Published in final edited form as:
Genomics. 2008 October ; 92(4): 226–234. doi:10.1016/j.ygeno.2008.06.004.

Heterogeneity in Gene Loci Associated with Type 2 Diabetes on Human Chromosome 20q13.1

J. L. Bento[a,b,*], **N. D. Palmer**[a,b,*], **M. Zhong**[b], **B. Roh**[a,b], **J. P. Lewis**[b,c], **M. R. Wing**[b,c], **H. Pandya**[c], **B. I. Freedman**[d], **C. D. Langefeld**[e], **S. S. Rich**[f], **D. W. Bowden**[a,b,c], and **J. C. Mychaleckyj**[f]

a*Department of Biochemistry, Wake Forest, University School of Medicine, Winston-Salem, North Carolina, 27157, USA*

b*Center for Human Genomics, Wake Forest, University School of Medicine, Winston-Salem, North Carolina, 27157, USA*

c*Molecular Genetics, Wake Forest, University School of Medicine, Winston-Salem, North Carolina, 27157, USA*

d*Department of Internal Medicine, Wake Forest, University School of Medicine, Winston-Salem, North Carolina, 27157, USA*

e*Department of Public Health Sciences, Wake Forest, University School of Medicine, Winston-Salem, North Carolina, 27157, USA*

f*Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia 22902, USA.*

Abstract

Human chromosome 20q12-13.1 has been linked to type 2 diabetes mellitus (T2DM) in multiple studies. We screened a 5.795Mb region for diabetes-related susceptibility genes in a Caucasian cohort of 310 controls and 300 cases with T2DM and end stage renal disease (ESRD), testing 390 SNPs for association with T2DM-ESRD. The most significant SNPs were found in the perigenic regions, *HNF4A* (hepatocyte nuclear factor 4-alpha), *SLC12A5* (potassium-chloride cotransporter member 5), *CDH22* (cadherin-like 22), *ELMO2* (engulfment and cell motility 2), *SLC13A3* (sodium-dependent dicarboxylate transporter member 3), and *PREX1* (phosphatidylinositol 3,4,5-triphosphate-dependent RAC exchanger 1). Haplotype analysis found 6 haplotype blocks globally associated with disease ($p<0.05$). We replicated the *PREX1* SNP association in an independent case control T2DM population, and inferred replication of *CDH22*, *ELMO2*, *SLC13A3*, *SLC12A5*, and *PREX1* using *in silico* perigenic analysis of two T2DM genome wide association study (GWAS) data sets. We found substantial heterogeneity between study results.

Keywords

Type 2 diabetes; end stage renal disease; SNPs; chromosome 20; linkage disequilibrium; haplotypes; association; heterogeneity

Corresponding Author: Josyf C. Mychaleckyj, Center for Public Health Genomics, University of Virginia, West Complex, Charlottesville, VA 22908, E-mail: jmychale@virginia.edu, Phone: (434) 982-1107, Fax: (434) 982-1815.
*These authors contributed equally to this study

INTRODUCTION

The human 20q12-13.1 region has been linked to T2DM in multiple studies. In our laboratory, efforts to identify diabetogenic genes have included fine mapping and physical mapping studies [1;2;3] and analysis of specific genes such as protein kinase C binding protein 1 (*PRKCBP1*), hepatocyte nuclear factor 4 alpha (*HNF4A*), *GLUT10* glucose transporter (*SLC2A10*)[2;3;4; 5], and multiple additional genes [4;5;6]. While other investigators have reported evidence for association of *HNF4A* with T2DM [7;8], we have observed limited evidence for a significant contribution of *HNF4A* to T2DM susceptibility in our population [9]. We have reported evidence of association with the *PTPN1* gene to T2DM and measures of glucose homeostasis [10;11]. Recently, efforts to identify novel diabetes susceptibility genes in the region have been reported in a Japanese sample [12] and in a combined United Kingdom/Ashkenazi Jewish sample [13]. Systematic map coverage of previously linked genomic regions provides a targeted strategy for determining genetic contributions to disease susceptibility.

As the availability of high throughput methods for genotyping has increased and costs have dropped, high-density SNP maps have emerged as a practical approach for identifying genes through systematic association analysis of extended chromosomal regions, and very recently, whole genome analysis. Systematic genotyping over contiguous regions has suggested that much of the genome is organized into discrete blocks of variable length defined by SNPs in high linkage disequilibrium (LD) with limited haplotype diversity [14;15]. The haplotype block structure of the genome in combination with dense SNP maps can potentially facilitate detection of disease-associated genes (e.g. [16;17;18]).

The use of extended, high density SNP maps enhances physical coverage of positional candidate genome segments and with judicious sampling of the marker allele frequency spectrum, can preserve nominal power to detect common variants across a significant proportion of the physical map. In general, little is known about the genetic variation or disease risk models that underlie a positional cloning region, whether the region has been identified through prior linkage or restricted association analysis. Pertaining to the 20q13.1 region, the question of whether the linkage and association signals result solely from variation in the *HNF4A* gene or include other gene locus variants is still undecided.

We have screened a 6Mb region on 20q13.1 for association with T2DM testing common SNPs identified in this region under multiple tests of association. We selected the most significant associated SNPs for replication typing in an independent Caucasian case control cohort, and performed *in silico* gene replication analysis of the region using two public GWAS data sets.

RESULTS

The goal of this study was to identify gene loci in the human chromosome 20q13.1 interval associated with susceptibility to T2DM. We developed a SNP map to screen for initial association in a rural Caucasian American population, recruited from Forsyth and adjacent counties, North Carolina, and then analyzed associated genes for replication. The screen employed 390 SNPs covering a 5.795Mb region (Build36:41.4–47.2 Mb), which we genotyped in a case-control population consisting of 300 Caucasian co-morbid individuals with T2DM and end-stage renal disease (T2DM-ESRD), and 310 Caucasian controls, Table 1. This region of 20q encompasses the highest likelihood region reported in our earlier genetic linkage study [1], overlaps regions of strongest evidence for linkage in other linkage studies [19;20], and includes multiple plausible T2DM candidate genes. Figure 1A and Figure 1B illustrate features of the resulting map. There are 98 annotated gene loci (NCBI Refseq 36.2), 77 of which are characterized, non-pseudogenes. There is a modest-sized gene desert (45.848–46.674Mb) in

the distal half of the map, flanked by *SULF2* and *PREX1* genes, containing a single pseudogene *SRMP1* (spermidine synthase pseudogene 1). SNPs were placed on the map at a greater density within specific candidate diabetogenic genes such as hepatocyte nuclear factor 4 alpha (*HNF4A*; 23 SNPs/111kb) and glucose transporter 10 (*SLC2A10*; 15 SNPs/50 kb). The map has a mean density of 1 SNP/14.9kb with maximum inter-SNP distance of 46.6kb (between rs1983 and rs244128, Figure 1B 42.556Mb) (Supplementary Table 1). Eleven of 390 SNPs were inconsistent with HWE in controls at a nominal exact $p<0.05$ level (exact min p=0.0035). One of these SNPs was also out of HWE in cases (rs3091566, 45.288Mb, p=0.0074) located in the *EYA2*/*PRKCBP1* region, an otherwise unremarkable region in this map (Supplementary Table 1).

Using the Gabriel block definition [15] with a haplotype frequency threshold of 1%, the map contains 63 haplotype blocks of membership of two or more SNPs, with mean size of 27.2kb (0.277-156.857kb) capturing 211/390 (54.1 %) of the markers or 29.6% of the map by genomic distance. This block structure is partially revealed in Figure 1A and Figure 1B at the track labeled |*D*'|, where the absolute value of the inter-SNP D' of contiguous SNPs from the control population is shown. The LD structure present in the cases and controls was compared by evaluating the differences between contiguous SNP pair D' statistics of cases and controls, as shown in Figure 1A and Figure 1B, track δ|*D*'|.

