traseR: TRait-Associated SNP EnRichment analyses

Li Chen, Zhaohui S.Qin

[1em]Department of Biostatistics and Bioinformatics
Emory University
Atlanta, GA 303022
[1em] `li.chen@emory.edu,zhaohui.qin@emory.edu`

May 1, 2024

Contents

Abstract

This vignette introduces the use of `traseR` (**TR**ait-**A**ssociated SNP**E**n**R**ichment analyses, which is designed to provide quantitative assessment whether a selected genomic interval(s) is likely to be functionally connected with certain traits or diseases. `traseR` consists of several modules, all written in R, to perform hypothesis testing, exploration and visualization of trait-associated SNPs(taSNPs). It also assembles the up-to-date taSNPs from dbGaP and NHGRI, SNPs from 1000 Genome Project CEU population with linkage disequilibrium greater than 0.8 within 100 kb of taSNPs, and all SNPs of CEU population from 1000 Genome project into its built-in database, which could be directly loaded when performing analyses.

1   Introduction

Genome-wide association study (GWAS) have successfully identified many sequence variants that are significantly associated with common diseases and traits. Tens of thousands of such trait-associated SNPs have already been cataloged which we believe are great resources for genomic research. However, no tools existing utilizes those resources in a comprehensive and convenient way. In this study, we show the collection of taSNPs can be exploited to indicate whether a query genomic interval(s) is likely to be functionally connected with certain traits or diseases. A R Bioconductor package named `traseR` has been developed to carry out such analyses.

2   Data collection

One great feature of `traseR` is the built-in database that collects various public SNP resources. Common public SNP databases include Association Result Browser and 1000 Genome Project. We briefly introduce the procedures to process those public available SNP resources

2.1   Obtain taSNPs

Association Results Browser (http://www.ncbi.nlm.nih.gov/projects/gapplusprev/sgap_plus.htm) combines identified taSNPs from dbGaP and NHGRI, which together provide 44,078 SNP-trait associations, 48,936 SNP-trait class associations, 30,553 unique taSNPs, 573 unique traits and 33 unique trait classes. This resource has been built into GRanges object `taSNP` and could be loaded into R console by typing `data(taSNP)`.

`traseR` need to specify the collection of trait-associated SNPs in particular format before we carry out enrichment analyses. The format starts with the columns,

1. Trait: Description of disease/trait examined in the study

2. Trait_Class: Trait class which is formed based on the phenotype tree. Close traits are grouped together to form one class.

3. SNP_ID: SNP rs number

4. p.value: GWAS reported p-values

5. seqnames: Chromosome number associated with rs number

6. ranges: Chromosomal position, in base pairs, associated with rs number

7. Context: SNP functional class

8. GENE_NAME: Genes reported to be associated with SNPs

9. GENE_START: Chromosome start position of genes

10. GENE_END: Chromosome end position of genes

11. GENE_STRAND: Chromosome strand associated with SNPs

Currently, the `traseR` package automatically synchronize trait-associated SNPs from Association Results Browser, which collects up-to-date GWAS results from dbGaP NHGRI GWAS catalog.

2.2   Obtain linkage disequilibrium taSNPs from 1000 Genome Project

We first download CEU vcf files from (ftp://share.sph.umich.edu/1000genomes/fullProject/2012.03.14/) that contain all sequence variants information. The followed two steps are used to identify linkage disequilibrium SNPs >0.8 and located within 100kb of taSNPs. Firstly, we use `vcftools` to convert the vcf file format to `PLINK` format. Then we use `PLINK` to call the LD taSNPs by specifying options that limit the linkage disequilibrium SNPs >0.8 (–ld-window-r2 0.8) and within 100kb of taSNP (–ld-window-kb 100). The detailed commands are listed below,

```
vcftools -vcf vcf.file -out plink.file -plink plink -file plink.file -r2 -inter-
chr -ld-snp-list snps.txt -ld-window-r2 0.8 -ld-window-kb 100 -out output.file -
noweb
```

Finally, we have 90,700 SNP-trait associations and 78,247 unique linkage disequilibrium trait-associated SNP. We also build linkage disequilibrium taSNP into another GRanges object `taSNPLD`, which could be loaded into R console by typing `data(taSNPLD)`.

