# Scout **Software for combined analysis of unrelated subjects and triads**

Emory University School of Medicine
Department of Human Genetics
615 Michael Street, Suite 301
Atlanta, GA 30322

Version 1.00, January 2006
Comments about this documentation should be sent to scout@genetics.emory.edu.

http://www.genetics.emory.edu/labs/epstein/software/scout/index.html

# Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Purpose

**SCOUT** is a software program for conducting combined association analysis of triads and unrelated subjects by utilizing the statistical methodology developed in Epstein et al. (2005). For a thorough description of this methodology, please review the paper (which is available for download here).

## 1.2 Background

With the public release of high-density maps of single-nucleotide polymorphisms (SNPs), genetic association approaches are becoming popular for identifying genetic variants that influence complex disease. Statistical methods for detecting association generally compare SNP alleles from a sample of affected cases with alleles from an appropriate sample of unaffected controls. The choice of the control sample often becomes a source of deliberation. Some studies collect controls that consist of the parents of the case subjects. The parental controls together with the affected case offspring form a unit commonly referred to as a triad. Using triads, one can test for association between a SNP and disease by assessing whether particular SNP alleles are preferentially transmitted from parent to affected child (Spielman et al., 1993). Such preferential transmission can be assessed using either a McNemar statistic (often defined in this situation as the Transmission/Disequilibrium Test) or a likelihood-based procedure called the conditional-on-parental genotypes (CPG) approach (Schaid and Sommer, 1993). As an alternative to using triads for analysis, one can also test for association using healthy unrelated subjects as a source of controls. In this situation, one can test for association by assessing whether SNP allele frequencies differ among unrelated cases and controls using standard goodness-of-fit statistics.

Each type of control sample has its own set of strengths and weaknesses, which are described in Epstein et al. (2005). Nevertheless, situations may arise where a study may have a collection of triads, unrelated controls, and (possibly) unrelated cases for genetic association analysis. In this situation, we developed a likelihood-based procedure that conducts joint association analyses of genetic data from both triads and unrelated subjects. Our procedure allows for flexible modeling of SNP allele effects on disease and relaxes assumptions of random mating and Hardy-Weinberg Equilibrium (HWE) that are sometimes made in genetic association analyses. Also, our likelihood procedure can account for missing genetic data in the triads using an Expectation-Maximization (EM) algorithm. Finally, our procedure provides formal tests on whether different data types (triads, unrelated controls, and unrelated cases) can be safely combined for inference (details on the importance of such tests are described in Epstein et al. (2005).

**SCOUT** is a software package that implements the above procedure. Given a user-defined genetic model, the program provides estimates of genotype relative risk (RR) parameters for the SNP of interest and also tests whether such RR parameters significantly differ from null values corresponding to no influence on disease. Based on the results of a **SCOUT** analysis, a researcher can identify genetic variants that are associated with disease, which hopefully will facilitate positional cloning efforts.

## 1.3 Assumptions

**SCOUT** makes the following assumptions:

1. all genetic markers are biallelic (e.g. SNPs), autosomal, and codominant,
2. genotyping and phenotyping are performed without error,
3. missing genotype data are missing at random.

## 1.4 System requirements and installation

**SCOUT** is a Win32 application written in Fortran 90 using numerical routines provided in the International Mathematical and Statistical Library (IMSL). This software has been tested on both Windows XP and 2000 machines. **SCOUT** has not been tested on earlier versions of the Windows OS or on UNIX platforms but is unlikely to perform correctly on these systems.

The **SCOUT** executable, documentation, and a few sample files are contained in both zipped and tar-gzipped packages that can be downloaded from http://www.genetics.emory.edu/labs/epstein/software/scout/index.html. When either package is uncompressed, the documentation and sample data will be in written to the respective directories, `doc/` and `sample_data/`. The sample data sets are contained in three files: one file of triad data, one file of control data only, and one file of both control and case data. These will be used in the example analyses described in this documentation. Also included in the distribution is a sample script file script.cpg [Something about script here once the examples are better thought out.]

