SimPhe: Simulate Multiple Phenotypes with Epistatic Interactions

Beibei Jiang, Benno Pütz, and Bertram Müller-Myhsok.
Statistical Genetics Group
Max Planck Institute of Psychiatry, Munich, Germany
beibei_jiang@psych.mpg.de; puetz@psych.mpg.de; bmm@psych.mpg.de

SimPhe: https://cran.r-project.org/web/packages/SimPhe
Background

- **genotype:**
  the set of heritable genetic identity in our DNA

- **phenotype:**
  a trait like disorder, height, brain size

- **epistasis:**
  Interaction between two or more genetic loci
Background

- Simulation tools developed for evaluating type I error rates for new statistical association tests or power comparisons. 

- The detection of dominance or the interactions it is involved in have been reported recently.
  *Hemani et al. 2014; Sztepanacz et al. 2015.*

- There are different models for modeling epistasis.
  *Fisher 1919; Cockerham 1954; Hayman and Mather 1955; Kempthorne et al. 1957.*
Background

- Simulation tools developed for evaluating type I error rates for new statistical association tests or power comparisons.
  

- The detection of dominance or the interactions it is involved in have been reported recently.
  
  *Hemani et al. 2014; Sztepanacz et al. 2015.*

- There are different models for modeling epistasis.
  
  *Fisher 1919; Cockerham 1954; Hayman and Mather 1955; Kempthorne et al. 1957.*
Background

- Simulation tools developed for evaluating type I error rates for new statistical association tests or power comparisons.
  

- The detection of dominance or the interactions it is involved in have been reported recently.
  
  Hemani et al. 2014; Sztepanacz et al. 2015.

- There are different models for modeling epistasis.
  
  Fisher 1919; Cockerham 1954; Hayman and Mather 1955; Kempthorne et al. 1957.

_Cockerham model has been reported as more appropriate than the other models for the study of epistasis between genes._ Kao and Zeng 2002.
Main function

Flowchart of main function Sim. Phe:

- simulate phenotype
  - sim.phe

- read parameters
  - read.simu.pars

- is.data.frame(fgeno)
  - F
    - read genotype
      - read.geno
        - T
          - calculate allele frequency
            - get.freq
              - get regression coefficients
                - get.gene.coef

  - T
    - exists
      - (heritability)
        - calculate noise variance
          - get.noise.var
        - calculate heritability
          - calc.herit

- get genetic effects
  - gene.effect

- add mean and random effect

- Output
  - pars.writer, phe.writer
Main function

Flowchart of main function Sim.Phe:

1. simulate phenotype
   - sim.phe

2. read parameters
   - read.simu.pars

3. is.data.frame(fgeno)
   - F
     - read genotype
       - read.geno
     - calculate allele frequency
       - get.freq
     - get regression coefficients
       - get.gene.coef
   - T
     - exists (heritability)
       - T
         - calculate noise variance
           - get.noise.var
         - calculate heritability
           - calc.herit
       - F
         - get genetic effects
           - gene.effect
         - add mean and random effect
         - Output
           - pars.writer, phe.writer
Main function

Flowchart of main function Sim.Phe:

- **simulate phenotype** `sim.phe`
- **read parameters** `read.simu.pars`
- **is.data.frame(fgeno)**
  - **read genotype** `read.geno`
  - **calculate allele frequency** `get.freq`
  - **get regression coefficients** `get.gene.coef`
- **exists (heritability)**
  - **calculate noise variance** `get.noise.var`
  - **calculate heritability** `calc.herit`
  - **get genetic effects** `gene.effect`
  - **add mean and random effect**
  - **Output** `pars.writer, phe.writer`
Multiple phenotypes

Conduct Correlation:

• Share same variants

• Convert via covariance matrix

\[ C = LL^T \quad \longrightarrow \quad Y_{\text{new}} = LY' \]

L is a lower triangular matrix with real and positive diagonal entries in a Cholesky decomposition.
install.packages("SimPhe")
library("SimPhe")

# get file path of simulation parameters
# (two shared SNP pairs and one independent SNP pair for each phenotype)
fpar <- system.file("extdata", "simupars.txt", package="SimPhe")

# get file path of genotype file: rows are individuals and columns are SNPs
fgeno <- system.file("extdata", "10SNP.txt", package="SimPhe")

phe <- sim.phe(fgenetic.pars = fpar,
               fgeno = fgeno,
               ftype = "snp.head",
               seed = 123,
               fwrite = FALSE)
install.packages("SimPhe")
library("SimPhe")

# get file path of simulation parameters
# (two shared SNP pairs and one independent SNP pair for each phenotype)
fpar <- system.file("extdata", "simupars.txt", package="SimPhe")

# get file path of genotype file: rows are individuals and columns are SNPs
fgeno <- system.file("extdata", "10SNP.txt", package="SimPhe")

phe <- sim.phe(fgenetic.pars = fpar,
               fgeno = fgeno,
               ftype = "snp.head",
               seed = 123,
               fwrite = FALSE)
Features

- Consider epistasis with dominance related genetic effects
- Simulate phenotype with given heritability
- Support single and multiple phenotype(s)
- Flexible genotype input formats
Features

- Consider epistasis with dominance related genetic effects
- Simulate phenotype with given heritability
- Support single and multiple phenotype(s)
- Flexible genotype input formats
Shortcomings

- Simple genetic model
- Fixed input format for simulation parameters
- Others

"Don’t worry if it doesn’t work right. If everything did, you’d be out of a job."
— Mosher’s Law of Software Engineering
We are just on the way.

Thank you.