# GATI-MFG User's Guide

Manyan Huang

2023-04-02

## 1. Introduction

The gene-based association test of interactions for maternal-fetal genotypes (GATI-MFG) is developed to test the joint effect of maternal and fetal genes on disease risk while allowing for maternal-fetal genotype interaction.

The proposed method has several advantages:

- It may account for the complex relationships between maternal and fetal genes during the embryogenesis.
- It is a gene-based test that can integrate the collective effect of multiple genetic variants, including both common and rare variants.
- It is especially suited for detecting the joint effects due to rare variants with a larger sample, which may be missed by the conventional single-variant-based association tests.

This document includes 1) an example pipeline for processing genetic data; 2) detailed description for GATI-MFG function, input and output data with an example; and 3) an example of simulating phenotype under various disease scenarios.

### 1.1 Dependencies

The GATI-MFG function requires the following R packages to be installed:

```
library(Matrix)
library(SPAtest)
library(BEDMatrix)
library(MGLM)
library(SKAT)
library(MASS)
library(survival)
library(compiler)
```

## 2. Data pre-processing

This document demonstrates the processing of genetic data for gene-based association tests. Depending on the specific requirements of their study, the user has the flexibility to select suitable data pre-processing strategies.

### 2.1 Genotype data processing

In our study, DNA samples of the infants and their mothers were collected by the National Birth Defects Prevention Study. Each sample was genotyped for approximately 5 million SNPs, including both common and rare variants, by using Illumina® Infinium HumanOmni5Exome BeadChip in the Hobbs Laboratory in Arkansas Children's Research Institute. Detailed information can be found elsewhere.[1]

For genetic data, we extracted maternal and fetal genotype in phase I and phase II from the genome data set. We removed the genetic variants with high missing rate. Software PLINK 1.9[2] was applied to process the genetic data in the instruction below. You can find comprehensive guidelines on how to use PLINK at https://www.cog-genomics.org/plink/.

```
s1=discv_mb
s2=rep_mb

d=aim1aim2.ACGT.final

out1=discv_mb_geno
out2=rep_mb_geno

plink \
      --bfile $d \
      --keep $s1 \
      --geno 0.05 \
      --keep-allele-order \
      --make-bed \
      --out $out1
plink
      --bfile $d \
      --keep $s2 \
      --geno 0.05 \
      --keep-allele-order \
      --make-bed \
      --out $out2
```

### 2.2 Extraction of genetic data based on gene regions

To conduct gene-based association tests, we used the UCSC Genome Browser (as-sembly GRCh37/hg19) to define gene regions as biologically meaningful units for association with CHD. An extended region from 7.5K upstream and downstream of the genomic region was defined as a testing unit.

```
discovery_sd=region_extracted_discovery
replication_sd=region_extracted_replication

p1=discv_mb_geno
p2=rep_mb_geno
```

[1] Rashkin SR, Cleves M, Shaw GM, Nembhard WN, Nestoridi E, Jenkins MM, Romitti PA, Lou XY, Browne ML, Mitchell LE, Olshan AF, Lomangino K, Bhattacharyya S, Witte JS, Hobbs CA; National Birth Defects Prevention Study. A genome-wide association study of obstructive heart defects among participants in the National Birth Defects Prevention Study. Am J Med Genet A. 2022 Aug;188(8):2303-2314. doi: 10.1002/ajmg.a.62759. Epub 2022 Apr 22. PMID: 35451555; PMCID: PMC9283270.

[2] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559-75. doi: 10.1086/519795. Epub 2007 Jul 25. PMID: 17701901; PMCID: PMC1950838.

```
while read chrstr begin_pos end_pos st gene
do

    chr="${chrstr:3}"

    # specify the output filename for each region
    out_discv=${discovery_sd}/${chr}_${begin_pos}_mb
    out_rep=${replication_sd}/${chr}_${begin_pos}_mb

    # extend the region 7.5kb upstream and downstream
    fr=$[begin_pos-7500]
    if [ $fr -lt 0 ]
    then
        fr=0
    fi
    to=$[end_pos+7500]

    # extract the genotype based on chromoson, start and end position
    plink --bfile $p1 --chr $chr --from-bp $fr --to-bp $to --make-bed --out $out_discv
    plink --bfile $p2 --chr $chr --from-bp $fr --to-bp $to --make-bed --out $out_rep

done < /geode2/home/u030/huanshan/Carbonate/manyan/GbyG/Rochy_application/refGene.merged.bed
```

The sample input of region information file (as refGene.merged.bed above):

```
##    Chr  Start    End Strand
## 1 chr1  11868  14409      +
## 2 chr1  14361  29370      -
## 3 chr1  30365  30503      +
## 4 chr1  34610  36081      -
## 5 chr1  69090  70008      +
## 6 chr1 134772 140566      -
##                                                                                                                                  Gene
## 1                                                                                                    NR_046018&DDX11L1,NR_148357&LOC102725121
## 2 NR_024540&WASH7P,NR_106918&MIR6859-1,NR_107062&MIR6859-2,NR_107063&MIR6859-3,NR_128720&MIR6859-4
## 3                NR_036051&MIR1302-2,NR_036266&MIR1302-9,NR_036267&MIR1302-10,NR_036268&MIR1302-11
## 4                                                                                  NR_026818&FAM138A,NR_026820&FAM138F
## 5                                                                                                       NM_001005484&OR4F5
## 6                                                                                                    NR_039983&LOC729737
```