

Genomic Data Analysis

Underpinning genomic research

Experimental scientists in genomics face two specific challenges. The volumes of data are enormous, and their experimental designs are subject to continual change. The GDA and GDS projects directly address both of these issues.

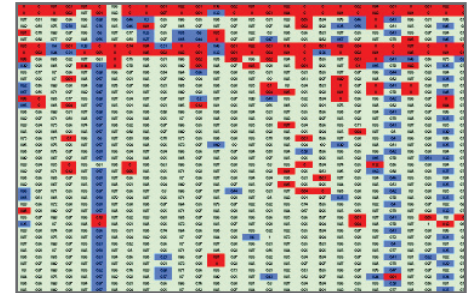
Overview

Genomic projects underpin almost all aspects of modern biology. This includes modern molecular biology, biodiversity studies, and medical research including but not limited to research into cancer, vaccines, antibiotics and drug development.

Many research institutions around NSW and elsewhere have purchased new generation DNA sequencing instruments and need to store, curate, access and analyse the immense

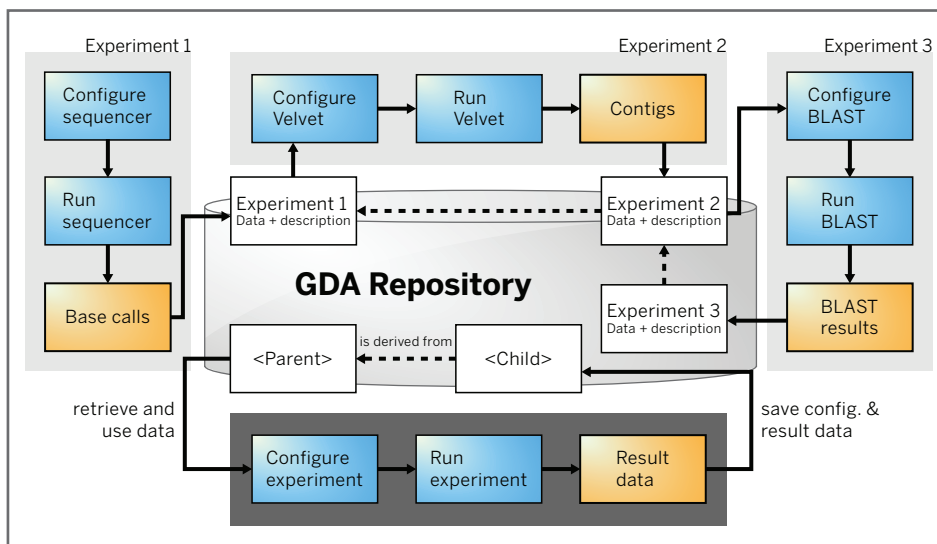
amount of DNA sequence data to be generated.

The new DNA sequencing equipment generates billions of base pairs worth of sequence data per day, and this will only rise. This new equipment shifts the bottleneck away from the generation of DNA data and onto the ongoing data processing, data management and data access to ensure that the information is readily available to support research.



This project centralises the effort of several major institutions in the scoping and development work necessary to make effective use of gene sequencing instruments, as well as ensuring centralised computational and data storage facilities can be used effectively in this research.

The project will benefit a wider user base including researchers at UNSW, Southern Cross University, the Australian National University and many others.



High level overview

“We can now store massive amounts of genomic data, share it with our colleagues and analyse it in a seamless manner.”

This is fundamental infrastructure that underpins genomic research. Without it, we just can't do the work.”

Professor Marc Wilkins
Ramaciotti Centre for Gene Function Analysis

One of the key benefits of the project is that it is designed for easy deployment at other, new, sites.

The project was completed in May 2010.

Features

GDA helps experimental scientists manage the data and design of their experiments. This includes wet-lab experiments like tissue sample preparation, next generation sequencing and base calling, and tertiary analyses such as Blast searching. Experimental scientists in genomics face two specific challenges. The volumes of data are enormous, and their experimental designs are subject to continual change. GDA directly addresses both of these issues.

The GDA user interface is designed around experiments and projects. Experimenters encode their experimental design as a set of input and output parameters (i.e. independent and dependent variables). The GDA user interface uses these definitions to allow easy and validated data entry of experiment configurations, as well as reuse and sharing of experimental designs.

Projects make it easy to allow controlled access to your experiments, including integrated support for inclusion of your data in the ANDS Australian Research Data Commons.

Intersect Australia Ltd

www.intersect.org.au

Level 12, 309 Kent St, Sydney
NSW 2000 Australia

enquiries@intersect.org.au
T +61 2 8079 2500

Experiment output data from all experiments are transparently compressed to minimise download times and server storage requirements.

Each result in the repository comprises the data for a single experiment (both its design and output data). Output data from one experiment is often used as input data to subsequent experiments, forming a sequence of interdependent experiments. GDA supports reuse of experiments, including full scientific auditing of results right back to their source.

Intersect's GDA system provides a repository for experimental results and metadata. The experimental metadata for any facility and type of experiment can be easily configured and the GDA system automatically generates the user interface to allow easy metadata entry.

Access to the GDA system is via a web based user interface. Upload of experimental results is handled via a Java applet running in the web browser and works over institutional fire-walls.

The system offers the following key features:

- a repository for results and metadata based on Fedora Commons;
- full user and group management;
- access management for results;

- grouping of results and users into projects;
- the ability to confer result ownership;
- the ability to create derived results and relate them to the original experiment;
- export of results to the ANDS Australian Research Data Commons.

Reusability: High ■■■

applicable to the gene sequencing community

Project Details

Start Date: September 2009

End Date: May 2010

Clients:

The Ramaciotti Centre for Gene Function Analysis, UNSW, the Centre of Plant Conservation Genetics, SCU

Technologies used:

Fedora Commons, GDA Repository, ANDS Research Data Commons, jQuery, PostgreSQL, Tomcat, Spring.

Related Links

www.scu.edu.au/research/cpcg

For enquiries, please contact: **Rodney Harrison**

E rodney.harrison@intersect.org.au

T +61 2 8079 2527