Genotype data from cases and controls was tested for association with T2DM-ESRD under additive trend, dominant, and recessive models. In Figure 1A and 1B the *–log(p)* track is a plot of the association analysis results for each statistic for every SNP in the map. Eleven SNPs (0.03%) fall in the "suggestive" category with single point test significance of $p\leq0.01$, and lie in 7 perigenic regions, Table 2. Repetition of the statistical association tests using a logistic regression model including a European population substructure marker of ancestry did not substantively change the results, and suggestive SNPs remained in this category after adjustment, Table 2.

The associated regions include: 1) the well-studied MODY *HNF4A* gene locus (hepatocyte nuclear factor 4 A) at 42.47Mb, a single SNP with suggestive association (rs1885088, dominant model p=0.0073); 2) a region at 44.12 Mb in a 37.5kb LD block encompassing exons 22–26 of *SLC12A5* (potassium-chloride cotransporter) and the complete *NCOA5* (nuclear receptor coactivator 5) gene, including a single associated SNP (rs9074, recessive model p=0.0016); 3) a concentration of 3 associated SNPs in *CDH22* (cadherin-like 22), located in 2 haplotype blocks (44.3Mb, rs6074069, rs966567, rs3904166, dominant model, p=0.0019, 0.0014, 0.004 respectively); 4) *ELMO2* (engulfment and cell motility 2) contains two SNPs in this map, one of which in intron 13 shows suggestive association (rs2257545, 44.44Mb, additive p=0.0059); 5) *SLC13A3* (sodium-dependent dicarboxylate transporter, member 3) with 2 intronic SNPs (rs126917, rs2425885, intron 1 and 5, 44.65–44.69Mb), although not formally located in the same block; 6) *SULF2*/gene desert (sufatase 2) with one SNP (rs2426039, 46.0Mb) 7) *PREX1*/gene desert (phosphatidylinositol 3,4,5-triphosphate-dependent RAC exchanger 1) with 3 SNPs (rs2426039, rs926693, rs3924220, 46.0-46.7Mb).

The region with the most compelling evidence for association ($p\leq0.001$) is shown in Figure 1B within the 46.0-46.7Mb region 3' to the *PREX1* (phosphatidylinositol 3,4,5-triphosphate-dependent RAC exchanger 1) gene. The most significant SNP is rs3924220 (recessive p=0.00099), located 32kb 3' distal to the *PREX1* gene locus (gene is coded on the reverse strand to the physical sequence map 'forward' direction). This is one of three SNPs in Table 2 that lie in the 800kb 'gene desert' region flanked centromerically by *SULF2* (sulfatase 2) and *PREX1* at the telomere end. The two SNPs closest to *PREX1* are singletons outside haplotype blocks (rs926693, rs3924220) and the SNP closest to *SULF2* is in a block that encompasses *SULF2*. The three SNPs mark independent suggestive point association test signals; they have

at most one flanking SNP with nominal (reduced) signal, and the association signal is attenuated at more distant SNPs. Linkage disequilibrium also decays in the two intervals between the 3 SNPs as shown qualitatively in Figure 1B and quantitatively by the observation that multiple distinct blocks separate rs2426039 and rs926693 and rs926693 and rs3924220.

Using a step-down permutation analysis, we tested the screening SNPs for map-level significance under strong control of the FWER. None of the markers satisfied the $p<0.05$ map-level significant criterion for any model class (additive, dominant, recessive). The effective pseudo-independent Bonferroni factor that would correct the nominal p-values in each class of model tests is ~114–252, very similar in magnitude to the n=390 total point tests, with assumed linkage equilibrium.

We tested the top two SNPs in the *CDH22* and *PREX1* perigenic regions in an independent replication case-control sample from the DHS cohort, Table 3. We found one SNP was significant in the DHS sample at $p<0.05$ (*PREX1* rs926693, DHS additive p=0.05, dominant p=0.031) while the other 3 SNPs were not significant. For the 4 SNPs, the differences between the T2DM-ESRD and DHS control minor allele frequencies was −1.1 – +2.9% while for cases it was −0.8 – −5.0%. The most comparable frequencies for cases and controls were for the replicated SNP.

Using the summary association statistics from the DGI and WTCCC studies we plotted the case-control p-values for tests of association together with our 11 leading SNP screen p-values in an integrated genetic association plot, Figure 2. The DGI p-values are meta-analysis results combining statistics from analysis of approx 1,000 matched cases and controls with discordant sibship family-based tests of association of approximately 830 families under a combined z-score model, corrected for genomic control inflation. The WTCCC p-values are for additive and genotype tests of association using 2,000 T2DM cases and 3,000 combined 1958BBC and UKBS controls [21]. The DGI leading candidate region (cluster of 3 SNPs, 44.224Mb) is located between *CD40* and *CDH22* gene loci, ~1.5kb 3' to *CDH22* (lead SNP rs1321000, p=0.00021). There is a single WTCCC SNP $p<0.01$ in the region between 44.0Mb and the 3 suggestive *CDH22* SNPs at 44.3Mb (rs16991026, p=0.0036), which lies in intron 17 of *SLC12A5*. In contrast, the lead candidate in WTCCC is in the distal promoter of *ELMO2*, approx 10kb 5' to exon1 (rs1033475, p=9.7e-05). Other SNPs in the WTCCC in the gene desert region were significant at $p<0.01$ (rs10211734, 120kb 5' to *SULF2*, p=0.0017; rs11699291, 300kb 5' to *PREX1*, p=0.0049). For our other leading perigenic regions, there is a pair of WTCCC SNPs at 44.739–44.805Mb, near our two *SLC13A3* SNPs. The first rs6017968, p=0.0065 is located in intron1 of *SLC13A3*, while the other (rs742321, p=0.0059) is ~60kb telomeric to the first and is 8kb 3' to *SLC2A10*. There were no SNPs significant at $p<0.01$ in the *HNF4A* perigenic region in either GWAS study. Two of our 11 leading candidate SNPs are present in the published GWAS results (rs2257545 DGI; rs966567 DGI+WTCCC), although neither was significant in these studies ($p>0.1$). We found measurable heterogeneity between the DGI and WTCCC study results in this map interval, with the leading 5 perigenic regions at $p<0.01$ differing between studies. The leading WTCCC and DGI SNPs are listed in Supplementary Table 2.

In addition to the tests of SNP main effects in the initial screen, we also performed haplotype analyses within the 63 haplotype block regions of high LD. Six of the 63 blocks (9.5%) showed nominal association at $p<0.05$, block numbers 22 (*SLC12A5*), 30 (*ELMO2*), 33 (*ZNF334*), 35 (*SLC13A3*), 48 (*SULF2*), 51 (gene desert region) (Table 3). Seven of the 11 most associated SNPs in Table 2 lie in 6 distinct haplotype blocks of at least size two, but only three of these blocks showed nominal association under a global model of homogeneity in cases and controls using simulation with haplo.score [22]: block 22 (rs9074), block 30 (rs2257545), and block 35 (rs126917). The other 3 nominally associated blocks (33: SNPs rs2780231-rs366860; 48:

SNPs rs425433-rs1889177; 51: SNPs rs6066563-rs2664588) do not contain a suggestive main effects SNP (Table 2, p<0.01). Block 48 is a 51.2kb block that spans intron 5–9 of *SULF2* (global sim p=0.043) containing SNP rs425433 which has a modest point association under dominant (p=0.02) and additive models (p=0.046) models.

Table 4 shows the detailed haplotype association results for two SNP LD block of *ZNF334*. The association (global sim. p=0.00105) is interesting in that neither of the two member SNPs (rs2780231, rs366860) show nominal point association with T2DM under any model (p>0.1). The maximum missing data for either of these two SNPs in cases or controls is 2.7% and both SNPs are in HWE in cases and controls. The global association is entirely driven by a strong effect in a rare risk haplotype (4% cases vs. 1% controls). The EM haplotype frequencies are robust to 'seeding' with copies of the rare haplotypes (data not shown). This block encompasses exon 1 – intron 2 of *ZNF334* (zinc finger protein 334) which is putatively a transcriptional factor or cofactor.

