Diabetologia (2011) 54:2038–2046
DOI 10.1007/s00125-011-2172-y

ARTICLE

Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American sample from Starr County, Texas

E. J. Parra · J. E. Below · S. Krithika · A. Valladares · J. L. Barta · N. J. Cox · C. L. Hanis · N. Wacher · J. Garcia-Mena · P. Hu · M. D. Shriver · Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium · J. Kumate · P. M. McKeigue · J. Escobedo · M. Cruz

Received: 17 January 2011 / Accepted: 6 April 2011 / Published online: 15 May 2011

Abstract
Aims/hypothesis We report a genome-wide association study of type 2 diabetes in an admixed sample from Mexico City and describe the results of a meta-analysis of this study and another genome-wide scan in a Mexican-American sample from Starr County, TX, USA. The top signals observed in this meta-analysis were followed up in the Diabetes Genetics Replication and

Electronic supplementary material The online version of this article (doi:10.1007/s00125-011-2172-y) contains supplementary material, including a full list of members of the DIAGRAM Consortium, which is available to authorised users.

E. J. Parra (✉) · S. Krithika · J. L. Barta
Department of Anthropology,
University of Toronto at Mississauga,
3359 Mississauga Road North,
Mississauga, ON, Canada L5L 1C6
e-mail: esteban.parra@utoronto.ca

J. E. Below · N. J. Cox
Department of Medicine, Section of Genetic Medicine,
University of Chicago,
Chicago, IL, USA

A. Valladares · M. Cruz (✉)
Unidad de Investigacion Medica en Bioquimica, Hospital de Especialidades, Centro Medico Nacional Siglo XXI, IMSS,
Av. Cuauhtemoc 330, Col. Doctores,
C.P. 06720, Mexico City, Mexico
e-mail: mcruzl@yahoo.com

C. L. Hanis
Human Genetics Center, School of Public Health,
The University of Texas, Health Science Center at Houston,
Houston, TX, USA

N. Wacher
Unidad de Investigacion Medica en Epidemiologia Clinica,
Hospital de Especialidades, Centro Medico Siglo XXI, IMSS,
Mexico City, Mexico

J. Garcia-Mena
Departamento de Genetica y Biologia Molecular, Cinvestav-IPN,
Mexico City, Mexico

P. Hu
The Centre for Applied Genomics,
The Hospital for Sick Children,
Toronto, ON, Canada

M. D. Shriver
Department of Anthropology, Penn State University,
University Park, PA, USA

J. Kumate
Fundacion IMSS,
Mexico City, Mexico

P. M. McKeigue
Public Health Sciences, University of Edinburgh Medical School,
Edinburgh, Scotland, UK

J. Escobedo
Unidad de Investigacion en Epidemiologia Clinica,
Hospital General Regional 1,
Dr Carlos McGregor, IMSS,
Mexico City, Mexico

Meta-analysis Consortium (DIAGRAM) and DIAGRAM+ datasets.

Methods We analysed 967 cases and 343 normoglycaemic controls. The samples were genotyped with the Affymetrix Genome-wide Human SNP array 5.0. Associations of genotyped and imputed markers with type 2 diabetes were tested using a missing data likelihood score test. A fixed-effects meta-analysis including 1,804 cases and 780 normoglycaemic controls was carried out by weighting the effect estimates by their inverse variances.

Results In the meta-analysis of the two Hispanic studies, markers showing suggestive associations ($p<10^{-5}$) were identified in two known diabetes genes, *HNF1A* and *KCNQ1*, as well as in several additional regions. Meta-analysis of the two Hispanic studies and the recent DIAGRAM+ dataset identified genome-wide significant signals ($p<5\times10^{-8}$) within or near the genes *HNF1A* and *CDKN2A/CDKN2B*, as well as suggestive associations in three additional regions, *IGF2BP2, KCNQ1* and the previously unreported *C14orf70.*

Conclusions/interpretation We observed numerous regions with suggestive associations with type 2 diabetes. Some of these signals correspond to regions described in previous studies. However, many of these regions could not be replicated in the DIAGRAM datasets. It is critical to carry out additional studies in Hispanic and American Indian populations, which have a high prevalence of type 2 diabetes.

Keywords Genome-wide association · Hispanics · Meta-analysis · Type 2 diabetes

Abbreviations

DIAGRAM	Diabetes Genetics Replication and Meta-analysis Consortium
GWA	Genome-wide association
LD	Linkage disequilibrium
QQ	Quantile-Quantile
SNP	Single nucleotide polymorphism

Introduction

Recent genome-wide association (GWA) studies have dramatically increased the number of known type 2 diabetes susceptibility loci. To date, approximately 40 loci have been confirmed using this approach [1, 2]. However, most have been identified in populations of European ancestry, and the number of GWA studies in non-European populations has been very limited [3–8]. It is critical to expand this type of study to other population groups because GWA scans in European populations may fail to identify some type 2 diabetes susceptibility variants in some situations, e.g. when: (1) the low-risk or high-risk alleles are nearly fixed in Europeans; (2) linkage disequilibrium (LD) with nearby tag single nucleotide polymorphisms (SNPs) is too weak in Europeans for the effects of risk alleles to be detected; or (3) there is too much allelic heterogeneity for associations with any single tag SNP to be detected. The identification of susceptibility variants within the gene *KCNQ1* in two Japanese studies is an excellent example. These variants were not identified in any of the previous, larger GWA studies in European populations, because the high frequencies observed for the risk alleles in Europe resulted in a dramatic reduction in statistical power [9]. Therefore, GWA studies in non-European populations are necessary in order to increase our understanding of susceptibility loci and aetiological pathways involved in type 2 diabetes.