The format of `taSNPLD` is,

1. seqnames: Chromosome number associated with rs number

2. SNP_ID: SNP rs number

3. ranges: Chromosomal position, in base pairs, associated with rs number

4. Trait: Description of disease/trait examined in the study

5. Trait_Class: Trait class which is formed based on the phenotype tree. Close traits are grouped together to form one class.

2.3 Obtain background SNPs from 1000 Genome Project

We use the command `plink -file plink.file -freq -out chr` to retrieve all SNPs with corresponding MAF (minor allele frequency) from the CEU vcf files downloaded. There are totally 6,571,512 SNPs (MAF>0.05) excluding variants on Y chromosome. Those SNPs could serve as background in hypothesis testing. We build those SNPs into the built-in GRanges subject `CEU` into the package.

The format of `CEU` is,

1. seqnames: Chromosome number associated with rs number

2. SNP_ID: SNP rs number

3. ranges: Chromosomal position, in base pairs, associated with rs number

3 Using `traseR`

To assess the enrichment level of trait-associated SNPs in given genomic interval(s) using `traseR`, one needs to follow the simple steps below.

1. Prepare the genomic intervals in R object of either data frame format with column names chr,start,end or GRanges object

2. Query a given a set of genomic interval(s) against all the taSNPs in the collection, perform statistical analyses

3. Explore genes/SNPs of particular interest

3.1 Background selection

3.1.1 Whole genome

The assumption is each base could be possibly be the taSNP. Based on the assumption, with the number of taSNPs inside and outside the genomic interval(s), the number of bases inside and outside of the genomic interval(s), we could classify all bases based on the fact that one base is taSNP or not and in genomic intervals or not.

3.1.2   All SNPs

The assumption is each SNP could possibly be the taSNP. Based on the assumption, with the number of taSNPs inside and outside the genomic interval(s), the non-taSNPs inside and outside of the genomic interval(s), we could classify all SNPs based on the fact that one SNP is taSNP or not and in genomic intervals or not.

3.2   Hypothesis testing

`traseR` provides differential hypothesis testing methods in core function `traseR`, together with other functions for exploring and visualizing the results. The genomic interval(s) could be a data frame with three columns as `chr`(chromosome), `start`(genomic start position) and `end`(genomic end position) or a GRanges object. `traseR` offers either including LD SNPs or excluding LD SNPs as the taSNPs and either using the whole genome or all SNPs as the background for hypothesis testing.

If using whole genome as background, the command line is:

```
> x=traseR(snpdb=taSNP,region=Tcell)
> print(x)
```

If including the LD SNPs, the command line is:

```
> x=traseR(snpdb=taSNPLD,region=Tcell)
> print(x)
```

If using all SNPs as background, the command line is:

```
> x=traseR(snpdb=taSNP,region=Tcell,snpdb.bg=CEU)
```

For the above commands, `region` is the data frame; `snpdb` is taSNPs or including LD SNPs; `snpdb.bg` is background SNPs; If `rankby` is set as "pvalue", all traits will be sorted by p-value in increasing order; if`rankby` is set as "odds.ratio", all traits will be sorted by odds ratio in decreasing order. There are four options for `test.method` including "binomial", "chisq", "fisher", and "nonparametric"to perform binomial test, χ^2 test, Fisher's exact test and nonparametric respectively. If `alternative` is set to "greater", `traseR` will perform hypothesis testing on whether genomic intervals are enriched of taSNPs than the background; If `alternative` is set to "less",`traseR` will perform hypothesis testing on whether genomic intervals are depleted of taSNPs than the background.

3.2.1 χ^2 test and Fisher's exact test

Based on which background we choose, we could construct the 2 by 2 contingency table. then, we could perform χ^2 test on the table to assess the difference of proportions of taSNPs inside and outside of genomic intervals(s). We could also assume taSNPs inside genomic intervals follows hypergeometric distribution and calculate p-value directly using Fisher's exact test.

