Phenome Wide Association Studies (PheWAS) in R

Robert J. Carroll
Department of Biomedical Informatics
Vanderbilt University School of Medicine
phewas@vanderbilt.edu

March 24, 2015

Packge **PheWAS** provides methods for the creation of PheWAS phenotypes, analysis, and plotting.While these methods are designed primarily for genetics based PheWAS analysis, they can perform GWAS or even phenotype only studies.

1 Data Input

There are many potential data sources and types; this necessitates that users handle the basic data i/o and formatting. Below are outlined some methods for importing common data into R.

1.1 Preparing plink data

Genome wide data is stored commonly stored in plink formats[1]. The simplest method to import data from plink is the `--recodeA` parameter in plink. Running the following in a terminal will get one started:

```
plink --recodeA --bfile example_data --extract interesting_snps
--out r_genotypes
```

This will recode the binary plink data ”example_data”, extracting the SNPs under investigation to the file ”r_genotypes.raw”. This raw data can be loaded into R with a single command[2]:

```
genotypes=read.table("r_genotypes.raw",header=TRUE)
```

Alternatively, assuming IIDs are unique, the following will load the data ready to be put into `phewas`.

```
> genotypes=read.table("r_genotypes.raw",header=TRUE)[,c(-1,-3:-6)]
> names(genotypes)[1]="id"
```

1.2 Data from file

R has robust methods for loading data from files[3]. For this section we will consider an example where the user may have exported their chart review data into a csv from a spreadsheet software. *example_phenotype.csv*:

[1] See `http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml` for plink data format details.
[2] See `http://pngu.mgh.harvard.edu/~purcell/plink/dataman.shtml#recode` for details
[3] See `?read.table` in R for the read methods discussed here.

```
id,T2D,max.a1c
1,T,10
2,F,NA
3,F,6
...
```

This can be loaded using `csv.phenotypes=read.csv("example_phenotype.csv")`. This table loaded into R is ready to be used in `phewas`-either as covariates or phenotypes (outcomes).

1.3 Data from database

The **RODBC** library contains great tools for importing data directly from electronic data warehouses. If one desired to use PheWAS codes in their analysis from an ICD9 billing code table, it might look like the following.

```
> library(RODBC)
> connection=odbcConnect("MyDSN")
> icd9.codes=sqlQuery(connection,"select id, icd9, count(distinct date)
    from icd9_codes group by id, icd9;")
> odbcClose(connection)
```

The `icd9.codes` data frame is ready to be used with the `createPhewasTable` function.

2 Data Transformation

The primary data transformation for this package is to convert and aggregate ICD9 codes into PheWAS codes. The function `createPhewasTable` allows for this conversion. Given the database data loaded from the above section, one can use the following code to create PheWAS phenotypes for use in `phewas`:

```
> phenotypes=createPhewasTable(icd9.codes)
```

There are some additional options for PheWAS code translation. Users can opt to forgo exclusions using `add.exclusions=F`; this increases the size of the control population, but at the cost of including potentially similar diagnoses in the control sets. The `min.code.count` parameter allows users to alter the specificity of case selection. It can also be set to `NA` to allow for continuous outcomes, the code count sum by default.

3 Phenome Wide Association Studies

The `phewas` function performs the PheWAS itself. Using the examples from above, one can directly pass the parameters.

```
> results=phewas(phenotypes=phenotypes,genotypes=genotypes)
```

If one wishes to speed up the analysis, a multi-threaded approach is available using **snowfall**.

```
> results=phewas(phenotypes=phenotypes,genotypes=genotypes,cores=4)
```

One can additionally provide covariates. In this case, we will consider an analysis adjusted by max.a1c.

```
> results=phewas(phenotypes=phenotypes,genotypes=genotypes,

+    covariates=csv.phenotypes[,c("id","max.a1c")])
```

An alternate method is to use the `data` parameter with name vectors in the `phenotype`, `genotype`, and `covariates` parameters.

```
> mydata=merge(phenotypes,genotypes)
> results=phewas(phenotypes=names(phenotypes)[-1],genotypes=c("rs1234","rs5678"),

+    data=mydata)
```

