

A Single Nucleotide Polymorphism in MGEA5 Encoding O-GlcNAc–selective *N*-Acetyl- β -D Glucosaminidase Is Associated With Type 2 Diabetes in Mexican Americans

Donna M. Lehman,¹ Dong-Jing Fu,² Angela B. Freeman,² Kelly J. Hunt,¹ Robin J. Leach,^{3,4} Teresa Johnson-Pais,⁴ Jeanette Hamlington,¹ Thomas D. Dyer,⁵ Rector Arya,¹ Hanna Abboud,¹ Harald H.H. Göring,⁶ Ravindranath Duggirala,⁶ John Blangero,⁶ Robert J. Konrad,⁵ and Michael P. Stern¹

Excess O-glycosylation of proteins by O-linked β -*N*-acetylglucosamine (O-GlcNAc) may be involved in the pathogenesis of type 2 diabetes. The enzyme O-GlcNAc–selective *N*-acetyl- β -D glucosaminidase (O-GlcNAcase) encoded by MGEA5 on 10q24.1-q24.3 reverses this modification by catalyzing the removal of O-GlcNAc. We have previously reported the linkage of type 2 diabetes and age at diabetes onset to an overlapping region on chromosome 10q in the San Antonio Family Diabetes Study (SAFADS). In this study, we investigated menangioma-expressed antigen-5 (MGEA5) as a positional candidate gene. Twenty-four single nucleotide polymorphisms (SNPs), identified by sequencing 44 SAFADS subjects, were genotyped in 436 individuals from 27 families whose data were used in the original linkage report. Association tests indicated significant association of a novel SNP with the traits diabetes ($P = 0.0128$, relative risk = 2.77) and age at diabetes onset ($P = 0.0017$). The associated SNP is located in intron 10, which contains an alternate stop codon and may lead to decreased expression of the 130-kDa isoform, the isoform predicted to contain the O-GlcNAcase activity. We investigated whether this variant was responsible for the original linkage signal. The variance attributed to this SNP accounted for ~25% of the logarithm of odds. These results suggest that this variant within the MGEA5 gene may increase diabetes risk in Mexican Americans. *Diabetes* 54:1214–1221, 2005

From the ¹Department of Medicine, University of Texas Health Science Center, San Antonio, Texas; the ²Lilly Research Laboratories, Indianapolis, Indiana; the ³Department of Cellular and Structural Biology, University of Texas Health Science Center, San Antonio, Texas; the ⁴Department of Pediatrics, University of Texas Health Science Center, San Antonio, Texas; the ⁵Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, Texas; and the ⁶Department of Nephrology, University of Texas Health Science Center, San Antonio, Texas.

Address correspondence and reprint requests to Donna M. Lehman, Department of Medicine/Clinical Epidemiology, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Dr., San Antonio, Texas 78229. E-mail: lehman@uthscsa.edu.

Received for publication 30 September 2004 and accepted in revised form 9 December 2004.

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

LD, linkage disequilibrium; LOD, logarithm of odds; MGEA5, menangioma-expressed antigen-5; O-GlcNAc, O-linked β -*N*-acetylglucosamine; O-GlcNAcase, O-GlcNAc–selective *N*-acetyl- β -D glucosaminidase; SAFADS, San Antonio Family Diabetes Study; 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.

Many nuclear and cytoplasmic proteins are glycosylated on serine or threonine residues by O-linked β -*N*-acetylglucosamine (O-GlcNAc) (1,2). This posttranslational modification is a dynamic and regulated process, much like protein phosphorylation (3), requiring the coordinated action of two enzymes: O-GlcNAc transferase, which uses the substrate uridine diphosphate–*N*-acetylglucosamine to attach a single O-GlcNAc residue and the enzyme O-GlcNAc–selective *N*-acetyl- β -D glucosaminidase (O-GlcNAcase), which catalyzes its removal (4). Aberrant protein glycosylation by O-GlcNAc may be involved in the pathogenesis of type 2 diabetes. Elevated levels of extracellular glucose, by providing more substrate for O-GlcNAc transferase, appears to lead to increased intracellular O-GlcNAc modification of proteins, which can perturb normal insulin signaling events (5,6). Pancreatic β -cells are particularly vulnerable to alterations in O-GlcNAc metabolism (rev. in 7). The β -cells are uniquely enriched with O-GlcNAc transferase and are therefore heavily dependent on the activity of O-GlcNAcase to regulate the O-glycosylation pathway. The pancreatic β -cell–specific toxin streptozotocin, an analogue of GlcNAc that is widely used to induce diabetes in animal models, irreversibly inhibits O-GlcNAcase (6,8). The resulting accumulation of glycosylated proteins in pancreatic β -cells may be the underlying mechanism causing β -cell death leading to diabetes in these models (9). Altered metabolism of O-GlcNAc has also been linked to insulin resistance (10,11), and potential substrates for O-GlcNAc involved in the mechanism include glycogen synthase (12,13) as well as proteins in the insulin signaling cascade (11,14). Recent studies (12,15) have demonstrated that the removal of O-GlcNAc residues either enzymatically or by introducing virally transmitted O-GlcNAcase is sufficient to normalize cell function despite continued exposure to elevated extracellular glucose. These data suggest that O-GlcNAcase has the ability to counteract the detrimental effects of exposure to hyperglycemia. Therefore, it follows that impairment of O-GlcNAcase enzymatic activity, via alterations of the gene encoding O-GlcNAcase (menangioma-expressed antigen-5 [MGEA5]), could influence susceptibility to diabetes.

MGEA5 has been localized to chromosome 10q24.1–24.3

(16). The MGEA5 gene consists of 16 exons spanning ~34 kb of genomic sequence and encodes a 130-kDa protein that is mainly localized in the cytoplasm. An alternatively spliced transcript (MGEA5s), consisting of exons 1–10 and part of intron 10, encodes a 75-kDa protein that has nuclear localization (17). The COOH-terminal region of MGEA5, which is missing in the splice variant, is predicted to contain the O-GlcNAcase activity (3,18).

