Software implementation of rare variant testing strategy in Ionita-Laza et al. 2011

Iuliana Ionita-Laza

May 7, 2011

Note for users: Please report any bugs/issues to: ii2135@columbia.edu. A more user-friendly version will be released soon.

This file contains documentation on the software implementation of the testing strategy for association with rare variants in a genetic region of interest, as described in Ionita-Laza et al. (Plos Genetics, 2011). The implementation is in C++, and therefore very efficient for large scale sequencing studies.

Input and Output Files

Input Files There are three input files:

1. **data.ped** is a standard pedigree file. The first 6 columns are:

   - Family_ID
   - Individual_ID
   - Father_ID
   - Mother_ID
   - Gender - 1 for male/ 2 for female
   - Phenotype - 1 for control/2 for case

   The remaining columns encode the genotypes at individual SNV positions. Each genotype is coded in an additive fashion: 0, 1 or 2 (i.e. the number of minor alleles). See Table 1 for a toy example of a pedigree file with 6 SNVs and 10 individuals: 5 cases and 5 controls.

   Note: all fields in the pedigree file MUST be integers.

2. **gene_position.txt** contains information about the grouping of individual SNVs into units to be tested together, such as genes. Each line in this file corresponds to one

1	1	0	0	1	2	2	0	0	0	0	0
2	1	0	0	2	2	0	0	0	1	0	0
3	1	0	0	1	2	0	1	0	0	0	0
4	1	0	0	2	2	1	0	0	0	0	2
5	1	0	0	2	2	1	0	0	0	0	0
6	1	0	0	1	1	0	0	1	0	1	0
7	1	0	0	2	1	0	0	0	0	1	0
8	1	0	0	1	1	0	1	1	0	0	0
9	1	0	0	1	1	0	0	0	1	0	0
10	1	0	0	2	1	0	0	1	0	0	0

Table 1: Toy pedigree file: 10 individuals (5 cases and 5 controls), and 6 SNVs.

gene and contains the name of the gene, the number of the first SNV in the gene, and the number of the last SNV in the gene. These start and end positions correspond to the ordering of SNVs in the pedigree file. For example for the pedigree file above we can have two genes, the first contains SNVs 1 and 2, and the second contains SNVs 3, 4, 5, and 6.

Gene1	1	2
Gene2	3	6

Table 2: File with start and end positions of each gene

3. **weights.txt** contains information about the weighs of individual SNVs, one weight per variant (therefore the total number of weights in the file is the number of SNVs in the pedigree file). In the simplest scenario, when all variants are tested each weight is 1. If only non-synonymous variants are tested then weights are 1 for non-synonymous variants and 0 for synonymous variants. More generally, weights can be any real number.

Output Files There is only one output file: **results.txt**. This file contains empirical P-values for each gene in the **gene_position.txt** file. One-sided P-values for risk and protective variants respectively are reported. In addition, two-sided P-values corresponding to the max-statistic and sum-statistic in the paper are reported.

Running the program

1. **Source Code.** The source code is in the file called: **test_rare.c**.
2. **Compilation.** To be able to compile it, it is necessary to have the The GNU Scientific Library (GSL) installed. This library can be downloaded freely from **http://www.gnu.org/software/gsl/**.

To compile the program type the following command:

g++ -lm -l -lgsl -lgslcblas -o test_rare test_rare.c

3. **Execution.** The resulting executable can be run by typing:

   ./test_rare nind nsnvs ngenes freq npermut

   where

   - **nind** is the number of individuals (number of lines) in the pedigree file, data.ped.
   - **nsnvs** is the number of SNVs in the pedigree file, data.ped.
   - **ngenes** is the number of genes in the gene file, gene_position.txt.
   - **freq** is the maximum MAF on the variants to be tested (e.g., 0.01, or 0.03).
   - **npermut** is the number of permutations to be performed. For computational efficiency we recommend a multiple-step approach, with 1000 permutations for the first step, followed by increased number of permutations only for the genes that seem promising (say have empirical P-values less than 0.05 after the first step).

References

[1] Ionita-Laza I, Buxbaum J, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease, PLoS Genetics, 2011.