Integrated Rule-based Data Management System for Genome Sequencing Data

A Research Data Management (RDM) "Green Shoots" Pilots

Project Report by Michael Mueller, Simon Burbidge, Steven Lawlor and Jorge Ferrer

Imperial College London

This project was funded as part of Imperial College's RDM "Green Shoots" Programme. In 2014, the Vice-Provost, Research, approved an allocation of £100K for academically-driven projects to identify and generate exemplars of best practice in RDM, specifically frameworks and prototypes that would comply with key funder RDM policies and the College position. The call for projects outlined that frameworks could be based either on original ideas or integrating existing solutions into the research process, improving its efficacy or the breadth of its usage. There was an expectation that solutions would support open access for data; solutions that supported Open Innovation were strongly encouraged.

Six projects were funded, covering different disciplines, faculties and research areas. The projects ran for six months, finishing at the end of 2014. Project reports were made available in 2015.

For more information on the programme and projects please visit:

http://www3.imperial.ac.uk/researchsupport/rdm/policy/greenshoots

Imperial College London

Integrated Rule-based Data Management System for Genome Sequencing Data

Michael Mueller, Simon Burbidge, Steven Lawlor, Jorge Ferrer; Imperial College London

Background

The advent and rapid evolution of next-generation DNA sequencing (NGS) technologies has revolutionised the field of genome research and has created new opportunities for the application of genomics in biomedical research and clinical practice. The ability to generate large amounts of information about individual genome sequences in a short amount of time at continuously falling costs has led to a widespread adoption of NGS technologies. Consequently, research institutes are now faced with increasingly large volumes of sequencing and associated data requiring more sophisticated approaches to data storage and management then commonly in place to date in order to ensure data integrity and security, avoid unnecessary data replication and facilitate data access for analysis and sharing.

While the implementation of large-scale data management systems is a challenge biomedical research has been confronted with only recently, other data-intense domains such as physics have been developing solutions to problems associated with the storage, management and sharing of large data collections for more than two decades. More recently, the genomics research community has started to adopt existing technologies that originated in these fields. Leading this trend are large-scale sequencing centres such as the Wellcome Trust Sanger Institute in the UK or the Broad Institute in the US, where data grid middleware such as the integrated Rule-Oriented Data management System (iRODS) has been successfully used for storing and managing large distributed sequencing datasets and associated metadata.

Requirements

The newly established DNA sequencing service at the NIHR Imperial BRC Genomics Facility will generate large amounts of genome sequencing data for users across Imperial College. The Illumina HiSeq2500 DNA sequencer installed in the Facility will output around 10TB of raw data every two weeks amounting to a total data throughput of 260TB per year. Manual processing, analysis and dissemination of sequencing data make data management time consuming, pose risks to data integrity and may result in unnecessary data replication. To address these issues the Facility will set up a data management system for the new sequencing service that should integrate with existing HPC infrastructure for processing, analysis and dissemination of raw sequencing data and analysis results. The system should i) optimise data storage usage ii) facilitate data sharing within the college and with external collaborators iii) comply with College and funder data preservation and protection policies iv) allow for a high degree of automation and v) be scalable. Figure 1 shows an overview of the data life cycle within the proposed system.

iRODS

The implementation of the system is based on the iRODS middleware. iRODS is a widely used, scalable, open source data grid system that has been successfully adopted for the management of next-generation sequencing data at other research institutions such as the Wellcome Trust Sanger Centre, the Broad Institute and the University of Uppsala. The iRODS platform implements a logical name space providing transparent access to distributed and disparate storage resources. iRODS features important for the implementation of the system are the rule engine and the metadata catalogue. The rule engine allows the invocation of pre-defined sequences of action (micro-services) in regular intervals or when particular events occur. State changes resulting from rule execution are stored and this information can be used to track and control rule execution. The rule engine will facilitate the effective management of data life cycles (file migration, file format conversion etc.), data preservation (consistency, replication and archiving) and data security (encryption). The metadata function allows to associate descriptive system and user defined metadata with files. This will enable search, management and tracking of data and data manipulation within the system.

