Loci on chromosomes 2 (*NIDDM1*) and 15 interact to increase susceptibility to diabetes in Mexican Americans

Nancy J. Cox[1], Mike Frigge[4], Dan L. Nicolae[2,4], Patrick Concannon[5], Craig L. Hanis[6], Graeme I. Bell[1,3] & Augustine Kong[2,4]

Complex disorders such as diabetes, cardiovascular disease, asthma, hypertension and psychiatric illnesses account for a large and disproportionate share of health care costs, but remain poorly characterized with respect to aetiology. The transmission of such disorders is complex, reflecting the actions and interactions of multiple genetic and environmental factors. Genetic analyses that allow for the simultaneous consideration of susceptibility from multiple regions may improve the ability to map genes for complex disorders[1,2], but such analyses are currently computationally intensive and narrowly focused. We describe here an approach to assessing the evidence for statistical interactions between unlinked regions that allows multipoint allele-sharing analysis to take the evidence for linkage at one region into account in assessing the evidence for linkage over the rest of the genome. Using this method, we show that the interaction of genes on chromosomes 2 (*NIDDM1*) and 15 (near *CYP19*) makes a contribution to susceptibility to type 2 diabetes in Mexican Americans from Starr County, Texas.

The correlation in scores assessing the evidence for linkage within families (the family specific $\overline{Z}$ scores calculated on the basis of allele-sharing[3], which we will refer to as family NPL scores) can be used to determine preliminary evidence for statistical interaction between unlinked regions. The use of the correlation has been suggested to assess potential interaction between unlinked regions, but in the context of parametric lod scores[4]. Lod scores were calculated assuming a specific parametric model, and thus both lod scores and correlations between lod scores were sensitive to misspecification of the underlying model. Calculating the correlation in family NPL scores from unlinked regions should provide a more appropriate test of the hypothesis that allele sharing at unlinked regions is independent within families—both because it is based on allele-sharing rather than a specific (and possibly incorrect) parametric model, and because the family NPL scores have a more appropriate distribution (mean 0, variance 1, with no missing information) than lod scores for calculation of correlations.

Unless the regions chosen for study actually contain loci that contribute susceptibility to disease, there is no expectation that family NPL scores from unlinked regions will be correlated, even if the regions are selected because they show some evidence for linkage. However, there is not always a simple correspondence between the biological interactions of genes and the statistical interactions that can be detected in the sample sizes commonly collected for linkage studies. For example, although some models of epistatic interaction can generate positive correlations between family NPL scores from the regions to which the loci map, many models of biological interaction would not generate detectable correlations. For substantial positive correlation to exist, not only do the two genes need to have some form of interaction, but there also need to be additional genetic or environmental factors potentially common to the affected family members that can independently lead to the disease. Many of the previously proposed two-locus models, whether parametric or non-parametric, have implicitly assumed that the phenotypes of affected pairs of relatives are conditionally independent—that is, given their genotypes at the two gene locations, their phenotypes are otherwise independent[5,6]. If other genes or environmental factors contribute to disease (beyond the two genes explicitly included in the models), the phenotypes of the relatives will often be dependent, even after conditioning on the genotypes of the two contributing loci in the model. This additional dependence between the phenotypes is analogous to what was called 'residual correlation' in the context of quantitative traits[7]. It has been shown for affected sib pairs that if conditional independence is assumed, then multiplicative penetrances (that can be used to model epistasis) lead to a correlation of 0 in allele sharing at the two susceptibility loci, which would also be a correlation of 0 for the family NPL scores. In contrast, without the restriction of conditional independence, one can construct examples that have substantial positive correlation. Moreover, negative correlations between regions can be generated when non-overlapping sets of families provide evidence for linkage due to genetic heterogeneity, in the absence of biological interactions between the susceptibility loci from these regions. Thus, finding significant correlations between family NPL scores at unlinked regions provides additional evidence that loci from those regions contribute to disease susceptibility and generates insight into the models most consistent with the type of correlation (positive or negative) observed.

