Association Mapping

Suppose a gene $A$ affects susceptibility to human disease.

Observations are available for disease status (affected or unaffected) or measurements are available for a risk factor (e.g. blood pressure).

Observations are also available for a genetic marker $M$.

If the marker frequencies are different for affected and unaffected people, or if the risk factor means are different in different marker classes, then loci $M$ and $A$ are associated. The expectation is that they are also close together on a chromosome. Association depends on the linkage disequilibrium between alleles at the marker and disease loci.
Genotype Frequencies

Suppose A has alleles A, a and M has alleles M, m. If there is random mating (Hardy-Weinberg) in the population, then the nine genotypic classes have frequencies

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM</td>
<td>$P_{AM}^2$</td>
<td>$2P_{AM}P_{aM}$</td>
<td>$P_{aM}^2$</td>
<td>$p_M^2$</td>
</tr>
<tr>
<td>Mm</td>
<td>$2P_{AM}P_{Am}$</td>
<td>$2P_{AM}P_{am} + 2P_{Am}P_{aM}$</td>
<td>$2P_{aM}P_{am}$</td>
<td>$2p_M p_m$</td>
</tr>
<tr>
<td>mm</td>
<td>$P_{Am}^2$</td>
<td>$2P_{Am}P_{am}$</td>
<td>$P_{am}^2$</td>
<td>$p_m^2$</td>
</tr>
<tr>
<td>Total</td>
<td>$p_A^2$</td>
<td>$2p_A p_a$</td>
<td>$p_a^2$</td>
<td>1</td>
</tr>
</tbody>
</table>
Marker allele frequencies

If people with genotypes $AA$, $Aa$, $aa$ have probabilities $G_{AA}$, $G_{Aa}$, $G_{aa}$ of being affected with the disease, then the marker allele frequencies are

$$\Pr(M|\text{Aff.}) = p_M + D_{AM}[p_A(G_{AA} - G_{Aa}) + p_a(G_{Aa} - G_{aa})]/\bar{G}$$
$$\Pr(m|\text{Aff.}) = p_m - D_{AM}[p_A(G_{AA} - G_{Aa}) + p_a(G_{Aa} - G_{aa})]/\bar{G}$$

$$\Pr(M|\text{Unaff.}) = p_M - D_{AM}[p_A(G_{AA} - G_{Aa}) + p_a(G_{Aa} - G_{aa})]/(1 - \bar{G})$$
$$\Pr(m|\text{Unaff.}) = p_m + D_{AM}[p_A(G_{AA} - G_{Aa}) + p_a(G_{Aa} - G_{aa})]/(1 - \bar{G})$$

where $\bar{G}$ is the chance that a random person has the disease, $\bar{G} = (p_A^2 G_{AA} + 2p_A p_a G_{Aa} + p_a^2 G_{aa})$. 
Case-control Test

There are many ways to conduct the case-control test. One is to compare marker allele frequencies between cases and controls. If the sample sizes were the same, the data could be set out as

<table>
<thead>
<tr>
<th></th>
<th>Affected</th>
<th>Unaffected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>x</td>
<td>y</td>
<td>x + y</td>
</tr>
<tr>
<td>m</td>
<td>n − x</td>
<td>n − y</td>
<td>2n − (x + Y)</td>
</tr>
<tr>
<td>Total</td>
<td>n</td>
<td>n</td>
<td>2n</td>
</tr>
</tbody>
</table>

A contingency table chi-square test statistic is

\[
X^2 = \frac{2n(x - y)^2}{(x + y)[n - (x + y)]}
\]

The size of this statistic depends on the size of the sample value of \( D_{AM} \). When there is no association, \( X^2 \sim \chi^2_{(1)} \).
Power of Case-control Test

Linkage disequilibrium can be expressed as

\[ r^2 = \frac{D_{AM}^2}{p_A p_a p_M p_m} \]

Disease susceptibility can be described with

\[ V_A = 2p_A p_a [p_A (G_{AA} - G_{Aa}) + p_a (G_{Aa} - G_{aa})]^2 \]

For the chi-square goodness of fit test statistic \( X^2 \):

\[ \mathcal{E}(X^2) \approx n \frac{r^2 V_A}{\bar{G}(1 - \bar{G})} \]

Dividing by \( n \) gives the non-centrality parameter of a Chi-square distribution with 1 degree of freedom. The non-centrality parameter is an increasing function of \( r^2 \), and thus the power increases as \( r \) increases.
**Trend Test**

The case-control test does not behave correctly if there are departures from Hardy-Weinber equilibrium. A better test uses the genotype counts

<table>
<thead>
<tr>
<th>Marker Genotype</th>
<th>MM</th>
<th>Mm</th>
<th>mm</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case counts</td>
<td>$r_0$</td>
<td>$r_1$</td>
<td>$r_2$</td>
<td>$R$</td>
</tr>
<tr>
<td>Control counts</td>
<td>$s_0$</td>
<td>$s_1$</td>
<td>$s_2$</td>
<td>$S$</td>
</tr>
<tr>
<td>Total counts</td>
<td>$n_0$</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

The linear trend test statistic is

$$X^2 = \frac{N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{SR[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]}$$

and, when there is no association, $X^2 \sim \chi^2_{(1)}$. 
Smyth et al. (Nature Genetics 38:617-619, 2006) presented data for association with Type I diabetes. Perform both the allelic case-control test and the linear trend test for these two datasets. Also test for Hardy-Weinberg equilibrium in cases, controls and cases plus controls for each dataset. Would you combine data sets 1 and 2 to increase the sample size?

<table>
<thead>
<tr>
<th></th>
<th>Set 1</th>
<th>Set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MM</td>
<td>Mm</td>
</tr>
<tr>
<td>Case</td>
<td>810</td>
<td>873</td>
</tr>
<tr>
<td>Control</td>
<td>513</td>
<td>720</td>
</tr>
</tbody>
</table>