Searching for genes in complex diseases: lessons from systemic lupus erythematosus

Commentary

See the related Letter to the Editor on pages 1501–1502

Neil Risch

Department of Genetics M322, Stanford University School of Medicine, Stanford, California 94305-5120, USA.
Phone: (650) 725-6006; Fax: (650) 725-1534; E-mail: risch@lahmed.stanford.edu.

An article appearing last year in the *JCI* (1) found evidence of a role for the gene for poly(ADP-ribose) polymerase, *PARP*, in susceptibility to systemic lupus erythematosus (SLE). Two letters now appear in the current issue, one, by Criswell et al. (2), refuting a role of *PARP* in SLE, and the other, from the original authors, responding to this letter (3). This exchange merits some additional discussion, because it illustrates many of the difficulties faced by researchers searching for genetic factors in complex diseases.

Linkage and linkage disequilibrium

The traditional approach for locating a disease gene in humans is linkage analysis. In this method, families with multiple affected relatives are scanned with roughly evenly spaced genetic markers — typically short tandem repeat markers, or "microsatellites," DNA sequences that show considerable variability among people but that have no functional consequences. The logic is that if a gene somewhere in the genome is responsible for the disease, affected family members are expected to have inherited the same disease-predisposing allele at the locus, and markers that lie physically near this disease gene will be transmitted along with the disease allele. If one uses enough markers — about 350 is usually sufficient — at least one of these markers will lie sufficiently close to the disease gene that its inheritance pattern will match the inheritance pattern of the disease in the family. Moreover, in a collection of families with different mutations in this same disease gene, the correlation between marker and disease status will hold even though more than one disease mutation (and more than one variant at the marker locus) is represented in the group. When single families or collections of families are large enough that this correlation can be shown to be highly unlikely to have occurred by chance, the marker is said to be linked to the disease gene. Linkage, so defined, indicates only that the gene for the disease lies in the vicinity of the marker, but tight linkage may permit the precise localization of the disease gene through the process of positional cloning.

Although a gene and a marker that are located near each other on a chromosome must demonstrate correlated inheritance patterns in families, this correlation is typically less than complete because of recombination. Recombination occurs when exchanges, or "crossovers," occur between a pair of homologous chromosomes in meiosis. The probability of crossing over between two loci increases with the physical distance between them. For this reason, recombination serves to define the physical location of a disease gene. After linkage is detected with an initial marker, many other markers nearby are also examined. Markers showing the strongest correlation with disease in families are assumed to be closest to the disease locus. Depending on the number of families available, this effort can limit the putative location of the disease gene to perhaps one million base pairs, or 10 to 30 genes.

Depending on a number of factors, such as mode of inheritance (dominant or recessive), penetrance (risk of being affected), and population under study, the number of different disease mutations represented in a collection of unrelated families will vary. Typically, greatest allelic diversity occurs for high-penetrance severe dominant diseases, because individual mutations are not expected to survive long in the population. By contrast, low-penetrance and/or mild dominant mutations or recessive mutations experience less selection pressure and thus tend to survive longer, leading to greater disease allele homogeneity. At one extreme, all families in a collection will carry a different allele of the disease gene, while at the other extreme, all families possess the same ancestral mutation. In this latter scenario, which is typical in genetic isolates where unique mutations can, by chance, become common, the families are actually related to each other, although the relationship may be too distant to be recognizable.

In such a group, the various collected families actually form one larger, extended pedigree and thus, in theory, provide much greater power for gene localization by linkage analysis. The problem, however, is that much of the pedigree is missing — i.e., the earlier generations that connect the various current families. Without knowing the connections between families, conventional linkage analysis offers no further benefit. However, another approach is possible, taking advantage of the fact that a single disease mutation, with a unique set of genetically linked markers, is responsible for all of the disease within a group of more or less distantly related families. The physical extent of the chromosomal region preserved around the mutation will depend on the number of generations (and hence the number of meiotic recombinations that have transpired) since the disease allele was transmitted by the most recent common ancestor of the affected families. Recent common ancestry implies a large amount of surrounding DNA preserved, while an older ancestry implies a small amount. Any genetic marker that lies within a preserved segment will demonstrate an identical allele among subjects that have inherit-

Markers showing the strongest correlation with disease in families are assumed to be closest to the disease locus.

