**STOPGAP2 Analysis Pipeline**

Judong Shen

Apr. 9th, 2015

STOPGAP (Systematic Target OPportunity assessment by Genetic Association Predictions) is a catalog of genetic associations and the genes that influence human traits. STOPGAP2 is a complete rebuild of the STOPGAP. This new version integrates more genetic association and functional genomics data to provide more comprehensive evaluation of the gene scores and variant-to-gene ranking hypotheses. The purpose of these data sets is primarily to support target validation, though there are a large variety of potential applications. STOPGAP2 uses various GWAS, LD data and functional annotation datasets, and some key algorithms for merging the GWAS datasets, generating LD r2 data, SNP locus clustering, variant-to-gene mapping, evaluating the gene scores and variant-to-gene ranking hypotheses to generate a set of STOPGAP2 datasets, which can be directly used for supporting target selection, validation, and repositioning decisions.

This STOPGAP2 analysis pipeline was implemented by R scripts, shell scripts, Python scripts and Perl scripts. The details of the scripts are as follows:

* STOPGAP2\_functions.R:
  + Most of the core functions are included in this file. These functions process and merge the three GWAS datasets, process various functional genomics datasets, identifying high LD SNPs and calculating the LD r2 from 1000 Genomes datasets (through the ./STOPGAP2\_LD/LDcalc.sh shell scripts), merging the LD SNPs with GWAS information and the functional genomics information, doing the LD SNP locus clustering, mapping variants to genes, evaluating the gene scores and variant-to-gene ranking hypotheses and generating multiple levels of STOPGAP2 datasets.
* STOPGAP2\_run.R:
  + The pipeline scripts to run STOPGAP2 R functions to generate STOPGAP2 datasets.
* ./STOPGAP2\_LD
  + LDcalc.sh: shell scripts to run the LD calculation, which run vcftools to calculate LD and python scripts to label LD.
  + Get\_rsID\_coord.py: python scripts to translate the rsids included in the GWAS association data to genomic coordinate.
  + Label\_LD\_results.py: python scripts to label the LD results with alleles and IDs.