

Robotic and Information Systems for Large Scale Population Genetic Studies: The North Shore LIJ Biorepository and the New York Cancer Project

Peter K. Gregersen and Robert Lundsten

Summary

This project grew out of the need for automated systems to enable the collection, preparation and storage of DNA and other blood derived elements from several hundred thousand human subjects. These subjects are being entered into a large longitudinal cohort study, the New York Cancer Project. Over the last 4 years, the North Shore LIJ Biorepository has designed and implemented robotic and information systems to support this project. The system is currently capable of automated storage and retrieval of approximately 250,000 specimens of genomic DNA with precise control of volume, concentration, and specimen tracking, in a 96 well format that is suitable for high throughput genotyping. The operations of the NS-LIJ Biorepository implement the latest technologies available for specimen processing and data management and is organized around various integrated systems, including Lab Processing Systems, Storage Systems, Lab Processing Systems, Inventory Systems, and Information Management Systems. A flexible modular robotics platform provides for dynamic robot protocol development and deployment to allow modifications in response to changing laboratory requirements. The system consists of the following components: a CRS robotic track arm which moves specimens to different processes along its table, a TECAN Genesis liquid handling workstation, a TECAN SpectraFluor Plus plate reader for specimen analysis via Fluorescence or UV spectroscopy, a CRS Storage Carousel, an Abgene ALPS100 plate sealer, a TECAN / GIRA MOLBANK, and other ancillary devices both purchased and custom designed. A Supervisory Control And Data Acquisition (SCADA) System provides freezer monitoring and data-logging services for all freezers at the NSHS – LIJ Biorepository, the MOLBANK, and the environments in which they are maintained. In the event of a freezer or environment discrepancy / failure, the system is capable of several programmable actions including, automatic voice phone alert system, pager alert system, fax alert system, and other application program launch via its connected computer and network. The principle components of the Information Management System utilizes a A Dell Power Edge Server - Redundant Independent Array Disks (RAID) system. Both a RAID 1 (disk mirroring) and RAID 5 (disk striping) disk management system have been setup to create disk redundancy in order to prevent data loss due to disk failure. The system also features a triple power supply and is powered by separate UPS systems and back-up power. A Microsoft SQL enterprise database receives data from robots, instruments, and laboratory technicians. Custom user applications and instrument interfaces are constructed from high-level languages “C”, “Visual Basic”, “RAPL-3”, and the Microsoft Visual Studio development suite. Microsoft IIS web server and user web services interfacing are currently being implemented in order to provide customized, web-based data access solutions for a variety of users.

Introduction

Ongoing developments in human genetics have focused attention on the requirement for large population datasets in order to investigate the role of genetic polymorphisms on disease susceptibility, drug responses, and normal phenotypic variation in the human population. In order to address this need for cancer and other phenotypes, a longitudinal cohort study was initiated in the New York metropolitan area. This study, the New York Cancer Project, was sponsored by the Academic Medicine Development Corporation (AMDeC). AMDeC is a consortium of all the major biomedical institutions in New York. The initial funding was provided by the City of New York under the leadership of Mayor Rudolf Giuliani. The recruitment of 20,000 subjects into the pilot phase of the New York Cancer Project was completed in the middle of 2002. The New York Cancer Project is specifically designed to include the collection of blood at enrollment, with a view to creating a Biorepository of genomic DNA on all study subjects. Therefore, we designed the North Shore LIJ Biorepository to accommodate the receiving, processing, aliquoting, storage and retrieval of these DNA specimens [1]. The completion of the human genome, the planned development of a human haplotype map, and the rapid advancements in high throughput genotyping are now expanding the need for reliable methods for storing and distributing large numbers of DNA samples. Therefore, we placed special emphasis on automation of sample retrieval, to allow for flexible ongoing access to these DNA specimens for both small and large-scale genetic studies.

Description, Methods and Materials

The North Shore Long Island Jewish Health System Biorepository (NSLIJ Biorepository) and any repository can be described by four main operations – Input, Processing, Storage, and Output. The collection of biological specimens, processing of biological specimens, storage and protection of specimens, and the distribution of specimens, make up the physical implementation of a Biorepository. In addition, a parallel system of information input, processing, storage, and output, closely parallels the physical implementation and is performed by computer hardware and software systems. Therefore the functions of the NSLIJ Biorepository are two simultaneously occurring processes – physical manipulation of specimens and management of information about those specimens. The methods and materials used to construct the NSLIJ Biorepository will be broken down into four main topics as described in the above paragraph.