ed the disease mutation across independent families. Typically, the closer the marker is to the disease locus, the greater the proportion of affected subjects carrying the identical allele at the marker. By this approach, linkage disequilibrium analysis, markers in the candidate region identified by linkage analysis can be used to narrow the location of the disease gene. In measuring the strength of linkage disequilibrium for a given marker, it is also important to have unaffected control subjects from the same population, because an allele shared among affected subjects may also be common in the general population and thus shared by chance rather than due to proximity to the disease locus. Linkage disequilibrium analysis can often pinpoint a gene of interest to a short segment of DNA — perhaps as few as 100,000 bp, which may contain only a few genes. This technique has been successfully used in the hunt for the genes underlying cystic fibrosis, Huntington's disease, diastrophic dysplasia, and primary torsion dystonia, among many others.

Dealing with complexity

Although positional cloning has been highly successful in identifying the loci underlying Mendelian diseases, it has been much less so for identifying genes for common, complex disorders. The reason is that Mendelian disorders are genetically simple: they feature a strong correspondence between the presence of a predisposing genotype at a single genetic locus and the phenotypic outcome. This correspondence produces a strong linkage signal in families and allows for localizing a disease gene by recombination events. The more common familial but complex disorders involve numerous loci, which may interact with each other to predispose to disease. Because the total genetic effect is partitioned among several or many loci, the correspondence between a predisposing genotype at one such locus and the disease outcome is weaker, greatly reducing the power of linkage analysis.

The power of linkage analysis to locate susceptibility loci for complex diseases is much greater in animal models than in humans, for a variety of reasons. First, inbred strains are used, which tends to reduce the genetic complexity and limit genetic effects to the loci that differentiate the original strains used in the breeding experiments. Second, by design, all matings are informative (e.g., all parents are heterozygous in an intercross). Third, all matings have the same phase (i.e., the genotypes at a presumed disease locus are known), and offspring from all matings can be combined into a single analysis. Thus, one strategy taken by human geneticists is first to localize a disease gene in an animal model, and then to examine the homologous region in human families. The assumption is that variation in the same gene exists in the human population and influences susceptibility to the analogous disease.

This approach has a mixed track record. There are some striking examples of homology, such as the similar role of the MHC in human type 1 diabetes and in the nonobese diabetic mouse. On the other hand, in many cases it has proved impossible to demonstrate such homologies. This approach may fail either because the gene of interest functions differently in humans and mice, or because functional variants comparable to those seen in the model system do not exist (or are uncommon) in the corresponding human gene.

For complex human diseases, a simple mode of genetic inheritance is not apparent, and indeed multiple contributing genetic loci are likely to be involved. In this setting, study designs that do not depend on the particulars of mode of inheritance are required. Typically, affected relatives provide most of the information for such analyses, and studies focus on searching for increased sharing of marker alleles above chance expectation among affected relatives. The simplest of such studies involves affected sibships, where allele sharing in excess of 50% (the expectation when there is no linkage) is sought.

One additional point relates to the issue of replication. The basis of all scientific research is the testing of hypotheses and validation of results by independent researchers. Independent replication, typically viewed as the *sine qua non* for accepting a hypothesis, has become an especially difficult issue in the genetic study of complex diseases. When a genetic effect is large, most independent researchers can readily obtain similar results with strong levels of statistical significance. Most genes for Mendelian disorders have lived up to this expectation, and there have also been occasional successes in more complex cases, such as Alzheimer's disease. In this disorder, the risk to carriers of the *APOE-ε4* allele is substantially higher than to noncarriers, an effect that has been seen in virtually all studies published to date. However, when genetic effects are weak and possibly context-dependent (e.g., they may vary by sex, ethnicity, or precision of diagnosis), replication may be particularly difficult, and very large samples may be required before confident conclusions can be drawn.

Is there an SLE locus on distal 1q?

With this background, we can examine the recent genetic studies involving chromosome region 1q42 and systemic lupus erythematosus (SLE), and the two letters published in the current issue. In 1997, Tsao et al. (4) published evidence for linkage on the long arm of chromosome 1 (1q41–42) in nuclear families with multiple cases of SLE. Prior linkage studies in mouse models of SLE had localized a gene or genes influencing specific SLE characteristics to the telomeric end of mouse chromosome 1. The homologous region in humans also lies on chromosome 1, in the segment 1q21 to 1q42. Thus, Tsao et al. (4) focused their initial human linkage studies on this location on chromosome 1 and used seven microsatellite markers that spanned about 27 centimorgans (cM). Two of these markers, D1S229 and D1S213, which are located at positions 238 and 242 cM on chromosome 1 according to a standardized genetic map (www.marshmed.org/genetics/), showed suggestive evidence of linkage; increased allele sharing of 64% and 61%, respectively, was observed for these two markers among 52 affected sib pairs from 43 multiplex families.

