GENE 210: Genomics and Personalized Medicine

How to do a GWAS

Konrad Karczewski

Genome-wide association studies are commonly used techniques to discover genetic factors that influence diseases. In these approaches, hundreds of thousands or millions of genetic loci are probed in a set of cases and controls of a disease. For each genetic locus, the association between genotype and phenotype is measured by a chi-squared (or Fisher-exact) test. A correction factor is then applied to the p-value of each association to discover "genome-wide" significant variants. We can also then use these data to determine the odds ratio of the variant.

Chi-squared (χ^2) test:

We will use the class results for earwax type as an example of how to do a GWAS. The results for the first SNP tested are as follows:

	Wet	Dry	Total
A	21	0	21
G	27	18	45
Total	48	18	66

Since there are a total of 48 wet and 18 dry (66 total), we have 72.7% wet and 27.2% dry. With 21 A alleles and 45 G alleles, we have 31.8% A and 68.1% G. If the genotype were independent of the phenotype, we would expect probability independences to apply (i.e. `p(AB) = p(A)*p(B)`). For instance, the expected proportion of wet & A alleles is:

```
p(Wet & A) = p(Wet)*p(A) = 72.7%*31.8% = 23.1%.
```

Since we have 66 individuals, we would expect 15.27 in the wet & A cell. The expected proportions (and expected counts in parentheses) are as follows:

	Wet	Dry	Total
A	23.1% (15.27)	8.6% (5.72)	21
G	49.5% (32.72)	18.5% (12.27)	45
Total	48	18	66

The chi-squared is defined by $\chi^2 = \Sigma(O-E)^2/E$ (O = observed, first table; E = expected, 2nd table). We sum over all cells in the table, taking the observed minus expected squared, divided by expected.

$$\chi^2 = (21-15.27)^2/15.27 + (0-5.72)^2/5.72 + (27-32.72)^2/32.72 + (18-12.27)^2/12.27 = 11.54$$

If we look up the p-value for this χ^2 value (with 1 degree of freedom) in a distribution table, we find that p = 0.000677. Alternatively, we could enter the values from the table into the `chisq.test()` command in R to find the same result. We can enter the 4 entries from our table as follows:

```
chisq.test(matrix(c(top-left, top-right, bottom-left, bottom-right), nrow=2), correct=FALSE)
```

In this case, we would use:

```
chisq.test(matrix(c(21, 0, 27, 18), nrow=2), correct=FALSE)

        Pearson's Chi-squared test

data:  matrix(c(21, 0, 27, 18), nrow = 2)
X-squared = 11.55, df = 1, p-value = 0.0006775
```

Fisher's exact test:

We typically use a χ^2 test as it is highly reliable and easy to calculate for large sample sizes. However, the χ^2 test is only an approximation of the significance of the results, since the sampling of these data are not exactly equal to a χ^2 distribution. Therefore, we would ideally use a Fisher's exact test, especially when sample sizes are small. Fisher's exact test is too complicated to calculate by hand for this exercise, but we could use a program such as R or MATLAB to do the test, or a website such as http://www.graphpad.com/quickcalcs/contingency1.cfm.

For the problem sets, as with χ^2, you may use the following command in R (do not add correct=FALSE):

```
fisher.test(matrix(c(21, 0, 27, 18), nrow=2))

        Fisher's Exact Test for Count Data

data:  matrix(c(21, 0, 27, 18), nrow = 2)
p-value = 0.0002895
```

Multiple hypothesis correction:

A p-value of 0.05 means that we would expect a result at least as significant as our result by chance 5% of the time. Thus, if we did 500,000 tests, we would expect about 25,000 of them to have a p-value of less than 0.05. Because of this, when we do a GWAS, we must correct our significance. One simple method is applying a Bonferroni correction, in which we adjust our threshold by the number of tests performed. So, for running a 500,000 SNP GWAS, our 0.05 threshold would be 0.05/500,000 = 1e-07. It should be noted that Bonferroni correction is a fairly conservative (stringent) correction method. Other methods, such as Benjamini or FDR (False Discovery Rate) corrections, as well as permutation tests, may be used to set a different threshold.

In this exercise, we're cheating a little bit, because we narrowed your search to only 5 options (SNPs we already knew were associated with one of these traits). Nevertheless, we'll correct the p-value using a Bonferroni correction of 5, and so, our new threshold is 0.01.