3.2.2 Binomial test

The assumption is the probability of observing a single base/SNP being a taSNP is the same inside and outside of genomic intervals. The probability of observing a single base/SNPs being a taSNP in genomic intervals could be estimated by using total number of taSNPs divided by the genome size/number of all SNPs. Then corresponding p-value could be calculated directly by Binomial test.

3.2.3 Nonparametric Test

Instead of imposing any assumption, the matched genomic interval(s) are generated by permuting the genomic intervals randomly N times and overlap with taSNPs in each time. Then we could calculate the empirical p-value directly by counting how many taSNP hits larger/smaller than the observed taSNP hits.

4 Choose appropriate statistical test method

Depending on the characteristics of the test statistics, we suggest to choose appropriate statistical test method under different scenarios,

- χ^2 test: the numbers in the contingency table is fairly large and balanced

- Fisher's exact test: the numbers of the contingency table is relatively small, e.g. no more than 20

- Nonparametric test: the number of query genomic intervals are small, e.g. no more than 1000

- Binomial test: default test method, not limited by sample size, distribution assumption and computational time

4.1 Example

To further illustrate the usage of `traseR` R package, we download H3K4me1 peak regions in peripheral blood T cell from Roadmap Epigenomics. Those peak regions are deemed the genomic intervals. Since the degree of enrichment level is measured by p-value, we could

rank traits/trait classes based on p-value in an increasing order. We choose Binomial test are the default option for `test.method`, use whole genome as background and over-enrichment as hypothesis testing direction.

```
> library(traseR)
> data(taSNP)
> data(Tcell)
> x=traseR(taSNP,Tcell)
> print(x)
   Trait        p.value odds.ratio taSNP.hits taSNP.num
1  All 3.788373e-233    2.134717       2625     30553
                           Trait      p.value      q.value odds.ratio
67                Behcet Syndrome 4.400406e-23 2.521433e-20   6.306579
172      Diabetes Mellitus, Type 1 1.704981e-11 4.884769e-09   5.045263
340 Lupus Erythematosus, Systemic 6.159346e-09 1.176435e-06   3.902195
49          Arthritis, Rheumatoid 1.442123e-07 2.065841e-05   5.126637
379             Multiple Sclerosis 1.644125e-05 1.884167e-03   2.905210
62             Autoimmune Diseases 5.201529e-05 4.967461e-03  15.892575
    taSNP.hits taSNP.num
67          59       274
172         33       185
340         32       223
49          20       112
379         26       236
62           6        15
                        Trait_Class      p.value      q.value odds.ratio
17               Immune System Diseases 3.729169e-35 1.143835e-33   3.658860
31 Skin and Connective Tissue Diseases 6.932335e-35 1.143835e-33   3.916319
32               Stomatognathic Diseases 1.041455e-22 1.145601e-21   5.675922
14                        Eye Diseases 3.479491e-18 2.870580e-17   3.313308
11             Digestive System Diseases 4.362324e-14 2.879134e-13   3.040672
7                 Cardiovascular Diseases 3.008551e-11 1.654703e-10   1.602762
13               Endocrine System Diseases 6.933337e-09 3.268573e-08   2.068149
24   Nutritional and Metabolic Diseases 4.763509e-08 1.964948e-07   2.068673
21               Musculoskeletal Diseases 3.118359e-05 1.143398e-04   2.716680
23               Nervous System Diseases 5.549981e-05 1.831494e-04   1.495744
16         Hemic and Lymphatic Diseases 1.649261e-04 4.947782e-04   3.596622
22                           Neoplasms 3.372076e-04 9.273210e-04   1.580636
30           Respiratory Tract Diseases 6.507770e-04 1.651972e-03   2.121839
```

```
     taSNP.hits taSNP.num
17          155      1122
31          142       970
32           63       318
14           87       689
11           74       633
7           253      3850
13           89      1076
24           79       956
21           27       260
23          122      1988
16           15       115
22           76      1181
30           29       349
```
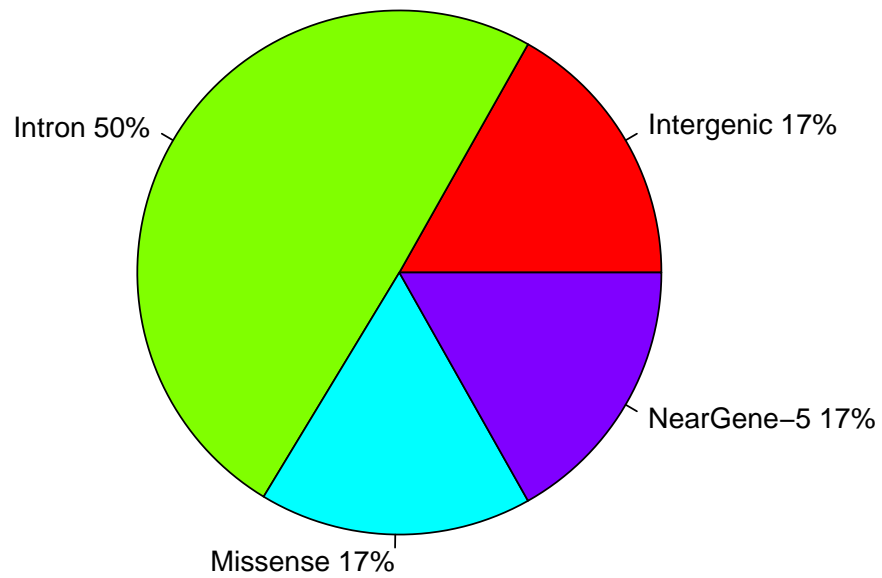
4.2   Exploratory and visualization functions