We report here the results of the first GWA study of type 2 diabetes in an admixed sample from Mexico. We also describe the results of a meta-analysis of our sample and another genome-wide scan in a Mexican-American sample from Starr County, Texas (see also Below et al. [10] in this edition of Diabetologia). Finally, we followed up some of the signals observed in the meta-analysis of the two Hispanic studies using the Diabetes Genetics Replication and Meta-analysis Consortium (DIAGRAM) dataset.

Methods

Study participants A total of 1,334 individuals participated in this study, 989 cases and 345 normoglycaemic controls. The cases were recruited in Mexico City, with selection based on type 2 diabetes age at diagnosis, preference being given to participants aged <55 years at diagnosis. Diabetes diagnosis was done according to the American Diabetes Association criteria, i.e. fasting plasma glucose ≥7.0 mmol/l or OGTT ≥11.1 mmol/l for the 2 h sample. The controls were 345 healthy blood donors (175 men, 170 women) from Mexico City without a family history of type 2 diabetes and with glucose concentrations of <7.8 mmol/l after an OGTT. After quality control, 24 participants (22 cases, two controls) were eliminated from the sample for different reasons (see below), so the final sample used in the statistical analysis comprised 967 cases (664 women, 303 men) and 343 controls (169 women, 174 men). The average age of diagnosis of type 2 diabetes in the final sample of cases was 43.7 years (range 20–60). The average age of the final sample of controls was 50.7 years (range 42–68). Data on BMI were available for cases and controls. Informed consent was obtained from each participant, and the research was approved by the ethical research boards of the Medical Center 'Siglo XXI' and the University of Toronto.

Genotyping Genotyping was carried out in the microarray analysis facility located in the Centre for Applied Genomics (Toronto, ON, Canada), using the Affymetrix Genome-wide Human SNP array 5.0 (Affymetrix, Santa Clara, CA, USA) and following standard protocols. Genotype calling was done with the Affymetrix Power Tools software package, including the full sample set and using two genotyping algorithms, the BRLMM-P and Birdseed algorithms (www.affymetrix.com, accessed 1 December 2010). To minimise genotyping errors, the PLINK version 1.06 (http://pngu.mgh.harvard.edu/~purcell/plink/, accessed 1 December 2010) [11] was used to merge the genotype results obtained with both algorithms, using the consensus call mode. The concordance for autosomal SNPs between the genotype calls obtained with the BRLMM-P and the Birdseed algorithms was 99.71%. The missing genotype rate in the sample of cases was 0.59% and the missing genotype rate in the sample of controls was 0.79%. We genotyped six individuals (four cases, two controls) in duplicate and the concordance rate was 99.88%.

Quality control On the basis of a preliminary analysis using 430,321 autosomal markers, one sample was excluded from the dataset due to high missing genotype rates (>5%) and 23 samples due to cryptic relatedness (n=20) or sample duplication (n=3). For pairs of individuals with identity by descent estimates >0.2, we removed the sample with the highest proportion of missing genotypes. From the initial list of autosomal markers, we additionally removed SNPs based on the following criteria: (1) minor allele frequency <5%; (2) Hardy–Weinberg p values <0.0001 in the control group; and (3) missingness >5% in the cases and the controls. After filtering, the final number of autosomal markers was 315,658. All the quality control procedures were completed using PLINK.

Inferring individual admixture proportions Individual admixture proportions were estimated with the program ADMIXMAP (http://homepages.ed.ac.uk/pmckeigu/admixmap/index.html, accessed 1 December 2010) using a genome-wide panel of 3,500 ancestry informative markers. This panel of ancestry informative markers was selected using genome-wide data available for European, West African and American Indian populations [12].

Population stratification To evaluate population stratification in the sample from Mexico City (e.g. potential differences in ancestry proportions between cases and controls), we used EIGENSOFT (http://genepath.med.harvard.edu/~reich, accessed 1 December 2010) to perform a principal component analysis of the Mexico City sample, after pruning markers showing short-range and long-range LD at a threshold of r^2=0.2. The first axis explains the majority of the variation in the sample (the eigenvalue for the first axis was 17.6 vs less than 2 for the remaining axes). The scores of the first axis show a very strong correlation with the American Indian admixture proportions estimated with ADMIXMAP (R^2=0.991) and were significantly associated with type 2 diabetes. For this reason, the scores of the first axis were incorporated as covariates in the association tests.

Imputation We imputed untyped genotypes with Impute version 2 (https://mathgen.stats.ox.ac.uk/impute/impute.html, accessed 1 December 2010), using the HapMap Phase 2 combined sample as a phased reference sample and the HapMap Phase 3 Mexican-American sample as an unphased reference sample. To evaluate the performance of the imputation, we randomly masked 5% of the markers (n=1,046) directly genotyped on chromosome 12 and compared the imputed genotypes with the Affymetrix genome-wide Human SNP array 5.0 genotype calls. For this analysis, imputed genotypes with probabilities <0.9 were set as missing data. Of the imputed genotypes, 85.7% had probabilities higher than 0.9 and the concordance rate observed between the Affymetrix 5.0 genotypes and the imputed genotypes was 99.4%. In an independent analysis, we masked 100 markers on chromosomes 4, 9, 16 and 20 from the inference files used for imputation and compared the imputed genotypes with the original genotypes obtained with the Affymetrix 5.0 microchip. This analysis resulted in 74.4% of imputed genotypes with probabilities higher than 0.9 and a concordance rate of 99.0%.

