Chapter 10

Gene Hunting

How do we find genes that contribute to Alzheimer’s disease, to schizophrenia, to cognitive ability, or to any number of traits? That is the topic of this chapter. In the past, this topic would have been divided into two broad sections—linkage analysis and association. Modern genotyping technology, however, has driven linkage analysis into obsolescence. The interested reader may consult any number of old genetics texts to learn about linkage analysis. Here, we focus on association.

This chapter first discusses the logic of association designs and points out the two major strategies—the population-based design and the family-based design. Then we discuss the two major purposes for association studies—the candidate gene study and the genome-wide association study or GWAS.

10.1 Logic of association

The logic as well as the statistical analysis of association studies depends on the type of phenotype. Almost all phenotypes fall into one of two major classes—dichotomous and quantitative. Dichotomous phenotypes are those in which “either you have it or you don’t have it.” Medical disorders and most forms of psychopathology are classic examples. Quantitative phenotypes are continuous traits like height. Everyone has the trait but not the same amount of it. We will discuss each type in turn.

10.1.1 Dichotomous phenotypes

There are several ways to gene hunt for medical disorders. One is to select a random sample from the general population. The participants are interviewed, tested, and medical histories and diagnoses catalogued. No general population studies have been initiated merely to hunt for genes. Instead, most have been running for many years, and in one of the follow-ups, DNA is gathered and tested. A classic example is the Framingham Heart Study, a longitudinal study...
that was initiated in 1948, five years before Watson and Crick unraveled the structure of DNA!

More commonly, gene hunting uses what epidemiologists call the *case-control design*. A series of “cases,” i.e., people with the disorder, are identified and then an appropriate group of people without the disorder are selected as controls. There are two major types of controls used in genetics. The first selects controls from the general population. This gives the *population-based association study*. The second is a genetic relative without the disorder. This gives the *family-based association study*. In the past there was considerable debate on the merits and disadvantages of the two strategies, mostly over the issue of population stratification (see Section X.X). Today, it is accepted that statistical methods can control adequately for this effect.

Let’s imagine a population-based approach for a single SNP with an A allele and a G allele. Suppose that the disorder is bipolar manic-depression, a form of psychopathology in which people have periods of excited overactivity often accompanied by feelings of grandiosity (the manic phase) and other periods of deep sadness feelings of worthlessness (depressive phase). Assume that there are 1,000 bipolars and an equal number of controls.

In the association design, one would genotype all 2,000 individuals on the SNP and create a two by three contingency table such as the one shown in Table 10.1 for fictitious data. If the SNP is not associated with bipolar disorder, then the number of bipolars with genotype, say, AA should equal the number of controls for that genotype. (If total sample sizes were unequal, then the proportion of bipolar should equal the proportion of controls with that genotype.) Similarly, the number (proportion) of bipolars with AG should equal that of controls, and the number (proportion) of bipolars with GG should be the same as controls.

On the other hand, if the SNP is associated with bipolar disorder, then the observed frequencies should differ from these chance expectations. Statistical tests can be performed to help us judge how well the observed data agree with the null hypothesis of chance association.

In the present case, it is highly unlikely we would observe the data in Table 10.1 if the association were purely due to chance. Inspection of the data reveals that the G allele is associated with risk for bipolar disorder.

Does this mean that the G allele is a *causal* factor in bipolar disorder? No. That could be the case, but we could not conclude that from these data. Why?

---

Table 10.1: Example of a population-based association test for a SNP.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Group</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bipolar</td>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>AG</td>
<td>GG</td>
<td>Total</td>
</tr>
<tr>
<td>Bipolar</td>
<td>339</td>
<td>510</td>
<td>151</td>
<td>1,000</td>
</tr>
<tr>
<td>Control</td>
<td>485</td>
<td>419</td>
<td>96</td>
<td>1,000</td>
</tr>
<tr>
<td>Total</td>
<td>824</td>
<td>929</td>
<td>247</td>
<td>2,000</td>
</tr>
</tbody>
</table>

---

1. The most common test would be a $\chi^2$ goodness-of-fit statistic.
2. $\chi^2 = 47.03$, df = 2, $p < 6E-11$. 

---
CHAPTER 10. GENE HUNTING  10.1. LOGIC OF ASSOCIATION

Table 10.2: Schematic for a discordant sib-pair design. Bp = bipolar sib; Un = unaffected sib.

