

## Patterns of Linkage Disequilibrium in the Type 2 Diabetes Gene Calpain-10

M. Geoffrey Hayes,<sup>1,2</sup> Laura del Bosque-Plata,<sup>3</sup> Takafumi Tsuchiya,<sup>3</sup> Craig L. Hanis,<sup>4</sup> Graeme I. Bell,<sup>1,2,3</sup> and Nancy J. Cox<sup>1,2</sup>

We investigated the patterns and extent of linkage disequilibrium (LD) in the vicinity of the type 2 diabetes gene calpain-10 (*CAPN10*) in Mexican Americans, European Americans, African Americans, and Chinese Americans. We found that *CAPN10* occurs within a single block of high LD and that LD decays rapidly outside of the gene. This reduces the likelihood that associations between *CAPN10* polymorphisms and type 2 diabetes could be attributed to variation at some distance from *CAPN10*. We also consistently observed that cases have more extensive LD than control subjects and that cases from families with evidence for linkage have more extensive LD than cases from families without evidence for linkage. These observations further suggest that there are one or more relatively common alleles increasing risk of type 2 diabetes in this local region. *Diabetes* 54:3573–3576, 2005

Linkage studies in Mexican Americans from Starr County, Texas, localized a susceptibility gene for type 2 diabetes to 2q37 (*NIDDM1*) (1,2), and subsequent positional cloning analyses suggested that calpain-10 (*CAPN10*; OMIM 605286) was the gene associated with type 2 diabetes. Individual studies have yielded mixed results, but pooled and meta-analysis of large numbers of case and control subjects (>5,000 each) have provided strong support for the involvement of two single nucleotide polymorphisms (SNPs), SNP-43 and -44, in affecting type 2 diabetes risk (3,4).

Studies subsequent to the original fine mapping and association study have typically genotyped only the three or four markers (SNP-43, -44, and -63 and Indel-19) showing prior association to type 2 diabetes, without regard for

From the <sup>1</sup>Department of Medicine, University of Chicago, Chicago, Illinois; the <sup>2</sup>Department of Human Genetics, University of Chicago, Chicago, Illinois; the <sup>3</sup>Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois; and the <sup>4</sup>Human Genetics Center, University of Texas Health Sciences Center, Houston, Texas.

Address correspondence and reprint requests to Dr. Nancy J. Cox, Department of Human Genetics, University of Chicago, 920 E. 58th St., CLSC 507, Chicago, IL 60637. E-mail: ncox@bsd.uchicago.edu.

Received for publication 5 June 2005 and accepted in revised form 26 August 2005.

L.D.B.-P. is currently affiliated with the Instituto Nacional de Medicina Genómica, Mexico, D.F., Mexico. T.T. is currently affiliated with the Department of Medicine and Molecular Science, Gunma University Graduate School of Medicine, Maebashi, Gunma, Japan.

Additional information for this article can be found in an online appendix at <http://diabetes.diabetesjournal.org>.

LD, linkage disequilibrium; NPL, nonparametric linkage; SNP, single nucleotide polymorphism.

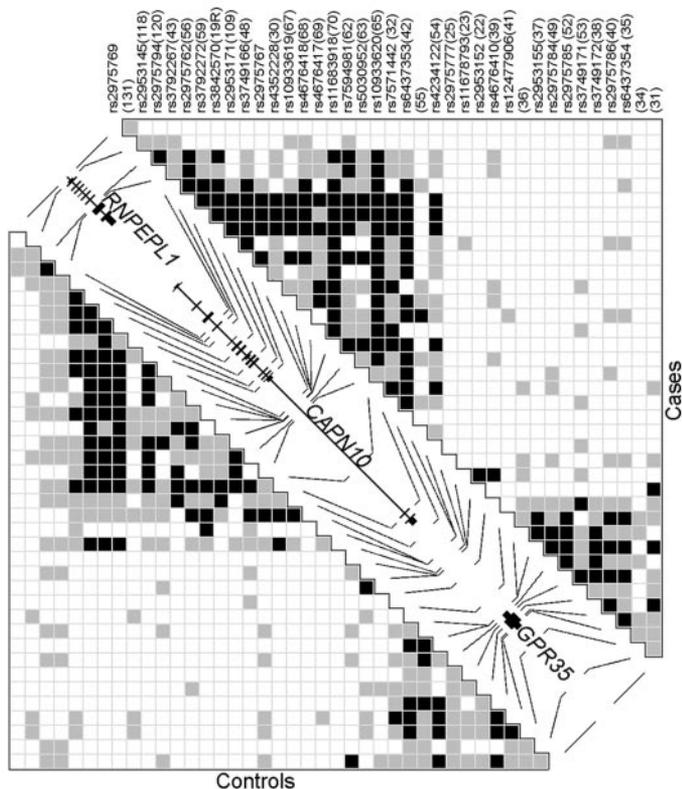
© 2005 by the American Diabetes Association.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