The program is a self-contained executable and does not require any complicated installation. The executable can be placed in and run from any location that is accessible to the user. A few dynamic link libraries may be necessary for the software to execute properly. Namely,

- DFORRT.DLL is the Fortran runtime library
- MSVCRT.DLL is the Microsoft C Runtime Library

These external libraries are included, legally, within the distribution. They either must be kept in the same directory as the executable or copied to the `/windows/system32` directory. If either of these files already exists in your directory, ensure that you are not replacing a newer version with an older version of the library.

The organization of the files within the uncompressed distribution are shown here:

```
|-- scout.dll
|-- DFORRT.DLL
|-- MSVCRT.DLL
|-- README
|-- scout_cl.exe
|-- doc
|   '-- scout.pdf
|-- sample_data
|   |-- case-control.dat
|   |-- control.dat
|   '-- triads.dat
'-- script.cpg
```

## 1.5 User support

While we have tried to carefully document **SCOUT** and make it easy to use, problems may occur. Some problems can likely be resolved by reading through this documentation. In the event that problems do occur with the software, we would appreciate being notified so that these problems may be rectified. Such problems or questions either about running **SCOUT** or the statistical methodology involved should be directed to .

## 1.6 How to cite this software

If you use **SCOUT** to analyze data for publication, we ask that you make the following citations:

- M.P. Epstein, C.D. Veal, R.C. Trembath, J.N. Barker, C. Li, G.A. Satten (2005). Genetic association analysis using data from triads and unrelated subjects. *Am. J. Hum. Genet.*, 76:592-608.

# 2 Inputting data

## 2.1 Data format: triads

**SCOUT** requires the sample data for triads to be in either a space- or tab-delimited text file. Within the file, one can insert a comment line by placing a hash symbol, #, at the beginning of the line. Otherwise, each row of the file must consist of three entries: a family index, a subject index and the subject's genotype code. These columns are summarized here:

| Column | Data value | Allowed values |
|---|---|---|
| 1 | Family index | up to 16 characters |
| 2 | Subject index | 1=father, 2=mother, 3=offspring |
| 3 | Genotype code | -1, 0, 1, 2 |

The family index can be any string of up to 16 characters in length and this index should be unique for each triad. The subject index specifies within the triad whether the subject is the father, mother, or affected offspring. The set of genotype codes is summarized in Table 2.1. Also, an example subset of data showing three triads is shown in Table 2.2.

Note also that if an affected offspring's genotype data is missing, then that trio will be ignored during the analysis. In this case, a warning about missing offspring genotype values will be reported to the console when the data are loaded.

Table 2.1: Summary of genotype codes for a given pair of alleles at a single locus. Missing data must be assigned a genotype code of -1. Column 3 of data file must contain exactly one of these values.

| Allele 1 | Allele 2 | **Genotype code** |
|---|---|---|
| 0 | 0 | **0** |
| 0 | 1 | **1** |
| 1 | 0 | **1** |
| 1 | 1 | **2** |
| missing data | | **-1** |

Table 2.2: Part of a sample data set showing three triads.

| Fam ID | Member | Genotype |
|---|---|---|
| 683b | 1 | 1 |
| 683b | 2 | 0 |
| 683b | 3 | 0 |
| 5167 | 1 | 0 |
| 5167 | 2 | -1 |
| 5167 | 3 | 0 |
| d3e0 | 1 | 1 |
| d3e0 | 2 | 0 |
| d3e0 | 3 | 1 |

## 2.2 Data format: unrelated subjects

The data for unrelated subjects (controls and/or cases) should also be in a space- or tab-delimited text file. Within the file, one can insert a comment line by placing a hash symbol, #, at the beginning of the line. Each row of data must contain three entries: a subject index, a subject disease index, and a subject genotype code. These columns are summarized in Table 2.3. The genotype codes for unrelated subjects are the same as those used for triads, which are summarized in Table 2.1. Missing genotype data in unrelated subjects are simply ignored in analysis, unlike missing parental data in the triads.

Table 2.3: Data specification for unrelated subjects.

| Column | Data value | Allowed values |
|---|---|---|
| 1 | Subject index | up to 16 characters |
| 2 | Disease status | 0=control, 1=case |
| 3 | Genotype code | -1, 0, 1, 2 |

A portion of a dataset for unrelated subjects is shown in Table 2.4. The top three rows in this set contain control data (second column value is zero) and the remaining two rows contain case data (second column value is one).

