

Information Management Systems for Molecular Biology

Todd M. Smith, Ph.D.

Geospiza, Inc.

July 18, 2000

Introduction

Technological advances in molecular biology have made it possible for laboratories to generate unprecedented amounts of data. DNA sequencing, for example, has seen an increase in throughput of over 400-fold in recent years. The accelerated production of data has produced a corresponding need for better data management.

The increased need for data management is partly due to institutional changes. Many institutions have been able to economize by establishing core facilities which offer laboratory services to researchers on a per sample or project basis. Core facilities are cost-effective because they provide several researchers with access to expensive equipment and skilled personnel. Because these core facilities handle a variety of data types and a large number of samples from multiple sources, their information management needs provide a worthwhile model for laboratory information management systems (LIMS).

As core facilities continue to add services and automate sample handling and data collection, the volume of data will continue to grow. Traditional methods of record keeping such as paper laboratory notebooks and recipe cards have already been overwhelmed by the current ability of laboratories to generate data. Today's information management systems must be able to accommodate different types of data, store information relevant to each sample, track sample processing, track laboratory workflow, and return data to scientists in a secure fashion that is compatible with further analysis. Since core facilities operate as businesses within institutions, they benefit from the ability of LIMS to track costs, maintain inventories, and schedule and record equipment maintenance.

The remainder of this article reviews general features of LIMS and presents a case study of Geospiza's *Finch-Suite* LIMS solution at Rocky Mountain Laboratories. The case study illustrates how a commercially available software system can be used to meet the information management needs of a DNA sequencing facility. Although the *Finch? -Server* was designed to facilitate data management in DNA sequencing facilities, the principles discussed here also apply to other high-throughput procedures such as array hybridization and mass spectrometry.

LIMS

LIMS (Laboratory Information Management Systems) are collections of software, communication devices, and computers that acquire, store, analyze, and present data and information about laboratory samples and their processing. LIMS are used to coordinate workflow and the movement of samples and information through different laboratory processes. LIMS centralize data storage, automate data analysis, and provide quality assurance reports for process monitoring. The central components of a modern LIMS are a relational database management system (RDBMS) running on a computer with one or more software interfaces with which users enter, view, and process data and information.

Features of LIMS

LIMS are integrated systems that encompass four functional areas: data and information capture, data analysis and reports, laboratory management, and system management (Table 1) (Liscouski 1995). Each of these areas can differ in complexity and cost depending on the laboratory environment.

The level of complexity needed in each area depends on several factors, such as the mission of the laboratory, the number of samples to be processed, analysis requirements, and workflow complexity. For example, a laboratory that sequences between 10,000 and 100,000 DNA samples per year will require less data entry and automated sample processing than a laboratory that sequences 10,000 samples per day. Similarly, a laboratory that sequences complete genomes will have different analysis requirements than a laboratory that focuses on genotyping or clone identification.

Table 1. Basic, Intermediate, and Advanced Features of LIMS Functions		
Function	Level	Features
Data and information capture	Basic: Intermediate: Advanced:	Manual data entry methods File transfers and simple barcode entries Two-way communication with laboratory devices such as data collection instruments or robotic devices Fully automated sample tracking
Data analysis and reports	Basic: Intermediate: Advanced:	Basic calculations, result verification, and preselected reports Automated procedures, reports, tables, and graphical presentation of results Integrated analytical procedures that link different types of experimental data Integrated external software systems Graphical data may be viewed with separate software tools (eg, chromatogram viewers for DNA sequencing) "Rich" reports, with links to specialized views and drill-down information Automatically generated reports, systems that notify users (e.g, e-mail)
Laboratory management	Basic: Intermediate: Advanced:	Tracks work requests Simple workflow organization Inventory, sample storage, and tracking systems Automated workflow scheduling and monitoring Automated decision-making, tracks revenue and costs Multi-site project management
System management	Basic: Intermediate: Advanced:	Disk backup and recovery, with some downtime allowed Redundant storage (eg, RAID) System is performance tuned Fault tolerant systems Dynamic performance tuning Advanced links to external communications

Types of DNA Sequencing Laboratories

Laboratories that sequence DNA can be divided in three main categories: production, research, and core service facilities. Although each type of laboratory performs similar activities, their objectives are different and their LIMS requirements will vary accordingly.

