Post-hoc estimation of genotype relative risk (GRR) and 95% confidence interval (CI)

To assess the increase in disease risk conferred by inheritance of 0, 1, or 2 copies of the putative risk allele A with population frequency q, the GRR and 95% CI were calculated for each highlighted SNP and haplotype based on observed T:NT results. Let B denote the set of all alleles other than A at the locus of interest. The probability of having the disease, conditioned on the number of A alleles and averaged over all other loci in the population was f0, f1, f2 for an individual with 0, 1, or 2 copies of A, respectively. The population frequency of the genotypes BB, AB, AA were denoted as q0, q1, q2 which, by Hardy-Weinberg Equilibrium, were p^2, 2pq, q^2 respectively, where p=1-q was the frequency of having an allele from the B set. It was assumed that the association study randomly sampled affected probands from the population without any biases related to the risk level of A itself. In a large (infinite) population of N unrelated individuals, the number of individuals having a total of i A alleles (i = 0, 1, or 2) would be qi*N, and since the probability of such an individual being affected was fi, the total number of affecteds with i A alleles would be Ni = fi*qi*N. Thus, when sampling n (finite) affected probands at random, i A allele individuals would be selected with probabilities pi given by the following proportions of Ni within the affected population N0+N1+N2, pi = Ni/(N0+N1+N2) = (fi*qi)/(f0q0+f1q1+f2q2). The results are subject to sampling variations that produce variations in the outcome. It is the impact of these sampling variations that must be quantified in order to properly describe the relation between genotype risk and association results. The absolute probabilities were divided out to obtain the result in terms of genotype relative risk factors, gi = pi/p0, (hence g0=1) which described the risk related to the total dose of A. Therefore, the
The overall goal was to express the result of the association study of allele A as a 
Z score relative to a standard normal distribution, so that higher Z values correspond to 
more significant association with the disease. The distribution depended on the 
parameters of the study and allele, namely the number of probands, n, the A genotype 
risks f_0, f_1, f_2, the frequency of A, q, and perhaps other parameters unique to the study 
design, which were collectively denoted by Q. Thus, the resulting Z implicitly depended 
on all these factors, \( Z = Z[g_1, g_2, n, q, Q] \). Given the observed \( Z_{\text{obs}} \) from an actual 
study, which would have definite values of n, q, and Q, but unknown risk levels, 
genotypic risks and/or confidence intervals can then be estimated. Specifically, the most 
likely values of g_1, g_2 would minimize the magnitude of Z, (i.e. given the least variation 
from the expected mean), \( \{g_1, g_2\} = \text{Minimizers of } |Z[g_1, g_2, n, q, Q]| \), or, in 
particular, values for which \( Z[g_1, g_2, n, q, Q] = 0 \).

The values compatible with a given confidence level, c, would be solutions of 
\( Z[g_1, g_2, n, q, Q] = Z[c] \) where \( Z[c] \) is the Z value corresponding to the desired 
confidence level. For example, for c = 85% confidence, \( Z[c] = 1.036 \), corresponding to a 
variation due to population sampling that is about one standard deviation above the 
expected association score. Or, for a symmetrical confidence interval covering the 
fraction c of possible results from all proband samplings, the boundary values of g_1, g_2 at 
the upper and lower bounds \( Z[.5+c/2], -Z[.5+c/2] \) could be similarly calculated.

In order to achieve this general form, the current genetic study was described by 
an additive test statistic, \( S = s_1 + s_2 + ... + s_n \) where the \( s_j \) is the contribution from the 
j^\text{th} proband, and the \( s_j \) are independent and identically distributed. Resampling the n
probands many times from the population, and thus redoing the study many times, results in an expected score \( S \) and a variance due to sampling variation for a given study design. The additive test statistic allows an approximation of a normal distribution of \( S \) under resampling, which can be further described in terms of the mean and variance of the individual proband scoring function \( s: E[S] = n E[s], \, \text{Var}[S] = n \text{Var}[s] \). This, in turn, allows formation of a normalized \( Z \) score equivalent to the original scoring statistic, \( Z = (S - n E[s]) / \sqrt{n \text{Var}[s]} \). The quantities \( E[s] \) and \( \text{Var}[s] \) depend on allele frequency \( q \) and the risks \( f_0, f_1, f_2 \)---or, more precisely, just the genotype relative risks \( g_1 = f_1 / f_0, g_2 = f_2 / f_0 \)---and thus the study outcome is described by a normal statistic parameterized by these quantities, as well as the study size \( n \), and any other study design parameters which might exist, \( Q \), as indicated above \( Z = Z[g_1, g_2, n, q, Q] \).