Table 2.4: Part of a sample data set showing three control subjects and two case subjects.

| Subject ID | Case-control indicator | Genotype |
|---|---|---|
| cbc1 | 0 | 0 |
| 1714 | 0 | 1 |
| fb90 | 0 | 0 |
| 231a | 1 | 2 |
| 5633 | 1 | 1 |

## 2.3 Data loading

Data are input into **SCOUT** using the **load** command as follows:

1. **load triads** *filename*
2. **load unrelateds** *filename*

These commands will read data from triads and unrelated subjects, respectively. If *filename* is not provided, a prompt requesting the name of the data file will be printed.

```
SCOUT > load triads ..\sample_data\triads.dat
  1000 subjects from ..\sample_data\triads.dat loaded
  triads:  964    dyads:  36    monads:  0
```

Figure 2.1: After loading triads data with the **load triads** command, a brief summary of the data is displayed.

Errors within the data, such as bad index or genotype values, will be reported on the console. If errors are detected, SCOUT will not load the data. These errors must be corrected before one can conduct analyses.

```
SCOUT > load triads ..\sample_data\triads-bad.dat
  2 errors found in data
  ERROR:  family ID 10 ---  G is not a valid member specifier
  ERROR:  family ID 56 ---  3 is not a valid genotype specifier
```

Figure 2.2: After the **load triads** command, any errors in the triad data will be summarized on the console.

## 2.4 Data summary

Once triad data are successful loaded, **SCOUT** can provide a data summary using the command **show triads**. An example summary is shown in Figure 2.3.

```
   Triad data summary

                    ---------- genotypes ----------
           family     father     mother   offspring
          --------  ---------  ---------  ---------
                980          1          1          1
                981          1          1          0
                982          0          0          0
                984          1          0          1
                985          1          0          0
                987          2          2          2
                988          2          1          1
                989          2          1          1
                990          1          1          1
                991          1          0          1
                992          1          1          1
                993          2          0          1
                994          2          1          2
                995          2          1          2
                996          1          1          1
                997          0          0          0
                998          1          0          1
                999          1          0          1
               1000          2          0          1
                 16          2         -1          2
                 32          1         -1          0
                 63          1         -1          0
                 64          1         -1          1
                144          2         -1          1
                166         -1          1          1
                173          1         -1          0
                199         -1          0          0
                219         -1          1          0
                295          2         -1          2
                312         -1          0          1
                315         -1          1          2
                325         -1          0          1
   ------------------------------------------------------
   Command summary:
   N - next triad   P - prev triad   T - top          B - bottom
   D - page down    U - page up      E - quit
```

Figure 2.3: Console display upon of using the **show triads** command.

In this summary, the first column displays the unique family index value followed by the Father, Mother, and offspring genotype values. Missing genotype data is indicated with the value -1. Within the console, complete triads will be shown in a green font, dyads (one parental genotype missing) are shown in yellow, and monads (both parental genotypes missing) are shown in red. (These colors do not appear in the console text shown in Figure 2.3.) Also, the

```
Case/control data summary

          id       genotype   disease
        ---------  ---------  ---------
              983      2          0
              984      0          0
              985      1          0
              986      2          0
              987      1          0
              988      1          0
              989      0          0
              990      1          0
              991      1          0
              992      0          0
              993      0          0
              994      1          0
              995      2          0
              996      1          0
              997      1          0
              998      0          0
              999      1          0
             1000      0          0
             1001      2          1
             1002      1          1
             1003      0          1
             1004      0          1
             1005      0          1
             1006      0          1
             1007      1          1
             1008      0          1
             1009      0          1
             1010      2          1
             1011      1          1
--------------------------------------------------
Command summary:
N - next         P - prev         T - top          B - bottom
D - page down    U - page up      E - quit
```

Figure 2.4: Console display upon of using the **show unrelateds** command.

displayed data are ordered with complete triads first followed by dyads, and then followed by monads.

Similarly, a summary of unrelated subject data can be printed to the console using the **show unrelateds** command. An example summary is shown in Figure 2.4.