Association tests Associations of the genetic markers with type 2 diabetes were tested with SNPTEST version 2 (http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html, accessed 1 December 2010) using the frequentist association tests implemented in the program, based on an additive model. To control for genotype uncertainty, we used the missing data likelihood score test (the score method). The statistical model incorporated sex, age and ancestry, and alternatively, sex, age, ancestry and BMI. In this case, ancestry corresponds to the coordinates of the first principal component of the Mexico City sample (see above, Population stratification).

Meta-analysis We used META (http://www.stats.ox.ac.uk/~jsliu/meta.html, accessed 1 December, 2010) to carry out an inverse variance fixed-effects meta-analysis of the Mexico City and the Starr County studies (for more details about the Starr County GWAs, see Below et al. [10] in this edition of Diabetologia). GWAMA (http://www.well.ox.ac.uk/gwama, accessed 1 December, 2010) was used to carry out an inverse variance fixed-effects meta-analysis of the top signals observed in the Hispanic studies and the 2008 DIAGRAM and DIAGRAM+ datasets.

Statistical power analysis A figure depicting the statistical power of the meta-analysis is available in the companion paper by Below et al. [10].

Results

Exploring admixture proportions and population stratification Analysis with ADMIXMAP indicated that the sample from Mexico City is an admixed sample with substantial American Indian and European contributions, and very limited West African influence. The average American Indian ancestry in the sample was estimated to be 63%, the average European ancestry 34% and the average West African ancestry 3%. Figure 1 shows the individual admixture proportions in the Mexico City sample, as well as the average admixture proportions for the case and control groups. There was considerable variation in admixture proportions in the American Indian–European axis. Additionally, a substantial difference in American Indian admixture proportions was observed in the case and control groups (66% vs 57%, respectively). The effect of population stratification (difference in admixture proportions between cases and controls) is clearly evident in the Quantile–Quantile (QQ) plot comparing p values of the score test conditioning on sex and age with the values expected under the null model of no association (electronic supplementary material [ESM] Fig. 1). Under this model, there was a strong departure of the observed and expected p values. In contrast, the score test conditioning on sex, age and ancestry produced a QQ plot with a good fit between observed and expected p values (ESM Fig. 2). Therefore, inclusion of ancestry in the model corrected for the effects of stratification in this sample.

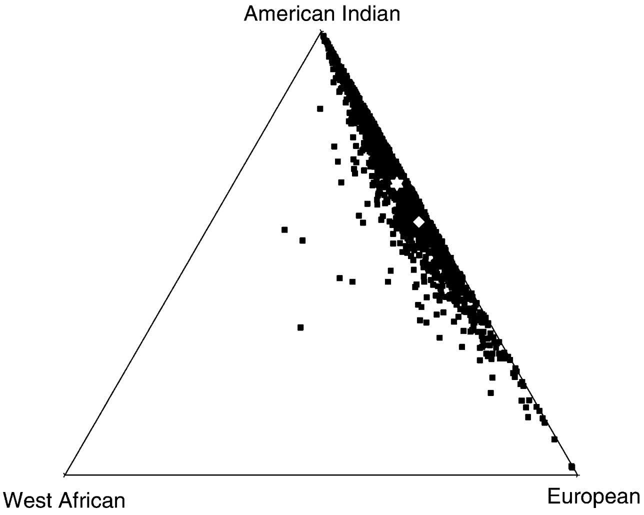

Fig. 1 Triangular representation of admixture proportions in the sample from Mexico City, estimated with ADMIXMAP. Individual admixture proportions are indicated by squares. Star, average admixture proportions in the case group; diamond, average admixture proportions in the control group

Statistical analysis, Affymetrix 5.0 and imputed markers The statistical analysis was carried out for approximately 2 million common markers (~300,000 directly genotyped and ~1.7 million imputed markers), using the missing likelihood score test implemented in SNPTEST version 2. The Manhattan plots summarising the results of the analyses and controlling for sex, age and ancestry, and sex, age, ancestry and BMI are presented as ESM Figs 3 and 4. ESM Tables 1 and 2 show the top signals ($p<10^{-5}$) observed in the analysis with sex, age and ancestry, and sex, age, ancestry and BMI as covariates, respectively, annotated using WGA Viewer [13].

