G-STRATEGY Software Documentation

Version 1.0
Aug 23, 2016

Miaoyan Wang[1,2] and Mary Sara McPeek[2,3]

Department of Mathematics[1]
The University of Pennsylvania, Philadelphia, PA, USA

Departments of Statistics[2] and Human Genetics[3]
The University of Chicago, Chicago, IL, USA

G-STRATEGY
A C program for selecting the optimal subset of individuals for sequencing based on phenotypes and pedigrees

Homepage: `http://www.stat.uchicago.edu/~mcpeek/software/index.html`
Release 1.0 Aug 23, 2016

We request that use of this software be cited in publications as follows:
M. Wang, J. Jakobsdottir, A.V. Smith and M. S. McPeek (2016). G-STRATEGY: Optimal Selection of Individuals for Sequencing in Genetic Association Studies. Genetic Epidemiology 40: 446-460.

Contents

1 Overview of G-STRATEGY 4

2 Installing G-STRATEGY 5

3 Running G-STRATEGY 5

4 Input 6

5 Output 10

6 Example 11

7 FAQ 11

8 Bug reports and feedback 11

9 Acknowledgements 12

1 Overview of G-STRATEGY

G-STRATEGY is a program, written in C, that implements the optimal selection of individuals for sequencing in genetic association studies. It uses simulated annealing to choose individuals for sequencing based on phenotypes, covariates and pedigrees in such a way that the statistical power to detect association is maximized.

G-STRATEGY can be thought of as an innovative extension of the classical "selective genotyping" strategy, in which a selected portion of the phenotyped individuals are genotyped. Unlike classical selective genotyping, G-STRATEGY is applicable to the full spectrum of study designs, ranging from combinations of unrelated individuals and small families to individuals sampled from a complex, inbreed pedigree. The key idea of G-STRATEGY is based on the retrospective association model, i.e., we condition on phenotype (and covariates) and treat genotype as random. As a result, G-STRATEGY determines whom to sequence by taking into account not only the phenotypes of the selected (and thus genotyped) individuals, but also the phenotypes of unselected individuals from the same pedigree. Another advantage of G-STRATEGY is that it also accounts for the effects, on the association test, of the statistical dependence among relatives' genotypes and among their phenotypes.

To understand the purpose of this program and how it works, we highly recommend users to read the paper [1]. In particular, G-STRATEGY has the following features:

1. G-STRATEGY would be particularly well-suited to situations in which power for an association test with a given phenotype is a priority. It would also be well-suited to the situation in which low-density genotype information is not necessarily available on all individuals.

2. G-STRATEGY has good control of type 1 error in both retrospective and prospective association analysis.

3. G-STRATEGY and its simplified version "Extreme Enrichment" strategy provide substantial power improvement over other selection methods. Both strategies have been shown efficiently and effectively for a broad range of disease models, including many binary trait models and quantitative trait models with covariates.

4. The software provides user-friendly options to allow users to specify (1) an initial subset of individuals who are already sequenced or genotyped (referred to as the "previously genotyped" individuals) and (2) a subset of individuals available to be selected for sequencing or genotyping (referred to as the "feasible" individuals), and it optimizes the selection of individuals for genotyping within these constraints.

5. Multiple runs of G-STRATEGY can provide multiple choices of subset to be genotyped that are expected to provide approximately equal power. This can be particularly useful if there are other considerations (such as cost or convenience) that arise in practice.

2 Installing G-STRATEGY

1. Download the G-STRATEGY package. This package contains GNU GPL license in file `gpl.txt` and four subdirectories:

   - `executables` contains a pre-compiled binary executable file for x86 64bit Linux and a binary executable file for MAC;
   - `src` contains the source code;
   - `doc` contains this document `G-STRATEGY_Documentation.pdf`;
   - `examples` contains example input and output files.

2. Read the file `G-STRATEGY_Documentation.pdf` carefully to understand the purpose of this program and how it works.

3. If none of the pre-compiled binaries are suitable for your use, follow the steps (1)-(3) to compile G-STRATEGY yourself:

   (1) Switch to the `src` directory by typing `cd src/`

   (2) Type `make`. This will build an executable program called `G-STRATEGY` and its auxiliary program `mastor_tpr`.