the underlying pattern of linkage disequilibrium (LD) in the region. An allele that causes an increase in disease risk can generate evidence of association (i.e., will be found more frequently in case than in control subjects), but even markers that do not directly affect susceptibility to disease may be associated with disease if the associated marker allele is in LD with the causal allele. LD is usually detectable only between alleles at tightly linked markers, and it decays proportionally to the recombination frequency between the sites. Thus, it is possible the 43–44–19–63 alleles and haplotypes in the *CAPN10* gene might not be the variants affecting susceptibility to type 2 diabetes but may simply be in strong LD with the true nearby disease-causing locus somewhere else in the *NIDDM1* region. To rigorously assess this possibility, we undertook a thorough examination of the patterns and extent of LD in the *NIDDM1* region in a variety of populations.

The pattern of LD in 108 Mexican-American cases and a random sample of 112 control subjects, all from Starr County, Texas, is shown in Fig. 1. Over the densely typed portion of the 2q region surrounding *CAPN10*, two regions of high LD are discernable. The larger contains all of *CAPN10* extending 5' to include the 3' end of *RNPEPL1*, and the smaller contains *GPR35* extending 3' of it. LD decays rapidly in both directions from *CAPN10*, clearly reducing the likelihood that associations between *CAPN10* polymorphisms and type 2 diabetes could be attributed to variation at some distance from *CAPN10*. The high LD within *CAPN10*, however, suggests that it will be difficult to distinguish the variants at *CAPN10* that affect risk of type 2 diabetes from those that are in strong LD. The observations on the association of SNP-44 with type 2 diabetes illustrate this problem. SNP-44 is in perfect LD ( $r^2 = 1$ ) with a missense mutation Thr504Ala (SNP-110) and several other polymorphisms in Mexican Americans and other populations that have been studied (5). Determining the unique relationship of each of these polymorphisms to the etiology of type 2 diabetes will require functional studies.

It was apparent from the beginning of our studies on LD in this region that cases had more extensive LD than control subjects. Figure 1 captures this subtle difference, as the proportion of pairwise comparisons with significant evidence for LD across the 70-kb region is significantly greater in case than control subjects (18.0 and 12.7%, respectively, Pearson's  $\chi^2_1 = 7.686$ ,  $P = 0.006$ ). To localize the observation, we calculated two related measures of LD using a sliding-window approach. The mean significance of LD between the markers "within" the window captures local LD patterns, while the mean significance of LD value

