Challenges in identifying genetic variation affecting susceptibility to type 2 diabetes: examples from studies of the calpain-10 gene

Nancy J. Cox*

The University of Chicago, Department of Human Genetics, 507H CLSC, 920 East 58th Street, Chicago, IL 60637, USA

Received July 26, 2001; Accepted July 31, 2001

Type 2 diabetes is a classic example of a complex disorder. It is strongly familial, but clearly arises as a consequence of the actions and interactions of many genetic and non-genetic factors. Type 2 diabetes is a common disorder, affecting 16 million Americans. It has a major impact on public health expenditures with more than 1 in 10 health care dollars spent on treating diabetes and its complications. Although a variety of therapies can be useful in treatment of type 2 diabetes, we remain sufficiently ignorant of the genetic risk factors to believe that identifying them will lead to better understanding of the primary physiology of the disorder, as well as to more specific and effective therapies. Moreover, identification of genetic risk factors may improve our ability to characterize more specific non-genetic risk factors for this disease that could be the targets for cost-effective prevention strategies. This manuscript reviews the challenges we face in moving from the linkage mapping of susceptibility genes for type 2 diabetes toward the identification of the genetic variation that actually affects risk to this disorder. I illustrate many of the challenges in designing, conducting and interpreting these studies by reviewing recent research conducted on the calpain-10 gene, implicated in positional cloning studies as a candidate gene for type 2 diabetes.

INTRODUCTION

Recent publications provide outstanding overviews of results of genetic studies of type 2 diabetes (1) and of the challenges inherent in genetic studies of complex disorders (2). In this manuscript, I focus more narrowly on recent studies of a positional candidate gene for type 2 diabetes. The calpain-10 gene (CAPN10) encodes a novel calpain-like cysteine protease, and was identified through positional cloning studies on the NIDDM1 region (3). Results of studies to date on CAPN10 provide a good preview of the challenges we will face in conducting and interpreting the studies that take us from linkage mapping of susceptibility loci for type 2 diabetes to the identification of the genetic variation that affects risk of disease.

IDENTIFYING CAPN10 AS A POSITIONAL CANDIDATE GENE FOR TYPE 2 DIABETES

The first published genome-wide screen to map susceptibility genes for type 2 diabetes identified a region at 2qter (NIDDM1) providing highly significant evidence for linkage in a sample of 330 Mexican-American affected sib pairs (ASPs) from Starr County, TX (4). Although no evidence for linkage was observed in samples from other racial/ethnic groups reported in this study, an independent replicate of Mexican-American ASPs provided nominally significant evidence for linkage in the same region (4). Subsequent studies reported evidence for a statistical interaction between the NIDDM1 region of chromosome 2 and the CYP19 region of chromosome 15, which increased the likelihood that diabetes susceptibility loci mapped to both regions and improved the resolution of the linkage localization (5). Fine-mapping and positional cloning studies on the NIDDM1 region focused on a 1.7 Mb region corresponding approximately to the 1-LOD confidence interval obtained for the NIDDM1 region when the chromosome 15 interaction was taken into account (3). Initial studies considered the association of genetic variation in this region with disease, by comparing allele and haplotype frequencies for polymorphisms (largely SNPs) from this region between unrelated patients and a random sample from Starr County, TX. The patients used in these studies included one member from each of 110 families from the original genome-wide screen. Thus, it was possible to classify the patients according to whether they came from a family that provided evidence for linkage in the NIDDM1 region, and in the CYP19 region of chromosome 15.

Results from these studies examining the association of the genetic variation in the NIDDM1 region with disease identified a region in which several markers showed such associations. Moreover, for some of the polymorphisms, the associations were stronger in the subset of patients from families with evidence for linkage in the NIDDM1 and the NIDDM1/CYP19 regions. Therefore, a region of ~70 kb centered on the first polymorphisms showing association with disease was resequenced in 10 Mexican-American patients in order to

*Tel: +1 773 834 1001; Fax: +1 773 834 0505; Email: ncox@genetics.bsd.uchicago.edu
identify all of the high frequency genetic variation in the associated region. Multiples of the additional polymorphisms identified in these studies also showed association with disease, but these markers spanned two genes and their intervening sequence. To better distinguish the genetic variation actually affecting susceptibility to disease, we sought to make more use of the original linkage data that generated the signal leading us to focus on this region. Instead of considering just the association of genetic variation with disease, we also considered the evidence for association of the genetic variation with the original evidence for linkage.