DISCUSSION

We have carried out an initial screen and replication analysis for T2DM-ESRD susceptibility genes in a 6Mb region of 20q13.1, a specific region implicated in previous linkage analyses which increases prior belief of association. We detected 53 of 390 SNPs (13.6%) that exhibited nominal association in one or more point tests of association (MIN3 statistic) under standard 1 df disease models (additive, dominant, recessive). At a more stringent significance level of p<0.01, 11 SNPs (0.03%) in 7 perigenic regions retained significance. The proportion of significant SNPs at α=0.05 was 6.4%, 7.9%, 5.6% (additive, dominant, recessive, respectively) and 1.5%, 1.3%, 0.8%, respectively, at α =0.01, comparable or very slightly elevated relative to the expectation under the complete null hypothesis for each class of independent tests assuming linkage equilibrium. The very slight increase could reflect an increased rate of false positives stemming from genotyping error, latent population substructure, or may represent common variants that contribute net modest additional genetic risk to T2DM-ESRD susceptibility. Inclusion of a marker of the major European axis of population ancestry (rs4988235) in our statistical models had little effect on our screening population association results, validating our belief that the strategy of simultaneous recruitment of cases and controls from the same rural regional Eastern Atlantic Caucasian population minimized systematic ancestry differences between cases and controls. None of the SNPs in the initial screen reached rigorous Bonferroni or permutation-based map-level significance in point tests of association corresponding to 390 tests for each genetic model or the composite MIN3 statistic. This is unsurprising since the screening study was not powered for a significance threshold of $p=10^{-5}$ under a common variant model with odds ratio ~ 1.3–1.5. We supplemented the screen by testing our leading SNPs in a replication case-control cohort, and integrated published GWAS data for this region also.

We tested 4 of our leading candidate SNPs from the most highly associated *PREX1*/gene desert and *CDH22* regions in a larger independent diabetic case control cohort (DHS) for replication. One SNP in the *PREX1*/gene desert locus was significant at p=0.05 increasing our confidence that this SNP signals a perigenic region that contains true variants associated with T2DM. By integrating analysis of the results from two separate GWAS data sets, we were able to infer replication in at least one GWAS within all of our perigenic regions with the exception (ironically) of *HNF4A*, the MODY risk locus and best studied of T2DM candidate genes in this interval. Two of our 11 leading candidate SNPs are present in the published GWAS results, although neither was significant in these studies. This suggests either that there is allelic heterogeneity at the associated gene loci and different case control populations are detecting different frequent gene variants, or that the true variants are the same but differing patterns of

linkage disequilibrium and SNP selection biases result in differing lead study SNPs for each of the predisposing genes.

We chose to perform multiple tests of association at each SNP to maximize the likelihood of detecting association under an unknown disease model of risk. The choice of association test statistic requires a balance between power and tolerance of type 1 error, and test performance will vary subtly with the latent genotype risk ratios, level of genotyping error, and point values of Hardy-Weinberg disequilibrium [23;24;25;26]. Allelic and genetic heterogeneity in the region may mean that no single test statistic is optimal for detecting all contributing alleles or genes, even if the variants are relatively common. The downside of using multiple tests is the increased probability of false positive results, albeit that the test statistics are mutually correlated.

The most significant SNP from the initial screen that was validated in the DHS replication cohort (rs926693, mean minor allele frequency 47.9% cases, 43.5% controls, additive model), lies in an 800kb 'gene desert' of the map, approx 160kb 3' to the *PREX1* gene with an odds ratio of 1.35 in the T2DM-ESRD screening cohort and 1.20 in the DHS replication. Two other SNPs also lie in this same gene desert region (45.9 – 46.7Mb, flanked by *SULF2* and *PREX1*) but the leading SNP from the screen was not significant in replication. The WTCCC GWAS analysis also found one SNP at $p<0.01$ in this desert region. *PREX1* does not have an obvious biological role with T2DM, but its protein product is involved in cell signaling pathways as a guanine nucleotide exchange factor for the RHO family of small GTP-binding proteins [27]. There are other clusters of spliced ESTs in this desert region that have not been characterized, which could also be functional sites of variation. We checked for the possibility of artifactual association and HWE results, but the mapped *PREX1* SNPs do not lie in any LINE or SINE human repeats or long segmental duplications [28].

Our screening map identified signals at two other perigenic regions that contained the most highly associated SNPs in this interval in both of the GWAS studies, *ELMO2* at 44.43Mb (WTCCC) and *CDH22* at 44.3Mb (DGI). The *CDH22* gene locus contains three suggestive SNPs in two blocks, although 2 of these were not significant in DHS replication. *CDH22* encodes a calcium-dependent cell adhesion protein that may play a role in morphogenesis and tissue development, and could be involved in the development of diabetes-related tissues, e.g. pancreas and liver [29].

HNF4A gene SNPs, the subject of recent reports of association with T2DM [7;8], generally showed modest association, although one SNP did reach suggestive significance (rs1885088) in the initial screen. Association of *HNF4A* with T2DM in this population has been explored in a previous publication [9] but was only tested under an additive model. The more extensive analysis here suggests that SNP dominant/recessive effect models are more important than purely additive in our population and screening map. *HNF4A* was not significant in either GWAS analysis at $p<0.05$.

The question of why our study results seem to corroborate genes in both GWAS studies separately and why there is noticeable heterogeneity between the GWAS analyses is still open. The lack of a stronger GWAS signal at *HNF4A* is also intriguing. While the GWAS analyses did find similar genes in the extreme tail of the whole genome statistics [30;31] this study suggests that there is detectable genetic heterogeneity lower in the statistical association ranking, concomitant with weaker effect sizes. This could be a result of the differing clinical characteristics of the case control populations or may be due to differing ancestral genetic background within the broad Caucasian race category, exposing the importance of different gene variants and shuffling of their relative importance. The DHS cases had a larger mean BMI (33.0kg/m2) than T2DM-ESRD screening cases (28.5), WTCCC(30.7) or DGI(28–29 range),

while the T2DM-ESRD cases had a mean diabetes duration of >15yrs compared to DHS (11.2yrs), WTCCC(8.3yrs) and DGI(5–9yrs range).

Other perigenic regions with suggestive association from the initial screen were exons 22–26 of the *SLC12A5* (potassium-chloride cotransporter) gene and the entire *NCOA5* (nuclear receptor coactivator 5) gene at 44.12Mb; and *SLC13A3* (sodium-dependent dicarboxylate transporter, member 3) at 44.65–44.69Mb. *SLC12A5* is a potassium chloride cotransporter that normally lowers intracellular chloride concentrations below the electrochemical equilibrium potential [32] but is not expressed in tissues of direct interest in T2DM or nephropathy. *NCOA5* encodes a coregulator for the alpha and beta estrogen receptors [33]. Alpha estrogen receptor knockout mice are insulin resistant, have impaired glucose tolerance, and are obese [34]. *NCOA5* acts as a corepressor and coactivator of the alpha estrogen receptor.

Our group previously reported evidence of association of the chromosome 20q *PTPN1* gene with T2DM and measures of glucose homeostasis [10;11]. The evidence of association to T2DM observed in this study, eg *CDH22*, are comparable to the odds ratios observed with *PTPN1* (1.3–1.5) and consistent with multiple genes in this region contributing risk to T2DM.

Our results are consistent with previous studies and replicate prior association reports with the genes *HNF4A*, *CDH22*, *ELMO2*, and *PREX1*. There is an extensive literature on association of *HNF4A* in MODY and older onset T2DM [7;8;9]. A recent study [12] of 675 Japanese cases of T2DM and 474 controls across a 17.7Mb region of 20q at a comparable density (1 per 15.4Mb vs. 14.7Mb here) found maximal association with the *LBP* gene (lipopolysaccharide binding protein) which is 5Mb centromeric to our map. However, they found secondary association of diabetes with *ELMO2* (rs2297056 p=0.010 and rs2257545 p=0.006) and *PREX1* (rs13044736 p=0.006) genes. Four recent GWAS studies found consistent 20q association at rs6030447 (odds ratios 0.99–2.53) which is located in the protein tyrosine phosphatase, receptor type T gene (*PTPRT*), 1Mb centromeric to our map [35;36;37;38]. Since our map boundaries were originally defined based on a LOD-1 maximal linkage criterion [2], it is possible that this gene could also contribute to our diabetes linkage signal, and association also.