In this summary, the first column displays the unique subject index value followed by the genotype index and the disease status. Controls are represented in green text whereas cases are represented in yellow text. (These colors do not appear in the console text shown in Figure 2.4.) Controls are displayed first, followed by cases.

Keys used to negotiate through either displayed triad and unrelated subjects lists are shown at the bottom of the console. These are summarized in the Table 2.5

Table 2.5: Displayed data navigation keys.

| Key | Action | Key | Action |
|---|---|---|---|
| N | scroll the list down by one subject | P | scroll the list up by one subject |
| space | scroll the list down by one page | R | scroll the list up by one page |
| B | scroll the list to the last subject | T | scroll the list to the first subject |
| E | exit display mode and return to **SCOUT** console | | |

# 3 SCOUT analysis

## 3.1 Association analysis

### 3.1.1 Association models

After data loading but prior to analysis, one must select a model that examines the effect of a reference SNP allele on disease. To model the SNP effect on disease, **SCOUT** requires specification of the relative risk (RR) parameters $\psi_1$ and $\psi_2$, where

$$\psi_g = \frac{P(\text{Disease}|\, g \text{ copies of reference SNP allele})}{P(\text{Disease}|\, 0 \text{ copies of reference SNP allele})}, \quad \text{where } g = 1,2.$$

A user specifies these RR parameters under one of four possible models. Here, $\psi$ ('Psi') denotes a scalar RR parameter which is estimated by **SCOUT** .

Table 3.1: Summary of models available in **SCOUT** .

| Model | RR parameter | user input |
|---|---|---|
| Multiplicative: | $\psi_1 = \psi, \quad \psi_2 = \psi^2$ | M, 1 |
| Additive: | $\psi_1 = \psi, \quad \psi_2 = 2\psi - 1, \quad (\psi > 0.50)$ | A, 2 |
| Dominant | $\psi_1 = \psi, \quad \psi_2 = \psi$ | D, 3 |
| Recessive: | $\psi_1 = 1, \quad \psi_2 = \psi$ | R, 4 |

### 3.1.2 Association analysis of triads

**SCOUT** can run an associaton analysis on triads only using the CPG approach of (Schaid and Sommer, 1993). To conduct such a CPG analysis, type the command **cpg** *MODEL*, where the value of *MODEL* is specified using one of the options shown in the third column of Table 3.1.

Once the command is executed, **SCOUT** will run two CPG-based analyses on the loaded triad dataset. The first CPG analysis allows the scalar RR parameter $\psi$ ('Psi') to be unconstrained and estimated from the data. The second CPG analysis constrains $\psi = 1$ (which is the null value corresponding to no association between SNP and disease). Using results from these two analyses, **SCOUT** then constructs and evaluates a likelihood-ratio (LR) statistic for testing the null hypothesis $H_0 : \psi = 1$, which is a test of association between the SNP and disease. A screen shot demonstrating both execution and output of the CPG analyses is shown in Figure 3.1.

Within the output, **SCOUT** first provides results for the CPG analysis that allows Psi to be unconstrained. Within these results, the program first states the chosen model and the AIC value for the analysis (which can be useful for model selection; models with smaller AIC values are preferable over models with larger values). Next, the program provides an estimate of the scalar RR parameter $\psi$. Finally, **SCOUT** provides estimates of parameters defined as $\mu_1$, $\mu_2$, $\mu_3$, $\mu_4$, $\mu_5$, $\mu_6$, which correspond to the population frequency of the different parental mating-types in the triads. The formal definition of these mating-type parameters are shown in Table 3.2.

