FORMAT_PED_PHENO Software Documentation

Version 1.0

Timothy Thornton[1] and Mary Sara McPeek[2,3]

Department of Biostatistics[1]
University of Washington

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

FORMAT_PED_PHENO

A C program that takes as input a pedigree/phenotype information file and outputs files that are in the required format for the MQLS-XM software, the KinInbcoef software, and the KinInbcoefX software.

Copyright(C) 2012 Timothy Thornton and Mary Sara McPeek

Homepage: http://galton.uchicago.edu/∼mcpeek/software/index.html

Release 1.0   July 1, 2012

===

License

This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY of FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program (see file gpl.txt); if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA.

To contact the first author:

Timothy A. Thornton
Department of Biostatistics
University of Washington
Health Sciences Building F-600
Box 357232
Seattle, WA 98195-7232

email: tathornt@u.washington.edu

Contents

1 Overview of FORMAT_PED_PHENO

FORMAT_PED_PHENO is a program, written in C, that takes as input a pedigree/phenotype information file and outputs formatted files that are compatible with (i.e. can be used as input to) the MQLS-XM, KinInbcoef, and KinInbcoefX software packages. These three software packages can be found at **http://galton.uchicago.edu/∼mcpeek/software/index.html**. The formatted output files for the FORMAT_PED_PHENO software are as follows:

(1) A formatted pedigree/phenotype information file that can be used as input to the MQLS-XM software

(2) A formatted pedigree information file and a formatted list file that can be used as input to the KinInbcoef software

(3) A formatted pedigree information file and a formatted list file that can be used as input to the KinInbcoefX software

Essentially, the function of FORMAT_PED_PHENO is to relabel all the families and the individuals in the dataset so that the family IDs and individual IDs conform to the requirements of the MQLS-XM, KinInbcoef and KinInbcoefX software.

2 Running FORMAT_PED_PHENO

Installation instructions:

1. Download the FORMAT_PED_PHENO package. This package contains documentation, source code, example input and output files, and a precompiled executable for Linux platforms.

2. Read the file FORMAT_PED_PHENO_Documentation.pdf carefully to understand the purpose of this program and how it works.

3. Edit the Makefile as necessary according to the instructions in the Makefile. You should only need to make sure that the correct compiler and compiler options for your machine are chosen.

4. Type "make". This will build an executable program called "FORMAT". If the message "make: 'FORMAT' is up to date" appears after typing "make", then to build the executable program you must first delete the precompiled binary FORMAT program that comes with the software by typing "rm FORMAT", and then type "make" to build the executable program FORMAT.

5. FORMAT is run from the command line via the command 'FORMAT' with all information specified by command line options. To run the executable program FORMAT: First, prepare the input pedigree/phenotype file (see Section 3 for more details).

Then, to run FORMAT_PED_PHENO with the default input and output filenames, one need only type

./FORMAT

Alternatively, to change the input or output filenames, use flags in the command line. The following flags are available:

./FORMAT   -f pedphenofile   -o outfile

-f pedphenofile   Allows the user to specify the name of the pedigree/phenotype information input file. (This file includes family ID numbers, individual ID numbers, parents' ID numbers, sex, and phenotype values.) The filename defaults to "pedphenofile" if this flag is not used. To specify a different filename, replace "pedphenofile" with the appropriate filename.

-o outfile   Allows the user to specify the name of the prefix for all of the output files. The prefix for the output files defaults to "outfile" if this flag is not used.

6. You can test the executable program FORMAT by running it with the sample input file "pedphenofile". You can then compare the resulting output, which will be printed to the files outfile.pedpheno, outfile.kinpedigree, outfile.kinlist, outfile.kinpedigreeX, outfile.kinlistX, and outfile.pedmatch with the correct output provided in the sample output files outfile.pedpheno.ex, outfile.kinpedigree.ex, outfile.kinlist.ex, outfile.kinpedigreeX.ex, and outfile.pedmatch.ex, respectively.

3 Input

Required Input File:

1. **pedigree/phenotype file**

   This file contains family IDs, individual IDs, parents' IDs, sex, and phenotype values. To illustrate the required format of the input pedigree/phenotype file, consider a study sample with 8 individuals from 3 families. The lines in the pedigree/phenotype file could be as follows:

F192	ID101	0	0	1	1
F192	ID102	0	0	2	2
F192	ID201	ID101	ID102	2	2
M012	205HG	0	0	1	2
M012	ID309	0	0	2	1
M012	ID2999	205HG	ID309	1	0
M012	ID9203	205HG	ID309	2	1
S230	LF6950	0	0	2	0
(1)	(2)	(3)	(4)	(5)	(6)

   Column (1) family ID

   Column (2) individual ID

   Column (3) father's ID (0 if the individual is a founder)

   Column (4) mother's ID (0 if the individual is a founder)

   Column (5) sex (1=male, 2=female)

   Column (6) phenotype value (should usually be 0, 1 or 2 — see below)

