Chapter 11: The Association Study

Introduction

Linkage locates the approximate position of genes. The association design, on the other hand, takes a known gene and tests whether individual differences in the gene are statistically associated with a phenotype. The logic of the design is straightforward—genotype individuals at the locus of interest and then perform a statistical test. For example, assume that a gene for a dopamine receptor is known to be polymorphic. Because many of the neuroleptic drugs for schizophrenia influence the dopamine system, it is natural to ask whether this locus may be associated with schizophrenia. Hence, a researcher may genotype 50 schizophrenics and 50 controls and cross-tabulate disease status with genotype at the locus. Table 11.1 presents fictitious data for this example.

[Insert Table 11.1 about here]

After gathering these data, the researchers would perform a statistical test[1] to see whether the distribution of the genotypes in the controls differs from the distribution in schizophrenics. For these hypothetical data the result would not be significant so the researchers would conclude that this polymorphism is not associated with schizophrenia[2].

[1] A χ^2 test is usually performed.

Advantages of the Association Study

Association designs have both advantages and disadvantages compared to linkage strategies. The two major advantages are statistical power[3] and knowledge. In terms of statistical power, it is much easier to find a gene contributing to a disease with an association design than with linkage. When several genes contribute to liability, then even very large linkage studies may not have sufficient statistical power to detect a gene. Much smaller samples are required to detect the effect in an association design.

The second advantage is the increase in knowledge. Linkage studies find the approximate location of a gene but cannot actually locate the gene itself or provide information about what the gene does. With an association study, on the other hand, one can begin to develop hypotheses about the pathophysiology of the disorder from knowledge about the locus and what it does. For example, the association between the APOE locus and Alzheimer's Disease has established a vigorous area of research into explaining why the E4 allele confers liability. This will assist us in understanding Alzheimer's much more than finding a positive linkage.

Disadvantages of the Association Study

With such advantages, one might logically ask why linkage designs are used at all. The reason is that there are three major disadvantages to association designs. The first is

---

[2] It is incorrect to conclude that the *gene* is not associated with schizophrenia. There may be other polymorphisms at the same locus that has associations with schizophrenia.

that a gene that has a theoretical reason to be connected to a disease must have already been identified beforehand. As the human genome project nears completion, more and more genes will be identified, so this disadvantage is not as serious as it once was. A second problem of association studies will arise when, paradoxically, the genome project has identified a large number of candidate loci. If researchers proceed in a shotgun fashion and try to associate a disorder with each and every known gene, then a large number of false positive findings can arise just by chance[4]. Again, this problem can be circumvented with some common sense. For example, those suspecting that dopamine may be associated with schizophrenia could test all genes known to have a direct influence on dopaminergic neurotransmission. The third disadvantage is population stratification. Because the effects of population stratification on association designs are still unclear, let us devote some time to discuss them.

Population stratification and the association design.

Population stratification comes about because allele frequencies are not evenly distributed across human populations. Hence, when people with a disorder and the control group are not carefully matched in terms of ethnicity, erroneous conclusions can

---

[3] Statistical power is the probability that a statistical test can reject an hypothesis when, in fact, the hypothesis is false.

[4] The more statistical tests that are performed, the greater the probability that at least one of then will be significant just by chance. Given that the number of peptide coding genes may approach 100,000, shotgun approaches to association will definitely uncover may false leads. One can, of course, adjust the statistical method to minimize the false positives, but this strategy robs the association design of its major advantage, statistical power.

be drawn about the causal role of genes. This is a difficult concept in the abstract, so let us consider a simple example.

Assume that science has identified a gene that influenced the rate of and amount of melanin production in the skin. Let allele *A* denote high production while allele *a* is associated with low production. Genotype *AA* would be darkly pigmented, genotype *aa* would be lightly pigmented, while *Aa* would have intermediate pigmentation. Suppose that the disease that we wished to study was sickle cell anemia and that we did a straightforward association design. From a clinic in the US, we select 50 patients with sickle cell anemia and then pick 50 random controls and genotype all 100 individuals on melanin locus. The design is given in Table 11.2. Note how this table is identical in form to Table 11.1. The only difference is that sickle cell anemia has been substituted for schizophrenia.

[Insert Table 11.2 about here]

Mentally try to fill in numbers for this table. The majority of sickle cell cases in the US are the descendents of Africans. Hence, there will be a high frequency of allele *A* among those with sickle cell. In a random sample of the US, however, the majority will be white. Hence there will be a high frequency of allele *a* among controls. This would give a positive statistic. Should one then conclude that there is an association between the hypothetical melanin locus and sickle cell anemia?

The answer to this question lies in the relationship between correlation and causality that should be well recognized by even the beginning social science student. There is a real statistical association--or correlation if you prefer that term--between

pigmentation and sickle cell anemia in the US. That association, however, is *not* causal. For obvious reasons, genes for melanin production have nothing to do with the β hemoglobin chain that causes sickle cell anemia.

Controlling for population stratification

This is an extreme example of the problem of population strategy and no contemporary researcher would ever consider using such a design. The question still remains as to how much population stratification might influence disorders with little ethnic association. At present, the answer is unknown, so most geneticists recommend two strategies to control for population stratification.

The first is simply to use ethnically matched controls. Imagine for the moment that hypothetical study used this tactic instead of randomly selecting controls from the general population. Every person entering the clinic would be asked to state his/her ethnicity. If the person responded as African-American, then the research would randomly sample from the African-American population as a control. With these types of controls, the data in Table 11.2 would no longer be statistically significant.