   (3) Type `make clean` to remove the `*.o` files.

   To compile G-STRATEGY on your own machine, you will need GCC including the standard C and Fortran libraries. If GCC is not already available on your machine, it is freely available at `https://gcc.gnu.org`.

3 Running G-STRATEGY

To run the executable program:
First, prepare the input files, see Section 4. Then, to run G-STRATEGY with the default input filenames, one need only type:

```
./G-STRATEGY -b
```

or

```
./G-STRATEGY -q
```

Alternatively, to change the input filenames or use other options, use flags in the command line. For example, the following commands are possible:

```
./G-STRATEGY -p phenotype.txt -k kinship.txt -s size.txt -b
./G-STRATEGY -p phenotype.txt -k kinship.txt -s size.txt -b -e
./G-STRATEGY -p phenotype.txt -k kinship.txt -s size.txt -q
```

```
./G-STRATEGY -p phenotype.txt -k kinship.txt -s size.txt -q -e
```

We briefly summarize the meaning of the flags below. More details can be found in Section 4.

- **-b** or **-q** This flag is required and instructs the program to choose between the MASTOR [2] and the M_QLS [3] objective functions. If **-b** is used, the program will use the M_QLS objective function, and if **-q** is used, the program will use the MASTOR objective function. See Section 4 for guidelines as to which one to choose.

- **-p phenotype.txt** This option allows users to specify the name of the phenotype input file. Each row of this file contains the information of an individual, including the family ID, individual ID, father ID, mother ID, sex, a 0/1 boolean value indicating whether or not this individual is previously genotyped, a 0/1 boolean value indicating whether or not this individual is feasible to be selected, the phenotypic value of interest, and covariates if any (see Section 4). The filename defaults to `phenotype.txt`.

- **-k kinship.txt** This option allows users to specify the name of the kinship coefficient input file. This file contains the kinship and inbreeding coefficients for all possible pairs of individuals within each family. Filename defaults to `kinship.txt`.

- **-s size.txt** This option allows users to specify the name of size input file. This file contains the total number of sampled individuals, the total number of families in the sample, the target number of individuals to be selected for genotyping, and the population prevalence (when `-b` is used) or the number of covariates (when `-q` is used). Filename defaults to `size.txt`.

- **-e** This flag is optional and instructs the program to run "Extreme Enrichment" strategy as an alternative to G-STRATEGY. Enrichment strategy can be viewed as a simplified version of G-STRATEGY, in which individuals with extreme enrichment values are prioritized to be selected. If **-e** is used, the program will only run Enrichment strategy, and the default output is in `g_enrichment.txt`. If **-e** is not used, the program will run G-STRATEGY, and the default output is in `g_strategy.txt`.

4 Input

1. **-b** or **-q** (required flag)

   This flag is used to choose between the M_QLS objective function and MASTOR objective function in the analysis. The `-b` flag indicates the M_QLS objective function, and the `-q` flag indicates the MASTOR objective function. Users must specify one of these two flags in the command. If neither of them is used, the program will report an error.

The M_QLS objective function would be well-suited to case-control studies with non-random ascertainment. The current version does not adjust for covariates. The MASTOR objective function would be well-suited to association studies with covariates and random additive polygenic effects, and it can be applied to a binary trait as well as a quantitative trait. The specific context to which the MASTOR or the M_QLS objective function is applicable can be found in [1].

2. **-p phenotype file**

   The phenotype input file contains the phenotype information for each sampled individual.