We have previously reported linkage of type 2 diabetes and age at onset of diabetes to a region overlapping the MGEA5 locus on chromosome 10q in the San Antonio Family Diabetes Study (SAFADS) (19). In addition, results from a number of genome scans for type 2 diabetes and measures of insulin sensitivity in other populations including the Pima Indians (20–23) have also implicated chromosome 10q as a region that might harbor a gene(s) influencing susceptibility to these traits. Therefore, in this study, we investigated MGEA5 as a positional and biological candidate gene for type 2 diabetes in the SAFADS, an extended pedigree study consisting of Mexican-American families.

RESEARCH DESIGN AND METHODS

Subjects used in this study were participants of the population-based SAFADS that has been described in detail elsewhere (19). Proband for SAFADS were low-income Mexican Americans with type 2 diabetes, and all first-, second-, and third-degree relatives of the probands, aged ≥ 18 years, were considered eligible for the study. The institutional review board of the University of Texas Health Science Center at San Antonio approved all procedures, and all subjects gave informed consent.

As part of a prior genome-mapping project, highly polymorphic markers providing coverage at 10- to 20-cM intervals on all autosomes were genotyped in a subset of 440 participants in the 27 most informative extended families. The genotyping methods and marker information have already been described (19,24). A genome-wide scan for type 2 diabetes susceptibility genes previously conducted using the genotypic and phenotypic data from these subjects revealed significant evidence for linkage to a region on chromosome 10q (19). Subsequently, a new genomic scan of 382 highly polymorphic markers distributed throughout the genome at ~10-cM intervals was performed on these participants by the Center for Inherited Disease Research. Genotype data were cleaned for both Mendelian and spurious double-recombinant errors using Simwalk2 (25). MultiMap/CRI-MAP (26,27) was used to construct sex-averaged marker maps using the cleaned genotype data. Allele frequencies were estimated by maximum likelihood methods (28) implemented in the computer program SOLAR (29), and multipoint identity-by-descent matrices were estimated using Markov Chain Monte Carlo methods implemented in LOKI (30). All analyses reported in this study were conducted using this new set of markers from the Center for Inherited Disease Research.

Molecular methods. We identified variants in MGEA5 genomic sequence by sequencing regions of interest in 22 diabetic and 22 nondiabetic individuals selected from 11 families who made the greatest contribution to the logarithm of odds (LOD) score in our first genome-wide scan described above (19). All exons, including the 5' and 3' untranslated regions and ~100 bp of flanking intronic sequence as well as 1 kb upstream of exon 1 (putative promoter), were sequenced in all 44 individuals. We also identified all intronic regions that exhibited >70% identity between the human and mouse genomic sequence using the global sequence alignment software tool VISTA (31), as automated on the Berkeley Genome Pipeline (available at <http://pipeline.lbl.gov>) (32). Large regions in introns 10 and 11 exhibited significant identity, so both introns were sequenced in their entirety. In total, 2.7 kb of coding sequence, 2.5 kb of untranslated region, and 9.6 kb of intronic and putative promoter regions were screened. The primers used for amplification and sequencing are provided in online appendix Table 1 (available at <http://diabetes.diabetesjournals.org>). The PCR amplification conditions and genetic variation discovery process have been described (33).

All variants identified using this sequencing strategy were genotyped in 436 individuals for whom DNA was available. Most single nucleotide polymorphism (SNP) assays were performed using the Applied Biosystems (Foster City, CA) TaqMan Allelic Discrimination methodology on an ABI Prism 7900HT Sequence Detection System. Others were genotyped using either restriction fragment–polymorphism assays, primer extension (ABI SNaPshot;

Applied Biosystems), or direct sequencing as listed in Table 1. SNP LLY-MGEA5-14 was genotyped using a restriction fragment–polymorphism assay, and subsequently all genotypes were confirmed by sequencing.

Statistical analysis. Diabetes was defined according to the current American Diabetes Association criteria (34,35). Participants who did not meet these criteria but who reported physician-diagnosed diabetes and who reported current therapy with either oral antidiabetic agents or insulin were also considered to have diabetes. We performed multipoint variance components linkage analysis using SOLAR on the discrete trait diabetes using a threshold model as described by Duggirala and colleagues (19,36). Age and age terms (2) were included in the model. In addition, we used SAS to model age of diabetes diagnosis as a proxy for age of diabetes onset with a Cox proportional hazards model. In the Cox proportional hazards models, for previously diagnosed diabetic participants, self-reported age of diagnosis was used as the time of the event; for diabetic individuals initially diagnosed at the SAFADS examination, the participants' reported age at that examination was used as the time of the event; and finally, nondiabetic participants were censored at their SAFADS examination age. Standard multipoint variance components linkage analysis was performed on the Martingale residual from the Cox proportional hazards model, a quantitative trait, using SOLAR. Since the SAFADS families were ascertained on the basis of type 2 diabetes probands, our analyses included ascertainment correction.

Linkage disequilibrium (LD) between each pair of SNPs was calculated by direct correlation ($|r|$) between SNP genotype vectors in which individual SNP genotypes were scored as 0, 1, or 2, depending on how many copies of the rarer allele an individual carried. This calculation is performed in SOLAR, which then produces a graphical plot of the absolute correlations among SNPs by nucleotide position. Haplotypes were estimated using the computer program MERLIN (37). For haplotypes with sufficient frequency (greater than five copies existing in the samples), haplotype score vectors were then generated with elements containing a 0, 1, or 2, depending upon the number of copies of a specific haplotype that an individual carried.

To test the association between each SNP or haplotype and the traits diabetes and age of onset of diabetes (i.e., the Martingale residual), a measured genotype approach (38) was used, with the allele counts at individual SNPs or the haplotype counts at all SNPs jointly serving as the measured genotypes. This method accounts for the relatedness among family members by estimating the likelihood of genetic models given the pedigree structure. The likelihood for a model in which the trait mean is allowed to vary according to genotype was compared with a nested model in which the genotypic means were restricted to be equal to each other. The significance of the association was tested by likelihood ratio tests, which compare the difference in the likelihoods of the full and nested models. Two times the difference between the logarithm of the likelihoods of the two models is distributed asymptotically as a χ^2 statistic with degrees of freedom equal to the difference in the numbers of parameters in the models being compared. The measured genotype method was implemented by using SOLAR, and a correction for multiple testing, which accounts for SNPs in LD with each other (39), was used. SOLAR also produces an estimate of the relative risk for the genotypes. To address possibilities of hidden population stratification in the SAFADS population, we used a pedigree test of transmission disequilibrium, specifically the quantitative trait disequilibrium test as described by Abecasis et al. (40).

