Gene Level Meta-analysis of Dichotomous Traits by Generalized Functional Linear Models

Ruzong Fan, NICHD/NIH, October 2015

1 Overview

This document describes a R package to implement the generalized functional linear models for gene level meta-analysis of complex diseases (Fan et al. 2015b). Section 2 briefly describes the installation of the program. Section 3 explains how to run the program using one example. Section 4 offers explanation of the results and warnings to use the programs. Section 5 provides some suggestions and parameter choices for real data analysis.

The theoretical basis for this program is given in our research papers in **References**. Please refer to the reference if you use the program in any published work. In case of suggestions and questions and/or problems, you can contact us via e-mail (fanr@mail.nih.gov).

2 Download and Installation

The package is written in statistical programming language R. Download "MetaGFLM_fixed_model.R", "MetaGFLM_beta_smooth_only.R", and an example file of "MetaSKAT_Example_for_MetaGFLM.R" from GFLM_meta.zip. In addition, we need "MetaFLM_additive_effect_model.R" from FLM_meta.zip, which is a R package to implement the models for functional linear models for meta-analysis of quantitative traits (Fan et al. 2015a). Put the files in a directory you may access.

3 How to Run the Program

The analysis needs libraries fda, MASS, Matrix, and MetaSKAT in R package. The datasets are from the package MetaSKAT. Make sure to install them before running our codes. Open the "MetaSKAT_Example_for_MetaGFLM.R" file on an R Console in a PC window. Please change the

paths leading to the directories of "MetaGFLM_fixed_model.R", "MetaGFLM_beta_smooth_only.R", and "MetaFLM_additive_effect_model.R" on your computer. Then you may run the codes. Please note that the following codes

```
y1 = rbinom( length(y.list[[1]]), 1, 0.5)
y2 = rbinom( length(y.list[[2]]), 1, 0.5)
y3 = rbinom( length(y.list[[3]]), 1, 0.5)
y = list(length = 3)
y[[1]] = y1
y[[2]] = y2
y[[3]] = y3
pheno = y
```

will generate 3 random samples of dichotomous trait values. Therefore, the results will be different from time to time. On October 27, 2015, I got the following results

```
> MetaGFLM_beta_smooth_only(L, is.homo = TRUE, y, mode = "Additive", geno, pos,
               order, bbasis, covariate, base = "bspline", interaction = FALSE)
$LRT
[1] 0.3173015
$Chisq
[1] 0.3173015
$Rao
[1] 0.3498529
......
> MetaFlm_add_effect(L, is.homo = FALSE, y, mode = "Additive", geno, covariate,
                     family = "binomial")
$LRT
```

```
[1] 0.08570803
$Chisq
[1] 0.08570803
$Rao
[1] 0.5783272
>
> ### The following function is not working since geno[[k]] have different
      number of columns ###
> MetaFlm_add_effect(L, is.homo = TRUE, y, mode = "Additive", geno, covariate,
                     family = "binomial")
Error in rbind(U, geno[[k]]) :
  number of columns of matrices must match (see arg 2)
```

To make "MetaFlm_add_effect(L, is.homo = **TRUE**, ...)" to run, one needs that each individual of the L studies is sequenced at the same variants. However, the function "MetaFlm_add_effect(L, is.homo = **FALSE**, ...)" can analyze different genotype data among multiple studies, i.e., individuals of different studies may be genotyped at different genetic markers. The details are provided in Fan et al. (2015a).

4 Explanation of the Results and Warnings

As shown in the Section 3, our program can output 3 p-values based on likelihood ratio test (LRT), χ^2, and Rao's efficient score test. The LRT is the same as χ^2, which have inflated type I error rates when sample size is smaller than or equal to 2,000 for single study (Fan et al. 2014). For large sample of multiple meta-analysis studies, LRT statistics (Hom-LRT) of the homogeneous genetic effect models have correct type I error rates but the LRT statistics (Het-LRT) of the heterogeneous genetic effect models inflate the type I error rates (Fan et al. 2015b).

The Rao's efficient score test has conservative and accurate type I error rates (Fan et al. 2014; 2015b).

If you use the R codes to analyze your data, we recommend to report the p-values of Rao's efficient score test.

5 Suggestions and Parameters for Real Data Analysis

We present two R functions "MetaGFLM_fixed_model.R" and "MetaGFLM_beta_smooth_only.R" to perform gene-based association analysis of dichotomous traits in this documentation. In practice, one may use one of them and either B-spline or Fourier spline basis functions for data analysis . We also suggest the following parameters for a data analysis:

```
order  =  4
bbasis = 10
gbasis = 10
fbasis = 11
```

6 References

Fan RZ, Wang YF, Mills JL, Wilson AF, Bailey-Wilson JE, and Xiong MM (2013) Functional linear models for association analysis of quantitative traits. *Genetic Epidemiology*, 37:726-742.

Fan RZ, Wang YF, Mills JL, Carter TC, Lobach I, Wilson AF, Bailey-Wilson JE, Weeks DE, and Xiong MM (2014) Generalized functional linear models for case-control association studies. *Genetic Epidemiology* **38**:622-637.

Fan RZ, Wang YF, Boehnke M, Chen W, Li Y, Ren HB, Lobach I, and Xiong MM (2015a) Gene level meta-analysis of quantitative traits by functional linear models. *Genetics* **200**:1089-1104.

Fan RZ, Wang YF, Chiu CY, Chen W, Ren HB, Li Y, Boehnke M, Amos CI, Moore J, and Xiong MM (2015b) Meta-analysis of complex diseases at gene level by generalized functional linear models. *Genetics*, in revision.