   It has the standard format:

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	1	0	0	0	0	1	0.2
1	2	0	0	1	0	0	1.1
1	3	1	2	1	0	1	−0.4
1	4	1	2	1	0	1	0
2	1	0	0	0	1	1	−1.5
2	2	0	0	1	0	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	

   (1) Family ID

   (2) Individual ID

   (3) Father ID (0=unknown)

   (4) Mother ID (0=unknown)

   (5) Sex (0=male, 1=female)

   (6) A 0/1 boolean value indicating whether this individual has been previously genotyped (Yes=1, No=0)

   (7) A 0/1 boolean value indicating whether this individual is available to be selected for genotyping (Yes=1, No=0)

   (8) Case-control status or the phenotypic value of interest

   (9) Covariate 1 (optional)

   (10) Covariate 2 (optional)

   (11) $\cdots$

   The requirements are following:

   - Families should be numbered by (not necessarily consecutive) integers.
   - Individuals at each family should be numbered by integers. The individual ID should be unique across all families.

- Sampled individuals who are unrelated to anyone else in the sample should be included in this file by giving each such person their own unique family ID.
- All father and mother IDs must be present in the file. If they are founders, they have zero for their father and mother IDs.
- If an individual has numerical value `1` in the column (6), he/she must also have numerical value `1` in the column (7). Namely, every perviously genotyped individuals should be regarded as being available to be selected for genotyping.
- If the user chooses the `-b` flag in the command, then the column (8) should be the case-control status (0=unknown, 1=unaffected, 2=affected). In this case, the user should not specify any covariates in the phenotype file.
- If the user chooses the `-q` flag in the command, then the column (8) should be the phenotypic value of interest, the column (9) should be the 1st covariate, and the column (10) should be the 2nd covariate, and so on. Intercept should NOT be included as a covariate in this case. The program will automatically add an intercept and adjust for the covariates in the analysis. The missing value code for phenotype and covariates is -9.0.

3. **-k kinship file**

The kinship file contains the kinship and inbreeding coefficients for each pair of individuals in each family with format:

(1)	(2)	(3)	(4)
1	1	1	0
1	1	2	0
1	1	3	0.25
1	1	4	0.25
⋮	⋮	⋮	⋮
1	4	4	0
2	1	1	0
2	1	2	0
2	1	3	0.25
⋮	⋮	⋮	⋮

(1) Family ID

(2) Individual ID1

(3) Individual ID2

(4) Kinship coefficient between ID1 and ID2 if ID1 is different from ID2; inbreeding coefficient if ID1 equals ID2

Everyone in the phenotype file should also be in the kinship file, and vice versa. The software program that can be used to obtain kinship and inbreeding coefficients is `KinInbcoef` program, which can be found at `http://www.stat.uchicago.edu/`

`~mcpeek/software/KinInbcoef/index.html`. The output file of the `KinInbcoef` program has the exact format required for this input.

4. **-s size file**

   This file contains four numbers. The first three numbers specify the problem size (as follows), and the fourth number specifies some additional information required by the program, depending on the `-b` or `-q` flag.

   The first three numbers are:

   (1) The total number of sampled individuals;

   (2) The total number of families in the sample;

   (3) The target number of individuals to be selected for genotyping (including those who are previously genotyped);

   The fourth number is:

   (4a) The population prevalence of the binary trait (when `-b` flag is used);

   (4b) The number of covariates (when `-q` flag is used).

   These four values can be either in a single row or a single column, where the order should correspond to the definition above. The requirements are the following:

   - The total number of sampled individuals should match with the number of individuals in the phenotype file.
   - The total number of families in the sample should match with the phenotype and kinship files.
   - The target number of individuals to be selected should be greater than the number of previously genotyped individuals, and fewer than the number of feasible individuals. These two boundary numbers should match with the column (6), and the column (7), respectively, in phenotype file.
   - If the users chooses the `-b` flag, then the fourth number should be the population prevalence of the binary trait. The population prevalence should be a decimal fraction between 0 and 1. We recommend using an estimate from previous studies or registry data from the population.
   - If the users chooses the `-q` flag, then the fourth number should be the number of covariates (excluding the intercept). The number of covariates should be a non-negative integer and match with the phenotype file. Note that intercept should NOT be counted as a covariate in this case. The program will automatically add an intercept and adjust for the covariates in the analysis.

   The software will check all these rules before starting optimization and stop if any of them are violated.

