A SEARCH FOR VARIANTS ASSOCIATED WITH YOUNG-ONSET TYPE 2 DIABETES IN AMERICAN INDIANS IN A 100K GENOTYPING ARRAY

Robert L. Hanson1
Clifton Bogardus1
David Duggan2
Sayuko Kobes1
Michele Knowlton2
Aniello M. Infante1
Leslie Marovich2
Deb Benitez2
Leslie J. Baier1
William C. Knowler1

1Phoenix Epidemiology and Clinical Research Branch
National Institute of Diabetes and Digestive and Kidney Diseases
Phoenix, Arizona, USA

2Translational Genomics Research Institute
Phoenix, Arizona, USA

Running Title: Genetic Variants and Young-Onset Type 2 Diabetes

Correspondence:
Robert L. Hanson
Phoenix Epidemiology and Clinical Research Branch
National Institute of Diabetes and Digestive and Kidney Diseases
550 E. Indian School Road
Phoenix, AZ 85014
rhanson@phx.niddk.nih.gov (e-mail)

Received for publication 3 April 2007 and accepted in revised form 5 September 2007.

Additional information for this article can found in an online appendix at http://diabetes.diabetesjournals.org.

Copyright American Diabetes Association, Inc., 2007
ABSTRACT

Objective: To identify genetic variants in linkage disequilibrium with those conferring diabetes susceptibility, a genome-wide association study for young-onset diabetes was conducted in an American Indian population.

Research Design and Methods: Data come from 300 cases with type 2 diabetes with onset < age 25 yrs and 334 nondiabetic controls who were > age 45 yrs. To provide for tests of within-family association, 121 nondiabetic siblings of cases were included along with 140 diabetic siblings of controls (172 sibships). Individuals were genotyped on the Affymetrix 100K array, resulting in 80,044 useable SNPs. SNPs were analyzed for within-family association and for general association in cases and controls and these tests were combined by Fisher’s method, with priority given to the within-family test.

Results: There were more SNPs with low p-values than expected theoretically under the global null hypothesis of no association, and 128 SNPs had evidence for association at p<0.001. The association of these SNPs with diabetes was further investigated in 1207 diabetic and 1627 nondiabetic individuals from the population study who were not included in the genome-wide study. SNPs from 10 genomic regions showed evidence for replication at p<0.05. These included SNPs on chromosome 3 near ZNF659, chromosome 11 near FANCF, chromosome 11 near ZBTB15 and chromosome 12 near SENP1.

Conclusions: These studies suggest several regions where marker alleles are potentially in linkage disequilibrium with variants that confer susceptibility to young-onset type 2 diabetes in American Indians.
INTRODUCTION

Type 2 diabetes is substantially influenced by genetic factors, as indicated by studies of familial aggregation (1-3) and twins (4-6); however the identity of most of the specific variants that influence diabetes susceptibility remains unknown. Consistent, albeit modest, associations have been observed with alleles at \textit{PPARG} and \textit{KCNJ11} (7-8). Recently, Grant et al identified an association of type 2 diabetes with alleles in \textit{TCF7L2} (9). This association, which is of greater magnitude than those for the other polymorphisms, has been widely replicated (e.g., 10-13). These variants explain only a small fraction of the genetic contribution to type 2 diabetes, so it is likely that additional variants remain unidentified. A number of genome-wide linkage studies have been conducted (14-15) and, while these have revealed several putative susceptibility loci, the variants responsible have not been definitively identified.

Recent advances in technology have produced methods for large-scale genotyping of dense panels of single nucleotide polymorphisms (SNPs). Thus, genome-wide association studies are feasible and these provide another, potentially powerful, approach for detection of novel variants that influence susceptibility to diabetes. The present study represents a genome-wide association study of type 2 diabetes in the Pima Indians- this population has a high prevalence of obesity and diabetes and when diabetes occurs it is overwhelmingly, if not exclusively, type 2, even in childhood (16). Analyses of the familial pattern of diabetes in this population show that young-onset diabetes is particularly familial and that genetic determinants are likely to influence age at onset of diabetes (3,17-18); therefore the present study was designed to detect variants associated with young-onset diabetes.

METHODS

Participants: The present data come from participants in a longitudinal study conducted in the Gila River Indian Community in central Arizona (19). In this study, community residents who are $\geq$ 5 years old are invited to a research examination every 2 years. These examinations include a 75g oral glucose tolerance test. Diabetes is diagnosed if the fasting plasma glucose concentration is $\geq 7.0$ mM, the two-hour plasma glucose concentration is $\geq 11.1$ mM (20) or if the
diagnosis is made during routine clinical care (19). DNA has been extracted from blood leukocytes. To detect variants associated with young-onset diabetes, individuals were selected from the extremes of the age-of-onset distribution. Thus, 300 “cases” were selected who had developed type 2 diabetes before the age of 25 years; for comparison, 334 “controls” were selected who were nondiabetic and ≥45 years old when last examined. All individuals were full-heritage American Indian and all individuals fulfilling criteria were selected regardless of affection status of other family members.

Although the case-control approach is potentially powerful for detection of associated variants (21,22), it is liable to spurious results due to population stratification. Therefore, to allow for within-family association tests that are robust to such confounding, siblings of these cases or controls who defined discordant sibling pairs were selected. Thus, 121 siblings of cases were included who were nondiabetic when last examined and whose age was older than that at which the youngest onset of diabetes in the family occurred. Likewise, 140 diabetic siblings of controls whose age of onset was younger than the age of the oldest control in the family were included. These individuals constituted 340 discordant sibling pairs in 172 potentially informative sibships.

