BIOTEAM
Enabling Science

Managing the Life Sciences Data Deluge
MiniLIMS 1.0 is Next-Gen Ready, Affordable, and Deploys Quickly

For organizations working with next generation sequencing (NGS), the most significant challenge is managing the enormous quantities of data and associated metadata produced by the new instruments. A lab with a single instrument such as the Illumina HiSeq 2000 can generate 700 Gigabases per run while a massive sequencing center such as BGI (formerly Beijing Genomics Institute) with more than 160 Next-Gen machines can generate around 5 Terabytes every day.

Clearly, the 'Big Data Era' has begun throughout life sciences. It requires new practical tools for managing and deciphering the data. BioTeam's MiniLIMS was developed to solve just this problem. Widely deployed in NGS labs for more than a year, and steadily improved by BioTeam, MiniLIMS 1.0 is a powerful, scientist-friendly laboratory data management platform suitable for virtually any experimental technology (e.g. proteomics, flow cytometry, etc.)

Next-gen data streams quickly become complex

"The data is the killer," says Mick Correll, associate director of the Center for Cancer Computational Biology (CCCB) at Dana Farber Cancer Institute (DFCI).

While operations on the scale of BGI are rare, the CCCB initiative is typical of a wave of Next-Gen sequencing centers sprouting in academia and the biopharmaceutical industry. They start modestly, usually with one early-Next-Gen machine, and upgrade quickly to keep pace with advancing technology and demand for increased sequencing capacity.

Case in point: CCCB upgraded its Illumina GAIIx to the HiSeq 2000 and its sequencing business is up. Correll says, "Next-Gen sequencing wasn't a initially focal point but in the past year or so it has become a major part of our operation."

Getting the most from these Next-Gen investments, both in terms of running the instruments efficiently and delivering solid science, requires robust data management, emphasizes Correll. Excel spreadsheets are a common starting spot but even small sequencing operations soon discover they're insufficient to reliably enter and retrieve sample, library prep and run data. Traditional LIMS products are powerful but they are also expensive, usually difficult to customize, and can take months or longer to deploy.

Even small sequencing labs need a system to organize data

"When organizations purchase LIMS, they are often buying more functionality than they need," says Michele Clamp, a senior consultant at BioTeam and the key developer of MiniLIMS (as well as Ensembl and Jalview).

BIOTEAM
Enabling Science

For years, the BioTeam has been developing LIMS solutions for customers using WikiLIMS, a framework based on the widely used Mediwiki software and the bundle of Semantic extensions. This framework has been a great foundation and continues to add value for customers with rich functionality requirements. However the complexity of the system is at odds with the goal of a many labs to simply get up and running. The BioTeam has distilled the key elements of WikiLIMS into a powerful web-based application, MiniLIMS.

MiniLIMS 1.0 is Next-Gen ready 'out-of-the-box' and can be deployed in a matter of days. "We are creating the core piece you need and we're starting from one part of the lab which is the DNA sequencing core," says Clamp, the prime architect of MiniLIMS.

Fundamental to MiniLIMS is the strategy to take onerous tasks previously handled by database administrators and programmers and put them into the hands of scientist by providing intuitive GUI-based tools that eliminate coding requirements. Making these improvements is an ongoing effort. Most recently, "We've made changes in three main areas of data input, data display, and data output to reflect customers' needs in a rapidly changing lab," says Clamp.

- **Report Builder.** Users can now easily create their own views of the data and sort, filter, and group the results as they wish. These can be used to generate reports or simply interrogate the data based on the lab users' questions. For example, data can be displayed over user-specified ranges for reports and filtered by properties such as lane yield to locate badly performing sequencers. Added output options (printing and excel) allow the user to export the data for further analysis in other software or to be shared.

- **Table Configuration.** In addition to the creation of new data tables all existing tables are also configurable. Users can add and remove columns to the display and sort and filter them how they wish. The changes can either be ad hoc or saved back to the database as a permanent change.