<table>
<thead>
<tr>
<th>Family</th>
<th>Genotype:</th>
<th>A alleles:</th>
<th>Difference: (3) - (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith</td>
<td>AG</td>
<td>AA</td>
<td>1</td>
</tr>
<tr>
<td>Jones</td>
<td>GG</td>
<td>AG</td>
<td>0</td>
</tr>
<tr>
<td>Brown</td>
<td>AG</td>
<td>AG</td>
<td>1</td>
</tr>
</tbody>
</table>

Remember linkage disequilibrium from Section X.X and haplotype blocks from Section X.X? The G nucleotide will be in linkage disequilibrium with several other polymorphisms close to it. Hence, we do not know if it is the G allele or one (or more) of the nearby polymorphisms that contributes to the disorder. In short, a positive finding in an association study suggests a correlation. Other data are required to establish causality.

In the family-based association, the controls would be one or more unaffected genetic relatives of the bipolar proband. (A proband is the person with a trait like a disorder through whom a pedigree is ascertained.) There is a wide variety of family based designs, and it is not possible to examine them all here. Let’s take a simple case where the control is an unaffected sibling. The data layout would resemble that in Table 10.2.

The family is the unit of analysis so there would be one row per family. The genotypes of the affected and unaffected sibs are recorded and then we tabulate the number of, say, A alleles in each sib. (Here, the selection of the A versus G allele is arbitrary.) The key data column is the difference in the number of A alleles between the bipolar and the unaffected sib. If there is no association between the SNP and bipolar disorder, then the average of this column should be within sampling error of 0. If statistics suggest that it is significantly different from 0, then there is evidence of an association. If the mean is positive, then the A allele is associated with the disorder; if negative, then the G allele is the risk allele. As is the case with the population-based association study, a significant result denotes a correlation, not necessarily causality.

10.1.1.1 Disclaimer

In reality, statistical analysis can be more complicated. For example, one may need to adjust prediction for demographic factors such as sex and age or for other phenotypes. In addition, some phenotypes may have “intermediate” categories. For example, a study of schizophrenia may also include people with a diagnosis of schizotypal personality disorder. In such circumstances, advanced statistical techniques such as logistic regression can be used for the analysis.
10.1. LOGIC OF ASSOCIATION

Figure 10.1: Mean height in boys and girls as a function of a SNP in the HMGA2 gene.

10.1.2 Quantitative phenotypes

A gene that contributes to a quantitative phenotype is termed a quantitative trait locus or QTL. As an example, let’s examine a real case. It will have more lessons for us that just the statistics.

HMGA2 is a gene that goes by the horrid name “high-mobility group AT-hook 2.” (Who said molecular biologists could not be inventive in naming things?) It codes a protein that binds to the DNA, especially in enhancer regions rich in AT nucleotides. (Hence, the “AT-hook: part of the name). Mutations in the gene are associated with various cancers, but the important part for our purposes here is that mice with mutations in these types of proteins are very small and called “pygmies” (Zhou et al., 1995; Benson and Chada, 1994). Recently, genetic association studies have implemented this locus in individual differences in height (Weedon et al., 2007). Let’s look at one part of the Weedon et al. (2007) data on height, that of seven year old boys and girls.

The relevant genotype is from a SNP (technically, rs1042725) within the HMGA2 gene that has two alleles—T and C. A plot of the mean heights of the boys and girls as a function of genotype on the SNP is given in Figure 10.1. The statistics used to assess draw a straight line of best fit (i.e., a regression line) through the three means for boys and another through the three means for girls and assess the slopes of these lines. If the slope is significantly different from 0, then there is evidence that the SNP correlates with height. Because the sample size was large (c. 6,000) the results in Figure 10.1 are highly significant ($p = 1 \times 10^{-6}$).