Results:

The results for earwax are as follows. *Try to calculate the χ^2 statistic for one of these by hand. Then, using statistical software such as R or MATLAB, run the χ^2 and Fisher's exact test for another one.*

earwax (4988235):

	Wet	Dry
A	21	0
G	27	18

earwax (7495174):

	Wet	Dry
A	41	7
G	7	11

earwax (713598):

	Wet	Dry
C	23	7
G	25	11

earwax (17822931):

	Wet	Dry
C	38	4
T	10	14

earwax (4481887):

	Wet	Dry
A	24	3
G	24	15

Our association study tested 5 SNPs for association with the earwax trait. For earwax type, if we run a χ^2 and Fisher's tests for all 5 SNPs, we find the results below.

SNP	$\chi 2$ (p-value)	Fisher's test (p)
rs4988235	11.55 (0.0006775)	0.0002895
rs7495174	12.038 (0.0005212)	0.0003776
rs713598	0.1432 (0.7051)	0.5861
rs17822931	15.966 (6.449e-05)	3.43E-05
rs4481887	0.4289 (0.5125)	0.3940757

In this case, we actually got three SNPs that showed significant association ($p<0.01$) with earwax type, but the strongest association is rs17822931 (4th row). This is likely due to a confounding race association, as the dry earwax trait is most often found among Asians, who also often have brown eyes and are lactose intolerant. A typical GWAS might control for this by only considering individuals from a single population.

Odds ratios:

Once we find a SNP that is significantly associated with the trait of interest, we can compute the odds ratio for that SNP. This is found using a simple application of Bayes' rule, where the probability of having a trait given a genotype `P(Trait|A)` is the probability of having the trait and the genotype, divided by the probability of having A:

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

Next, we can convert this probability to an odds, by dividing it by 1-itself (Think "Vegas": 75% chance means a 3:1 odds of it happening). So:

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

Then, we find the odds ratio by dividing the odds of both genotypes:

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

Thus, for our most significant SNP (rs17822931), we have:

`P(Wet | C) = 38/(38+4) = 0.904` `P(Wet | T) = 10/(10+14) = 0.416`
`Odds(C) = 9.5` `Odds(T) = 0.714`

Thus, the odds ratio is: `Odds(C)/Odds(T) = 13.3.`

This does not exactly mean that people with the C allele are 13.3 times more likely to have the trait. We'll discuss the exact application of these onto a personal genome (including different risk models such as likelihood ratios), but for now, just consider that the C allele is significantly associated with and increases your likelihood of having wet earwax.

Full results:

The full results of the class association study are shown below. *Try to figure out which SNPs are associated with which trait.*

```
eyes (4988235):
      Brown      Blue/Green
A      9         12
G      29         16
eyes (7495174):
      Brown      Blue/Green
A      20         28
G      18         0
eyes (713598):
      Brown      Blue/Green
C      18         12
G      20         16
eyes (17822931):
      Brown      Blue/Green
C      17         25
T      21         3
eyes (4481887):
      Brown      Blue/Green
A      14         13
G      24         15
```

```
bitter (4988235):
      Yes       No
A      16         5
G      30         15
bitter (7495174):
      Yes       No
A      32         16
G      14         4
bitter (713598):
      Yes       No
C      15         15
G      31         5
bitter (17822931):
      Yes       No
C      31         11
T      15         9
bitter (4481887):
      Yes       No
A      22         5
G      24         15
```

```
asparagus (4988235):
      Yes       No
A      12         9
G      28         17
asparagus (7495174):
      Yes       No
A      31         17
G      9          9
asparagus (713598):
      Yes       No
C      17         13
G      23         13
asparagus (17822931):
      Yes       No
C      26         16
T      14         10
asparagus (4481887):
      Yes       No
A      19         8
G      21         18
```

```
lactose (4988235):
      Yes       No
A      1          20
G      9          36
lactose (7495174):
      Yes       No
A      9          39
G      1          17
lactose (713598):
      Yes       No
C      4          26
G      6          30
lactose (17822931):
      Yes       No
C      5          37
T      5          19
lactose (4481887):
      Yes       No
A      2          25
G      8          31
```