GENE 210: Genomics and Personalized Medicine

Applying GWAS to Personal Genomes

Konrad Karczewski

A major challenge of personal genomics lies in the application of known disease-associated variants in a clinical context. When a variant is discovered by a GWAS (as in the primer on "How to do a GWAS"), we might know that the "G" allele has a high odds ratio for some disease, but if we are "GG" at this locus, what does that mean for our risk for our disease? Here, we'll discuss another way to do GWAS using genotypes, as well as other statistics, including odds ratio* (OR*), likelihood ratio, and percent variance explained.

Genotype-based GWAS:

Let's go back to a simple bitter taste GWAS example. Before, we had grouped phenotypes (bitter taste yes vs. no) by allele, counting 2 for homozygotes (i.e. a CC individual that tasted bitter would contribute 2 C alleles in the yes column) and 1 each for heterozygotes (a CG with yes would add 1 C to yes, and 1 G to yes). Now, if we reconsider this problem in a diploid context, we can simply add up each genotype for each phenotype independently (i.e. a CG with yes adds 1 CG to yes). For rs713598, the class distribution is as follows:

rs713598	**Yes**	**No**
CC	1	6
CG	13	3
GG	10	1

Now, the issue here is that with genotypes, we have a 2x3 table, which is often less-powered to detect differences and sometimes more difficult to interpret (e.g. what if all CG were "Yes", but CC and GG were all "No": we would see a significant result, but it would be difficult to interpret). One way to get around these problems is to group 2 of the genotypes, which we can interpret in a dominant or recessive model. For example, if we assume C is a recessive allele, we can compare CC individuals to those with CG and GG. Then, we have a table like:

rs713598	**Yes**	**No**
CC	1	6
CG and GG	23	4

Now, we can do the association test in the usual way, using a chi-squared or a Fisher's exact test. In this case, we would find a significant association ($\chi^2 = 13.4591$, $p = 0.0002438$; Fisher's $p = 0.0009592$) with an odds ratio of 29.31.

Different models:

In this case, we assumed the C allele was recessive (and by proxy, the G allele is dominant). The dominant/recessive models are the simplest and easiest to interpret, but the choice of which allele is dominant is not always known beforehand. One option is to run the association tests in both directions (once with one allele as dominant, and then with the other as dominant), which can be effective to discover these associations, but since there are twice as many tests being done, our significance threshold must be adjusted accordingly.

Alternatively, we could analyze these data based on alleles (as we did in the previous primer) and assume an additive or multiplicative model. In an additive model, the presence of each allele would add to the odds ratio, where having 'Aa' might have an odds ratio of x, and 'AA' would have an odds ratio of 2x. In a multiplicative model, 'Aa' might have an odds ratio of x, and 'AA' would have an odds ratio of x^2. We won't calculate these models for this example, but we could use any number of these methods which may provide slightly different results, depending on the assumptions we make.

Odds Ratio*:

Now that we have a significant result, we would like to apply these results to a personal genome to adjust our probability of having a trait or disease. One method involves the use of OR*, which is similar in spirit to an odds ratio, but now comparing the odds of having the disease with a certain genotype to the background odds of having the disease. Specifically,

$$\text{Odds Ratio*} = \frac{\text{Odds(A)}}{\text{Odds(Background)}}$$

We get the odds of the background using the probability of having the trait given the background population (0.75 for Europeans, but these vary across populations and so, the OR* will be different for each population).

$$\text{Odds(European)} = \frac{0.75}{0.25} = 3$$

P(Yes | CC) = 1/(6+1) = 0.142 P(Yes | CG/GG) = 23/(23+4) = 0.851
Odds(CC) = 0.167 Odds(CG/GG) = 5.75

So, for Europeans, the odds ratio* for CG and GG is: Odds(CG/GG)/Odds(Background) = 1.916 and the odds ratio* for CC is: Odds(CC)/Odds(Background) = 0.0556.

Then, we can apply this to our genotype, by multiplying the pre-test odds by all the relevant OR*s (if we are considering across multiple loci), which will give us the post-test

odds for our personal genome. In this case, we have only 1 SNP (let's take rs713598:CG as an example), and so the post-test odds is `3*1.916 = 5.75.`

We can convert this back into a probability by inverting the odds function, or:

$$P(\text{Trait} \mid A) = \frac{\text{Odds}(A)}{1 + \text{Odds}(A)}$$

In this case, we have `P(Trait | CG) = 5.75/6.75 = 0.8514`, so our probability of being a bitter taster was adjusted from 75% to 85.14%.