Meta-analysis of two type 2 diabetes GWA studies in Hispanics We performed a fixed-effects meta-analysis of two GWA studies carried out in individuals of Mexican ancestry, i.e. our Mexico City study (967 type 2 diabetes cases, 343 normoglycaemic controls) and a study in Starr County, Texas, USA (837 cases, 437 normoglycaemic controls). Table 1 lists the markers showing suggestive associations ($p<10^{-5}$) in the meta-analysis without and with BMI as a covariate. Only markers that were nominally significant ($p<0.05$) in both studies are included, with only markers with the lowest p values in each region being reported. The meta-analysis was carried out with META, setting the information content exclusion threshold for the imputed markers (score information) at 0.5. The lowest p values were observed for two variants within and near the gene *HNF1A* on chromosome 12 ($p=3.5\times10^{-7}$ lowest p value). Eight additional regions with SNPs showing values of $p<10^{-5}$ were identified in the meta-analysis, corresponding to the genes *RP11-354K1.1* on chromosome 1, *LSAMP* on chromosome 3, *FGF12* on chromosome 3, *CSMD1* on chromosome 8, *NIPAL2* on chromosome 8, *CIT* on chromosome 12 and *DEPDC5* on chromosome 22, as well as an expressed sequence tag located on chromosome 9. In the analysis incorporating BMI as a covariate, the markers located within or near the genes *LSAMP* and *NIPAL2* were not among the top signals, and there were several new regions with values of $p<10^{-5}$ in this analysis, including the genes *EPHB2* on chromosome 1, *ANK2* on chromosome 4, *MCPH1* on chromosome 8 and *KCNQ1* on chromosome 11.

Follow-up of the top signals of the Hispanic meta-analysis in the DIAGRAM dataset The signals observed in the samples of Mexican ancestry were further explored by meta-analysis using the 2008 DIAGRAM consortium dataset [14], which included three GWA studies in individuals of European ancestry (Wellcome Trust Case–Control

Table 1 Top signals observed in the meta-analysis of the Mexico City and Starr County studies

SNP	*p* value	Chr	Pos	Type	Closest gene	NEA	EA	OR	OR_95L	OR_95U	*p*-SC	*p*-MC
Without BMI												
rs7516189[c]	7.5×10^{-6}	1	211828134	Upstream	*RP11-354K1.1*[d,e]	G	A	1.46	1.24	1.72	1.9×10^{-3}	7.6×10^{-4}
rs938911[c]	8.6×10^{-6}	3	116420202	Intronic	*LSAMP*	G	A	2.05	1.50	2.82	7.9×10^{-5}	3.1×10^{-2}
rs9831045[b]	2.4×10^{-6}	3	192475259	Intronic	*FGF12*[e]	T	A	1.41	1.22	1.62	2.3×10^{-3}	1.9×10^{-4}
rs9773092[a]	9.4×10^{-7}	8	3216025	Intronic	*CSMD1*[e]	T	G	1.42	1.23	1.63	2.6×10^{-3}	7.2×10^{-5}
rs896416[b]	7.7×10^{-6}	8	99199066	Downstream	*NIPAL2*[e]	C	T	1.37	1.19	1.57	4.3×10^{-3}	3.6×10^{-4}
rs10809674[a]	5.6×10^{-6}	9	12015741	Intergenic	*RP11-74C3.1*[d]	A	G	1.47	1.25	1.74	5.9×10^{-5}	2.6×10^{-2}
rs202983[c]	8.9×10^{-6}	12	120162346	Intronic	*CIT*	A	G	1.52	1.26	1.82	1.7×10^{-3}	1.4×10^{-3}
rs7305618[b]	2.8×10^{-7}	12	121402932	Intergenic	*HNF1A*[e]	T	C	1.69	1.38	2.06	7.5×10^{-5}	9.9×10^{-4}
rs5998144[b]	4.5×10^{-6}	22	32248244	Intronic	*DEPDC5*[e]	T	C	1.56	1.29	1.89	3.2×10^{-2}	1.6×10^{-5}
With BMI												
rs10465543[c]	9.40×10^{-6}	1	23134218	Intronic	*EPHB2*[e]	G	A	1.52	1.26	1.83	1.2×10^{-3}	2.4×10^{-3}
rs7516189[c]	5.73×10^{-6}	1	211828134	Upstream	*RP11-354K1.*1[e,d]	G	A	1.48	1.25	1.75	1.1×10^{-3}	1.2×10^{-3}
rs9831045[c]	9.16×10^{-6}	3	192475259	Intronic	*FGF12*	T	A	1.39	1.20	1.61	7.3×10^{-3}	1.9×10^{-4}
rs413019[a]	6.71×10^{-6}	4	114079127	Intronic	*ANK2*[e]	G	A	1.41	1.21	1.64	2.6×10^{-4}	7.9×10^{-3}
rs9773092[a]	6.11×10^{-7}	8	3216025	Intronic	*CSMD1*[e]	T	G	1.43	1.24	1.65	2.1×10^{-3}	5.7×10^{-5}
rs4841333[a]	3.70×10^{-6}	8	6453786	Intronic	*MCPH*1[e]	A	G	1.50	1.26	1.78	3.2×10^{-5}	2.7×10^{-2}
rs10809674[a]	1.93×10^{-6}	9	12015741	Intergenic	*RP11-74C3.1*[e,d]	A	G	1.52	1.28	1.80	2.7×10^{-5}	1.9×10^{-2}
rs2237892[c]	6.75×10^{-6}	11	2839751	Intronic	*KCNQ1*[e]	T	C	1.47	1.24	1.74	5.0×10^{-3}	3.4×10^{-4}
rs202983[c]	2.71×10^{-6}	12	120162346	Intronic	*CIT*	A	G	1.57	1.30	1.90	8.3×10^{-4}	8.6×10^{-4}
rs7305618[b]	3.91×10^{-7}	12	121402932	Intergenic	*HNF1A*[e]	T	C	1.69	1.38	2.08	7.0×10^{-5}	1.6×10^{-3}
rs5998144[b]	1.68×10^{-6}	22	32248244	Intronic	*DEPDC5*[e]	T	C	1.61	1.32	1.96	2.1×10^{-2}	9.3×10^{-6}

