Queensland Centre for Medical Genomics

SGI Solutions Help Scientists Accelerate Cancer Research

Key Facts

Organization:
Queensland Centre for Medical Genomics

Primary Location:
Brisbane, Australia

Application:
Data management

The University of Queensland's Institute for Molecular Bioscience (IMB) is internationally recognized as a leading centre for molecular bioscience research. Established in 2000, it is located in the Queensland Bioscience Precinct.

The IMB is a multi-disciplinary research institute with 500 research staff and students who focus on a range of strategic programs in mammalian systems biology, and are supported by some of the finest facilities in the world.

The major focus of IMB research is to improve human health with the development of new pharmaceuticals, cell therapies and diagnostics through the understanding of information contained in the genes, proteins and molecules of plants and animals.

Business Challenge

The Queensland Centre for Medical Genomics (QCMG) within the IMB has a mission to sequence and characterize pancreatic and ovarian cancer to an unprecedented depth, enabling a more complete understanding of the mechanisms that lead to genetic instability and ultimately, cancer itself.

The QCMG is a member of the International Cancer Genome Consortium, whose members will together sequence the genetic codes of 25,000 tumours from 50 different types of cancer.

Scientists at the QCMG are specifically studying pancreatic and ovarian tumours, two of the most common causes of cancer death in the developed world. Pancreatic cancer has a 95% mortality rate within 12 months of diagnosis. Ovarian cancer, while less deadly in its primary form, currently has no screening test and is usually not discovered until it has spread, making treatment difficult.

QCMG has 11 Life Technology ABI SOLiD V4 genome sequencers and a Life Technology NGS 5500xL producing 0.5TB of summarized data per week per sequencer (6TB per week), with the volume doubling in early 2012 with upgraded sequencer technology.

The data has to be catalogued with metadata from the scanners and LIMS systems, and then routed across networks to the HPC clusters and storage for processing and analysis. Managing the volume of data and workflow are major practical challenges for the QCMG.

> "We need to keep the operations side of the QCMG lean so we can concentrate on research, and that means automating as many of our workflows as possible. That's where we're using LiveArc[1]. To complete an analysis, we have to manage sequencing, storage and computational resources, as well as move raw and derived datasets from resource to resource."

John Pearson
Senior Bioinformatics Manager at QCMG

Technology Solution

QCMG researchers use SGI LiveArc digital content management software to handle the data, metadata and workflow processes in their processing pipeline. LiveArc provides the supporting framework for automated data management and storage through each step of the QCMG workflow. The key design principles for the solution are to:

- Manage large amounts of raw sequence data, and data from secondary and tertiary analyses, with automated data and metadata handling, repository dispatch and lifecycle curation functions.
- Automate processing of the data and collection of the results from the HPC clusters.
- Transparently move data between scratch storage and long term storage on the Hierarchical Storage Management system.
- Provide a framework for automation of other regularly repeated analysis processes.
- Ensure that the infrastructure provides flexibility for scientific endeavours without constraints from rigid processes, permitting the evolution of schemas, processes and system interactions.

SGI LiveArc is specifically designed to allow for faster data discovery, one step sharing and close collaboration, enabling efficient data management and reduced infrastructure costs. As data volumes continue to grow, so do the challenges for IT managers to cost effectively provide higher levels of utilization across different workflow and data types.

LiveArc manages the QCMG metadata and data throughout its entire lifecycle and is responsible for:

- Ingesting data and metadata from the genome sequencers and Laboratory Information Management System (LIMS) for the wet lab
- Replication of metadata and data from the local data store to a highly available long term tiered data store using SGI DMF™ tiered storage virtualization solution with geographically distributed tape libraries
- Sophisticated search facilities to allow researchers to select the data required for transformation and analysis
- Automatic creation of workflow jobs, moving data to high speed scratch storage for processing on the SGI high performance computing (HPC) cluster system
- Re-ingesting the resulting secondary and tertiary data and metadata from the HPC analysis process back into the LiveArc managed data repository

SGI LiveArc is built on an extensible service-oriented architecture (SOA) framework, providing a digital content management system with version and revision control along with workflow management and web serving in a single package. LiveArc simplifies the data administration and management while providing the interfaces and tools to connect with other applications (such as the LIMS) to collect the whole lifecycle of metadata. LiveArc provides a common interface across multiple application systems and typically incompatible data silos and metadata schemas without the need for major applications development or infrastructure change.

At its core, LiveArc provides a comprehensive platform to rapidly develop and deploy workflow-specific digital content management solutions. Applications interface with LiveArc through industry-standard protocols. It is platform neutral and can be deployed on laptops, desktops and enterprise-class servers using standard operating and file systems.

LiveArc uses a very efficient binary XML object database to enable orders of magnitude better performance than relational and other databases, while using a much smaller memory and database footprint. Other benefits include near zero administration while retaining open standards data portability.

LiveArc is closely integrated with the SGI DMF tiered storage virtualization solution, providing seamless online visibility to multiple types of disk and tape based storage pools. LiveArc can replicate data across multiple sites, and provides a parallel federated search and data administration capability across multiple repositories. This results in a system that is highly scalable as both data volume and file count grow.