After presenting results from the unconstrained analysis, **SCOUT** next provides results from the constrained CPG analysis where $\psi = 1$. These results consist only of the estimates $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6$. Finally, **SCOUT** provides a LR statistic for testing $H_0 : \psi = 1$ and the related $p$-value.

```
SCOUT > cpg 1

  ---------------------------------------------------------------
    CPG: Psi Unconstrained
  ---------------------------------------------------------------
  Model :  1 (Multiplicative)          AIC   :  4287.916022

  Relative risk parameters:

  Psi   :  0.734280

  Mating type parameters:

  mu1:  0.069152    mu2:  0.295492    mu3:  0.082517
  mu4:  0.154234    mu5:  0.291039    mu6:  0.107566


  ---------------------------------------------------------------
    CPG: Psi = 1
  ---------------------------------------------------------------
  Mating type parameters:

  mu1:  0.048767    mu2:  0.246424    mu3:  0.079472
  mu4:  0.152134    mu5:  0.331790    mu6:  0.141412


  ---------------------------------------------------------------
    Hypothesis Tests
  ---------------------------------------------------------------

    Null Hypothesis         LR Statistic             P-value
  ------------------     ------------------     ------------------
               Psi=1             20.510280              5.93E-06
```

Figure 3.1: Results displayed after running the **cpg** command.

### 3.1.3 Association analysis of triads and unrelated controls

**SCOUT** can also conduct a joint association analysis of triads and unrelated controls using a version of the CPG approach that is augmented to allow for unrelated control information. To implement this analysis, type the command **cpgc** *MODEL* at the prompt, where the value of *MODEL* is selected from one of the options shown in Table 3.1. Once the command is executed, **SCOUT** will run three analyses on the loaded data from triads and controls. First, the program will run an unconstrained analysis that estimates the scalar RR parameter from the triad information $\Psi$ ('Psi') and the control information $\Psi_c$ ('Psi_c') as separate parameters. Second, the program will run an analysis that constrains $\Psi = \Psi_c$ but estimates the RR parameter from the data. Finally, the program will run a null association analysis that constrains $\Psi = \Psi_c = 1$.

Using results from these three analyses, **SCOUT** then constructs and evaluates likelihood-ratio statistics for testing two hypotheses of interest. The first hypothesis tested is $H_0^{(c)} : \Psi = \Psi_c$, which is a test for assessing whether the controls can be safely combined with the triads for association analysis. Rejection of $H_0^{(c)}$ suggests the two samples cannot be combined, so inference should be based on the CPG analysis of triads only described in Section 3.1.2.

Table 3.2: Summary of mating type parameters used in **SCOUT** .

| |
|---|
| $\mu_1 = P(\text{Each parent in triad has 2 copies of reference allele})$ |
| $\mu_2 = P(\text{One parent has 2 copies of reference allele, other parent has 1 copy of reference allele})$ |
| $\mu_3 = P(\text{One parent has 2 copies of reference allele, other parent has 0 copies of reference allele})$ |
| $\mu_4 = P(\text{Each parent has 1 copy of reference allele})$ |
| $\mu_5 = P(\text{One parent has 1 copy of reference allele, other parent has 0 copies of reference allele})$ |
| $\mu_6 = P(\text{Each parent has 0 copies of reference allele})$ |

The second hypothesis tested is $H_0 : \Psi = \Psi_c = 1$, which is a test of association between SNP and disease. Note that the testing of $H_0$ is only valid if one fails to reject $H_0^{(c)}$.

A screen shot demonstrating both execution and output of the analyses is shown in Figure 3.2.

```
SCOUT > cpgc 1

  ---------------------------------------------------------------
    CPG/C: Psi, Psi_c Unconstrained
  ---------------------------------------------------------------
  Model:  1 (Multiplicative)          AIC   :  6394.397981

  Relative risk parameters:

  Psi   :  0.734354
  Psi_c :  0.726082

  Mating type parameters:

  mu1:  0.065686    mu2:  0.298702    mu3:  0.078875
  mu4:  0.163121    mu5:  0.290604    mu6:  0.103011


  ---------------------------------------------------------------
    CPG/C: Psi = Psi_c
  ---------------------------------------------------------------
  Model:  1 (Multiplicative)          AIC   :  6392.407523

  Relative risk parameters:

  Psi   :  0.731421

  Mating type parameters:

  mu1:  0.065404    mu2:  0.297955    mu3:  0.078806
  mu4:  0.163152    mu5:  0.291229    mu6:  0.103454


  ---------------------------------------------------------------
    CPG/C: Psi = Psi_c = 1
  ---------------------------------------------------------------
  Mating type parameters:

  mu1:  0.053157    mu2:  0.265658    mu3:  0.077038
  mu4:  0.162275    mu5:  0.318378    mu6:  0.123493


  ---------------------------------------------------------------
    Hypothesis Tests
  ---------------------------------------------------------------

    Null Hypothesis         LR Statistic              P-value
  ------------------     ------------------     ------------------
           Psi=Psi_c              0.009542              9.22E-01
               Psi=1             32.768735              1.04E-08
```

Figure 3.2: Results displayed after running the **cpgc** command.

