the sample tracking and data management frees up internal resources to focus on continued development of custom software for bioinformatic and QC related analyses.**

"What is nice about using the Geneus system is that it frees up our in-house software developers time from writing sample tracking software to work on more customized quality control reports, data normalization methods and biological analyses. These are the things we want to spend our time working on."

- **System scalability ensures the system will be adopted long term as next generation sequencing throughput and data generation sky rockets.**

"Our microarray projects actually stress the GenoLogics system much more than the Genome Analyzer does, just because the number of samples at this point is orders of magnitude more than what we process on our Genome Analyzer (although anyone who works with next-gen sequencing knows that the downstream processing is still a challenge). With the Geneus system, it really keeps up with our sample and data tracking needs quite easily. We have three Genome Analyzers today but we feel it could scale to many more."

In order to realize great results from their investment in next generation sequencing technologies, institutions like the USC Epigenome Center are investing in information and data management systems for three key reasons:

1. To streamline the process for lab managers and researchers with comprehensive sample traceability, smoother hand-offs of experiments between groups, instrument integrations, and the automation of many manual steps such as data analysis pipelining.
2. To keep up with the constant changes in next generation sequencing technology. This is cutting-edge research and it is changing and scaling rapidly. Classic LIMS approaches with fixed workflows cannot assist in a quickly changing environment.
3. To enable the sharing of information through the Geneus/LabLink interface to direct experiments and help lab managers and researchers with day-to-day decision making.

In summary, bioinformaticians, lab managers and researchers alike can realize great time and cost savings benefits, from implementing a strong lab and data management system with built-in ability to integrate to industry leading platforms such as Illumina's Genome Analyzer, and unparalleled flexibility that enables the use of API's to create added value on top of the base LIMS/ data management platform.

References

1. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. 2008. Nature 455:1061-1068.

Listen to Dr. Berman discuss his lab workflow and custom integration work in detail and to view his slides. Visit: www.genologics.com/forms/usc-illumina-webinar to download the USC webinar.

For more information on the Geneus next gen lab and data management solution visit: www.ienablenextgen.com