Production laboratories are large-scale facilities with 100 or more DNA sequencing instruments and production volumes of millions of sequences per year. Their primary goal is to collect data, not to analyze its biological significance. These facilities require highly advanced, usually customized, LIMS that are maintained by teams of computer professionals.

Research laboratories, on the other hand, are more concerned with analysis because they focus on the biological significance of DNA sequences. Information management is usually controlled by the individual researchers involved in a project. These laboratories typically sequence on a small scale, producing between 10,000 and 100,000 sequences annually. Their throughput capacity, however, is

continually improving along with the technology, and is likely to result in increased data management needs in the future.

Core facilities operate on a scale between research and production facilities, producing between 10,000 and 500,000 sequences annually. Like production laboratories, their emphasis is on collecting data. Unlike production laboratories, core facilities offer a variety of services to researchers in their host institutions or companies and operate as small businesses within institutions. Core facilities process a variety of samples in varying states from many different users. Although core facilities have a great need for LIMS, their business structure makes the long-term cost of a dedicated informatics support team, around \$250,000 per year, prohibitive. As molecular biologists utilize more high-throughput technologies, the services offered by core facilities will continue to increase, leading to a continued demand for information management. Core facilities, then, serve as an excellent model for examining the benefits offered by commercial LIMS.

LIMS Options for Core Facilities

Molecular biology core facilities provide three general services: DNA and peptide synthesis, management of shared equipment, and sample analysis. This discussion focuses on DNA sequencing, which is considered an analytical service.

The process of DNA sequencing in a core facility begins when a researcher sends DNA samples to the laboratory along with a work order describing the samples and listing the services that are needed. Samples may be submitted in a variety of states, such as tissue or blood, and may need a varying number of purification steps. The quantity of samples will also vary from one sample in a tube to thousands of samples in 96 or 384 well trays. The laboratory manager organizes work orders for processing by sample type and the type of service requested. Sequencing services can range from simply loading completed sequencing reactions on a gel to full service where DNA is purified, added to sequencing reactions, and loaded on gels or into capillaries. Once sequencing is completed, data are sent to the researcher who made the request. Finally, billing and cost accounting statements must be created for invoicing and budget accounting.

In many laboratories, data management is paper-based, with requests made on forms and stored in laboratory notebooks. Requests are organized for processing by re-entering information into new worksheets which are then retyped into forms and re-entered in spreadsheets required by data collection software. Data are collected, manually sorted on the computer, and delivered to researchers on a disk, as a paper printout, or over a network by email, through ftp sites, or via shared file systems. Paper-based data management is labor intensive and burdensome for any laboratory that processes more than a few

samples per day. Further, with a potential for error every time data is reentered, the probability of mix-ups is high.

Software that supports core facilities must allow researchers to enter information about their samples and select a service offered by the laboratory. The software needs to provide mechanisms to organize requests for sample processing and to track samples' progress through the laboratory. The software must capture and store the resulting data and distribute the data to researchers in a secure fashion. Value can be added if the software incorporates different types of analyses to allow assessment of the quality of laboratory operations or to prepare data for further analysis. Other programs such as Phred (Ewing and Green 1998; Ewing, Hillier et al. 1998), Phrap, CrossMatch (P. Green, unpublished), and BLAST (Altschul, Madden et al. 1997) may be integrated into the data processing pipeline to further expedite data analysis.

Buy vs Build

Once data volume exceeds a certain threshold, laboratories benefit from having a computerized system for managing operations and cost accounting. Laboratory managers must then decide whether to purchase a commercial system or build a customized LIMS.

