ldblock package: linkage disequilibrium data structures

Vincent J. Carey, stvjc at channing.harvard.edu

February 2015

Contents

1 Introduction

There is a nice vignette in *snpStats* concerning linkage disequilibrium (LD) analysis as supported by software in that package. The purpose of this package is to simplify handling of existing population-level data on LD for the purpose of flexibly defining LD blocks.

2 Import of HapMap LD data

The `hmld` function imports gzipped tabular data from hapmap's repository hapmap.ncbi.nlm.nih.gov/downloads/ld_data/2009-02_phaseIII_r2/.
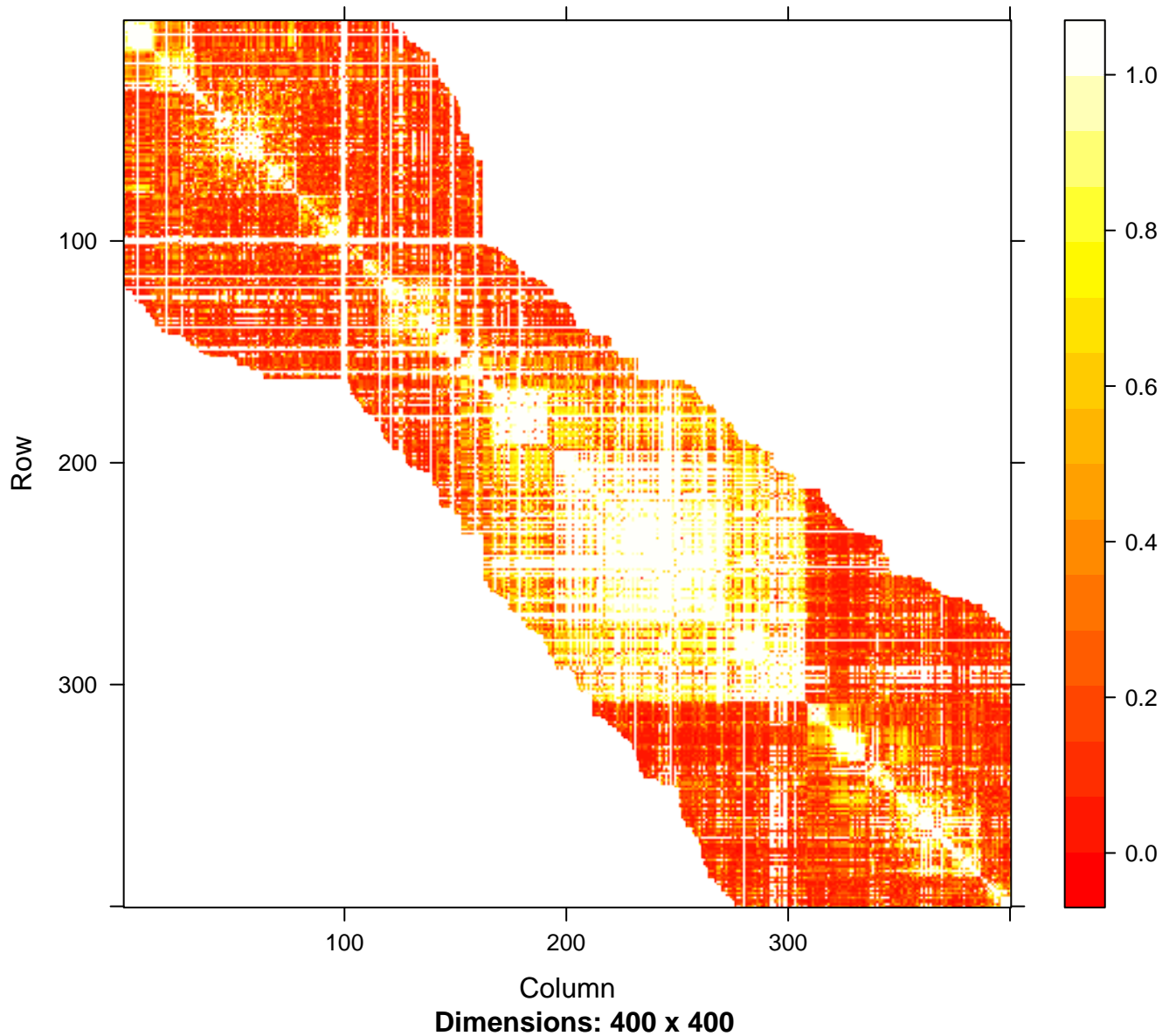
```
suppressPackageStartupMessages(library(ldblock))
path = dir(system.file("hapmap", package="ldblock"), full=TRUE)
ceu17 = hmld(path, poptag="CEU", chrom="chr17")
## importing /tmp/RtmpFVWmBo/Rinst32b1728aa0d8/ldblock/hapmap/ld_chr17_CEU.txt.gz
## done.
```

```
ceu17
## ldstruct for population CEU, chrom chr17
## dimensions: 36621 x 36621 ; statistic is Dprime
## object structure:
##       ldmat       chrom      genome      allpos      poptag    statInUse
## "dsCMatrix" "character" "character"   "numeric" "character" "character"
##       allrs
## "character"
```