Once preliminary studies provide evidence for statistical interaction between regions, it is possible to incorporate linkage evidence from one region in assessing evidence for linkage at a second region (or multiple regions) by weighting families according to their evidence for linkage. The multipoint allele-sharing approach described by Kruglyak *et al.*[3] and extended by Kong and Cox[8] to efficiently use incomplete information was designed to allow families to be weighted individually, but these original implementations assigned each family equal weight. Although traditional non-parametric analyses usually weighted pairs (rather than families) equally, the 'optimal' weighting function is dependent on the true model of disease susceptibility transmission[8]. As a general strategy, we choose the scoring function and weighting factors to be robust over as broad a range of models of transmission as possible. In the absence of information, such as mutation status at a known contributory gene that differentiates families (or pairs), equal weighting may be appropriate, but when such information is available, it can be used by altering the weights for families. Our newest extension (GENEHUNTER-PLUS v2.0) allows users to

Departments of [1]Medicine, [2]Statistics and [3]Biochemistry and Molecular Biology, and Howard Hughes Medical Institute, The University of Chicago, 5841 South Maryland Avenue, Chicago, Illinois 60637, USA. [4]deCODE Genetics, Lynghals 1, 110 Reykjavik, Iceland. [5]Virginia Mason Research Center, 1000 Seneca Street, Seattle, Washington 98101, and Department of Immunology, University of Washington School of Medicine, Seattle, Washington 98195, USA. [6]Human Genetics Center, The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA. Correspondence should be addressed to N.J.C. (e-mail: nancy@hhmi.bsd.uchicago.edu).

specify individual weights for each family based, for example, on pedigree structure, number of affecteds and/or their evidence for linkage at a particular location. We can model positive interactions (such as epistasis) by assigning weight 0 to families with 0 or negative linkage scores and weight 1 to families with positive linkage scores (weight$_{0\text{-}1}$), and can model heterogeneity by assigning weight 1 to families with negative linkage scores and weight 0 to families with 0 or positive linkage scores (weight$_{1\text{-}0}$). We can also construct more complex family specific weights proportional to the evidence for linkage (weight$_{\mathrm{PROP}}$).

A previous genome-wide screen for type 2 diabetes genes in Mexican Americans localized a major susceptibility gene, *NIDDM1*, to the *D2S125–D2S140* region of chromosome 2 (multipoint lod score=4.03; refs 9,10). This was the only region in the primary analyses to meet genome-wide criteria for significance[11]. Animal studies have suggested that type 2 diabetes may result, at least in part, from epistatic interactions between genes[12,13]. In addition, some alleles at genes associated with monogenic forms of diabetes such as maturity onset diabetes of the young (MODY, a genetically heterogeneous form of diabetes characterized by autosomal dominant inheritance, with onset usually before age 25 and pancreatic β-cell dysfunction) may cause a form of diabetes that resembles type 2 diabetes[14,15]. Thus, we examined the evidence for statistical interactions between *NIDDM1* and the ten other autosomal regions providing nominal evidence for linkage (P<0.05; ref. 9), as well as five regions containing genes for MODY (Table 1). Two regions, *CYP19* on chromosome 15 and the gene encoding hepatocyte nuclear factor (HNF)-1α/MODY3 (*TCF1*) on chromosome 12, showed significant correlations between their family NPL scores and the family NPL scores at *NIDDM1*, even after Bonferroni correction for the number of correlations examined.

Determining the significance of apparent interactions requires care. The nominal P-values associated with the sample correlations are calculated using the Pearson's correlation test (t-test). The significance associated with the increased lod, when evidence for linkage at a particular location is taken into account using family-specific weights, can be determined either by simulation or by using a conservative χ^2 test with 1 degree of freedom (d.f.) as follows. If we consider a more general 1-d.f. family of weights in which weight$_{0\text{-}1}$ and weight$_{1\text{-}0}$ are the two extremes, then the increase over baseline of the lod for the family weighting yielding the maximum lod multiplied by 2 log(10) is asymptotically distributed as a χ^2 with 1 d.f. under the null hypothesis of no interaction. The test is conservative because we are not actually maximizing the lod with respect to the weighting factors and currently consider only a few family specific weights. Interpretation of such studies, however, still requires taking multiple comparisons into account. To limit the Bonferroni adjustment, it seems prudent to focus on the top signals from the primary linkage analysis and perhaps a small number of candidate regions. Even with this adjustment, such secondary analyses may increase the overall false positive rate because they are designed to strengthen the support for regions that do not themselves meet genome-wide criteria for significance. Given that, and the absence of information on the *a priori* likelihood of such interactions, it is appropriate to use more stringent criteria for determining significance, that is, P-values must be lower than 0.01 instead of 0.05. The evidence for interaction between the *CYP19* and *NIDDM1* regions meets these criteria after the Bonferroni adjustment, but that between *NIDDM1* and *TCF1* does not (Table 1). More research will be necessary to determine whether such statistical interactions will be common in complex traits, and how criteria that have been suggested for assessing genome-wide significance[11] should be modified when the evidence for linkage at multiple susceptibility loci is considered simultaneously. In addition, the methods described in this manuscript are best suited for data sets consisting of many small families, such as these MA sibships. For example, the P-values (Table 1) are based on large sample approximations, which can be shown by simulation to be very accurate. P-values obtained in a similar manner can be much less reliable, however, when the data consist of only a moderate number of families, especially if these families vary in size and structure. Moreover, in such situations this method may lack power. We are currently working on refinements of the method that make it more suited for data of the latter type.

