Diabetologia (2011) 54:2047–2055
DOI 10.1007/s00125-011-2188-3

ARTICLE

Genome-wide association and meta-analysis in populations from Starr County, Texas, and Mexico City identify type 2 diabetes susceptibility loci and enrichment for expression quantitative trait loci in top signals

J. E. Below · E. R. Gamazon · J. V. Morrison · A. Konkashbaev · A. Pluzhnikov · P. M. McKeigue · E. J. Parra · S. C. Elbein · D. M. Hallman · D. L. Nicolae · G. I. Bell · M. Cruz · N. J. Cox · C. L. Hanis

Received: 18 January 2011 / Accepted: 14 April 2011 / Published online: 7 June 2011

Abstract
Aims/hypothesis We conducted genome-wide association studies (GWASs) and expression quantitative trait loci (eQTL) analyses to identify and characterise risk loci for type 2 diabetes in Mexican-Americans from Starr County, TX, USA.
Method Using 1.8 million directly interrogated and imputed genotypes in 837 unrelated type 2 diabetes cases and 436 normoglycaemic controls, we conducted Armitage trend tests. To improve power in this population with high disease rates, we also performed ordinal regression including an intermediate class with impaired fasting glucose and/or glucose tolerance. These analyses were followed by meta-analysis with a study of 967 type 2 diabetes cases and 343 normoglycaemic controls from Mexico City, Mexico.
Result The top signals (unadjusted p value $<1\times10^{-5}$) included 49 single nucleotide polymorphisms (SNPs) in eight gene regions (*PER3*, *PARD3B*, *EPHA4*, *TOMM7*, *PTPRD*, *HNT* [also known as *RREB1*], *LOC729993* and

Electronic supplementary material The online version of this article (doi:10.1007/s00125-011-2188-3) contains supplementary material, which is available to authorised users.

J. E. Below · D. L. Nicolae · G. I. Bell · N. J. Cox
Department of Human Genetics, University of Chicago,
Chicago, IL, USA

E. R. Gamazon · J. V. Morrison · A. Konkashbaev · A. Pluzhnikov · D. L. Nicolae · N. J. Cox
Section of Genetic Medicine, KCBD 3220, University of Chicago,
900 E 57th Street,
Chicago, IL 60637, USA

P. M. McKeigue
Public Health Sciences Section, Division of Community Health Sciences, University of Edinburgh Medical School,
Edinburgh, UK

E. J. Parra
Department of Anthropology, University of Toronto,
Mississauga, ON, Canada

S. C. Elbein
Section on Endocrinology and Metabolism,
Wake Forest University School of Medicine,
Winston-Salem, NC, USA

D. M. Hallman · C. L. Hanis (✉)
Human Genetics Center, University of Texas Health Science Center at Houston,
P.O. Box 20186, Houston, TX 77225, USA
e-mail: Craig.L.Hanis@uth.tmc.edu

D. L. Nicolae
Department of Statistics, University of Chicago,
Chicago, IL, USA

D. L. Nicolae · G. I. Bell · N. J. Cox
Department of Medicine, University of Chicago,
Chicago, IL, USA

M. Cruz
Unidad de Investigacion Medica en Bioquimica, Hospital de Especialidades, Centro Medico 'Siglo XXI', IMSS,
Mexico City, Mexico

N. J. Cox (✉)
Section of Genetic Medicine, University of Chicago,
KCBD 3220, 900 E 57th Street,
Chicago, IL 60637, USA
e-mail: ncox@medicine.bsd.uchicago.edu

IL34) and six intergenic regions. Among these was a missense polymorphism (rs10462020; Gly639Val) in the clock gene *PER3*, a system recently implicated in diabetes. We also report a second signal (minimum p value 1.52×10^{-6}) within *PTPRD*, independent of the previously implicated SNP, in a population of Han Chinese. Top meta-analysis signals included known regions *HNF1A* and *KCNQ1*. Annotation of top association signals in both studies revealed a marked excess of *trans*-acting eQTL in both adipose and muscle tissues.

Conclusions/Interpretation In the largest study of type 2 diabetes in Mexican populations to date, we identified modest associations of novel and previously reported SNPs. In addition, in our top signals we report significant excess of SNPs that predict transcript levels in muscle and adipose tissues.

Keywords Expression quantitative trait loci · Genome-wide association study · Meta-analysis · Mexican-American · Type 2 diabetes

Abbreviations

CQC	Contrast quality control
CRLMM	Corrected robust linear model with maximum likelihood classification
DIAGRAM+	Diabetes Genetics Replication and Meta-analysis Consortium (expanded dataset)
eQTL	Expression quantitative trait loci
GWAS	Genome-wide association study
IBD	Identical-by-descent
IFG	Impaired fasting glucose
IGT	Impaired glucose tolerance
LCL	Lymphoblastoid cell lines
LD	Linkage disequilibrium
MAF	Minor allele frequency
NHGRI	National Human Genome Research Institute
PCA	Principal component analysis
QC	Quality control
SNP	Single nucleotide polymorphism
SQC	Skewness quality control