The `phewas` function can be used for more than just generic PheWAS. In the following example, `outcomes` and `predictors` are used for a phenotype only analysis. Note that these parameters are simply alternate names for `phenotypes` and `genotypes`, respectively.

```
> max.a1c.results=phewas(outcomes=phenotypes,

+    predictors=csv.phenotypes[,c("id","max.a1c")])
```

The `phewasMeta` method can assist in meta-analysis of multiple PheWAS, e.g., if one has multiple genotype platforms of data to analyze. It wraps the `metagen` method of the **meta** package.

```
> results.omni1=phewas(phenotypes=phenotypes.omni1,genotypes=genotypes.omni1)
> results.omni1$study="Omni 1"
> results.omni.express=phewas(phenotypes=phenotypes.omni.express,

+    genotypes=genotypes.omni.express)
> results.omni.express$study="Omni Express"
> results.merged=rbind(results.omni1,results.omni.express)
> results.meta=phewasMeta(results.merged)
```

4 Plotting

Three methods for plotting data are included, `phewasManhattan`, `phenotypeManhattan`, and `phenotypePlot`, which wrap each other. `phewasManhattan` is the highest level method, and can plot PheWAS results directly from `phewas`.

```
> phewasManhattan(results)
```

This method returns a **ggplot2** object, which can be further manipulated using methods from that package[4]. The `...` parameter will pass further options into `phenotypeManhattan` and `phenotypePlot`. These lower level plot functions can be used in a stand-alone fashion for different types of data. For example, `phenotypePlot` can display information about the count for every individual of each ICD9 code.

```
> id.phenotype.value=icd9.codes
> names(id.phenotype.value)=c("id","phenotype","value")
> phenotypePlot(id.phenotype.value,use.color=F,x.group.labels=F)
```

[4] See `http://docs.ggplot2.org/current/` for the web documentation of **ggplot2**

5 Package Example

The following is the complete example from the **PheWAS** package.

```
> library(PheWAS)
> example(PheWAS)
```

```
PheWAS> #Install the recommended packages, if necessary
PheWAS> #install.packages(c("snowfall","shiny","MASS","meta"))
PheWAS> #Load the PheWAS package
PheWAS> library(PheWAS)

PheWAS> #Set the random seed so it is replicable
PheWAS> set.seed(1)

PheWAS> #Generate some example data
PheWAS> ex=generateExample()

PheWAS> #Extract the two parts from the returned list
PheWAS> id.icd9.count=ex$id.icd9.count

PheWAS> genotypes=ex$genotypes

PheWAS> #Create the PheWAS code table- translates the icd9s, adds
PheWAS> #exclusions, and reshapes to a wide format
PheWAS> phenotypes=createPhewasTable(id.icd9.count)

PheWAS> #Run the PheWAS
PheWAS> results=phewas(phenotypes,genotypes,cores=1,
PheWAS+   significance.threshold=c("bonferroni"))

PheWAS> #Plot the results
PheWAS> phewasManhattan(results, annotate.angle=0,
PheWAS+   title="My Example PheWAS Manhattan Plot")

PheWAS> #Add PheWAS descriptions
PheWAS> results_d=addPhewasDescription(results)

PheWAS> #List the significant results
PheWAS> results_d[results_d$bonferroni&!is.na(results_d$p),]
    phewas_code phewas_description       snp adjustment      beta      OR
495         335 Multiple sclerosis rsEXAMPLE       <NA> 0.4942269 1.63923
            SE           p     type n_total n_cases n_controls HWE_p
495 0.06611966 7.73601e-14 logistic    4416    1777       2639     1
    allele_freq n_no_snp note bonferroni
495   0.4987545        0            TRUE

PheWAS> #List the top 10 results
PheWAS> results_d[order(results_d$p)[1:10],]
     phewas_code               phewas_description       snp adjustment
```

```
495          335               Multiple sclerosis rsEXAMPLE       <NA>
414          293 Symptoms involving head and neck rsEXAMPLE       <NA>
456        313.2              Tics and stuttering rsEXAMPLE       <NA>
1301       694.1                         Vitiligo rsEXAMPLE       <NA>
924        527.2                    Sialoadenitis rsEXAMPLE       <NA>
1698         994                  Sepsis and SIRS rsEXAMPLE       <NA>
1700       994.2                           Sepsis rsEXAMPLE       <NA>
1441       736.5     Acquired deformities of knee rsEXAMPLE       <NA>
486        333.1                 Essential tremor rsEXAMPLE       <NA>
548       362.26      Macular puckering of retina rsEXAMPLE       <NA>
           beta        OR         SE            p     type n_total n_cases
495   0.4942269 1.6392305 0.06611966 7.736010e-14 logistic    4416    1777
414   1.3523545 3.8665187 0.36437457 2.060831e-04 logistic    4426      29
456  -0.9695479 0.3792545 0.26934689 3.186761e-04 logistic    4781      56
1301 -0.9887376 0.3720461 0.29622485 8.444620e-04 logistic    4300      46
924   0.9507643 2.5876867 0.30809623 2.029145e-03 logistic    4559      43
1698 -0.7829288 0.4570654 0.25514548 2.150942e-03 logistic    5000      64
1700 -0.9036474 0.4050894 0.29708980 2.352742e-03 logistic    4982      46
1441  0.8309801 2.2955675 0.29174434 4.395125e-03 logistic    4502      49
486   0.7225387 2.0596554 0.25386960 4.425805e-03 logistic    2709      70
548  -0.6966698 0.4982418 0.24841660 5.040388e-03 logistic    4113      71
     n_controls HWE_p allele_freq n_no_snp note bonferroni
495        2639     1   0.4987545        0            TRUE
414        4397     1   0.4957072        0           FALSE
456        4725     1   0.4953985        0           FALSE
1301       4254     1   0.4945349        0           FALSE
924        4516     1   0.4950647        0           FALSE
1698       4936     1   0.4957000        0           FALSE
1700       4936     1   0.4958852        0           FALSE
1441       4453     1   0.4971124        0           FALSE
486        2639     1   0.4785899        0           FALSE
548        4042     1   0.4989059        0           FALSE
```

```
> phewasManhattan(results, annotate.angle=0)
```