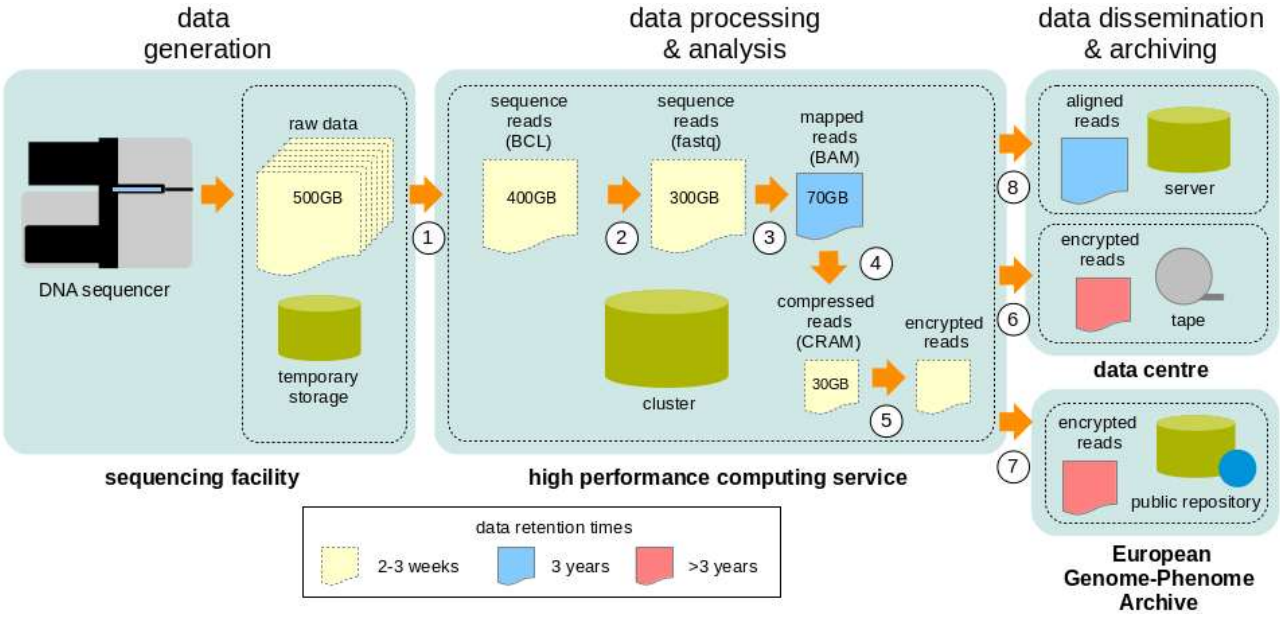

Figure 1 Data management system for genome sequencing data. Raw sequencing data is transferred from a local storage server at Hammersmith Campus to the HPC Service at South Kensington (1), read data is converted from the vendor specific BCL format to the platform independent fastq format (2), sequencing reads are mapped to a reference genome sequence (3), mapped read data is compressed further using reference based read compression (4), compressed reads data is enrypted (5), encrypted read data is archived on tape (6) and remotely at a public repository (7), aligned read data is disseminated locally to users (8).

Implementation

We have implemented a prototype of an iRODS-based data management system that comprises all steps of the data processing required to disseminate sequencing data to internal users of the facility (steps 1 – 4 and 8 in Figure 1): (1) upon completion of a sequencing run raw data required for the conversion of raw sequencing data from the vendor-specific BCL format to the platform-independent fastq format is transferred to a storage server at the HPC Service in South Kensington. (2) Reads are converted from BCL to fastq format using the vendor-supplied bcl2fastq software.

Splitting of read data by sample and project also occurs during this step. Subsequently a quality control report is generated with FastQC. (3) Reads in fastq format are then mapped to a reference genome sequence with BWA MEM which outputs mapping results in BAM format resulting in a >3-fold reduction in file size. (4) Further compression is achieved by applying a reference based compression algorithm using samtools v1.0 to convert BAM to CRAM files. (8) Aligned read data in BAM format is transferred to a webserver and made accessible for download through the web-based iRODS client iDrop Web 2.

The initial proposal envisaged management of the entire data processing workflow through the iRODS system. Figure 2A outlines the iRODS setup originally proposed with iRODS resource servers running on all storage resources. The webserver is also an iCAT enabled resource server that hosts the iRODS meta data catalogue. In this setup the entire data life cycle can be managed and tracked by iRODS.

However, concerns were raised by the HPC service team regarding user authentication and file ownership management. The default iRODS setup requires iRODS user authentication separately from the system authentication. Apart from the additional administrative overhead the trustworthiness of the iRODS authentication system was questioned by the HPC service team and considered a potential security risk. With regard to file ownership management, the fact that all files that are put into iRODS will be owned by a single user (the user the iRODS service is running as) was considered a problem as this creates complete dependence on the iRODS file ownership management.

Figure 2 Proposed vs implemented setup of the data management system. A) Proposed iRODS setup in which all storage resources in the system are iRODS resource servers belonging to the same iRODS 'zone' (igfZone). B) Implemented iRODS setup. Only storage resources administrated by the Genomics Facility run iRODS resource servers. The storage resources administrated by the HPC service are outside the iRODS system.

Firstly, to address the concerns regarding user authentication it was suggested to enable iRODS PAM authentication which allows the use of system passwords for iRODS authentication. Secondly, the issues relating to file ownership could be overcome by implementing the HPC iRODS resource as a 'direct access resource'. Direct access data resources are accessible both through iRODS and through the file system. iRODS acts as an "overlay" for the UNIX file system providing meta-data annotations for the files in the file system. However, in this operation mode iRODS runs as a root process which means that iRODS could potentially access/write any data on the system. The complete reliance on the authentication and access control in iRODS was not considered acceptable. Due to the concerns regarding data security and integrity the system was implemented without the integration of the HPC storage resource into iRODS (see Figure 2B). Instead data is fetched from iRODS via an iRODS client installed on the HPC system. After processing of the data on the cluster results and metadata are pushed back into iRODS.

Conclusion

We have successfully implemented a prototype of a data management system for sequencing data generated by the Imperial BRC Genomics Facility. The use of iRODS to manage and track the data life cycle within the system constitutes a compromise between the powerful functionality of iRODS and the sacrifice of file system-based control over data access and ownership. This becomes an issue when shared resources such as HPC systems need to be integrated where iRODS-based file management might conflict with the existing setup of the system. These issues would need to be addressed if the iRODS system was to be i) extended to manage data generated by secondary data analysis workflows and ii) used more widely across the college to manage sequencing and other research data.