The second way to control for population stratification is to use *genetic relatives* as controls. Because of the uncertainties about population stratification, this has become the gold standard for genetic research in genetically heterogeneous populations such as the US. There are several different research designs and tactics that use genetic relatives as controls, each with its own clever name and statistical jargon. For our purposes here,

let us refer to these designs as *within-family segregation studies* because the common denominator is that they all trace the segregation of alleles within pedigrees. Two examples follow.

First, consider the use of unaffected siblings as controls. Siblings are perfectly matched on ethnicity, so differences between siblings on a locus cannot be due to population stratification. The unit of analysis in this design is the *sibship*, not the individual. Any row in the data matrix consists of a sibship with separate columns for the affected and unaffected sib in the pair[5]. The hypothetical data in Table 11.3 illustrate how the design would be implemented for the schizophrenic phenotype.

[Insert Table 11.3 about here]

By recording the number of *A* alleles in the affected and in the unaffected sibs, we can then perform a statistical test to test for an association[6]. If the test is significant and positive, then the schizophrenic has more *A* alleles than his/her normal sib does and allele *A* is associated with liability to schizophrenia. If the test is significant and negative then, then the schizophrenic has fewer *A* alleles than his/her normal sib, so allele *a* is related to liability. A nonsignificant test, of course, implies that the locus has no association with schizophrenia.

A second example of a within-family segregation design is the classic description of the *transmission disequilibrium test* or TDT. For example, suppose that there is good

---

[5] It is also possible to deal with sibships greater than size two. However, the mathematics of doing so is beyond the scope of this text.

[6] Choosing allele *A* is arbitrary. One would come to the same substantive conclusion by recording the number of *a* alleles. The specific statistical test would be a paired *t*-test involving the last two columns of Table 11.3.

reason to suspect that a polymorphic locus might be associated with bipolar disorder[7] and that allele *a* is likely to contribute the liability of the disorder. To keep the example simple, suppose that we identified a large group of bipolar mothers who were genotype *Aa* on the locus and who were married to unaffected men, all of whom had genotype *AA*. We then study the children by genotyping them at the locus and by determining whether or not they suffer from bipolar disorder.

The expected genotypic frequencies in the offspring of *Aa* X *AA* matings are simple to compute--half will be *Aa* and the other half *AA*. All of the *Aa* offspring will have received the *a* allele from their affected mother. Hence, if allele *a* contributes to bipolar disorder, they will be at elevated risk for having a bipolar phenotype. All *AA* offspring received allele *A* from mom, and if there is a real association, then they should be at reduced risk for bipolar disorder. If allele *a* does not influence bipolar disorder, then *Aa* offspring are at equal risk with *AA* offspring. According to this hypothesis, half of all affected offspring would be *Aa* and half *AA*.

Table 11.4 presents hypothetical data from this study in which 200 offspring were studied. As expected half of them are genotype *Aa* and the other half are *AA*. Of the 200 offspring, 28 or 14% have bipolar disorder, a figure that is within range of the overall empirical literature for this phenotype. However, all but three of these affected offspring are genotype *Aa* on a locus suspected to be associated with bipolar disorder. These data would strongly support the hypothesis that allele *a* confers risk for bipolar disorder.

[7] Bipolar disorder is a mood disorder in which individuals experience episodes of mania (markedly elevated mood with a series of concomitant symptoms like overactivity) and episodes of deep depression.

Conclusions

In the near future, linkage and association studies will be used in tandem to search for genes contributing to liability in DCG. In fact, the large pedigrees collected in linkage studies are very suitable for within-family segregation designs used to test for associations. For DCG, especially those involving behavior, there should be no argument as to which strategy is superior—both are currently needed. As the human genome becomes better characterized by dense marker loci, it will be easier to detect the influence of genes of moderate effect size with linkage. On the other hand, as more and more polymorphisms are identified that influence protein structure and/or production, then association designs, with their greater statistical power, will become increasingly important.

In the future, more insight will be gained into the problem of population stratification. Perhaps (and hopefully) there will be little difficulty as long as one uses common sense in matching for ethnicity. On the other hand, future data might show that stratification is a real problem and within-family segregation designs are the only way to achieve satisfactory scientific control.

Table 11.1. Example of a genetic association study for schizophrenia.

	Genotype:			
Disease Status:	*aa*	*Aa*	*AA*	Total
Control	4	17	29	50
Schizophrenia	5	22	23	50
Total	9	39	52	100

Table 11.2. Example of a hypothetical association study between a melanin production locus and sickle cell anemia.

	Genotype:			
Disease Status:	*aa*	*Aa*	*AA*	Total
Control				50
Sickle Cell				50
Total				100

Table 11.3. Example of a genetic association design using siblings as controls to avoid the problem with population stratification.

	Genotypes:			**Number of *A* alleles:**	
Family	Schizophrenic Sib	Normal Sib		Schizophrenic Sib	Normal Sib
Smith	*aa*	*Aa*		0	1
Jones	*Aa*	*AA*		1	2
.	.	.		.	.
Williams	*aa*	*Aa*		0	1

Table 11.4. Example of a genetic association study using the transmission disequilibrium test. Entries are the number of offspring who fall into each cell.

	Phenotype:		
Offspring Genotype:	Bipolar	Unaffected	Total
a	25	75	100
A	3	97	100
Total	28	172	200