The *p* values of the meta-analysis are based on the results of the Mexico City and Starr County score tests; none of the markers showed evidence of heterogeneity of effects between studies (Cochran's Q, $p>0.1$)

All imputed markers had score information >0.8, except rs938911 and rs7305618, which had score information >0.7

[a] Directly genotyped in SC and MC samples; [b] directly genotyped in SC and imputed in MC; [c] imputed in SC and MC samples; [d] ENCODE Genecode gene annotations; [e] More than one marker in this region showed a *p* value of $p<10^{-5}$; only the marker with the lowest *p* value is shown

Chr Chromosome; Pos Position; NEA Non-effect allele; EA Effect allele; OR 95L Lower 95% confidence limit of odds ratio; OR 95U Upper 95% confidence limit of odds ratio; *p*-MC *p* value in Mexico City; *p*-SC *p* value in Starr County

Consortium [WTCCC], Diabetes Genetics Initiative [DGI] and the Finland–United States Investigation of NIDDM Genetics [FUSION] studies). Table 2 describes the markers with global values of $p<10^{-4}$ that show the same direction of effect in the Hispanic and European studies. The list includes markers in five regions that have been identified in previous GWA studies: *IGF2BP2* on chromosome 3 (lowest *p* value $p=9.5\times10^{-6}$), *PTPRD* on chromosome 9 (lowest *p* value $p=4.4\times10^{-7}$), a region near *CDKN2A/CDKN2B* on chromosome 9 (lowest *p* value $p=9.6\times10^{-10}$), *KCNQ1* on chromosome 11 (lowest *p* value $p=5.6\times10^{-6}$) and *HNF1A* on chromosome 12 ($p=4.4\times10^{-5}$). There was also one SNP in a region not previously identified in GWA studies, *C14orf70* on chromosome 14 ($p=4.0\times10^{-6}$).

Recently, the DIAGRAM consortium expanded their meta-analysis to include eight GWA studies in individuals of European ancestry (DIAGRAM+) [1]. We did not have access to the full DIAGRAM+ dataset, but we were able to follow up the signals observed in our meta-analysis of the Hispanic (Mexico City and Starr County) and 2008 DIAGRAM studies in this expanded dataset. Table 3 reports the results of this analysis. Two of the previously described regions achieved genome-wide significance ($p<5\times10^{-8}$): *CDKN2A/CDKN2B* ($p=5.7\times10^{-10}$) and *HNF1A* ($p=1.9\times10^{-8}$). *IGF2BP2*, *KCNQ1* and *C14orf70* also showed suggestive associations with type 2 diabetes ($p=1.3\times10^{-7}$, $p=3.8\times10^{-6}$ and $p=8.0\times10^{-6}$, respectively). However, the *p* value for *PTPRD* decreased substantially ($p=0.047$), driven by the changes observed in the 2008 DIAGRAM and the DIAGRAM+ datasets.

Of the top signals ($p<10^{-5}$) observed in the meta-analysis of the Mexico City and Starr County studies (Table 1), the *HNF1A* and *KCNQ1* regions were confirmed in the European dataset. Also, the marker rs10465543 within the *EPHB2* gene was nominally significant in the DIAGRAM+ dataset ($p=2.5\times10^{-2}$). None of the remaining regions (we followed up only the markers with the lowest *p*

Table 2 List of markers with combined *p* values of $p<10^{-4}$ in the Hispanic and 2008 DIAGRAM dataset

SNP	Chr	NEA	EA	*p*-META	OR	OR 95L	OR 95U	Type	Closest gene	*p*-DIAG	*p*-Hisp
rs1374910	3	C	T	9.5×10^{-6}	1.30	1.16	1.45	Intronic	*IGF2BP2*	6.1×10^{-3}	9.6×10^{-5}
rs1544056	9	C	A	4.4×10^{-7}	1.18	1.11	1.26	Intronic	*PTPRD*	2.3×10^{-4}	1.2×10^{-4}
rs1333051	9	T	A	9.6×10^{-10}	1.32	1.21	1.44	Intergenic	*CDKN2A/CDKN2B*	3.0×10^{-7}	2.3×10^{-4}
rs2237892	11	T	C	5.6×10^{-6}	1.26	1.14	1.40	Intronic	*KCNQ1*	1.4×10^{-2}	2.3×10^{-5}
rs7305618	12	T	C	4.4×10^{-5}	1.16	1.08	1.24	Intergenic	*HNF1A*	1.5×10^{-2}	2.8×10^{-7}
rs730570	14	A	G	4.0×10^{-6}	1.21	1.12	1.31	Downstream	*C14orf70*	9.7×10^{-4}	7.0×10^{-4}

Only markers that show the same direction of effect in the Hispanic and European studies are included

Chr Chromosome; NEA Non-effect allele; EA Effect allele; *p*-META *p* value Hispanic/2008 DIAGRAM meta-analysis; OR Odds ratio; OR 95L Lower 95% confidence limit of odds ratio; OR 95U Upper 95% confidence limit of odds ratio; *p*-DIAG *p* value 2008 DIAGRAM dataset; *p*-Hisp *p* value Hispanic meta-analysis

value in each region) were nominally significant in the DIAGRAM+ dataset.