ASSOCIATION OF GENETIC VARIATION WITH THE EVIDENCE FOR LINKAGE

This concept is illustrated in Figure 1A, which summarizes the patterns of genetic variation across the entire NIDDM1 region in all patients routinely typed for DNA polymorphisms identified in the fine-mapping and positional cloning studies. The polymorphisms are ordered left to right from the most centromeric to the most telomeric. Each row summarizes the genotypic data for a single patient, with blue indicating the patient is homozygous for the common allele at that site, red indicating the patient is heterozygous at that site, and yellow indicating the patient is homozygous for the rare allele at that site. The patients are ordered top to bottom according to the NPL score (6) obtained using the original microsatellite marker data in the family from which the patient was drawn. Studies focusing on the association of genetic variation with disease could be illustrated by contrasting the color patterns observed for patients (Fig. 1A) with the color patterns observed for the random sample (not shown). In contrast, studies focused on the association of genetic variation with the evidence for linkage are illustrated by contrasting the color patterns observed for patients from families with evidence for linkage in the NIDDM1 region (Fig. 1A and B, top) with those observed for patients from families with evidence against linkage in the NIDDM1 region (Fig. 1A and B, bottom). Figure 1B illustrates observed patterns of variation for the same patient sample (in the same order from top to bottom) for only the polymorphisms typed within CAPN10. While there is an obvious color difference between the top and bottom parts of this Figure for the variation in CAPN10, there is little notable difference over the rest of the NIDDM1 region that was studied. This includes both genes adjacent to CAPN10, RNPEPL1 and GPR35, in which a number of variants showed some association with disease (3).

While it is possible to see the association of the genetic variation in CAPN10 with the evidence for linkage when illustrated in this way, more formal approaches to this problem were also described (3). The simplest approach can be applied to the data from the patient sample by classifying each family according to the genotype of the single member from the family typed for the polymorphism being assessed. Under the null hypothesis of no linkage disequilibrium (LD) between the polymorphism being studied and the variation in the region that actually affects susceptibility to disease, the evidence for linkage (obtained using the original genome screen data in all members of the family) is expected to be proportionally distributed among the genotypic groups according to the number of families in that group. When one or more of the genotypic groups shows disproportionate evidence for linkage, it suggests that the study polymorphism is in LD with the variation affecting susceptibility to disease.

A second approach was developed to address this problem for the subset of markers typed in all members of all families. In these studies, the observed evidence for linkage obtained in the ASPs concordant for the at-risk genotype(s) was contrasted with that expected, given the original linkage signal and the fact that only pairs concordant for a particular genotype are included in analysis. Under the null hypothesis of no LD between the tested polymorphism and the genetic variation affecting susceptibility to disease, there should be no increase in the evidence for linkage observed in pairs concordant for the at-risk genotype, once the original linkage signal and ascertainment of concordant pairs are taken into account.

As noted above, a number of polymorphisms across two genes (CAPN10 and GPR35) and their intervening sequence showed some association with disease, but a much smaller number located in CAPN10 or the adjacent intervening sequence showed significant association with the evidence for linkage. These observations, coupled with results from subsequent functional studies, and replication of observed associations in populations of Northern European descent led Horikawa et al. (3) to propose that variation at CAPN10 affects susceptibility to type 2 diabetes.