To assess whether a SNP accounted for the linkage signal, linkage on chromosome 10q was reevaluated conditional on the measured genotype effects. By including a genotype-based covariate in the model of the trait mean, the variance attributed to it is removed from the linkage model. If the measured genotype is the sole functional variant in this region of linkage that is influencing the trait, then identity-by-descent allele sharing should provide no additional information, and the LOD score in the conditional linkage analysis should drop to nearly zero. If the genotyped variant is one of several functional variants or is in LD with the true functional variant, not all of the quantitative trait loci variance will be absorbed into the mean effects model, and some evidence for linkage should remain in the conditional analysis. This method and background are described in Almasy and Blangero (41).

RESULTS

The data from a total of 436 individuals, aged 17–97 years, were used for this study. The characteristics for these subjects by diabetes status are presented in Table 2. The age and age-adjusted (2) heritability ($h^2 \pm SE$) for diabetes was 0.63 ± 0.16 ($P < 0.0001$), while the heritability for the Martingale residual was 0.23 ± 0.081 ($P = 0.0002$). The Martingale residuals for diabetes age of onset meet the

TABLE 1
Methods and primers used for genotyping of SNPs in MGEA5

SNP LLY-	Sequence	Assay method	Primers	Probes
MGEA5-19	GCTCCTTGGARTGGTGTCTTT	TaqMan	F-GGGAAAATGATTTCAAGGGCAGAGG R-CCCAAGGCTTGGTTGGTAAA	VIC-CTCCTTGGAAATGGTGTG FAM-TCCCTTGGAGTGGTGTG
MGEA5-1	AGCCTATGAAYGTGGCCCTGTT	TaqMan	F-CCAAGAAGAGGAAATGGAGATTGA R-TCCCATTTCTGGAGCCTTAAAGGA	VIC-CAGGCAACGTTTCATAG FAM-CAGGCAACATTCATAG
MGEA5-2	TTGCCCTGTTTWTACATATGTA	TaqMan	F-GGAAATGGAGATTGACAGCCTATGA R-TCCCATTTCTGGAGCCTTAAAGGA	VIC-AGAAAAATTACATATGTATAAACACAG FAM-AAATTTACATATGTAAAAACACAG
MGEA5-22	TTTTGCGACCARCCCTGGGCAGC	SNaPshot	F-GTAGGAGGCCCTTTTATCAATT R-CAAGCAGCTGGGACTACAGG	GTTTACCACATGCTGCCACAGG
MGEA5-23	AACAAACAAAASAAAAGAGAAAA	TaqMan	F-CTGGGCACTAAGAGTGAGACT R-GCCACAGTATTCTAAATGTAGGGAGAA	VIC-AAGAAAAACAAAACAAAACAAAAGA FAM-AACAAAACAAAACAAAAGAAAAGA
MGEA5-24	TTTTACTTTTTCYCCCTACATTT	Sequencing	F-CAAAAAACCCATGGCCCTTTTG R-CAAAAGCATTCAAATCCAGG	
MGEA5-3	GTACTTTGATCYTTCACCTTGG	TaqMan	F-TGGAGTGAAGAAAAGATGTAGTGTAGT R-GCATACTTTCCTACTGGCTCTAGTTA	VIC-CAAAGGTGAAGGATCAAG FAM-AAGGTGAAGGATCAAG
MGEA5-12	CAAAAACATATTKTTGTGCTTTT	TaqMan	F-TCTCATTGTGCAGAAAGGCTTTTGT R-GTACTATCTTCACTGTCCAGCTGCAA	VIC-AAAAAGCAACAACAATAGTT FAM-AAAAAGCAACAACAATAGTT
MGEA5-4	ATCTGTGTGAMCAATTTTATTT	TaqMan	F-GCATCTTTCTGAATCTAATGTGTTCTATCTCT R-TGTGCAACTTGCCTACCTACAAAA	VIC-CTGTGTGACCAATTTTA FAM-CTGTGTGAAACAATTTTA
MGEA5-5	ACCTTCCCTTAYGAGCATGGAC	TaqMan	F-GGAGATTGTCAGTACTTCTGCTGAT R-CGTAAGCATCTGTGCTCTTTG	VIC-CATGCTCGTAAAGGAA FAM-CCATGCTCATAAAGGAA
MGEA5-6	AAATGAGTACRCTGTAACTGA	TaqMan	F-CCCATACTGTGTAATGTGAAAAAG R-ACCGGCCTAAAAGAACATATTTAAAAACA	VIC-TTCTTAAATGAGTACACTGTAAAC FAM-CTTAAATGAGTACACTGTAAAC
MGEA5-7	GAGAAAAACAYTTPAGGGAAAT	Sequencing	F-TGTGCAACAGGACAATTC R-CTGAATTAGCCCAAGGGAGA	
MGEA5-8	TGTTTACCAGRGATFAGAAAT	RFLP	F-CGCATGCGAGCTTTAGTGTA R-CTGGCTGTGTGCTGTAT	
MGEA5-9	TTTTTTTTTTTTTWAATTTGTCAAA	BstNI	F-CGCATGCGAGCTTTAGTGTA R-CTGGCTGTGTGCTGTAT	
MGEA5-27	GGCGTGGTGGYGCATGCCTGT	RFLP	F-GAGAGTATTTTGACCTCATG R-GTACAGGAGCTATACAGAGC	
MGEA5-16	ACACAAGTTTTRTAAAACTTTCT	Dra I	F-GCAGTGTGAGCAATTTATGGTGGT R-TCTCCCTGTGCTGAGCTTT	R-CAAAGTAGCTGGGATTCAGGCATGTC
MGEA5-17	ACTATCATTTASTGTATTTTAT	SNaPshot	F-GCAGTGTGAGCAATTTATGGTGGT R-TCTCCCTGTGCTGAGCTTT	
MGEA5-18	AAACCAGATASAAAACCTCTTAA	Sequencing	F-GTACCCCTACCCCTTGAAAAATGACA R-GACAGTCCCAAAAGCAACCAATTT	VIC-AAACCAGATAGAAAACCTCT FAM-CCAGATACAAAACCTCT
MGEA5-14	TAGTTACTATWTTTTTATGACA	Sequencing	F-TGAATTCCTTCATGGGCTGT R-CGGCATTTCAATACTAATGTTCC	
MGEA5-13	AGACAAAACCTT(G)GGACTTACTTT	TaqMan	F-GAATTCCTTCATGGGCTGTG R-AAGCGGAGAACGTTACGAGA	
MGEA5-20	GAGATGTGCTGCYAAATCCCCCCC	Sequencing	F-GAATTCCTTCATGGGCTGTG R-AAGCGGAGAACGTTACGAGA	
MGEA5-10	TTTTTTTTTTTTTTTTTTTTTTA	Sequencing	F-CTTATGTTGGCCTCAGATG R-GTTTCATGGAGCTTCCACTG	CTTTCACAGTAATTTTCTGAAAGGAGAT GCTGC
MGEA5-11	GGAAAGAACRTRTATATTTTTTG	SNaPshot	F-TCAGCCTGGATTACTGCTTT R-AAATATCAAAATAGTACTCCA	ATTTACTATGTTAAATGGTCTTTTTTATTT VIC-ATAGGAAAAAGAACTATATATT
MGEA5-25	TTTTGTCTTTTTTGTGAAGTTTTA	TaqMan	F-GGTGGAAAAAGCAGTAPAAAATTTACAAITCA R-GGTGGAAAAAGCAGTAPAAAATTTACAAITCA	FAM-AGGAAAAAGAACTGTATATT Reverse

