Diabetologia (2009) 52:600–608
DOI 10.1007/s00125-008-1254-y

ARTICLE

Risk prediction of prevalent diabetes in a Swiss population using a weighted genetic score—the CoLaus Study

X. Lin · K. Song · N. Lim · X. Yuan · T. Johnson · A. Abderrahmani · P. Vollenweider · H. Stirnadel · S. S. Sundseth · E. Lai · D. K. Burns · L. T. Middleton · A. D. Roses · P. M. Matthews · G. Waeber · L. Cardon · D. M. Waterworth · V. Mooser

Received: 17 July 2008 / Accepted: 3 December 2008 / Published online: 13 January 2009

Abstract
Aims/hypothesis Several susceptibility genes for type 2 diabetes have been discovered recently. Individually, these genes increase the disease risk only minimally. The goals of the present study were to determine, at the population level, the risk of diabetes in individuals who carry risk alleles within several susceptibility genes for the disease and the added value of this genetic information over the clinical predictors.
Methods We constructed an additive genetic score using the most replicated single-nucleotide polymorphisms (SNPs) within 15 type 2 diabetes-susceptibility genes, weighting each SNP with its reported effect. We tested this score in the extensively phenotyped population-based cross-sectional CoLaus Study in Lausanne, Switzerland (n=5,360), involving 356 diabetic individuals.
Results The clinical predictors of prevalent diabetes were age, BMI, family history of diabetes, WHR, and triacylglycerol/HDL-cholesterol ratio. After adjustment for these variables, the risk of diabetes was 2.7 (95% CI 1.8–4.0, p=0.000006) for individuals with a genetic score within the top quintile, compared with the bottom quintile. Adding the genetic score to the clinical covariates improved the area under the receiver operating characteristic curve slightly (from 0.86 to 0.87), yet significantly (p=0.002). BMI was similar in these two extreme quintiles.
Conclusions/interpretation In this population, a simple weighted 15 SNP-based genetic score provides additional

X. Lin
Discovery Analytics, GlaxoSmithKline,
Collegeville, PA, USA

K. Song · N. Lim · X. Yuan · L. T. Middleton · L. Cardon · D. M. Waterworth · V. Mooser (✉)
Genetics Division, GlaxoSmithKline,
709 Swedeland Road, UW2111,
King of Prussia, PA 19406, USA
e-mail: Vincent.2.mooser@gsk.com

T. Johnson
Division of Medical Genetics, CHUV University Hospital,
Lausanne, Switzerland

T. Johnson
Swiss Institute of Bioinformatics,
Lausanne, Switzerland

A. Abderrahmani · P. Vollenweider · G. Waeber
Department of Medicine, CHUV University Hospital,
Lausanne, Switzerland

H. Stirnadel
Worldwide Epidemiology, GlaxoSmithKline,
London, UK

S. S. Sundseth · E. Lai · D. K. Burns · A. D. Roses
Genetics Division, GlaxoSmithKline,
Research Triangle Park, NC, USA

A. D. Roses
R. David Thomas Center, Duke University,
Durham, NC, USA

P. M. Matthews
Clinical Imaging Center, GlaxoSmithKline,
London, UK

Present address:
L. T. Middleton
Department of Clinical Neurosciences,
Imperial College,
London, UK

information over clinical predictors of prevalent diabetes. At this stage, however, the clinical benefit of this genetic information is limited.

Keywords Diabetes · Genetics · Obesity · Population · Prediction · SNP

Abbreviations

IDI	Integrated discriminant index
QC	Quality control
ROC	Receiver operating characteristic
SNP	Single-nucleotide polymorphism

Introduction

Several clinical variables are associated with incident type 2 diabetes including age, sex, obesity, family history of diabetes, dyslipidaemia, smoking, hypertension and lack of physical activity [1–5]. Recently a series of susceptibility genes for this disease have been discovered [6–10]. Taken individually, the size of the effects of the risk alleles is generally modest, in the order of 5–20% [7, 8]. To what extent the accumulation of risk alleles within multiple susceptibility genes predisposes to diabetes, the added value of the information derived from genetic markers over clinical predictors, and the way to derive from risk alleles a genetic score with an optimal predictive value are of major importance, as a better prediction of diabetes would allow preventative behavioural and pharmacological measures to be deployed in a more targeted, efficient and cost-saving manner [11, 12]. These questions are not specific to diabetes, but are equally relevant to other diseases for which genome-wide scans have now revealed several susceptibility genes [13].