Table 1 • Interactions between the NPL scores at *NIDDM1* and other regions

Region	Correlation	Corrected *P*-value*	Baseline lod	*NIDDM1*- weighted lod
CYP19	0.288	2.1×10^{-3}	1.27	4.00 (weight$_{0\text{-}1}$)
D7S502	0.180	0.29	0.76	1.31 (weight$_{0\text{-}1}$)
D3S3054	0.098	ns	0.81	–
D2S377	0.085	ns	1.28	1.50 (weight$_{0\text{-}1}$)
D15S104	0.066	ns	0.93	1.20 (weight$_{0\text{-}1}$)
D3S2452	0.031	ns	1.24	–
D2S441	0.027	ns	0.78	–
D12S379	–0.012	ns	0.68	–
D11S1314	–0.059	ns	0.78	–
D17S1298	–0.172	0.39	0.73	1.21 (weight$_{1\text{-}0}$)
GCK	0.124	ns	0.01	0.26 (weight$_{0\text{-}1}$)
TCF1	–0.228	0.04	0.01	1.03 (weight$_{1\text{-}0}$)
TCF2	0.010	ns	0.00	–
HNF4A	0.003	ns	0.38	–
IPF1	–0.187	0.24	0.32	1.11 (weight$_{1\text{-}0}$)

*P-values corrected by multiplying the nominal P-value by the number of correlations examined (15). Numerical P-values are given only for those loci in which the uncorrected P-values were nominally significant (P<0.05). Weighted lod scores are given only for those loci in which a weighted lod score was greater than the baseline lod score. The marker used for *TCF1* was GATA32A10, for *TCF2* was *D17S1788*, for *HNF4A* was *ADA* and for *IPF1* was *D13S221*.

The lod in the *CYP19* region was 1.3 in baseline analysis but increased to 4.0 when the families were weighted by their evidence for linkage at *NIDDM1* using weight$_{0\text{-}1}$, and to 4.1 when families were weighted by their evidence for linkage using weight$_{\mathrm{PROP}}$ (Fig. 1*a*). Note that the more distal region of chromosome 15 with similar baseline evidence for linkage does not show a comparable increase in lod when the evidence for linkage at *NIDDM1* is taken into account. The lod score at *NIDDM1* rises from 4.0 in the baseline analyses to 5.6 when families were weighted by their evidence for linkage at *CYP19* using weight$_{0\text{-}1}$ and to 7.3 using weight$_{\mathrm{PROP}}$ (Fig. 1*b*). In simulations conducted to determine the significance of the increase in the lod at *CYP19* from 1.3 to more than 4.0, we found that none of 10,000 replicates from a simulation in which 95 families (the number of families in these data with positive NPL scores at *NIDDM1*) were randomly chosen and analysed for the actual chromosome 15 data had a lod score as large as 4.0, although 4 (of 10,000) yielded lods between 3.5 and 4.0. Thus, a reasonable estimate of the nominal significance of the increase in lod from 1.3 to 4.0 based on simulation is 0.0001, or 0.0015 corrected for the number of regions examined. The conservative χ^2 test described above would be calculated as 2 log(10) (4.0–1.3)=12.4, giving a P-value of 0.0004, before adjusting for multiple comparisons. The P-value obtained in this way is indeed comparable with the P-values obtained from the correlation test and the simulations, but is more modest (conservative) because we have not actually maximized the evidence for linkage over a range of family specific weights (for example, the lod score for weight$_{\mathrm{PROP}}$ is 4.1).