5. **-e** (optional flag)

   When `-e` is used, the program will run the Enrichment strategy as an alternative to G-STRATEGY. Both G-STRATEGY and the Enrichment strategy make use of the enrichment principle and have been shown to consistently outperform other common strategies. The Enrichment strategy selects the subset of individuals that have extreme enrichment values. G-STRATEGY takes one step further, and in prioritizing individuals for genotyping, it also accounts for the effects, on the association test, of the dependence among relatives' data. As shown in [1], G-STRATEGY typically outperforms the Enrichment strategy in terms of power for association, though the differences can be small.

5 Output

By default, program will output four files, one result file and three diagnostic files.

1. `g_strategy.txt` The result file contains the subset of individuals selected using G-STRATEGY. It has the following format:

(1)	(2)	(3)	(4)
1	1	0.076317	0
1	3	−0.129716	0
2	1	−0.594978	1
2	2	0.349898	0
⋮	⋮	⋮	⋮

   (1) Family ID

   (2) Individual ID

   (3) Enrichment value

   (4) A 0/1 boolean value indicating whether this individual has been previously genotyped (Yes=1, No=0).

   All previously genotyped individuals will be automatically included in this output file.

2. The three diagnostic files are

   - `objective_trajectory.txt` This file records the objective values in the simulated annealing run.
   - `accept_rates.txt` This files records the acceptance rate at each temperature in the simulated annealing run.
   - `objetive_up.txt` This file records the objective values in the post-processing stage.

3. `g_enrichment.txt` This output file is optional and is only generated when the user chooses the `-e` flag. The file contains the subset of individuals selected using Enrichment strategy. It has the exactly the same format as `g_strategy.txt`.

6 Example

Example input and output files are provided in the directory `G-STRATEGY/examples`. The input files are: `phenotype_ex.txt`, `kinship_ex.txt` and `size_ex.txt`. You can move the example input files to the same directory as the binary executables `G-STRATEGY mastor_tpr`, and then run the program by typing:

```
./G-STRATEGY -p phenotype_ex.txt -k kinship_ex.txt -s size_ex.txt -q
```

This generates four output files `g_strategy.txt`, `objective_trajectory.txt`, `objective_up.txt`, and `accept_rates.txt`. Compare them to the example output files `*_ex.txt`. Note that your results might not be identical to the supplied results because of the probabilistic nature of the algorithm. Actually, running `G-STRATEGY` multiple times on a large data set will typically produce multiple different choices of the subset of individuals for sequencing. These different solutions are expected to be approximately equal in power, which provides great flexibility to users if there are other considerations (such as cost or convenience) that arise in practice.

7 FAQ

1. **MZ twins** The current version of the program does not handle the co-occurrence of MZ twin pair in the phenotype file.

2. **What if the pedigree structure is not available?** The G-STRATEGY depends on the pedigree structure only through the input kinship and inbreeding coefficients. The father IDs and mother IDs in the pedigree and phenotype data file are not used by the program, and thus their values can be arbitrarily assigned for the sake of this program only.

8 Bug reports and feedback

We welcome comments and suggestions and if you do encounter a bug in the G-STRATEGY software please send us a message. Please include in your message the program version (printed out when the program is run), platform (windows, mac, linux, etc.), description of your problem, and if possible example files (in a zip folder) that caused the problem.

9 Acknowledgements

1. Numerical Recipes in C. We use the utility functions in `nrutil.h`.

2. LAPACK. We used matrix computation subroutines from `CLAPACK`.

References

[1] Wang M., Jakobsdottir J., Smith A. V., McPeek M. S. (2016). G-STRATEGY: Optimal Selection of Individuals for Sequencing in Genetic Association Studies. Genetic Epidemiology. 40: 446-460.

[2] Jakobsdottir, J., and McPeek, M. S. (2013). MASTOR: Mixed-model association mapping of quantitative traits in samples with related individuals. American Journal of Human Genetics 92, 652-666.

[3] Thornton, T., and McPeek, M. S. (2007). Case-control association testing with related individuals: A more powerful quasi-likelihood score test. American Journal of Human Genetics 81, 321-337.