Commercial vs Custom

There are a number of advantages to purchasing a commercial LIMS. A commercial system can be delivered within a few months, whereas custom systems require a long process of development and a significant time commitment on the part of laboratory personnel. Commercial systems eliminate the need to "reinvent the wheel," having completed the processes of design, review, and testing. Commercial systems are subjected to beta testing by multiple users, making these systems more robust and adaptable than computer programs written for personal use. Training, documentation, and support also tend to be better with commercial systems because companies that produce LIMS have more experience in training new users. The overall costs of commercial LIMS, therefore, tend to be lower than custom solutions in terms of development time and training time, and because they're designed for a larger market.

Despite the many advantages of a commercial system, custom solutions are the best choice in some situations. Commercial LIMS may not be available for rapidly changing technologies. Some laboratories may have highly specific requirements or may not want to adapt their procedures to meet the constraints of a commercial system. The higher cost of building a custom LIMS requires careful research and planning before a decision is made.

If a laboratory chooses to build a customized LIMS, the next decision is whether to contract a team of professional developers or hire personnel to develop the software internally. Planning and building a LIMS is analogous to building a custom home, with the software development team acting as the architect and construction crew. After hiring an architect, the first step in building a house is for the prospective homeowners to define their needs in terms of square footage, number of bathrooms, and other physical parameters. Before building a LIMS, laboratory personnel also need to specify their requirements, including what types of data will be captured, who will be allowed to access data, and numerous other performance criteria. They must list the functions that the LIMS needs to perform and communicate that information to their software development team.

In building a home, the architect prepares a blueprint based on the client's requirements. In building a LIMS, the software development team creates a blueprint, or functional specification, describing how the software will be built and used. Choices of data model, software architecture, hardware platforms,

and software tools are based on the requirements identified by the laboratory. The functional specification includes data flow diagrams, descriptions of data objects, and prototypes of user interfaces. Functional specifications or blueprints are then reviewed by the clients and further modified. During a LIMS construction project, unanticipated events may occur that require further planning and changes in the specifications.

The following sections in this report discuss points that need to be considered when planning a data management system. A case of study of Rocky Mountain Laboratories and their choice to purchase Geopiza's *Finch-Server* will be used to illustrate the decision process and the implementation of a LIMS.

Cost

Whether purchasing a LIMS, building one in-house, or outsourcing a custom project there are two types of expenses associated with a LIMS: acquisition and ownership. In addition, delays in full implementation of the system may have a significant impact on the competitiveness of a core facility.

Acquisition costs are the initial costs for the hardware, software, training and documentation needed. For a commercial LIMS these costs are straightforward to calculate and are supplied by a quote from the vendor. When a custom system is being developed, acquisition costs also include the planning time required of the laboratory personnel, building the software, testing it, and implementing it.

Ownership costs involve the long-term support of the software. For commercial systems these costs are factored into support contracts. Depending on the system however, ownership costs can also include personnel costs for individuals needed to maintain the hardware and software components such as the RDBMS. Other ownership costs include continuing education to learn how to use new components and training new staff.

LIMS costs are offset, however, by reducing expenses associated with maintaining a paper-based system and reducing errors. LIMS further lower sequencing costs by improving data entry, integrating quality control analyses, and streamlining data distribution to sequencing laboratory clients.

Depending on the product and the degree of support required, a core facility can expect to pay between \$40,000 and \$300,000 for commercial software, installation, and initial training. Service contracts can cost between \$8,000 and \$50,000 annually. Custom solutions, whether built in-house or outsourced, generally cost between \$200,000 and \$1,500,000, with ongoing maintenance costs from \$80,000 to \$250,000 annually.

Geospiza, Inc.

Geospiza, Inc. ("Geospiza"), founded in 1997, provides bioinformatics products (software and systems) and custom design services for corporate, academic, and government laboratories. Geospiza's systems are used as LIMS by production centers, research laboratories, and core facilities worldwide. Geospiza employs a multidisciplinary team with extensive experience in computer science, software engineering, and biological research, and a thorough understanding of the challenges of managing biological data.