Introduction

Genome-wide association studies (GWASs) of type 2 diabetes in large samples of recent European descent have identified a number of susceptibility loci; it is widely accepted that, to further understand the genetic aetiology of type 2 diabetes, it will be necessary to replicate these findings and identify new loci using other populations. Mexican-Americans have disproportionately high rates of type 2 diabetes. In order to identify genetic variants contributing to type 2 diabetes susceptibility in Mexican-Americans we conducted both binary trait and ordinal GWA on 1618 individuals from Starr County with diabetes (cases), impaired glucose tolerance (IGT) or impaired fasting glucose (IFG), or normal fasting glucose and glucose tolerance (controls), using nearly 2 million directly interrogated and imputed markers. These analyses were combined in meta-analyses with the results from GWASs of admixed Mexicans from Mexico City [1]. This sample is described in detail in the companion paper. These cohorts comprise the largest type 2 diabetes analysis in Mexican and Mexican-American samples to date. In addition, we followed our association and meta-analyses by annotating SNPs with expression quantitative trait loci (eQTL) information from lymphoblastoid cell lines (LCLs) as well as adipose and muscle tissues, to characterise possible functional roles for SNPs associated with type 2 diabetes.

Methods

Study population and phenotype assessment Using the 1979 National Diabetes Data Group criteria [2], cases (n=990) and controls (n=990) were selected from genetic and epidemiological resources developed in the Mexican-American population in Starr County, TX, USA (these also meet definitions set out in the 1997 ADA criteria [3]). Of the 931 successfully genotyped case samples, we selected 837 unrelated individuals for our analyses.

A representative type 2 diabetes control sample was selected from a systematic survey of the primary population centres of Starr County conducted from 2002 to 2006. Blocks were randomly selected within primary population centres in Starr County, and all individuals in each household on selected blocks were enumerated. A random individual aged 20 years and above with no history of diabetes was selected from each household for a physical evaluation including an oral glucose tolerance test with a 75 g glucose load. From 1,345 oral glucose tolerance tests, 350 individuals were classified as having diabetes based on their fasting or 2 h post-load glucose, and were excluded from consideration as controls.

From the remaining 995 individuals, 990 were selected for genotyping, and after genotype quality control and pruning related samples, the control sample included 781 unrelated Starr County residents. Of these, 345 individuals had IGT (defined as a blood glucose level of 7.8–11.0 mmol/l after a 2 h oral glucose tolerance test) or impaired fasting glucose (defined as fasting blood glucose level of 5.6–6.9 mmol/l after overnight fast) and were included in the IGT/IFG category as an intermediate phenotype in the ordinal analysis. Table 1 summarises phenotypes measured in the case, IGT/IFG and control groups.

Table 1 Characteristics of the study population

Characteristic	Cases	IFG/IGT	Controls
n	837	345	436
Sex (*n*, female)	504	259	299
Age (years) (age at diagnosis in cases)	46.7±10.9	40.9±10.5	37.6±9.0
BMI (kg/m^2)	31.8±6.4	31.0±5.8	29.5±6.5
HbA_{1c} (%)	10.9±3.7	NA	NA

Data are mean ± SD, unless otherwise indicated

NA, not available

Genotyping and quality control Genomic DNA was isolated from whole blood of 1,980 unique individuals from Starr County, and quantified by picogreen. The Affymetrix Genome-Wide SNP Array 6.0 assay was performed according to manufacturer protocols (Affymetrix, Santa Clara, CA, USA) at the Center for Inherited Disease Research. Case, IGT/IFG, and control groups were randomised to plates. Genotypes were called for 1,890 samples passing quality control (QC) using both the Affymetrix supported Birdseed v2 [4] calling algorithm as well as the corrected robust linear model with maximum likelihood classification (CRLMM) [5] and only calls in complete agreement between the two algorithms were used in analyses.

No samples uniformly failed contrast QC (CQC) and skewness QC (SQC) [6] in the Affymetrix enzyme-specific subsets as well as the full set of enzyme-specific SNPs. Plate effects were examined [6] and three plates were identified as having an excess of SNPs with low plate-association *p* values (132, 413, and 1,080 SNPs with *p* value $> 1\times10^{-7}$ on these plates). The plate showing the most severe plate effects was comprised of samples that had failed genotyping previously and had been repeated. Since no systematic bias was apparent from the battery of quality measures and no single set of individuals consistently clustered together in a genotype class in the cluster plots of the most plate-biased SNPs (rather poor clustering and miss-calling were frequently observed in these cluster plots), we chose to recall the data with CRLMM without removing any individuals. Only autosomal genotypes with perfect match calls in both the Birdseed and CRLMM datasets ('consensus calls') were carried forward, resulting in a modest reduction in overall call rate (from 0.997 to 0.996 across SNPs and samples). Plate effect analysis was repeated using the consensus calls; the number of SNPs with plate effect *p* values $<1 \times 10^{-7}$ for the three plates noted above are 14, 111 and 77, respectively. That a modest plate effect remains in the data may reduce our power to detect true phenotype associations, but the lack of evidence of differential bias between cases and controls suggests that the bias will not increase the false positive rate. We annotated all SNPs with their quality scores from the QC phase, and checked for flags, such as strong association with a plate, in all top signals.

