A bioinformatics workflow for detecting signatures of selection in genomic data

Murray Cadzow[1,2], James Boocock[1,2], Hoang Tan Nguyen[1,2], Phillip Wilcox[2,3], Tony R Merriman[1] and Michael A Black[1]

[1]Department of Biochemistry, University of Otago
[2]Department of Mathematics and Statistics, University of Otago
[3]Scion Research, Rotorua, New Zealand

September 9, 2013

Contents

1 Introduction

Mik to fill this in

2 Getting Started

2.1 Prerequisites

The selection pipeline requires some basic libraries and tools are installed on your system. These are python3, Zlib and others.

2.2 Installation

To install the package standalone, requiring manual configuration of the config file, run this as root.

```
./install.sh --standalone
```

The rest of this section will be dedicated to the automatic installation. To perform an automatic installation of the selection pipeline run as root.

```
./install.sh
```

Installation creates a default config file located in the base directory of the pipeline. Installation adds a program called selection_pipeline to the system path. To test the program is installed correctly run the following command at a terminal prompt.

```
selection_pipeline -h
```

The pipeline uses VCF tools, VCF tools requires a perl library to work correctly. Assuming you are using bash as your default shell open your .bashrc file in your home directory and append to the PERL5LIB environment variable.

```
# Change the path to match the location of your selection pipeline bin folder.
export PERL5LIB=\${PERL5LIB}:/home/sfk/selection_pipeline/lib/perl5
```

To avoid having to change the default config file append the same folder to you PATH environment variable.

```
export PATH=\${PATH}:/home/sfk/selection_pipeline/bin
```

2.3 Genetic Maps and Impute Haplotypes

To use the phasing and imputation features of the pipeline requires both genetic map files and haplotype files. For humans these files that conform to the format required for shapeit and impute2 can be found here. Download and extract the reference files of your choice and extract them. To use the files with the selection pipeline requires setting a few options in the config file you will use these are as follows an example config file is avaliable the base directory of the selection pipeline. The question mark character "?" in the config is substituted by the chromosome number, this is used for reference files that are split on chromosomes.

```
[shapeit]
genetic_map_dir = # genetic map file directory #
genetic_map_prefix = # full file name with ? where chromosome number changes #
[impute_executable]
impute_map_dir= # genetic map file directory location #
impute_reference_dir = # impute reference directory location #
impute_map_prefix = # the full file name with a ? for where the chromosome number is #
impute_reference_prefix = # The file name with a ? for chromosome name and the extension left out.
```

2.4 Ancestral Fasta Files

To generate results for iHS requires assigning the ancestral allele. The selection pipeline uses the 6-way EPO (Enredo-Pecan-Ortheus) alignment pipeline. The files can be downloaded from here. To setup the selection pipeline open your config files and change the following fields.

```
[ancestral_allele]
ancestral_allele_script= /home/smilefreak/MerrimanSelectionPipeline/selection_pipeline/aa_annotate.py
ancestral_fasta_dir = # directory you downloaded alignment to #
ancestral_prefix=human_ancestor_?.fa # default extracted file chromosome substitution #
```

3 Tutorial

3.1 Selection Signatures at the Lactase Locus

3.1.1 Getting the Data

The lactase gene is located on Chromosome 2 between 136,545,410-136,594,750 positions. For the example we will use a 10 megabase region containing the Lactase gene and the CEU and YRI populations from the 1000 genomes. In order to demonstrate how to use the pipeline we will use the chromosome 2 region 130,000,000-140,000,000. To download the example dataset enter the command below.

```
wget http://tutorial_file_location.com
```

Extract the example data into a new folder.

3.2 Setting up Pipeline Run

To perform the analysis the first step will is to navigate to the folder you extracted the tutorial data.

3.3 Population Files

Population files are required for any cross population comparisions. The commands below will initiate the data generation step.

```
multipop\_selection_pipeline -p POP1.txt -p POP2.txt -i input.vcf --c defaults.cfg
```

The generated folders and current folder have all the data required to perform further selection analysis. Within each population folder 4 output files are generated these contain tajima's D, iHH, an updated VCF and Fey and Wu's H statistic these files are located in the results folder inside each population subfolder. Between each each population Fst are generated and located in the fst folder. Fst calculations are generated using the weir and cockerham estimator.

