Selecting a LIMS for next-generation sequencing research

Genomics has revolutionised the life sciences industry by combining human ingenuity with right-place/right-time serendipity. Advances in computer processing and storage have provided the bandwidth and throughput to enable the visionary science imagined by those pioneering the Human Genome Project. The revolution continues today. No other industry has seen processing speeds rise and costs drop as dramatically as genomics (**Figure 1**). And with next-generation sequencing (NGS) providing the ability to sequence entire genomes in less than a day for pennies per base pair, organisations are now wondering how they will handle the data these techniques generate

By Bruce Pharr and Dr Michael Kuzyk

Organisations understand the true promise of genomics lies not in the production of data, but in how well scientists exploit the data produced. The challenge is more daunting in light of recent commentaries which note that while the cost of sequencing has decreased, analysis costs remain high – more than half again as much as sequencing alone. Simply gathering and storing the reams of data generated is not enough – data needs to be considered in tandem with contextual sample and project information in order to inform downstream analysis and critical research decision points.

Lab information management systems (LIMS) are a mature class of life science software introduced in the 1980s to manage such tasks as sample management, experimental monitoring and data collection, analysis and reporting. Commercial LIMS are now available specifically for genomics. The best of these systems offer the following advantages to modern sequencing facilities:

- End-to-end sample traceability.
- Scalability so that labs can get up, running and producing results quickly.
- Adaptability to help labs accommodate changing technologies and methodologies.
- Data analysis, workflow management and operational reporting tools to ensure labs run efficiently and collaboratively.

Achieving these benefits requires labs to assess available LIMS against specific experimental needs and research workflows. This article reviews three criteria that next-generation labs should evaluate when selecting a LIMS. The choice will depend on:

- How well the system supports best practices in instrument configuration out of the box.
- How easy the system is to configure and customise.
- Whether the system provides user specific interfaces to streamline the work performed by the various types of users who will need to interact with the LIMS.

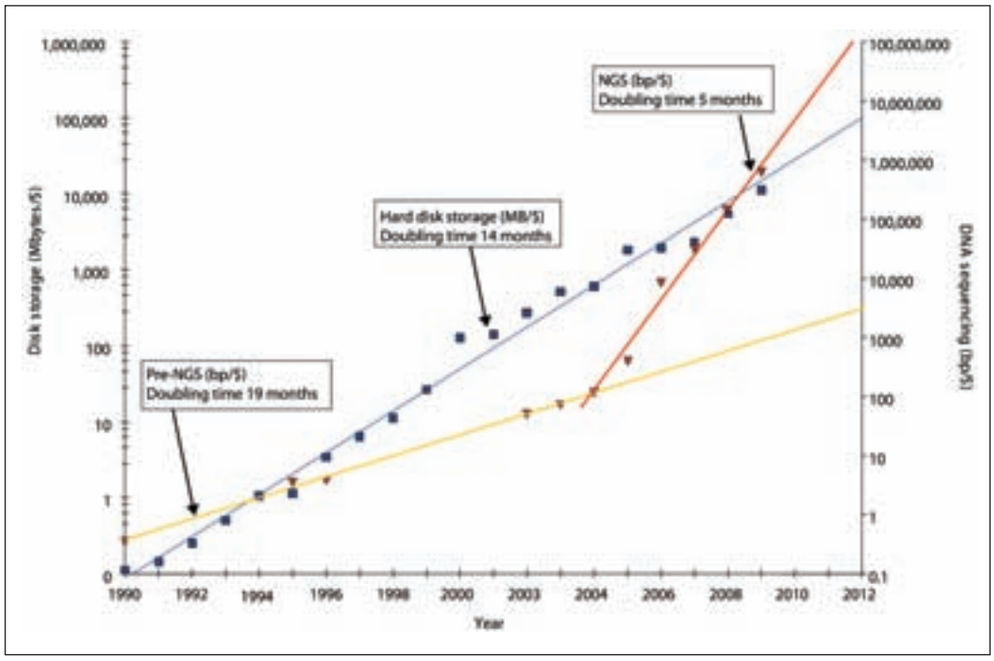

Figure 1
A logarithmic plot of historical trends in storage prices versus DNA sequencing costs. The advent of next-generation sequencing (NGS) in 2004 causes an inflection (red) in the exponential curve associated with DNA sequencing costs (yellow) to a doubling time of less than six months. *(Source: Stein, Genome Biology 2010)*

Data and lab information management for NGS

Today, next-generation genomics labs can easily produce more data than they can effectively manage or analyse. Industry analysis that once focused on the costs associated with sequencing genome data now focus on the challenges of managing it. In a JP Morgan report conducted in 2010, lab directors cited data storage, data management and informatics as the biggest collective hurdle to expanding NGS operations.