The scoring of a given allele \( A \) is commonly described as counting transmissions or non-transmissions of \( A \) from heterozygous, i.e. \( AB \) genotype, parents to the proband, as a means to assess the extent to which \( A \) confers risk. \( BB \) and \( AA \) parents do not contribute to the scoring, as their transmissions provide no information. Let \( T \) denote the total count of \( A \) transmissions observed in the \( n \) trios, and \( NT \) the total number of non-transmissions. Then, \( T > NT \) is evidence of risk from the \( A \) allele, and the significance of this bias relative to the null hypothesis of Mendelian transmission (\( A \) and \( B \) equally likely to be sent from an \( AB \) parent) can be judged exactly using the binomial distribution or commonly approximated using a chi square test assuming a normal distribution.

The standard test statistics was modified to the additive form described above. The statistic \( T - NT \) itself was used, which was broken down into a \( t-nt \) contribution from each proband trio, \( S = T - NT = (t-nt)_1 + \ldots + (t-nt)_n \). Thus the per-proband statistic
was \( s = t - nt \), the difference between total transmissions and non-transmission of A within
the given trio. Note that for a given trio, \( t = 0, 1, \) or 2 and \( nt = 0, 1, \) or 2, so \( t - nt \)
necessarily has an integer value between -2 and +2. To systematically compute the mean
an variance for a given trio, the probability distribution of \( t - nt \) was computed directly. We
consider each outcome of \( t - nt \) in detail, and add the probabilities of all ways in which
each value can be achieved:

\( t - nt = 2 \): this requires drawing an AA proband, and both parents must draw a B
allele as their unsent allele.

\[ P[t - nt = 2] = p_2 \cdot (1-q)^2 = p_2 \cdot q_0 \]

\( t - nt = -2 \): this requires drawing an BB proband, and both parents must draw an A
allele as their unsent allele.

\[ P[t - nt = -2] = p_0 \cdot q^2 = p_0 \cdot q_2 \]

\( t - nt = 1 \): this can be achieved by drawing an AA proband, and drawing one B and
one A for the unsent parental alleles, OR by drawing an AB proband, and having both
parents draw B alleles as their unsent alleles.

\[ P[t - nt = 1] = p_2 \cdot 2 \cdot q \cdot (1-q) + p_1 \cdot (1-q)^2 = p_2 \cdot q_1 + p_1 \cdot q_0 \]

\( t - nt = -1 \): this can be achieved by drawing an BB proband, and drawing one B and
one A for the unsent parental alleles, OR by drawing an AB proband, and having both
parents draw A alleles as their unsent alleles.

\[ P[t - nt = -1] = p_0 \cdot 2 \cdot q \cdot (1-q) + p_1 \cdot q^2 = p_0 \cdot q_1 + p_1 \cdot q_2 \]

\( t - nt = 0 \): this can be achieved with an AA proband, and both parents draw an A as
unsent; or an AB proband, and one parent draws an A and the other draws a B, or a BB
proband, and both parents draw a B. However, this term drops out of the mean and
variance due to the zero value and therefore the formula is not needed.

From this, one can readily compute the expectation and variance of the per-proband score:


$$= 2*p^2*q0 + p^2*q1 + p1*q0 - (p0*q1 + p1*q2) - 2*p0*q2$$

and


$$= 4*p^2*q0 + p^2*q1 + p1*q0 + (p0*q1 + p1*q2) + 4*p0*q2.$$  

These moments in turn provide the variance

$$Var[s] = E[s^2] - E[s]^2$$