Using PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/) [7], SNP and individual level call rates were calculated as well as sex incompatibilities (one mismatch identified and removed using markers on the X and Y chromosomes). All individuals passed per sample missingness of <0.1 and 7124 SNPs were removed as a result of missingness >0.1. Identical-by-descent (IBD) estimations were calculated in PLINK and an 'unrelated' cohort of 1,618 samples was created with a proportion of pairwise sharing IBD less than 0.28, weighting inclusion of cases and higher call rate. Individual and plate level heterozygosity was estimated in PLINK and no outliers observed. SNPs with minor allele frequency (MAF) <0.01 were removed. Departures from Hardy–Weinberg equilibrium were calculated and flagged in the full sample, cases, and controls.

Major ancestry group components for the unrelated admixed sample in a SNP set pruned of local and long-distance linkage disequilibrium (LD) were calculated in EIGENSTRAT [8] (Fig. 1a); however, neither of the first two principal components were found to be significantly correlated with type 2 diabetes (mean regression *p* values were 0.18, 0.39, with Pearson correlation coefficients between covariate and phenotype of 0.005 and 0.021, respectively), similar to our previous findings [9]. Although we did not see significant correlation of type 2 diabetes and European admixture in the Starr County sample, strong correlation was observed in the Mexico City study (Fig. 1b). To prevent excessive false positive rates, all analyses in the Mexico City sample controlled for the distribution along the primary axis of variation. For additional details of covariates in this sample please refer to the companion paper [10].

Imputation We performed imputation in MACH 1.0 [11, 12] from consensus call genotypes with MAF >0.01 in 1,618 Starr County Mexican-Americans to the full panel of CEU (Centre d'Etude du Polymorphisme [Utah residents with northern and western European ancestry]), YRI (Yoruba in Ibadan, Nigeria), and JPT (Japanese in Tokyo, Japan) + CHB (Han Chinese in Beijing, China) HapMap2 phased polymorphic SNPs from release 22 [13]. In total 137,532 SNPs with an estimate of the squared correlation between imputed and true genotypes, $r^2 \leq 0.7$ (as estimated by MACH) were removed. An additional 2,616 SNPs with MAF <0.01 were removed, resulting in 1,829,586 genotyped or imputed markers for analyses.

Type 2 diabetes association analyses We used SNPtest [14] to calculate the Armitage trend test using covariates and the

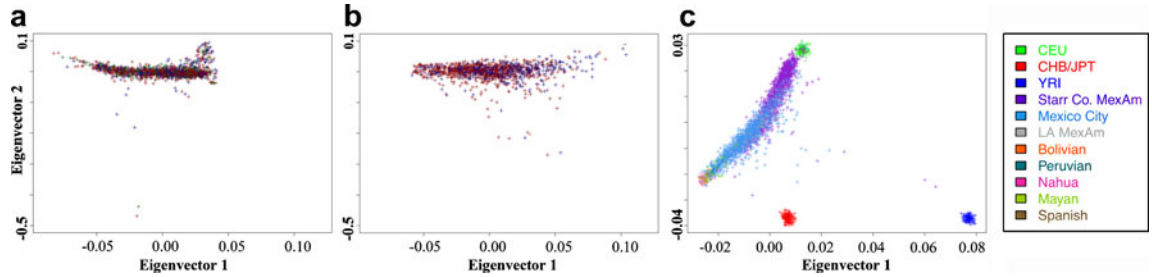

Fig. 1 PCA analysis of (**a**) study population from Starr County, TX, USA, and (**b**) study population from Mexico City. Cases are shown in red, IGT/IFG samples are in green, gluconormal controls are in blue. Both populations analysed together with a number of ancestral populations are shown in (**c**)

option `score`, which specifies a missing-data likelihood test. Additionally, the missing-data likelihood test was extended to a three-category (cases, IGT/IFG, and controls) ordered logistic regression using the same covariates, and implemented in R [15]. We tested the significance of the correlation of covariates in PLINK. In Starr County, 56% of people aged 70 years and older have diabetes, suggesting that substantial numbers of those with non-normal or even normal glucose tolerance will ultimately develop diabetes. Therefore, to generate a more meaningful covariate for age, we used the age at diagnosis (dxage) for cases and the age at enrolment for the controls. In the case–control analysis as well as in the data including the IGT/IFG category, age/dxage, sex and BMI were significantly correlated with phenotype (mean regression p values across directly interrogated SNPs were $<1\times10^{-34}$, $<1\times10^{-4}$ and $<1\times10^{-3}$, with Pearson correlation coefficients between covariate and ordinal phenotype of 0.36, 0.08 and 0.15 respectively).

Meta-analysis Type 2 diabetes association results from SNPtest in Mexicans from Mexico City were combined with results from the associations in the Starr County sample using SNPtest's companion program META [16]. We performed principal component analysis (PCA) in EIGENSTRAT on a local and long-distance LD-pruned SNP set in the Mexico City sample as well as together with parental populations, to check homogeneity between the samples (Fig. 1b, c). Population substructure of the samples within the context of HapMap CEU, YRI, CHB + JPT, and Mexican-American samples from Los Angeles, as well as samples from Bolivians, Peruvians, and Nahua, Mayan, and Spanish populations [17], was largely overlapping, suggesting the samples are derived from similar parental populations; the Starr County sample exhibited more European ancestry. The score method in SNPtest generates a relative information measure on the scale of (0,1) of the observed statistical information from the imputation for the estimate of allele frequency of the SNPs, and was used to assess confidence in the meta-analysis p values. Reported p values have a combined score statistic ≥0.5, MAF>0.05 and p value <0.05 in each study, and a meta-analysis p value <0.001 showing the same direction of effect.