This hurdle, however, is more of steeplechase than a single, easily cleared obstacle because the bottlenecks continually shift. Storage was the first critical concern for most organisations as they confronted the reality that new machines running at capacity could generate in a single year more information than was deposited in GenBank by the beginning of 2008 (**Figure 2**). Some labs changed their initial data handling strategies midstream to free up space; image files, for instance, are by far the bulkiest data types produced by sequencing, and some labs have opted not to store these files types. Other labs, realising that processing and storage power are relatively cheap, have opted to store everything and figure out afterwards what they need for analysis.

The 'store everything' approach, though, shifts the bottleneck to analysis, which explains why analysis costs remain high even as the total cost of sequencing a human genome has significantly decreased. The most generous estimates put analysis at half again as much as the cost of sequencing. Researchers at the National Center for Genome Resources said that the bulk of the costs in a quarter-million dollar sequencing project in 2009 comprised analysis expenses. "An awful lot of manual analysis is required", according to this report. "It's a very large amount of human effort".

Clearing the analysis hurdle requires more than an investment in hardware, infrastructure and bioinformatics expertise. Organisations must completely revamp the workflows that support sequencing, many of which are based on manual, one-at-a-time processes and information stored in disconnected silos such as spreadsheets, emails or document-based communications and paper lab notebooks. Sample preparation often emerges as a critical area of emphasis for organisations seeking to streamline operations. The most prestigious grants and research projects often require labs to be able to guarantee sample traceability – it is essential when dealing with the often limited DNA supplies associated with certain clinical sample cohorts. Nevertheless, busy labs often struggle to ensure that samples received from clients and collaborators are appropriately labelled and that all vital experimental context is passed on efficiently and accurately to bioinformaticians. Clear sample taxonomy, tracked from the moment a sample enters a lab to the point at which results are reported, makes it easier for research scientists and bioinformaticians to set up

and validate experimental runs. It also speeds downstream analysis by ensuring that a sample's history and origin is tied directly to the results obtained (**Table 1**).

The unprecedented throughput, experimental complexity and changeability associated with NGS create unique challenges for traditional LIMS. The rapid timescales associated with sequencing require LIMS that can be quickly implemented and easily configured by lab staff with no programming experience. Additionally, bioinformaticians and scientific programmers must have the power to make changes through the software's application programming interfaces (APIs) in order to accommodate the unique workflows that drive next-generation genomics research. Finally, NGS requires iterative, collaborative work that is performed by many different types of scientists. User-specific interfaces can ensure that these workers have access to all and only the information they need to do their jobs effectively.

Selection criterion #1: Does the LIMS enable labs to get up and running quickly?

One of the primary selling points of LIMS since its inception has been its ability to integrate with laboratory instrumentation. As recently as 2009, LIMS users across all industries cited instrument integration as the most desired capability in a LIMS, with fully 70% of academic laboratories ranking it number one. In NGS, however, LIMS must do more than simply run and interface with instrumentation – it must provide a framework to appropriately capture data and streamline and automate mundane, routine tasks to eliminate the bottlenecks that can slow or even stall sequencing workflows and analysis. Data management and experimental tracking is even more difficult for labs using DNA indexing (also known as barcoding or tagging) to pool and multiplex samples from diverse, unrelated sources on a single flow cell lane. These techniques have in fact created a bottleneck at the library preparation step, where sheer throughput combined with the need to track which samples have been pooled in which runs, delays the rate at which labs can get samples on to sequencers.

While each type of NGS instrumentation comes with vendor-specified kits and protocols to optimise use and performance, the generalised task of integrating with sequencing instrumentation encompasses three primary phases, each of which should be supported out of the box by a LIMS.

First, organisations must consider how they are collecting information about samples and associating them with runs. Traditionally, scientists have spent many hours poring over Excel spreadsheets to check sample preparation and run assignments. Any LIMS should provide some level of integration with major NGS instrumentation. How the LIMS integrates with the instrumentation may differ: some LIMS may integrate more tightly with particular instrumentation, and organisations should verify the connectivity between LIMS and their preferred instrumentation. For instance, some integrations may require scientists to specify the samples they wish to run so that the LIMS can generate the appropriate files for the lab's sequencing equipment. Conversely, NGS instruments can be configured to hand off information on performed runs directly to the LIMS, reducing hands-on time for lab staff.