RFLP, restriction fragment-length polymorphism.

TABLE 2
Characteristics of individuals with and without type 2 diabetes in the SAFADS

	Diabetic subjects	Nondiabetic subjects
<i>n</i>	97	339
Male (%)	41.2	40.3
Age (years)	56.9 ± 14.2	38.7 ± 15.9
Age of diagnosis (years)*	48.4 ± 13.7	—
Fasting glucose (mg/dl)	199 ± 73.7	90 ± 11.6
BMI (kg/m ²)	31.8 ± 6.4	29.6 ± 6.8
Waist circumference (mm)	1041 ± 146.7	954 ± 184.3

Data are means ± SD. *Actual age at diagnosis (not Martingale residual value).

prerequisites of the variance component method used (all phenotypes are within 4 SDs of the mean, kurtosis of 0.49, residual kurtosis of 0.52, and skewness of 0.37).

Sequencing of the 44 selected SAFADS subjects identified SNPs in this locus, of which 19 are novel. The minor allele frequency ranged from 0.02 to 0.25, as described in Table 3. Pairwise LD tests were conducted with all SNP genotypes. As can be seen from Fig. 1, there is weak or no association between LD and physical distance in this particular gene, suggesting that LD in this region is unpredictable. For example, SNPs LLY-MGEA5-4 and LLY-MGEA5-12 have a minor allele frequency >15% and are <1 kb apart but exhibit very little LD. In contrast, common SNPs LLY-MGEA5-23 and LLY-MGEA5-16 are >10 kb apart and are in complete LD. Also, rare (minor allele frequency <5%) SNPs LLY-MGEA5-22, LLY-MGEA5-3, and LLY-MGEA5-5 span >7 kb and are in near complete LD,

while rare SNPs LLY-MGEA5-1 and LLY-MGEA5-2 are only 12 bp apart and exhibit no LD. The average absolute correlation among the 24 SNPs was 0.133. This is quite low. However, four sets of SNPs showed high intraset correlation. The highly correlated sets are LLY-MGEA5-4, -9, and -23; LLY-MGEA5-10, -12, and -13; and LLY-MGEA5-3, -5, and -22. The members of each of these sets exhibit a correlation of at least 0.95 with every other member of the set. The 24 SNPs behave statistically like 20.8 independent SNPs using the method described by Nyholt (39). This level of observed nonindependence among SNPs requires, using Bonferroni's correction, that we observe a *P* value <0.002460 (or negative log *P* > 2.6091) to obtain an experiment-wide *P* value ≤0.05.

Individual association tests indicated association of two SNPs with the traits diabetes age of onset or diabetes as shown in Fig. 2 and Table 3. Significant association of SNP LLY-MGEA5-14 was observed with the traits diabetes age of onset (*P* = 0.0017) and diabetes (*P* = 0.0128). The risk for diabetes was 2.77 times greater for subjects carrying one copy of the T allele for SNP LLY-MGEA5-14 compared with those subjects with two A alleles. No homozygotes for the T allele were observed. In addition, SNP LLY-MGEA5-20 was moderately associated with diabetes age of onset (*P* = 0.0336). Both SNPs were only present in the heterozygous state; therefore, only association tests using additive (in this case, equivalent to dominance) models were done. Using the quantitative trait disequilibrium method of Abecasis et al. (40), we observed no evidence for hidden stratification for these SNPs (data not shown). Haplotype analyses revealed that the rare variants for MGEA5-14 and MGEA5-20 reside on distinct haplotypes.

TABLE 3
Allele frequencies and evidence for association for SNPs in MGEA5 genomic sequence detected in the SAFADS population