eQTL analysis The top type 2 diabetes association signals from SNPtest and from the three-category ordered logistic regression were evaluated for their effect on transcript levels in LCLs, muscle, and adipose tissue. The set of eQTLs in lymphocytes among the type 2 diabetes associated SNPs were retrieved from SCAN [18], which holds the results of eQTL mapping of HapMap SNPs from 90 CEU and 90 YRI samples, to transcriptional expression evaluated in LCLs from these individuals by the Affymetrix GeneChip Human Exon 1.0 ST Array (Affymetrix). For the analyses of muscle and adipose eQTLs, samples were chosen from among 184 study participants for availability of both tissues, and are described in Sharma et al. [19]. Sixty-two study participants at the tails of the distribution of insulin sensitivity (p value $<1\times10^{-5}$) adjusted for age, sex, BMI and per cent fat were genotyped on the Illumina 1M platform (Illumina, San Diego, CA, USA) using standard Illumina protocols. Transcript levels were assayed using Human Whole Genome 4×44 arrays (Agilent Technologies, Santa Clara, CA, USA). eQTL mapping in each tissue was done by testing each marker for the additive effect of allele dosage on normalised probe-level expression intensity. In our annotations, the best proxy SNPs were determined for those type 2 diabetes associated SNPs not directly typed on the Illumina platform used in the eQTL analyses of muscle and adipose tissues.

Using SNP eQTL annotations in LCLs, and adipose and muscle tissues, we conducted simulations to test for an enrichment of eQTLs (p value $<1\times10^{-4}$) among the top signals (p value $<1\times10^{-3}$) associated with type 2 diabetes. We generated 1,000 random SNP sets, each of the same size (n=722) as the original list of LD-pruned top signals (r^2>0.30), containing variants matched on MAF distribution, sampled without replacement from the set of typed SNPs on the Illumina 1M platform; MAF matching was

done by drawing from the platform SNPs, which had been grouped into discrete MAF bins, each spanning 5% of the allele frequency spectrum. These simulations (n = 1,000) yield an empirical p value, calculated as the proportion of simulations in which the number of eQTLs exceeds the observed number, Q, in the list of top type 2 diabetes signals.

Results

Association and meta-analyses All analyses were performed using the combined age/dxage measurement and sex as covariates and association tests were conducted with and without correcting for BMI. We estimated the inflation factors (the ratio of the trimmed means of the observed and expected values), $\lambda_{age/dxage+sex}$=1.034 and $\lambda_{age/dxage+sex+BMI}$=1.026 in the case–control analysis and 1.038 and 1.023 (respectively) in the three-category ordered logistic regression. The Manhattan and quantile–quantile (QQ) plots summarising results from the ordinal logistic regression without BMI as a covariate are presented in Fig. 2a, b (others are in ESM Figs 1–3). Forty-nine high quality SNPs, with an uncorrected p value $<1 \times 10^{-5}$ in at least one analysis, fell within 14 genomic regions, and are listed, with p values from the analysis, in Table 2 and depicted in Fig. 2a (p values for these markers in each analysis can be found in ESM Table 1). These top signals fall within eight genic regions: *PER3*, *PARD3B*, *EPHA4*, *TOMM7*, *PTPRD*, *HNT* (also known as *RREB1*), *LOC729993* (also known as *SHISA9*), *IL34*, and six intergenic regions. In general, the SNPs within these genic and intergenic regions represent single blocks of LD ($r^2 > 0.7$). Aside from the SNPs in *PARD3B* and *EPHA4*, which have MAF of about 0.05, these signals are driven by common variation (MAF > 0.1) (Table 2). We calculated heterogeneity in META (www.stats.ox.ac.uk/~jsliu/meta.html); we observed no excess of heterogeneity genome-wide and all Starr County top signals have heterogeneity p values $>1 \times 10^{-4}$ (Table 2) [20]. Results for SNPs in LD with previously reported associated type 2 diabetes SNPs from the National Human Genome Research Institute (NHGRI) catalogue [21], Diabetes Genetics Replication and Meta-analysis Consortium (expanded dataset) (DIAGRAM+) [22], as well as top signals identified in the previous GWAS in Starr County [9], were examined in detail (ESM Table 2). We also examined our results for variation within *CAPN10*, previously reported as a type 2 diabetes gene in Mexican-Americans, but key SNPs in the associated *CAPN10* haplotypes were neither directly interrogated nor imputed.

The top meta-analysis signals fell within and near the genes *CSMD1* and *HNF1A*; additional signals implicated *LSAMP*, *FGF12*, *RP11-354K1.1*, *KCNQ1*, *ANK2*, *NIPAL2*, *RP11-74C3.1*, *CIT*, *C22orf30* (also known as *PRR14L*), *EPHB2*, *MCPH1* and *DEPDC5*. Follow-up meta-analysis of top signals from the combined Mexican/Mexican-American studies with the recent DIAGRAM+ dataset identified genome-wide significant signals (p value $<5\times10^{-8}$) in or near susceptibility genes *HNF1A* and *CDKN2A/CDKN2B*, and suggestive evidence for *IGF2BP2*, *KCNQ1* and the previously unreported *C14orf70*. For additional details of the meta-analyses of these samples and follow-up of top signals in DIAGRAM+, please refer to the companion paper.