Products

Geospiza's *Finch-Suite* of software components provides a complete, integrated data management solution for high-throughput production centers, research laboratories, and core service facilities.

Geospiza's *Finch-Servers* are data management systems composed of one or more software components built upon Geospiza's *Finch? -Core* software platform (Fig. 1). By combining software components, customers can purchase a *Finch-Server* that meets their specific needs.

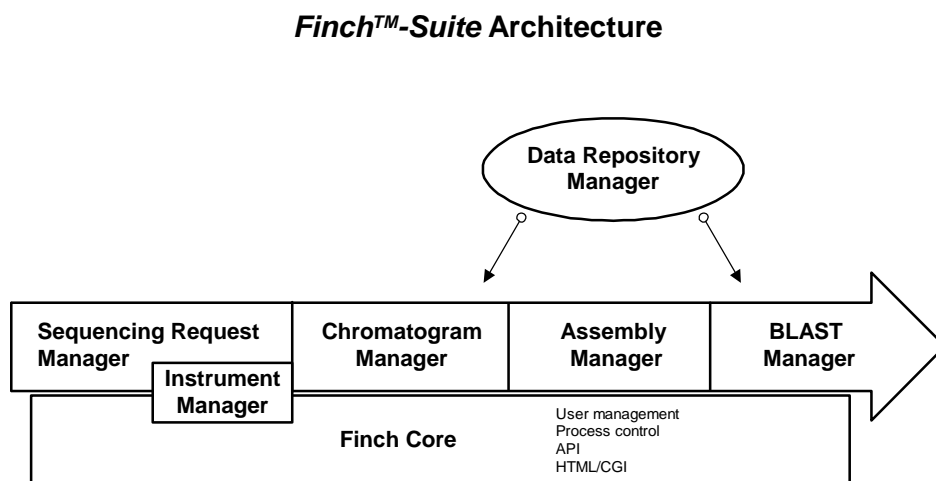


Figure 1. Organization of *Finch-Suite* components.

The *Finch-Core* is a scalable, client-server system designed to maintain all aspects of a LIMS. At the heart of the *Finch-Core* is a fully SQL compliant, embedded RDBMS (relational database management system) that is used to store sample information, sequence data, and analysis results. Access to the database is provided through a Web-based interface that allows secure access from any type of computer.

Layered on top of the *Finch-Core* are a number of integrated software components that cover specific domains in bioinformatics (Table 2).

Table 2. <i>Finch-Server</i> Components	
Software Component	Activity
<i>FinchTM-Sequencing Request Manager</i>	Provides online data entry and automates sample tracking
<i>FinchTM-Chromatogram Manager</i>	Manages chromatogram files and monitors data quality
<i>FinchTM-Instrument Manager</i>	Tracks information about sequencing instruments and sets instrument-specific parameters automatically
<i>FinchTM-Assembly Manager</i>	Assembles sequence data into contigs and automatically updates assembly sets
<i>FinchTM-BLAST Manager</i>	Compares query sequences to sequences in public or private databases and provides data in a variety of linked reports
<i>FinchTM-Data Repository Manager</i>	Retrieves sequences from data repositories and maintains secure local copies of continually changing databases such as GenBank

Services

In addition to commercial software products, Geospiza offers custom software development, scientific consulting services, and training.

Custom software is built by working with customers to identify requirements and write a functional specification for custom LIMS. The functional specification serves as the architectural blueprint and is used to prepare accurate cost estimates for software development.

Geospiza's scientific consulting encompasses bioinformatics planning, training, and custom report preparation. Organizations benefit from Geospiza's expertise in developing bioinformatics plans for preparing budgets and obtaining grant funding.

Specialized training services help customers take full advantage of Geospiza's systems. Geospiza also assists clients in using SQL to create custom data reports.

