

# The **GLIDE** package: Global and Individual Tests for Direct Effects in Mendelian randomization studies

Xiaoyu Wang, James Y. Dai

October 12, 2021

## 1 Introduction

“Mendelian randomization”, an inferential method that uses germline genotypes as instrumental variables (IVs) to study the environmentally modifiable cause of human disease, has been increasingly popular in genetic epidemiology. Exploiting the random assortment of genes from parents to offsprings at the time of gamete formation, likened to “Nature’s randomized experiment.” Mendelian randomization holds promise for removing confounding and reverse causation that loom over observational epidemiology. While conceptually appealing, skepticism has persisted on feasibility of strong assumptions required by Mendelian randomization, stated informally, that genetic variants are independent of confounders of the exposure disease relation, that genetic variants are associated with the exposure, preferably strongly, and that there is no direct effect from genetic variants to the disease other than the pathway through the exposure, also known as no “pleiotropy”. Perhaps the Achilles’ heel for Mendelian randomization is potential pleiotropic effects among candidate genetic IVs, which can be sometimes plausible given complex biological pathways and networks. This threat to the validity of Mendelian randomization is heightened by the very practice of including a large number of risk variants to boost the strength of IVs. The **GLIDE** package performs the diagnostic tests for assessing global and individual variant pleiotropy in Mendelian randomization studies.

## 2 Main functions

We show an example analysis using simulation data that resemble the BMI analysis for GECCO presented in the paper [1]. Briefly the dataset contains genotype data for 75 SNPs and 20,000 subjects, as well as demographic variables including age, gender and 3 principal components for the parent genome-wide association study. First we load the example dataset:

```

> data(simdata)
> #The example dataset is a list composed of two dataframes.
> #simat stores 20,000 observations of 81 variables,
> #including outcome, 5 adjusting covariates, and 75 SNPs.
> simdat=simdata$simdat
> dim(simdat)

[1] 20000      81

> head(colnames(simdat),n=10)

[1] "outcome" "pc1"      "pc2"      "pc3"      "age"      "sex"      "SNP1"
[8] "SNP2"    "SNP3"    "SNP4"

> #coeff stores the 75 external regression coefficients.
> coeff=simdata$coeff
> head(coeff)

      coeff
SNP66 0.0192
SNP56 0.0230
SNP10 0.0229
SNP11 0.0307
SNP38 0.0174
SNP24 0.0402

>

```

We define the columns in simdat that contain genotype data:

```

> genotype_columns=which(grepl("^SNP",colnames(simdat)))

```

We next define the regression formula for outcome and adjusting covariates (variables to be adjusted for in addition to genotype variables, typically include age, gender, and principal components), as in other generic R regression functions:

```

> formula=as.formula("outcome~age+sex+pc1+pc2+pc3")
> formula

```

```

outcome ~ age + sex + pc1 + pc2 + pc3

```

Much of computation time is spent in calculating the correlation matrix of the surrogate direct effects as defined in the paper [1]. When there are many SNPs to be evaluated as candidate instrumental variables, parallel computing can be deployed to speed up the glide function by setting the corenumber argument. If the requested number of cores

is greater than number of cores available, GLIDE will use the latter number. GLIDE allows users to define a q-value cut-off that declares the significance of evaluating individual pleiotropy accounting for multiple testing. The default is `qcutoff=0.2`, to be conservative about validity of individual SNPs. Users can set the number of permutations for q-value computation. The default number is `np=100,000`. We then run the `glide` function with inputs as:

```
> out=glide(formula=formula,exposure_coeff=coeff,genotype_columns,data=simdat,
+           np=100000,qcutoff=0.2,parallel=TRUE,corenumber=1,verbose=TRUE)
```

GLIDE program starts at: 2021-10-12 10:06:34

Warning message: 3 rows were removed due to missing data in age

Compute the correlation matrix at: 2021-10-12 10:06:46

Total 75 SNPs...

There are 4 cores available in the machine.