In addition to physical proximity, we considered the roles that SNPs in the top signals play in transcript prediction. Annotation of LD-pruned top signals (p value $<1\times10^{-3}$, $r^2<0.3$) from the ordinal regression analysis in Starr County, with sex and age/dxage as covariates, identified 258 SNPs predicting transcript levels in LCLs, 289 in muscle, and 359 in adipose tissue. The mean number of frequency matched SNPs that predict transcript levels in LCLs in 1,000 simulations was 252 (SD 12.17), suggesting no enrichment. However, the maximum number of muscle and adipose eQTLs observed in 1,000 simulations was 173 and 193, respectively. The distributions of eQTL counts established by simulations, as well as the actual observed counts in muscle and adipose tissues, are depicted in Fig. 3a, b (see ESM Fig. 4 for LCLs). Notably, nearly all

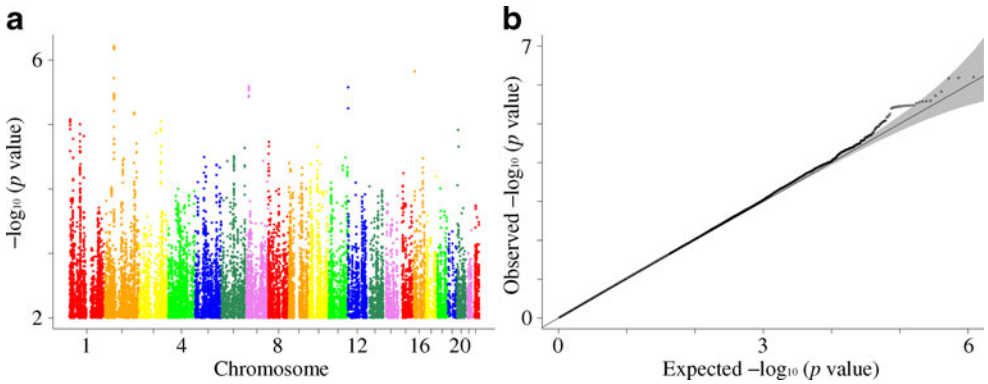

Fig. 2 Manhattan plot of $-\log_{10}$ (p values) at 1.8 million autosomal SNPs (**a**) and QQ plot of observed vs expected p values (**b**) of results of ordinal regression analysis on type 2 diabetes adjusted for age/dxage and sex

Table 2 Summary of top regions from GWAS of type 2 diabetes in Starr County Mexican-Americans

SNP	Chr	Pos	Risk allele	Non-risk allele	Ordinal analysis, age- and sex-adjusted							Meta-analysis *p* value	Gene	Location
					OR (95% CI)	*p* value	Freq (all)	Freq (cases)	Freq (IFG/IGT)	Freq (controls)	Imputation r^2			
rs10462020	1	7803270	T	G	1.43 (1.27–1.59)	8.39×10^{-6}	0.83	0.85	0.8	0.8	1	8.57×10^{-3}	*PER3*	Missense
rs6728803	2	66001970	G	A	1.38 (1.26–1.51)	6.80×10^{-7}	0.52	0.56	0.5	0.48	0.83	9.16×10^{-3}	*KRT18P33-LOC729348*	Intergenic
rs849230	2	205969274	A	G	1.83 (1.57–2.1)	6.86×10^{-6}	0.93	0.95	0.92	0.91	0.88	–	*PARD3B*	Intron
rs1879976	3	149826207	A	G	1.34 (1.21–1.47)	8.91×10^{-6}	0.59	0.61	0.59	0.54	0.9	–	*LOC344741-RPL38P1*	Intergenic
rs2240727	7	22819037	T	C	1.37 (1.24–1.5)	2.79×10^{-6}	0.34	0.38	0.29	0.31	0.9	7.72×10^{-3}	*TOMM7*	3′ UTR
rs4766119	12	3419017	A	G	1.65 (1.44–1.85)	2.65×10^{-6}	0.87	0.89	0.85	0.84	0.72	1.25×10^{-3}	*LOC100128253-LOC100129223*	Intergenic
rs149228	16	13211975	C	A	1.43 (1.28–1.57)	1.49×10^{-6}	0.24	0.27	0.21	0.2	0.89	1.61×10^{-3}	*LOC729993*	Intron
rs4522658	2	134203633	G	A	1.28 (1.17–1.4)	3.40×10^{-5}	0.46	0.49	0.46	0.42	0.95	1.27×10^{-4}	*NAP5-MGAT5*	Intergenic
rs16862811	2	222109330	G	T	1.8 (1.53–2.08)	2.65×10^{-5}	0.94	0.95	0.95	0.91	0.94	4.12×10^{-4}	*EPHA4*	Intron
rs10192201	2	237468147	G	C	1.26 (1.13–1.4)	4.44×10^{-4}	0.64	0.65	0.67	0.58	0.93	1.54×10^{-3}	*LOC100128709-LOC93463*	Intergenic
rs649891	9	10420602	C	T	1.29 (1.16–1.43)	1.69×10^{-4}	0.41	0.43	0.44	0.35	0.78	5.80×10^{-6}	*PTPRD*	Intron
rs3099797	11	131626596	C	T	1.3 (1.18–1.43)	4.60×10^{-5}	0.64	0.67	0.63	0.58	1	3.38×10^{-5}	*HNT*	Intron
rs8033124	15	89712478	C	A	1.3 (1.16–1.44)	1.83×10^{-4}	0.42	0.44	0.44	0.37	0.74	–	*LOC100128403-LOC441732*	Intergenic
rs3813905	16	69238351	C	G	1.41 (1.25–1.57)	3.35×10^{-5}	0.75	0.77	0.73	0.73	0.73	9.73×10^{-4}	*IL34*	5′ UTR