Rocky Mountain Laboratories

Rocky Mountain Laboratories (RML), located in Hamilton, Montana, is a national laboratory within the National Institute of Allergy and Infectious Disease (NIAID), one of the National Institutes of Health (NIH). RML has a core facility that serves 12 scientists, with a sequencing throughput of approximately 300,000 sequences per year.

RML needed a LIMS that could:

- Organize data by project (organism) and subproject (folders and different DNA libraries), allow for comments and annotations, and link sample information to raw chromatogram data

- Allow scientists to monitor their samples and project status and access the system through a variety of computer platforms

- Create both sample sheets and plate records for DNA sequencing instruments

- Process data to identify and mask vector, common repeats, and other sequences from various cloning artifacts and to allow laboratory personnel to monitor sequence quality on a daily basis

- Assemble overlapping sequences into contiguous sequence strings (contigs) and use individual reads and contigs to search public and private sequence databases with BLAST

- Meet RML's performance requirements for speed, storage capacity, and security

- Reformat and import approximately 70,000 "legacy" chromatogram files that had already been collected

Before choosing a LIMS, RML had to decide what the system would be expected to do in terms of workflow, scale and performance, data processing, and security. Decisions were made about the type of communications interfaces that would be needed, and how the system would be implemented, administered, and maintained. Because a LIMS changes the day to day operations of a laboratory and its customers, training, documentation, and support were critical aspects of RML's decision.

The next sections highlight issues that were of concern to RML and show how a Sun computer system, in conjunction with a *Finch-Server* that included Geospiza's entire *Finch-Suite* was used to meet RML's sequence data management needs.

Workflow

Workflow encompasses the day to day operations of the laboratory. Genome sequencing, expressed sequence tag (EST) projects, and confirmatory sequencing have different data collection, processing, and management needs.

RML's sequencing projects involve shotgun sequencing of bacterial genomes between 2 Mbp and 5 Mbp in length, studies of genetic variation, and clone confirmation. Since RML is engaged in multiple projects, they needed a LIMS capable of organizing data by project and subproject. They also desired a LIMS that would allow scientists to record information about their samples and automatically build sample names that follow defined naming conventions required by certain analysis software. The LIMS had to allow laboratory personnel to organize sequencing requests by type of sequencing project and prepare plate records and sample sheets for ABI PRISM[®] 3700 and 377 DNA sequencing instruments. RML's laboratory managers also wanted to track sequencing failures and monitor sequencing progress. Timely identification of equipment malfunctions and/or problems with reagents is important in running a core facility. It is also important to identify sequencing artifacts such as cloning contaminants and overrepresentation of vector sequences, since early identification of these problems minimizes the loss of time and money that results from sequencing poor quality samples.

RML's scientists use the Sequencing Request Manager to enter sample information and designate folders for storage of their completed sequences. It tracks request and sample status as samples move through the sequencing process, and allows users to obtain status information through Web browsers running on their local computers. Configurable parameters in the *Finch-Server* allow laboratory managers to define services, DNA type, and other sample attributes that are used to organize requests for processing. Finally, variable tags used in different sequencing request interfaces are used to automatically build filenames for different naming conventions.

In conjunction with the Instrument Manager, the Sequencing Request Manager is used to create sample sheets and plate records, special files used by sequencing data collection software to name and add information to chromatogram files. The Instrument Manager is also used by the laboratory to track maintenance and repairs to sequencing instruments.

Once data are collected and transferred to the *Finch-Server*, the Chromatogram Manager processes the data through a variety of data processing programs. Laboratory managers and technicians browse reports to view data quality and check for artifacts. Graphical reports integrated into the *Finch-Server* show sequencing quality over time grouping data by day, month, and year. Data quality can also be viewed by folder and links within these reports lead to drill-down views of the data. Scientists use the

Chromatogram Manager to view data quality for their projects and retrieve their data for additional analysis on local computers.

In addition to predefined reports, users also create specialized reports using the SQL interface. Through this feature, resource accounting and project statistics can be obtained. An added benefit of this system is the ability to correlate data quality with laboratory procedures, thus helping managers choose between different protocols.