- **Easier Data Input.** Data input forms now allow autocomplete on all fields. This allows the forms to be filled in much faster and removes the need to remember multiple group or sample names. Additionally, new pages can be input on the fly while a form is being edited without having to go to a new page. For example when creating a new sample, a new species page can be quickly created without moving away from the sample form.

Says Aaron Kitzmiller, another BioTeam consultant and veteran of sequencing instrument companies, "We are constantly finding places in the system where we can get rid of the need to code and make graphical user interfaces. We want to be able to put MiniLIMS into the hands of scientists at the bench. That person may not have programming skills, but wants and needs the ability to manage his or her data."

Linked data representation eliminates the need for custom code

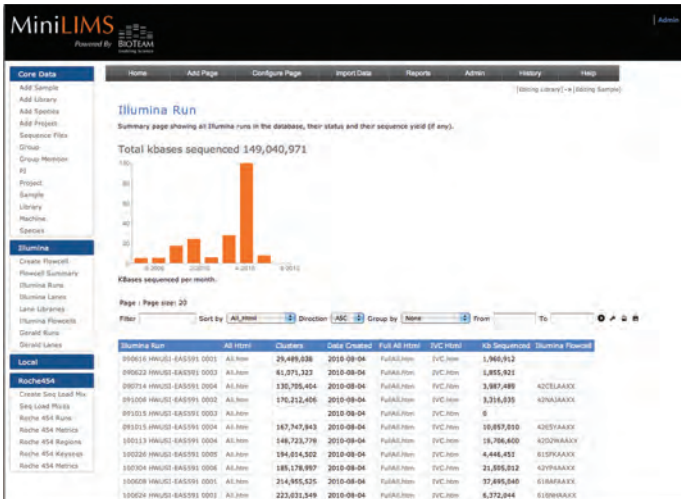
The ability to the make these kinds of powerful, yet user-friendly improvements stems in part from BioTeam's early decision to build MiniLIMS using linked data representation. Doing so speeds and simplifies customization (coding is generally not necessary). "The whole thing is structured in a semantic way, which is basically a single four column table. As there is no complex schema, you can change what you want to store without ever having to know about what's happening behind the scenes," says Clamp.

BIOTEAM
Enabling Science

A column-oriented database structure was used to speed performance. Unlike traditional row-oriented DBs, which are optimized for online transactional processing (OLTP), column-oriented DBs are inherently more efficient for analytics applications performing many queries against large data sets.

'Plug-ins' let people only buy the functionality they need

BioTeam adopted a 'plug-in' strategy to providing functional features and mapped early features to the workflow and data requirements that characterize Next-Gen sequencing. Core plug-ins currently part of MiniLIMS include forms for tracking people, projects, groups, samples, and libraries.



"When you put samples into the system they usually belong to someone and to a project. We've seen that adding sample data is the bulk of what people will do," says BioTeam's Clamp. "Once the samples are in they then need to be formatted to go on the sequencer. For example for Illumina you have to create a flow cell and we provide an easy way to do that."

Illumina plug-ins for working with GAIIx, HiSeq, base-calling, and QC data are available now and take advantage of the autocomplete functionality to fill in libraries. A further improvement is the ability to filter the libraries to those in a specific project or species (or any other library property - even user created ones) so making flowcells becomes much easier.

For people who like to create input data offline a new flowcell upload form will upload multiple samples and create a flowcell in one go. On the data loading side MiniLIMS now has support for the CASAVA 1.8 base-calling output and can import fully demultiplexed runs.

A new Roche 454 plug-in is now available that lets users associate libraries with runs and regions and automatically uploads output data. Prototypes for PacBio/SOLiD platforms are in final development stages and MiniLIMS' near-term plans include plug-ins for Ion Torrent, reagents, invoicing, analysis, and graphs.

Clients like the flexibility and easy installation

"I like the fact that when you look at a particular sample, you have a link to every page on which the sample is mentioned. It's a good way of navigating the database," says Timothy Read, Associate Professor, Emory University School of Medicine, and Director, Emory GRA Genomics Center (GGC).

