Linkage Disequilibrium: Lab 2 SOlutions

Introduction

Linkage Disequilibrium with Genotype Data

```
library(genetics)

## Loading required package:  combinat
##
## Attaching package:  'combinat'
##
## The following object is masked from 'package:utils':
##
##     combn
##
## Loading required package:  gdata
## gdata:  Unable to locate valid perl interpreter
## gdata:
## gdata:  read.xls() will be unable to read Excel XLS and XLSX files
## gdata:  unless the 'perl=' argument is used to specify the location
## gdata:  of a valid perl intrpreter.
## gdata:
## gdata:  (To avoid display of this message in the future, please
## gdata:  ensure perl is installed and available on the executable
## gdata:  search path.)
## gdata:  Unable to load perl libaries needed by read.xls()
## gdata:  to support 'XLX' (Excel 97-2004) files.
##
## gdata:  Unable to load perl libaries needed by read.xls()
## gdata:  to support 'XLSX' (Excel 2007+) files.
##
## gdata:  Run the function 'installXLSXsupport()'
## gdata:  to automatically download and install the perl
## gdata:  libaries needed to support Excel XLS and XLSX formats.
##
## Attaching package:  'gdata'
##
## The following object is masked from 'package:stats':
##
##     nobs
##
## The following object is masked from 'package:utils':
##
##     object.size
##
## Loading required package:  gtools
## Loading required package:  MASS
## Loading required package:  mvtnorm
##
##
## NOTE: THIS PACKAGE IS NOW OBSOLETE.
```

```
##
##
##
##   The R-Genetics project has developed an set of enhanced genetics
##
##   packages to replace 'genetics'.  Please visit the project homepage
##
##   at http://rgenetics.org for informtion.
##
##
##
##
## Attaching package:  'genetics'
##
## The following objects are masked from 'package:base':
##
##     %in%, as.factor, order
```

```
load("genotypes_chr1.RData")
```

Exercise 1

Compare the following r^2 values:

1. sites from columns 75 and 76 to the r^2 values we found in lab 1 for these site combinations.

   ```
   genosite75 <- as.genotype.allele.count(geno1[, 75])
   genosite76 <- as.genotype.allele.count(geno1[, 76])
   LD(genosite75, genosite76)$"R^2"

   ## [1] 0.9592
   ```

2. sites from columns 75 and 77

   ```
   genosite77 <- as.genotype.allele.count(geno1[, 77])
   LD(genosite75, genosite77)$"R^2"

   ## [1] 0.01919
   ```

3. sites from columns 75 and 80

   ```
   genosite80 <- as.genotype.allele.count(geno1[, 80])
   LD(genosite75, genosite80)$"R^2"

   ## [1] 0.003511
   ```

In each case the r^2 values are close by not exactly the values found in Lab 1 for these site combinations. Contrary to Lab 1 LD between sites 75 and 77 is greater than 77 and 80.

```r
convertgenos <- function(start, end) {
    as.genos <- function(x) {
        as.genotype.allele.count(geno1[, x])
    }
    siteS <- as.matrix(c(start:end))
    sitekeep <- which(!colSums(geno1[, start:end], na.rm = TRUE) == 0)
    siteS <- as.matrix(siteS[sitekeep, ])
    genos <- (apply(siteS, 1, as.genos))
    colnames(genos) <- colnames(geno1)[which(!colSums(geno1[, start:end], na.rm = TRUE) ==
        0)]
    genos <- makeGenotypes(genos)
}
```

Exercise 2

We can now use the `LDheatmap()` function to create heat maps of r^2 values. The use of this function is demonstrated below:

