

# Functional Linear Models for Gene-based Association Analysis of Quantitative Traits

Ruzong Fan and Yifan Wang, NICHD/NIH, December 2014

## 1 Overview

This document describes a R package to implement the functional linear models for association analysis of quantitative traits. 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 paper 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 R. Download R codes “FLM\_fixed\_model.R”, “FLM\_beta\_smooth\_only.R”, “FLM\_FPCA.R”, and “FLM\_FPCA\_no\_position.R”, and example files of “Example\_FLM.R” and “Example\_FLM\_multiple\_gene\_analysis.R” from FLM\_web.zip. Plus, you will need datasets from data.zip to run the examples. Put the files in a directory you may access.

## 3 How to Run the Program

### 3.1 One Gene Analysis

The analysis needs libraries fda, MASS, and Matrix in R package. Make sure to install them before running our codes. Open the “Example\_FLM.R” file on an R Console in a PC window. Change the paths leading to the directories of the package “FLM\_fixed\_model.R”, “FLM\_beta\_smooth\_only.R”,

“FLM\_FPCA.R”, “FLM\_FPCA\_no\_position.R”, and the datasets on your computer. Then, you may run the program. The following results are based on the datasets in data.zip by “R i386 3.1.2”.

```
> flm_fixed_model(pheno, mode = "Additive", geno, pos, order, bbasis,  
                  fbasis, gbasis, covariate, base = "bspline", interaction = FALSE)
```

```
$LRT
```

```
[1] 1
```

```
$Chisq
```

```
[1] 1
```

```
$F
```

```
[1] 1
```

```
> flm_fixed_model(pheno, mode = "Additive", geno, pos, order, bbasis,  
                  fbasis, gfbasis, covariate, base = "fspline", interaction = FALSE)
```

```
$LRT
```

```
[1] 0.6598117
```

```
$Chisq
```

```
[1] 0.6598117
```

```
$F
```

```
[1] 0.660154
```

```
> flm_beta_smooth_only(pheno, mode = "Additive", geno, pos, order, bbasis,  
                       covariate, base = "bspline", interaction = FALSE)
```

```
$LRT
```

```
[1] 0.774038
```

```
$Chisq
```

```
[1] 0.774038
```

```

$F
[1] 0.7732432

> flm_beta_smooth_only(pheno, mode = "Additive", geno, pos, order, fbasis,
                        covariate, base = "fspline", interaction = FALSE)

$LRT
[1] 0.6598117

$Chisq
[1] 0.6598117

$F
[1] 0.6601542

flm_fpca_no_position(pheno, mode = "Additive", geno, covariates = covariate,
                      kz = 20, kb = 10, smooth.cov=FALSE, family = "gaussian")

$LRT
[1] 0.774038

$Chisq
[1] 0.774038

$F
[1] 0.7732432

> flm_fpca(pheno, mode = "Additive", geno, covariates = covariate, pos,
            kz = 20, kb = 10, smooth.cov=FALSE, family = "gaussian")

$LRT
[1] 0.6621036

$Chisq

```

```
[1] 0.6621036
```

```
$F
```

```
[1] 0.6624255
```

## 3.2 Multiple Gene Analysis

The analysis needs libraries `fda`, `MASS`, `SKAT`, and `Matrix` in R package. Make sure to install them before running our codes. Open the “`Example_FLM_multiple_gene_analysis.R`” file on an R Console in a PC window. Change the paths leading to the directories of the package “`FLM_fixed_model.R`”, “`FLM_beta_smooth_only.R`”, “`FLM_FPCA.R`”, “`FLM_FPCA_no_position.R`”, and the datasets on your computer.

Then, you may get one csv file named “`y_mode=Additive_order=4_bbasis=15_fbasis=25.csv`” after running “`Example_FLM_multiple_gene_analysis.R`” file. Note that only two genes are analyzed, but you may add more for multiple gene analysis.

## 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),  $\chi^2$ , and  $F$ -distributed test. The LRT is the same as  $\chi^2$ , which may inflate type I error rates when sample size is smaller than or equal to 1,000 (Fan et al. 2013, p733, top of the left column). The  $F$ -distributed test has conservative and accurate type I error rates (Fan et al. 2013). If you use the R codes to analyze your data, we recommend to report the  $p$ -values of  $F$ -distributed test. If you analyze large sample data, both LRT and  $F$ -distributed tests can be used.

## 5 Suggestions and Parameters for Real Data Analysis

In this documentation, we present four R functions to perform gene-based association analysis of quantitative traits. In practice, one may use one of them for data analysis. We suggest to use `flm_fixed_model` by either B-spline or Fourier spline basis functions. We also suggest the following

parameters for a data analysis:

`order = 4`

`bbasis = 15`

`gbasis = 15`

`fbasis = 25`

`gfbasis = 25`

## 6 References

1. 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.