NON-CODING SEQUENCE VARIATION AND THE COMPLEXITY OF GENETIC MODELS

The CAPN10 variation implicated in susceptibility to type 2 diabetes was located in non-coding sequence. Although gel-shift and luciferase expression assays confirmed the possibility that at least some of the implicated variation could expression of CAPN10 (3), more direct support for this possibility came in studies of CAPN10 variation in non-diabetic Pima Indians. Results of these studies demonstrated that variation in CAPN10 was associated with a significant reduction in skeletal muscle calpain-10 mRNA levels as well as measures of insulin resistance in non-diabetic Pima Indians (7). The reduction in mRNA levels was observed for all calpain-10 isoforms, an observation more consistent with the hypothesis that the tested variation affects expression than that it affects alternative splicing of the calpain-10 mRNA (8).

The genetic model proposed to explain how variation at CAPN10 affects susceptibility to type 2 diabetes is complex. Combinations of variants are more strongly associated with disease than individual polymorphisms. Combining results obtained for both the original and replication samples of Mexican-Americans from Starr County, TX, indicates that individuals heterozygous for the two different haplotypes that form the high-risk combination have $\sim$3-fold increased risk relative to all other combinations of haplotypes, and $\sim$7-fold increased risk relative to the lowest risk haplotype combination. Individuals homozygous for either of the haplotypes in the high-risk combination were not at increased risk of developing type 2 diabetes.

The magnitude of the difference in risk between the lowest- and highest-risk haplotype combinations probably contributed to the ability to detect the evidence for linkage in this region. Simulation studies suggest that genetic models capable of generating the observed patterns of risk for all haplotype
combinations yield LOD scores with both higher mean and variance than were obtained for genetic models that merely replicated the observed increased risk of the high-risk combination relative to all other combinations (3). Thus, genetic models yielding the same risk for the high-risk variation at a locus relative to all other variation can still yield substantially
different expected LOD scores, depending on whether additional variation at the locus reduces risk of disease. This pattern, with some variation increasing risk of disease and other variation decreasing risk of disease, is similar to what has been observed for other regions with variation affecting susceptibility to complex traits (e.g. HLA in autoimmune disease, ApoE in Alzheimer’s disease). Given the variability in allele frequencies across populations, it is possible that some of the difficulties in replicating linkage signals across populations might be attributable to variability in the frequencies of polymorphisms that affect risk of disease in this way. Even when some of the contributing variation is found at similar frequency in two populations, if the variation that affects risk differs in frequency between the populations, the ability to detect effects of the susceptibility locus via linkage mapping might be considerably altered.

RESULTS OF FOLLOW-UP STUDIES

Although calpain proteases are generally thought to cleave proteins at a single site to activate or inactivate the protein, the primary function of calpains remains unclear, and they have been characterized as proteases ‘in search of a function’ (9). Results of recent studies suggest that calpains may play a role in insulin secretion and insulin action (10). These studies found that short-term administration of calpain inhibitors to mouse pancreatic islets increases the insulin secretory response to glucose. Exposing skeletal muscle strips or adipocytes to calpain inhibitors reduces insulin-mediated glucose transport and incorporation of glucose into glycogen. These results are consistent with the possibility that calpain proteases play a role in glucose homeostasis, but because none of the tested calpain inhibitors is a specific inhibitor of calpain-10 (i.e. it is presumed that they inhibit all calpains), it is uncertain what role inhibition of calpain-10 played in these observations.

Studies on variation at \( \text{CAPN10} \) in a British population (11) suggest that replication and extension of the studies on \( \text{CAPN10} \) will be every bit as challenging as has been characterized for other diabetes susceptibility genes (12). Results of these studies in British trios and ASPs found no evidence for linkage in the \( \text{NIDDM1} \) region or for over-transmission of the high-risk haplotypes implicated in the studies of Mexican-Americans and Northern European populations. However, they did find that variation in \( \text{CAPN10} \), suggested in the research of Horikawa et al. (3) to have a functional consequence in expression of calpain-10 (SNP-44), was in perfect LD with an amino acid polymorphism at \( \text{CAPN10} \) (T504A). The rare alleles at these sites were significantly over-transmitted to affected individuals from complete trios (11). Although the frequencies of the rare alleles at these sites were twice as high in Mexican-American patients as in the random sample, the variation was sufficiently rare (\( q = 0.05 \) in the random sample) that it could not have accounted for the observed evidence for linkage in the Mexican-American data. Nevertheless, the possibility that variation at these sites contributed to the overall evidence for linkage cannot be ruled out. Moreover, it should be noted that the trios in which this variation was found to be over-transmitted had a significantly younger age at onset (by \( \sim 15 \) years) than individuals with type 2 diabetes in this population who were not confirmed to have two living parents. Thus, it is possible that the variation at one or both of these sites (SNP-44 or T504A) increases risk of a form of type 2 diabetes having an earlier age at onset.