3 A view of the block structure

For some reason knitr/render will not display this image nicely.

```r
library(Matrix)
image(ceu17@ldmat[1:400,1:400],
    col.reg=heat.colors(120), colorkey=TRUE, useRaster=TRUE)
```


Dimensions: 400 x 400

This ignores physical distance and MAF. The bright stripes are probably due to SNP with low MAF.

4   Collecting SNPs exhibiting linkage to selected SNP

We'll use ceu17 and the gwascat package to enumerate SNP that are in LD with GWAS hits.

```r
library(gwascat)
## Loading required package: Homo.sapiens
## Loading required package: AnnotationDbi
## Loading required package: stats4
```

```
## Loading required package: BiocGenerics
## Loading required package: parallel
##
## Attaching package: 'BiocGenerics'
##
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
##
## The following object is masked from 'package:Matrix':
##
##     as.vector
##
## The following objects are masked from 'package:stats':
##
##     IQR, mad, xtabs
##
## The following objects are masked from 'package:base':
##
##     Filter, Find, Map, Position, Reduce, anyDuplicated, append,
##     as.data.frame, as.vector, cbind, colnames, do.call,
##     duplicated, eval, evalq, get, grep, grepl, intersect,
##     is.unsorted, lapply, lengths, mapply, match, mget, order,
##     paste, pmax, pmax.int, pmin, pmin.int, rank, rbind, rownames,
##     sapply, setdiff, sort, table, tapply, union, unique, unlist,
##     unsplit
##
## Loading required package: Biobase
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
##
## Loading required package: IRanges
## Loading required package: S4Vectors
##
## Attaching package: 'IRanges'
##
## The following object is masked from 'package:Matrix':
##
##     expand
##
## Loading required package: OrganismDbi
## Loading required package: GenomicFeatures
## Loading required package: GenomeInfoDb
## Loading required package: GenomicRanges
## Loading required package: GO.db
## Loading required package: DBI
##
## Loading required package: org.Hs.eg.db
##
```

```
## Loading required package: TxDb.Hsapiens.UCSC.hg19.knownGene
## Now getting the GODb Object directly
## Now getting the OrgDb Object directly
## Now getting the TxDb Object directly
## gwascat loaded.  Use data(ebicat38) for hg38 coordinates;
##   data(ebicat37) for hg19 coordinates.
```

```
data(ebicat37)
e17 = ebicat37[ which(seqnames(ebicat37) == "17") ]
```

Some dbSNP names for GWAS hits on chr17 are

```
rsh17 = unique(e17$SNPS)
head(rsh17)
## [1] "rs4986172" "rs2270518" "rs3785574" "rs757608"  "rs2854160" "rs4794665"
```

We will use expandSnpSet to obtain names for SNP that were found in HapMap CEU to have which $D' > .9$ with any of these hits. These names are added to the input set.

```
length(rsh17)
## [1] 495
```

```
exset = expandSnpSet( rsh17, ldstruct= ceu17, lb=.9 )
## Warning in expandSnpSet(rsh17, ldstruct = ceu17, lb = 0.9): dropping 138
## rsn not matched in ld matrix
```

```
length(exset)
## [1] 12738
```

```
all(rsh17 %in% exset)
## [1] TRUE
```

Not all GWAS SNP are in the HapMap LD resource. You can use your own LD data as long as the format agrees with that of the HapMap distribution.