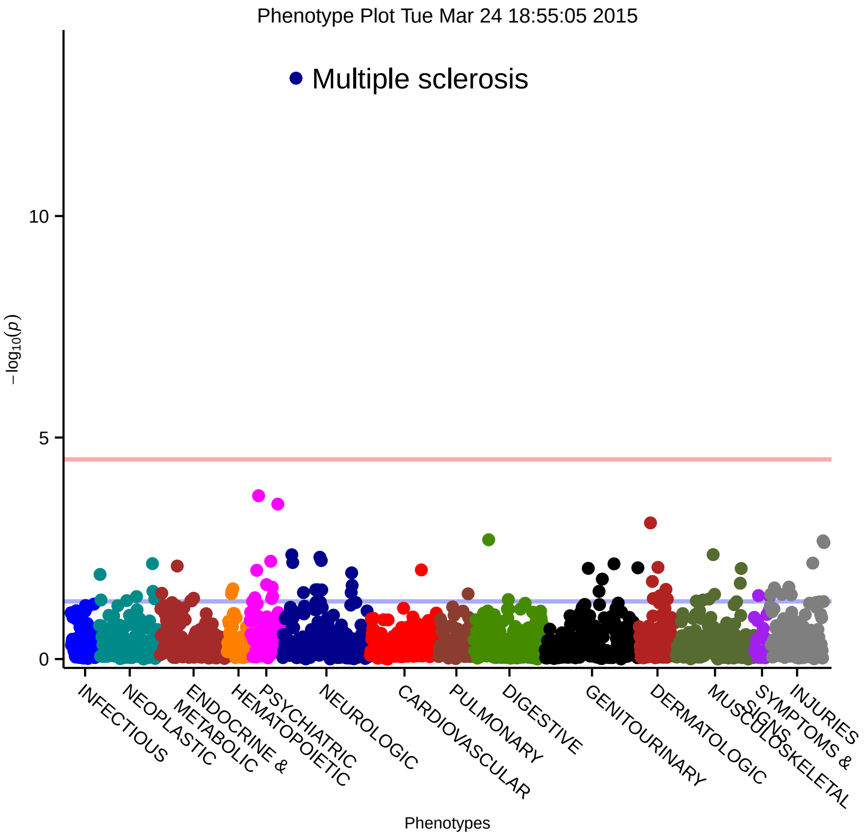

Figure 1: Example PheWAS Manhattan plot