Haplotype association analysis was performed for all 63 haplotype blocks in the map. Six of 63 blocks showed nominal global association ($p<0.05$) although no block was significant at the n=63 Bonferroni-corrected level. Only seven of the 11 associated SNPs fell in haplotype blocks, and only three of these were nominally associated. This suggests that the association signals of the four remaining SNPs were partially and completely scrambled in the three non-associated haplotype blocks. The 10 SNP block 22 containing rs9074 (SLC12A5-NCOA5) showed evidence of a frequent risk and a protective haplotype, the risk haplotype association being largely determined by the rs9074 allele tag. The two other nominally associated blocks contain associated SNPs displaying haplotype association patterns that suggest the association is largely driven by the singly associated SNP. We also found a 2-SNP haplotype block (rs2780231, rs366860) contained in the *ZNF334* gene locus for which neither SNP had suggestive or nominal point levels of significance, but which displayed the most significant block association. The association appears to result from a rare risk haplotype (4% cases versus 1% controls). There were also two other blocks devoid of associated SNPs, that showed more modest haplotype association (p=0.03, 0.04).

We compared our map coverage with another recently published T2DM GWAS [39] which used a HumanHap300 multiplex assay (Illumina, San Diego, CA) containing 317,503 SNPs. Based on our map interval ([2], Build36, chr20:41415109:47243708) the HumanHap300 assay contains 701 SNPs giving a relative physical coverage of 1 per 8.3kb (5.828Mb/701) versus our 1 per 14.9kb for 40% greater physical density. The HumanHap300 coverage is 10.7%

higher than the 633 expected SNPs from the mean non-redundant autosomal density of HumanHap300 (308,330 SNPs spanning 2.868Gb). We estimated our map tag coverage using HapMap data. In release22 of HapMap (Mar 2007) the same map interval contains 11,947 typed SNPs; 336/390 of our map SNPs were typed in this release. Haploview identifies a minimal pairwise tagSNP set of 1343 in this interval for the 30 CEPH families, tagging 4965 SNPs with MAF>0.05 at r^2>0.8. Of the 1343 tagSNPs, 98 are included in our map. At r^2>0.8, the 336 HapMap-included SNPs in our map tag 2020 HapMap SNPs at MAF >0.05, for a relative map coverage of 40.7% compared to release22 HapMap, while at r^2>0.5, the relative coverage is 56.5% (2804/4965). Since our map contains 54 additional non-HapMap typed SNPs, the map coverage will be larger than these estimates.

Power to detect variants is an important consideration during the design of any genetic epidemiological study. The power here is sufficient to reliably detect variants with relatively large effect sizes of odds ratio ~1.5, comparable to effect size for the recently discovered T2DM susceptibility gene TCF7L2 transcription factor 7-like 2 [40] or very frequent variants (>10%), this is a screening study designed to identify variants for replication testing. For odds ratio ~1.3 (comparable to the functional candidate gene *PPARG*, [41]), nominal power is low, but this is likely to be an underestimate of effect size of one or more loci in this region, given the prior replication of linkage signals in multiple studies, which are typically underpowered for frequent variants of modest effect. Furthermore the study design herein also assumes that there are common variants (MAF >10%) that contribute to disease risk. It will not detect genes where the susceptibility is distributed amongst rare variants with a high degree of allelic heterogeneity. The controls in this study were not tested for diabetes, and the small proportion of undiagnosed T2DM cases may have reduced our power and led to underestimation of odds ratios.

Since the case population was ascertained by diagnosis of T2DM with ESRD with mean duration of diabetes >15 years, there is a possibility that we have detected diabetic nephropathy/ESRD predisposing alleles. Since the DHS replication T2DM cases were recruited with exclusion of nephropathy, the replication of the *PREX1*/gene desert region suggests that at least this region is associated with the primary T2DM clinical phenotype. Based on the expression and known biology of these gene loci, *SLC13A3* (solute carrier family 13, sodium-dependent dicarboxylate transporter, member 3) is the most likely of the genes to affect renal function, although we cannot absolutely exclude any of the genes. This is a high-affinity transporter of Krebs cycle dicarboxylate intermediates, which localizes to the basolateral membrane of human renal proximal tubule [42]. Further study in a diabetic population with clear adjudicated lack of nephropathy will be necessary to fully answer the question of disease phenotype specificity. Other recent large GWAS case-control population studies of T2DM did not enforce case exclusionary criteria based on a clinical diagnosis of diabetic nephropathy.

In summary, we have analyzed an important linkage region for type 2 diabetes, and have replicated and extended the analysis of genes previously reported to be associated with T2DM. We have also identified new loci that warrant further detailed analysis. While these genes may not have the strongest effect genome-wide, their consistent replication suggests that more than one gene in this region contributes to T2DM-ESRD pathology.

MATERIALS AND METHODS

Subjects

Samples for initial screen consisted of 300 unrelated Caucasian T2DM patients with end-stage renal disease (ESRD), and 310 randomly-ascertained unrelated Caucasian subjects without known diabetes at recruitment. Cases and controls were recruited simultaneously from the 10 county region around Forsyth County, North Carolina. All subjects had parents who were born in North Carolina. The diabetes cases will be referred to as “T2DM-ESRD”; their ascertainment

and recruitment have been previously described in detail [10]. Ascertainment of cases based on a diagnosis of diabetes-related ESRD results in a group of cases with one of the most prevalent and lethal complications of diabetes. While random ascertainment of the control group may lead to some loss of power due to undiagnosed or future development of diabetes, this ascertainment scheme is more robust to potentially unperceived population stratification. The subjects in this case-control sample are independent from, though ascertained and recruited in an identical manner to, family samples described by [1]. T2DM-ESRD subjects had a mean age at diagnosis of diabetes of 46.5 ± 12.8 years, mean BMI at recruitment of 28.5 ± 7.0, mean duration of diabetes greater than 15 years, and mean HbA1c of 8.6%. The Caucasian control individuals had a mean age of 45.8 years with a mean BMI of 25.7.

Replication genotyping of selected SNPs was performed in T2DM-affected probands from the Diabetes Heart Study (DHS) [43]. The DHS is a single-center family-based study of atherosclerosis in T2DM-enriched families. Only the probands of each family were used as a case subject. Four hundred seventy T2DM-affected proband cases were genotyped (mean age of diagnosis of diabetes 51.6 years, mean BMI 33.0 kg/m2, mean duration of diabetes 11.2 years, mean HbA1C 7.5%). Unlike the T2DM-ESRD cases, DHS participants were recruited with an exclusion of diabetic nephropathy. As a control population for this group we genotyped 442 unrelated healthy self-declared Caucasian subjects (mean age 57.3 years, mean BMI of 28.1).

SNP Selection and Genotyping

The 20q13.1 region mapped in this study is identical to that described in [2] and encompasses the highest likelihood linkage region from the combined analyses carried out by the NIDDK-supported International Type 2 Diabetes Genetic Mapping Consortium (data not shown). SNPs were selected from HapMap, NCBI dbSNP, and Appelera/Celera Discovery System to map the build36 region (identical in build35) chr20:41415109-47243708; NT_011362.9:7034610-12863209 (5,828,600bp) with a bias for validated SNPs. Minor allele frequencies ranged 0.030–0.50 in controls (median=0.30), and 0.027–0.50 in cases (median=0.31) (Supplementary Table 1). All SNP amplification primers were designed for amplicon uniqueness and mapped using NCBI BLAST/UCSC BLAT. Primers that mapped to low copy number segmental duplications or interspersed repeats were redesigned, or alternative SNPs selected if necessary. Gene annotations were extracted from NCBI contig NT_011362, version 9. The SNP map (Figures 1A, 1B) was drawn using the perl program MAPPLOT (unpublished work). Within-European stratification was tested using SNP rs4988235 [44] which is a perfect marker of lactase persistence and highly informative for European North-South ancestral gradient.

Genotyping was performed on a Sequenom MassArray Genotyping System using methods previously described [10;45;46]; 331 SNPs were typed with hME chemistry, 50 with iPLEX, and 9 with both. Specific primer sequences are available upon request.