Statistical significance assesses the likelihood that there is, in fact, no correlation and the observed data occurred simply by chance. A second statistical measure is effect size. It measures the magnitude of the relationship. For a SNP QTL, an appropriate measure of effect size deals with the predicted change in mean height by substituting one allele for another. That is, if I took a genotype
and substituted a C for a T, what is the predicted difference in mean height? The answer is about half a centimeter (0.2 inches), about the length of a housefly.

A second measure of effect size is the percent of phenotypic variance explained by genetic variance. That is abstract and we will learn more about it later, but for the moment think of it in these terms: If I look around at all of the individual differences in height among seven year olds, what percent of those differences can I explain by this gene? The answer is about 0.5%, one half of one percent.

You are probably thinking to yourself something along the lines of “that’s not very much,” and you would be correct. This is the emerging pattern for not only QTLs but for many of the genes that contribute to disease. That is, we can detect genes, but the typical effect of a single gene is usually very small. Most geneticists suspects that for complex phenotypes this means that there are many, many genes that contribute to individual differences.

You may also be thinking, “so what?” What is the big deal about finding a gene that does not do very much? If you have such an impression, then congratulations. This is indeed an important question. We will have more to say about this later in the text.

The Weedon et al. (2007) study was a population-based association study. It is also possible to use family-based designs to search for QTLs. Once again, there are many ways to do this, all of which address the same question: Within a family, are relatives with more C alleles taller than those with fewer C alleles? Once, again, let us examine a sib-pair study.

Let us continue with the HMGA2 gene and height and, for simplicity, ignore the sex difference in height. From the genotypes on the sib-pairs we would calculate the difference in the number of C alleles for each pair. For example, in the Smith family, if sib 1 has genotype TC and sib 2 is CC, then sib 1 and one C allele and sib 2 has two. Hence, the difference is 1 - 2 = -1. We would also calculate the difference in height. Again, if Smith sib 1 is 130.19 cm and sib 2 is 132.50 cm, then the difference is 130.19 - 132.50 = -2.31. Continuing with this logic gives us the schematic presented in Table 10.3.

We then calculate the correlation between the difference in C alleles and the difference in height. If the correlation is significant and positive then we would conclude that the gene is correlated with height.

As in association studies for disorders, a significant QTL finding indicates
10.2. TWO PURPOSE FOR ASSOCIATION STUDIES. GENE HUNTING

a correlation. It does not necessarily imply causality. There are several other polymorphisms in and around the HMGA2 gene that were not genotypes in the Weedon et al study. Other types of data and research designs must be used to determine causal variants from those that happen to be in linkage disequilibrium with those variants.

10.2 Two purpose for association studies

Today there are two major uses for the association strategy—the candidate gene study and the genome-wide association study or GWAS. A candidate gene approach examines a known gene to see if it is associated with a phenotype. GWAS, on the other hand, is a “shotgun” approach to association. There are no prior hypotheses (usually). Instead, one examines polymorphisms throughout the whole genome.

10.2.1 Candidate gene study

The candidate gene study (CGS) starts with one or more genes that have strong theoretical relevance for a phenotype and/or have shown significant association with the phenotype in past research. A CGS may involve a single gene or several hundreds of genes. Unfortunately, there is no rigorous definition of “strong theoretical relevance.” In the early days of DNA polymorphisms in the 1980s and 1990s, virtually any polymorphism involving a neuronal transmission was a candidate for any psychiatric disorder. Today, matters are more disciplined, and the list is usually compiled from known biochemical and signaling pathways often involving bioinformatic databases. Let us examine two candidate gene studies on smoking and nicotine dependence.

10.2.1.1 Smoking and nicotine dependence

Contrary to popular opinion, smoking is not a “mild” addiction. Instead, “the evidence clearly identifies nicotine as a powerful drug of addiction, comparable to heroin, cocaine and alcohol” (Stolerman and Jarvis, 1995). Twin data show that nicotine dependence, like all behavioral phenotypes, has a moderate heritability. In addition, there is something special about nicotine use and dependence that makes it attractive to the geneticist—we have brain receptors that respond specifically to nicotine!\(^3\)

Our first study, Saccone et al. (2007), was a case-control design of nicotine dependence, involving a total sample size close to 2,000. The candidate genes were selected from gene families for nicotinic receptors, the enzymes involved in the metabolism of nicotine and in two neurotransmitter systems previously associated with various types of chemical dependencies—dopamine and \(\gamma\)-aminobutyric acid (GABA). The candidates were subdivided into a “A” list of 55 loci.