Location in gene (kb position)*	Gene region	Eli Lilly SNP-ID	dbSNP loc	Major/minor allele	Minor allele frequency	Association <i>P</i> value	
						Diabetes	Age at onset†
-0.862	promoter	LLY-MGEA5-19	NA	A/G	0.006	0.576	0.783
4.854	intron 2	LLY-MGEA5-1	rs3758557	T/C	0.040	0.234	0.415
4.866	intron 2	LLY-MGEA5-2	NA	A/T	0.030	0.359	0.721
11.605	intron 5	LLY-MGEA5-22	NA	A/G	0.030	0.162	0.557
11.819	intron 5	LLY-MGEA5-23	rs3740422	C/G	0.230	0.706	0.856
11.862	intron 5	LLY-MGEA5-24	NA	T/C	0.003	0.245	0.062
14.309	intron 7	LLY-MGEA5-3	NA	T/C	0.030	0.154	0.549
17.593	intron 7	LLY-MGEA5-12	rs2305194	A/C	0.230	0.572	0.646
18.535	intron 8	LLY-MGEA5-4	NA	C/A	0.160	0.457	0.383
19.078	exon 9	LLY-MGEA5-5	NA	C/T	0.030	0.160	0.552
19.263	intron 9	LLY-MGEA5-6	NA	A/G	0.030	0.768	0.252
20.121	intron 10	LLY-MGEA5-7	NA	T/C	0.020	0.540	0.866
20.598	intron 10	LLY-MGEA5-8	NA	G/A	0.010	0.995	0.669
20.684	intron 10	LLY-MGEA5-9	NA	A/T	0.150	0.622	0.569
21.301	intron 10	LLY-MGEA5-27	NA	G/A	0.012	0.075	0.052
22.168	intron 10	LLY-MGEA5-16	rs11191116	G/A	0.250	0.367	0.475
22.236	intron 10	LLY-MGEA5-17	NA	G/C	0.060	0.109	0.309
23.002	intron 10	LLY-MGEA5-18	NA	C/G	0.015	0.879	0.600
23.693	intron 10	LLY-MGEA5-14	NA	A/T	0.040	0.013	0.002
23.769	intron 10	LLY-MGEA5-13	rs5787445	G/'-'	0.230	0.571	0.648
24.300	intron 11	LLY-MGEA5-20	NA	C/T	0.032	0.126	0.034
27.175	intron 14	LLY-MGEA5-10	rs4244349	A/G	0.230	0.611	0.709
30.712	intron 15	LLY-MGEA5-11	NA	A/G	0.006	0.469	0.775
32.409	3' untranslated region	LLY-MGEA5-25	NA	T/C	0.040	0.363	0.465

*Kilobase position from the translation initiation site according to NM_012215. †Trait is the Martingale residual.

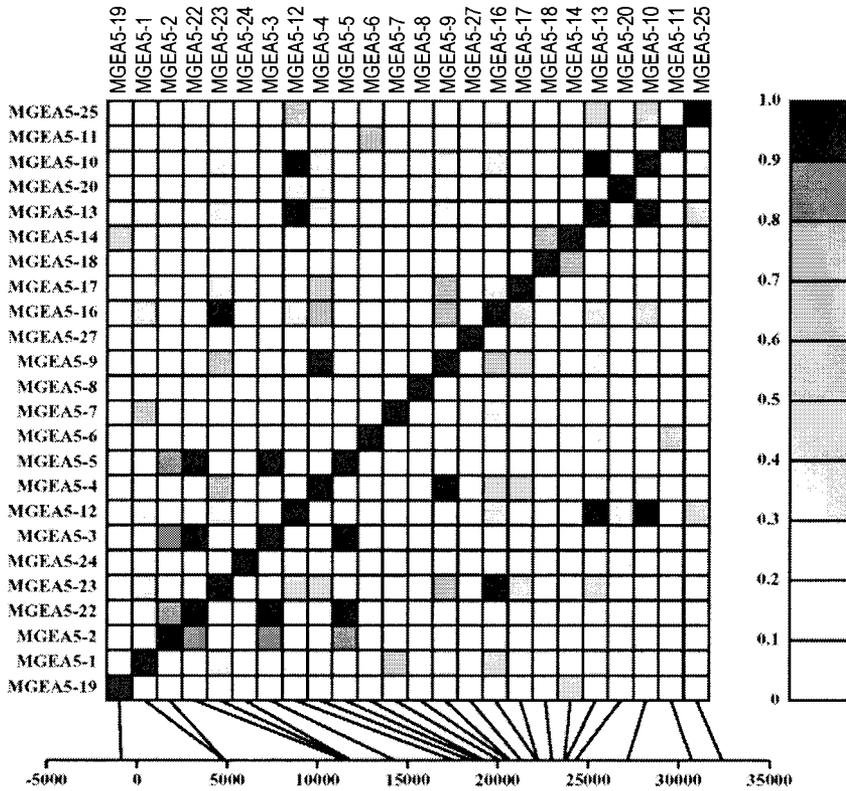


FIG. 1. LD within the MGEA5 gene in the SAFADS population. SNPs are labeled on both axes. SNP locations within the gene are shown on the lower x-axis. Figure depicts the measure $|r|$ between pairs of SNPs by intensity of shaded box as shown in legend. The diagonal represents a comparison of each SNP against itself (i.e., $|r| = 1.0$).

Association tests using haplotype information for all SNPs did not reveal any stronger association than the individual SNPs themselves.

After correcting for multiple testing as described above, the association of SNP LLY-MGEA5-14 with the trait diabetes age of onset remained significant, so next we investigated whether this variant was responsible for the original linkage signal. Variance components linkage anal-

ysis conditional on the SNP genotypes as fixed effects was conducted. As shown in Fig. 3A, the variance attributed to SNP LLY-MGEA5-14 accounted for ~25% of the LOD score for the trait diabetes age of onset. The LOD dropped from

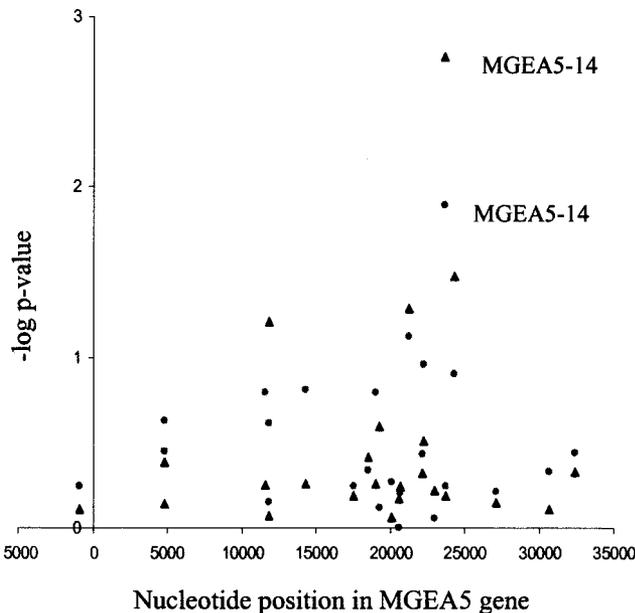


FIG. 2. Association of SNPs in the MGEA5 gene with type 2 diabetes and age of diabetes onset in SAFADS. The negative log of the P value for association for individual SNPs is plotted against their location within the MGEA5 gene where 1 denotes the translation initiation site according to NM_012215. ▲ denotes association with the trait age of diabetes onset; ● denotes association with the trait diabetes.