More recently, the same group, studying 26 additional affected sib pairs (for a total of 78) and six additional markers, found increased evidence for linkage, deriving a maximum lod score of 3.3 near a marker located approximately midway between the two markers noted above. They further defined a 5-cM suggested confidence interval in which the predisposing gene would be expected to reside (1). These authors also argued that their linkage findings are supported by independent studies by Gaffney et al. (5) and Moser et al. (6). However, close examination of these

latter two studies reveals some disquieting inconsistencies in the linkage results. First, the study of Gaffney et al. (5) reported a maximum lod score of 1.51 in region 1q42 in 105 SLE-affected sib pairs at a marker positioned 10 cM distal to the far end of the confidence region described by Tsao et al. (1). At this location, Tsao et al.'s linkage evidence is much weaker; conversely, using multiple markers in the 1q41–1q42 region, Gaffney et al. (5) found no linkage evidence in the confidence interval described by Tsao et al. (1). Similarly, although Moser et al. (6) reported a lod score of 3.5 at 1q41 in 31 African-American families, their best evidence for linkage was for markers located outside Tsao et al.'s confidence interval. Furthermore, the high lod score was obtained only using these 31 African-American families and could not be duplicated in a larger collection of 55 Caucasian families. In a follow-up study, Moser et al. (7) studied 13 markers spanning from 235 to 245 cM, including Tsao et al.'s confidence interval. Nearly all of the markers gave no or little evidence for linkage in this region, either in the combined data set or in the separated ethnic groups. Although an analysis of all markers simultaneously was not performed, the impression is that the linkage evidence in this region is weak, at best.

The lesson from these linkage results is that localization of genes with modest effects by linkage analysis may be difficult and, if it is to be possible at all, may require the analysis of a large number of families. In this case, it is likely that Tsao et al. (1) were overly optimistic in constructing a 5-cM confidence interval on chromosome 1q42 for an SLE susceptibility locus. Furthermore, replication of linkage results requires careful examination of the exact same chromosomal locations. Positive lod scores some distance away on the same chromosome cannot be considered as a valid replication; the linkage evidence for an SLE locus in this region of chromosome 1q42 appears not to be confirmed.

Do *PARP* mutations underlie SLE?

In positional cloning, the usual next step in identifying a disease gene is to search the linkage confidence interval for candidate genes. Depending on the size of the interval, there may be a handful to hundreds of genes to examine. Once genes in the interval are identified, they are sequenced to find allelic variation that could be causally related to the disease in question. In this case, the confidence interval was defined as 5 cM (1). Assuming a total human genetic map of 3500 cM length and a total of 100,000 genes, we would estimate that this interval contains 100,000 × (5/3500) = 143 genes. Tsao et al. (1) focused on three genes known to lie in this interval based on prior knowledge of gene function. Of the three candidate genes that were tested, one, *PARP*, appeared to show significant evidence of linkage disequilibrium. *PARP* also appeared to be a good candidate for relation to SLE susceptibility, because it shows reduced levels of expression and activity in SLE patients (8) and because its product participates in DNA repair and apoptosis, two events that may go awry in SLE (1). The marker used was a microsatellite located 906 bp upstream from the transcription initiation site. In a study of 124 affected offspring, Tsao et al. (1) found excess transmission of one allele (the 85 bp allele) to offspring from parents heterozygous for this allele; the statistical evidence was very strong, even after adjusting for the number of loci and alleles tested. The same pattern was observed in Caucasian and non-Caucasian (primarily Hispanic and Asian) subgroups. The authors interpreted these results to indicate either that this particular variant, located in the promoter region of *PARP*, is directly involved in disease susceptibility or that it is in linkage disequilibrium with the actual causative allele, which might be located in *PARP* or another gene close by.

Subsequent to the publication of these results, an independent group presented results from a study of French Caucasian SLE patients and ethnically matched controls (9). A total of 171 SLE patients and 193 controls were examined for allele frequencies at the same *PARP* microsatellite studied by Tsao et al. These authors found no evidence for association of the 85 bp allele with SLE. In fact, the frequency of the 85 bp allele was slightly lower in the SLE patients (61.1%) than in controls (65.5%). Similarly, in a study of 90 African-American SLE patients and 100 ethnically matched controls, the 85 bp allele was not found to be at increased frequency in the SLE patients (10).