Population-Based Association Studies: To examine the potential importance of associated markers on a population basis a population-based sample of individuals from the longitudinal study was selected for genotyping with selected markers. This sample consisted of all participants from the longitudinal study with available DNA whose heritage was full Pima and/or Tohono O’odham (a closely related tribe); 1561 of these individuals had diabetes, while 1940 were nondiabetic. There were 2834 individuals who were not included in the genome-wide association study and analyses in these individuals were used to provide a replication of results from the genome-wide study. Analyses in all 3501 individuals from the population study (who constituted 1880 sibships) were used to determine population-based parameters. Because of differences in selection criteria, there were 228 individuals in the genome-wide study who were not in the population-based study. Characteristics of individuals in the various studies are shown in Table 1.
Genotyping: DNA was isolated using a proteinase K high salt ethanol precipitation method. Prior to genotyping for the genome-wide association study, DNA was purified using Montage PCR plates (Millipore). Individuals were genotyped using the Affymetrix 100K Human Mapping array, which contains 115,810 SNPs with known positions on the autosomal and X chromosomes, according to the manufacturer’s protocol (Affymetrix, Santa Clara, CA). Genotypes were generated using the BRLMM algorithm (23). 50 individuals were genotyped in duplicate and 5122 SNPs were excluded from analysis which had discrepancy rates of >2.7% or which produced valid genotypes in <85% of individuals. The mean genotypic concordance rate among the 50 pairs of duplicates was 99.5%. Since spurious associations may occur more frequently with very rare alleles, a further 28,215 SNPs with minor allele frequency <1% were excluded. Hardy-Weinberg equilibrium was tested among all genotyped individuals, using a continuity correction to produce a better approximate statistic for rarer alleles (24). Since systematic genotyping errors may produce severe violations of Hardy-Weinberg equilibrium, a further 2429 SNPs were excluded that deviated from Hardy-Weinberg equilibrium at p<0.001. Thus, the present genome-wide association analyses include results for 80,044 SNPs. Family relationships (parents, siblings) were confirmed by comparison of the observed proportion of alleles identical by state for these markers with that expected. SNPplex (Applied Biosystems, Foster City, CA) was used to genotype individuals in the follow-up population-based studies. Physical positions are given according to NCBI Build 35.

Statistical Analysis: Analyses were performed using the SAS package (SAS Institute, Cary, NC). The association between genotype and case-control status was assessed with logistic regression. Genotype was analyzed as a numeric variable representing the number (0,1,2) of copies of a given allele. For X chromosome markers outside the pseudoautosomal regions men were coded as homozygous. To account for the inclusion of multiple individuals in the same sibship, data were analyzed using a modified regressive model in which, for each individual, the prevalence of case status among all of his or her siblings was included as a covariate (25). This produces an approximation to the regressive model of Bonney, in which covariates
are used to model the residual phenotypic correlation among siblings (26). To provide for a specific test of within-family association in sibships discordant for diabetes, data were also analyzed with conditional logistic regression (27). All models included sex as a covariate and the likelihood ratio test was used to assess statistical significance. Odds ratios were calculated from regression coefficients. In the event that one allele is absent from cases or controls, parameter estimates will not reliably converge, but the likelihood ratio statistic is still approximately correct. In these cases, the estimated odds ratio is infinite; therefore we report an odds ratio of infinity for these SNPs.

To summarize the results of case-control and within-family analyses, p-values for the two tests were combined with Fisher’s method (28). To maintain robustness to population stratification, priority was given to the p-value for the within-family test (P_{WI}). This was accomplished by calculating a one-sided p-value for the case-control test (P_{CC}) for an association in the same direction as that observed in the within-family analysis. The contribution of the case-control result (P_{CC*}) to the combined test statistic was taken as the maximum of P_{WI} and 1-(1-P_{CC})^2, where the latter term employs a correction to partially account for the fact that the two tests have been performed in some of the same individuals. The combined test statistic was calculated as:

\[ \chi^2_{CC-WI} = -2\ln(P_{WI}) - 2\ln(P_{CC*}) \]

Simulation studies show that this method augments power of the within-family test by using information from the general test, while maintaining robustness to population stratification (29). If P_{WI} and P_{CC*} were independent, the p-value associated with \( \chi^2_{CC-WI} \) on 4 df would have a uniform distribution under the null hypothesis (28). However, because P_{CC*} is “truncated” to be \( \geq P_{WI} \) and because of correlation among the tests, this “nominal” p-value for \( \chi^2_{CC-WI} \) does not have a uniform distribution. To generate a p-value corrected for these distributional deviations, 50 million pairs of standardized test statistics, representing within-family and case-control logarithms of the odds ratio divided by their standard errors, were simulated from a bivariate distribution with correlation of 0.32 (the observed correlation among these statistics for 80,044 SNPs). In these simulated data, \( \chi^2_{CC-WI} \) was calculated for each
pair of tests and the p-value for the value of $\chi^2_{CC-WI}$ observed for a given SNP was taken as the proportion of replicates at which the test statistic for these simulated values exceeded the observed value.

**Permutations:** To assess experiment-wide statistical significance, affection status was permuted across individuals and analyses of the association of genotypes with permuted affection status were repeated genome-wide. 200 permutations were thus analyzed. To maintain the familial nature of the study, all data for a sibship were permuted in tandem (across sibships of the same size) and affection status of individuals, including covariates, was then further permuted within the sibship. The proportion of permutations in which the maximum of $\chi^2_{CC-WI}$ exceeded the value observed for a given SNP was taken as the experiment-wide p-value for that SNP. To compare the global distribution of test statistics with that expected under the null hypothesis of no association with any marker, the signed Kolmogorov-Smirnov deviation ($d$) was calculated for the observed distribution in comparison with the null distribution from each permutation (30). This statistic is the maximum deviation, over the full distribution of p-values, of the observed cumulative distribution function for any given p-value ($f(p)_{obs}$) from the value expected under the null ($f(p)_{null}$):

$$d = i \times \max(|f(p)_{obs} - f(p)_{null}|)$$

where $i$ is an indicator variable that is 1 if $f(p)_{obs} \geq f(p)_{null}$ and -1 if $f(p)_{obs} < f(p)_{null}$ at the point of maximum deviation. Under the global null hypothesis, the expected value of $d$ is 0; thus the test that the mean $d=0$ across all permutations provides an empirical test of the global null hypothesis. Values of $d>0$ indicate a shift toward lower p-values in the observed compared with the null distribution. Since this analysis assumes exchangeability among men and women, it was restricted to the 78,568 autosomal SNPs.