Plot the distribution of SNP functional class

```
> plotContext(snpdb=taSNP,region=Tcell,keyword="Autoimmune")
```
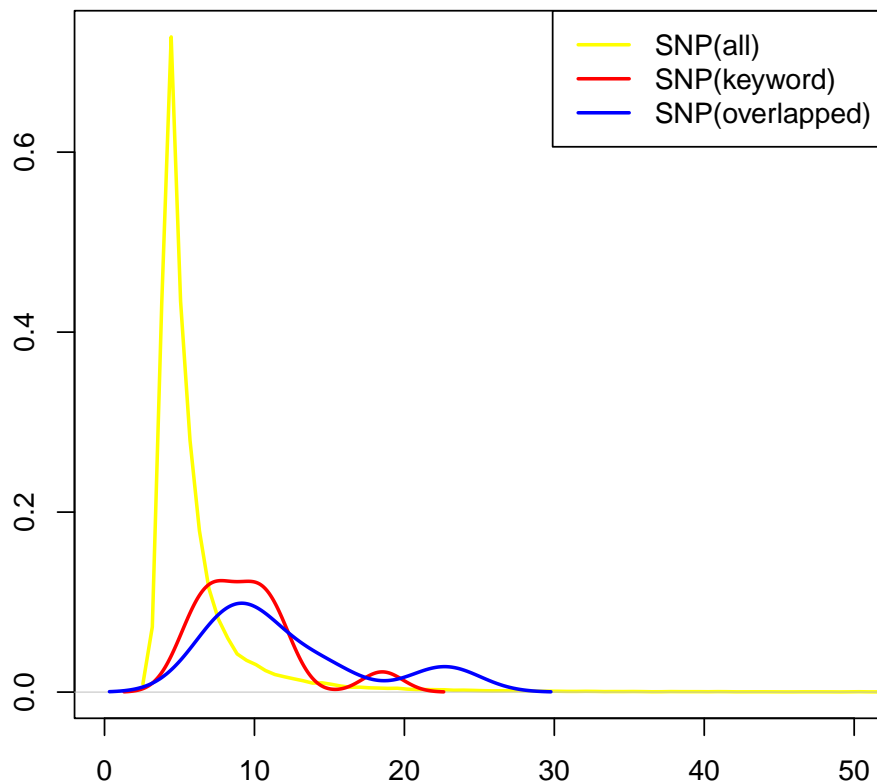
Pie Chart of Context