The *CYP19* region of chromosome 15 was the only location besides *NIDDM1* to be replicated (P<0.05) in a smaller, independent sample of Mexican American families[9]. This, as well as the evidence for statistical interaction between these regions, suggests that in collections of Mexican American families similar in size to that in the original genome scan, the evidence for linkage in analyses of

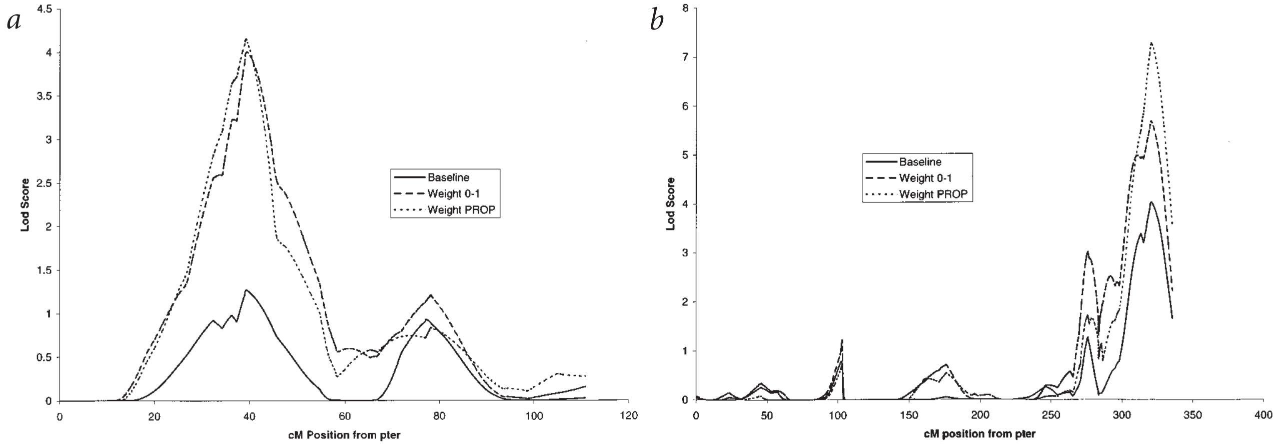

Fig. 1 Interaction between *NIDDM1* and *CYP19*. ***a***, Multipoint allele-sharing analysis of chromosome 15 weighted by the evidence for linkage at *NIDDM1* on chromosome 2. ***b***, Multipoint allele-sharing analysis of chromosome 2 weighted by the evidence for linkage at *CYP19* on chromosome 15.

chromosome 15 might sometimes be more prominent than that for *NIDDM1*, and that in many such collections, the signals from both regions may be comparable and only modest unless the interaction is properly taken into account. Thus, it is possible that some of the difficulties recognized in replicating results obtained in genome scans for complex disorders[16] might be alleviated by conducting analyses to identify potential interactions. Finally, the improvement in localization offered by linkage analyses that allow for the contributions of multiple susceptibility loci may be critical to the successful positional cloning of genes for complex disorders.

Methods

We report results of analyses on genome scan data from 524 autosomal markers genotyped in 408 individuals from 170 Mexican American sibships previously described[9]. These data included 121 sibships with 2 affected sibs, 34 sibships with 3 affected sibs, 12 sibships with 4 affected sibs, 2 sibships with 5 affected sibs and 1 sibship with 6 affected sibs. A region near *D2S140* provided strong evidence for linkage to type 2 diabetes in Mexican Americans (*NIDDM1*, lod=4.03, $P<8\times10^{-6}$). We calculated correlations using the family NPL scores from this region with each of the other ten autosomal regions providing nominally significant ($P<0.05$, lod>0.59) evidence for linkage. We also calculated correlations between the family NPL scores at *NIDDM1* and five regions from which MODY genes have been characterized (*GCK*, ref. 17; *TCF1*, ref. 18; *TCF2*, ref. 19; *HNF4A*, ref. 20; and *IPF1*, ref. 21). In addition, we weighted the contribution from families according to their evidence for linkage at *NIDDM1* in linkage analyses on these 15 regions. In constructing the weight$_{0\text{-}1}$ family weighting, we assigned families weight 0 if their NPL score at *NIDDM1* (*D2S140*, the location providing the strongest evidence for linkage in the *NIDDM1* region) was 0 or negative and weight 1 if their NPL score at *NIDDM1* was positive. In constructing the weight$_{1\text{-}0}$ family weighting, we assigned families weight 1 if their NPL score at *NIDDM1* was negative and weight 0 if their NPL score at *NIDDM1* was 0 or positive. In constructing the weight$_{\mathrm{PROP}}$ family weighting, the weight for families with positive NPL scores was the family NPL score, and the weight for families with negative NPL scores was 0. In all of these analyses, we used the exponential model[8] to calculate the lod scores. We used simulation studies to assess the significance of the increase in lod score at *CYP19* with the weight$_{0\text{-}1}$ family weighting on *NIDDM1*. At *D2S140* there were 95 families with positive NPL scores and 75 families with 0 or negative NPL scores. Simulations based on the weight$_{0\text{-}1}$ or weight$_{1\text{-}0}$ family weighting can be rapidly conducted using the extension which allows families to be weighted individually. We conduct the basic GENEHUNTER analysis only once on the actual data (in this case, from chromosome 15), and then many replicate weighting files are generated randomly (in this example, 95 randomly chosen families are given weight 1 and the remaining 75 families are given weight 0) and used to calculate the final lod scores. The software is distributed as GENEHUNTER-PLUS (version 2.0 or later) and is available via anonymous ftp (galton.uchicago.edu on the /pub/kong directory). The allele-sharing method used has been described[8] and version 2.0 introduces an option to provide a family specific weight in the lod score computation, but does not include simulation.