IMPLICATIONS OF THE OBSERVATIONS AT \( \text{CAPN10} \) FOR FUTURE STUDIES

If variation at \( \text{CAPN10} \) is indeed solely responsible for both the original linkage signal at \( \text{NIDDM1} \) and the observed associations of variation at \( \text{CAPN10} \) with type 2 diabetes and related phenotypes, the genetic contribution of \( \text{CAPN10} \) to type 2 diabetes susceptibility is quite complex at the molecular level. Variation in non-coding sequence that would not traditionally have been considered functional appears to affect expression of the gene, and it is possible that additional variation (including amino acid polymorphisms) may also affect risk of disease. Combinations of variants may affect risk in ways that are difficult to predict based on the effects of the individual variants. If such molecular complexities routinely contribute to complex disorders, there are many implications for study design and data analysis. It will add considerably to the work and expense these studies entail if non-coding sequence variation must be tested and evaluated, and the analytic component to the identification of the variation affecting susceptibility to complex traits will be considerably more challenging if combinations of variants must be evaluated. Moreover, it is possible that we have prematurely excluded some of the functional candidate genes examined previously, because of simplistic assumptions about the nature of the variation likely to affect susceptibility to complex disorders.

Alternatively, variation at \( \text{CAPN10} \) may affect susceptibility to type 2 diabetes, but additional variation in the \( \text{NIDDM1} \) region at one or more other loci might also have contributed to the original evidence for linkage. The possible confounding of LD among variants at multiple, tightly linked contributing loci, each with modest effect on risk, could make it quite difficult to identify the other loci in the \( \text{NIDDM1} \) region affecting susceptibility to disease.

Finally, it is possible that the variation at \( \text{CAPN10} \) plays no role in susceptibility to type 2 diabetes, but is merely in LD with causal variation that is located elsewhere. Both of the flanking genes (\( \text{RNPPL1} \) and \( \text{GPR35} \)) were resequenced in 10 Mexican-American patients; however, and all informative variation (0.1 < \( q < 0.9 \)) with unique patterns in the 10 individuals was examined to determine whether it could better account for the observations than the variation at \( \text{CAPN10} \). None of the tested variations at these other loci, whether amino acid polymorphism, synonymous substitution or non-coding sequence variation, could account for the observed evidence for linkage, and none was as strongly associated with the evidence for linkage as the variation at \( \text{CAPN10} \). None of the variation at these other loci was in perfect LD with the variation at \( \text{CAPN10} \) implicated in either the original studies or the successful replication studies in other populations, and the LD in this region does fall off rapidly with physical distance. Moreover, if it is variation elsewhere that accounts for the original evidence of linkage, with the variation at \( \text{CAPN10} \) apparently implicated only because of its LD with the actual causal variation, the fact that the key variation at \( \text{CAPN10} \) affects expression of calpain-10 mRNA levels is merely a coincidence. Similarly, results of the physiological studies indicating
that inhibition of calpain proteases may have effects on insulin would also have to be coincidental.

There can be little doubt that we have much yet to learn about the interpretation of results of fine-mapping and positional cloning studies of complex disorders. The naïveté evident throughout our short history of linkage mapping studies of complex disorders served us poorly, and we can ill afford to be complacent about the challenges we face in moving these studies forward. Still, it is premature to assume these challenges will be insurmountable, and surely the rewards of identifying new pathways in complex disease etiology are worthy of our best efforts to overcome these challenges.

ACKNOWLEDGEMENTS

This work was supported in part by United States Public Health Service grants DK-55889, DK-47486 and DK-20595.

REFERENCES