Quality Control

1. Manual Review. Each genotype spectrum was manually reviewed by a lab member very experienced with the Sequenom platform. During review, doubtful genotypes were preferentially changed to missing rather than risk misclassification. ***2. QC Duplicates and Sample Blanks*** In control plates, 2 samples were duplicated at 5 separate well addresses, while in cases, 8 samples were duplicated at 2 well addresses. Each plate also contained 2 water blanks. If there was >1 discordant genotype in QC replicates, or genotype signals were seen in the negative control blank wells, the SNP genotyping was rejected and repeated. Depending on the type of QC error, SNPs were repeated in single or multiplex mode, and typed /reviewed by a second experienced lab member. ***3. Hardy-Weinberg Equilibrium (HWE).*** Genotyped

SNPs were tested for departures from Hardy-Weinberg equilibrium (HWE) using an exact test [47]. SNPs not in HWE were repeated. After repeat typing, we rejected any SNP where the exact test was significant at $p<0.001$ in controls, resulting in rejection of 11 SNPs. In two SNPs (rs733379 controls and rs2297201 cases) the HWE p-value was close to the 0.001 cutoff, and the SNP genotype patterns were manually reviewed. ***4. Missing Data.*** For any SNP missing >10% of data or Hardy-Weinberg p-value <0.001 in cases or controls, an experienced lab staff member reviewed the 2-dimensional genotype cluster plots to verify patterns and cluster separation. The maximum missing data in controls was 14.5% (mean 2.7%) and in cases 13% (mean 2.9%), with 16 SNPs missing >10% data in cases. Of these 16, only 1 SNP had a HWE test significance of $p<0.05$ in cases (rs1984076), and only 1 showed evidence of association $p<0.05$ (1 df recessive model, rs2425941). In situations where there was any doubt in the genotype calling, SNP cluster plots were manually reviewed for adequate cluster separation. While departure from HWE may indicate unresolved problems with assay accuracy, it may also reflect true disease association. The association test statistics chosen are somewhat robust to departures from HWE [24], but may still lead to inflated type 1 error rates under circumstances where genotypes are differentially misclassified between cases and controls. All of the most significant 'suggestive' associated SNPs were reviewed.

Statistical Analysis

The pairwise linkage disequilibrium statistics D' and r^2 were calculated using Haploview [48] and the haplo.em function in the R package haplo.stats (version 1.3.1, Mayo Clinic/Foundation, http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm). All SNPs were tested for association in the case-control study under the following 1 df models: Cochran-Armitage additive trend (1 df, scores = 0,1,2), dominant (1 df), and recessive (1 df). We ranked SNPs based on the minimum p-value of the preceding 3 tests, i.e. for the *kth* SNP:

$$\mathrm{MIN3}(k)=\min\{p^{(k)}_{\mathrm{add}},p^{(k)}_{\mathrm{dom}},p^{(k)}_{\mathrm{rec}}\}$$

This robust statistic was motivated by the MAX3 statistic used by [23] for case-control association tests using the Cochran-Armitage test, where the optimal relative score for heterozygote in the 3 genotype classes (AA, Aa, aa) is unknown. We estimated the power to detect nominal levels of association under Cochran-Armitage score test statistics using the refinement described in [23] derived by Slager and Schaid [22;25]. The parameters used were selected to be realistic of a T2DM complex disease common risk locus under a multiplicative disease risk model. Assuming a T2DM disease prevalence of 10% and multiplicative genotype risk ratios (GRR) of 1.5 or 1.3, the power of the screen to detect association at a nominal SNP significance level of 5% is respectively 72.9% or 36.5% under an additive trend test with risk allele frequency of 10%, and respectively 91.7%, or 56.3% for risk allele frequency equal to 20%. Under different disease models, the dominant and recessive score tests can outperform the additive trend.

We tested whether population stratification could account for our case control association results using a logistic regression model incorporating the within-European marker of substructure (rs4988235) as a covariate to the SNP main effect term.

Two methods were used to adjust for type 1 error rates across multiple single SNP tests of association. We filtered based on the heuristic criterion of $MIN3 \leq 0.01$ to obtain a set of SNPs "suggestive of association", i.e. at least one point test of association significant at the 0.01 level. To correct for map-level multiple testing we evaluated the association results using a permutation analysis to simulate the null joint distribution of marker statistics and applied step-down analysis on the ordered, adjusted p-values for each separate class of tests. This procedure adjusts for multiple map tests while maintaining control over the Family Wise Error Rate (FWER) at any desired level of significance.

We identified haplotype blocks in the combined case+control cohort under the null hypothesis of no differences in haplotype structure between cases and controls, for map SNPs with a minor allele frequency > 1% and a minimum haplotype frequency threshold of 1%, using the Gabriel 95% confidence interval algorithm [15] implemented in Haploview [48]. Haplo.score was used to test for case-control association of haplotypes within these haplotype blocks in a manner similar to that described in Bento et al [10] skipping haplotypes with frequencies <0.01. The map-wide significance for a single haplotype block global test of significance was set to the Bonferroni-corrected value of 0.05/63 total blocks = 0.0008. R version 2.3 (http://www.r-project.org) was used for all statistical analyses other than where explicitly mentioned.

GWAS data sets

We downloaded two publicly available T2DM case control data sets to test replication of our results in additional populations of larger sample size. The Diabetes Genetics Initiative (DGI) is a collaboration between Lund University, Broad Institute and Novartis Institutes for Biomedical Research. Full details of patient recruitment and the primary analysis are available at “Whole Genome Scan for Type 2 Diabetes in a Scandinavian Cohort” http://www.broad.mit.edu/diabetes/scandinavs/index.html [31]. We also used results from the Wellcome Trust Case Control Consortium (WTCCC) http://www.wtccc.org.uk/. Full study and analysis details are available at this site and in the companion publication [30]. We applied a genome coordinate filter to the data sets to identify the relevant region.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

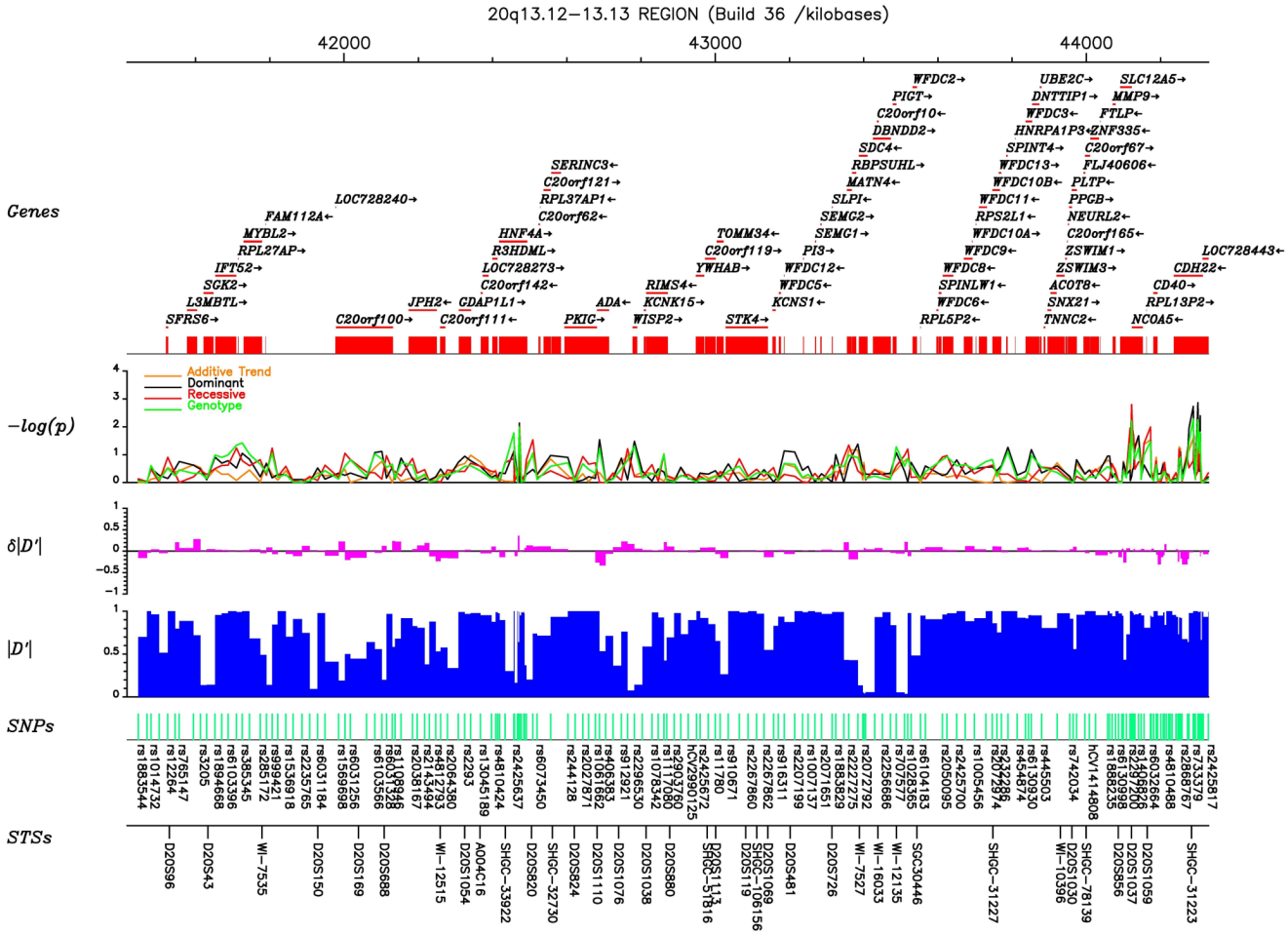
This work was supported by NIH grant R01 DK56289 to DWB.