**Analysis of Population Study:** SNPs that had the strongest associations in the genome-wide study were genotyped and analyzed in the population sample to examine the association on a population basis. To assess replication in a separate group of individuals, the association of genotype with prevalence of diabetes at the last available examination was analyzed among the 2834 individuals who were not included in the genome-wide association study. These analyses were conducted using
logistic regression models which were fit with generalized estimating equations to account for correlation among siblings (31). Within-family tests of association were also conducted using a modification of the method of Abecasis et al which partitions the association into between- and within-family components (32). In this method, the sibship mean of the numeric variable representing genotype is used to assess the between-family effects and each individual’s deviation from this mean is used to assess within-family effects. The p-value for this within-family coefficient (PWI) was further combined with the general association result using the modification of Fisher’s method described above to produce a summary test of these two effects - with the difference that PWI was calculated in a one-sided fashion to ensure that claims of replication would only be made if the direction of association was the same as that observed in the genome-wide association study. The distribution corrected p-value was calculated as described above by simulation of 50 million pairs of test statistics from a bivariate distribution in which the correlation was 0.52.

Analyses of the potential importance of these SNPs in the population were conducted among the 3501 individuals from the population study. The hazard rate ratio (HRR) for diabetes associated with each copy of the marker allele was estimated. In this model, individuals who developed diabetes were considered to have developed the disease at the age of onset observed in the longitudinal study, while nondiabetic individuals were considered to be at risk for diabetes until the age at last examination. To account for familial resemblance among siblings, these analyses were conducted with a generalized estimating equations model in which diabetes incidence rates were represented as a Poisson function. The HRR was used to calculate population attributable risk (PAR) for each marker (33). If PLL, PLL and PHH represent, respectively, the proportion of individuals homozygous for the low-risk genotype, heterozygous and homozygous for the high-risk genotype, the PAR, under a multiplicative model, is:

$$\text{PAR} = \frac{1}{[P_{LL} + P_{LL} \times \text{HRR} + P_{HH} \times \text{HRR}^2]}$$

The PAR represents the proportion by which diabetes incidence would decrease if all individuals had the same risk as individuals with the low-risk genotype.
RESULTS

The distribution of p-values for all 80,044 SNPs is shown in Figure 1, along with the distribution expected under the global null hypothesis. Overall, the observed distribution of p-values was similar to that expected under the null hypothesis; however, there was a slight excess of low p-values beyond that expected and this is more apparent if the portion of the distribution at p<0.05 is examined (Figure 1, Panel B). The average Kolmogorov-Smirnov deviation comparing the observed and expected distributions in 200 permutations was significantly greater than 0 (average d=0.012, se=0.00036, p<0.0001). This indicates that the observed distribution contains a statistically significantly greater proportion of low p-values than would be expected under the global null hypothesis of no association with any SNP. Similar results were obtained with the within-family test alone (d=0.006, se=0.00035, p<0.0001). To explore genotyping artifact as a potential source of bias, analyses were repeated with more stringent thresholds for Hardy-Weinberg equilibrium, genotype success rate and minor allele frequency with similar results. For the 25,915 SNPs with p>0.05 for Hardy-Weinberg, genotype success rate >0.99 and minor allele frequency >0.1, results were very similar (e.g., for the within-family test d=0.008, se=0.00058, p<0.0001).

Results of tests for individual SNPs for association with young-onset type 2 diabetes are shown in Figure 2. There were several regions where ≥1 SNP showed fairly strong evidence for association. The SNPs with the lowest p-values were rs686989 (p=2.7 x 10^{-6}, which corresponds to an experiment-wide p-value of 0.11), rs672849 (p=1.5 x 10^{-5}, experiment-wide p=0.55), both on chromosome 11 at 113.54 Mb, rs2164000 on chromosome 9 at 18.75 Mb (p=2.3 x 10^{-5}, experiment-wide p=0.69) and rs10520926 on chromosome 5 at 25.36 Mb (p=2.6 x 10^{-5}, experiment-wide p=0.73). The two chromosome 11 SNPs were highly concordant (r^2=0.99). There were 128 SNPs with p<0.001 (~80 expected under the null) and these SNPs were further genotyped in individuals from the larger population. The 128 SNPs are listed in the Supplemental Online Table (available at http://diabetes.diabetesjournals.org).

There were 11 of these SNPs (out of 119 successfully genotyped) which also showed evidence for association (p<0.05) with diabetes in individuals from the population who were not included in the genome-wide
Genetic Variants and Young-Onset Type 2 Diabetes

study and for which the direction of association was the same as that in the genome-wide study. One would expect ~6 such replications by chance at this level of significance. Results of analyses for these SNPs are shown in Table 2. The chromosome 11 SNPs at 113.54 Mb (rs686989 and rs672849) and the chromosome 5 SNP at 25.36 Mb (rs10520926) were among those that were replicated. In general, the HRR in the population study was lower than the odds ratio from the original study; this is expected given the selection procedure and the fact that the odds ratio only approximates the HRR under certain assumptions [e.g. incident cases are sampled or the disease is rare (34)]. The HRR for most alleles is modest, and those with high PAR are largely those for which the risk alleles are at high frequency.

To assess the extent to which SNPs identified as associated with type 2 diabetes in the present study replicated in other populations, data from three other studies conducted with the Affymetrix 100K array were analyzed for all 646 SNPs with p<0.007 in the present study. These data were provided by investigators from the Amish Family Diabetes Study (35), the Framingham Heart Study (36) and the Starr County Diabetes Study of Mexican-Americans (37). P-values for an association in the same direction observed in the present study were combined across all 3 other studies by Fisher’s method (28). 88.2% of the SNPs had valid data for all three other studies, 6.0% for two other studies, 5.6% for one other study and 0.2% (1 SNP) for none. There were 30 SNPs, shown in Table 3, which replicated the results of the present study at p<0.05 (~32 expected under the null). None of these overlap with those shown in Table 2. The overall analysis strategy and the number of SNPs selected at various stages of analysis are summarized in Figure 3.

