Linkage Disequilibrium, Relative Risk, and Tag SNP Selection

1. Problem 1

Consider loci 1 and 2. Suppose the allele for locus 1 is either A or a, and the allele for locus 2 is either B or b. Suppose A and B are associated such that:

$$\begin{aligned} p_{AB} &= p_A p_B + D_1 \\ p_{Ab} &= p_A p_b - D_2 \\ p_{aB} &= p_a p_B - D_3 \\ p_{ab} &= p_a p_b + D_4 \end{aligned}$$

Show that $D_1 = D_2 = D_3 = D_4$.

Solution 1

We begin by noting the marginal allele frequencies:

$$\begin{aligned} p_A &= p_{AB} + p_{Ab}, \quad p_a = p_{aB} + p_{ab} \\ p_B &= p_{AB} + p_{aB}, \quad p_b = p_{Ab} + p_{ab} \end{aligned}$$

We are given that $D_1 = p_{AB} - p_A p_B$. Now, consider:

$$D_2 = p_A p_b - p_{Ab} = p_A(1 - p_B) - p_{Ab} = p_A - p_A p_B - p_{Ab}$$

Substituting $p_A = p_{AB} + p_{Ab}$, we get:

$$D_2 = (p_{AB} + p_{Ab}) - p_A p_B - p_{Ab} = p_{AB} - p_A p_B = D_1$$

Now evaluate D_3:

$$D_3 = p_a p_B - p_{aB} = (1 - p_A)p_B - p_{aB} = p_B - p_A p_B - p_{aB}$$

Again using $p_B = p_{AB} + p_{aB}$, we find:

$$D_3 = (p_{AB} + p_{aB}) - p_A p_B - p_{aB} = p_{AB} - p_A p_B = D_1$$

Finally, consider D_4:

$$D_4 = p_a p_b - p_{ab} = (1 - p_A)(1 - p_B) - p_{ab}$$

Expanding:

$$D_4 = 1 - p_A - p_B + p_A p_B - p_{ab}$$

Now substitute $p_{ab} = 1 - p_{AB} - p_{Ab} - p_{aB}$, so:

$$D_4 = 1 - p_A - p_B + p_A p_B - (1 - p_{AB} - p_{Ab} - p_{aB}) = -p_A - p_B + p_A p_B + p_{AB} + p_{Ab} + p_{aB}$$

Now simplify using $p_A = p_{AB} + p_{Ab}$, $p_B = p_{AB} + p_{aB}$:

$$D_4 = -(p_{AB} + p_{Ab}) - (p_{AB} + p_{aB}) + p_A p_B + p_{AB} + p_{Ab} + p_{aB} = -p_{AB} + p_A p_B = D_1$$

Thus, we conclude:

$$\boxed{D_1 = D_2 = D_3 = D_4}$$

2. Problem 2

Consider the following SNP genotype distributions:

SNP1: Case group: 40 A, 160 C Control group: 20 A, 180 C

SNP2: Case group: 20 A, 180 C Control group: 10 A, 190 C

Compute the relative risk (RR) and odds ratio (OR) for both SNPs. Which SNP has a higher risk?

section*Solution 2

For SNP1:

$$RR = \frac{40/(40+160)}{20/(20+180)} = \frac{0.2}{0.1} = 2.0$$

$$OR = \frac{40/160}{20/180} = \frac{0.25}{0.1111} \approx 2.25$$

For SNP2:

$$RR = \frac{20/(20+180)}{10/(10+190)} = \frac{0.1}{0.05} = 2.0$$

$$OR = \frac{20/180}{10/190} = \frac{0.1111}{0.0526} \approx 2.11$$

Conclusion: Both SNPs have the same relative risk ($RR = 2.0$), but SNP1 has a higher odds ratio ($OR = 2.25 > 2.11$). Therefore, SNP1 shows a stronger association with disease risk.

3. Problem 3

Given the following six haplotypes over 11 SNPs:

$$H_1 = (0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)$$
$$H_2 = (0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0)$$
$$H_3 = (0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1)$$
$$H_4 = (1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1)$$
$$H_5 = (1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1)$$
$$H_6 = (1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1)$$

Determine the minimal subset of tag SNPs that distinguish all haplotypes.

Solution 3

We divide the 11 SNPs into three blocks:

Block 1: SNPs 1–3 Block 2: SNPs 4–7 Block 3: SNPs 8–11

For Block 1: If $r_1 = 0$, we observe tag patterns 010, 011, and 000 $\rightarrow$ covering three haplotypes. If $r_1 = 1$, the tag 101 distinguishes the remaining three haplotypes.

For Block 2: If $r_1 = 0$, tag 0101 covers the three $r_1 = 0$ haplotypes. If $r_1 = 1$, tag 1010 distinguishes the others.

For Block 3: If $r_1 = 0$, tag 0111 identifies three haplotypes. If $r_1 = 1$, tag 1101 identifies the rest.

All three blocks successfully partition the six haplotypes. Therefore, selecting one informative SNP from each block is sufficient. The minimal set is:

$$\boxed{\{r_1, r_4, r_8\}}$$

These SNPs form a minimal subset of tag SNPs that can distinguish all six haplotypes.