Data Processing

DNA sequencing laboratories require a variety of specialized programs for data analysis. High-throughput shotgun sequence data are analyzed with the Phred basecalling algorithm to get sequence reads and their corresponding quality values. These data are combined by the Phrap algorithm to assemble reads obtained from overlapping clones. Phred provides a metric for quality assessment and allows for the use of full length reads in the assembly, increasing the amount of usable data. However, before data can be assembled, the short regions of vector that commonly precede each sequence must be removed (masked) so that they do not interfere with assembly.

Since RML's major focus is bacterial genome projects, they needed a LIMS that would be able to automatically process data, allowing them to observe sequence quality and assemble overlapping sequences into contiguous strings (contigs). Geospiza met this requirement by using SPS-Phrap (<http://www.spsoft.com>) to perform high-throughput assemblies in concert with the *Finch-Assembly Manager*.

These and other projects also require the ability to search public and private databases of DNA and protein sequences with the BLAST search algorithms. Geospiza's *Finch-BLAST Manager* integrates search capabilities into the *Finch-Server*, allowing researchers to query databases using sets of query sequences organized by folder, sequence ID, contigs, or through an SQL query. Once a search is complete, the results are stored in the RDBMS and are presented through Web pages as high level summaries containing links to drill-down views and alignments.

Scale and Performance

Storage issues must be considered along with plans to prevent accidental data loss by backing up data on a regular basis. RML required a data storage system able to store and backup up the 1,000,000 chromatogram files they anticipate generating in the next three years. The data storage system needed to be scalable, allowing additional storage and backup media can be added to the system over time. It was also important to minimize downtime in the event of a disk failure.

The performance of a LIMS is related to the volume of information, the quantity of data that need to be stored, the amount of data processing, and the speed required for data processing. These features impact both hardware and software design decisions. RML had to be able to manage an annual volume of 300,000 sequences and required a high-throughput data processing system able to accommodate sequence assemblies with up to 60,000 reads. The data processing time was also important. Since new data are assembled daily, assemblies must be completed within two or three hours.

In order to meet the above performance requirements, the *Finch-Server* was installed on a Sun E450 computer with four 450 MHz processors, 4 GB of RAM, over 200 GB of redundant disk storage and a tape backup system. The system was installed by Dynamic Systems, Inc. a Sun distributor located in Pasadena, California (<http://www.dyansys.com>). Since 1991, Dynamic Systems has sold and supported UNIX-based solutions into government agencies and commercial accounts.

Communication Between the LIMS, Sequencing Instruments, and Laboratory Computers

A LIMS must be able to interface with users, equipment, and other software programs. For sequencing laboratories, this may include a variety of sequencing instruments in addition to the computers used by the laboratories and the computers used by their customers.

The use of multiple types of sequencing instruments may necessitate the ability to capture data that have different file formats and change the formatting to accommodate analysis programs. For example, the Phrap algorithm (P. Green, unpublished) is used to assemble overlapping sequences from random sequencing projects. Phrap requires that sequencing chemistry be specified in either the name of the sequence or with a specific header in the file. Phrap also requires information about the orientation of the sequence relative to the template from which it was derived.

RML's customers needed to access data from different types of computers. Scientists at RML use platforms that include PC's running Windows and Macintosh computers. Although Geospiza's *Finch-Server* was developed to run on UNIX-based computer platforms, its Web-based interface allows for platform-independent access and operation. Thus, RML scientists can access data via any computer (Macintosh, PC, UNIX) that has a Web browser.

Training, Documentation, and Support

The ultimate success of the LIMS depends on high-quality training. Sequencing facilities have many different users with different objectives and training requirements. Customers and collaborators will use the system to make requests, view the status of their samples and requests, and retrieve data and

information from the LIMS. Personnel working in the sequencing laboratory will use the system to organize requests for processing, confirm that samples match requests, and track sample processing. Laboratory personnel will add raw data to the system and review quality reports. Laboratory managers will use the LIMS to view laboratory operations, account for costs, and issue invoices and billing statements. Administrators must be thoroughly trained in system operations and troubleshooting. Due to turnover in clients and laboratory personnel, core facilities have ongoing training needs.

