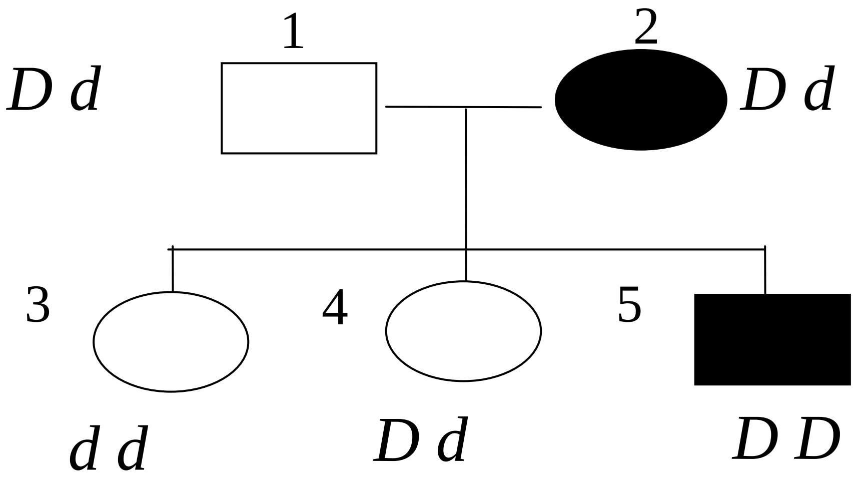
Homework 1

1. Some short proofs of the recombination process.
   a. Prove the Mether's formular: p(R(J) = 0.5*p(X(J)>0).
   b. Explain what is a first-division segregation (FDS) and a second-division segregation (SDS). Why they are interesting?
   c. Prove the inductive formula for the second-division segregation: $S_{k+1}$=4$F_k$+2$S_k$. (note that in class we proved that a similar formula for the FDS, mimic the technique we used there!)
   d. Discuss the relationship and difference between Haldane's mapping function p(R(d))=(1/2)(1-exp(-2d)) and the mapping function for the SDS frequency s(d)=(2/3)(1-exp(-3d)).

2. Genetic Linage Analysis
   a. What does "Hardy-Weinberg equilibrium" mean in terms of inheritance and mating outcome?
   b. Consider the following pedigree on slide 12, lecture 8:



   What is the conditional probability of the genotype of the individual 2, i.e., p($G_2$=Dd|$G_1$=Dd,$G_3$=dd,$G_4$=Dd,$G_5$=DD)? Use the allele frequencies given in the previous slides.

   c. What LODs of linkage are additive across independent pedigrees?
   d. Explain for *n* QTLs there are *$2^n$* distinct genotypes for BC and $3^n$ distinct genotypes for IC?

3. A common strategy to infer the haplotype of multiple SNPs is to use a parsimony criteria to evaluate the results and devise algorithm that greedily optimize this criteria over possible solutions. Now suppose we have the genotype of 3 SNPs of individual 1: $G_1$={1/0, 1/0,1/0}, and we also know that the genotype of an individual 2 is: $G_2$={1/1, 0/0,1/1}. Can you guess what is a "parsimonious" solution of individual 1's haplotypes? What if $G_2$={0/0, 0/0,0/0}?
   a. Why haplotype is advantageous over single SNPs for linkage analysis?

b. Why is the computational and biological meaning for finding “blocks” in long sequences of SNPs?

c. Implement the EM algorithm for Haplotype inference and test it on the data to be posted.

To open the zipped file, use command “tar zxvf  hap.data.tgz”. In the data directory, you will see a .genos and a .haplos file. The format is as follows:

Genotype:

```
1603 0  2 -1 1  1 0 0 1  1 1 1 2 1  0
2103 0 -1  2 1 -1 0 0 1 -1 1 1 0 1 -1
```

…

The first column is the sample id, starting from the second column, you see the genotype of 14 SNPs in each sample. “0” denote the genotype (0,0), “1” denote the genotype (1,1), “2” denote a genotype (0,1), i.e., a heterogeneous genotype, and finally, “-1” denote an artifact in the measurement of that SNP, and can be ignored.

Haplotype:

```
1603 0 0 1 -1 1 1 0 0 1  1 1 1 1 1 0
1603 1 0 0 -1 1 1 0 0 1  1 1 1 0 1 0
2103 0 0 0  1 1 1 0 0 1 -1 1 1 0 1 -1
2103 1 0 -1 0 1 0 0 0 1  1 1 1 0 1 0
```

…

The first column is the sample id; the second column is the index of paternal and maternal allele; starting from the third column, you see the haplotype of 14 SNPs in each of the two alleles of every sample. Note that “-1” denote an artifact in the measurement of that SNP, and can be ignored.

You need to infer the haplotype of the SNPs of all the samples using your program. Report both the haplotype and the overall error ratio (i.e., the number of wrongly inferred loci over all heterozygous loci).