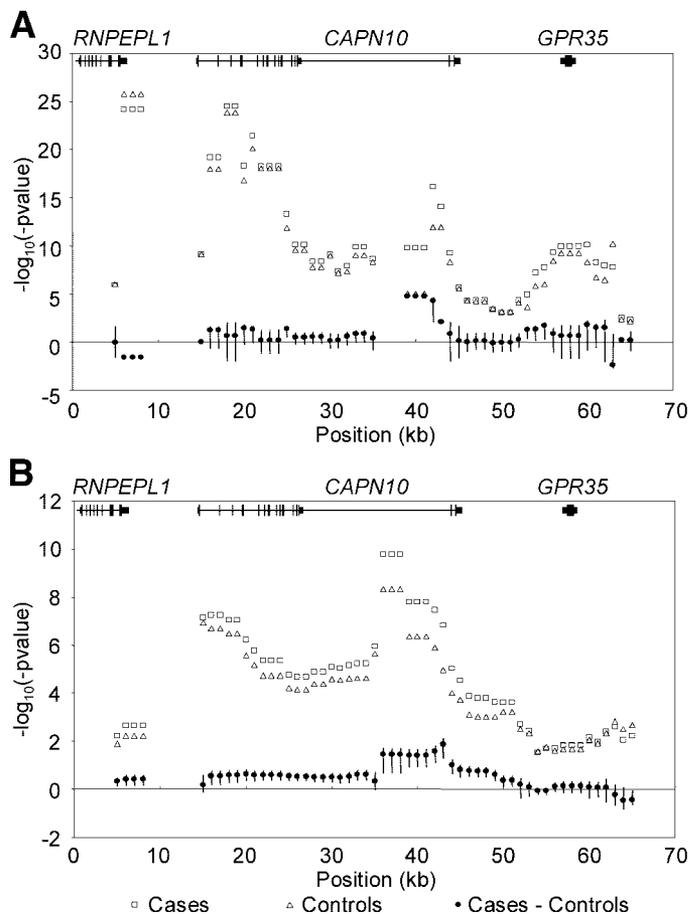


**FIG. 1.** LD comparison of Mexican-American case and control subjects in the 70-kb region surrounding *CAPN10*. Cases are in the upper right, control subjects in the lower left, and an intron/exon map of the 3' end of *RNPEL1*, *CAPN10*, and *GPR35* indicating polymorphism locations is depicted across the diagonal. Marker designations and original UCSNP labels (10) in brackets are listed at the top. Following the definitions of Gabriel et al. (15), black squares represent strong evidence of LD between the pairs of markers, white squares represent strong evidence for recombination between the two markers, and light gray squares represent the remainder of the pairwise comparisons that are uninformative for LD.

between any marker within the window and any marker "outside" of the window informs longer-range LD between the window in question and other nearby regions.

For the within-window significance of LD (Fig. 2A), cases generally have more LD than control subjects across the region, although this difference is significantly different only in the large intron (intron 12) between exons 12 and 13 of *CAPN10*. For the outside-window significance of LD (Fig. 2B), cases again generally have more LD than control subjects across the region, but for this measure, the difference is significant throughout much of *CAPN10*, including the regions containing SNP-44, SNP-43, Indel-19, and SNP-63. The maximal difference is again located in the intron 12. No bias in the number of pairwise comparisons was detected (i.e., no significant correlation exists between number of markers in the window and the case-control LD difference).

The sliding-window analysis of the cases partitioned by nonparametric linkage (NPL) scores (Fig. 3) shows an even greater difference between the two groups. The subset of cases from families with  $NPL \geq 0.7$  at *NIDDM1* have more significant LD within the 10-kb windows in *CAPN10* than cases from families with  $NPL \leq -0.7$ . The greatest difference between the two groups is again found in intron 12. That this difference in the extent of LD between case and control subjects is also observed within the cases alone when partitioned for the evidence of

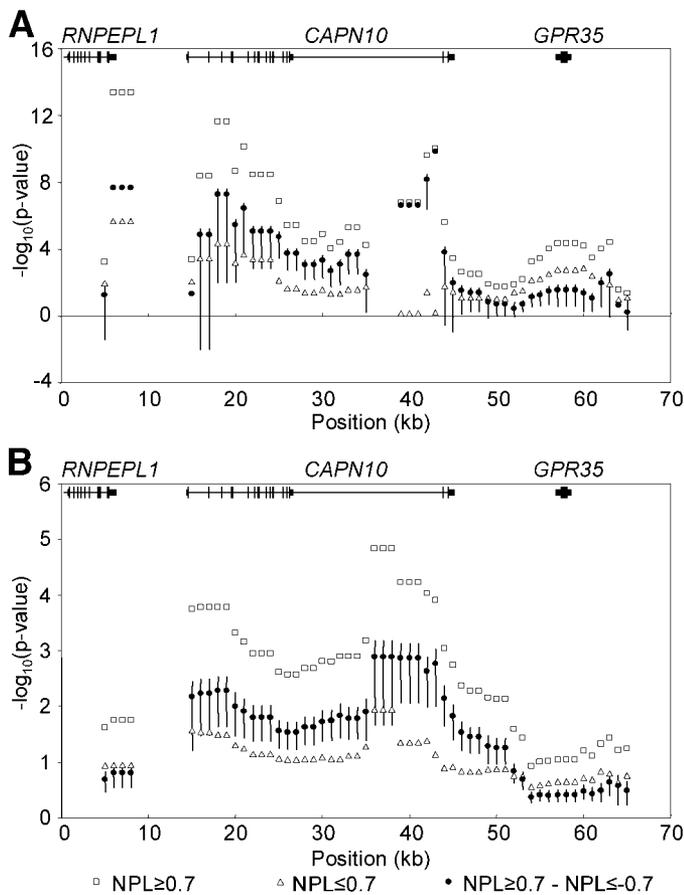


**FIG. 2.** Sliding-window LD difference between case and control subjects in the vicinity of *CAPN10*. Mean significance for all pairwise comparisons within the 10-kb window (A) and between any marker inside the window and any marker outside the window (B). The error bars represent 95% CIs of the difference between the case and control subjects. An intron/exon map of the region is shown at the top of each section.

linkage allows us to rule out the possibility that the former observation is simply an artifact of different amounts of admixture in the case and control subjects, which is consistent with previous studies indicating that genome-wide microsatellite data do not reveal any population substructure in this population (6).