Good documentation can minimize the frustration experienced by new users as they adapt to a new system, thus software companies devote considerable resources to providing documentation that is usable and appropriate for its intended audiences.

Geospiza sponsored a three-day training session at RML to introduce laboratory personnel and new users to the new LIMS. Training sessions provided opportunities for users to review and modify procedures in order to make the best use of the new LIMS. By the end of the training, laboratory personnel were able operate all facets of the *Finch-Server*. They could create sequencing requests, use requests to create instrument runs for the different sequencing instruments, and transfer chromatogram data to the server. In addition, small groups of scientists were trained on how to use the *Finch-Server* to create requests, and view and retrieve their data. Ongoing support is provided in form of technical support, online and printed user manuals, and continued options for on-site training.

Security

Core facilities have many clients, and data viewing must be restricted to appropriate users. However, sequencing laboratory personnel must have the ability to access all data in order to monitor the quality of laboratory processes. A LIMS must have the capacity to control access to different levels of system information.

The *Finch-Server* at RML supports 17 users (scientists and technicians) and four lab groups. Other core facilities have as many as 100 or more lab groups with over 300 users. The *Finch-Server* uses lab groups and user roles to control access to information. Roles are based on different functions a user may perform, ie, their roles in collecting and viewing data. For example, scientists are primarily customers of the core facility, thus they can make sequencing requests, view, and retrieve their data. Scientists can only see data and requests within their lab group. Technicians and lab managers, on the other hand, need access to more information and therefore have a wider range of permissions.

System Administration and Maintenance

After a LIMS is installed there must be a system in place to allow for repairs, upgrades, and specialized data manipulations. LIMS rely on external hardware and software products, including the computers' operating systems and data processing programs. Most LIMS operate on top of an RDBMS and use features of the RDBMS software to control system access and data processing. Software bugs can be uncovered long after the system is in daily use. In addition, programs used to process data are often refined and change on a frequent basis. On occasion, these changes negatively impact other components of the LIMS. Finally, even in the most automated laboratory environment, sample processing and data tracking errors can occur, requiring that data be manipulated within the RDBMS by a database administrator.

Geospiza offers a range of support contracts to accommodate this needed systems and database administration. With RML, Geospiza uses secure shell connections to install new software components and other updates via remote connections. Geospiza also monitors the database and database backups to ensure that the system is running properly. Providing these services through remote connections reduces travel expenses and reduces the customer's need for in-house experts.

Additional systems administration support can be provided by companies such as Dynamic Systems. As an expert in installing and operating Sun computers, Dynamic Systems offers a range of services including operating system tuning, planning tape backup strategies and their implementation, and system monitoring and alerting. Dynamic Systems also offers general systems administration (adding and modifying user accounts, adding peripherals, and adjusting system parameters) and detailed reporting of system events through weekly email reports and monthly teleconferences. In laboratories, like RML, where a single system is installed for a specific purpose there is not enough work to warrant a full time system administrator. Therefore, support packages that offer remote access services are worth considering when planning the LIMS.

Summary

The increasing use of high-throughput technologies is changing the way that molecular biology is performed. As laboratories generate larger volumes of data and automate more processes, they must devote more resources to systems for information management. LIMS requirements can be extensive and vary considerably among laboratories. Table 3 summarizes key aspects and a fraction of the questions that need to be considered when planning a LIMS. As Table 3 suggests, LIMS development presents many challenges and may take years to complete. Many labs are not adequately prepared to tackle the challenges of building a custom LIMS solution and most labs do not need a fully customized solution.

RML was chosen for this case study because it is representative of core facilities that need LIMS that can operate in high-throughput environments. In addition to laboratories like RML, the *Finch-Server* also supports laboratories that have less extensive data processing needs, but greater user management needs. In some labs the *Finch-Server* is used by more than 60 scientists who make sequencing requests, view, and retrieve their data on a daily basis.

