Video Tutorial 10.1: Understanding Manhattan plots and genome-wide association studies

The most common method to identify candidate genes that contribute to a complex trait in humans is to use a genome-wide association study – abbreviated to GWAS - as described in Chapter 10 in the book. Let's look at an example to see how data from a genome-wide association study are displayed and interpreted. Eosinophilic esophagitis (EoE) is a chronic inflammatory disorder associated with allergic hypersensitivity to food. It is diagnosed by the presence of eosinophils—a type of white blood cell—in the esophagus. Affected individuals often have difficulty swallowing, and in the most severe cases, the esophagus can become so narrow that food gets stuck.

It clearly has a genetic component but it is not inherited in any simple Mendelian fashion. It is a complex trait, with an estimated frequency of about one person is 200, so it is not exceptionally rare. But the symptoms vary, as expected for many complex traits. Most affected individuals also have other inflammatory conditions or allergic sensitivities to other stimuli in addition to food, but not all affected people do.

A genome-wide association study to find candidate genes associated with EoE was published by Kottyan et al in 2014, and will be our example for examining the data from a GWAS. They used information from 736 affected individuals, and 9246 unaffected control individuals of European ancestry. A total of about 1.5 million SNPs were tested throughout the genome.

Those data are shown in this figure known as a Manhattan plot since it resembles a city skyline. This type of plot is used for nearly all GWAS, with only minor variations in the way the data are presented.

Manhattan plots used in GWAS look at the frequency or association of each haplotype as defined by these SNPs in affected individuals compared to those without the condition. This type of analysis can be used to find loci on the chromosome that are likely to have genes affecting a particular condition.

The X-axis on the figure shows haplotypes from each region of the genome that was tested, organized by chromosome, shown in different colored blocks. Although the data from each chromosome look like a block with an irregular surface, this area actually is composed of thousands of dots, each dot being one haplotype. The apparent vertical lines arise from haplotypes that cover the same genetic location. The haplotypes are plotted as they occur from left to right on the chromosome, so the first line represents the left-most haplotypes on at the left end of chromosome 1 and the last line represents the haplotypes at the right end of X chromosome.

For each haplotype, the relative association with EoE is plotted as the height of the line. Thus, the Y axis represents the relative frequency or association of a haplotype in that region in affected individuals compared to control individuals. This number on the Y axis is not always computed in precisely the same way, but the data are interpreted the same way—the higher the peak, the stronger the association with the trait.

In the language of statistics, the experiment is testing the hypothesis that a particular locus is <u>not</u> associated with an increased occurrence of the disease, or that the haplotype or locus is segregating independently from EoE. The Y axis is the P value or the probability that the association was observed by

chance. The plots are often done as a negative log scale. In other words, this haplotype is very unlikely not to be associated with the increased occurrence of EoE.

The haplotypes that have a statistically significant association with EoE are shown as dots above the horizontal line. Haplotypes in 4 different regions of the genome are found to be associated with an increased occurrence or risk for EoE, as detected for the population in this study, on chromosomes 2, 5, 8, and 15. These regions have a P value that is less than 10^{-8}. The best candidate gene that maps to this region is also shown.

There are six other regions for which variants are suggestive of an association, with a P value of 10^{-7}. These other associations are found on chromosomes 1, 5 (distinct from the previous one), 10, two distinct regions on chromosome 11, and one on chromosome 21. Genes that are located in some of these regions and that are known or suspected from other analysis to be strong candidates are shown above the region.

What does this tell us about the genetics and biology of EoE? This study re-enforces what we know about pleiotropy, the recognition that the same gene can contribute to different phenotypes, as discussed in Chapter 8. Remember that many individuals with EoE also exhibit other inflammatory conditions or allergic sensitivity. Some of these regions and candidate genes have also been identified as being associated with those conditions, suggesting that therapies used for these conditions may be helpful for EoE as well.

Most significantly, it tells us that several different genes can contribute to the disease, and these genes are found throughout the genome. Two individuals affected with EoE could have different underlying molecular and genetic causes for the disease. Recall that the nature of a complex trait is that the same phenotype can arise from different genotypes. This difference in the underlying molecular basis for the disease could have a dramatic effect on the recommended therapy or course of the disease.

For example, in animal models and in a pilot study in humans, an antibody that blocks the activity of the TSLP protein, shown here, appears to be an effective therapy. Because this therapy is specific to this particular protein, it is not expected to be as effective if a person with EoE has one of the other haplotypes but not this one.

This study also identifies the gene CAPN14, which encodes a protein known as calpain 14, as a very strong candidate for contributing specifically to many cases of EoE. This paper followed up on that genes and found that that CAPN14 is highly expressed in the esophagus, and its expression is upregulated among individuals with EoE, so this could be also be a specific target for therapy.

As we said, EoE is extremely variable in its manifestations. The GWAS suggests that there are at least 4 and possibly as many as 10 different loci in the genome that can cause EoE. Some of the alleles at these loci could be dominant in their effects, others could be recessive. The loci could interact with each other in complicated ways or have complicated interactions with different environmental factors. Until now, EoE is the best label that we have for this condition because affected people have eosinophils in their esophagus. But with so many different underlying genetic factors, perhaps “EoE” is really not a specific enough description of the phenotype.