The second phase of instrumentation integration is configuring the LIMS to track the quality of sequencing data coming off instruments. Many sequencing instruments run for days on end, making it wasteful and inefficient for organisations to wait until runs are completed before evaluating the quality of the data obtained. In addition to monitoring the status of runs in progress, LIMS can also collect metrics, such as the total bases yielded from a run or the percentage of base calls with a Phredquality score of more than Q30. Over time, these metrics can aid in assessing instrument performance. With data from sample runs archived and searchable in a centralised LIMS, labs can make better, more informed decisions about which samples to rework, whether to request more samples for further experimentation, or how much time to spend on additional analysis.

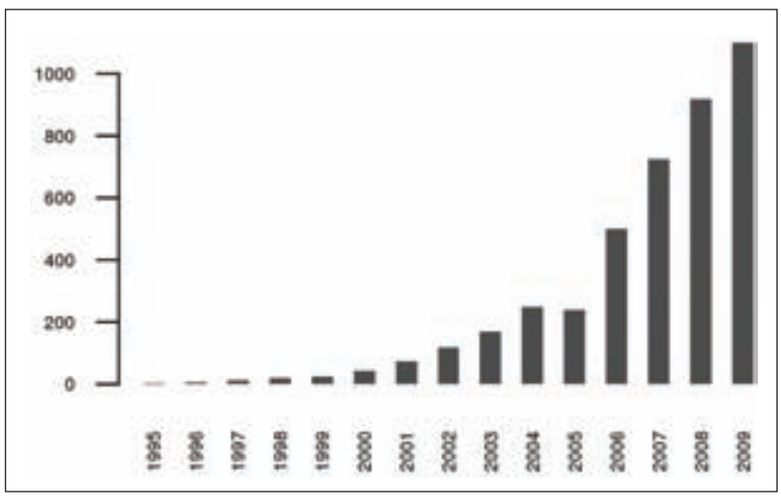

Figure 2
Number of genomes entered into GenBank by year as of September 2009

The final consideration in integrating instrumentation effectively with a LIMS is results tracking. Most labs today have accumulated massive directories on their local area network dedicated to storing information associated with sequencing runs. Often this high level information appears in reports and summaries, while the underlying, granular information is stored for future reference. Unfortunately, locating necessary detail can take staff hours or even days, leading some labs to rerun experiments rather than sift through directories for archived files. Multiplexing can also require an additional data management step; in some cases, those pooled samples must be 'unpooled' or 'demultiplexed' before the results can be analysed and interpreted.

A LIMS can eliminate some of the most tedious aspects of NGS for lab managers and bioinformaticians. Intuitive query tools enable labs to quickly collect information on sequencing runs, whether it was obtained last week or last year. Best-in-class LIMS also provide simple ways to create automated workflows that can demultiplex reads, create sample sheets for sequencing instrumentation, or incorporate specific open source and commercial analysis pipelines. Freed from the need to sort through and organise data, lab staff can spend more time on analysing data, making decisions, publishing results and envisioning creative research projects.

Selection criterion #2: How easy is the LIMS to configure and customise?

Change is the operative word in next-generation genomics labs. Methods used one day are practically obsolete the next. Or they may not exist at all – analytics and methods created specifically to handle unique science are also key outputs of next-generation genomics labs. In this environment, labs succeed by pushing the boundaries of innovation—and they cannot afford to be constrained in their vision by the software they

PRIMARY PEOPLE INVOLVED	STAGE IN WORKFLOW	KEY INFORMATION TO TRACK

External collaborators
Principal investigators
Bioinformaticians
	Project Initiation & Sample Submission	▪ Defined research goals
		▪ Agreed upon experimental approach number of samples
		  ▪ types of samples
		  ▪ sample taxonomy
		  ▪ type of sequencing analysis (single or paired end read, read length)
		  ▪ data analysis strategy
		▪ Quotes, statements of work
		▪ Contact and payment information