Collection of Biological Specimens

The first of the four main areas, collection of specimens by the NSLIJ Biorepository and collection of data related to specimens, is performed at remote collection sites around the New York City area. Human blood specimens are collected from subjects using standard blood collection tubes (Becton Dickenson) and standard blood collection techniques (professional phlebotomists). Six tubes (ten milliliter each) are collected from each subject from remote sites around the New York City area. In addition to the blood collection, data is collected via IRB approved questionnaire submitted to and completed by the subject. Data is collected about subjects/specimens using portable laptop computers and automated graphical user interfaces (GUI), which enforce pre-determined data collection rules. Software was developed using Borland Delphi GUI development tools. Data related to each

specimen is then downloaded to a secure Oracle database running on a Unix OS. Family history, demographic information, drug usage, ethnic background, and other disease relevant data are collected from subjects. Specimens collected are linked to subject data via barcode labels designed at the NSLIJ Biorepository and produced by Digi-Trax Corporation; Buffalo Grove, IL Specimens that arrive at the Biorepository are accessioned into a Microsoft SQL 2000 enterprise database upon arrival using custom GUI programs developed in MS Visual Basic. The accessioning client software is part of the NSLIJ Biorepository Laboratory Information Management System (Biorep LIMS). The Biorep LIMS governs all of the manual processing in the lab and is integrated with Biofile, which is the robotic governing software. It also scripts to Zebra bar code printers directly producing custom labels on demand while simultaneously time-stamping these events in the SQL database.

Processing of Biological Specimens

The processing operations of the Biorepository are performed both robotically and manually. The main function of the NSLIJ Biorepository is to extract and purify DNA from a large number (200 specimens per day) of human whole blood specimens. A method was developed from previously described aqueous extraction methods [2,3,4] and scaled up to meet the needs of this project (1-2 mg of DNA from 50-55 ml whole blood). Once the chemistries and procedures were tested and defined, reagents were made in bulk at the Biorepository and others were purchased from Roche Molecular (bulk agreement) to meet our large-scale processing requirements.

Since the functions of the processing operations make up the most important and largest area, our focus was to automate as much of the functions here as possible limited by expense or time consumption. We grouped the major tasks of the processing operation into two main functions; 'processing of specimens' and 'formatting and analysis'. Processing of specimens - which includes removal of specimens from shipping containers, movement to storage directly for some specimen types, processing (DNA purification), and preparation prior to formatting of specimens – was developed as non-robotic operations but automated using computer programs and other equipment developed to drive the laboratory technicians actions. Formatting and analysis – which includes plating of specimens in a 96 well plates designed to be handled by robotic systems, spectral and fluorescence analysis of specimens, and data gathering – was developed completely using robotic devices.

Semi-Automation of the NSLIJ Biorepository (non-robotic processing operations)

Programs developed using MS Visual Basic drive all specimen and DNA purification processes performed by laboratory technicians. Data entry is primarily performed by bar code scanning equipment from Symbol Technologies or by equipment linked via serial ports reducing keyboard errors. Specimen barcode labels for tracking are generated on the fly in response to specimen label scanning using Zebra barcode printers (Zebra Technologies). Custom scripts were developed in MS Visual Basic that sends ZPL (Zebra Programming Language) commands to barcode printers. Error handling source code was developed for these operations which checks all data entry against predetermined set of

rules and alerts the technician when an inappropriate value is entered requiring additional corrective action. The VB client software writes all data to the Microsoft SQL 2000 enterprise database, which runs on Microsoft Windows 2000 Server. The data is automatically backed-up using snapshot replication – that is the database is duplicated (snapshot) locally and then sent to three other server machines (replication) also running SQL Server one of which is the Web Server machine. This is performed every 2 hours. Process data is also protected by tape backup and a Redundant Array of Inexpensive Disk (RAID) system. This system implements a RAID 1 (mirroring of disks) and RAID 5 (disk striping) configuration on a Dell Power Edge Server.