References

1. Bowden DW, et al. Linkage of genetic markers on human chromosomes 20 and 12 to NIDDM in Caucasian sib pairs with a history of diabetic nephropathy. Diabetes 1997;46:882–886. [PubMed: 9133559]
2. Fossey SC, et al. A high-resolution 6.0-megabase transcript map of the type 2 diabetes susceptibility region on human chromosome 20. Genomics 2001;76:45–57. [PubMed: 11549316]
3. Price JA, et al. A physical map of the 20q12-q13.1 region associated with type 2 diabetes. Genomics 1999;62:208–215. [PubMed: 10610714]
4. Bagwell AM, Bailly A, Mychaleckyj JC, Freedman BI, Bowden DW. Comparative genomic analysis of the HNF-4alpha transcription factor gene. Mol Genet Metab 2004;81:112–121. [PubMed: 14741192]
5. Dawson PA, et al. Sequence and functional analysis of GLUT10: a glucose transporter in the Type 2 diabetes-linked region of chromosome 20q12-13.1. Mol Genet Metab 2001;74:186–199. [PubMed: 11592815]
6. Bento JL, et al. Genetic analysis of the GLUT10 glucose transporter (SLC2A10) polymorphisms in Caucasian American type 2 diabetes. BMC Med Genet 2005;6:42. [PubMed: 16336637]
7. Love-Gregory LD, et al. A common polymorphism in the upstream promoter region of the hepatocyte nuclear factor-4 alpha gene on chromosome 20q is associated with type 2 diabetes and appears to contribute to the evidence for linkage in an ashkenazi jewish population. Diabetes 2004;53:1134–1140. [PubMed: 15047632]
8. Silander K, et al. Genetic variation near the hepatocyte nuclear factor-4 alpha gene predicts susceptibility to type 2 diabetes. Diabetes 2004;53:1141–1149. [PubMed: 15047633]

9. Bagwell AM, et al. Genetic analysis of HNF4A polymorphisms in Caucasian-American type 2 diabetes. Diabetes 2005;54:1185–1190. [PubMed: 15793260]
10. Bento JL, et al. Association of protein tyrosine phosphatase 1B gene polymorphisms with type 2 diabetes. Diabetes 2004;53:3007–3012. [PubMed: 15504984]
11. Palmer ND, et al. Association of protein tyrosine phosphatase 1B gene polymorphisms with measures of glucose homeostasis in Hispanic Americans: the insulin resistance atherosclerosis study (IRAS) family study. Diabetes 2004;53:3013–3019. [PubMed: 15504985]
12. Takeuchi F, et al. Search of type 2 diabetes susceptibility gene on chromosome 20q. Biochem Biophys Res Commun 2007;357:1100–1106. [PubMed: 17466274]
13. Sandhu MS, et al. Common variants in WFS1 confer risk of type 2 diabetes. Nat Genet 2007;39:951–953. [PubMed: 17603484]
14. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. Nat Genet 2001;29:229–232. [PubMed: 11586305]
15. Gabriel SB, et al. The structure of haplotype blocks in the human genome. Science 2002;296:2225–2229. [PubMed: 12029063]
16. Chumakov I, et al. Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. Proc Natl Acad Sci U S A 2002;99:13675–13680. [PubMed: 12364586]
17. Hugot JP, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 2001;411:599–603. [PubMed: 11385576]
18. Rioux JD, et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 2001;29:223–228. [PubMed: 11586304]
19. Ghosh S, et al. Type 2 diabetes: evidence for linkage on chromosome 20 in 716 Finnish affected sib pairs. Proc Natl Acad Sci U S A 1999;96:2198–2203. [PubMed: 10051618]
20. Ji L, et al. New susceptibility locus for NIDDM is localized to human chromosome 20q. Diabetes 1997;46:876–881. [PubMed: 9133558]
21. Zeggini E, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 2007;316:1336–1341. [PubMed: 17463249]
22. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002;70:425–434. [PubMed: 11791212]
23. Freidlin B, Zheng G, Li Z, Gastwirth JL. Trend tests for case-control studies of genetic markers: power sample size and robustness. Hum Hered 2002;53:146–152. [PubMed: 12145550]
24. Sasieni PD. From genotypes to genes: doubling the sample size. Biometrics 1997;53:1253–1261. [PubMed: 9423247]
25. Slager SL, Schaid DJ. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. Hum Hered 2001;52:149–153. [PubMed: 11588398]
26. Zheng G. Use of max and min scores for trend tests for association when the genetic model is unknown. Stat Med 2003;22:2657–2666. [PubMed: 12898550]
27. Weiner OD. Rac activation: P-Rex1 - a convergence point for PIP(3) and Gbetagamma? Curr Biol 2002;12:R429–R431. [PubMed: 12123595]
28. Eichler EE. Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet 2001;17:661–669. [PubMed: 11672867]
29. Sugimoto K, et al. Molecular cloning and characterization of a newly identified member of the cadherin family, PB-cadherin. J Biol Chem 1996;271:11548–11556. [PubMed: 8626716]
30. Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–678. [PubMed: 17554300]
31. Saxena R, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 2007;316:1331–1336. [PubMed: 17463246]
32. Sallinen R, et al. Chromosomal localization of SLC12A5/Slc12a5, the human and mouse genes for the neuron-specific K(+)-Cl(−) cotransporter (KCC2) defines a new region of conserved homology. Cytogenet Cell Genet 2001;94:67–70. [PubMed: 11701957]