DISCUSSION

Type 2 diabetes is partially genetically determined (1-6), but susceptibility variants identified conclusively to date explain only a small portion of the familial risk. It is, therefore, likely that there are additional susceptibility variants that have not yet been identified. Genome-wide association studies are a potentially powerful way to detect these variants. The present analysis represents a genome-wide association study in Pima Indians, a population with a high prevalence of type 2 diabetes and obesity (3,19). Diabetes in this population is highly familial, particularly when
it occurs at young ages (3,17-18,38); therefore the present study was designed to detect variants associated with young-onset diabetes. Although case-control and related association tests can be powerful, they are liable to confounding by population stratification (39,40). In contrast, within-family tests, though less powerful, are robust to such confounding (40). In the present study both case-control and family-based designs were employed and results were combined so that the case-control results augment the power of the within-family test, while maintaining robustness to population stratification (29). Thus, markers identified with this approach are likely to reflect both association and linkage with diabetes-susceptibility variants.

Many of the associations observed in the present study have p-values that are quite strong by conventional criteria. However, the interpretation of p-values in genetic association studies is problematic. Because the prior probability that any given variant is in linkage disequilibrium with a disease susceptibility allele is low, many statisticians recommend very stringent criteria for declaration of statistical significance, e.g., p-values $10^{-5}$ to $10^{-8}$ (41-43). The problem is compounded in genome-wide studies, where because of multiple tests, one would expect some strong associations to occur by chance. Classical methods of adjustment for multiple comparisons, such as Bonferroni, are too stringent for genome-wide association studies, because they assume that tests are independent and ignore linkage disequilibrium among markers. Furthermore, while corrections for multiple testing proceed on the assumption that the global null hypothesis of no association with any marker is of interest (44), the corrections are applied on marker-wise basis and, thus, may not efficiently utilize information from the full distribution of p-values relevant to this global null.

In the present analysis, a permutation procedure was used to assess experiment-wide statistical significance. This procedure maintains the linkage disequilibrium structure among markers. In addition the Kolmogorov-Smirnov test was used to compare the overall distribution of p-values observed in the actual data with that observed in data permuted under the null hypothesis. Although no single marker achieved experiment-wide statistical significance, analysis of the distribution of p-values resulted in strong rejection of the global null hypothesis.
of no association with any marker. This result is expected if there are multiple susceptibility variants, because, in such a situation, a test of the full distribution can be more powerful than a test of a single variant. A systematic bias is an alternative possibility that is difficult to completely exclude, but the present techniques are robust to bias from population stratification. The present results were also unmodified when restricted to SNPs with increased stringency of Hardy-Weinberg equilibrium, allele frequency and genotype success rate, and this suggests that they are not explained by bias due to genotypic artifacts that can be captured by these measures. The number of true functional susceptibility variants is difficult to determine, given that modest linkage disequilibrium may extend over long distances and that many associations may be due to chance. Thus, while the present results strongly suggest that some of the low p-values observed reflect linkage disequilibrium with diabetes-susceptibility loci, the distinction between “true” and “false” positive results is difficult.

Replication of the association in a separate group of individuals provides additional evidence that a marker is associated with diabetes. In the present study, SNPs with the strongest evidence for association in the genome-wide study were further evaluated in a separate set of individuals from the population. (However, some of these individuals were related to those in the genome-wide study.) Several SNPs showed nominal evidence for association (p<0.05) in this replication set. This analysis is limited in power, because the vast majority of individuals from the extremes of the age-of-onset distribution—who provide much of the statistical power—were excluded due to participation in the genome-wide study. In addition, one would expect some of the markers to show association by chance. Thus, some false positives are likely to remain among the SNPs that replicated, and some true positives among those that did not. However, the probability that SNPs showing replication are in linkage disequilibrium with diabetes-susceptibility variants is enhanced above those for which replication was not observed. The result for rs686989 is of particular interest because it is in a region that was identified as linked to diabetes and obesity in a previous genome-wide linkage study in this population (15); however, the association with rs686989 itself does not explain the
linkage signal (data not shown).

The PAR, calculated in a group representative of the full population, provides a measure of the potential importance of each associated SNP, and this information can be used to prioritize regions for follow-up studies. The PAR is calculated on the assumption that the observed association is causal; therefore it may be underestimated if the marker is not highly concordant with a functional allele, or overestimated if confounded by population stratification. In the present analysis, the PAR was calculated from longitudinal data observed in the population study. Thus, in contrast to estimates derived from case-control studies that are based on prevalent cases, the present results do not depend on the questionable assumptions of a rare disease or of sampling only incident cases (34).

Studies of these SNPs in other populations may also be relevant in prioritization of regions for follow-up studies. The present results were further compared with those obtained from three other genome-wide association studies of type 2 diabetes on the Affymetrix 100K array (35-37). Comparison among studies is complicated by the fact that all had different study designs; the present study focused on young-onset diabetes and gave priority to within-family tests. (Characteristics of each study are presented in Joint Supplementary Table 1.) Nonetheless, several of the SNPs identified in the present study had some evidence for association in the other studies (overall p<0.05). These regions are also high priority for follow-up studies. It is noteworthy that rs516415, which is also in the diabetes-linked region on chromosome 11q, showed replication, as did rs1886004, which is in a region on chromosome 1q linked to diabetes in Pimas and other populations (14-15). However, none of the SNPs with evidence for replication in the Pimas at p<0.05 also had a combined p<0.05 in these other studies. This may reflect low power across the studies to detect alleles of modest effect.