Lab managers Lab techs	**Library Preparation**	**Project-Specific experimental d etails** ▪ Sample identity ▪ Library strategy (genomic, mRNA-seq, ChIP-seq, mate paired, indexed) ▪ Average fragment length ▪ Gel images ▪ Quantitation results ▪ Quality measurements **Information to ensure and Improve Quality** ▪ Kit versions ▪ Reagent lot numbers ▪ Protocol information ▪ Number of PCR cycles
	Cluster Generation	▪ Sample loading pattern ▪ Kit versions ▪ Reagent lot numbers ▪ Protocol information ▪ Flow cell ID
	Sequencing	▪ Location of data on network ▪ Kit versions ▪ Reagents lots numbers ▪ Protocol information ▪ Flow Cell ID
Bioinformaticians Lab managers Technicians	**Primary data Analysis**	▪ Run quality metrics (%PF, first cycle, intensity, cluster density, etc.) ▪ Base calling algorithm ▪ De-multiplexed reads
Bioinformaticians PIs / External collaborators Lab managers	**Secondary and tertiary data Analysis**	▪ Assembly / alignment algorithms ▪ Algorithm parameters ▪ Location of output result files (SAM/BAM files) ▪ Summary tables of SNP counts, InDels, etc.
Bioinformaticians PIs / External collaborators Lab managers	**Results Reporting & Invoicing**	▪ Collating all work performed on samples ▪ Summarizing results and quality metrics ▪ Invoicing for work performed

Table 1: Overview of information required to track next-generation sequencing research

implement to manage sample and experiment data and workflows.

The adaptable, extensible systems required by next-generation genomics labs are hard to buy and even harder to build. Many commercial LIMS designed for NGS research can be rigid and prescriptive about how work proceeds – and changes to the out-of-the-box configuration are discouraged and often impossible. Labs also have the option to work with broad enterprise LIMS vendors, who will build tailored systems, but at a cost – these systems take time and money to develop and when a lab needs change (and in next-generation sequencing, change is guaranteed), the vendor will need to update the system.

Given that commercial software vendors have trouble building flexible, adaptable LIMS, it is surprising how many labs opt to build their own LIMS. The promise of implementing exactly the system they want is appealing, but these labs ultimately discover that maintaining and updating software over time drains critical resources. Does a leading-edge NGS lab also want to become an expert in software design and development?

The best answer to the build/buy question is both. Effectively implementing this hybrid approach requires that labs first select a LIMS that addresses their specific science and workflows needs. They then build the parts they are best suited to build. Achieving this requires software that can be configured by scientists and customised by scientific programmers and bioinformaticians using modern, familiar software development tools and APIs.

Software marketing often conflates 'configuration' and 'customisation', but engineers understand they are distinct. Configuration refers to changes in existing software that can be made via the user interface by any user. As mentioned above, some systems offer preconfigured, out-of-the-box set-ups that scientists can use to add new lab methods, collect records off a new instrument, or specify a particular sample preparation procedure. Easy configuration empowers scientists – who best understand laboratory requirements and how the system needs to work – to make critical changes to the LIMS.

More importantly, configuration frees programmers and bioinformaticians to focus on more high-value projects, which typically require customisation. Scientific programmers understand that customisation is quite simply changing the actual code of software so that it can do something new or different. Anyone armed with the appropriate programming expertise, software tools and APIs can make these types of modifications. Bioinformaticians and scientific programmers work best when they have the power and control afforded through systems that let them use familiar tools (such as standard-based architectural styles and scripting languages such as Python, PERL or Groovy) to adapt software to accommodate the unique needs to their labs. Software that supports both configuration by scientists and lab technicians and customisation by bioinformaticians and scientific programmers enables labs to efficiently implement systems that better match their current and future informatics requirements.

Selection criterion #3: Does the LIMS accommodate different users and workflows?

NGS labs require varied expertise to accomplish their objectives. Principal investigators, lab managers, lab technicians, scientists, bioinformaticians and scientific programmers all contribute to keep experiments running quickly and efficiently. All of these individuals have different responsibilities, priorities and correspondingly different ways that they wish to view and act on data.

It sounds clichéd, but in a leading-edge next-generation lab, one user interface does not fit all. To work effectively, users require access to all and only the information relevant to their job. Targeted user interfaces in an NGS LIMS should provide a dashboard of relevant activities while also pulling appropriate data from the larger system and displaying it to those who need to act on it. Intelligent, targeted user interfaces can aid the following types of users:

- **Lab technicians.** Scientists and technical staff require fast, efficient access to data that helps them track sample status, determine which samples can be prepared together, simplify creation of library pools for multiplexed sequencing runs and access and review past work. Dashboards should help them answer such questions as: "What experiments do I need to carry out today?" "What work is coming my way so I can plan ahead?" "Which libraries can I pool together for a multiplexed run?" or "Am I getting good quality data off that run I just started?"
- **Lab managers.** Management dashboards can provide a high level summary of everything happening in the lab—an overview of active project status and instrument performance with the ability to drill down into activities to look at more specific results or metrics. Managers also require reporting and project management tools to manage client communications, invoicing, and administrative reporting. Interfaces

