Problems with measuring LD

All measures of LD depend on the allele frequencies of the pair of SNPs measured.

In different parts of the genome SNPs have different allele frequencies

'measure ?'

Dependence of the alleles frequencies of LD measures is a real bottleneck!

Doing better, more robust with → D:

$$P_{11}, P_{12}, P_{21}, P_{22} \geq 0$$

$$P_{11} \geq 0 \Rightarrow D = P_{11} - p_1 q_1 \Rightarrow D \geq -p_1 q_1$$

$$P_{22} \geq 0 \Rightarrow D \geq -p_2 q_2$$

} MIN

$$D_{MIN} = MAX \{-p_1 q_1, -p_2 q_2\}$$

$$P_{12} \geq 0 \Rightarrow D \leq p_1 q_2$$

$$P_{21} \geq 0 \Rightarrow D \leq p_2 q_1$$

} MAX

$$D_{MAX} = MIN \{p_1 q_2, p_2 q_1\}$$

We introduce : D'

$$D' = \begin{cases} \frac{D}{D_{MAX}} & \text{if } D > 0 \\ \frac{D}{D_{MIN}} & \text{if } D < 0 \end{cases}$$

Another measure of LD: r^2

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2}$$

correlation coefficient measure

The two measures r^2 and D' are used together: they have desirable complementary type of information they measure

---

Both D' and r^2 are having values in $[0, 1]$

r^2 has the property that the intermediary values

$0 < \rho < 1$

can be interpreted !!

HAPMAP PROJECT

Based on this "interpretability"

D' does not have the interpretability property

Three axioms / desiderata from the decades of deep research in Human Genetics:

Ax 0. Independence of allele frequencies

AX1. The curse of the pairwise: we want LD measures to be extendible to multi markers/SNP in a conservative way

AX3. Interpretability of intermediate values

BIG OPEN PB: To find an LD measure satisfying the above three ~~desiderata~~ axioms simultaneously.

Interpretability:

(1) If two SNPs are in equal frequency and create

between them only 2 haplotypes, there $r^2 = 1$, $D' = 1$

A C
A C $r^2 = 1$
T G $D' = 1$
T G

---

② If the two SNPs are in unequal frequency, and <u>at least</u> three haplotypes are present then:

Exactly three haplotypes are present, then $D' = 1$, $r^2 < 1$

A C
A G $r^2 < 1$
T G $D' = 1$
T G

---

③ when all four haplotypes are present then $D' < 1$, $r^2 < 1$

AC
TC $r^2 < 1$
AG
TG $D' < 1$

Population

- Sample of individuals: a random sample from the population we are studying
- Computing LD for the sample we think of it as an approximation of LD for the population.

Testing for LD

is a statistical process

LD is an empirical pattern

statistical significance of LD

Are the two sites significantly associated? I.e. whether the LD detected is statistically significant.

The Null Hypothesis:

H_0: they are in Linkage Equilibrium

: $P_{ij} = P_i q_j$

We use the counts of each of

the alleles occuring together

There are four haplotypes at the two sites/SNPs.

The Null hypo. states that the two sites are independent.

The natural statistical test is a test for independence of 2x2 contingency table with elements n_{ij}, where:

n_{ij} = the number of loops carrying allele i at position 1 and allele j at position 2

$$\sum_{j=1}^{2} \sum_{i=1}^{2} n_{ij} = n = \text{total sample size}$$

The most precise test is Fisher's exact test of independence.

However, Fisher's exact test is complicated to calculate and when n is large, we may calculate instead

$$\chi = \sum \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

e_{ij} = the estimated expected of n_{ij} under the Null Hypothesis

i.e. assumption of no association

If e_{ij} are sufficiently large (>5) then χ is approximately χ^2 distributed with one degree of freedom.

