DAP-G User Manual

Contents

1 Compilation

The following C/C++ libraries are required for compiling the source code

- GNU GSL library
- OpenMP compiler (many popular compilers are compatible)

Simply run `make` to compile the executable named `dap`.

Run `make static` to compile an executable with static linked library.

2 Command-line Syntax

```
dap-g -d data_file |  -d_z zvalue_file -d_ld ld_file |
 -d_est effect_estimate_file -d_ld ld_file -d_n sample_size -d_syy syy
 [-g grid_file] [-p prior_file] [-msize K] [-ld_control r2_threshe]
 [-converg_thresh thresh_value]
 [-t nthread]
 [-o output_file] [-l log_file] [--output_all]
```

Important Tip: Run adaptive DAP algorithm with multi-thread option (`-t nthread`) whenever possible! It will significantly speed up the computation.

3 Input Data Options

DAP-G accepts three different types of input data:

- Individual-level genotype and phenotype data in sbams format. For this option, use `-d sbams_data_file` to speify the single input file.
- Z-scores for each candidate SNP from single-SNP analysis and an LD matrix. Two input files are required for this option, use `-d_z zval_file -d_ld LD_file` to specify the z-score and LD (genotype correlation) files, respectively.
- Estimated effect size (beta-hat) and corresponding standard error for each candidate SNP from single-SNP analysis, an LD matrix, plus two numbers: sample size and total sum of squares of the outcome variable (denoted by Syy). The two input files containing single SNP association estimates and LD information are specified by `-d_est estimate_file -d_ld LD_file`, respectively; use `-d_n sample_size -d_syy Syy` to provide numeric values for the sample size and Syy.

3.1 Individual-level Data

The input data file for phenotype-genotype data are in "sbams" format. It is a text file originally designed to represent data from a meta-analytic setting including data from multi ple subgroups. Note that, the *current* implementation of the DAP algorithm only supports a single subgroup (the multi-group feature is maintained in version 1). We will add the multi-group feature back in the near feature.

The file contains a phenotype section, a controlled covariate section and a genotype section.

In the phenotype section, each line contains a list of measured traits of all individuals in a subgroup, i.e.,

```
pheno  pheno_id group_id exp_ind_1 exp_ind_2 ... exp_ind_n
```

The leading "pheno" is a keyword that indicates the line encodes phenotype measurements. The pheno_id field contains a character string that denotes the name of the phenotype. The group_id field is a character string that uniquely labels a specific subgroup. The following entries are numerical values of expression levels of individual 1 to n for the target gene in the indicated subgroup. Note, because each subgroup my have different number of individuals, the length of each line in this section can differ

```
geno snp_id  group_id geno_ind_1 geno_ind_2 ... geno_ind_n
```

The leading "geno" is a keyword that indicates the line encodes genotypes. The snp_id field contains a character string that denotes the ID of a SNP. The additional group_id field is not

used in the current implementation of dap-g. The remaining entries are genotypes in dosage format (i.e, 0,1 or 2, or any fractional number in [0,2] if genotypes are imputed).

The controlled covariates are coded in the similar format lead by the keyword "controlled", i.e.,

```
controlled variable_name group_id value_ind_1 value_ind_2 ... value_ind_n
```

Use command line option `-d sbams_data_file` to inform dap-g that inidividual-level data is supplied.

Example (Individual-level data):

```
dap-g -d sample_data/sim.1.sbams.dat
```

3.2 Sufficient Summary Statistics

The sufficient summary statistics refer to the following information

1. a p-vector of estimated effect size from single SNP testing, b_i for each SNP i.
2. estimated standard error for b_i for each SNP i.
3. a p x p correlation matrix of p SNPs, R (typically obtained from an appropriate population panel)
4. sample size, n
5. total sum of squares (SST) for the quantitative trait, SST

To use sufficient summary statistics, two input text files and two numeric values are required to supply via the command line.