Compute the null p-values at: 2021-10-12 10:07:23

Compute the FWER and FDR values at: 2021-10-12 10:07:30

GLIDE program ends at: 2021-10-12 10:07:34

```
> head(out)
```

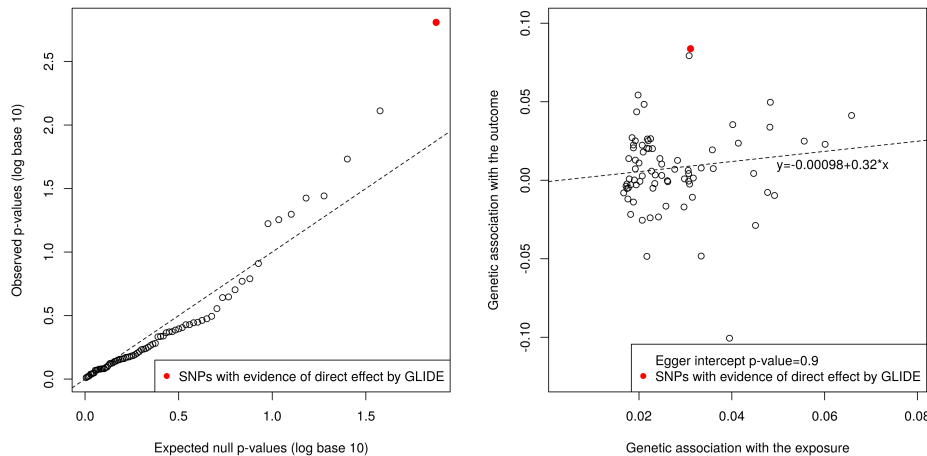
|       | observed_pvalue    | expected_pvalue | fwcr    | q_value   | g_outcome   |
|-------|--------------------|-----------------|---------|-----------|-------------|
| SNP60 | 0.001555538        | 0.01330080      | 0.11085 | 0.1177500 | 0.08377947  |
| SNP48 | 0.007731389        | 0.02646498      | 0.43795 | 0.2896350 | -0.04846745 |
| SNP45 | 0.018526927        | 0.03962386      | 0.75021 | 0.4623733 | -0.10059746 |
| SNP49 | 0.036145821        | 0.05284620      | 0.93488 | 0.5604800 | 0.05428275  |
| SNP43 | 0.037575697        | 0.06595674      | 0.94161 | 0.5604800 | 0.07930513  |
| SNP3  | 0.050387996        | 0.07912879      | 0.97797 | 0.5604800 | 0.04835978  |
|       | g_outcome_variance |                 |         |           |             |
| SNP60 | 0.0005712880       |                 |         |           |             |
| SNP48 | 0.0004120537       |                 |         |           |             |
| SNP45 | 0.0022840654       |                 |         |           |             |
| SNP49 | 0.0005370443       |                 |         |           |             |
| SNP43 | 0.0011821317       |                 |         |           |             |
| SNP3  | 0.0004534040       |                 |         |           |             |

### 3 plot result

GLIDE provides two functions for plotting the results: the first shows the q-q plot of individual variant's p-value for evaluating direct effects, the other shows the Egger regres-

sion plot with point estimates of genetic association with the exposure and the outcome. SNPs that were detected to have evidence of pleiotropy (for example,  $FDR < 0.2$ ) will be shown in both plots.

We draw the q-q plot for the example dataset: `plot.glide(out)`. We draw the Egger plot for the example dataset: `plot.egger(out, exposure_coeff=coeff)`. The figures are shown as follows: There is one SNP showing evidence of pleiotropy that may be deleted



in the subsequent Mendelian randomization analysis.

## 4 session information

The version number of R and packages loaded for generating the vignette were:

```
R version 3.5.3 (2019-03-11)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19043)
```

```
Matrix products: default
```

```
locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=Chinese (Simplified)_China.936
[3] LC_MONETARY=Chinese (Simplified)_China.936
[4] LC_NUMERIC=C
[5] LC_TIME=Chinese (Simplified)_China.936
```

```
attached base packages:
```

```
[1] stats      graphics  grDevices utils      datasets  methods   base
```

other attached packages:

```
[1] GLIDE_1.0.3
```

loaded via a namespace (and not attached):

```
[1] MASS_7.3-51.1      compiler_3.5.3      parallel_3.5.3      tools_3.5.3
```

```
[5] codetools_0.2-16   doParallel_1.0.16   iterators_1.0.13     foreach_1.5.1
```

## References

- [1] J. Y. Dai, U. Peters, X. Wang, J. Kocarnik, J. Chang-Claude, M. L. Slattery, A. Chan, M. Lemire, S. I. Berndt, G. Casey, M. Song, M. A. Jenkins, H. Brenner, A. P. Thrift, E. White, and L. Hsu. Diagnostics of pleiotropy in mendelian randomization studies: Global and individual tests for direct effects. *The American Journal of Human Genetics*, 2017, submitted.