Likelihood Ratio:

Another method we can use involves likelihood ratios, which have their roots in evidence-based medicine. This method flips the dependence of the expression above, noting the probability of having a genotype given the disease (i.e. instead of p(Trait|A), we consider p(A|Trait)). Typical lab tests in a clinic would use these methods to adjust a probability of having a disease (e.g. if we observed a high level of PSA, we would adjust the likelihood of having prostate cancer, along with other factors such as age and race). Because of the details in the probability calculations, this model assumes the independence of tests, not the independence of risk factors (i.e. genotypes, race, etc). Here, the question answered is not "what is the probability the patient has this disease?" but "what is the probability of seeing these results if the patient has the disease, compared to if the patient is healthy?" (or "what is the likelihood of disease given these results?").

In the case of GWAS, the likelihood ratio is just as easy (if not easier) to calculate as the odds ratio. The likelihood ratio is defined as:

$$\text{Likelihood Ratio*} = \frac{P(A \mid \text{Trait})}{P(A \mid \text{No trait})}$$

Now, to calculate `P(A|Trait)` we flip the calculation to correspond with the dependence. That is,

$$P(A \mid \text{Trait}) = \frac{P(\text{Trait} \,\&\, A)}{P(\text{Trait})}$$

The challenge comes in *post hoc* application of published association studies, which may not report the allele frequencies of cases and controls in the study. However, since we have that information for the bitter taste trait, we can calculate the likelihood ratios easily. Assuming the same model as before, we find:

`P(CG/GG | Yes) = 23/(23+1) = 0.9583` `P(CG/GG | No) = 4/(6+4) = 0.4`

Then, the likelihood ratio is simply `LR(CG/GG) = 0.958/0.4 = 2.395`. Now, we can convert this to a post-test probability as before, where we multiply the likelihood ratio by the pre-test odds to get the post-test odds. So, for Europeans, we have a post-test odds of `3*2.395 = 7.187.` Then, we convert back to probabilities to get our post-test

probability of 7.187/8.187 = 0.877. So using this likelihood ratio method, our probability of being a bitter taster changes from 75% to 87.7%.

Percent Variance Explained:

Now that we know the effect of a given variant on a trait, we'll want to know how much of the variation in the trait can be explained by known genetics. For this, we'll need to calculate how much variation exists in the phenotype of interest, which we can quantify using a "sum of squares" statistic. These methods are generally applicable to quantitative traits, but we can use it for a binary trait if we choose a value of 1 for yes, and 0 for no. The total sum of squares is given as:

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

where y_i is the actual label of each individual i (0 and 1 in a binary trait) and $\bar{y}$ is defined as the mean of y (in this case, the overall frequency of the trait). We'll sum over all both columns in the table, putting in the actual label (yes or no; 1 or 0) into y_i. In the case of rs713598 for bitter taste (with the grouping of the distributions with C as recessive, as above), we have:

$$\bar{y} = 24/34 = 0.7058$$

Now,

$$SS_{tot} = \sum_{24}(1-0.7058)^2 + \sum_{10}(0-0.7058)^2$$

$$SS_{tot} = 24(0.2941)^2 + 10(-0.7058)^2$$

$$SS_{tot} = 7.0588$$

Then, we calculate the remaining sum of squares after we take into account the genetic prediction. This is given by:

$$SS_{\text{err}} = \sum_i (y_i - f_i)^2$$

where f_i is our prediction for individual i assuming we predict ALL individuals with the genotype as having the trait. This statistic can be thought of how many times we would be wrong if we just guessed based on genetics. We sum over all 4 cells of the table:

$$SS_{err} = \sum_{23}(1-1)^2 + \sum_{4}(0-1)^2 + \sum_{1}(1-0)^2 + \sum_{6}(0-0)^2$$

$$SS_{err} = 4(-1)^2 + 1(1)^2$$

$$SS_{err} = 5$$

The percent variance explained is then calculated by:

$$R^2 \equiv 1 - \frac{SS_{\text{err}}}{SS_{\text{tot}}}.$$

In this case, we have `R2 = 1 - 5/7.0588 = 0.2916`. This may seem low, but notice that there is not a ton of variation in the trait (70.5% of the class were bitter tasters, so if you just guessed "yes" for everyone, you'd be right 70.5% of the time). Thus, to explain more of the trait by genetics, you'd already need to do better than that. In our case, predicting if someone is not a bitter taster is somewhat difficult (40% of "no" individuals are CG or GG).