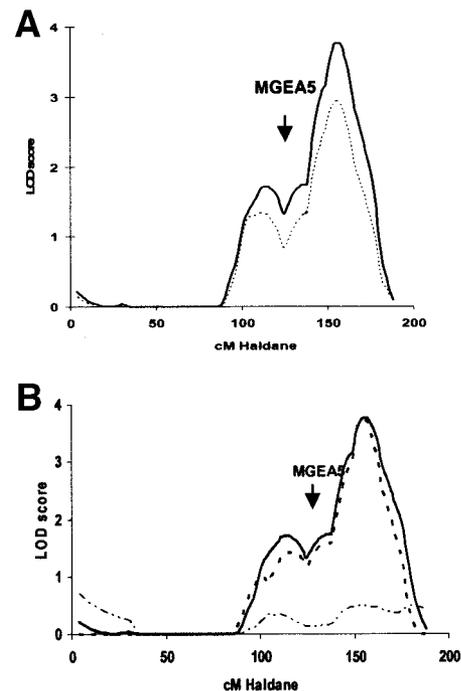


FIG. 3. Linkage evidence for age of diabetes onset on chromosome 10. A: Linkage with all 27 families ($n = 436$) (solid line). Dotted line indicates linkage conditional on SNP LLY-MGEA5-14 as a fixed effect. B: Linkage by presence/absence of the risk allele (T) for SNP LLY-MGEA5-14 in at least one family member. Twenty-seven SAFADS families (solid line), 12 at-risk families (dashed line), and 15 not-at-risk families (heavy dotted line).

TABLE 4
Characteristics of individuals with and without T allele at SNP MGEA5-14

	With T allele	Without T allele
<i>n</i> (families)	12	20
Average number of members per family	21.2 ± 12.3	12.6 ± 5.7
<i>n</i> (subjects)	32	404
Age (years)	43.7 ± 15.9	42.7 ± 17.4
BMI (kg/m ²)	29.6 ± 7.5	30.1 ± 6.7
Age at diabetes onset (diabetic subjects only)	43.5 ± 8.9	49.3 ± 14.3*
Have diabetes (%)	46.9	20.4†

Data are means ± SD. **P* = 0.0017 when correcting for family structure. †*P* = 0.0128 when correcting for family structure.

3.77 to 2.84 when SNP LLY-MGEA5-14 was in the model as a fixed effect. As an exploratory analysis, we separated our families based on the presence of the T allele in at least one family member. Following the nomenclature of Silander et al. (42), those families carrying the T allele were identified as “at risk” (12 families), and those in whom the T allele was not present were identified as “not at risk” (15 families). Linkage analysis of the at-risk and not-at-risk families indicated that nearly all of the evidence for linkage on chromosome 10q was observed in the former (Fig. 3B). The peak LOD in all 27 families was 3.77. The peak LOD in the at-risk families was 3.75, and the LOD in the not-at-risk families was 0.48. The average number of family members for whom both phenotypic and genotypic information were available in the at-risk and not-at-risk families was 21.2 (range 2–41) and 12.6 (range 4–23), respectively. Therefore, the at-risk families are larger on average and may be contributing more to linkage simply due to the greater number of relative pairs for which identity-by-descent information is available.

Characteristics of the subjects who carried a T allele at SNP MGEA5-14 were compared with those subjects who do not. As shown in Table 4, the mean age and BMI were not statistically different between the groups (*P* = 0.82 for age, *P* = 0.40 for BMI) when using a measured genotype approach to account for family relations. However, as stated above, age of diabetes onset was significantly lower and the prevalence of diabetes was significantly higher in the individuals that carried a T allele at this SNP.

DISCUSSION

The gene encoding O-GlcNAcase, MGEA5, is an appealing candidate gene for type 2 diabetes. Accumulating evidence using animal models suggests that impairment of the enzyme activity may impair pancreatic β-cell function and/or lead to insulin resistance, thereby enhancing susceptibility to diabetes. We have therefore investigated MGEA5 as a positional and biological candidate gene for type 2 diabetes and age of onset of diabetes in the SAFADS, an admixed population of European and Native-American origin.

We identified variants in the gene by resequencing the coding and potential regulatory regions of the locus in diabetic and nondiabetic subjects. Twenty-four SNPs were

identified in the ~14.7 kb of sequence that were screened. No missense or nonsense mutations were observed, and only one synonymous SNP was detected, confirming that this gene is highly conserved. Using a measured genotype analysis, we observed significant evidence of association for SNP LLY-MGEA5-14 with the quantitative trait age of onset of diabetes and the discrete trait diabetes. No significant difference was observed in mean age or BMI between individuals with and without SNP LLY-MGEA5-14, so the association is not confounded by these variables. SNP LLY-MGEA5-14 is located within intron 10, which contains an alternate stop codon, so it could conceivably affect relative expression of MGEA5 isoforms. Decreased expression of the 130-kDa isoform, which is predicted to contain the O-GlcNAcase activity, may sufficiently alter O-GlcNAc metabolism to lead to impaired β-cell function and/or insulin resistance. Allele-specific functional studies will be necessary to determine whether this variant affects expression of the protein.