The results of these joint analyses of triads and unrelated controls are presented in a similar format to the results described for the CPG analysis of triads described in section 3.2. At the bottom of the results, **SCOUT** provides two LR statistics; one for testing $H_0^{(c)} : \Psi = \Psi_c$ and one for testing $H_0 : \Psi = \Psi_c = 1$. The software also provide the related $p$-values for these LR statistics.

### 3.1.4 Association analysis of triads, unrelated controls, and unrelated cases

**SCOUT** can also conduct a joint association analysis of triads, unrelated controls, and unrelated cases. In this case, a CPG approach is augmented to allow for both unrelated case and control information. To implement this analysis,

type the command **cpgcd** *MODEL* at the prompt, where the value of *MODEL* is selected from one of the options shown in Table 3.1. Once the command is executed, **SCOUT** will run four analyses on the loaded data from triads, controls, and cases. First, the program will run an unconstrained analysis that estimates the scalar RR parameter from the triad information $\Psi$ ('Psi'), the control information $\Psi_c$ ('Psi_c'), and the case information $\Psi_d$ ('Psi_d') as separate parameters. Second, the program will run an analysis that constrains $\Psi = \Psi_c$ but estimates the RR parameter from the data. Third, the an analysis will be run constraining $\Psi = \Psi_c = \Psi_d$, again while estimating the RR parameter from the data. Finally, the program will run a null association analysis that constrains $\Psi = \Psi_c = \Psi_d = 1$.

Using results from these four analyses, **SCOUT** constructs and evaluates likelihood-ratio statistics for testing three hypotheses of interest. The first two hypotheses considered by **SCOUT** collectively assess whether one can combine controls and cases with triads for association analysis. **SCOUT** first considers whether controls can be combined with triads by testing $H_0^{(c)} : \Psi = \Psi_c$. **SCOUT** tests $H_0^{(c)}$ first because, if this null hypothesis is rejected, then it is difficult to justify combining the data of triads with either controls or cases (see Epstein et al., 2005, page 596). Therefore, if **SCOUT** rejects $H_0^{(c)}$, one should base inference on the CPG analysis of triads only described in Section 3.1.2.

If **SCOUT** fails to reject $H_0^{(c)}$, one next considers the second hypothesis $H_0^{(d)} : \Psi = \Psi_c = \Psi_d$, which is a test for whether the cases can be safely combined with the triads and controls. Rejection of $H_0^{(d)}$ suggests that cases cannot be combined with the triads and controls. Therefore, in this situation, one should base inference on the association analysis of triads and controls considered in Section 3.1.3.

The third hypothesis tested is $H_0 : \Psi = \Psi_c = \Psi_d = 1$, which is test of association between SNP and disease. Note that the testing of $H_0$ is only valid if one fails to reject both $H_0^{(c)}$ and $H_0^{(d)}$.

