

Connecting Genetics Researchers to NESI.

James Boocock
Department Of Biochemistry
University of Otago

***Supervisors: Dr. Mik Black, Associate Professor
Tony Merriman, Dr. David Eyers and Dr. Phil
Wilcox***

About Me

- Computer Science Graduate.
- Currently pursuing a Diploma for Graduates in Genetics and Statistics.
- Research Assistant in the Merriman Lab at Otago University
- Currently working on a selection pipeline.
(+ this)

UNIVERSITY
of
OTAGO



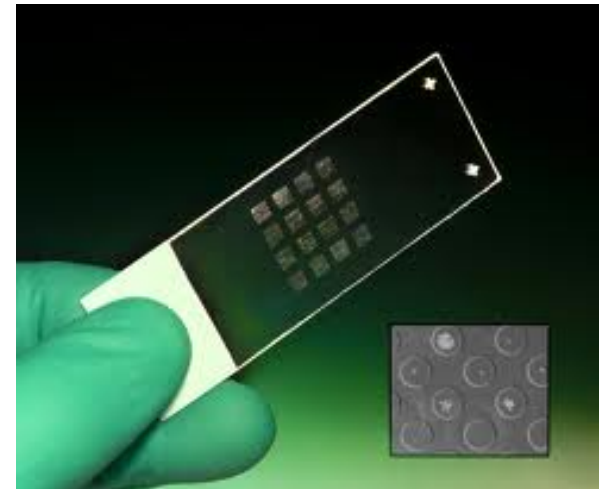
Te Whare Wānanga o Otago

NEW ZEALAND

eResearch as an Enabler

Genomic Data Explosion!

- Genomic data is coming from everywhere.
- Microarray data
- Sequencing data (both RNA and DNA sequencing)



Motivation

1. Genetics researchers are required now more than ever to use computers as part of their research. This usually requires the command-line
2. As the scale of the data increases (e.g high density genotyping) the computational cost is increasing and researchers are required to recruit the use of shared computing resources.

Previous Work

- For a Summer of eResearch and subsequent work Edward Hills and I maintain a galaxy instance (still do) for the Merriman lab. My role involves adding the most common command-line processes to galaxy.
- Recently over summer (another Summer of eResearch project) attempted to tackle both problems 1 and 2.





Galaxy – A Screenshot

The screenshot displays the Galaxy web interface with the IMPUTE2 (version 1.0) tool selected. The interface is divided into three main sections: Tools, the main tool configuration area, and History.

Tools Panel (Left): Contains a search bar and a list of tool categories including Get Data, Send Data, Otago - Selection Tools, Otago - Variant Effect Predictors, Otago - Filters, Otago - Calculations, Otago - VCF file Manipulation, Otago - Work with 1000 Genomes Data, Otago - Query Databases, Otago - Plink tools, Otago - File converters, SnpEff tools, ENCODE Tools, Lift-Over, Text Manipulation, and Filter and Sort.

IMPUTE2 (version 1.0) Configuration (Center):

- Chromosome Number:** A text input field.
- For X chromosome please just enter x**
- Are your haplotypes phased or unphased?:** A dropdown menu set to "Unphased".
- Unphased data will produce slightly more accurate results but take longer to run**
- Unphased .gen file:** A text input field.
- Allow larger imputation regions:** A checkbox.
- This is not a recommended option but for imputing large regions this will be required**
- Buffer Size:** A text input field set to "250".
- Length of buffer region (in kb) to include on each side of the analysis interval. Using a buffer region helps prevent imputation quality from deteriorating near the edges of the analysis interval. Larger buffers may improve accuracy for low-frequency variants (since such variants tend to reside on long haplotype backgrounds) at the cost of longer running times.**
- Effective Population Size:** A text input field set to "20000".
- Commonly denoted as N_e in the population genetics literature from which your dataset was sampled. This parameter scales the recombination rates that IMPUTE2 uses to guide its model of linkage disequilibrium patterns.**
- Execute** button.
- Available Grids:** A dropdown menu set to "NewPan".
- Parallelism Type:** A dropdown menu set to "Base Pair".
- Parallelism String:** A text input field.
- This tool uses IMPUTE2 a program for imputation. It can accept pre-phased data (e.g. from ShapeIT) or unphased data**

History Panel (Right): Shows a list of previous jobs. The top job is "Nesi Test" (14.6 GB). Below it is a job titled "3: 18:0-100000000 region extracted from ALL.chr18.phase1.projectConsensus.genotypes.vcf" (Job is currently running). Below that is a job titled "2: ALL.chr18.phase1.projectConsensus.genotypes.vcf". At the bottom is a job titled "1: ImputeData".