To determine whether the variance attributed to SNP LLY-MGEA5-14 accounted for our linkage signal, we re-evaluated linkage on chromosome 10, conditional on the measured genotype effect. By including LLY-MGEA5-14 genotypes as a covariate in the model, the LOD score for the trait diabetes age of onset dropped by 25%, indicating that this variant accounts for a considerable part of the observed LOD. The point estimate of the residual LOD is still significant (*P* = 0.00015), however suggesting that other as yet unidentified variants, such as those in non-conserved introns or more distant regulatory regions, in this gene may be involved as well. We are currently expanding our resequencing efforts to screen more regions of the gene for further investigation. It is also possible that a cluster of genes underlies the linkage signal, and variation in those genes may account for the remaining LOD score. Alternatively the LLY-MGEA5-14 variant may be in LD with another variant in this region. It is also interesting to note that this SNP is only present in the 12 families that contributed nearly all of the evidence for linkage to chromosome 10q, yet conditional linkage results indicate that variation at this SNP clearly does not account for all of the linkage. In this case, identity-by-descent allele sharing among the relatives of those 12 families is providing additional information for linkage to indicate that additional variation at this locus is influencing the trait. That is, although the at-risk families are responsible for nearly all of the evidence for linkage, the LLY-MGEA5-14 SNP itself is not.

Farook et al. (43) previously examined the MGEA5 gene as a candidate gene in the Pima Indians and reported no evidence to support its involvement in susceptibility to diabetes or insulin resistance. Based on SNP discovery efforts conducted on 30 subjects, that study identified only two variants in the gene regions that were screened. One SNP was located in the putative promoter with a minor allele frequency of only 2% and was not analyzed. We did not observe this SNP in any of the 436 subjects in our study (data not shown). The second SNP corresponded to dbSNP entry rs2305194 and showed no association with any indexes of insulin resistance. SNP rs2305194 was observed in the SAFADS subjects with a minor allele frequency of only 23% compared with 40% in the Pima

Indians, and likewise, no association was observed with this SNP and the traits diabetes or diabetes age of onset. While Farook et al. investigated similar regions of MGEA5 and found only two SNPs, this study identified many additional SNPs. This could be due to a variety of reasons, such as the use of different methods for variant detection (we did not pool samples), differences in population admixture (the SAFADS population consists of an admixed population of Native and European Americans, while the Pimas are primarily of Native-American descent), and the screening of additional intronic regions in this study. Eight of the SNPs identified in this study were located in conserved regions of introns 10 and 11 that were not screened in the previous study. Interestingly, the SNPs exhibiting association with diabetes traits in this study are located in this region of the gene.

In conclusion, this study provides the first evidence that the gene encoding O-GlcNAcase may be a susceptibility locus for type 2 diabetes in humans. The relative risk for diabetes attributed to having the rare allele for the associated SNP LLY-MGEA5-14 is substantial in this population. Future functional studies are planned to determine whether this variant located in intron 10 results in impairment of the O-GlcNAcase enzyme activity.

ACKNOWLEDGMENTS

This research was supported by grants from the National Institutes of Health (R01-DK-42273, R01-DK-47482, R01-DK-53889, MH-59490, and P50DK061597) and a Junior Faculty Award from the American Diabetes Association (to D.M.L.).

We thank the participants of SAFADS and are grateful for their participation and cooperation. We are also very appreciative of the support from Dr. Jude Onyia for this project.