33. Sauve F, et al. CIA, a novel estrogen receptor coactivator with a bifunctional nuclear receptor interacting determinant. Mol Cell Biol 2001;21:343–353. [PubMed: 11113208]
34. Heine PA, Taylor JA, Iwamoto GA, Lubahn DB, Cooke PS. Increased adipose tissue in male and female estrogen receptor-alpha knockout mice. Proc Natl Acad Sci U S A 2000;97:12729–12734. [PubMed: 11070086]
35. Florez JC, et al. A 100K genome-wide association scan for diabetes and related traits in the Framingham Heart Study: replication and integration with other genome-wide datasets. Diabetes 2007;56:3063–3074. [PubMed: 17848626]
36. Hanson RL, et al. A search for variants associated with young-onset type 2 diabetes in American Indians in a 100K genotyping array. Diabetes 2007;56:3045–3052. [PubMed: 17846125]
37. Hayes MG, et al. Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. Diabetes 2007;56:3033–3044. [PubMed: 17846124]
38. Rampersaud E, et al. Identification of novel candidate genes for type 2 diabetes from a genome-wide association scan in the Old Order Amish: evidence for replication from diabetes-related quantitative traits and from independent populations. Diabetes 2007;56:3053–3062. [PubMed: 17846126]
39. Sladek R, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 2007;445:881–885. [PubMed: 17293876]
40. Grant SF, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet 2006;38:320–323. [PubMed: 16415884]
41. Altshuler D, et al. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 2000;26:76–80. [PubMed: 10973253]
42. Bai X, et al. Identification of basolateral membrane targeting signal of human sodium-dependent dicarboxylate transporter 3. J Cell Physiol 2006;206:821–830. [PubMed: 16331647]
43. Wagenknecht LE, et al. Familial aggregation of coronary artery calcium in families with type 2 diabetes. Diabetes 2001;50:861–866. [PubMed: 11289053]
44. Enattah NS, et al. Identification of a variant associated with adult-type hypolactasia. Nat Genet 2002;30:233–237. [PubMed: 11788828]
45. Beaulieu, M.; Hong, P. Application Note. 2004. Multiplexing the Homogeneous MassEXTEND Assay.
46. Oeth, P., et al. Application Note. 2005. iPLEX™ Assay: Increased Plexing Efficiency and Flexibility for MassARRAY. System Through Single Base Primer Extension with Mass-Modified Terminators.
47. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 2005;76:887–893. [PubMed: 15789306]
48. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263–265. [PubMed: 15297300]



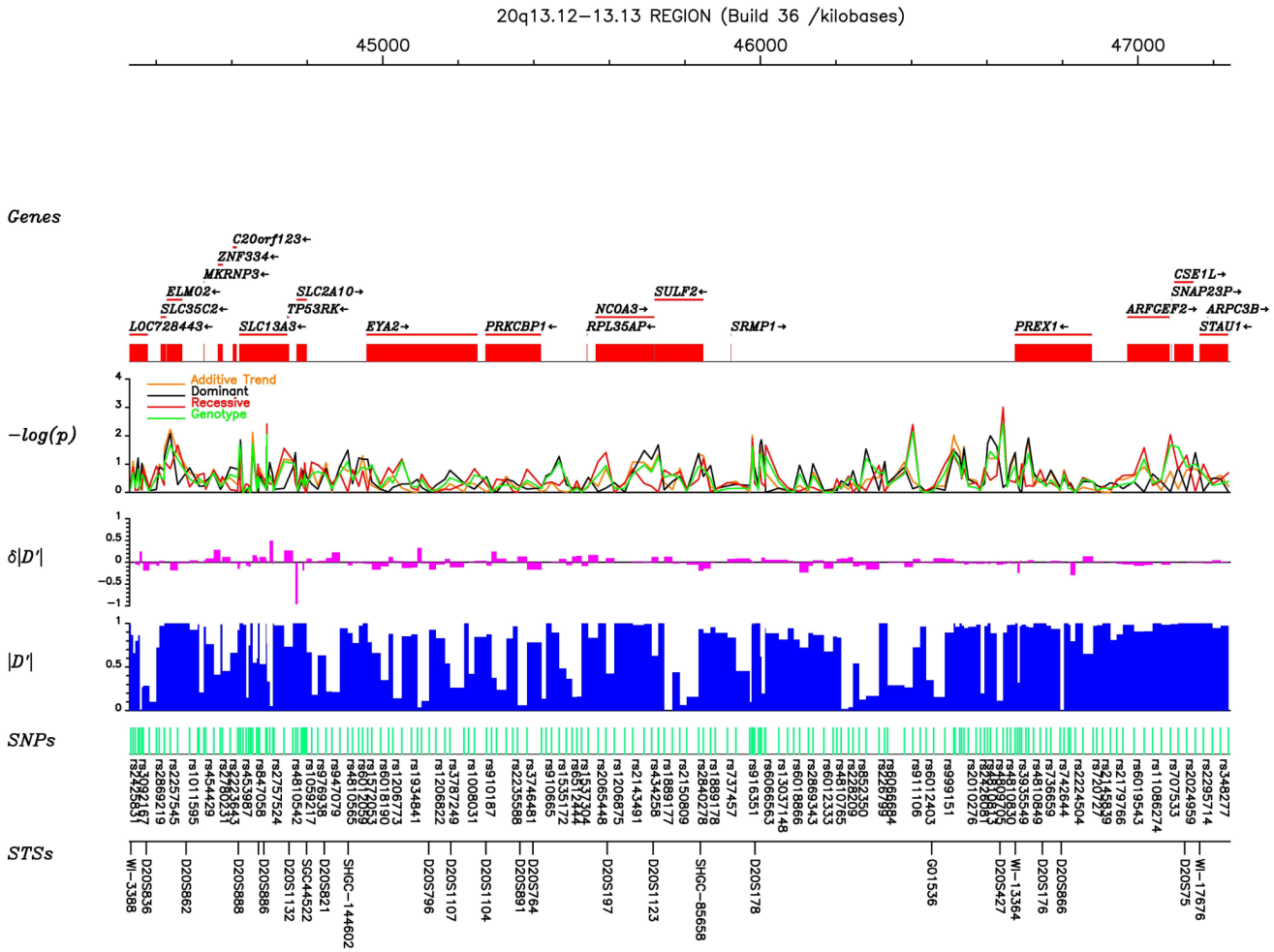


Figure 1. SNP map of 20q13.1

The contiguous 6 Mbases are shown in proximal and distal 3 Mbase panels. Chromosome 20 physical map distance (NCBI Build 36) is denoted along the top of the figure in kilobases, followed by the annotated genes in the *Genes* track, the association analysis for each individual SNP located in the *–log(p)* track, the difference between inter-SNP D' values in T2DM-ESRD cases and controls in the δ|*D'*| track, the inter-SNP D' values in this region as observed in the control population, the location of SNPs and other markers located in the region. Some SNP marker names have been omitted for clarity. A full listing of all markers can be found in Supplementary Table 1.