- Estimates file: contains effect size estimates and corresponding standard errors for all p SNPs. The required format is as follows

```
snp_name_i  b_i  se(b_i)
```

Note that each line represents a single SNP. Use the command line option `-d_est estimate_file_name` to specify the estimates file.

- LD file: contain the ***correlation matrix*** between p SNPs. The order of the SNP is required to match the order listed in the estimates file. Use the command line option `-d_ld LD_file_name` to specify the LD file.

- Sample size: a numeric value. Use option `-d_n sample_size` to specify the sample szie.
- Total sum of squares of phenotype: a numeric value. Use option `-d_syy SST_value` to specify SST.

Example (sufficient summary statistics):

```
dap-g -d_est sample_data/sim.1.est.dat -d_ld sample_data/sim.1.LD.dat -d_n 343 -d_syy
```

Note that the using sufficient summary statistics and individual-level should yield the same results (except small numerical difference).

3.3 Z-statistics and LD Information

DAP-G also accepts input in the format of z-scores (from single-SNP testing) and LD matrix. In particular, two input files are required for this format

- z-score file: contain z-scores for all p candidate SNPs. The text file has two columns, the specific format is as follows:

```
snp_name_i  z_i
```

Use option `-d_z z_score_file` to specify the z-score file.

- LD file: contain the ***correlation matrix*** between p SNPs. The order of the SNP is required to match the order listed in the z-score file. Use the command line option `-d_ld LD_file_name` to specify the LD file.

Example (z-score + LD information)

```
dap-g -d_z sample_data/sim.1.zval.dat -d_ld sample_data/sim.1.LD.dat
```

Note the output from this type of input will differ from the other two input types, and the corresponding results are typically more conservative.

3.4 Extracting Summary Statistics from sbams File

Both types of summary satistics can be extracted from the sbmas-format individual-level data, if it is already available.

- Extract z-scores and LD info

```
dap-g -d sbams_file --dump_summary
```

A text file `LD.dat` containing correlation matrix and a text file `zval.dat` will be generated.

- Extract sufficient summary statistics

```
dap-g -d sbams_file --dump_summary2
```

A text file `LD.dat` containing correlation matrix and a tex file `est.dat` containing all (b_i, se(b_i)) will be generated. Sample size and SST information will be output to standard error.

4 Command-line Arguments

4.1 Input File Options

- `-d sbams_file`: specify sbams file with individual-level phenotype and genotype data
- `-d_est effect_est_file`: specify effect size estimates file
- `-d_ld  LD_file`: specify file containing correlation matrix between SNPs
- `-d_n sample_size`: specify the sample size
- `-d_syy SST`: specify SST

4.2 Prior Specifications

The following options are used to specify the exchangeable prior probability that a candidate SNP being causally associated (denoted by pi1).

- `-ens expected_number_of_signal`: specify prior expected number of signals. Note that, pi1 = expected_number_of_signal/p
- `-pi1 probability`: specify the exchangeable prior probability.

By default, `ens` is set to 1, or equivalently, pi1 = 1/p.

To specify non-exchangeable prior, use a text file with 2 columns

```
snp_name   prior_probability
```

This option is best used if genomic annotation information is available. Program like `[TORUS](https://githu` can estimate such priors from data.

- `-p prior_file`: specify the non-exchangeable prior

4.3 Run Options

- `-t thread_number`: specify the number of parallel threads to run DAP algorithm
- `-ld_control r2_threshold`: specify the LD threshold to be considered within a single signal cluster. SNPs within a signal cluster should be correlated, the threshold here determines how strong the genotype correlation between member SNPs in a signal cluster. If `ld_control` is set to 0, the behavior of the DAP-G algorithm is similiar to the previous version of DAP, where signal cluster is not rigorously defined. Higher threshold should typically reduce the size of inferred signal cluster and speed up the computation. **By default, the threshold is set to 0.25**.

