Multi-Phenotype Association Script Guidelines

Deborah Weighill

February 22, 2019

1 Calculating Proportional Similarity (PS)

The starting point for this method is a tab-delimited GWAS profile matrix of significant associations, in which rows represent SNPs, columns represent phenotypes and entries being 0's (no significant association) or 1's (significant GWAS association).

The script `find_num_jobs_from_size.pl` assists in determining how many jobs to split the PS calculation into, depending on the size of the input matrix. This script generates a `job_index.txt` file.

Example usage:

```
perl find_num_jobs_from_size.pl [Input File: GWAS profile matrix] [Input Integer:
    approximate job size].
```

The `PS_mpi_efficient.pl` script uses MPI to calculate the similarity between all pairs of SNPs in a parallel fashion, using the number of jobs determined above. This script reads in the `job_index.txt` file.

Example submit script:

```
#!/bin/bash
#PBS -A BIF102
#PBS -V
#PBS -l nodes=9,walltime=2:00:00
#PBS -N czek
cd $PBS_O_WORKDIR
aprun -n 134 /usr/bin/perl PS_mpi_efficient.pl
```

The PS results are then thresholded and pruned to only include SNPs within MPA genes. This is done using the `mpi_prune_to_snps_in_pleiotropic_genes_and_thresh.pl` script.

Example submit script:

```
#!/bin/bash
#PBS -A BIF102
#PBS -V
#PBS -l nodes=9,walltime=1:00:00
#PBS -N prune_pleiotropy_thresh
cd $PBS_O_WORKDIR
aprun -n 134 /usr/bin/perl mpi_prune_to_snps_in_pleiotropic_genes_and_thresh.pl [Input
    File: list of PS results files] [Input File: list of SNPs in MPA genes]
```

Finally, we need to concatenate back in "self loops" for each SNP in MPA genes, so that SNPs with unique phenotype associations are not excluded from modules, as they should be assigned to their own, unique module. After this, one has the `PS_results_thresholded_SL` file.

A list of SNPs in MPA genes as well as the self loops can be constructed using the `get_snps_in_MPA_genes.p` script. This script uses as input the GWAS results which have been mapped to genes in a tab-delimited file with the following column order:

```
phenotypeID   Pvalue snpID  GeneID PhenotypeAnnotation  BetaValue
```

Example usage:

```
perl get_snps_in_MPA_genes.pl [Input File: SNP-gene GWAS file] [Output File: SNPs in
    MPA genes] [Output File: self loops for SNPs in MPA genes]
```

2 Construct GP Matrix

The GP matrix is easily constructed from GWAS results by mapping SNPs to genes, and pruning the results to maintain genes which contain SNPs collectively associating with multiple phenotypes, i.e. multi-phenotype association genes (MPA genes). This uses as input a list of MPA genes and the GWAS results which have been mapped to genes in a tab-delimited file with the following column order:

```
phenotypeID   Pvalue snpID  GeneID PhenotypeAnnotation  BetaValue
```

Example usage:

```
perl make_GP_network.pl [Input File: list of MPA genes] [Input File: SNP-gene GWAS
    file] [Output File: GP network SIF file]
```

3 Construct Modules

Modules are constructed by converting the `PS_results_thresholded_SL` file to SIF format and clustering with MCL (Available from `https://micans.org/mcl/`), which efficiently extracts connected components when the PS threshold set was 1. The `assign_module_ids.pl` script assigns module IDs to the MCL clusters.

Example usage:

```
cut -f1-3 [Input File:PS_results_thresholded_SL] >[Output File:
    PS_results_thresholded_SL.sif]

mcl [Inpur File: PS_results_thresholded_SL.sif] --sif -I 2 -o [Output File: mcl output
     file]

perl assign_module_ids.pl [Input File: mcl output file] [Output File: modules output
    file]
```

4 Construct GM Matrix

The GM matrix is constructed using the `make_GM_network.pl` script. This requires a tab-delimited SNP-gene-phenotype map file from the GWAS results with the following column order:

```
phenotypeID   Pvalue snpID  GeneID PhenotypeAnnotation  BetaValue
```

The script also requires a file containing a list of MPA genes with one ID on each line, as well as the modules file constructed previously.

Example usage:

```
perl make_GM_network.pl [Input File: SNP-gene GWAS file] [Input File: list of MPA
   genes] [Input File: modules file] [Output File: GM network SIF file]
```

5 Construct MP Network

Example usage:

```
perl make_MP_network.pl [Input File: self loops file] [Input File: modules file] [
   Output File: MP network SIF file]
```

These SIF file networks can be visualized in Cytoscape.