In the current issue of the *JCI*, we now have a third attempt at replication, this time in a large sample (n = 448) of SLE patients and parents (2). These authors likewise find no significant excess of transmission of the 85 bp allele to affected offspring from heterozygous parents. In response to this latest failure at replication, the original authors postulate that lack of replication is due to the tested polymorphism in *PARP* not being the causative allele but, rather, being in linkage disequilibrium with the actual, nearby gene (3). However, this conjecture cannot explain why the same association does not appear in two large studies of the same ethnicity, Caucasians (2, 9). Since the original observed association apparently cannot be replicated by independent investigators, it may have been a false positive; or the effect of variation in *PARP* may be too weak to be detected with adequate power without extremely large study populations.

The difficulty in replicating the initial linkage and linkage disequilibrium findings, as seen here for SLE and chromosome 1q42, is disappointing but has a considerable precedent in the study of other genetically complex disorders. For many such conditions — multiple sclerosis, autism, schizophrenia, and the like — the evidence for a strong genetic contribution is compelling, but specific genes have proved elusive. As in the current case, many of the candidate genes analyzed have enjoyed considerable attention because, based on functional information, it was possible to draw a plausible biological model that explained how the pathology might result from missing or altered function of the gene of interest. Despite such information, however, the ultimate evidence for a role in disease susceptibility must come from human association studies. In the face of conflicting data on association, it is often difficult either to confirm or to refute that the initially promising candidate indeed contributes in a minor way to the development of the disease. On the other hand, moderate to large gene effects should be readily replicable, as has been the case, for example, of *APOE* and Alzheimer's disease.

Complex disorders are believed to arise from the interaction of many genes; these genes have so far largely proved refractory to conventional positional

cloning. To date we know the identity and functionality of only a small proportion of the total complement of human genes that could be considered as candidates for a disease of interest. However, as the human genome project progresses, it will not only provide a catalog of all human genes and their functionality but also elucidate the naturally occurring genetic variation in those genes in the human population. With these data in hand, geneticists anticipate developing powerful new approaches to the study of complex diseases.

1. Tsao, B.P., et al. 1999. PARP alleles within the linked chromosomal region are associated with systemic lupus erythematosus. *J. Clin. Invest.* **103**:1135–1140.
2. Criswell, L.A., et al. 2000. PARP alleles and SLE: failure to confirm association with disease susceptibility. *J. Clin. Invest.* **105**:1501–1502.
3. Tsao, B.P., et al. 2000. Letter to the editor. *J. Clin. Invest.* **105**:1501–1502.
4. Tsao, B.P., et al. 1997. Evidence for linkage of a candidate chromosome 1 region to human systemic lupus erythematosus. *J. Clin. Invest.* **99**:725–731.
5. Gaffney, P.M., et al. 1998. A genome-wide search for susceptibility genes in human systemic lupus erythematosus sib-pair families. *Proc. Natl. Acad. Sci. USA.* **95**:14875–14879.
6. Moser, K.L., et al. 1998. Genome scan of human systemic lupus erythematosus: evidence for linkage on chromosome 1q in African-American pedigrees. *Proc. Natl. Acad. Sci. USA.* **95**:14869–14874.
7. Moser, K.L., et al. 1999. Confirmation of genetic linkage between human systemic lupus erythematosus and chromosome 1q41. *Arthritis Rheum.* **42**:1902–1907.
8. Sibley, J.T., Haug, B.L., and Lee, J.S. 1989. Altered poly(ADP-ribose) metabolism in the peripheral blood lymphocytes of patients with systemic lupus erythematosus. *Arthritis Rheum.* **32**:1045–1049.
9. Delrieu, O., et al. 1999. Poly(ADP-ribose) polymerase alleles in French Caucasians are associated neither with lupus nor with primary antiphospholipid syndrome. *Arthritis Rheum.* **42**:2194–2197.
10. Tan, F.T., Stivers, D.N., Reveille, J.D., Tsao, B.P., and Arnett, F.C. 1999. Polymorphisms of the poly(ADP-ribose) polymerase gene (PARP) in African American SLE patients. *Arthritis Rheum.* **42**(Suppl.):S311. (Abstr.)