Other major processing equipment includes; automatic dispensing equipment purchased from BrandTech Technologies and Nunc/Nalgene (pump dispensers and tubing), six high-speed centrifuges from Sorvall, and four automatic heated shaking water baths from VWR Scientific.

Automation of the NSLIJ Biorepository (robotic formatting, analysis, and storing operations)

The NSHS –LIJ Biorepository with the Medical Automation Research Center (MARC) @ <http://marc.med.virginia.edu> developed a flexible modular robotics platform that performs specimen processing in an efficient and error-free manner [6]. The system allows dynamic robot protocol development and deployment to allow changes in response to changing laboratory requirements. The system consists of the following components, Biofile custom scheduling software (MARC@UVA), CRS robotic track arm on a 4 meter track which schedules to move specimens to different processes along its 5 meter by 2 meter table, TECAN Genesis liquid handling workstation, TECAN SpectraFluor Plus plate reader for specimen analysis via Fluorescence or UV spectroscopy, CRS Storage Carousel, Abgene ALPS100 plate sealer, TECAN / GIRA MOLBANK, and other ancillary devices both purchased and custom developed. This system is capable of handling any biological fluid for aliquoting, storage, and robotic retrieval. In the case of DNA, automated measurement and adjustment of concentration with flexible aliquoting into 96 deep well storage formats is done in an error free, fully bar coded fashion. These “master plates” of high concentration DNA are archived in a –30 environment until needed to make “daughter plates”. In addition, replicate “daughter plates” are produced for short-intermediate term storage in a 4C environment in the MOLBANK. This system allows for completely robotic retrieval of samples by “cherry picking” from up to 250,000 separate DNA samples that can be housed in the MOLBANK. This process is fully automated, with full back up and computer monitoring of all events. A movie of this sample retrieval and send-out process can be viewed at www.biorep.org. or at the MARC web site http://marc.med.virginia.edu/videos/pro_biorep.html

Storage and Protection of Specimens

Storage Systems of the Biorepository provide short-term accessible and long-term archival storage of all specimens processed. The following components make up the Storage Systems of the Biorepository.

TECAN / GIRA MOLBANK is an automated storage and retrieval system that is linked to the Biorepository track robot arm. The MOLBANK is a robotic freezer and specimen management system that is incorporated into the robotic operations at the NSHS - LIJ Biorepository. It holds 2574 storage plates each containing 96 DNA specimens for a total system capacity of 247,104 specimens. The Biofile software communicates with the MOLBANK robotic system producing physical movement of bar coded specimens between instruments and communication of all tracking data to the linked SQL Server database. The MOLBANK system provides accurate hands-off processing of specimen storage and retrieval operations.

Revco (-80 degrees), Revco (-30 degrees), Walk-in (-20 degrees) freezers are externally alarmed and temperature monitored. Freezers are maintained in an air-conditioned environment, which is also monitored to provide the most optimal freezer-operating environment. Freezers have been fitted with efficient space utilizing inventory rack systems to permit easy storage and retrieval of bar coded specimens. Freezer power requirements are backed up by emergency generator systems.

Supervisory Control And Data Acquisition (SCADA) System provides freezer monitoring and data-logging services for all freezers at the NSHS – LIJ Biorepository, the MOLBANK, and the environments in which they are maintained. Data-logging is performed every 10 mins and is recorded in the database providing report capabilities. In the event of a freezer or environment discrepancy / failure, the system is capable of several programmable actions including, automatic voice phone alert system, pager alert system, fax alert system, and other application program launch via its connected computer and network. The power required for this system is backed-up by emergency battery power.

Emergency diesel generator back-up system supports all of the mission critical operations of the Biorepository. In addition, the Biorepository systems are backed-up by UPS / Battery or diesel generator power and in some cases both.

Inventory Systems - Rack and box systems for commonly used specimen formats have been developed in order to efficiently make use of freezer space. Bar code labeling systems for tracking all specimens throughout the lab are labeled and tracked using bar code technologies. The Biorepository maintains and programs Zebra bar code printers and Symbol bar code laser scanners.