   Each family ID is assumed to be a character string that does not contain a space, newline or tab. The same assumption holds for each individual ID. Individuals who have the same Family ID should appear together consecutively in the file, i.e, individuals from the same family must appear in a single cluster, though there is no requirement on the order of the individuals within a family. There is no limit on the number of individuals or the number of families, but the number of individuals in a single family is set to be smaller than 500. To increase this limit, just change the value of MAXFAMILYSIZE in the FORMAT_PED_PHENO_SOURCE.c source file and recompile the program. The phenotype value can be any character string that does not contain a space, newline or tab, and this same character string will appear in the output file. **However, note that in order for the output to be compatible with the MQLS-XM software, the input phenotype values should be coded as 2=affected, 1=unaffected, 0=unknown.**

4 Output

1. **outfile.pedpheno** is the default name of the formatted pedigree/phenotype information file, which contains the same six columns as the original pedigree/phenotype file, except that the families and the individuals have been relabeled with new ID numbers to conform to the format requirements of the MQLS-XM software. Specifically, the families are now numbered from 1 to F, without gaps, where F is the number of unique families in the input pedigree/phenotype file, and the new individual IDs for the sample individuals are positive integers. This file can be used as input to the MQLS-XM software. If the -o flag is used, the name of this output file changes according to the prefix specified by the user. For example, the command

   ./FORMAT -o newfile

   specifies that the name of the formatted pedigree/phenotype output file should be "newfile.pedpheno"

2. **outfile.kinpedigree** is the default name of the output file containing the pedigree information (first four columns of outfile.pedpheno). It is in the format required by, and can be used as input to, the KinInbcoef software. If the -o flag is used, the name of this output file changes according to the prefix specified by the user. For example, the command

   ./FORMAT -o newfile

   specifies that the name of the output file containing the pedigree information will be "newfile.kinpedigree"

3. **outfile.kinlist** is the default name of the output file containing a list of the new family and individual IDs (first two columns of outfile.pedpheno). This file is in the format required by, and can be used as input to, both the KinInbcoef and KinInbcoefX programs. If the -o flag is used, the name of this output file changes according to the prefix specified by the user, in a similar way to that described above.

4. **outfile.kinpedigreeX** is the default name of the output file containing the pedigree and sex information (first five columns of outfile.pedpheno). It is in the format required by, and can be used as input to, the KinInbcoefX software. If the -o flag is used, the name of this output file changes according to the prefix specified by the user, in a similar way to that described above.

5. **outfile.pedmatch** is the default name of the output file that lists both the new and old family and individual IDs for each individual. This file is purely for the convenience of the user who might want to be able to match up the original IDs with the software-generated ones. The file consists of four columns: original family ID, original individual ID, software-generated family ID, software-generated individual ID. If the -o flag is used, the name of this output file changes according to the prefix specified by the user, in a similar way to that described above.

6. **outfile.err** is an error file that may contain warnings

- when the ID for the parent of an individual in the input pedigree/phenotype file does not match any individual IDs in the same family.

- when individuals in the same family do not appear consecutively in the file.

If the -o flag is used, the name of this output file changes according to the prefix specified by the user, in a similar way to that described above.

5 Tips

1. Input

   The program will stop if an error is detected in the format of the input pedigree/phenotype file. Please read Section 3 carefully and make sure the input file is in the correct format.

2. Phenotype coding

   The program does not make any changes in the phenotype coding. If the resulting pedigree/phenotype file is to be used as input to MQLS-XM, then the phenotype should be coded as 2=affected, 1=unaffected, 0=unknown.

6 EXAMPLE

1. Consider a pedigree/phenotype file named "EXAMPLE.file" that contains the following text:

FAM1	ID101	0	0	1	1
FAM1	ID102	0	0	2	2
FAM1	ID201	ID101	ID102	2	2
M012	205HG	0	0	1	2
M012	ID309	0	0	2	1
M012	ID2999	205HG	ID309	1	0
M012	ID9203	205HG	ID309	2	1
S230	LF6950	0	0	2	0
(1)	(2)	(3)	(4)	(5)	(6)

2. To obtain a formatted pedigree/phenotype file, the following command can be used:

   ./FORMAT   -f EXAMPLE.file   -o newfile

3. The above command will create the output file "newfile.pedpheno" which contains the following text:

1	1	0	0	1	1
1	2	0	0	2	2
1	3	1	2	2	2
2	1	0	0	1	2
2	2	0	0	2	1
2	3	1	2	1	0
2	4	1	2	2	1
3	1	0	0	2	0

.

4. The output files "newfile.kinpedigree", "newfile.kinlist", and "newfile.kinpedigreeX" will also be created using the above command and these files are in the format required for the KinInbcoef and KinInbcoefX software packages, which can be found at **http://galton.uchicago.edu/~mcpeek/software/index.html**

   - To obtain autosomal kinship coefficients using the KinInbcoef software, download the KinInbcoef software, and type the following command:

     ./KinInbcoef newfile.kinpedigree newfile.kinlist final.kinship

     This command creates the output file "final.kinship" containing autosomal kinship and inbreeding coefficients.

- Similarly, to obtain X-chromosome kinship coefficients, download the KinInbcoefX software, and then type the following command:

  ./KinInbcoefX newfile.kinpedigreeX newfilefile.kinlist final.kinshipX

  This command creates the output file "final.kinshipX" containing X-chromosome kinship and inbreeding coefficients.