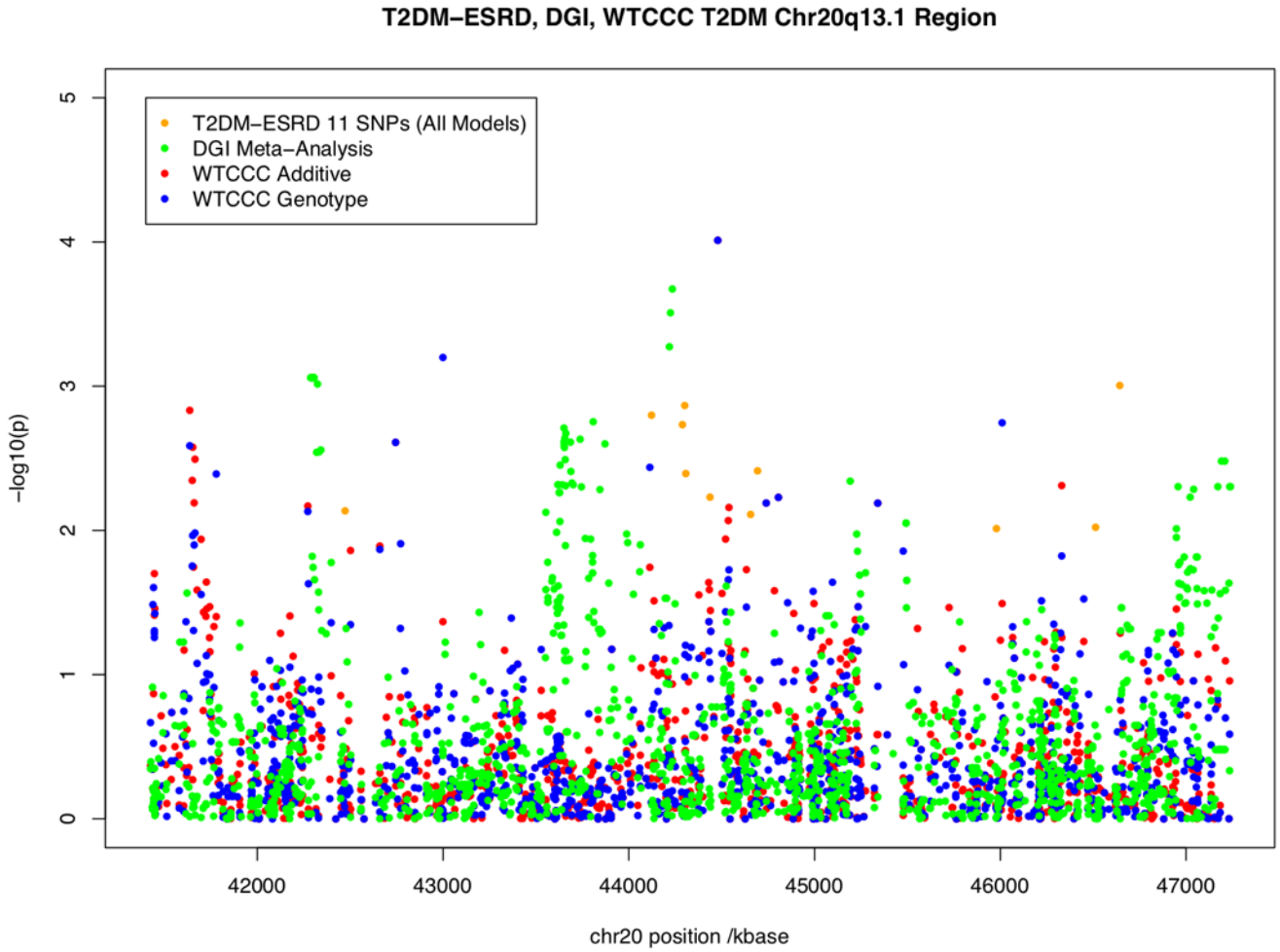


Figure 2. Integrated genetic association analysis of 20q13.1

This shows the $-\log_{10}$(p-value) for 3 separate T2DM case control studies plotted against chromosome 20 physical map position in kilobases. The current study (T2DM-ESRD 11 SNP) shows the top 11 SNPs from Table 2; DGI Meta-Analysis are the meta-analytical results from the Diabetes Genetics Initiative (Lund-Broad-Novartis), 1520 SNPs; and WTCCC are the Wellcome Trust Case Control Consortium additive (1df) and genotype (2df) results for 1074 SNPs. See also Supplementary Table 2.

Table 1

Characteristics of the chromosome 20q13.1 SNP map.

Total region surveyed	5.795Mb
Gene loci in region	98
	77 characterized, non-pseudo
	14 predicted genes
	7 pseudogenes
Number of SNPs	390
Mean SNP density (kb)	1 SNP/14.9kb
(Min - Max inter-SNP)	(0.033–46.6kb)
Number of blocks	63
Mean length of blocks	27.2kb
(Min - Max)	(0.277 - 156.857kb)
Total number of SNPs in blocks	211 (54.1%)
Sum length of all blocks	1714kb (29.6%)

Table 2

The most significant association signals detected in the 390 SNP map screen using additive, dominant, and recessive 1 df tests with p<0.01. The most significant test p-value is shown, although other models may also be significant at this threshold. The adjusted p-value model includes a covariate SNP for European population stratification (rs4988235). The annotation of the location of the SNP is relative to the nearest characterized gene. SNPs in the 'gene desert' region spanned by characterized genes *SULF2 – PREX1* are reported with the closer of the 2 flanking genes. The gene model position is relative to the gene model transcript (NM sequence accession) shown. Blocks are counted with 2 or more SNP members; singleton SNPs are not considered 'blocks'.

SNP	Position (kb)	Nominal p-value	Adjusted p-value	Model	Gene Locus / Map Region	Gene Model Location	Gene Model Transcript
rs1885088	42472454	0.00731	0.00518	Dominant	*HNF4A*	Intron 3	NM_178850
rs9074	44122072	0.00159	0.00181	Recessive	*SLC12A5*	3'UTR	NM_020708
rs6074069	44288133	0.00185	0.00090	Dominant	*CDH22*	Intron 3	NM_021248
rs966567	44300632	0.00136	0.00078	Dominant	*CDH22*	Intron 2	NM_021248
rs3904166	44306828	0.00404	0.00237	Dominant	*CDH22*	Intron 1	NM_021248
rs2257545	44437195	0.00589	0.00808	Additive	*ELMO2*	Intron 13	NM_133171
rs126917	44655342	0.00774	0.00580	Additive	*SLC13A3*	Intron 5	NM_022829
rs2425885	44693038	0.00386	0.00070	Recessive	*SLC13A3*	Intron 1	NM_022829
rs2426039	45978568	0.00971	0.00828	Additive	*SULF2*; Gene Desert	-	-
rs926693	46512364	0.00950	0.00868	Additive	*PREX1*; Gene Desert	-	-
rs3924220	46642852	0.00099	0.00107	Recessive	*PREX1*; Gene Desert	32kb 3' dist	NM_02820

Table 3

Comparison of the DHS cohort replication genotyping results for the most significant *CDH22* and *PREX1* perigenic SNPs from the T2DM-ESRD case-control screening analysis with empirical odds ratios under the most significant test model.

SNP	Position (kb)	Gene Locus / Map Region	Cohort	Major /Minor Alleles	Model	Allele Frequency Controls , Cases	p-value	OR (95% CI)
rs6074069	44288133	*CDH22*	DHS	G / A	Dominant	0.212 , 0.199	0.681	0.89 (0.45–1.74)
			T2DM-ESRD	G / A	Dominant	0.197 , 0.249	0.00185	4.87 (1.63–14.58)
rs966567	44300632	*CDH22*	DHS	G / C	Dominant	0.232 , 0.229	0.885	1.04 (0.60–1.82)
			T2DM-ESRD	G / C	Dominant	0.229 , 0.276	0.00136	3.51 (1.56–7.92)
rs926693	46512364	*PREX1*; Gene Desert	DHS	A / G	Additive	0.430 , 0.475	0.05	1.20 (0.99–1.44)
			DHS	A / G	Dominant	as above	0.031	1.37 (1.03–1.83)
			T2DM-ESRD	A / G	Additive	0.441 , 0.483	0.00950	1.35 (1.08–1.70)
rs3924220	46642852	*PREX1*; Gene Desert	DHS	T / C	Recessive	0.224 , 0.244	0.1284	1.17 (0.96–1.44)
			T2DM-ESRD	T / C	Recessive	0.195 , 0.265	0.00099	1.73 (1.25–2.40)

Haplotype blocks in relative map order with nominal significance (p<0.05) under a global score test of association using simulation. Association tests were performed using haplo.score (17), retaining haplotypes with frequency of 1% or greater. Positions of SNPs that start (5') or end (3') a block are for chromosome 20 forward strand physical map coordinates, genome build 36. Global simulated p-values marked with (*) contain at least one SNP with significance p<0.01 in one or more marginal tests of association.

Block #	5' SNP	3' SNP	5' SNP Position	3' SNP Position	Block Size	Block Length (kb)	Global Sim p-value
22	rs2297200	rs2868764	44117933	44155418	10	37.486	0.027*
30	rs1044369	rs2257545	44420725	44437195	3	16.471	0.038*
33	rs2780231	rs366860	44569851	44574807	2	4.957	0.0011
35	rs6066029	rs126917	44650935	44655342	2	4.408	0.012*
48	rs425433	rs1889177	45729687	45744925	2	15.239	0.043
51	rs6066563	rs2664588	46012126	46014041	2	1.916	0.034

Table 4

Haplotype analyses for the 2-SNP haplotype block (rs2780231, rs366860) of the ZNF334 gene. T2DM-ESRD vs. controls (cases; n=300, controls; n=310).

Haplotype	Frequency			Hap-Score	p-value	
	Cases	Controls	Combined		Simulated	Global
CT	0.45	0.47	0.46	−0.88	0.382	0.00105
TG	0.51	0.51	0.51	−0.11	0.922	
CG	0.04	0.01	0.02	3.64	0.0002	