Discussion

We carried out a GWA study of type 2 diabetes in a sample from Mexico City. We observed strong evidence of population stratification due to variation in admixture proportions within the sample (Fig. 1). These results are in agreement with our previous research, in which we had described an association between American Indian ancestry and type 2 diabetes in a sample from Mexico City [15]. However, it is important to note that socioeconomic status was inversely correlated with diabetes risk and American Indian ancestry in that study. A strong association between socioeconomic status and genetic ancestry has also been reported in other admixed samples in the Americas [16]. Unfortunately, we do not have information on socioeconomic status for all the samples included in the present analysis. However, an analysis of the subset of samples for which we have information on education showed, in agreement with our previous research, that American Indian ancestry was inversely correlated with education (data not shown). Thus, some of the association between ancestry and disease risk observed in our sample could have been mediated by the effects of socioeconomic status. Irrespective of the factors responsible for the strong population stratification observed in the sample from Mexico City, inclusion of ancestry in the statistical model corrects for the effects of population stratification (ESM Figs 1 and 2). Our study emphasises the need to include ancestry in GWA studies in admixed populations, in order to control for confounding. The comparison of our sample from Mexico City with other samples of Mexican ancestry from the US (Mexican–Americans from Los Angeles and Starr County) highlights the large variation in admixture proportions (and also in the extent of population stratification) that may be present in populations of Mexican ancestry living in different geographic areas. For example, based on a genome-wide panel of markers, we estimated that the average American Indian contribution to the Mexican–American HapMap and Starr County samples is substantially lower than that observed in Mexico City (approximately 20% and 30% lower, respectively). Additionally, in contrast to the strong stratification observed in Mexico City, the scores of the top axes of the principal component analysis were not

Table 3 Follow-up of markers listed in Table 2 in the recent DIAGRAM+ dataset

SNP	Chr	NEA	EA	*p*-META	OR	OR 95L	OR 95U	Type	Closest gene	*p*-DIAG+	*p*-Hisp
rs1374910	3	C	T	1.3×10^{-7}	1.24	1.15	1.34	Intronic	*IGF2BP2*	4.1×10^{-5}	9.6×10^{-5}
rs1544056	9	C	A	4.7×10^{-2}	1.05	1.00	1.10	Intronic	*PTPRD*	4.5×10^{-1}	1.2×10^{-4}
rs1333051	9	T	A	5.7×10^{-10}	1.22	1.15	1.30	Intergenic	*CDKN2A/CDKN2B*	7.1×10^{-8}	2.3×10^{-4}
rs2237892	11	T	C	3.8×10^{-6}	1.20	1.11	1.29	Intronic	*KCNQ1*	2.7×10^{-3}	2.3×10^{-5}
rs7305618	12	T	C	1.9×10^{-8}	1.14	1.09	1.20	Intergenic	*HNF1A*	6.0×10^{-6}	2.8×10^{-7}
rs730570	14	A	G	8.0×10^{-6}	1.14	1.08	1.21	Downstream	*C14orf70*	6.6×10^{-4}	7.0×10^{-4}

Chr Chromosome; NEA Non-effect allele; EA Effect allele; *p*-META *p* value Hispanic/DIAGRAM+ meta-analysis; OR Odds ratio; OR 95L Lower 95% confidence limit of odds ratio; OR 95U Upper 95% confidence limit of odds ratio; *p*-DIAG+ *p* value DIAGRAM+ dataset; *p*-Hisp *p* value Hispanic meta-analysis significantly associated with type 2 diabetes in the Starr County sample.

In our statistical analysis of the Mexico City sample, we identified several regions showing suggestive associations with type 2 diabetes (ESM Tables 1 and 2). We also performed a meta-analysis of two type 2 diabetes GWA studies in Hispanics (Mexico City and Starr County, Texas, USA), further exploring the SNPs that showed nominal significance and similar direction of effect in those analyses in samples of European ancestry (2008 DIAGRAM and DIAGRAM+ datasets). Two of the top signals ($p<10^{-5}$) observed in the Mexico City/Starr County meta-analysis correspond to regions identified in previous GWA studies: *HNF1A* (hepatocyte nuclear factor 1-alpha) and *KCNQ1* (potassium voltage-gated channel, KQT-like subfamily, member 1).

Variants within and near *HNF1A* showed the lowest *p* values in our Hispanic meta-analysis. Functional mutations in this gene underlie MODY3, (MIM number 600496), a familial form of non-insulin-dependent diabetes mellitus with early age of onset and autosomal dominant inheritance. Additionally, a marker near *HNF1A* reached genome-wide significance in the joint analysis of stage 1 and 2 data in the recent DIAGRAM+ study (rs7957197, $p=2.4\times10^{-8}$) [1]. This meta-analysis in European samples added *HNF1A* to an increasing list of loci involved in monogenic/syndromic forms of diabetes and in the more common, multifactorial type 2 diabetes, which currently includes *GCK*, *HNF1B*, *PPARG*, *WFS1* and *KCNJ11* [2]. The SNP with the lowest *p* value in our Hispanic meta-analysis (rs7305618) (Table 1) also showed suggestive *p* values in the DIAGRAM+ study ($p=6.0\times10^{-6}$) and achieved genome-wide significance in the combined Hispanic/DIAGRAM+ meta-analysis ($p=1.9\times10^{-8}$, OR 1.14) (Table 3). The sentinel SNP reported in the DIAGRAM+ study (rs7957197) is located less than 60 kilobases apart from rs7305618 on chromosome 12. Unfortunately, data for rs7957197 were not available for the Starr County sample. However, these two SNPs are in relatively strong LD in the HapMap European sample ($r^2=0.54$, D′=0.923) and they show similar *p* values in the Mexico City sample (rs7957197 $p=0.0012$; rs7305618 $p=0.001$), suggesting that the Hispanic and European studies may be capturing the same signal.