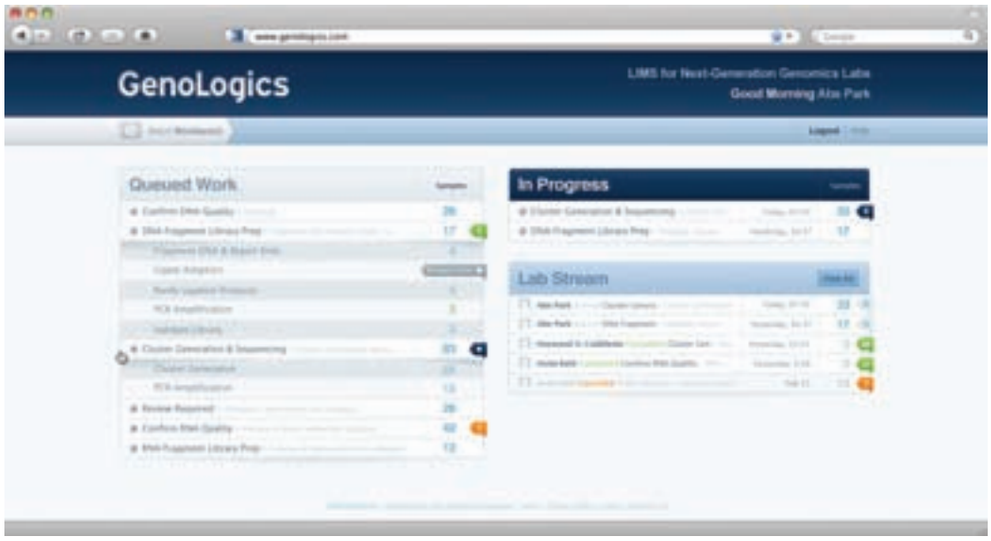

Figure 3
Example of a role-based user interface for lab technicians. A sample dashboard that provides at-a-glance information to the lab tech

should help managers answer such questions as, "What the quality of data coming off that new sequencer?" "What's the status on the project we've been running for our new collaborator?" or "Where are the results from that experiment we did six months ago?"

- **External collaborators.** A secure portal ensures that outside collaborators have immediate access to data relevant to their projects, while protecting the broader project data accumulated by the servicing lab. The portal should provide a centralised way for collaborators to initiate work requests, inquire about project status and view project summaries. Through the interface, collaborators should be able to answer such questions as: "Is my project finished yet?" "Are there any results available to download?" or "I've got some additional details to provide – how can I get them to you?"

Selecting the right tool for the task

NGS may be the newest, hottest technology in the lifescience space, but the software needed to manage and communicate NGS data is mature and well understood. LIMS have proven themselves across a range of industries for more than 30 years. Many options exist – what system is best for a given facility will depend on that facility's size, scope and research goals. The unique demands of NGS, however, make certain issues imperative. Will the LIMS be easy for a lab to initially implement? How easy can the system be adapted when lab needs change? Does the LIMS provide actionable information to specific users so that they can do their jobs better and faster? A thorough examination of these questions will help organisations select a LIMS that meets their lab and data information management needs now and in the future. DDW

Bruce Pharr *has increased the value of half a dozen companies by generating incremental income, building brand equity and helping execute successful M&A strategies. Most recently he served as VP, Marketing, at Symyx Technologies where he helped transform the company from chemical research to lifescience software, positioned it as a thought-leader in the emerging, fast-growing electronic laboratory notebook (ELN) market and facilitated a merger with Accelrys. Mr Pharr has held lead marketing positions at several technology-based companies and he founded and led a marketing consultancy firm for more than a decade. He is currently VP Worldwide Marketing for GenoLogics.*

Dr Michael Kuzyk *brings a wide range of expertise in both genomic and proteomic research and lab management experience to GenoLogics where he currently holds the position of Senior Product Manager. His responsibilities include looking after the genomics product line with a focus on the next generation sequencing market segment. Most recently, Dr Kuzyk has held the positions of Assistant Research Professor with the University of Victoria and Staff Scientist at Canada's Michael Smith Genome Science Centre, where he managed the day-to-day operations of the core services labs. Dr Kuzyk is a graduate from the University of Victoria where he completed his doctorate in biochemistry.*