The power of genetic association studies depends on the frequency of the functional polymorphisms and on the magnitude of their effects (22). Given that cases represent the lower 10% of the age of onset distribution and controls the upper 45%, we estimate that the present sample size has ~75% power to detect a commona (maf>0.1) functional allele at p<0.001 that explains 3% of the variance in age at onset of diabetes. Power also depends the extent
to which one of the typed markers is strongly concordant with an allele that influences disease susceptibility (21,22).

Since the 100K array used in the present study does not exhaustively capture allelic variation, it is possible that additional regions that contain diabetes susceptibility variants were not identified. The pattern of linkage disequilibrium among SNPs identified in the HapMap project suggests that ~30% of common variants have $r^2>0.8$ with a marker on this array in non-African populations (45). American Indians are not included in the HapMap and they tend to share relatively few common haplotypes with HapMap populations (46). This may be reflected in the fact that a larger proportion of markers were nearly monomorphic in the present study than in the other studies that used the 100K array (35-37), and this could result in lower power for this array to detect associations in American Indians. On the other hand, surveys of various populations suggest that linkage disequilibrium is higher among American Indians than in many other populations so that fewer variants are required to capture common haplotypes (46). Thus, fixed marker sets, such as the 100K array, may capture common variation in American Indians to a similar extent as in other non-African populations.

Recently, several high-density genome-wide association studies have been conducted in northern European populations, and these have identified six gene regions, apart from TCF7L2, PPARG and KCNJ11, which contain putative diabetes susceptibility variants (47-50). None of the SNPs most strongly associated in the present study was in any of these regions. Furthermore, none of the SNPs consistently associated across these other genome-wide studies is well-captured by the 100K array. However, as shown in Table 4, some SNPs in these regions had modest evidence for association in the present study ($p<0.05$). These putative diabetes susceptibility variants have been largely identified in northern Europeans and it is not clear if they play a significant role in diabetes in American Indians or other high-risk populations. A more exhaustive survey of variation in these regions is required to quantify their role in diabetes susceptibility in the Pima population. The variants in TCF7L2 that are most strongly associated with type 2 diabetes in other populations (9-13) have been typed in the Pimas in whom there is little
Genetic Variants and Young-Onset Type 2 Diabetes

evidence that they influence diabetes risk (51).

Genome-wide mapping studies are typically only an initial step in the elucidation of susceptibility variants. The present analyses have identified several regions that may harbor genetic variants that influence susceptibility to young-onset diabetes in American Indians. Several of the associated SNPs are in or near genes, including \textit{ZNF659} (chromosome 3, 21.47 Mb), \textit{FANCF} (chromosome 11 22.60 Mb), \textit{ZBTB15} (chromosome 11 113.54 Mb) and \textit{SENP1} (chromosome 12 46.71 Mb). Fine-mapping studies of these regions are needed to confidently localize the signals to specific genes. Confirmation of the role of genes in the regions identified in the present study will require replication studies in other populations, and, ultimately, functional studies.

ACKNOWLEDGMENTS

The authors thank Jose Florez for review of the manuscript. They also thank the staff of the Diabetes Epidemiology and Clinical Research Section and of the Diabetes Molecular Genetics Section for assistance. They further thank members of the Gila River Indian Community who participated in the study. This research was supported by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases.
REFERENCES:


29. Hanson RL, Knowler WC: Design and analysis of genetic association studies to finely map a locus identified by linkage analysis: assessment of the extent to which an association can account for the linkage. Ann Hum Genet in press.


43. Manly KF: Reliability of statistical associations between genes and disease. *Immunogenetics* 57:549-558, 2005


Genetic Variants and Young-Onset Type 2 Diabetes

79 (Suppl):312, 2006
### Table 1. Characteristics of Individuals in the Genome-Wide and Population-Based Studies

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>% Men</th>
<th>Mean Age* (SD) [Range]</th>
<th>Mean BMI† (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genome-Wide Association Study</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Case-Control Study</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>300</td>
<td>38</td>
<td>19.2 (4.5) [5.5-24.9]</td>
<td>38.9 (8.4)</td>
</tr>
<tr>
<td>No Diabetes</td>
<td>334</td>
<td>48</td>
<td>55.5 (9.8) [45.1-87.9]</td>
<td>35.4 (8.0)</td>
</tr>
<tr>
<td><strong>siblings of Cases/Controls</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>140</td>
<td>34</td>
<td>41.0 (8.3) [25.0-67.4]</td>
<td>38.9 (9.3)</td>
</tr>
<tr>
<td>No Diabetes</td>
<td>121</td>
<td>47</td>
<td>27.8 (7.9) [12.3-42.1]</td>
<td>36.9 (8.9)</td>
</tr>
<tr>
<td><strong>Population Study</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not in Genome-Wide Study</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>1207</td>
<td>38</td>
<td>39.7 (10.6) [10.7-78.0]</td>
<td>38.3 (8.2)</td>
</tr>
<tr>
<td>No Diabetes</td>
<td>1627</td>
<td>46</td>
<td>27.7 (11.6) [5.5-79.7]</td>
<td>35.6 (8.2)</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>1561</td>
<td>37</td>
<td>37.2 (12.1) [5.5-78.0]</td>
<td>38.5 (8.4)</td>
</tr>
<tr>
<td>No Diabetes</td>
<td>1940</td>
<td>46</td>
<td>31.1 (14.5) [5.5-85.3]</td>
<td>35.7 (8.2)</td>
</tr>
</tbody>
</table>

*Age is the age at last examination for nondiabetic individuals and age at onset of diabetes for diabetic individuals.

