Computing LD

Alencar Xavier and Bill Muir

September 7, 2015

Suppose you have two alleles from a population of 196 individuals. The example below are the SNPs *Gm01_3321482* and *Gm01_39533385* from chromosome 1 in soybeans (source: R package NAM).

```
Allele1
```

```
##   [1] 2 0 2 0 1 0 2 2 0 0 1 2 0 2 2 0 0 0 1 2 1 0 2 2 0 0 2 0 0 2 0 0 0 2 2
##  [36] 0 0 0 0 0 0 2 1 1 2 0 2 2 2 0 2 0 2 2 2 2 2 2 2 1 2 2 2 2 0 2 0 2 2 2
##  [71] 2 0 2 2 2 2 2 2 2 2 2 0 0 0 0 0 0 0 2 2 0 2 0 2 2 2 0 0 1 0 0 2 2 2 0
## [106] 0 2 0 0 2 0 0 1 2 0 0 0 2 2 0 0 2 1 2 2 2 2 2 0 0 0 1 0 2 2 2 1 2 2 2
## [141] 1 1 1 0 2 0 0 0 0 2 0 0 2 0 2 2 2 0 0 1 2 0 0 0 0 2 0 2 0 2 0 0 0 2 2
## [176] 2 2 0 0 2 0 2 2 2 2 0 2 2 0 0 2 2 0 0 2 0
```

```
Allele2
```

```
##   [1] 0 2 0 2 0 2 0 0 2 2 1 0 2 0 0 2 2 2 1 0 0 2 0 0 2 2 0 2 2 1 2 2 2 0 0
##  [36] 2 2 2 2 2 2 2 0 0 1 0 2 0 0 0 2 0 2 2 0 0 0 0 0 0 1 0 0 1 2 2 2 2 0 0 0
##  [71] 0 2 0 0 0 0 0 0 0 0 0 2 2 2 2 2 1 2 0 0 2 0 2 2 0 2 2 2 1 2 2 0 2 2 2
## [106] 2 0 2 2 1 2 2 2 0 2 2 2 0 0 2 2 0 1 0 0 0 0 0 2 2 2 1 2 0 1 2 1 0 0 0
## [141] 2 1 0 2 0 2 2 2 2 0 2 2 0 0 2 0 0 2 2 0 0 0 0 2 2 0 2 0 2 0 0 2 2 0 0
## [176] 0 0 2 2 2 2 0 2 0 0 2 0 2 2 2 0 0 2 2 2 2
```

To compute LD it is necessary to phase the haplotypes. Phasing is commonly done through the Expectation-Maximization algorithm proposed by Marita Olsson (1996). I personaly coded the EM algorithm as follows (for 012 genotypes):

```
EM=function(A,B,n=12){
  Z=suppressWarnings(matrix(c(table(paste(A,B,sep=""))),3,3))
  # Initial guess
  COUP = 0.5 ; REPU = 0.5
  # Function to estimate haplotypes
  HapProb=function(Z,Co,Re){
    AB = 2*Z[1,1] + Z[2,1] + Z[1,2] + Co*Z[2,2]
    Ab = 2*Z[3,1] + Z[2,1] + Z[3,2] + Re*Z[2,2]
    aB = 2*Z[1,3] + Z[1,2] + Z[2,3] + Re*Z[2,2]
    ab = 2*Z[3,3] + Z[3,2] + Z[2,3] + Co*Z[2,2]
    props = data.frame(AB,Ab,aB,ab)
    haps=(data.matrix(props)/2)/sum(Z)
    rownames(haps)="Hap"
    return(haps)}
  HHH=c()
  # Loop
  for(i in 1:n){
    H=HapProb(Z,COUP,REPU)
    # cat(cbind(H,COUP,REPU),'\n')
    H2=matrix(H,2,2); dia=H2[1,1]*H2[2,2];
    off=H2[2,1]*H2[1,2]; tt=dia+off;
    oC = COUP; oR = REPU #for while loop
```

```
    COUP = dia/tt; REPU = off/tt
    HHH=rbind(HHH,rbind(cbind(H,COUP,REPU)))
    diff= abs(COUP-oC)+abs(REPU-oR) #for while loop
  }
  rownames(HHH)=1:n
  EM=round(tail(HHH,1)[1:4],4)
  names(EM)=colnames(HHH)[1:4]
  return(EM)
}

# TESTING
EM(Allele1,Allele2)
```

```
##     AB     Ab     aB     ab
## 0.0361 0.4384 0.4435 0.0820
```

To compute the LD matrix, it is necessary to calculate the LD between each pair of *PHASED* markers. The LD is obtained as follows:

$D_{ab} = p_{AB} - p_A p_B$

Then, D has to be scaled int D' (pronounced as 'D prime')

$D' = D/D_{max}$

Where D_{max} is:

- $\min\{p_A p_B, (1-p_A)(1-p_B)\}$ when $D < 0$
- $\max\{p_A(1-p_B), p_B(1-p_A)\}$ when $D > 0$

And alternatively represented as r^2, calculated as:

$r = D/sqrtp_A(1-p_A) \times p_B(1-p_B)$

To arrange this terms in R, I coded the following function:

```
EM_LD = function(A,B){
phase=EM(A,B)
if(det(matrix(phase,2,2))<0) phase=phase[c(2,1,4,3)]
X = matrix(phase,2,2)
D=X[1,1]-sum(X[,1])*sum(X[1,])
if(D<0){
  Dp=min((sum(X[,1])*sum(X[1,])),sum(X[,2])*sum(X[2,]))
  }else{
  Dp=min((sum(X[,1])*sum(X[2,])),sum(X[,2])*sum(X[1,]))
}
r=sqrt(prod(X))
r2=r**2
ld = as.vector(data.frame(D,Dp,r,r2))
return(ld)
}

example_LD = EM_LD(Allele1,Allele2)
example_LD
```

```
##           D        Dp          r           r2
## 1 0.1914702 0.2275702 0.02399068 0.0005755529
```

Now, the final step is simplified to the computation of pairwise D' across a given region. It can be done as follows:

```
LDmat = function(gen,type.of.LD){
  n = ncol(gen)
  LD = matrix(NA,n,n)
  colnames(LD) = rownames(LD) = paste('SNP',1:n,sep='')
  for(i in 1:n){
    for(j in 1:n){
      if(j>i){
        LD[i,j] = as.numeric(EM_LD( gen[,i], gen[,j])[type.of.LD])
        LD[j,i] = as.numeric(EM_LD( gen[,i], gen[,j])[type.of.LD])
      }
    }
  }
  LD = round(LD,3)
  return(LD)
}
# type.of.LD can be: 1 = D, 2 = D', 3 = r, 4 = r2
example_LD_matrix = LDmat(gen[,1:5], type.of.LD = 2)
example_LD_matrix
```

```
##       SNP1  SNP2  SNP3  SNP4  SNP5
## SNP1    NA 0.228 0.227 0.232 0.238
## SNP2 0.228    NA 0.155 0.245 0.239
## SNP3 0.227 0.155    NA 0.245 0.239
## SNP4 0.232 0.245 0.245    NA 0.112
## SNP5 0.238 0.239 0.239 0.112    NA
```