"We looked at some other systems but my feeling was they were designed for centers that were much bigger. We didn't need the enterprise scale but were moving up to the HiSeq, which was going to be effectively four times the throughput we were able to do on the GAIIx. We knew we needed

BIOTEAM
Enabling Science

to get something a little more formal in place," says Read. CCCB had similar concerns. "We needed something quick, almost what I would call disposable soft-ware," says Correll. "We wanted it to be up and running within a month to three months, knew that we weren't going to use it beyond three years, and didn't want to spend six months doing an exhaustive analysis."

Three groups interact with CCCB's MiniLIMS: the lab team which checks samples in; after the run the informatics team conducts a first-pass analysis and prepares the data files to be returned to the client; and the business office which tracks the status of projects to determine who can be billed and who can't.

"The problem was this steady stream of samples coming in from 40 labs. We needed to connect the dots of what went where, if it had been processed, had the data gone back to the customer, and just being able to organize it all. There were a number of more comprehensive LIMS solutions out there," says Correll, "but the bottom line was that MiniLIMS solved 85 percent of our major pain points and did it in a really simple way. They positioned it as something that would literally be installed in a couple of days and that's pretty much how it went."

Key System Benefits

Most operations are quickly done using the MiniLIMS web interface but it also has powerful API and query engine for custom pages.

Here are just a few MiniLIMS' capability highlights:

- Linked Data Technology.
  MiniLIMS uses the power of Linked data representation to let you store the data you want and automatically generate hyperlinked tables, web forms, and reports. Check out Tim Berners-Lee's presentation on Linked Data at the 2009 TED conference: http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

- Open Source Tools.
  MiniLIMS uses open source tools (Mysql, PHP, Apache). Users can go from a vanilla Linux install to a customized site in under an hour.

- All Fields Are Configurable.
  Names of data field can be edited and new fields can be quickly added using the web interface. Restricted vocabularies and restrictions by page type are also easily configured.

- Batch Data Upload.
  Columns of data (tab, space or comma delimited) can be uploaded. Defining which columns to import can be done on the fly. Column data can be renamed and reformatted using the web interface.

- Alerts & Edit History.
  Configure an email to be sent when any value on any page changes. Use a Google voice account to send a text message. MiniLIMS also captures the full history of changes. "We store the history for logins, inserts and deletes and we can roll back to previous incarnations of pages if we need to, and see who made the changes," says Clamp.

BIOTEAM
Enabling Science

- API and Web Access.
  Access data and create new web pages using the object–oriented PHP API. Alternatively fetch and store data via the web using the language of your choice. "Programmatic access to MiniLIMS is powerful" says Clamp. "We've tried to make the API easy to use yet powerful enough to create new web pages. Whether you're automating your pipeline or making new web reports MiniLIMS provides an easy way to store and display data".

Virtually all Next-Gen sequencing centers serve multiple constituents, whether internal, external, or a mix of the two. For that reason, managing appropriate access to the data is critical.

The Emory Genome Center, for example, is the core lab for work throughout the university. Read says, "We have 6-10 projects at any time. It's all internal work. We have the added complication where we'll have one run with multiple projects and with different PIs. " Emory also has a Roche/454 instrument.

Client Experiences

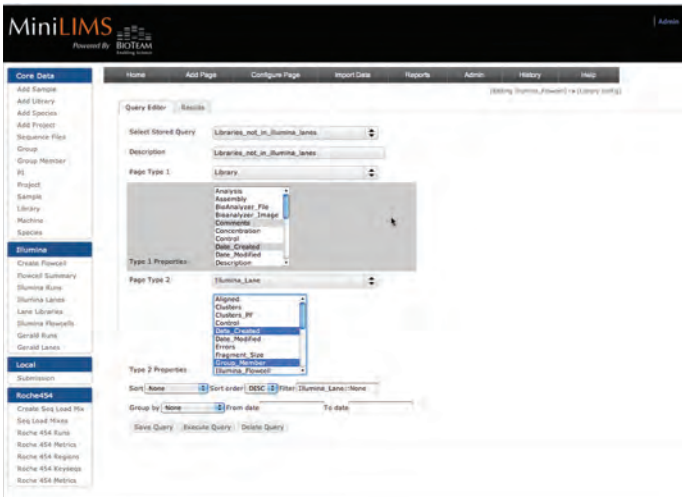
The CCCB experience is a little different but also has multiple constituencies. "We're primarily a computational group and we run like small business. So we do bioinformatics consulting for researchers within Dana Farber and across the campus. We develop software of our own and we do some primary genomics data analysis," explains Correll.