†Maximum BMI observed after age 15 years in the longitudinal study.
Table 2. Association of Single Nucleotide Polymorphisms Detected in the Genome-Wide Association Study with Evidence for Replicated Association in Additional Individuals.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chrom</th>
<th>Mb</th>
<th>Allele</th>
<th>Freq*</th>
<th>OR†</th>
<th>P-val</th>
<th>OR†</th>
<th>P-val</th>
<th>HRR‡</th>
<th>PAR§</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs10493685</td>
<td>1</td>
<td>81.33</td>
<td>C/T</td>
<td>0.98</td>
<td>∞</td>
<td>1.6 x 10^-4</td>
<td>3.65</td>
<td>0.0043</td>
<td>1.93</td>
<td>--</td>
</tr>
<tr>
<td>rs6994019</td>
<td>8</td>
<td>89.34</td>
<td>C/A</td>
<td>0.98</td>
<td>∞</td>
<td>3.8 x 10^-4</td>
<td>3.49</td>
<td>0.0189</td>
<td>1.68</td>
<td>--</td>
</tr>
<tr>
<td>rs1500415</td>
<td>3</td>
<td>20.80</td>
<td>A/G</td>
<td>0.98</td>
<td>∞</td>
<td>8.6 x 10^-4</td>
<td>3.69</td>
<td>0.0489</td>
<td>1.63</td>
<td>--</td>
</tr>
<tr>
<td>rs1859441</td>
<td>12</td>
<td>46.71</td>
<td>T/C</td>
<td>0.88</td>
<td>2.72</td>
<td>1.7 x 10^-4</td>
<td>1.86</td>
<td>0.0027</td>
<td>1.17</td>
<td>0.25</td>
</tr>
<tr>
<td>rs424695</td>
<td>3</td>
<td>21.47</td>
<td>G/A</td>
<td>0.88</td>
<td>2.33</td>
<td>2.6 x 10^-4</td>
<td>1.63</td>
<td>0.0149</td>
<td>1.15</td>
<td>0.23</td>
</tr>
<tr>
<td>rs10520926</td>
<td>5</td>
<td>25.36</td>
<td>T/A</td>
<td>0.55</td>
<td>2.17</td>
<td>2.6 x 10^-5</td>
<td>1.36</td>
<td>0.0335</td>
<td>1.07</td>
<td>0.07</td>
</tr>
<tr>
<td>rs10500938</td>
<td>11</td>
<td>22.60</td>
<td>A/G</td>
<td>0.20</td>
<td>2.14</td>
<td>4.6 x 10^-4</td>
<td>1.65</td>
<td>0.0011</td>
<td>1.14</td>
<td>0.05</td>
</tr>
<tr>
<td>rs672849</td>
<td>11</td>
<td>113.54</td>
<td>A/G</td>
<td>0.19</td>
<td>2.84</td>
<td>1.5 x 10^-5</td>
<td>1.34</td>
<td>0.0164</td>
<td>1.14</td>
<td>0.05</td>
</tr>
<tr>
<td>rs686989</td>
<td>11</td>
<td>113.54</td>
<td>A/G</td>
<td>0.19</td>
<td>3.26</td>
<td>2.7 x 10^-6</td>
<td>1.27</td>
<td>0.0333</td>
<td>1.13</td>
<td>0.05</td>
</tr>
<tr>
<td>rs9290075</td>
<td>3</td>
<td>162.50</td>
<td>A/C</td>
<td>0.25</td>
<td>2.47</td>
<td>7.0 x 10^-5</td>
<td>1.39</td>
<td>0.0134</td>
<td>1.10</td>
<td>0.04</td>
</tr>
<tr>
<td>rs10506855</td>
<td>12</td>
<td>80.88</td>
<td>T/G</td>
<td>0.02</td>
<td>6.70</td>
<td>5.7 x 10^-4</td>
<td>1.79</td>
<td>0.0335</td>
<td>1.36</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Listed in order of descending PAR.

* Frequency of the allele listed first, which gave OR > 1.
† Odds ratio per copy of the allele listed first in the within-family analysis.
‡ Hazard rate ratio per copy of the allele listed first.
§ Population attributable risk; not shown for SNPs for which the number of low-risk homozygotes was too small for reliable estimation.
Table 3. SNPs with evidence for replicated association with type 2 diabetes (combined p-value<0.05) in Amish, Framingham and Starr County studies