\(^3\)These are called nicotinic receptors and endogenously they respond to the neurotransmitter acetylcholine.
SNPs within these gene regions were extensively genotyped. The remaining 296 genes in the “B” list were also genotyped but were given less weight in the data analysis. The top “hits” in the study were SNPs in two different genes, both of which coded for proteins that are subunits of the nicotinic receptor.  

The second candidate gene study was that reported by Caporaso et al. (2009). The sample size was approximately 4,500. The interesting feature of this study was that, with minor exception, it used the same list of candidate genes as the Saccone study. This permits us to examine the replicability of the findings. Only one gene reached statistical significance—monoamine oxidase A, an enzyme that breaks down certain neurotransmitters.

If we let matters as is, it looks as if the two studies disagree. There are indeed areas of disagreement. The top hit in the Saccone study did not give any indication of significance in the larger Caporaso study. But further examination suggests important areas of replication. One of the important genes reported by Saccone, CHRNA5 which codes for a subunit of the nicotinic receptor, just missed being significant in Caporaso. Also, the MAOA gene, significant in Caporaso but not Saccone, was among a list of “top signals” that gave important evidence for association but did not meet the criterion for significance. It is possible that both of these genes may be associated with smoking and nicotine dependence but the stringent criteria for significance makes it difficult to detect them in all circumstances. Variations in defining the phenotype may also contribute to the differences.

10.2.2 Genome-wide association study (GWAS)

The genome-wide association study (GWAS) examines a very large number of common polymorphisms scattered throughout the human genome and tests for an association locus by locus. Although it is possible to use any type of polymorphism in GWAS, the usual modus operandi is to use DNA arrays and SNPs (see Sections X.X and X.X). A second common strategy is to assay for CNVs (see section X.X). Unlike candidate gene studies, a GWAS is agnostic about polymorphisms.

A GWAS is very expensive and time consuming. As a result, most of the phenotypes studies are medical diseases, including psychopathology, and phenotypes correlated with those diseases (e.g., blood pressure, cholesterol levels). GWAS results have been reported for IQ and some personality traits, but the data here are often gathered in the context of medical syndromes.

The number of loci assessed in GWAS had exploded—from tens of thousand in about 2005 to over a million as of this writing. The number of studies is growing exponentially. Statistics from the National Human Genome Research Institute (NHGRI) website (http://www.genome.gov/gwastudies/) show one or two published GWAS reports per quarter in 2005 growing to 1350 per quarter

4The relevant genes were CHRNA3 and CHRNA5. The nicotinic receptor contains five proteins or subunits. At least 17 different genes code for the various subunits that can make up one of the many forms of a nicotinic receptor.

5This study also contained a genome-wide association component.
10.2. TWO PURPOSE FOR ASSOCIATION STUDIES. GENE HUNTING

Figure 10.2: Summary of genes discovered in published GWAS.

Published Genome-Wide Associations through 12/2012
Published GWA at ps5x10⁻⁸ for 17 trait categories


in 2012. Given the cost of GWAS and the voluminous data gathered from them, there are public data bases for GWAS studies and a searchable NHGRI web site (http://www.genome.gov/gwastudies/; see Hindorff et al., 2009) that catalogs all published GWAS studies, noting the genes associated with various phenotypes. A synopsis of their findings through 2012 is shown in Figure 10.2.

Note the phenotypes listed in the lower right-hand side of the figure. The categories “nervous system disease” and “other trait” are the closest ones for behavior. All others except for body measurement are medical disorders or allied phenotypes.

Gazing at Figure 10.2, you may find it surprising to learn that GWAS is not really in its infancy, but somewhere in childhood. Initial studies used small samples. After several years, it was recognized that sample sizes need to be dramatically increased. Today, sample sizes for complicated traits like psychopathology are in the tens of thousands. Often individual studies are pooled to achieve an adequate sample size for analysis.