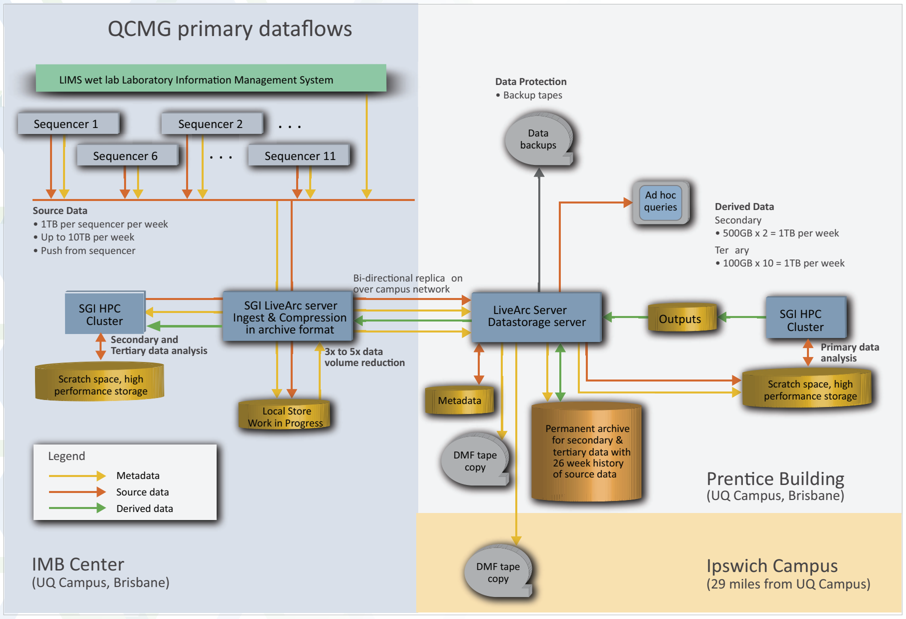
Business Results

The fundamental goal of the project is to use the combined resources of laboratories worldwide to create a map of the genetic changes that lead to cancer. The genetic map will act as a large information resource for medical researchers worldwide, allowing for more rapid and personal treatments for cancer sufferers in the future.

QCMG staffers have fully automated over 80% of the data and metadata capture, analysis, processing and long term archival requirements from the LIMS all the way to HPC processing and archival storage. Quality data and metadata from the QCMG research is made available to researchers around the world to further advance the global understanding of cancer and progression to improved treatment options.

> “LiveArc allows us to automate work flows and verify data transfers and archiving, to query our LIMS to determine the appropriate type of analysis based on the run type, and then trigger our cluster-based analysis tools and generate reports. LiveArc is the glue that holds our analytical pipeline together,”

John Pearson
Senior Bioinformatics Manager at QCMG


QCMG primary dataflows
LIMS wet lab Laboratory Information Management System
Sequencer 1
Sequencer 2
. . .
Sequencer 6
. . .
Sequencer 11
Source Data
• 1TB per sequencer per week
• Up to 10TB per week
• Push from sequencer
Bi-directional replica on over campus network
SGI HPC Cluster
SGI LiveArc server Ingest & Compression in archive format
Secondary and Tertiary data analysis
3x to 5x data volume reduction
Scratch space, high performance storage
Local Store Work in Progress
Legend
Metadata
Source data
Derived data
IMB Center
(UQ Campus, Brisbane)
Data Protection
• Backup tapes
Data backups
Ad hoc queries
Derived Data
Secondary
• 500GB x 2 = 1TB per week
Ter ary
• 100GB x 10 = 1TB per week
LiveArc Server Datastorage server
Outputs
SGI HPC Cluster
Primary data analysis
Scratch space, high performance storage
Metadata
DMF tape copy
Permanent archive for secondary & tertiary data with 26 week history of source data
Prentice Building
(UQ Campus, Brisbane)
DMF tape copy
Ipswich Campus
(29 miles from UQ Campus)

Each sequencer run generates around 10,000 data files organized in large directory trees. These files are automatically compressed into LiveArc archive files and stored in the system as single files. The single archive file holding the many data files is easy to manage, reducing communication transfer times between sites, as well as processing load on the SGI DMF systems, storage space and cost.

Even though many files produced by the sequencers are already in a compressed binary (BAM) format, the LiveArc compressed archive format further reduces the storage requirements by a factor of between three to five times.

There are over 3,000 managed assets in the QCMG LiveArc genome data repository, including 1,300 compressed archive assets with total managed data of 80TB.

[1] SGI LiveArc is the result of a partnership between SGI and Arcitecta Pty Ltd of Australia. SGI has deployed LiveArc in media, science and research institutions, government and academia in the Asia Pacific region under the name Mediaflux™. LiveArcsolutions are designed and delivered worldwide by SGI. Mediaflux is a trade-mark of Arcitecta Pty Limited in Australia.

The University of Queensland is the leading Australian Life Science research institute as determined by the World University Rankings, 2011-2012.

SGI Solutions

- SGI® LiveArc™
- SGI® DMF
- SGI® InfiniteStorage
- SGI® Altix® XE HPC clusters
- SGI® Rackable™ servers and HPC cluster
- Geographically distributed tape libraries

SGI Products for Life Sciences

- SGI® LiveArcTM: Workflow and data management
- SGI® DMF: Tiered storage virtualization
- SGI® ArcFiniti: Integrated data archive solution
- SGI® InfiniteStorage RAID, MAID and Tape platforms
- SGI® RackableTM rackmount and SGI® ICE blade-based HPC clusters
- SGI® UV: Shared memory system
- SGI® HTC (High Throughput Computing) Bioinformatics Wrapper

SGI® LiveArc™ Turning Data into Knowledge

- Multi-tiered storage
- Fast free-text search
- Drag and drop files
- Automatic metadata extraction
- Arbitrary collections
- Federated search
- Collaboration tools
- Version control
- Policy driven replication
- Data validation
- User defined dictionaries
- Security classes
- Roles and access control lists
- Audit trails
- Geospatial search
- Automatic proxy generation
- Integrated file manager
- Application launcher
- Compressed downloads
- XSLT translations
- Workflow automation
- DICOM support
- Hierarchical classifications
- Integrated with SGI DMF
- POSIX File System Interface

Global Sales and Support: sgi.com/global