Polymorphisms within the gene *KCNQ1* were originally associated with susceptibility to type 2 diabetes in two relatively small GWA studies in Japanese samples [3, 4] and have been confirmed in many subsequent studies, including two recent GWA studies [7, 8]. Interestingly, one of the intronic variants identified in our Hispanic meta-analysis, rs2237892 ($p=4.4\times10^{-5}$) is the same variant described in one of the original Japanese studies [3]. This SNP also had a nominally significant *p* value in the DIAGRAM+ dataset ($p=2.7\times10^{-3}$), while the *p* value of the Hispanic/DIAGRAM+ meta-analysis was 3.8×10^{-6} (OR 1.20) (Table 3). Similarly to the situation described in East Asians, the frequency of the rs2237892 risk allele (C allele) is substantially higher in Europeans than in American Indian populations (data not shown). Importantly, it should be noted that there is now evidence of two independent signals in *KCNQ1*. The recent DIAGRAM+ study identified a genome-wide signal in this gene (lead SNP, rs231362, $p=2.8\times10^{-13}$), which is independent of the markers identified in East Asian scans [1]. The SNPs rs231362 and rs2237892 are not correlated in the European HapMap ($r^2=0.019$), East Asian HapMap ($r^2=0.016$) or the Mexico City samples ($r^2=0.001$). The marker rs231362 is not significant in the Mexico City sample ($p=0.195$), and the *p* value in the Starr County study was $p=0.040$ (meta-analysis $p=0.016$).

Other loci identified in the Hispanic/DIAGRAM+ meta-analysis and reported in previous GWA studies are *CDKN2A/CDKN2B* (cyclin-dependent kinase inhibitor 2A and 2B) on chromosome 9 and *IGF2BP2* (insulin-like growth factor 2 mRNA binding protein 2) on chromosome 3. Several GWA studies in European and Japanese populations have identified genome-wide associations between markers near *CDKN2A/CDKN2B* and type 2 diabetes [7, 14, 17–19]. In our Hispanic meta-analysis, the SNP rs1333051 had a *p* value of 2.3×10^{-4}. The *p* value in DIAGRAM+ for this marker was $p=7.1\times10^{-8}$ and the combined Hispanic/DIAGRAM+*p* value reached genome-wide significant levels ($p=5.7\times10^{-10}$, OR 1.22). In the Mexico City sample, rs1333051 is in strong LD with the lead SNPs reported in Japanese and European studies, e.g. rs2383208 [7] and rs10811661 [19] ($r^2>0.75$, D′>0.95). SNPs within the *IGF2BP2* gene have also been associated with type 2 diabetes in many GWA studies [1, 4, 7, 14, 17–19]. The intronic variant rs1374910 located within this gene showed suggestive evidence of association in the Hispanic/DIAGRAM+ meta-analysis ($p=1.3\times10^{-7}$, OR 1.24). In the Mexico City sample, rs1374910 showed weak LD with the sentinel SNPs reported in previous studies, e.g. rs1470579 [1], rs6769511 [4] and rs4402960 [17–19] ($r^2<0.10$, D′<0.45), while the *p* value for rs1374910 ($p=9.6\times10^{-5}$) in the Hispanic meta-analysis was substantially lower than those observed for the other three markers ($p=0.01$ approx.). The opposite is the case for the DIAGRAM+ study (rs1470579, $p=2\times10^{-9}$ and rs1374910, $p=4.1\times10^{-5}$). It would be interesting to explore this region in more detail in additional populations.

Finally, we observed another region on chromosome 14 showing suggestive association with type 2 diabetes in the Hispanic/DIAGRAM+ meta-analysis. The SNP rs730570, which is very close to *C14orf70* (chromosome 14 open reading frame 70), showed a combined *p* value of $p=8.0\times$

10^{-6} in this analysis (OR 1.14) (Table 3). To our knowledge, this region has not been reported in previous GWA studies. It and other markers in this region should therefore be followed up in additional studies in order to confirm its potential involvement in type 2 diabetes risk.