Galaxy

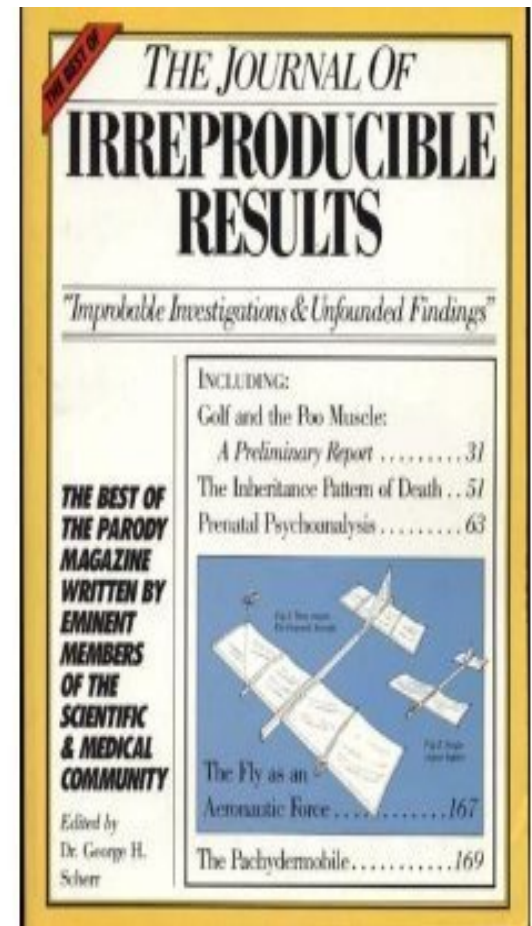
- Galaxy is a web interface for large scale computational biomedical analysis. It is widely accepted by the community.
- Galaxy is easy to use and has a very slight learning curve.
- Enabling command line tools to be integrated under the one interface makes it easier for anyone who is not familiar with the command line. Running in a web browser gives users an interface they are likely to be acquainted.

Galaxy's Goals

- Galaxy aims to be the web-based platform for accessible, reproducible, and transparent computational biomedical research.
- **Accessible:** Users without programming experience can easily specify parameters and run tools and workflows.
- **Reproducible:** Galaxy captures information so that any user can repeat and understand a complete computational analysis.
- **Transparent:** Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis.