In each of these analyses, the largest LD difference occurs in the middle, or toward the 3' end, of intron 12. None of the polymorphisms in this region yielded significant allelic associations between cases and the random sample or between cases partitioned for the evidence of linkage to *NIDDM1* and random control subjects. Therefore, the LD signal observed here indicates something other than a simple allelic frequency difference between the two groups (either case versus control subjects or  $NPL \geq 0.7$  cases vs.  $NPL \leq -0.7$  cases).

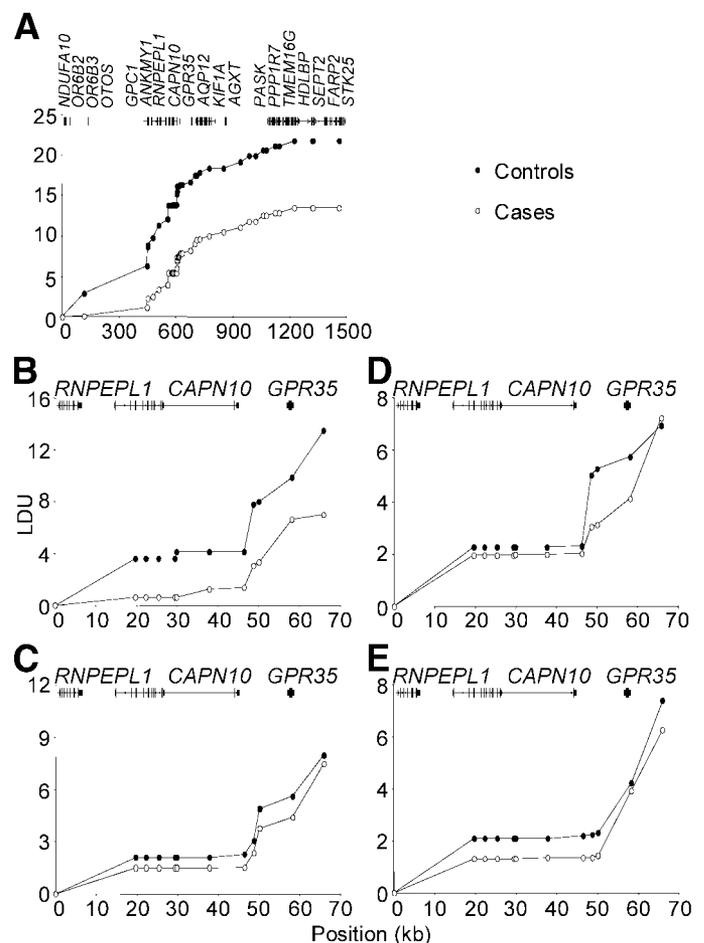
Multilocus LD mapping methods can provide increased power and resolution over single-locus methods in association studies often conducted for the fine-mapping stage of a positional cloning effort (7). Such methods have been applied to *NIDDM1*, implicating the same general region we highlight here (8). However, these approaches are quite computationally intensive. Since we could detect differences in the extent of LD between the case and random samples using even simple pairwise measures of LD, we examined additional approaches utilizing multiple mark-



**FIG. 3.** Sliding-window LD difference between cases partitioned by NPL score in the vicinity of *CAPN10*. Mean significance for all pairwise comparisons within the 10-kb window (A) and between any marker inside the window and any marker outside the window (B). The error bars represent 95% CIs of the difference between the two case sets. An intron/exon map of the region is shown at the top of each section.

ers simultaneously. LD unit plots are a particularly useful way to visualize and quantify differences in the extent of LD. Figure 4A summarizes the extent of LD in the same Mexican-American case and random samples for the 2q region using the LDMap approach, which considers LD between all SNP pairs simultaneously. The LD unit scale on the *y*-axis increases as the extent of LD decreases and is a cumulative unit (as are units of genetic map distance, such as centimorgans) (9). The plot is characterized by a series of plateaus and steps. The former represent regions of high LD and low haplotype diversity (e.g., the *CAPN10* LD block), and the latter are areas of low LD and high recombination. LD drops off dramatically in both directions from this *CAPN10* LD block, indicating the 43–44–19–63 disease-associated polymorphisms are in high LD with markers within this block but are not in LD with other markers in the *NIDDM1* region outside of *CAPN10*. Also, we again observe that cases possess dramatically more LD and therefore less haplotype diversity than the random samples. The significance ( $P = 0.032$ ) of this magnitude in case-control LD unit difference was assessed by permuting the case and control labels.