It is a distinct advantage that GWAS can hunt across the whole genome with hundreds of thousands of polymorphisms. That is also its major disadvantage. The problem comes about because each polymorphisms involves a single statis-
tical test for an association. That means, hundreds of thousands of statistical tests. To see the problem, imagine that there are 100,000 statistical tests. If we use the conventional statistical criterion of a 5% false positive rate to determine significance, then \( .05 \times 100,000 = 5,000 \) of them will be false positives just due to chance. Suppose that we observe 6,000 significant results. Which 5,000 of those are just due to chance and which 1,000 are real associations? (The advanced Section 10.3 below gives more information about the statistics used when many genes are analyzed.)

To overcome for this problem, GWAS sets the significance level to something much lower than .05. Typical values are 10E-6 to 10E-8. Such a low significance level, however, comes with a price. It is necessary to increase sample size to achieve adequate statistical power to detect an effect that is, in fact, present. Currently, most GWAS studies involve samples in the tens of thousands and/or involve testing of multiple samples. As a result, research interested in a certain phenotype have formed large consortiums that pool data on different samples. Let us examine one such consortium and, along the way, learn more about GWAS.

10.2.2.1 The Tobacco and Genetics Consortium

Because cigarette smoking, the most prevalent form of nicotine addiction, is associated with cancer, pulmonary, cardiovascular and psychiatric disorders, considerable data have been gathered on it in a wide variety of studies. In 2009, a group of research called the Tobacco and Genetics Consortium (2010) pooled their resources to examine GWAS data from 16 different samples. There were seven studies involving subjects from the general community, the Framingham Heart Study being one of them. The remaining nine studies were case control studies of atherosclerosis, diabetes, cancer, anxiety and depression. Individual sample sizes ranges from 856 to over 22,000. The number over all studies was just shy of 75,000.

Such collaborations present special issues. Not all studies used the same genotyping technology nor did they assess the same polymorphisms. But because all 16 assessed polymorphisms across the whole genome, they group could use *imputation* (Li et al., 2009) to fill in polymorphisms that were not assessed in a single study.\(^6\) With imputation, the group had data on approximately 2.5 million SNPs. To control for false positives, the researchers set the significance level at \( 5 \times 10^{-8} \) or .00000005. This is called the *genome-wide significance level*. This is used when scanning the whole genome. Other significance levels may be used for other criterion—e.g., a gene previously suspected to be associated with the phenotype.

The typical manner of presenting the overall results from such a study is a graphic called a *Manhattan plot*, so named after the skyline of the epony-

---

\(^6\)The concept of imputation is straightforward. Suppose one study assays for SNP1 but another study records a nearly polymorphism, SNP2, that is in linkage disequilibrium with SNP1, but does not directly genotype SNP1. Given the known statistics for linkage disequilibrium, we predict the genotypes for SNP1 in the second study from the SNP2 genotypes.
10.2. TWO PURPOSE FOR ASSOCIATION STUDIES

mous borough of New York City. The four Manhattan plots from the original publication are reproduced in Figure 10.3.

The horizontal axis of the figure gives position of the SNPs in linear order over the 22 autosomal chromosomes. Because there are so many SNPs, they are banded according to their location in the chromosome.

The vertical axis for all four panels in Figure 10.3 is the $p$ level, expressed as $-\log_{10}(p)$. Hence, a value of 6 for $-\log_{10}(p)$ is a decimal point, followed by 5 zeros, followed by a 1. The larger the value of $-\log_{10}(p)$, the more statistically significant the result. The two vertical lines correspond to a $p$ value of $5 \times 10^{-8}$ (the line just below $-\log_{10}(p) = 8$) which is the group’s standard for genome-wide significance and $p = 0.0001$ ($-\log_{10}(p) = 8$) which is an arbitrary criterion for suggestive significance. Note that the values of the vertical axis differ among the four phenotypes.

