MANTRA Documentation
Trans-ethnic meta-analysis of genome-wide association studies

Version 1: 2 February 2012

Introduction

The detection of loci contributing effects to complex human traits, and their subsequent
fine-mapping for the location of causal variants, remains a considerable challenge for the
genetics research community. Meta-analyses of genome-wide association studies, primarily
ascertained from European descent populations, have made considerable advances in our
understanding of complex trait genetics, although much of their heritability is still
unexplained. With the increasing availability of genome-wide association data from diverse
populations, trans-ethnic meta-analysis may offer an exciting opportunity to increase the
power to detect novel complex trait loci and to improve the resolution of fine-mapping of
causal variants by leveraging differences in local linkage disequilibrium structure between
ethnic groups. However, we might also expect there to be substantial genetic heterogeneity
between diverse populations, both in terms of the spectrum of causal variants and their
allelic effects, which cannot easily be accommodated through traditional approaches to
meta-analysis. In order to address this challenge, MANTRA takes account of the expected
similarity in allelic effects between the most closely related populations, whilst allowing for
heterogeneity between more diverse ethnic groups. This approach yields substantial
improvements in performance, compared to fixed-effects meta-analysis, both in terms of
power to detect association, and localisation of the causal variant, over a range of models of
heterogeneity between ethnic groups. Furthermore, when the similarity in allelic effects
between populations is well captured by their relatedness, this approach has increased
power and mapping resolution over random-effects meta-analysis.

Reference. Morris AP (2011). Transethnic meta-analysis of genomewide association studies.
Genet Epidemiol 35: 809-22.

Questions or problems?

If you have any problems with compiling or running the MANTRA software, please contact
Andrew Morris via e-mail: amorris@well.ox.ac.uk
1. Installation

Begin by unpacking the software using the command:

```
tar -xvzf mantra.tar.gz
```

This package includes the MANTRA software, additional data formatting scripts and example data files. You should compile the software and scripts using the command:

```
gfortran dmatcal.f95 -o dmatcal
gfortran mantra.v1.f95 -o mantra.v1
```

2. Data formats

The MANTRA software requires two input files. The first, `mantra.in`, contains a label for each study contributing to the meta-analysis, one label per row. Labels are limited to 20 characters and must form a single word. For example:

Finland
Germany
Italy
Japan
China

The second file, `mantra.dat`, contains association summary statistics for each study, for each SNP. The data for each SNP is arranged over a series of consecutive rows in the file. The first row gives details about the SNP: identifier, chromosome (23 for X), position (bp), effect allele and other allele. Each subsequent row then provides summary statistics for each study, in the order in which they appear in the `mantra.in` file. The statistics provided are: presence or absence indicator (1 if the SNP is reported for the study, 0 otherwise), the sample size for that study at the SNP, the effect allele frequency, the allelic effect (beta) aligned to the effect allele (i.e. log-odds ratio for a binary phenotype), and the standard error. Then repeat this process for all SNPs. For example, for the `mantra.in` file above:

```
rs1 1 100000 A C
1 2000 0.25 0.180 0.020
1 1500 0.27 0.160 0.030
1 1200 0.32 0.175 0.035
1 2000 0.48 0.010 0.020
1 2500 0.52 -0.020 0.015
rs2 1 100500 T G
0 0 0 0 0
1 1500 0.10 0.010 0.025
1 1200 0.18 -0.020 0.040
1 2000 0.03 0.010 0.065
1 2500 0.05 -0.010 0.050
```

In this example, SNP rs1 is reported for all studies, whilst SNP rs2 is not reported for the Finland study.
3. Assessing relatedness between studies

MANTRA makes use of a prior model of relatedness between studies in the clustering process. The “distance” between studies is given in the file `dmat.out`, and could correspond to FST or mean effect allele frequency differences between populations. As an example, for the `mantra.in` file above:

\[
\begin{align*}
0.000 & \text{ <- distance between Finland and Finland (will be 0)} \\
0.017 & \text{ <- distance between Finland and Germany} \\
0.050 & \text{ <- distance between Finland and Italy} \\
0.123 & \text{ <- distance between Finland and Japan} \\
0.143 & \text{ <- distance between Finland and China} \\
0.017 & \text{ <- distance between Germany and Finland} \\
0.000 & \text{ <- distance between Germany and Germany (will be 0)} \\
0.060 & \text{ <- distance between Germany and Italy} \\
\ldots \\
0.000 & \text{ <- distance between China and China (will be 0)}
\end{align*}
\]

If you have your own distances between populations (based on external data), format the `data.out` file yourself. Please note that studies should be arranged in the same order as in the `mantra.in` file. However, if you would like to calculate mean allele frequency differences across SNPs in the `mantra.dat` file, you can make use of the `dmatcal` script to generate the `dmat.out` file automatically. Simply type the command:

```
./dmatcal
```

4. Running MANTRA

To run the MANTRA analysis, simply type the command:

```
./mantra.v1
```

You will be prompted to enter the name of an output file (for example, `mantra.out`) and an output file in which parameter estimates will be stored (for example, `mantra.bet.out`), and a seed for the random number generator (just enter a random large integer). MANTRA is computationally intensive: if you wish to run the software in the background, create a file called `fname.in`, containing the output filenames and seed, and then use the command:

```
./mantra.v1 < fname.in &
```

5. Interpreting MANTRA output

The output file (for example, `mantra.out`) contains the following information for each SNP, one per row: SNP ID, chromosome, position, effect allele, other allele, number of studies, log10 Bayes’ factor in favour of association, posterior probability of association, total sample size and effect directions (coded as ? for missing, + for positive allelic effect for effect allele, and - for negative allelic effect for effect allele). The parameter estimate output file contains the following information for each SNP, again one per row: SNP ID, chromosome, position,
number of studies, and then for each study reporting data for that SNP, the posterior mean allelic effect (aligned to the effect allele) and posterior standard deviation.

6. Example analysis

When you unpack the MANTRA software, you will find an example data set for analysis, incorporating data for four SNPs from five studies. Begin by generating the dmat.out file on the basis of mean allele frequency differences using the command:

./dmatcal

Then run the MANTRA analysis using the command:

./mantra.v1

The program will take a minute or so to run. Looking in the output file, you should see strong evidence of association for SNPs rs1, rs3 and rs4, each having large log10 Bayes' factors. You should also see evidence of heterogeneity in allelic effects (posterior probability greater than 0.5) for SNPs rs1 and rs4. Looking at the parameter estimate output file, you should see that for SNP rs1, the European studies have large posterior mean allelic effects, whilst the East Asian studies have posterior mean allelic effects close to zero. This would suggest that this effect is specific to Europe, or that rs1 is a poor tag for the effect in East Asian populations because of differences in LD structure between ethnic groups. For rs4, however, the allelic effects are in opposite directions in European and East Asian studies, which could reflect multiple causal variants, or differing LD structure between ethnic groups.

7. Using MANTRA with large data sets

MANTRA is computationally demanding, and will take many hours to run for genome-wide data sets with hundreds of thousands of SNPs, particularly if the number of studies is large. If multiple computing processors are available, it would be sensible to break the genome into chunks, and to analyse each chunk independently, in parallel. To do this:

a) Construct the dmat.out file on the basis of genome-wide data (in a single mantra.dat file).

b) Break the genome into chunks of 50,000 SNPs, and create a directory for each chunk. This directory should contain the genome-wide dmat.out file, mantra.in file, and a new mantra.dat file consisting only of the 50,000 SNPs.

c) Run the MANTRA analysis for each chunk in each directory in parallel.

d) On completion, combine output files from all chunks for genome-wide output file.