- `-msize maximum_model_size`: specify the maximum size of model dap-g explores. Valid maximum model size ranges from 1 to p. **By default, it is set to p**, i.e., there is no restriction on how large the true association model can be. If it is specified, the DAP-G runs DAP-K algorithm and stops at the specified maximum model size.
- `-converg_thresh thresh_value`: specify the stopping condition for model exploration. DAP-G computes log10 normalizing constant up to the current model size, denoted by log10_NC(K). Exploration of larger model size ends if log10_NC(K) - log10_NC(K-1) is no greater than `thresh_value` and we consider the model is saturated. **By default, `thresh_value = 0.01`.**

4.4 Output Options

- `-o output_file`: specify the output file name. **By default, the output goes to standard out, i.e., screen**.
- `-l log_file`: specify the log file name. **By default, the messages during the run output to standard error.**
- `--output_all` (or `--all`): Output information for all SNPs and all signal clusters. **By default, only SNPs with PIP $>$ 0.001 and signal clusters with SPIP $>$ 0.25 are displayed**

4.5 Miscellaneous Options

We do not recommend users to change the following setting, but these options may be useful for specific situations (e.g., customized applications, speeding up computation).

- `-g grid_file`: specify the text file containing effect size priors to be integrated out. By default, the marginal likelihood/Bayes factor is computed by averaging over a default grid of prior effect sizes. The default prior size is typically sufficient for both QTL mapping and GWAS analysis. User can specify an alternative set of effect size priors (in terms of squared expected effect sizes) in a text file with a single column, and each line represent a unique prior expected squared effect size.
- `-size_limit maximum_cluster_member_size`: specify the maximum number of variants allowed in a signal cluster. By default, there is no constraint and the size of each signal cluster is completely data determined. Setting a small number of maximum_cluster_member_size forces DAP-G to cap the number of variants into each cluster and reduces computation.

DAP-G can also be used to perform following data processing tasks besides multi-SNP association analysis

- `--scan`: perform single-SNP analysis (instead of multi-SNP analysis) for the given data. log10 Bayes factors for all candidate SNPs are computed and output.

- `--dump_summary`: extract z-scores and LD correlation matrix from the individual-level data. Must use with `-d sbams_file` option./output
- `-dump_summary2`: extract effect size estimates (b_i and se(b_i)) for each candidate SNP along with LD correlation matrix, sample size and SST information. Must use with `-d sbams_file` option.

5 Output Format

There are roughly three sections in the DAP-G output. They correspond to the summaries of association models, individual SNPs and signal clusters, respectively.

5.1 Model Summary

An example line from model summary section is shown below:

```
    1   4.1846e-02    3     10.938   [rs54] [rs927] [rs986]
```

Specifically, the first column denotes the rank of the model; the second column shows the posterior probability of the corresponding model; the third column indicates the size (i.e., the number of SNPs) of the model; the fourth column shows the un-normalized posterior score of the model (defined as log10[model prior]+log10[BF]); and the last column gives the exact configuration of the model.

To extract the model summary portion of the output, use the following command

```
grep "\[" dap_output_file
```

5.2 SNP Summary

An example line from SNP summary section is shown below:

```
((1))              rs54 5.74985e-01   8.167 1
```

Specifically, the first column denotes the rank of the SNP (measured by PIP); the second column indicates the SNP name; the third column shows the corresponding posterior inclusion probability (PIP); and the last column provides the corresponding signal cluster ID ("-1" is shown if a SNP is not considered a member of any signal cluster).

To extract the SNP summary portion of the output, use the following command

```
grep "((" dap_output_file
```

5.3 Signal Cluster Summary

An sample output for signal cluster summary is shown below

```
cluster     member_snp  cluster_pip      average_r2
   {1}          4        9.964e-01      0.934             0.934   0.003   0.003
   {2}          7        9.964e-01      0.915             0.003   0.915   0.001
   {3}         10        9.943e-01      0.855             0.003   0.001   0.855
```

The last three columns represent the average LD measures (r^2) for SNPs within a cluster and between clusters.

To extract the cluster summary portion of the output, use the following command

```
grep "{" dap_output_file
```