The observations on LD in case and control subjects in the *NIDDM1* region are, perhaps, not surprising given the very strong evidence for linkage in this region and the fact that case and control subjects are known to have significantly different distributions of haplotype frequencies in



**FIG. 4.** LD unit maps of the *NIDDM1* region. Plotted are the case and control LD units for the 1.5-Mb *NIDDM1* region in Mexican Americans (A) and the 70-kb region surrounding *CAPN10* in four different ethnic populations: African Americans (B), European Americans (C), Chinese Americans (D), and Mexican Americans (E). An intron/exon map of the region is shown at the top of each section.

the local region. To determine whether our observations on the *NIDDM1* region in Mexican Americans could readily be generalized to other populations, the same markers were typed in cases with type 2 diabetes and control subjects in a variety of populations with no reported evidence for linkage to type 2 diabetes in the *NIDDM1* region. The pattern of LD across this region of 2q in these populations is in excellent agreement with that observed in similar populations in the HAPMAP (data not shown). Fig. 4B–E summarizes the extent of LD difference between case and control samples for each of these populations. Although the magnitude of the difference varies, all populations appear to cumulatively have more extensive LD in case than control subjects in this overall region, even though the loci have relatively modest effects on population-specific risk (insufficient to generate evidence for linkage in the population generally and insufficient in the population to generate significant evidence for association at any of the polymorphisms considered individually). Although no single population generated a significant case-control difference in LD units, collectively these observations are significant ( $P < 0.001$ ) for the area between the case and control LD unit plots as assessed by permuting the case and control labels, even when the Mexican-American subjects are excluded ( $P = 0.034$ ).

Our results are consistent with the hypothesis that there

are one or more relatively common alleles increasing risk of type 2 diabetes in the region and that the increased frequency of these alleles in cases, relative to random or control samples, reduces the haplotype diversity in cases (a high proportion of the haplotypes are historically related to each other in the vicinity of the alleles increasing risk and thus increases the extent of LD). The difference in extent of LD between the case and random samples is readily quantified using simple pairwise measures of LD. The use of the LD units approach also clearly highlights CAPN10 as a region with more extensive LD in case than in control subjects in multiple populations. This difference in the extent of LD likely contributes to the localization of a susceptibility gene for type 2 diabetes to this region using more computationally intensive multipoint LD mapping methods (8). The results also suggest that it may be possible to exploit differences in the extent of LD to identify regions harboring susceptibility genes for complex phenotypes.

## RESEARCH DESIGN AND METHODS

The Mexican-American case and random samples include individuals from the vicinity of Starr County, Texas. The 108 case samples represent one affected individual chosen at random from each affected sibpair from the original linkage and positional cloning studies (2,10). The unrelated random samples ( $n = 112$ ) were collected from the same location, and the same random sample was used in the original linkage and positional cloning studies. The European-American, African-American, and Chinese-American case samples ( $n = 87, 66,$  and  $89,$  respectively) and unaffected control samples ( $n = 73, 56,$  and  $53,$  respectively) are from individuals residing in the San Francisco, California, area (11,12).

**Genotyping.** In the original studies in Mexican Americans (10), all polymorphisms were identified through resequencing 10 subjects; once the extent of LD in the region was noted, all polymorphisms with unique patterns and minor allele frequencies  $>0.2$  were genotyped (also by resequencing) in the larger case-control set. Additional markers were chosen for the studies in other populations based on physical coverage and informativeness across the populations and were typed by direct sequencing and Taqman-based assays. **Statistical analyses.** We calculated pairwise measures of LD ( $D'$  and  $r^2$ ) for 69 biallelic Indels/SNPs spanning 1.5 Mb in the Mexican-American case and random control subjects using GOLD (<http://www.sph.umich.edu/csg/abecasis/GOLD/> [13]) and Haploview (<http://www.broad.mit.edu/mpg/haploview/index.php> [14]). We limited our investigation to polymorphisms that, for both the case and control samples, were 1) in Hardy-Weinberg equilibrium, 2)  $\geq 0.10$  in minor allele frequency, and 3) genotyped in  $\geq 70\%$  of the individuals. Some analyses were further restricted to include only those markers ( $n = 38$ ; rs2975769 through UCSNP-31) meeting these criteria from the most densely typed 70-kb region spanning the 3' end of *RNPEPL1*, *CAPN10*, and *GPR35* (online appendix Table A1 [available at <http://diabetes.diabetesjournals.org/>]).