Representative SNPs from each genomic region with *p* values below 1×10^{-5} in at least one analysis (regression on the ordinal or binary phenotype, with/without covariate BMI). Reported statistics are derived from the ordinal regression on type 2 diabetes adjusted for age/dxage and sex. Reported meta-analysis *p* values are derived from the binary case–control analysis combined with the analogous analysis in the Mexico City dataset. Any evidence for association is driven solely by the Starr County study, as all *p* values in the Mexico City sample for these SNPs were >0.15 (SNPs in the table without high quality Mexico City genotypes are labelled with a dash); *p* values for heterogeneity are greater than 1×10^{-4} for all meta-analysed SNPs.

Chr, chromosome; Freq, frequency; Pos, position; UTR, untranslated region

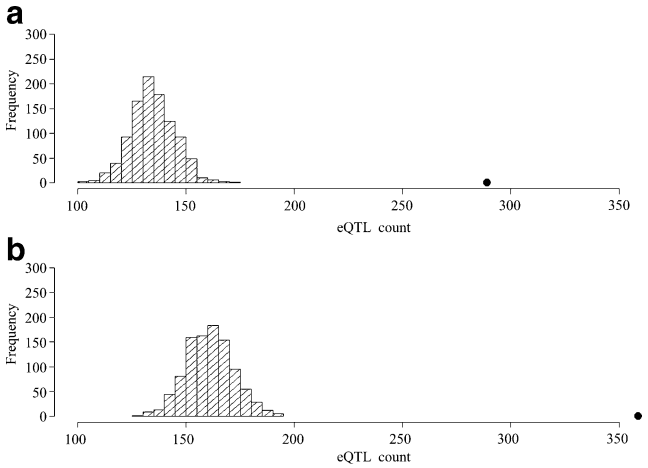

Fig. 3 Simulated distribution of eQTL counts for SNPs with allele frequencies matched to signals with p value $<1 \times 10^{-3}$ from the ordinal regression on type 2 diabetes adjusted for age/dxage and sex in the Starr County study population. The black dot represents the actual eQTL count observed in SNPs with p value $<1 \times 10^{-3}$ in this analysis. Results are shown for transcription prediction in adipose (**a**) and muscle (**b**) tissue

of the enrichment in muscle and adipose tissue is driven by *trans*-acting eQTLs; only three of the eQTLs observed in muscle, and eight in adipose tissue, were *cis*-acting. In addition, we replicated the observed enrichment in muscle and adipose tissues in the top signals (defined as described above) from the analysis in the Mexico City sample (Fig. 4a, b). A number of SNPs showing association for type 2 diabetes in our studies predicted gene expression across multiple tissues; 14 SNPs predict expression in all three tissues tested (ESM Fig. 5).

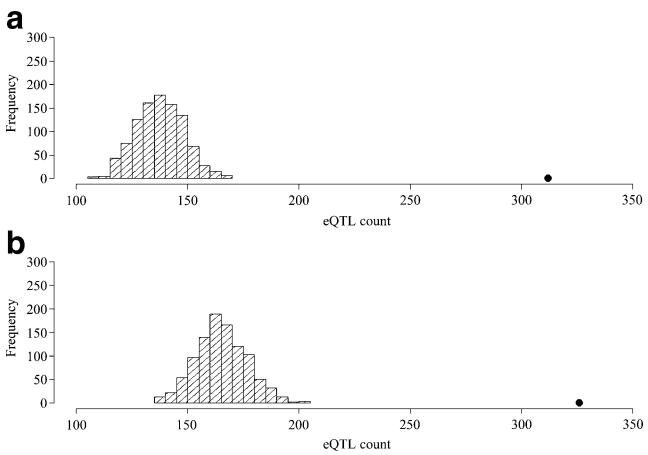

Fig. 4 Simulated distribution of eQTL counts for SNPs with allele frequencies matched to signals with p value $<1 \times 10^{-3}$ from the ordinal regression on type 2 diabetes adjusted for age/dxage and sex in the Mexico City study population. The black dot represents the actual eQTL count observed in SNPs with p value $<1 \times 10^{-3}$ in this analysis. Results are shown for transcription prediction in adipose (**a**) and muscle (**b**) tissue

Discussion

As studies of genetic association in type 2 diabetes are pursued in non-European populations, it is important to recognise that, in populations with markedly increased disease rates, there may be room to improve on the case–control study design. In this case, we favour an approach that uses more information and recognises the staging of the disease, enabling us to recover a substantial number of individuals. In general, the results of the ordinal and binary analyses are very similar, but including the intermediate class allows us to include more of the individuals in the sample, and should, therefore, provide a more complete picture of the association of genetic variation with the spectrum of phenotypes related to type 2 diabetes. Among the top signals in the ordinal analysis is a functional missense polymorphism in the clock gene *PER3*. Substantial research has implicated circadian clock function in metabolic diseases, coronary function and obesity [23–25], but this signal is not replicated in the Mexico City sample (p value ~0.78).