Intron 50%

Intergenic 17%

NearGene−5 17%

Missense 17%

Plot the distribution of p-value of trait-associated SNPs

```
> plotPvalue(snpdb=taSNP,region=Tcell,keyword="autoimmune",plot.type="densityplot")
```
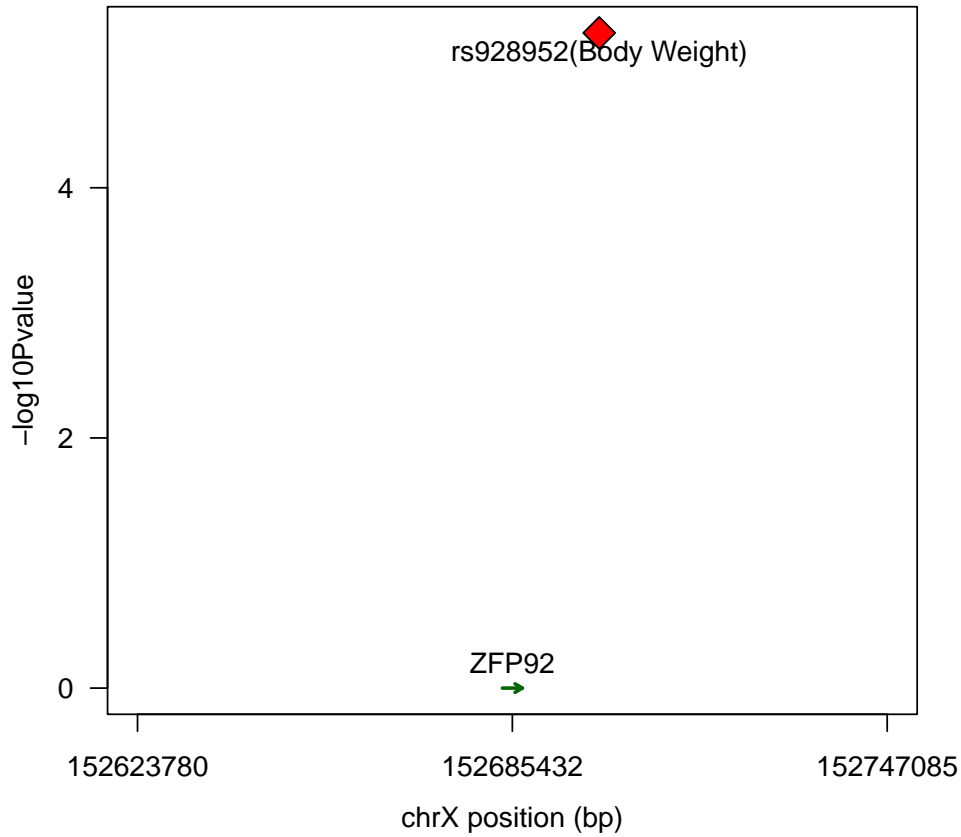
–log10Pvalue Distribution

Plot SNPs or genes given genomic interval

```
> plotInterval(snpdb=taSNP,data.frame(chr="chrX",start=152633780,end=152737085))
```

Query trait-associated SNPs by key word,

```
> x=queryKeyword(snpdb=taSNP,region=Tcell,keyword="autoimmune",returnby="SNP")
> head(x)

          SNP_ID   Chr  Position Trait.num          Trait.name
4343 rs11203203 chr21  43836186         1 Autoimmune Diseases
4341  rs1876518  chr2  65608909         1 Autoimmune Diseases
4348  rs1953126  chr9 123640500         1 Autoimmune Diseases
4342  rs2298428 chr22  21982892         1 Autoimmune Diseases
4345  rs7579944  chr2  30445026         1 Autoimmune Diseases
4338   rs864537  chr1 167411384         1 Autoimmune Diseases
```

Query trait-associated SNPs by gene name,

```
> x=queryGene(snpdb=taSNP,genes=c("AGRN","UBE2J2","SSU72"))
> x

GRanges object with 3 ranges and 5 metadata columns:
      seqnames            ranges strand |   GENE_NAME Trait.num
         <Rle>         <IRanges>  <Rle> | <character> <integer>
  [1]     chr1   955502-991491      + |        AGRN         1
  [2]     chr1 1477052-1510261      - |       SSU72         1
  [3]     chr1 1189291-1209233      - |      UBE2J2         1
              Trait.name taSNP.num  taSNP.name
             <character> <integer> <character>
  [1]     Body Mass Index         1   rs3934834
  [2]             Glucose         1    rs880051
  [3] Waist Circumference         1  rs11804831
  -------
  seqinfo: 23 sequences from an unspecified genome; no seqlengths
```