Within this smaller region, we used sliding windows (10-kb window length sliding 1 kb) to compare the difference between mean LD values of the case and control subjects. Mean significance of LD ( $r^2 = \chi^2/N$ ) was calculated for all pairwise comparisons within a window and for all pairwise comparisons in which one marker was in the specified window and the other was outside the specified window. An identical approach was used to examine the LD patterns within the cases partitioned by the NPL score in the family of the affected case from the original genome scan. We focused on cases from families in which the maximum NPL score for the *NIDDM1* region was  $\geq 0.7$  or  $\leq -0.7$ , as 0.7 is half the maximum possible NPL score for an affected sibpair.

LD was also compared between the case and control subjects in several additional ethnic populations using LDMap (<http://cedar.genetics.soton.ac.uk/pub/PROGRAMS/LDMap/> [9]). These analyses were conducted using the full set of 69 markers in the Mexican Americans and with a set of 12 polymor-

phisms meeting the three criteria above for case and control subjects in all four populations studied (online appendix Table A1).

## ACKNOWLEDGMENTS

This research was supported in part by U.S. Public Health Services Grants DK-20595, DK-47486, DK-47487, DK-55889, DK-58026, and HL-007605.

## REFERENCES

- Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A: Loci on chromosomes 2 (*NIDDM1*) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nat Genet* 21:213–215, 1999
- Hanis CL, Boerwinkle E, Chakraborty R, Ellsworth DL, Concannon P, Stirling B, Morrison VA, Wapelhorst B, Spielman RS, Gogolin-Ewens KJ, Shepard JM, Williams SR, Risch N, Hinds D, Iwasaki N, Ogata M, Omori Y, Petzold C, Rietzch H, Schroder HE, Schulze J, Cox NJ, Menzel S, Boriraj VV, Chen X, Lim LR, Lindner T, Mereu LE, Wang YQ, Xiang K, Yamagata K, Yang Y, Bell GI: A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nat Genet* 13:161–166, 1996
- Weedon MN, Schwarz PE, Horikawa Y, Iwasaki N, Illig T, Holle R, Rathmann W, Selisko T, Schulze J, Owen KR, Evans J, del Bosque-Plata L, Hitman G, Walker M, Levy JC, Sampson M, Bell GI, McCarthy MI, Hattersley AT, Frayling TM: Meta-analysis and a large association study confirm a role for calpain-10 variation in type 2 diabetes susceptibility. *Am J Hum Genet* 73:1208–1212, 2003
- Song Y, Niu T, Manson JE, Kwiatkowski DJ, Liu S: Are variants in the *CAPN10* gene related to risk of type 2 diabetes? A quantitative assessment of population and family-based association studies. *Am J Hum Genet* 74:208–222, 2004
- Fullerton SM, Bartoszewicz A, Ybazeta G, Horikawa Y, Bell GI, Kidd KK, Cox NJ, Hudson RR, Di Rienzo A: Geographic and haplotype structure of candidate type 2 diabetes susceptibility variants at the calpain-10 locus. *Am J Hum Genet* 70:1096–1106, 2002
- Voight BF, Cox NJ, Pritchard JK: Detection of population structure in Starr County Mexican-Americans (Abstract). *Am J Hum Genet* 71 (Suppl. 4):352, 2002
- Jorde LB: Linkage disequilibrium and the search for complex disease genes. *Genome Res* 10:1435–1444, 2000
- Zöllner S, Pritchard JK: Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* 169:1071–1092, 2005
- Maniatis N, Collins A, Xu CF, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke X, Morton NE: The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc Natl Acad Sci U S A* 99:2228–2233, 2002
- Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE, del Bosque-Plata L, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL, Bell GI: Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet* 26:163–175, 2000
- Bell GI, Horita S, Karam JH: A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* 33:176–183, 1984
- Xiang KS, Cox NJ, Sanz N, Huang P, Karam JH, Bell GI: Insulin-receptor and apolipoprotein genes contribute to development of NIDDM in Chinese Americans. *Diabetes* 38:17–23, 1989
- Abecasis GR, Cookson WO: GOLD: graphical overview of linkage disequilibrium. *Bioinformatics* 16:182–183, 2000
- Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265, 2005
- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: The structure of haplotype blocks in the human genome. *Science* 296:2225–2229, 2002