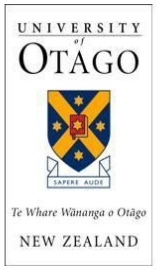

Galaxy, 1000 Genomes and the GATK

Overview

Table of Contents

Introduction

Galaxy is a web interface for genomic analysis. Galaxy enables users to peform various operations on genomic data under one easy to use interface. It gives you the ability to create workflows, annotate datasets, histories and everything in between, use the in-built visualiser and have the ability to look at others work and follow their experimental procedure they used to achieve the results.

The flexibility of Galaxy gives you the ability to create your own tools and simply plug them into Galaxy with the use of some handy XML syntax. This is due to Galaxys underlying system which more or less just simply calls a command line program. This flexibility allows your own instance of Galaxy to run anything that you can ordinarily run through the command line with few exceptions.

Galaxy is needed by scientists who lack the computer skills necessary to use these complicated command line applications and find the learning curve cumbersome. The utilization of Galaxy is of importance to people as productivity is often hampered as a result of the skill disparity. Because of the needs of the people and the general community acceptance of Galaxy we were instructed to help the Otago Genetics department in utilizing this vast resource.

We have constructed a wide range of tools as well as better utilized ones that were already in place. This coupled with our local install of Galaxy on Ubuntu 10.10 allows for easy use and understanding, as well as the freedom to modify or carry on with the work we have started.

The 1000 Genomes Project is a large effort with the hopes of sequencing more than 1000 people at good coverage and finding all genetic variants. The 1KG project is an

mbitious and large project but are widely regarded as the leaders in genetic sequencing.

Becase of this we have chosen to adopt as much information from the data they have collected as possible and integrate into the tools we use.

The Genome Analaysis Toolkit is a set of software created by MIT and the Broad Institute. The GATK focuses on being able to go straight from the raw data produced by sequencing machines and straight trough to a VCF format which can be viewed and analysed. These tools are very powerful and simple to use. Due to Galaxy being able to run command line programs such as the GATK we have tried to implement this inside Galaxy. However, the Broad institute are also working on it and have it in a Beta stage so we decided not to bother as it would be best left up to them. We do however use one of their tools which we felt was too complicated for our purposes.

Installation and Updating

1) Installation

In our scripts folder we have a number of scripts used for installing our various tools as well as setting up a production envrionment identical to the one we run our Galaxy installation on.

We have two main setup scripts:

1) **full_install.sh**

   This script goes from a clean installation of Ubuntu 10.10 and goes right through until you have a clean install of Galaxy, ProFTP, Postgres and our tools. We recommend this choice as it sets everything up identical to our setup which we find is the best practice. This is mainly due to the fact that setting up Galaxy and all her dependencies can be a time consuming and difficult task. This way we can gurantee we have everything working.

2) **tool_install.sh**

   This script can be used by ***setup_script.sh*** or standalone. This script reads *three* parameters from the config file ***setup_config.cfg*** and installs all of our tools into a directory specified by the config file. When running standalone there are a few dependencies which may be needed as well as certain web server configurations for some tools too work.

2) Updating

Because our tools rely heavily on third party vendors, some files need to be updated on a semi-regular basis. Things like **dbSNP** and **1000 Genomes Project** regularly bring out new or revised versions of their databases so to stay up to date and have the lastest information and genome references, they will need to be updated. Other tools that we rely on do not necessarily need to be updated but can be if you so choose. We have provided a readme for updating called ***UPDATING*** in the ***docs/*** folder which details how and where to update.

Tools

3.1) Ensembl Variant Effect Predictor

Input: Un-annotated VCF

Input: Filter Options

Output: Annotated VCF against Ensembl database with effects

This tool is an inplementation in Galaxy of the Ensembl Variant Effect Predictor perl script provided by the Ensembl website. We call this script and pass it the options which have been selected by the user. It then queries the downloaded Ensembl cache, runs its effect predictor algorithms and returns an annotated VCF file that has gone through Ensembl.