An example: NM blood group in a British population

A sample of 1000 people with genotype frequencies as follows:

Gene NM — Two alleles: N, M

genotypes { MM 298 "blood group M"
MN 489 " —11— MN"
NN 213 " —11— N"

sum = 1000

Estimate allele frequencies:

M: $\hat{p}$, N: $\hat{q}$

$\hat{p} = \frac{1085}{2000} = 0.54$ | $1085 = 2 \times 298 + 489$

$\hat{q} = \frac{915}{2000} = 0.45$

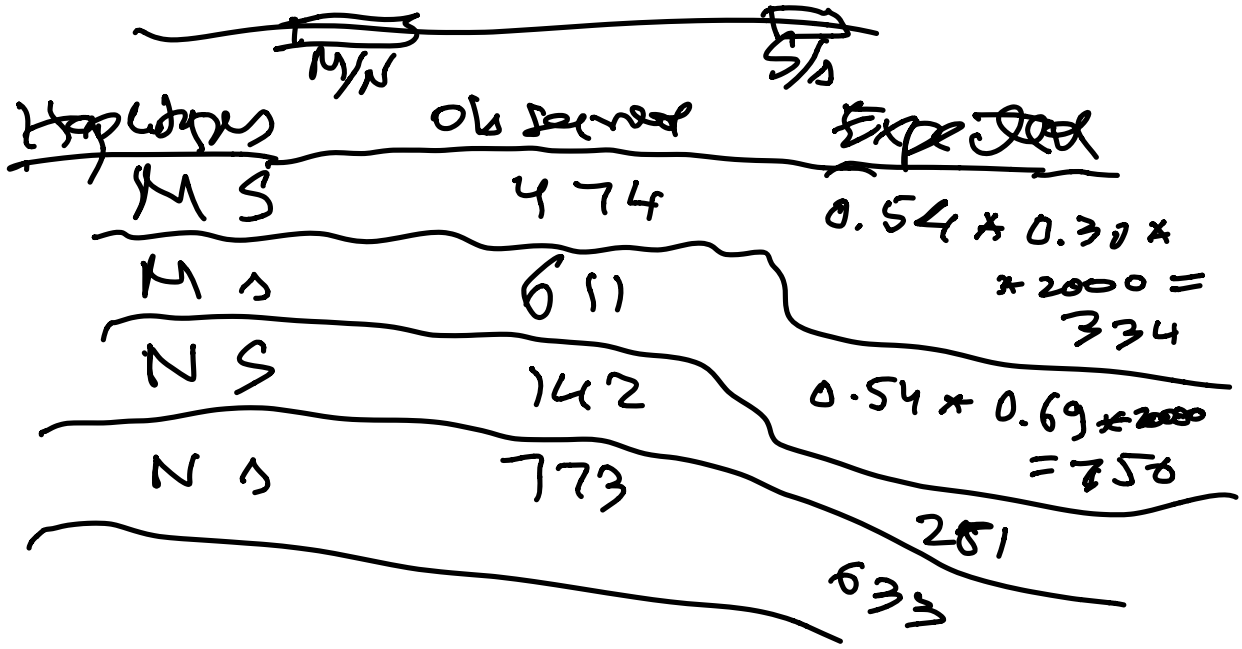
Gene S,s

SS	99
Ss	418
ss	483

S: $\hat{q}_1 = 0.30$

s: $\hat{q}_2 = 0.69$

2 Genes : haplotypes



$X^2 = 184.7$ with

df = 1

$P_1 P_2 P_3 P_4$: 4 − 3 = 1

$-1, -1$ $\quad -1$

$p_1 + q_1 = 1$ $\quad p_2 + q_2 = 1$

P-value assoc with $X^2 = 184.7$

very small P-value < 0.0001

Reject Null Hypothesis

<u>there is LD.</u>

How much LD:

To quantify the amount of LD, we use D'.

$P_{11}, P_{12}, P_{21}, P_{22}$

MS: $\hat{P}_{11} = \frac{474}{2000}$

MS: $\hat{P}_{12} = \frac{611}{2000}$

NS: $\hat{P}_{21} = \frac{142}{2000}$

NS: $\hat{P}_{22} = \frac{773}{2000}$

$$\hat{D} = \hat{P}_{11}\hat{P}_{22} - \hat{P}_{12}\hat{P}_{21} = 0.07$$

$$\hat{D}_{MAX} = \min \begin{Bmatrix} \hat{P}_1\hat{q}_2 = 0.38 \\ \hat{P}_2\hat{q}_1 = 0.14 \end{Bmatrix} = 0.14$$

$$\hat{D}' = \frac{\hat{D}}{\hat{D}_{MAX}} = \frac{0.07}{0.14} = 50\% \text{ of}$$

if theoretical max