Getting into Next-Gen sequencing, he says, "was really smart for a computational group because to really understand how best to analyze this stuff you've go to know the whole chemistry." The growth of the sequencing business has been a bit of a surprise. CCCB routinely does two runs (16 samples) a week on the HiSeq.

MiniLIMS has a password/login function with multiple user roles settings to control who sees what. "I think the multiple layer viewing capability is impressive and it's also very intuitive to set up," says Emory's Read.

Change MiniLIMS as your process changes

Although MiniLIMS has been carefully architected to be useable out-of-the-box, BioTeam also designed the software to be easily configurable by users, generally without requiring additional coding. This is particularly useful for groups with limited IT expertise or resources, but even when those resources are present as at CCCB, BioTeam's approach speeds change-making process.



"You can add pages or modify them yourself, mainly through the web interface," says Clamp. "We've tried to eliminate as much coding as

BIOTEAM
Enabling Science

possible and still keep the power of the semantic query available." The process is straight forward, and the use of semantic technology mean data are linked automatically.

The column based database approach also means MiniLIMS is extensible by the user. "MiniLIMS doesn't mind what kind of data it stores and is easily extensible for new data types." Clamp explains. "If you want to add the ability to store new things such as funding information or new instrument output this can be done through the web interface in a matter of seconds."

Being able to make changes is an important feature for most users. "I like that I am not dependent on BioTeam to make every tweak. If I want to add another field, if I want to pop in a link for an html report, our guys can actually do that themselves," says Correll.

Conversely some changes require help. "Our workflow is not complicated. Samples come into the lab and we enter them into the system. We've been talking to Bioteam about changing this to an integrated barcode reader and iPod Touch," says Correll. "I'd love nothing more than at check-in to slap a bar code on a sample, scan it, and scan the badge on the technician so we know who checked it in at what time."

Currently MiniLIMS is focused on Next-Gen applications, but at its core, it is a powerful & flexible metadata management platform that could be customized to accommodate virtually any experimental technology. One client is currently configuring it to manage a microarray workflow.

BioTeam's technology roadmap for MiniLIMS calls for introducing more core functionalities plus plug-ins for other technologies and applications (reports, analysis tools).

MiniLIMS™ Requirements and Feature List

Hardware Requirements

- CPU 2 GHz. Preferably >1 processor for large installs.
- RAM 4 Gbytes
- Disk space (100 Mb for base install, 2 Mb per 100 samples).
- Mac OSX or Linux.

Software Requirements

- Java (for the installer).
- Apache2 + php5
- Mysql 5
- Mysql 5 extension for php5
- Root access (or at least read/write access to web directories).
- Mysql create/drop database access.

Other Requirements

- A server that has read/write access to the Illumina run directory.

Feature List

- Stores samples and associated data (species, library type, project, owner).
- Stores details of library preparation and uploads QC images.
- Organizes people into PIs, groups, projects and group members.
- Groups can only view and edit their own data.
- All data can be entered via configurable web forms.
- Username and password access to the site.
- Combines samples into multiplexed flowcells.

BIOTEAM
Enabling Science

- Tracks status of Illumina GAIIx runs.
- Imports Bustard summary data.
- Imports GERALD alignment summary data.
- Provides links to Illumina output QC files.
- Provides links to sequence files.
- Pie charts and bar charts summarizing Illumina run data.
- Can configure, add and delete properties stored with data.
- Can add new data types and automatically generate pages and forms.
- Can add/delete new authenticated users and assign to roles.
- Stores the history of edits to the data.
- Can read/write data using php API.

To schedule a live demo and to receive a demo account please contact:

Stan Gloss, Managing Director
BioTeam Inc.
stan@bioteam.net
978-304-1222