Query trait-associated SNPs by SNP name,

```
> x=querySNP(snpdb=taSNP,snpid=c("rs3766178","rs880051"))
> x

GRanges object with 2 ranges and 9 metadata columns:
        seqnames    ranges strand |       Trait      SNP_ID   p.value
           <Rle> <IRanges>  <Rle> | <character> <character> <numeric>
  42234     chr1   1478180      * |     Glucose   rs3766178  3.26e-05
  42127     chr1   1493727      * |     Glucose    rs880051  6.44e-05
             Context   GENE_NAME GENE_START   GENE_END GENE_STRAND
         <character> <character>  <integer>  <integer> <character>
  42234      Intron       SSU72    1477052    1510261           -
  42127      Intron       SSU72    1477052    1510261           -
                 Trait_Class
                 <character>
  42234 Chemicals and Drugs ..
  42127 Chemicals and Drugs ..
  -------
  seqinfo: 23 sequences from an unspecified genome; no seqlengths
```

5   Conclusion

`traseR` provides methods to assess the enrichment level of taSNPs in a given sets of genomic intervals. Moreover, it provides other functionalities to explore and visualize the results.

6   Session Info

```
> sessionInfo()

R version 4.4.0 RC (2024-04-16 r86468)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 22.04.4 LTS

Matrix products: default
BLAS:   /home/biocbuild/bbs-3.20-bioc/R/lib/libRblas.so
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB              LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

time zone: America/New_York
tzcode source: system (glibc)

attached base packages:
[1] stats4    stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] traseR_1.35.0                    BSgenome.Hsapiens.UCSC.hg19_1.4.3
 [3] BSgenome_1.73.0                  rtracklayer_1.65.0
 [5] BiocIO_1.15.0                    Biostrings_2.73.0
 [7] XVector_0.45.0                   GenomicRanges_1.57.0
 [9] GenomeInfoDb_1.41.0              IRanges_2.39.0
[11] S4Vectors_0.43.0                 BiocGenerics_0.51.0
```

```
loaded via a namespace (and not attached):
 [1] Matrix_1.7-0            BiocStyle_2.33.0
 [3] jsonlite_1.8.8          compiler_4.4.0
 [5] BiocManager_1.30.22     rjson_0.2.21
 [7] crayon_1.5.2            SummarizedExperiment_1.35.0
 [9] Biobase_2.65.0          GenomicAlignments_1.41.0
[11] Rsamtools_2.21.0        bitops_1.0-7
[13] parallel_4.4.0          BiocParallel_1.39.0
[15] yaml_2.3.8              fastmap_1.1.1
[17] lattice_0.22-6          R6_2.5.1
[19] S4Arrays_1.5.0          curl_5.2.1
[21] knitr_1.46              XML_3.99-0.16.1
[23] DelayedArray_0.31.0     MatrixGenerics_1.17.0
[25] GenomeInfoDbData_1.2.12 rlang_1.1.3
[27] xfun_0.43               SparseArray_1.5.0
[29] cli_3.6.2               zlibbioc_1.51.0
[31] grid_4.4.0              digest_0.6.35
[33] evaluate_0.23           codetools_0.2-20
[35] abind_1.4-5             RCurl_1.98-1.14
[37] restfulr_0.0.15         rmarkdown_2.26
[39] httr_1.4.7              matrixStats_1.3.0
[41] tools_4.4.0             htmltools_0.5.8.1
[43] UCSC.utils_1.1.0
```

References

[1] Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L et al (2010). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acid Research*, **42**, D1001-1006.

[2] Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al. (2015). Integrative analysis of 111 reference human epigenomes *Nature*, **7539**, 317-330

[3] http://www.ncbi.nlm.nih.gov/projects/gapplusprev/sgap_plus.htm *Association Results Browser*

[4]   ftp://share.sph.umich.edu/1000genomes/fullProject/2012.03.14/ *1000Genome EUR*