3.2) SnpEff Variant Effect Predictor

Input: Un-annotated VCF

Input: Filter Options

Output: Annotated VCF against SnpEff database with effects

Output: HTML file containing graphs about stats

SnpEff is Variant Effect Predictor created by P. Cingola and is the preferred VEP by the Broad Institute (and the creators of the GATK). This tool allows you to enter a VCF file and have it run through SnpEffs VEP. This tool produces two outputs, a annotated VCF as well as a HTML file. The HTML file contains a lot of useful graphs and tables to display visually the information stored in the VCF.

3.3) GATK Variant Annotator

Input: Snp Effect output VCF

Input: Original VCF that the SnpEff output was generated from

Output: Original VCF plus SnpEff information embedded in info column.

This tool is not the GATK Variant Annotator (which already exists inside Galaxy), but instead our implementation of it. Our one is slightly different as its sole purpose is to re-annotate VCF files with their SNP effect output. It is much simplier and eaed by his father and in 1982sier to use as the purpose is for this one task. Although this severly limits the functionality, the original tool still exists within Galaxy so if need be that is available.

3.4) Region Effect Filter

Input: SnpEff annotated VCF

Input: Filter Options

Output: Filtered SnpEff annotated VCF

This tool allows you to filter your annotated VCF based on the options selected. Some of these options can filter NON_SYNONYMOUS coding regions and well as SPLICE_SITES etc. By limiting the fields displayed it can make life much easier to focus on a particular thing at a time.

3.5) Get Allele Frequencies

Input: VCF containing allele frequencies

Output: Txt file containing easy to read allele frequencies

This tool returns the allele frequencies and a few other important columns from a chosen VCF. Its output is a standard txt file.

3.6) Get (Full or Novel) SNPs

Input: VCF file

Output: Same VCF file or VCF containing SNPs without rsIDs

This tool will return all SNPs or SNPs which it deems are novel. These 'novel' SNPs are just SNPs that don't have a rsID. So could be novel or could be a false positive etc.

3.7) Quality Filter

Input: VCF file

Input: Quality threshold

Output: VCF filtered based on quality threshold

This tool will filter out all records in a given VCF that have a quality of less than what the user has entered.

3.8) Indel Quality Filter

Input: VCF file

Input: Quality threshold

Output: VCF filter based on quality threshold and is an indel

This tool will filter out all records in a given VCF that have a quality of less than what the user has entered and is an indel.

3.9) Multiple Homozygous Count Filter

Input: VCF file

Output: VCF filtered by homozygous count

This tool will filter out all records in a given VCF that have a homozygous SNP that is present in more than a specified number of SNPs.

3.10) dbSNP rsID Annotator

Input: VCF file

Output: VCf file annotated with rsIDs from dbSNP

This tool annotated a given VCF file with rsIDs from dbSNP.

3.11) Mulitple Region Selection

Input: VCF file

Input: Txt file containing one region per line (chr:startPos-endPos)

Output: VCF file containing only regions specified in txt file

This tool allows a user to select a multitude of regions from a VCF given a list of entries in a txt file, one line per region.

3.12) Read Depth Filter

Input: VCF file

Input: Read Depth threshold

Output: VCF file filtered by a read depth threshold

This tool will filter out all records in a given VCF that have a read depth less than the threshold value specified by the user.

3.13) RSQ LD Plotter

Input: Hapmap formatted file

Output: R-Squared LD Plot

This tool produces an R Squared LD Plot based on a hapmap file downloaded from the hapmap website.

3.14) TS/TV Ratio Calculator

Input: VCF file

Output: TS/TV Ratio from VCF file

This tool calculates the Transition/Transversion Ratio from a VCF.

3.15) VCF Merge

Input: VCF file

Input: VCF file (as many as you like)*

Output: VCF file with all inputs aggregated

This tool allows two or more VCFs to be merged into one single file and the data accumulated.