You may be surprised to read one the major results: “only one locus contained SNPs that exceeded genome-wide significance for one of the four phenotypes” (p. 441). Indeed if you examine the four panels, you will see that only one, panel (a) for cigarettes per day (CPD), shows an area with a $p$ value smaller than $5 \times 10^{-8}$. You read that correctly. Despite the enormous sample size and the vast number of polymorphisms, only one gene achieved significance at the genome-wide level, and that was for only one phenotype.

The area of significance is in 15q25, and coincides with the a cluster of genes (CHRNA5, CHRNA3, CHRN4B) which code for proteins that comprise part of the nicotinic receptor. Exactly what polymorphism within the gene might be responsible for CPD? This is problematic. There were 130 SNPs within the region that reached genome-wide significance, so teasing the causal variant(s) out of this pool is not possible. Once again, this demonstrates the correlational nature of a positive GWAS finding.

Gazing at spike at chromosome 15 in panel (a) of Figure 10.3, you may think that CHRNA3 has a large effect on the number of cigarettes smoked per day. Recall, however, that statistical significance does not necessarily imply a large effect size. The strongest predictive SNP in this region\(^7\) has a $p$ value of $2.8 \times 10^{-73}$. That is not a misprint, the $p$ value starts with a decimal point and is followed by 72 zeros! Yet that SNP accounts for only 0.5% of the variance in CPD, a difference of less than one cigarette per day.

What of the other “spikes” in Figure 10.3 that are suggestive but do not pass the threshold for genome-wide significance? The authors examined these in light of other published GWAS studies on smoking and reported several other genes that may be associated with smoking behaviors. One notable gene was BDNF which was associated with smoking initiation. BDNF stands for brain-derived neurotrophic factor and is an important protein in the survival and growth of neurons.

\(^7\)rs1051730
Figure 10.3: Manhattan plots for four smoking phenotypes from the Tobacco and Genetics Consortium analysis.

From The Consortium on Tobacco and Genetics (2010); CPD in panel (a) is cigarettes per day.
10.3 Multiple statistical tests*

Most modern gene hunting strategies assay a large number of genes and, in doing so, encounter the problem of multiple statistical tests. The problem, outlined in Section 10.2.2, is that when many statistical tests are preformed, a large number of them may be false positives. To minimize the false positive rate, one alters the \( p \) level at which a finding is termed significant. Instead of the traditional \( p \) value of 0.05 used in the behavioral sciences, one may use a \( p \) level of \( 10^{-7} \) or lower.

There are several customary ways of adjusting the \( p \) value (Bush and Moore, 2012; Clarke et al., 2011). The first is the Bonferroni correction. This takes the traditional \( p \) value and divides it by the number of statistical tests. For example, if the traditional value is 0.05 and there are 100,000 tests, then the Bonferroni corrected \( p \) value is \( 0.05/100,000 = 5E^{-7} \). The Šidák adjustment (Šidák, 1967) is a close cousin to the Bonferroni.

A second correction is the false discovery rate or FDR (Benjamini and Hochberg, 1995). The mechanics of calculating the FDR are too complicated to present here, but its logic is straightforward. Again, start with the traditional significance value of 0.05. This is the probability of a false positive for any single statistical test. An FDR applies this not to a single statistical test, but to the set of all results determined to be statistically significant. In other words, find a \( p \) value such that 5\% of all results called significant will be false positives.

A third, and currently the “gold standard” (Clarke et al., 2011) is the permutation test. In a case control design, the assignment of an observation to “case” versus “control” is made randomly. Repeated simulations are used to establish a distribution of \( p \) values. This is the distribution that is expected under the null hypothesis of no association. The cutoff would be the \( p \) value that separates the lower 5\% of the distribution from the upper 95\%. A problem with the permutation test is that it is computationally very demanding and must be performed study by study.

Several other procedures, including Bayesian approaches and graphical methods have also been proposed (e.g., Moskvina and Schmidt, 2008; Verzilli et al., 2006; Zhang, 2012).

The major problem with adjustments to avoid false positives is that they increase the frequency of false negatives. That is, they increase the number of genes that do in fact have an effect but fail to meet the criterion for statistical significance. The only way to avoid this “damned if you do and damned if you don’t” scenario is to increase sample size.

10.4 References