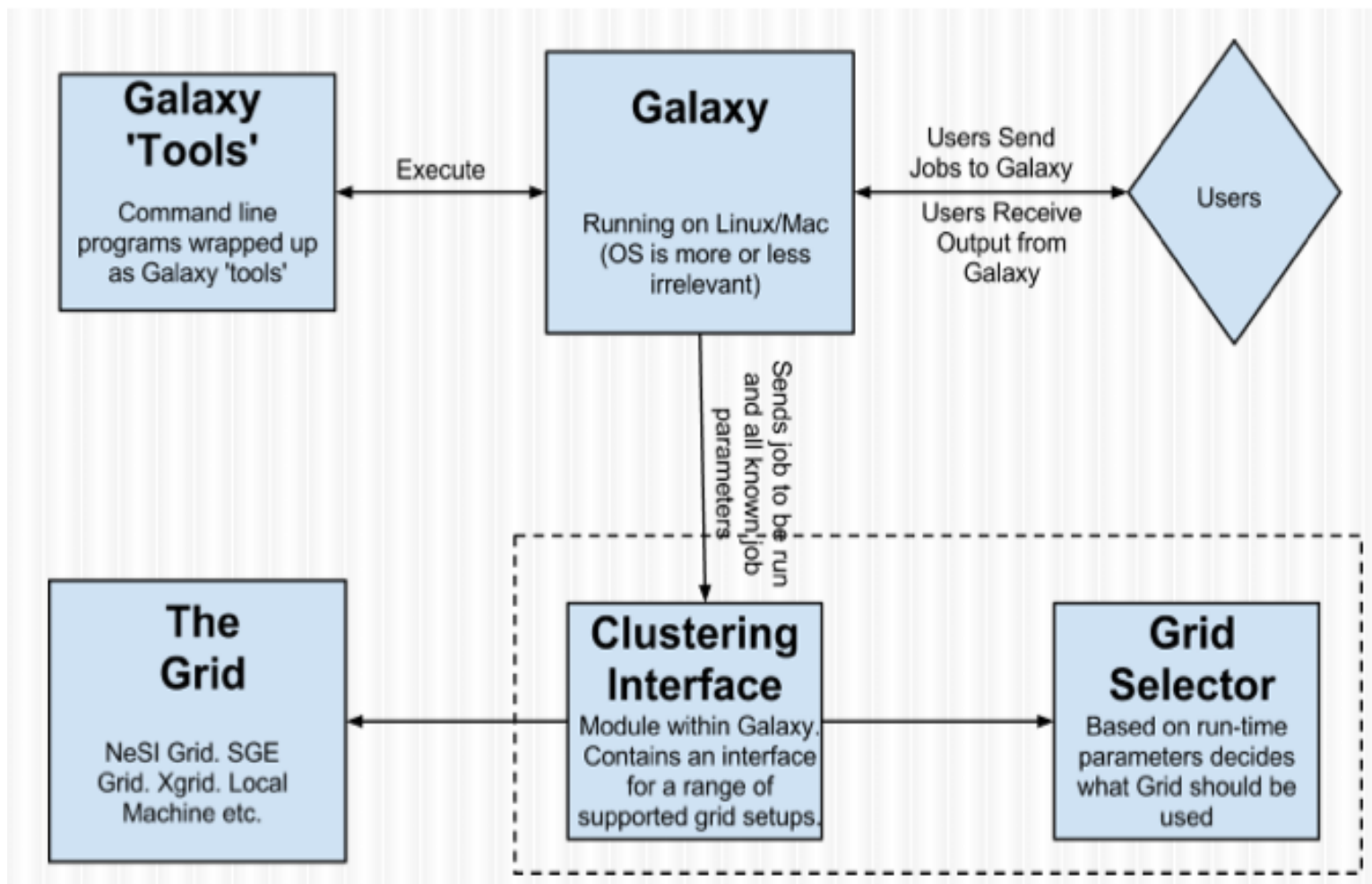
Reproducible Research

- It is important that scientific results are reproducible.
- Large computational analyses performed on the command-line tend to be lazily documented.
- Galaxy documents this for you using histories, workflows and pages





Galaxy – Grid Infrastructure



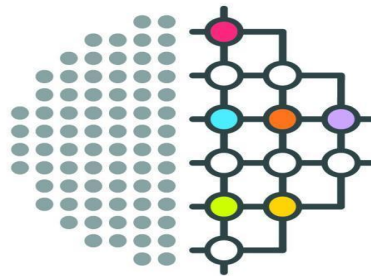


Galaxy HPC support

- Galaxy has inbuilt interfaces to many common grid interfaces from TORQUE, PBS, SGE and many more.
- The recommended setup specifies that shared file system be used.
- To take advantage of the HPC jobs should be split into smaller tasks where possible.

NeSI and Galaxy

- A major goal and motivation was to create an interface to go between galaxy and NeSI infrastructure using the standard nesi toolkit.



NeSI
New Zealand eScience
Infrastructure

Imputation

- estimation of unmeasured genotypes

Typical imputation scenario

HapMap or 1,000 Genomes	0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
	0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
	1	0	1	1	0	0	0	1	1	1	1	1	0	0	1
Cases and controls typed on SNP chip	1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
	1	?	?	?	1	?	0	?	?	?	?	?	0	?	0
	0	?	?	?	1	?	1	?	?	?	?	1	0	?	1
	1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
	?	?	?	?	2	?	0	?	?	?	?	0	0	?	0
	1	?	?	?	1	?	1	?	?	?	?	1	0	?	?
	0	?	?	?	2	?	0	?	?	?	?	0	1	?	1
	1	?	?	?	1	?	1	?	?	?	?	1	1	?	2