The *Finch-Server* installed at RML is used for bacterial genome sequencing, genotyping, and other sequencing applications. RML needed a system that could support sample information tracking and create plate records and sample sheets for ABI PRISM[®] 3700 and 377 DNA sequencing instruments. In addition, the system was required to automatically assemble large data sets of overlapping sequences and provide reports monitoring sequencing progress and overall data quality. SPS-Phrap was installed in the *Finch-Assembly Manager* to perform sequence assembly on data sets of up to 60,000 sequences within a short time period. Throughput and data analysis needs were met by installing all *Finch-Suite* components on a Sun E450 computer system with four 450 MHz processors and 4 GB of RAM and 200 GB of disk storage.

Geospiza's *Finch-Suite*, while designed for core facilities, is well suited for different DNA sequencing laboratory environments. It accommodates a wide range of data production scales and operates on computer systems, like Sun, that are designed to support large amounts of data and high-throughput data processing. Customers who purchase a *Finch-Server* benefit from Geospiza's extensive skills in laboratory management, bioinformatics, and professional software development. They also benefit from the experience Geospiza continues to gain from a strong and diverse user base that includes customers in production, research, and core facilities.

Table 3. Key Considerations When Planning a LIMS for DNA Sequencing.	
Workflow Logistics	
Workflow	Which steps of the process must be tracked?
Sample management	What types of samples will the laboratory handle? How are they named?
Identifying problems	How are sequencing failures identified and documented?
Inventory and ordering	Will reagent and disposable plasticware inventories be tracked?
Cost accounting/Billing	What types of accounting reports are needed?
Scale and Performance Issues	
Data volume	How much data is expected and what is the anticipated growth rate for data acquisition?
Data processing time	Can data processing keep pace with data production?
Response times	How long will users wait for summary reports or data entry forms to be returned from the system?
Network	Can the network adequately support data transfers?
Data Considerations	
Acquisition	What types of data must be captured? Are there legacy data?
Data entry	How will sample information enter the system?
Analysis	Do sequences need to be assembled or checked for contamination? Will sequences be searched against databases with BLAST?
Delivery	How are data delivered to clients?
Storage and backup	How long is client data retained on the server?
Communication Issues	
Equipment	Does equipment (eg, robots, sequencing instruments) need to communicate directly with the LIMS?
User interfaces	How will users interact with the system?
Third-party software	Is additional software needed to support the system?
Training, Documentation, and Support	
Tutorials	Are online tutorials or other materials available?
Laboratory personnel	Will seminars or practical hands-on work sessions be provided?
Continuing education	What follow-up training will be provided?
System documentation	Does the LIMS have adequate system, installation and maintenance instructions?
Technical support	What types of technical support are required?
User documentation	Are user manuals available for clients and collaborators?
Security	
Data access	How will the LIMS control access to data and information?
Network/computer	Is the computer secure from unwanted access?
System permissions	What mechanism is used to control which users are allowed to read, write, or delete data from the LIMS?
Administration and Maintenance	
System administration	How will overall administration of the system be coordinated?
Database administration	How frequently does the RDBMS require monitoring and performance tuning?
Network administration	How does laboratory equipment communicate with the LIMS?
Software maintenance	How will LIMS software and third party updates be carried out?
Hardware maintenance	How will disks and backup systems be maintained? Will backups be verified?

References

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs." *Nucleic Acids Res.* **25**(17): 3389-3402.

Ewing, B. and P. Green (1998). "Basecalling of automated sequencer traces using phred. II. Error probabilities." *Genome Res.* **8**: 186-194.

Ewing, B., L. Hillier, et al. (1998). "Basecalling of automated sequencer traces using phred. I. Accuracy assessment." *Genome Res.* **8**: 175-185.

Liscouski, J. (1995). *Laboratory and Scientific Computing a Strategic Approach*. New York, John Wiley and Sons, Inc.