White Paper 23-01

Estimating Genotype-Specific Incidence for One or Several Loci

Authors:

Mike Macpherson
Brian Naughton
Andro Hsu
Joanna Mountain

Created: September 5, 2007
Last Edited: November 18, 2007
Summary:

We wish to estimate the incidence for a given trait based on an individual's genotype at one or more SNPs associated with the trait and any available phenotypic data for that individual. Here we describe the methods used to estimate these incidences, first for the case of a single SNP, then for the case of multiple SNPs.

Single Locus Calculation

In practice, we provide an estimate of trait incidence conditional on the individual's genotype and available phenotypic information. The calculations below, however, hold whether incidence or prevalence estimates are computed. We henceforth use the generic term "risk" instead of incidence to indicate this fact.

We assume a binary trait D and a single associated locus at which there are three possible genotypes G_1, G_2, and G_3, ordered so that G_1 is the lower-risk homozygote and G_3 is the higher-risk homozygote. The individual for whom we wish to estimate the incidence has genotype $G_m : m \in \{1, 2, 3\}$. The quantity we wish to compute is $\Pr(D|G_m)$, or the probability that the individual is affected given their genotype.

The unconditional risk for the trait, denoted $\Pr(D)$, is assumed known for a subpopulation of which the individual is a member. For instance, this might be an estimate of Type 2 diabetes prevalence for Asian subjects between the ages of 20 and 40. We further assume that estimates of the three genotype frequencies $\Pr(G_i)$ are available for the same subpopulation. Lastly, we assume that estimates of the three genotype-specific odds ratios OR_1, OR_2, and OR_3 are available, where $OR_1 = 1$ by definition.

Under these assumptions, we may compute the $\Pr(D|G_i)$ by solving the following system of equations:

$$\Pr(D) = \Pr(D|G_1)\Pr(G_1) + \Pr(D|G_2)\Pr(G_2) + \Pr(D|G_3)\Pr(G_3) \quad (1)$$

$$OR_2 = \frac{\Pr(D|G_2)/(1 - \Pr(D|G_2))}{\Pr(D|G_1)/(1 - \Pr(D|G_1))} \quad (2)$$

$$OR_3 = \frac{\Pr(D|G_3)/(1 - \Pr(D|G_3))}{\Pr(D|G_1)/(1 - \Pr(D|G_1))} \quad (3)$$

Equation 1 follows from basic probability theory, and Equations 2 and 3 follow from the definition of an odds ratio. The estimated genotype-specific risk at the single locus is then $\Pr(D|G_m)$.

To convey the genotype-specific risk relative to the average risk in the population, we compute the quantity $OR^*_m = odds(D|G_m)/odds(D)$, where, for brevity, we have introduced the function $odds(X) = \Pr(X)/(1 - \Pr(X))$. The inverse odds function will be used later on, and is $odds^{-1}(X) = \Pr(X)/(1 + \Pr(X))$. The superscript asterisk on OR^*_m is a convention we adopt to distinguish an odds ratio computed relative to the average odds, rather than relative to the odds of lower-risk homozygote, which is the standard definition.

Complications

Odds Ratio Estimates: In the association study literature, odds ratios are commonly estimated via logistic regression assuming an additive model. This means that the logarithm of the odds ratio is assumed to relate linearly to the number of copies of the higher-risk allele (*cf.* Jewell, 2003). When the odds ratio is estimated in this way, the higher-risk homozygote's estimated log odds ratio is, by definition, exactly twice that of the heterozygote's log odds ratio, from which it follows that the higher-risk homozygote's odds ratio is the square of the heterozygote's odds ratio. Thus the odds ratio reported when the additive model is assumed is that associated with one copy of the higher-risk allele, which could be called an *allele-specific* odds ratio. In such cases we set the allele-specific odds ratio equal to the heterozygous genotype's odds ratio, OR_2, and set the higher-risk homozygote's odds ratio $OR_3 = (OR_2)^2$. In cases for which genotype-specific odds ratios are estimated separately, we use those estimates directly.

Prevalence v. Incidence: We typically report genotype-specific risk in terms of incidence. However, the formulas in this white paper apply equally well to prevalence and incidence data. Specifically, the value $\Pr(D)$ may represent either a prevalence or an incidence datum.