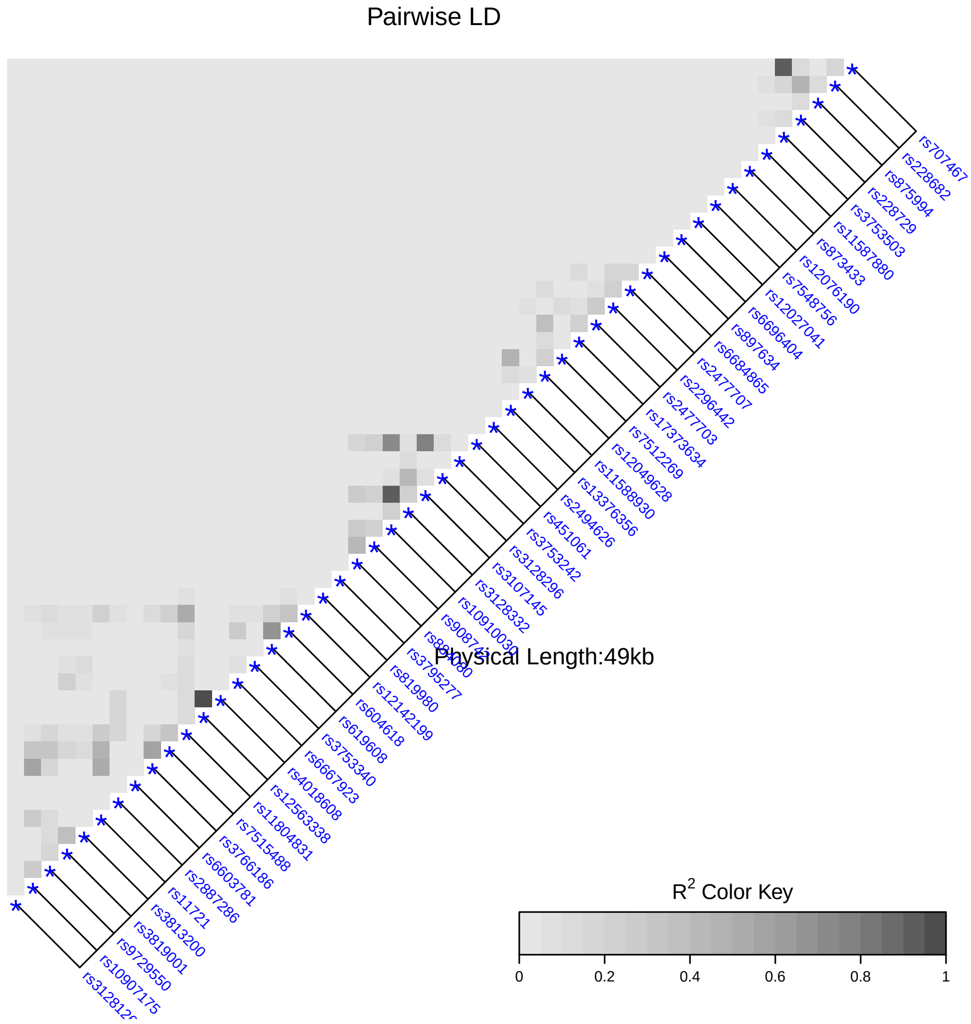
```r
require(LDheatmap)

## Loading required package:  LDheatmap
## Loading required package:  grid

# LDheatmap(genos, LDmeasure='r', SNP.name=colnames(genos))
```

1. Create a new genotype object of the first 50 sites from the `geno1` object and create a heatmap from this object. *Both steps may take a little while.*

```r
genos50 <- convertgenos(1, 50)
LDheatmap(genos50, LDmeasure = "r", SNP.name = colnames(genos50))
```



2. What pairs of sites appear to have the greatest LD?

   There appears to be high LD between 3 pairs of sites: rs12563338 and rs4018608; rs10910030 and rs3107145; rs707467 and rs3753503 .

3. Are there groups of sites that you notice to have higher LD than others? Between which sites does these occur? There appears to be four clusters of noticeable LD. The first is between sites: rs10907175 and rs12142199. The second is between sites: rs884080 and rs451061. The third is between sites: rs13376356 and rs6684855. The last is between sites: rs3753503 and rs707467(the final site).

4. What pair of sites have the greatest LD?

```
genomat50 <- LD(genos50)$"R^2"
maxld <- which(genomat50 == max(genomat50, na.rm = TRUE), arr.ind = TRUE)
rownames(genomat50)[maxld[1]]

## [1] "rs12563338"

colnames(genomat50)[maxld[2]]

## [1] "rs4018608"
```