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr</th>
<th>Mb</th>
<th>Allele</th>
<th>Present Study</th>
<th>Amish</th>
<th>Framingham</th>
<th>Starr County</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OR</td>
<td>P-val</td>
<td>OR</td>
<td>P-val</td>
<td>HRR</td>
</tr>
<tr>
<td>rs1775368</td>
<td>1</td>
<td>72.70</td>
<td>T/A</td>
<td>5.15</td>
<td>6.7 x 10^-5</td>
<td>0.84</td>
<td>0.9356</td>
<td>1.20</td>
</tr>
<tr>
<td>rs1886004</td>
<td>1</td>
<td>165.06</td>
<td>A/G</td>
<td>∞</td>
<td>8.8 x 10^-4</td>
<td>2.16</td>
<td>0.0158</td>
<td>1.19</td>
</tr>
<tr>
<td>rs10496191</td>
<td>2</td>
<td>73.59</td>
<td>G/A</td>
<td>1.88</td>
<td>5.5 x 10^-3</td>
<td>1.13</td>
<td>0.1484</td>
<td>1.44</td>
</tr>
<tr>
<td>rs10496193</td>
<td>2</td>
<td>73.69</td>
<td>C/T</td>
<td>1.90</td>
<td>5.3 x 10^-3</td>
<td>1.14</td>
<td>0.1457</td>
<td>1.39</td>
</tr>
<tr>
<td>rs10497681</td>
<td>2</td>
<td>189.38</td>
<td>A/G</td>
<td>12.97</td>
<td>6.8 x 10^-5</td>
<td>1.13</td>
<td>0.1457</td>
<td>1.39</td>
</tr>
<tr>
<td>rs2366760</td>
<td>3</td>
<td>21.15</td>
<td>A/G</td>
<td>1.57</td>
<td>5.0 x 10^-3</td>
<td>1.33</td>
<td>0.0079</td>
<td>1.13</td>
</tr>
<tr>
<td>rs10517403</td>
<td>4</td>
<td>37.05</td>
<td>G/A</td>
<td>2.02</td>
<td>2.9 x 10^-3</td>
<td>1.17</td>
<td>0.0690</td>
<td>0.99</td>
</tr>
<tr>
<td>rs1072389</td>
<td>4</td>
<td>75.36</td>
<td>T/C</td>
<td>2.02</td>
<td>2.9 x 10^-3</td>
<td>1.17</td>
<td>0.0690</td>
<td>0.99</td>
</tr>
<tr>
<td>rs460610</td>
<td>4</td>
<td>103.51</td>
<td>A/G</td>
<td>2.02</td>
<td>3.7 x 10^-4</td>
<td>1.22</td>
<td>0.0337</td>
<td>1.39</td>
</tr>
<tr>
<td>rs543862</td>
<td>5</td>
<td>177.97</td>
<td>T/C</td>
<td>1.74</td>
<td>3.2 x 10^-3</td>
<td>1.15</td>
<td>0.1088</td>
<td>1.27</td>
</tr>
<tr>
<td>rs10488205</td>
<td>7</td>
<td>11.92</td>
<td>C/T</td>
<td>2.20</td>
<td>2.3 x 10^-3</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>rs1486171</td>
<td>7</td>
<td>45.98</td>
<td>A/T</td>
<td>1.80</td>
<td>5.5 x 10^-3</td>
<td>1.18</td>
<td>0.0633</td>
<td>1.00</td>
</tr>
<tr>
<td>rs3847109</td>
<td>7</td>
<td>130.56</td>
<td>A/G</td>
<td>1.88</td>
<td>2.2 x 10^-3</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>rs3780825</td>
<td>9</td>
<td>132.19</td>
<td>A/G</td>
<td>4.92</td>
<td>3.3 x 10^-3</td>
<td>0.99</td>
<td>0.5110</td>
<td>2.19</td>
</tr>
<tr>
<td>rs2130154</td>
<td>11</td>
<td>86.50</td>
<td>C/T</td>
<td>2.34</td>
<td>2.1 x 10^-3</td>
<td>NA</td>
<td>NA</td>
<td>1.38</td>
</tr>
<tr>
<td>rs2048973</td>
<td>11</td>
<td>86.52</td>
<td>T/G</td>
<td>2.11</td>
<td>3.8 x 10^-3</td>
<td>1.19</td>
<td>0.0507</td>
<td>1.59</td>
</tr>
<tr>
<td>rs586421</td>
<td>11</td>
<td>94.50</td>
<td>C/T</td>
<td>1.65</td>
<td>4.0 x 10^-3</td>
<td>1.28</td>
<td>0.0136</td>
<td>1.00</td>
</tr>
<tr>
<td>rs1789369</td>
<td>11</td>
<td>109.50</td>
<td>T/C</td>
<td>1.91</td>
<td>4.6 x 10^-3</td>
<td>0.94</td>
<td>0.7015</td>
<td>1.01</td>
</tr>
<tr>
<td>rs516415</td>
<td>11</td>
<td>121.56</td>
<td>A/G</td>
<td>2.63</td>
<td>5.0 x 10^-5</td>
<td>1.26</td>
<td>0.0179</td>
<td>0.90</td>
</tr>
<tr>
<td>SNP</td>
<td>Chromosome</td>
<td>Position</td>
<td>Allele</td>
<td>Minor Allele</td>
<td>Minor Allele Odds Ratio</td>
<td>Log <em>10</em> Minor Allele Odds Ratio</td>
<td>Major Allele</td>
<td>Major Allele Odds Ratio</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------</td>
<td>----------</td>
<td>--------</td>
<td>--------------</td>
<td>------------------------</td>
<td>-------------------------------</td>
<td>--------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>rs7301434</td>
<td>12</td>
<td>91.36</td>
<td>C/A</td>
<td>3.39</td>
<td>1.5 x 10^-3</td>
<td>0.94</td>
<td>0.6263</td>
<td>1.35</td>
</tr>
<tr>
<td>rs54939</td>
<td>13</td>
<td>54.45</td>
<td>C/A</td>
<td>1.64</td>
<td>3.2 x 10^-3</td>
<td>1.21</td>
<td>0.0360</td>
<td>0.87</td>
</tr>
<tr>
<td>rs1537779</td>
<td>13</td>
<td>81.92</td>
<td>C/T</td>
<td>1.73</td>
<td>2.2 x 10^-3</td>
<td>0.94</td>
<td>0.7160</td>
<td>1.50</td>
</tr>
<tr>
<td>rs9285325</td>
<td>13</td>
<td>81.95</td>
<td>C/G</td>
<td>1.74</td>
<td>3.1 x 10^-3</td>
<td>0.98</td>
<td>0.5880</td>
<td>1.40</td>
</tr>
<tr>
<td>rs1959225</td>
<td>14</td>
<td>40.11</td>
<td>A/T</td>
<td>1.95</td>
<td>1.2 x 10^-3</td>
<td>0.98</td>
<td>0.5885</td>
<td>1.41</td>
</tr>
<tr>
<td>rs10498547</td>
<td>14</td>
<td>80.19</td>
<td>C/A</td>
<td>2.57</td>
<td>6.6 x 10^-4</td>
<td>1.17</td>
<td>0.0901</td>
<td>1.07</td>
</tr>
<tr>
<td>rs958435</td>
<td>20</td>
<td>6.52</td>
<td>G/C</td>
<td>2.76</td>
<td>5.7 x 10^-3</td>
<td>1.28</td>
<td>0.0341</td>
<td>1.44</td>
</tr>
<tr>
<td>rs6140410</td>
<td>20</td>
<td>7.69</td>
<td>A/G</td>
<td>2.98</td>
<td>6.0 x 10^-3</td>
<td>0.89</td>
<td>0.8319</td>
<td>2.33</td>
</tr>
<tr>
<td>rs1223271</td>
<td>20</td>
<td>13.24</td>
<td>G/A</td>
<td>2.21</td>
<td>5.3 x 10^-3</td>
<td>1.60</td>
<td>0.0039</td>
<td>1.20</td>
</tr>
<tr>
<td>rs477184</td>
<td>20</td>
<td>13.25</td>
<td>A/G</td>
<td>2.32</td>
<td>1.2 x 10^-3</td>
<td>1.00</td>
<td>0.5092</td>
<td>1.55</td>
</tr>
<tr>
<td>rs6030447</td>
<td>20</td>
<td>40.79</td>
<td>C/T</td>
<td>2.53</td>
<td>3.6 x 10^-3</td>
<td>1.16</td>
<td>0.0821</td>
<td>0.91</td>
</tr>
</tbody>
</table>

*Odds ratio per copy of the allele listed first (which gave OR > 1 in the present study); in the present study this result is for the within-family analysis.