Acknowledgements

This research was supported in part by National Institutes of Health grants GM-46800, DK-20595, DK-47481, DK-47486, DK-47487 and DK-47494.

Received 16 June 1998; accepted 6 January 1999.

1. Schork, N.J., Boehnke, M., Terwilliger, J.D. & Ott, J. Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *Am. J. Hum. Genet.* **53**, 1127–1136, 1993.
2. Buhler, J., Owerbach, D., Schaffer, A.A., Kimmel, M. & Gabbay, K.H. Linkage analyses in type I diabetes mellitus using CASPAR, a software and statistical program for conditional analysis of polygenic diseases. *Hum. Hered.* **47**, 211–222 (1997).
3. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**, 1347–1363 (1996).
4. MacLean, C.J., Sham, P.C. & Kendler, K.S. Joint linkage of multiple loci for a complex disorder. *Am. J. Hum. Genet.* **53**, 353–366 (1993).
5. Risch, N. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am. J. Hum. Genet.* **46**, 229–241 (1990).
6. Cordell, H.J., Todd, J.A., Bennett, S.T., Kawaguchi, Y. & Farrall, M. Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of *IDDM2* and *IDDM4* with *IDDM1* in type 1 diabetes. *Am. J. Hum. Genet.* **57**, 920–932 (1995).
7. Bonney, G.E. On the statistical determination of major gene mechanisms in continuous human traits. *Am. J. Med. Genet.* **18**, 731–749 (1984).
8. Kong, A. & Cox, N.J. Allele-sharing models: LOD scores and accurate linkage tests. *Am. J. Hum. Genet.* **61**, 1179–1188 (1997).
9. Hanis, C.L. *et al.* A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nature Genet.* **13**, 161–166 (1996).
10. Bell, G.I. *et al.* Genetics of NIDDM in the Mexican American of Starr County, Texas: an update. *Diabetes Rev.* **5**, 277–283 (1997).
11. Lander, E.S. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* **11**, 241–247 (1995).
12. Terauchi, Y. *et al.* Development of non-insulin-dependent diabetes mellitus in the double knockout mice with disruption of insulin receptor substrate-1 and β cell glucokinase genes. *J. Clin. Invest.* **99**, 861–866 (1997).
13. Brunning, J.C. *et al*. Development of a novel polygenic model of NIDDM in mice heterozygous for IR and IRS-1 null alleles. *Cell* **88**, 561–572 (1997).
14. Mahtani, M.M. *et al.* Mapping of a gene for type 2 diabetes associated with an insulin secretion defect by a genome scan in Finnish families. *Nature Genet.* **14**, 90–94 (1996).
15. Iwasaki, N. *et al.* Mutations in the hepatocyte nuclear factor-1α/MODY3 gene in Japanese subjects with early- and late-onset NIDDM. *Diabetes* **46**, 1504–1508 (1997).
16. Suarez, B.K., Hampe, C.L. & Van Eerdewegh, P. Problems of replicating linkage claims in psychiatry. in *Genetic Approaches to Mental Disorders* (eds Gershon, E.S. & Cloninger, C.R.) 23–46 (American Psychiatric Press, London, 1994).
17. Froguel, P. *et al.* Familial hyperglycemia due to mutations in glucokinase. Definition of a subtype of diabetes mellitus. *N. Engl. J. Med.* **328**, 697–702 (1993).
18. Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-1α gene in maturity onset diabetes of the young (MODY3). *Nature* **384**, 455–458 (1996).
19. Horikawa, Y. *et al.* Mutation in hepatocyte nuclear factor-1β gene (*TCF2*) associated with MODY. *Nature Genet.* **17**, 384–385 (1997).
20. Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-4α gene in maturity onset diabetes of the young (MODY1). *Nature* **384**, 458–460 (1996).
21. Stoffers, D.A., Ferrer, J., Clarke, W.L. & Habener, J.F. Early-onset type-II diabetes mellitus (MODY4) linked to *IPF1. Nature Genet.* **17**, 138–139 (1997).