Another top signal from the Starr County sample, as well as in the meta-analysis with Mexico City, implicates a previously reported type 2 diabetes risk gene *PTPRD*. *PTPRD*, protein tyrosine phosphatase receptor type delta gene [26], was recently identified by GWAS in Han Chinese populations [26]. Although the SNPs are both intronic, our result is 1.6 Mb downstream and represents an independent signal from within the gene (the SNPs are in linkage equilibrium in HapMap CHB + JPT as well as CEU samples). The role of *PTPRD* in type 2 diabetes is not yet understood; however, when ablated in mice, genes in this subfamily induce defects in glucose homeostasis and insulin sensitivity [27, 28].

There is little literature suggesting roles for *IL34*, *TOMM7*, *HNT* or *PARD3B*, nor the intergenic regions, in diabetes; however, in HapMap Europeans, LD extends from several of these top regions to include candidate genes for type 2 diabetes. For example, two intergenic SNPs on chromosome 12 are in high LD with *PRMT8* (r^2=0.764), a gene known to influence lipid levels [29]. Yet, LD with a type 2 diabetes-related gene alone is not sufficient evidence of a SNP's functional role. To assess functional consequences of associated variants, we explored our findings in the context of gene regulation.

Previous studies have identified enrichment of eQTLs in trait associated SNPs [30]. A recent study into the functional role of type 2 diabetes associated SNPs found an excess of eQTLs among signals from the Wellcome Trust Case Control Consortium and the Diabetes Genetics Initiative in muscle and adipose [31, 32]. The enrichment in these cases was driven almost exclusively by *trans*-acting eQTLs. This enrichment was robust to thresholds of eQTL

p value (1×10^{-6} was also considered and a similar level of enrichment is still observed) and threshold of significance of type 2 diabetes association (top 1,000 and top 10,000 signals were analysed and enrichment was seen in both). We find similar enrichment in both the Starr County and Mexico City datasets. The lack of enrichment of eQTLs in LCLs, despite significant enrichment of eQTLs from muscle and adipose tissue, is likely to be attributable to the fact that LCLs as a cell type are not as relevant as muscle and adipose tissue to type 2 diabetes risk, and only a proportion of eQTLs are shared across tissues [31, 32]. The number of SNPs predicting transcript level in both muscle and adipose tissue is markedly higher (178) than those overlapping LCLs (33), suggesting concordance of regulatory action in muscle and adipose tissues in a large proportion of top type 2 diabetes associated SNPs in our study. Previous studies of eQTLs in LCLs in top type 1 diabetes associated loci have found significant enrichment [30], suggesting that lymphoblasts are a relevant tissue for type 1 diabetes; this is reassuring given the role of autoimmunity in the disease.

Of the 53 previously reported SNPs considered (drawn from the NHGRI catalogue, DIAGRAM+, and Hayes et al. [9], and interrogated or represented by proxy SNP in the Starr County analyses) we identified seven markers with p value <0.05 and odds ratios in the same direction. These markers are in or near *IGF2BP2*, *WFS1*, *JAZF1*, *OR13D1-OR13D3P*, *KCNQ1*, *VPS33B*, and *EPB41L3-LOC645355*. An additional 20 SNPs do not reach marginal significance in our dataset, but do show odds ratios in the same direction. Seventeen SNPs have previously reported confidence intervals for the odds ratios which overlap those observed in the Starr County dataset (all data shown in ESM Table 2).

As in all GWASs of type 2 diabetes published to date, we lack power to detect the kinds of small effect we now know are likely to contribute to disease risk (see ESM Fig. 6). However, the absence of support for a number of established type 2 diabetes genes is notable in this dataset. Of particular interest is the lack of association of SNPs within *TCF7L2*, the gene with the strongest associations with type 2 diabetes observed in Europeans. While the allele frequency of the variant most strongly associated in Europeans is lower in Asian populations, estimates of the odds ratios were quite similar. In contrast, the frequency of the risk allele is similar in Mexican-Americans and Europeans, but the odds ratios are lower in both the Mexican-American and Mexico City datasets; in a fixed effects meta-analysis of SNP rs7903146 in our data and that of Voight et al. [22], the heterogeneity p value is 0.05. Our confidence intervals for rs7903146 also overlap, but just barely, according to DIAGRAM estimates.

The winner's curse, the tendency of a biased overestimate of effect in the first study to report an association and be successfully published, may explain the somewhat higher overestimates of effect in the top signals in the Starr County and Mexico City studies [33]. In addition, ascertainment can contribute to these effects. For example, in Starr County, patients from families with two or more affected members were specifically targeted for the GWAS, which improves power, but leads to estimates of effect size that are larger than would be estimated in an unselected case series. In general, however, lack of replication is likely to be due to the fact that all studies of type 2 diabetes to date are under-powered to detect small effects, and therefore only detect association signals from some (probably different) subset of the contributing variants.

To identify, confirm, and characterise additional genetic risk factors for type 2 diabetes it is likely that we will need to continue studies in non-European populations, to explore new study designs and methods of analysis in high-risk groups, and to enhance GWAS with eQTL and other functional studies.

