Documentation for *SLAT* software: Version 1.0

Xiaoquan Wen and Dan L. Nicolae
Department of Statistics
The University of Chicago
5734 S. University Avenue
Chicago IL 60637, USA

July 9, 2009

1 Introduction

SLAT stands for "Set Level Association Testing". The method implemented in *SLAT* is designed to assign a measure of significance to pre-defined sets of markers in the genome. The sets can be genes, conserved regions, or groups of genes such as pathways. The method is particularly advantgeous when there are multiple modest signals presented in the marker set of interest. For the detailed description of the methodology, please refer to De la Cruz et al.(2009).

SLAT is implemented using C++ programing language and GNU scientific library (GSL). The released executables have been optimized to achieve the maximum computational efficiency. The source code is also available upon request.

The current release of *SLAT* (version 1.0) implements the method of testing genotyped SNPs for a pre-defined set of SNPs in case-control settings of genomewide association study. The method for testing untyped SNPs will be implemented in the next release of *SLAT*.

This document provides information and demonstrations on input data formats, program usage and program outputs.

2 Input file format

SLAT requires three types of input data files

- A genotype data files for case-control samples.
- A map file for genotyped SNPs.
- A SNP set definition file.

All data files should be in plain text (ASCII) format.

2.1 Genotype file

In genotype data files, genotypes are coded by the nucleotide letters (e.g. AG, AA or GG). Missing genotype should be denoted with symbol "??". Each individual sample is listed in a row with following format:

Individual_ID (string) *affection_status (binary)* G_1 ... G_k

In field of affection status, code "0" and "1" denote control and case respectively.

2.2 Map file

Map files annotate the SNPs used in genotype files. It is also cross-referenced by the program to decide if a particular marker is located in a pre-defined region. The format of the map file is as following

RS_Number *Chromosome* *Position* *Strand* *Allele1* *Allele2*

An example of map file entry is shown below:

rs4949165 1 5103122 1 C G

Note, the strand orientation information is not used by current version (v1.0) of the software (users can set strand orientations of all SNPs as "1" for using of current release of *SLAT*), it will be used in the future release for correctly imputing the missing genotypes. It is important that the order of the SNPs listed in the map files matches the order in genotype files.

2.3 Set definition file

Set definition file defines the "sets" of SNPs that of interest. Each entry of this file defines a set to be tested by *SLAT*. There are two ways to define a set: either by genomic positions or by SNP rsnumbers.

To define a set by genomic positions, the format is as follows

Set_ID Set_Name Chromosom Start_Position End_Position

Both region ID (integer) and region name (string) are user specified identifiers for the set of to be defined. If the sets are genes, then region IDs and region names are naturally gene IDs and gene names, e.g. for Lactase Gene, the following entry can be used.

3938 LCT 2 136379147 136428482

The other way to define a set is to enumerate all SNP rsnumers/IDs in the set, i.e.

Set_ID Set_Name rs1 rs2 ....

There is no limitations on number of SNPs in a particular set. However if a given SNP is not annotated by the map file, it will be ignored from the analysis by *SLAT*.

3 Running *SLAT*

The command line syntax for running *SLAT* is:

slat -d genotype_data_file -m map_file -s set_definition_file [-o output_file] [-p permutation_iteration]

There are three mandatory arguments expected for *SLAT* in the command line. Those are "-d" followed by the genotype data file location, "-m" followed by the SNP map file name and "-s" followed by the set definition file.

The two optional flags are "-p" which controls the number of permutations performed by the program and "-o" which controls output destination. If not specified, *SLAT* will run 1,000 permutations and output result to standard out (e.g. screen).

4 Getting results

Upon successfully running, *SLAT* outputs measures of significance of association signals for each specified set.

The output contains following information:

- Set name as defined in set definition file.
- Chromosome, starting and ending genomic position of the set. These fields are shown only if the set is defined by genomic positions.
- Number of SNPs tested.
- Minimum p-value of the region by permutation test. Minimum p-value of the region is an alternative measure of significance for a defined set.
- Set p-value as proposed by (De la Cruz et al 2009).

A sample output from testing a set of genes is shown below,

```
-------------------------------------------------------------------------------
   Name   Chr     Start       End    SNP Number  Minimum p-value  Set p-value
   AP1B1  22    28048223   28109123    15         2.102100e-01    1.971970e-01
  RFPL1S  22    28157560   28162672    12         1.671670e-01    5.205200e-02
   RFPL1  22    28159126   28162998    12         1.671670e-01    5.205200e-02
    NEFH  22    28200773   28211833    17         1.941940e-01    9.009000e-02
   THOC5  22    28228710   28274290    15         3.753750e-01    3.023020e-01
NIPSNAP1  22    28275354   28301882    15         1.761760e-01    1.691690e-01
-------------------------------------------------------------------------------
```

5 References

- De la Cruz O, Wen X, Ke B, Song M and Nicolae DL (2009) "Gene, region and pathway level analyses in whole-genome studies" Genetic Epidemiology, (accepted).