A screen shot demonstrating both execution and output of the analyses is shown in Figure 3.3.

```
SCOUT > cpgcd 1

  ---------------------------------------------------------------
    CPG/CD: Psi, Psi_c, Psi_d Unconstrained
  ---------------------------------------------------------------
  Model :  1 (Multiplicative)          AIC   :  8445.180057

  Relative risk parameters:

  Psi   :  0.734402
  Psi_c :  0.726868
  Psi_d :  0.740938

  Mating type parameters:

  mu1:  0.063094    mu2:  0.299014    mu3:  0.076646
  mu4:  0.169022    mu5:  0.291212    mu6:  0.101011


  ---------------------------------------------------------------
    CPG/CD: Psi = Psi_c
  ---------------------------------------------------------------
  Model:  1 (Multiplicative)          AIC   :  8443.187998

  Relative risk parameters:

  Psi   :  0.731720
  Psi_d :  0.742694

  Mating type parameters:

  mu1:  0.062853    mu2:  0.298320    mu3:  0.076587
  mu4:  0.169050    mu5:  0.291789    mu6:  0.101402


  ---------------------------------------------------------------
    CPG/CD: Psi = Psi_c = Psi_d
  ---------------------------------------------------------------
  Model :  1 (Multiplicative)          AIC   :  8441.244875

  Relative risk parameters:

  Psi   :  0.737121

  Mating type parameters:

  mu1:  0.062887    mu2:  0.298311    mu3:  0.076450
  mu4:  0.169168    mu5:  0.291762    mu6:  0.101422


  ---------------------------------------------------------------
    CPG/CD: Psi = Psi_c = Psi_d = 1
  ---------------------------------------------------------------
  Mating type parameters:

  mu1:  0.046875    mu2:  0.251830    mu3:  0.074631
  mu4:  0.166031    mu5:  0.330743    mu6:  0.129891


  ---------------------------------------------------------------
    Hypothesis Tests
  ---------------------------------------------------------------

    Null Hypothesis         LR Statistic              P-value
  ------------------     ------------------     ------------------
           Psi=Psi_c              0.007942              9.29E-01

     Psi=Psi_c=Psi_d              0.056877              8.12E-01

                Psi=1             45.354595              1.64E-11
```

Figure 3.3: Results displayed after running the **cpgcd** command.

# 4 Program options and output

## 4.1 Program options

EM Algorithm parameters:

- Convergence parameter
- Number of iterations
- Number of function calls per iteration

**EM algorithm parameters** This group contains numerical parameters specific to the EM algorithm routines.

- **Number of restarts,** $N_{\text{restart}}$: Parameter estimates for the EM algorithm may converge to satellite points rather than true maxima, which can yield unreliable results. To protect against convergence to satellite points, one should restart the EM algorithm multiple times at random initial parameter values. **SCOUT** performs this procedure assuming 10 restarts by default. Decreasing the number of restarts decreases the computation time but may increase the possibility of convergence to satellite points.
- **Convergence parameter,** $\varepsilon$: Used to help declare convergence of the EM algorithm. If the square root of the sum of squares of the parameter values at successive iterations is less than $\varepsilon$, **SCOUT** assumes the EM has converged. The default value of the parameter is $\varepsilon = 10^{-10}$. Increasing this value will speed up computation time but may lead to convergence of parameter values to satellite points and not the true maxima.
- **Max number of iterations,** $iter_{\max}$: This parameter sets an upper bound on the number of iterations used in the successive-approximations scheme. In particular, the subroutine will, regardless of convergence considerations, cease to iterate after this number has been exceeded. This is fundamentally different from the $N_{\text{restart}}$ parameter discussed above.

## 4.2 Saving output

Two commands are necessary to save results to a file: **set output** and **save**. A file in which to print output is chosen by using the **set output** *FILE* command. If the *FILE* argument is missing, a separate prompt will ask for the output file name. Alternatively, invoking **SCOUT** from a command line using the option **–output** *FILE* is equivalent to specifying an output filename with the **set output** command.

```
SCOUT > set output sample.out
  Output file set: sample.out
```

Figure 4.1: Use the **set output** *FILE* command to specify an output filename.

Once analyses have been performed, a **save** command will save the results into the output file.

# Glossary

This is a summary list of abbreviations and acronyms used in this document.

| | |
|---|---|
| AIC | Akaike information criterion |
| EM | expectation-maximization algorithm |
| GUI | graphic user interface |
| IMSL | International Mathematical and Statistical Libraries |
| LR | likelihood ratio |
| SNP | single nucleotide polymorphism |
| RR | relative risk |

# Bibliography

Epstein, M., C. Veal, R. Trembath, J. Barker, C. Li, and G. Satten (2005). Genetic association analysis using data from triads and unrelated subjects. Am. J. Hum. Genet. 76, 592–608.

Schaid, D. and S. Sommer (1993). Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am. J. Hum. Genet. 53, 1114–1126.

Spielman, R., R. McGinnis, and W. Ewens (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). Am. J. Hum. Genet. 52, 506–516.