Reference
haplotypes

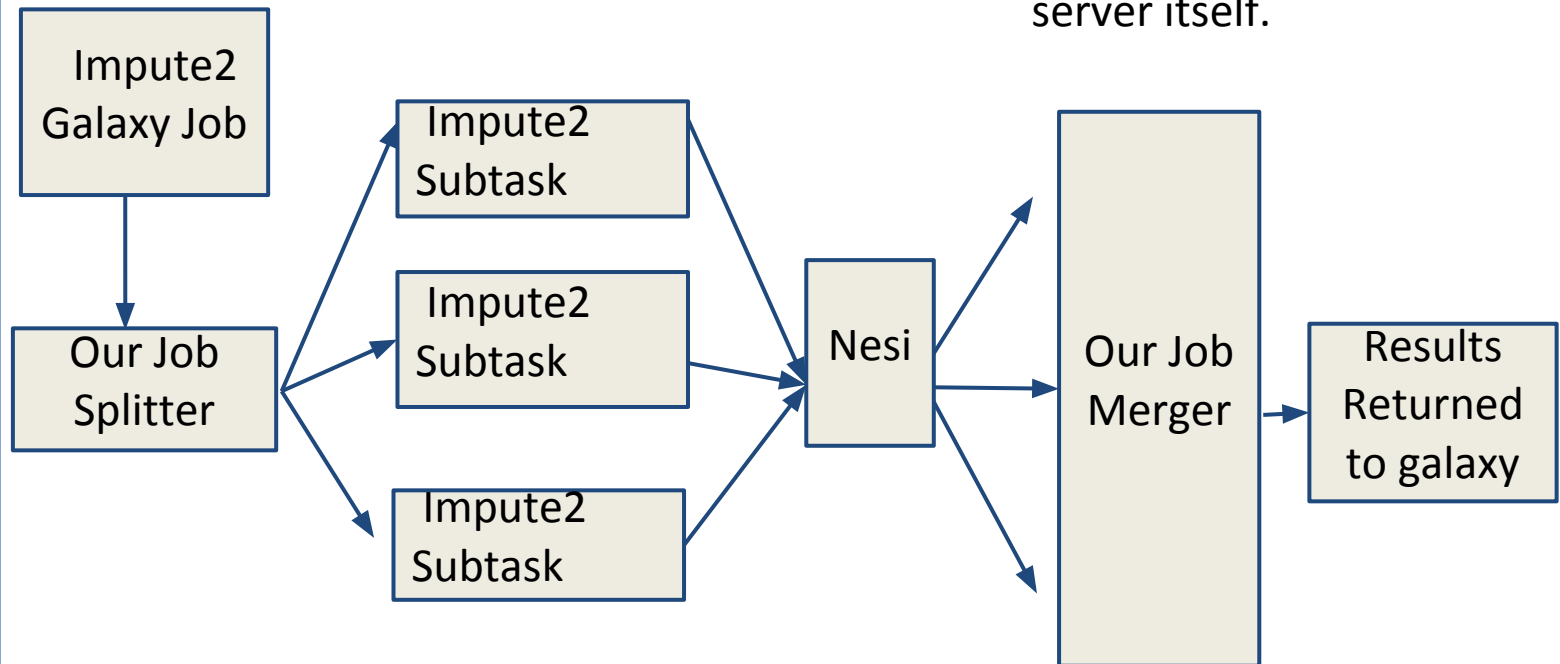
Study
genotypes



Nesi and Galaxy - Impute 2 Run

Subtasks would be sent a grid available to the galaxy instance.

Job Merging and Job Splitting is run on the galaxy server itself.



Module development

- Git was used to track the changes to the codebase and github.com was used to store the code base.
- Sphinx a python documentation generator will be used for documentation.



The Nerdy Details

- Galaxy is written in the python programming language so to ease integration all the code we wrote is in python.
- This is apart from the submission scripts which use the Grisu client library and Jython which is a java package that lets you write in python syntax.



Unfinished Business

- The project is not complete various bugs need to be ironed out before the project could be deployed (retrieving impute2 job results is a folder download away).
- Authentication is done using the users local copy of the standard Nesi Tools.
- Shibboleth has been investigated to authenticate the galaxy users to Nesi but further investigation is required.



Shibboleth.

Unfinished Business

- Edwards Hills investigated efficiency measures for parallelization or some bioinformatics tools.
- Building on his project a machine-learning approach to job submission could be implemented. The job options could be set based on information about how previous runs progressed.





Acknowledgements

- Edward Hills
- Murray Cadzow
- Hoang Ngyuen
- Markus Binsteiner
- Vlad Mencil
- Richard Hosking
- Sina Masoud-Ansari,
- Tim McNamara
- Nick Jones



Acknowledgements

- Dr David Evers
- Dr Mik Black
- Associate Professor Tony Merriman
- Phil Wilcox

UNIVERSITY
of
OTAGO



Te Whare Wānanga o Otago

NEW ZEALAND

Thanks and Questions

