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 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/bin
```

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
```

2.2 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 in the base directory of the selection pipeline.

```
[shapeit]
genetic_map_dir = # genetic map file directory location #
genetic_map_prefix = # full file name with ? where chromosome number changes #
[impute_executable]
genetic_map_dir = # genetic map file directory location #
impute_map_prefix = # the full file name with a ? for where the chromosome number is #
impute_reference_prefix = # the full file name for the reference haplotype files with a ?
```

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. Usually iHS would be done chromosome wide but 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
```

3.2 Setting up Pipeline Run

To perform the analysis we will only