3.16) Get Sample IDs

Input: VCF file

Output: Sample IDs from input VCF

This tool returns all sample Ids from a VCF. Usually used in a workflow with the sample id subsetter.

3.17) Sample ID Subsetter

Input: VCF file

Input: Sample IDs

Output: VCF file with only information from input sample IDs

This tool returns a subset of the given VCF based on a number of a sample IDs given to it.

3.18) VCF to CSV Converter

Input: VCF file

Output: CSV file (can be opened in Excel etc.)

This tool converts a VCF into a CSV which can be opened by external programs such as Excel and other spreadsheet software.

3.19) Get Population Specific Allele Frequencies

Input: Chromosomal Region

Output: VCF file from 1000 Genomes Project for inputted region

Output: CSV file (can be opened in Excel etc.)

This tool queries the 1000 Genomes Project and returns the allele frequencies per population for a given region. It returns two files, the first is a VCF containing the

output and the second is a summary which contains minimal information and allele frequencies.

3.20) Exome Variant Server Querier

Input: Gene Name, Chromosomal Region or Gene ID

Output: VCF file for inputted region or gene

Output: TXT file containing all sites information for region or gene

Output: TXT file containing summary stats for region or gene

This tool queries the Exome Variant Server based on the region, gene id or gene name and returns three files. One is the outputted VCF, the second is a All_Sites file and the third is a summary_stats file.

3.21) Region Extraction

Input: VCF file

Input: Region

Output: VCF file with only selected region present

This tool extracts a region from a VCF file based on a region the user has entered.

The Future

Looking ahead there are multitude of options to extend and improve the work we have started. In the future our mentors hope to have a New Zealand based galaxy distribution server on a national compute cluster. The possibilty of having local instances of the large genomic databases (1000genomes, DBSNP, hapmap etc) was something we also wanted to implement but due to storage limits related to server administration meant this was not possible. Setting up our Galaxy installation on a cluster would also be worthy of investigating if we had the time.

Citation and Thanks

Galaxy web interface.

Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010 Aug 25;11(8):R86.

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. "Galaxy: a web-based genome analysis tool for experimentalists". *Current Protocols in Molecular Biology*. 2010 Jan; Chapter 19:Unit 19.10.1-21.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. "Galaxy: a platform for interactive large-scale genome analysis." *Genome Research*. 2005 Oct; 15(10):1451-5.

Galaxy web interface Visualisation

Jeremy Goecks, Kanwei Li, Dave Clements, Team, The Galaxy, James Taylor, The Galaxy Track Browser: Transforming the genome browser from visualization tool to analysis tool. Biological Data Visualization (BioVis), 2011 IEEE Symposium on (October 2011), pp. 39-46.

SnpEff and SnpSift were used frequently throughout our project.

Cingolani, P. "snpEff: Variant effect prediction", http://snpeff.sourceforge.net, 2012.

Gatk Variant Annotator is used for Adding snpEff genomic annotations to a VCF file.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep; 20(9):1297-303. Epub 2010 Jul 19.

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl., C., Philippakis, A., del Angel, G., Rivas, M.A, Hanna, M., McKenna, A., Fennell, T. Kernytsky, A., Sivachenko, A, Cibulskis, K., Gabriel, S., Altshuler, D. and Daly, M. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011 Apr; 43(5):491-498.

Ensembl's VEP perl script was how galaxy interfaced with the locally installed ensemble cache.

www.ensembl.org/

EVS

Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [date (Febuarary, 2012) accessed].

Samtools's Tabix

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11.http://samtools.sourceforge.net/tabix.shtml

VCF-tools

The Variant Call Format and VCFtools, Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group, Bioinformatics, 2011

Haploview

Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005 Jan 15 [PubMed ID: 15297300]

1000Genomes

Nature 467, 1016-1073 (28th October 2010).

DBSNP

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11.