Mismatched Datasets: In practice, we rely on association studies that are most often based on individuals of European descent for our odds ratio estimates. We obtain genotype frequency estimates primarily from HapMap, which provides one population of European descent, one of Yoruban descent, and one of Asian (Chinese & Japanese) descent. We obtain prevalence/incidence estimates from a variety of sources, and these estimates often pertain to individuals of European descent alone.

Ideally, we would have perfectly-matched odds ratio, genotype frequency, and prevalence estimates, meaning that, if the prevalence estimate pertains to Asian subjects between 20 and 40, we would also have genotype frequency estimates and odds ratio estimates for Asians between 20 and 40. It is most often the case that we do not have such matched estimates. At this writing, we assume that the overall HapMap genotype frequency estimates apply across all age ranges, *i.e.* that the genotype frequencies within any age range are identical to the overall frequencies. We also assume that odds ratios apply across age ranges. We record whether an odds ratio estimate derives from a European, African, or Asian population, and do the same for prevalence estimates. By 'derives' we mean that we consider the population for which a given odds ratio or prevalence estimate, and (subjectively) decide whether that population is near enough to one of the three

HapMap populations for us to use the estimate. We only report genotype-specific risk estimates when the underlying estimates match at the population level. For example, when an odds ratio estimate exists for an Asian population, a genotype frequency estimate exists for an Asian population, and a prevalence estimate exists for an Asian population, we will report the genotype-specific risk estimates, even if the prevalence estimate is age-range specific. At this writing, we have not studied how this policy affects the accuracy of the estimates.

Higher Moments: At this writing, we do not provide estimates of our certainty in the reported genotype-specific risk point estimate. It is standard practice to provide confidence intervals for odds ratios reported in the literature. We have sample sizes for the HapMap frequency data, and often have sample sizes for prevalence estimates, so it should be relatively straightforward to produce risk confidence intervals, and informative to our users.

Multiple Locus Calculation

For many traits, associations exist between the trait and several SNPs. We wish to combine the genotype-specific risk estimates from multiple loci into a single, genotype-specific *composite* risk estimate for an individual's genotype.

Here we assume that the composite odds ratio, OR_C, is given by the product of the individual's odds ratios at each locus. This is very similar to what is done in multiple logistic regression under an additive model, where each copy of a risk allele at a given locus is assumed to add one unit of the log odds ratio specific to that locus to the overall log odds ratio (*cf.* Jewell, 2003; Risch, 1990). It is not exactly the same because a multiple regression would be performed simultaneously on all the data, where we use individual odds ratios obtained from separate studies, which may, for instance, differ widely in sample size. We assume that there are K loci of interest, and denote the kth odds ratio of the ith genotype $OR_{i,k}$. Similarly, the recentered odds ratios are denoted $OR^*_{i,k}$. We denote the genotypes at the kth locus $G_{1,k}, G_{2,k}, G_{3,k}$, and denote the individual's genotype at the kth locus $G_{m_k,k}$. Thus

$$OR^*_C = \prod_{k=1}^{K} OR^*_{m_k,k} \tag{4}$$

The quantity OR^*_C has the interpretation

$$OR_C^* = odds(D|\cap_{k=1}^K G_{m_k,k})/odds(D), \tag{5}$$

where $odds(D|\cap_{k=1}^K G_{m_k,k})$ is the odds of the individual's multilocus genotype.

In computing the product OR_C^*, we implicitly assume that the point where $\log(odds(D))$ intersects the respective logistic regression line for each locus is the same point in each of the individual regression calculations, as would be true if a multiple logistic regression had been performed. Then

$$\Pr(D|\cap_{k=1}^K G_{m_k,k}) = odds^{-1}\left[OR_C^*\ odds(D)\right]. \tag{6}$$

Complications

Higher Moments: As in the single-locus case, we do not calculate confidence intervals for the estimates we provide. Provided confidence intervals for individual risk at a single locus, it is straightforward to compute them for the composite risk.

References

Jewell, Nicholas P. 2003. *Statistics for Epidemiology*. Chapman & Hall/CRC.

Risch, N. 1990. Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet*, **46**(2), 222–228.