Many of the top signals observed in our meta-analysis of the Mexico City and Starr County studies could not be confirmed in the DIAGRAM+ dataset. The markers identified near or within the genes *RP11-354K1.1*, *LSAMP*, *FGF12*, *ANK2*, *CSMD1*, *MCPH1*, *NIPAL2*, *SLC5A8*, *CIT*, *CHRNA7* and *DEPDC5*, as well as the intergenic region on chromosome 9 (Table 1), were not nominally significant in the DIAGRAM+ meta-analysis. Similarly, one of the loci initially identified in the Hispanic/2008 DIAGRAM meta-analysis (*PTPRD*) could not be confirmed in the DIAGRAM+ dataset. It is possible that some of these signals are just false positives. Alternatively, these discrepancies may be due to population differences between the Hispanic and DIAGRAM datasets. The two samples of Mexican ancestry show a substantial American Indian component (in fact, the sample from Mexico City has primarily American Indian ancestry). It is crucial to follow up these regions in additional Hispanic and/or American Indian samples in order to clarify these issues. Interestingly, some of the genes mentioned above have been previously associated with type 2 diabetes or diabetes-related traits in previous studies, and/or are relevant from the functional point of view. For example, the gene *PTPRD* (protein tyrosine phosphatase, receptor type, D) has been previously associated with type 2 diabetes in a recent GWA study in Han Chinese [8]. However, the variant identified in the Chinese study (rs17584499) was not significant in our meta-analysis (p=0.517). A recent study exploring the genetic determinants of metabolic syndrome in the Framingham Heart Study (Offspring Cohort, Exam 7) identified an association between *CSMD1* (CUB and Sushi multiple domains 1) and metabolic syndrome using two different approaches [20]. Finally, recent studies have highlighted the potential role of the product of *ANK2*, ankyrin-B, in metabolic regulation [21, 22]. Kline et al. [21] demonstrated that the C-terminal domain of Kir6.2, a critical component of K_{ATP} channels in many cells including pancreatic cells, contains a C-terminal ankyrin-B-binding motif that is important in K_{ATP} channel trafficking and membrane metabolic regulation. Healy et al. [22] reported that ankyrin-B is involved in parasympathetic potentiation of glucose-stimulated insulin secretion, and that the R1788W mutation of *ANK2* impaired its function in pancreatic islets and is associated with type 2 diabetes in individuals of European descent and Hispanics.

In summary, we report the results of a GWA study of type 2 diabetes in an admixed sample from Mexico City and of a meta-analysis combining this study with another genome-wide survey in a Mexican-American sample from Starr County, Texas, USA. It is critical to carry out additional GWA studies in Hispanic and American Indian populations, which have a high prevalence of type 2 diabetes. Most of the recent advances in our knowledge of the genetic architecture of type 2 diabetes have been driven primarily by GWA studies in European populations, which have sampled tens of thousands of individuals. Further advancement will require a similar effort in other population groups in order to obtain a more complete understanding of the common genetic factors involved in type 2 diabetes risk.

Acknowledgements We thank all those individuals who participated in the study. In Canada this work was supported by the Canadian Institutes of Health Research, the Banting and Best Diabetes Centre, the Canada Foundation for Innovation and The Ontario Innovation Trust. In Mexico this work was supported by the following grants: CONACYT SALUD-2005-C02-14412, Proyectos Estrategicos, Apoyo Financiero Fundacion IMSS and Fundacion Gonzalo Rio Arronte I, and by A. P. Mexico and Apoyos a la Vinculacion del Cinvestav con el Sector Salud 2006. M. Cruz is a recipient of a Fundación IMSS Scholarship, Mexico. We thank the members of the DIAGRAM consortium for kindly sharing the summary data of their meta-analysis.

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

References

1. 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
2. McCarthy MI (2010) Genomics, type 2 diabetes, and obesity. N Engl J Med 363:2339–2350
3. Yasuda K, Miyake K, Horikawa Y et al (2008) Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat Genet 40:1092–1097
4. Unoki H, Takahashi A, Kawaguchi T et al (2008) SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet 40:1098–1102
5. 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
6. Hanson RL, Bogardus C, Duggan D et al (2007) A search for variants associated with young-onset type 2 diabetes in American Indians in a 100 K genotyping array. Diabetes 56:3045–3052
7. Takeuchi F, Serizawa M, Yamamoto K et al (2009) Confirmation of multiple risk loci and genetic impacts by a genome-wide association study of type 2 diabetes in the Japanese population. Diabetes 58:1690–1699
8. 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
9. McCarthy MI (2008) Casting a wider net for diabetes susceptibility genes. Nat Genet 40:1039–1040
10. Below JE, Gamazon ER, Morrison JV et al (2011) Genome-wide association and meta-analysis in populations from Starr County, Texas and Mexico City identify type 2 diabetes susceptibility loci and enrichment for eQTLs in top signals. Diabetologia doi:10.1007/s00125-011-2188-3

11. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81:559–575
12. 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
13. Ge D, Zhang D, Need AC et al (2008) WGAViewer: software for genomic annotation of whole genome association studies. Genome Res 18:640–643
14. 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
15. 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
16. Florez JC, Price AL, Campbell D et al (2009) Strong association of socioeconomic status with genetic ancestry in Latinos: implications for admixture studies of type 2 diabetes. Diabetologia 52:1528–1536
17. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, Novartis Institutes of BioMedical Research et al (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316:1331–1336
18. Zeggini E, Weedon MN, Lindgren CM et al (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316:1336–1341
19. Scott LJ, Mohlke KL, Bonnycastle LL et al (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341–1345
20. Nock NL, Wang X, Thompson CL et al (2009) Defining genetic determinants of the metabolic syndrome in the Framingham Heart Study using association and structural equation modeling methods. BMC Proc 3(Suppl 7):S50
21. Kline CF, Kurata HT, Hund TJ et al (2009) Dual role of K ATP channel C-terminal motif in membrane targeting and metabolic regulation. Proc Natl Acad Sci U S A 106:16669–16674
22. Healy JA, Nilsson KR, Hohmeier HE et al (2010) Cholinergic augmentation of insulin release requires ankyrin-B. Sci Signal 3: ra19