3.4 Data Analysis

The pipelines purpose is to generate signiatures of selection from a VCF formatted input file. In order to express the usefulness of the pipeline it is pertinent to illustrate the effectiveness of the pipeline. The next section describes some basic plotting of these data using the R programming language.

3.4.1 Fst

3.4.2 Fey and Wu's H

3.4.3 iHS

3.4.4 Tajima's D

3.4.5 rSb

4 Command line Arguments

The selection pipeline contains three major scripts selection_pipeline aa_annotate and multipopulation. The selection pipeline does all the within population statistics calculations. The multipopulation program calculates all the between population statistics and calls the selection pipeline. The aa_annotate program annotates a haplotype file or a phased vcf file with the ancestral allele from the 6-way EPO alignment, for other species or alternative ancestral annotation the feauture will be added promptly.

4.1 Multipopulation

4.1.1 Input Files

- -i <vcf input file> VCF file containing all the populations you want to analyse from one chromosome or a part of a chromosome only.

4.1.2 Output Files

- FST

  Fst results are stored in the fst folder with the chromosome number followed by the two populations. e.g 2CEUYRI.fst

- Selection Pipeline Results

  All single population pipeline results are stored in the subdirectory of the population in a folder named results. These contain the ihh, tajimasD and a population VCF file.

4.1.3 Other parameters (Compulsory)

- -c <Chromosome>

  Chromosome name used for labelling outputs.

- -a <Arguments to the selection pipeline>

  Quoted string containing any extra arguments to the selection_pipeline program. e.g "–imputation"

- -C <path to config file>

  Path to the selection pipeline config file an example config file is located in the base directory of the extracted package.

4.1.4 Other parameters (Optional)

- –fst-window-size <FST window size>

  Argument is passed directly to the VCF tools command line.

- –fst-window-step <FST window step>

  Argument is passed directly to the VCF tools command line.

4.2 Selection Pipeline

4.2.1 Input Files

- -i <VCF input file>

  Single population single chromosome VCF input file.

4.2.2 Output Files

The Results directory contains all the output files.

- .ihh file

  The outputted iHH data for each SNP

- .taj_d file

  Tajima's D output

- .vcf file

  Single population VCF updated by the pipeline, can contain.

4.2.3 Other parameters(Compulsory)

- –config-file <Config File path>

  Path to the selection pipeline config file an example config file is located in the base directory of the extracted package.

4.2.4 Other parameters(Optional)

- –maf <minimum MAF>

  Minor allele frequency filter threshold any SNPs below this threshold will be discarded from the analysis.

- –hwe <hardy-weinberg minimum p-value>

  A hardy weinberg test is performed on every snp any snps failing the test will be discarded.

- –daf <Minimum derived allele frequency>

  Derived allele frequencies below this minimum will be discarded.

- –remove-missing <Inclusion threshold for missing genotypes>

  Inclusion criteria for SNPs with missing data. SNPs with less than this value will be removed from analysis.

- –TajimaD <tajimas D bin size>

  Tajima's D statistic bin size.

4.3 Ancestral Annotation

4.3.1 Input Files

4.3.2 Output Files

4.3.3 Other parameters

4.4 Configuration File

The selection pipeline requires a configuration file, by default the program looks in the current working directory for a file named defaults.cfg but you can point the program to any file using command line arguments. There are two main programs in selection pipeline namely *selection_pipeline* and *multi_population*. These programs share a config file but certain configuration parameters can be ommitted when using the *selection_pipeline* program exclusively. A clean install of the program generates an example configuration file containing default arguments for all the compulsory parameters.

4.5 Argument Description

5 Extra Features

5.1 Genome Wide Selection Scans

The script is designed to work on one chromosome starting with a VCF file containing one or more populations. The pipeline was designed to be applied genome wide in the genome_wide/ directory we have provided a script that can perform a genome wide analysis. Read the README in this folder for more information.

5.2 Ancestral Allele Annotation

Needs to be fixed to take any set of human fasta files as the ancestral allele combination.

5.3 Galaxy Intergration

The galaxy folder contains the scripts required to add the selection pipeline to your local galaxy installation. The pipeline is also avaliable on the galaxy toolshed at galaxy_url