†One-sided p-value for association in the direction observed in present study

‡Hazard rate ratio per copy of the allele listed first.

¶P-value from the generalized estimating equation analysis in Framingham Heart Study (36).

‖P-value combined from all three populations (except present study) by Fisher’s method.

NA- not analyzed
Table 4. Strongest association results in present study for SNPs in regions identified as associated with type 2 diabetes in multiple high-density genome-wide association studies.

<table>
<thead>
<tr>
<th>GENE</th>
<th>Sentinel SNP*</th>
<th>Chrom</th>
<th>Mb</th>
<th>SNP in Present Study†</th>
<th>Mb</th>
<th>Allele</th>
<th>Freq‡</th>
<th>OR§</th>
<th>P-val</th>
<th>N SNPs¶</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPARG</td>
<td>rs1801281</td>
<td>3</td>
<td>12.38</td>
<td>rs10510422</td>
<td>12.51</td>
<td>T/C</td>
<td>0.92</td>
<td>2.31</td>
<td>0.0384</td>
<td>19</td>
</tr>
<tr>
<td>IGF2BP2</td>
<td>rs4402960</td>
<td>3</td>
<td>187.99</td>
<td>rs6763887</td>
<td>187.97</td>
<td>G/A</td>
<td>0.28</td>
<td>1.34</td>
<td>0.0945</td>
<td>3</td>
</tr>
<tr>
<td>CDKAL1</td>
<td>rs10946398</td>
<td>6</td>
<td>20.77</td>
<td>rs7758851</td>
<td>20.65</td>
<td>C/T</td>
<td>0.94</td>
<td>1.82</td>
<td>0.2085</td>
<td>15</td>
</tr>
<tr>
<td>SLC30A8</td>
<td>rs13266634</td>
<td>8</td>
<td>118.25</td>
<td>rs10505309</td>
<td>118.31</td>
<td>C/T</td>
<td>0.14</td>
<td>2.18</td>
<td>0.0332</td>
<td>28</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>rs1081161</td>
<td>9</td>
<td>22.12</td>
<td>rs2025798</td>
<td>22.27</td>
<td>G/A</td>
<td>0.97</td>
<td>1.93</td>
<td>0.0864</td>
<td>12</td>
</tr>
<tr>
<td>HHEX</td>
<td>rs1111875</td>
<td>10</td>
<td>94.45</td>
<td>rs2096177</td>
<td>94.50</td>
<td>T/G</td>
<td>0.79</td>
<td>1.63</td>
<td>0.0446</td>
<td>6</td>
</tr>
<tr>
<td>TCF7L2</td>
<td>rs7903146</td>
<td>10</td>
<td>114.75</td>
<td>rs10509970</td>
<td>114.91</td>
<td>G/T</td>
<td>0.53</td>
<td>1.30</td>
<td>0.0719</td>
<td>9</td>
</tr>
<tr>
<td>KCNJ11</td>
<td>rs5219</td>
<td>11</td>
<td>17.34</td>
<td>rs2190454</td>
<td>17.49</td>
<td>C/T</td>
<td>0.10</td>
<td>1.27</td>
<td>0.4600</td>
<td>4</td>
</tr>
<tr>
<td>FTO</td>
<td>rs8050136</td>
<td>16</td>
<td>52.37</td>
<td>rs10521308</td>
<td>52.42</td>
<td>T/C</td>
<td>0.05</td>
<td>1.27</td>
<td>0.3299</td>
<td>13</td>
</tr>
</tbody>
</table>

*The “sentinel” SNP is the one with the strongest association with diabetes across multiple genome-wide association studies (47-50).

†The SNP with the smallest p-value within a 200 kb window on either side of the sentinel SNP is reported.

‡Frequency of the allele listed first, which gave OR > 1.

§Odds ratio in the within-family analysis per copy of the allele listed first.

¶Number of SNPs from the present study in the 400 kb window surrounding the sentinel SNP.
FIGURE LEGENDS

Figure 1. Cumulative distribution of p-values observed for all 80,044 SNPs in the genome-wide association study compared with the expected distribution under the null hypothesis of no association with any marker. Panel A: entire distribution. Panel B: distribution below p=0.05.

Figure 2. P-values for association with diabetes across all chromosomes. The x-axis represents the SNPs position on each chromosome. The y axis is the negative of the base 10 logarithm of the p-value (higher values represent greater statistical significance). For ease of presentation, only SNPs with p<0.01 are shown.

Figure 3. Diagrammatic representation of the present study, along with molecular and analytic strategies to assess replication.
Figure 1.
Figure 2.
Figure 3.

Selection of SNPs for further genotyping
P < 0.001 in genome-wide study
128 SNPs

Case-Control Group
300 Cases/334 Controls

Affected/Unaffected Family Members-340 Discordant Sib Pairs

Genome-wide Association SNPs
80,044 SNPs Analyzed after Quality Control Procedures

Selection of SNPs for comparison studies
P < 0.007 in genome-wide study
646 SNPs

Population-Based Group
N = 3501

Replication Group
N = 2834 not in genome-wide study

Replication at P < 0.05
11 SNPs

Replication in Other 100K Populations
Amish, Starr County, Framingham

Replication at P < 0.05
30 SNPs