Distribution of Specimens

The distribution system is a composition of previously described components. Based on request search criteria, a list of specimens to be sent to a requestor is generated by the Oracle database (holds the subject data separately). The request list is sent to the MSSQL database and the Biorep LIMS automatically manages and presents this list to Biofile. When activated by a lab operator, Biofile automatically directs the hit-picking operation performed by the robotic platform. This hit-picking operation can process up 3000 sendout specimens per week in a non-stop fashion. VB source code and SQL stored procedures

govern the choice and order of specimens that the system chooses in order to maximize efficiency of the system. This is a complex set of algorithms looking at specimen availability, volume remaining, current concentration as compared with request concentration, specimen status, organization of plate retrievals from the MOLBANK based on the number of hits existing in that plate, and many more. Upon completion of the sendout operation, 96 well plates are presented to the operator to be distributed to the requestor. Biofile generates paper reports using Crystal Reports to accompany the sendout plates. The requestor also has the option to view data about specimens via www.biorep.org. Microsoft Internet Information Server running on a Compaq Proliant Server manages the NSLIJ Biorepository web site. Data is moved to a SQL Server database on this machine via replication and is updated every 2 hours.

Results and Discussion

The NSLIJ Biorepository project is made up of hardware and software that provides the user of this resource with an easy to use system for requesting and receiving ready to use DNA and other specimens stored at the Biorepository. All of the automated operations of the NSLIJ Biorepository remove time-consuming and expensive processes from a research project. A typical user only supplies specimen search criteria and receives specimens and data at the other end. Although the NSLIJ Biorepository is still in the design and construction phase of development – it is currently operational. We have collected, processed, and stored over 70,000 DNA, RNA, plasma, and cell specimens combined in the past two years and have distributed approximately 5,000 DNA or plasma specimens more recently. The NSLIJ Biorepository is now providing snap-in support to a number of different research projects as it was intended. Much the same way the individual modular components of the Biorepository makeup the Biorepository as a whole, the Biorepository now provides modular support to research projects in an automated manner. The NSLIJ Biorepository provides user-friendly front-end interfacing via electronic specimen requests and back-end robotically friendly interfacing via standard robotic specimen plates and web interfacing for electronic data transfer data at www.biorep.org. The components that make up the operations of the NSLIJ Biorepository have proven useful in reducing technician specimen handling errors. Labeling or mishandling of specimens leading to the non-useable specimens is less than a 1 percent error rate. When considering the large number of specimens handled at this facility this fact then becomes significant. We believe it is the most sophisticated robotic biorepository currently in existence in an academic or hospital setting.

Acknowledgements

The authors would like to acknowledge the contributions of the Medical Automation Research Center (MARC), including; B. Sean Graves, Theodore A. Mifflin, Robin A. Felder, Steve Kell, Jim Gunderson, Chris Estey, Sarah Geddy, Catherine Piche, Greg Wasson, and Kevin Bowman. The authors would also like to acknowledge the development team at North Shore – LIJ Health System including; Houman Khalili, Jubal Dias, and Rodney Coe.

References

- [1] Gregersen, P.K., Felder, R.A. (2000) Searching for Gene-Environment Interactions in Cancer. *JALA*. 5(5):37-39
- [2] S. A. Miller, Dykes, D. D., and H. F. Polesky . (1988) " A simple salting out procedure for extracting DNA from human nucleated cells" *Nucleic Acids Research* 16: 1215.
- [3] Sambrook, J, Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* 2nd. ed. Vol. 2. Cold Spring Harbor Laboratory Press, New York. E10-E14.
- [4] Hofstetter, J.R., Zhang, A., Mayeda, A.R., Guscar, T., Nurnberger J.I., and Lahiri, DK (1997) Genomic DNA from Mice: A Comparison of Recovery Methods and Tissue Sources. *Biochem Mol Med Dec*; 62(2):197-202
- [5] Lahiri DK, Bye S, Nurnberger JI Jr, Hodes ME, Crisp M. (1992) A non-organic and non-enzymatic extraction method gives higher yields of genomic DNA from whole-blood samples than do nine other methods tested. *J Biochem Biophys Methods* 1992;25:193-205.
- [6] Graves, B.S., Mifflin, T.A. (2001) A Biological Repository for Human Genetic Material: Overview and Software Implementation *JALA* 5:6 (2001) pp. 106-108.