Acknowledgements Supported in part by DK073451, HL084715, and DK20595, the FOFOI IMSS-2004/014, 2005/2/I/363, 2006/1B/1/002, Fundacion IMSS A. C., and Fundacion Gonzalo Rio Arronte I. A. P. Mexico. Also supported in part by National Institutes of Health grant T32GM007197 and grant CTSA T32 UL1 RR024999. M. Cruz is a recipient of a Fundación IMSS Scholarship, Mexico. We also gratefully acknowledge the willing participation of the residents of Starr County and the efforts of the field staff in Starr County responsible for the collection of these data. Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C.

Duality of interest The authors declare that there is no duality of interest associated with this manuscript.

References

1. Martinez-Marignac VL, Valladares A, Cameron E et al (2007) Admixture in Mexico City: implications for admixture mapping of type 2 diabetes genetic risk factors. Hum Genet 120:807–819
2. National Diabetes Data Group (1979) Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. Diabetes 28:1039–1057
3. The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus (1997) Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Diabetes Care 20:1183–1197
4. Korn JM, Kuruvilla FG, McCarroll SA et al (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40:1253–1260
5. Carvalho B, Bengtsson H, Speed TP, Irizarry RA (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics 8:485–499
6. Pluzhnikov A, Below JE, Konkashbaev A et al (2010) Spoiling the whole bunch: quality control aimed at preserving the integrity of high-throughput genotyping. Am J Hum Genet 87:123–128
7. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

8. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909
9. Hayes MG, Pluzhnikov A, Miyake K et al (2007) Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. Diabetes 56:3033–3044
10. Parra EJ, Below JE, Krithika S et al. (2011) Genome-wide association study of type 2 diabetes in a sample from Mexico City and meta-analysis with a Mexican American sample from Starr County, TX. Diabetologia. doi:10.1007/s00125-011-2172-y
11. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406
12. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2006) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34:816–834
13. Huang L, Li Y, Singleton AB et al (2009) Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet 84:235–250
14. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
15. Valenta Z, Pitha J, Poledne R (2006) Proportional odds logistic regression—effective means of dealing with limited uncertainty in dichotomizing clinical outcomes. Stat Med 25:4227–4234
16. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913
17. Mao X, Bigham AW, Mei R et al (2007) A genomewide admixture mapping panel for Hispanic/Latino populations. Am J Hum Genet 80:1171–1178
18. Gamazon ER, Zhang W, Konkashbaev A et al (2010) SCAN: SNP and copy number annotation. Bioinformatics 26:259–262
19. Sharma NK, Das SK, Mondal AK et al (2008) Endoplasmic reticulum stress markers are associated with obesity in nondiabetic subjects. J Clin Endocrinol Metab 93:4532–4541
20. Zeggini E, Scott LJ, Saxena R et al (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40:638–645
21. Hindorff LA, Junkins HA, Hall PN, Mehta JP, Manolio TA (2011) A catalog of published genome-wide association studies. www.genome.gov/gwastudies. Accessed 6 April 2011
22. Voight BF, Scott LJ, Steinthorsdottir V et al (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42:579–589
23. Young ME, Wilson CR, Razeghi P, Guthrie PH, Taegtmeyer H (2002) Alterations of the circadian clock in the heart by streptozotocin-induced diabetes. J Mol Cell Cardiol 34:223–231
24. Maury E, Ramsey KM, Bass J (2010) Circadian rhythms and metabolic syndrome: from experimental genetics to human disease. Circ Res 106:447–462
25. Marcheva B, Ramsey KM, Buhr ED et al (2010) Disruption of the clock components CLOCK and BMAL1 leads to hypoinsulinaemia and diabetes. Nature 466:627–631
26. Tsai FJ, Yang CF, Chen CC et al (2010) A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. PLoS Genet 6:e1000847
27. Ren JM, Li PM, Zhang WR et al (1998) Transgenic mice deficient in the LAR protein-tyrosine phosphatase exhibit profound defects in glucose homeostasis. Diabetes 47:493–497
28. Chagnon MJ, Elchebly M, Uetani N et al (2006) Altered glucose homeostasis in mice lacking the receptor protein tyrosine phosphatase sigma. Can J Physiol Pharmacol 84:755–763
29. Aulchenko YS, Ripatti S, Lindqvist I et al (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41:47–55
30. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6:e1000888
31. Elbein SC, Gamazon E, Kern PA, Rasouli N, Tikhomirov A, Cox NJ (2010) Mapping genes near type 2 diabetes (T2D) associated loci as expression quantitative trait loci (eQTL) in human adipose and muscle. ADA Abstract no. 205-OR. http://professional.diabetes.org/Abstracts_Display.aspx?TYP=1&CID=79149
32. Elbein SC, Gamazon ER, Das SK, Rasouli N, Kern PA, Cox NJ (2010) High proportion of transcripts associated with insulin sensitivity in fat and muscle are associated with expression quantitative trait loci (eQTL). American Society of Human Genetics Abstract (program no. 195). www.ashg.org/cgi-bin/2010/ashg10s?author=Elbein&sort=ptimes&sbutton=Detail&absno=20639&sid=559883
33. Goring HH, Terwilliger JD, Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 69:1357–1369