5. What is the average LD of the first 50 sites?

```
mean(genomat50, na.rm = TRUE)

## [1] 0.02077
```

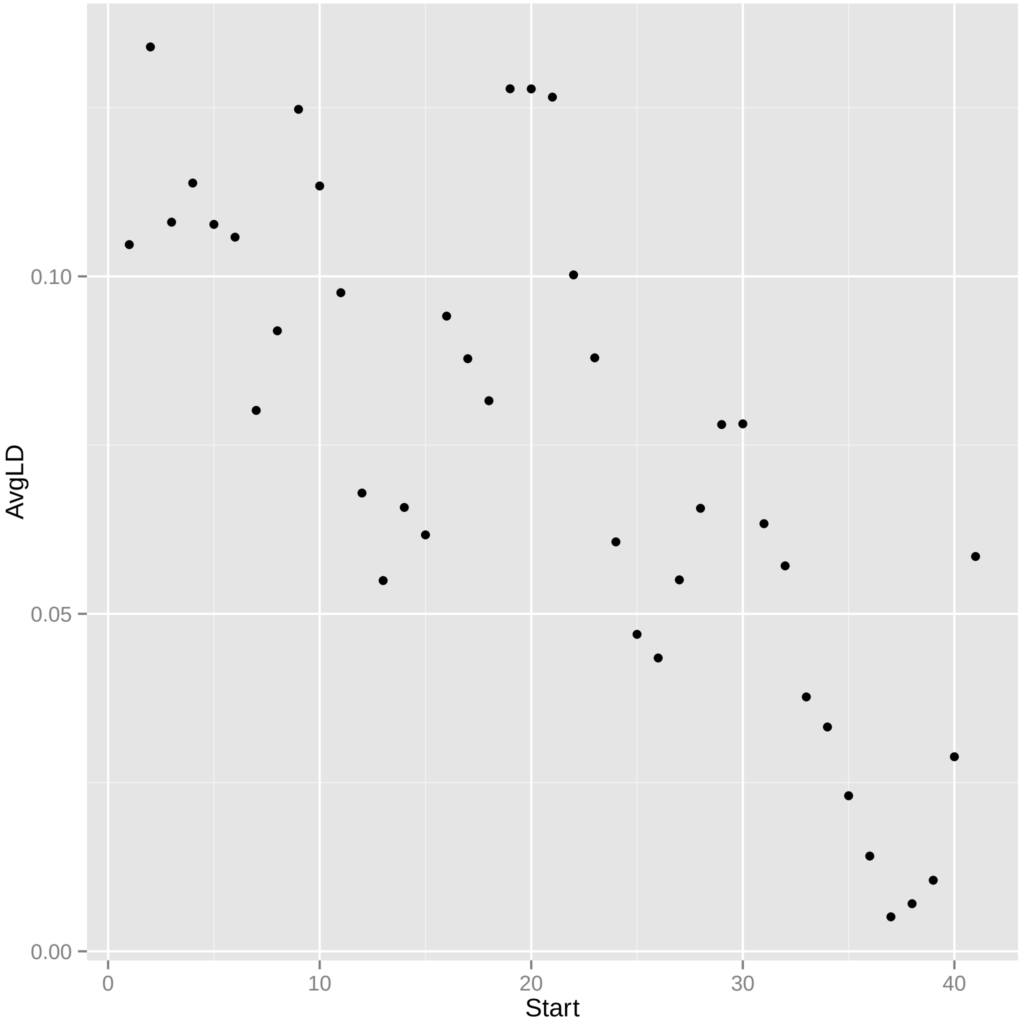
```
avgld <- function(genos, start, end, length) {
    lo <- start:(end - length + 1)
    hi <- length:end
    vec <- rbind(lo, hi)
    avgld <- function(vec) {
        start <- vec[1]
        end <- vec[2]
        genomat <- genos[start:end]
        mat <- LD(genomat)$"R^2"
        avg <- mean(mat, na.rm = TRUE)
        cbind(avg, start, end)
    }
    avgLD <- t(apply(vec, 2, avgld))
    colnames(avgLD) <- c("AvgLD", "Start", "Stop")
    avgLD
}
```

6. Plot the moving average of the first 50 sites, with a window length of 10? Is it consistent with your results from *question 2*?

```
require(ggplot2)

## Loading required package:  ggplot2

movavg50 <- avgld(genos50, 1, 50, 10)
movavg50 <- as.data.frame(movavg50)
qplot(Start, AvgLD, data = movavg50)
```



In the plot there appears to be four distinct peaks around a starting points of 2, 20, 30, and 41, which is consistent with question 2.