REFERENCES

- Hanover JA: Glycan-dependent signaling: O-linked N-acetylglucosamine. *FASEB J* 15:1865–1876, 2001
- Wells L, Vosseller K, Hart GW: Glycosylation of nucleocytoplasmic proteins: signal transduction and O-GlcNAc. *Science* 291:2376–2378, 2001
- Wells L, Gao Y, Mahoney JA, Vosseller K, Chen C, Rosen A, Hart GW: Dynamic O-glycosylation of nuclear and cytosolic proteins: further characterization of the nucleocytoplasmic beta-N-acetylglucosaminidase, O-GlcNAcase. *J Biol Chem* 277:1755–1761, 2002
- Gao Y, Wells L, Comer FI, Parker GJ, Hart GW: Dynamic O-glycosylation of nuclear and cytosolic proteins: cloning and characterization of a neutral, cytosolic beta-N-acetylglucosaminidase from human brain. *J Biol Chem* 276:9838–9845, 2001
- Konrad RJ, Janowski KM, Kudlow JE: Glucose and streptozotocin stimulate p135 O-glycosylation in pancreatic islets. *Biochem Biophys Res Commun* 267:26–32, 2000
- Liu K, Paterson AJ, Chin E, Kudlow JE: Glucose stimulates protein modification by O-linked GlcNAc in pancreatic beta cells: linkage of O-linked GlcNAc to beta cell death. *Proc Natl Acad Sci U S A* 97:2820–2825, 2000
- Konrad RJ, Kudlow JE: The role of O-linked protein glycosylation in beta-cell dysfunction. *Int J Mol Med* 10:535–539, 2002
- Roos MD, Xie W, Su K, Clark JA, Yang X, Chin E, Paterson AJ, Kudlow JE: Streptozotocin, an analog of N-acetylglucosamine, blocks the removal of O-GlcNAc from intracellular proteins. *Proc Assoc Am Physicians* 110:422–432, 1998
- Konrad RJ, Mikolaenko I, Tolar JF, Liu K, Kudlow JE: The potential mechanism of the diabetogenic action of streptozotocin: inhibition of pancreatic beta-cell O-GlcNAc-selective N-acetyl-beta-D-glucosaminidase. *Biochem J* 356:31–41, 2001
- Vosseller K, Wells L, Lane MD, Hart GW: Elevated nucleocytoplasmic glycosylation by O-GlcNAc results in insulin resistance associated with defects in Akt activation in 3T3-L1 adipocytes. *Proc Natl Acad Sci U S A* 99:5313–5318, 2002
- Arias EB, Kim J, Cartee GD: Prolonged incubation in PUGNAc results in increased protein O-linked glycosylation and insulin resistance in rat skeletal muscle. *Diabetes* 53:921–930, 2004
- Parker GJ, Lund KC, Taylor RP, McClain DA: Insulin resistance of glycogen synthase mediated by o-linked N-acetylglucosamine. *J Biol Chem* 278:10022–10027, 2003
- Parker G, Taylor R, Jones D, McClain D: Hyperglycemia and inhibition of glycogen synthase in streptozotocin-treated mice: role of O-linked N-acetylglucosamine. *J Biol Chem* 279:20636–20642, 2004
- Patti ME, Virkamaki A, Landaker EJ, Kahn CR, Yki-Jarvinen H: Activation of the hexosamine pathway by glucosamine in vivo induces insulin resistance of early postreceptor insulin signaling events in skeletal muscle. *Diabetes* 48:1562–1571, 1999
- Clark RJ, McDonough PM, Swanson E, Trost SU, Suzuki M, Fukuda M, Dillmann WH: Diabetes and the accompanying hyperglycemia impairs cardiomyocyte calcium cycling through increased nuclear O-GlcNAcylation. *J Biol Chem* 278:44230–44237, 2003
- Heckel D, Comtesse N, Brass N, Blin N, Zang KD, Meese E: Novel immunogenic antigen homologous to hyaluronidase in meningioma. *Hum Mol Genet* 7:1859–1872, 1998
- Comtesse N, Maldener E, Meese E: Identification of a nuclear variant of MGEA5, a cytoplasmic hyaluronidase and a beta-N-acetylglucosaminidase. *Biochem Biophys Res Commun* 283:634–640, 2001
- Schultz J, Pils B: Prediction of structure and functional residues for O-GlcNAcase, a divergent homologue of acetyltransferases. *FEBS Lett* 529:179–182, 2002
- Duggirala R, Blangero J, Almasy L, Dyer TD, Williams KL, Leach RJ, O'Connell P, Stern MP: Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. *Am J Hum Genet* 64:1127–1140, 1999
- Pratley RE, Thompson DB, Prochazka M, Baier L, Mott D, Ravussin E, Sakul H, Ehm MG, Burns DK, Foroud T, Garvey WT, Hanson RL, Knowler WC, Bennett PH, Bogardus C: An autosomal genomic scan for loci linked to prediabetic phenotypes in Pima Indians. *J Clin Invest* 101:1757–1764, 1998
- Ghosh S, Watanabe RM, Valle TT, Hauser ER, Magnuson VL, Langefeld CD, Ally DS, Mohlke KL, Silander K, Kohtamaki K, Chines P, Balow JJ, Birznieks G, Chang J, Eldridge W, Erdos MR, Karanjawala ZE, Knapp JI, Kudelko K, Martin C, Morales-Mena A, Musick A, Musick T, Pfahl C, Porter R, Rayman JB: The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. I. An autosomal genome scan for genes that predispose to type 2 diabetes. *Am J Hum Genet* 67:1174–1185, 2000
- Vionnet N, Hani E, Dupont S, Gallina S, Francke S, Dotte S, De Matos F, Durand E, Lepretre F, Lecocoeur C, Gallina P, Zekiri L, Dina C, Froguel P: Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21–q24. *Am J Hum Genet* 67:1470–1480, 2000
- Wiltshire S, Hattersley AT, Hitman GA, Walker M, Levy JC, Sampson M, O'Rahilly S, Frayling TM, Bell JI, Lathrop GM, Bennett A, Dhillo R, Fletcher C, Groves CJ, Jones E, Prestwich P, Simecek N, Rao PV, Wishart M, Bottazzo GF, Foxon R, Howell S, Smedley D, Cardon LR, Menzel S, McCarthy MI: A genomewide scan for loci predisposing to type 2 diabetes in a U.K. population (the Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. *Am J Hum Genet* 69:553–569, 2001
- Duggirala R, Stern MP, Mitchell BD, Reinhart LJ, Shipman PA, Uresandi OC, Chung WK, Leibel RL, Hales CN, O'Connell P, Blangero J: Quantitative variation in obesity-related traits and insulin precursors linked to the OB gene region on human chromosome 7. *Am J Hum Genet* 59:694–703, 1996
- Sobel E, Papp JC, Lange K: Detection and integration of genotyping errors in statistical genetics. *Am J Hum Genet* 70:496–508, 2002
- Lander ES, Green P: Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* 84:2363–2367, 1987
- Matise TC, Perlin M, Chakravarti A: Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. *Nat Genet* 6:384–390, 1994
- Boehnke M: Allele frequency estimation from data on relatives. *Am J Hum Genet* 48:22–25, 1991
- Almasy L, Blangero J: Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 62:1198–1211, 1998
- Heath SC: Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* 61:748–760, 1997

31. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I: VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* 16:1046–1047, 2000
32. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* 13:721–731, 2003
33. Zhang EY, Fu DJ, Pak YA, Stewart T, Mukhopadhyay N, Wrighton SA, Hillgren KM: Genetic polymorphisms in human proton-dependent dipeptide transporter PEPT1: implications for the functional role of Pro586. *J Pharmacol Exp Ther* 310:437–445, 2004
34. Expert Committee on the Diagnosis and Classification of Diabetes Mellitus: Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care* 20:1183–1197, 1997
35. American Diabetes Association: Summary of revisions for the 2002 Clinical Practice Recommendations. *Diabetes Care* 25 (Suppl. 1):S3, 2002
36. Duggirala R, Williams JT, Williams-Blangero S, Blangero J: A variance component approach to dichotomous trait linkage analysis using a threshold model. *Genet Epidemiol* 14:987–992, 1997
37. Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 30:97–101, 2002
38. Boerwinkle E, Chakraborty R, Sing CF: The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. *Ann Intern Med* 50:181–194, 1986
39. Nyholt DR: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 74:765–769, 2004
40. Abecasis GR, Cookson WO, Cardon LR: Pedigree tests of transmission disequilibrium. *Eur J Hum Genet* 8:545–551, 2000
41. Almasy L, Blangero J: Exploring positional candidate genes: linkage conditional on measured genotype. *Behav Genet* 34:173–177, 2004
42. Silander K, Mohlke KL, Scott LJ, Peck EC, Hollstein P, Skol AD, Jackson AU, Deloukas P, Hunt S, Stavrides G, Chines PS, Erdos MR, Narisu N, Conneely KN, Li C, Fingerlin TE, Dhanjal SK, Valle TT, Bergman RN, Tuomilehto J, Watanabe RM, Boehnke M, Collins FS: Genetic variation near the hepatocyte nuclear factor-4 α gene predicts susceptibility to type 2 diabetes. *Diabetes* 53:1141–1149, 2004
43. Farook VS, Bogardus C, Prochazka M: Analysis of MGEA5 on 10q24.1-q24.3 encoding the beta-O-linked N-acetylglucosaminidase as a candidate gene for type 2 diabetes mellitus in Pima Indians. *Mol Genet Metab* 77:189–193, 2002