For type 2 diabetes, analyses of three susceptibility genes in the Botnia Study [14], in the UK Diabetes Study [15] and in the French Data from an Epidemiological Study on the Insulin Resistance syndrome (DESIR) Study [16] indicate that accumulation of multiple risk alleles markedly increases the risk of diabetes, and evidence has been provided that incorporation of additional genes should improve the diabetes risk prediction [17]. Recently Lango et al. [18] have shown that, in a case–control study involving 2,309 diabetic and 2,598 non-diabetic individuals, the risk of diabetes rose in proportion to the added number of risk alleles within 18 susceptibility genes for diabetes, and that diabetic patients with a large number of risk alleles had an earlier age of onset of the disease. In this particular study though, the discriminatory power of the genetic score over age, BMI and sex was small, and limited phenotypic information was available. The relevance of these findings at a population level and the value added by the genetic information over family history of diabetes and other clinical predictors of the disease remained unknown. Very recently, the ability of these risk alleles to predict incident disease has been evaluated in two prospective populations, the Framingham [19] and the combined Malmö and Botnia studies [20]. Both studies showed that once all non-genetic predictors were taken into account, the addition of the genetic score provided only a marginal increase in prediction. The fact that risk alleles within susceptibility genes vary somehow in their effect was not taken into account in the generation of the additive genetic score in any of the above studies, with the exception of that of Meigs et al. [19]. In this particular study, weighting the genetic score did not change substantially its predictive value.

In the present study, we first compared the clinical characteristics of 356 diabetic individuals and 5,004 non-diabetic controls in the extensively phenotyped CoLaus population-based study in Lausanne, Switzerland [21, 22]. We then performed a multivariate logistic regression analysis to identify the clinical variables that were independently associated with the presence of the disease. Subsequently, we constructed a genetic score using the most replicated single-nucleotide polymorphism (SNP) within 15 susceptibility genes for diabetes (two SNPs were not available in this cohort). As the amplitude of the effect varies somewhat between diabetes-susceptibility genes, and in an attempt to penalise those SNPs with a less reliable OR estimate, we weighted here each SNP using the log lower boundary of the reported 95% CI. We used an additive model consistent with the additive risk described for most of the susceptibility genes for diabetes and the absence of interactions between these genes [18]. We next compared the predictive value of this weighted genetic score with the unweighted score. Finally, we examined the discriminatory power of the weighted additive genetic score over the clinical predictors of the disease using receiver operator characteristic (ROC) curve and integrated discriminant index (IDI) analyses [20, 23].

Methods

Design of the CoLaus Study and phenotypic assessment The CoLaus Study has been described previously [21]. Briefly, 6,200 white individuals aged 35–75 years were randomly selected from the general population in Lausanne, Switzerland. These individuals underwent a detailed phenotypic assessment including measurement of several metabolic markers in fasting blood samples. Diabetes was defined as fasting blood glucose ≥7.0 mmol/l or prescription of glucose-lowering drugs. The study was sponsored in part by GlaxoSmithKline and each participant was duly informed about, and consented to, the use of their data and

samples by this company and/or its affiliates. The protocol was approved by the local ethics committee.

Genotyping, quality controls and imputation The list of 17 susceptibility genes for diabetes, the corresponding SNPs and the sizes of their effects were extracted from published meta-analyses [8, 9]. Eleven of these SNPs were included in the Affymetrix 500K SNP chip which had been used to genotype 6,000 CoLaus participants. Genotype quality control (QC) was performed to detect genotype inconsistencies and check the genotyping efficiency. The first step covered QC of individuals. Samples were removed from the analysis if: (1) sex was inconsistent with genetic data from X-linked markers; (2) the returned genotype call rate was <90%; or (3) genotypes were inconsistent compared with control markers. We next performed a QC analysis for SNP markers. Markers were removed if: (1) they were monomorphic in all samples; (2) the genotype call was <95%; or (3) the p value for Hardy–Weinberg equilibrium was $<10^{-7}$. Based on these criteria, a total of 640 individuals were removed from the analysis so that the full set of phenotypic data and these 11 genotyped SNPs were available for 5,360 individuals.

The remaining six SNPs were imputed. For that purpose, we applied the IMPUTE version 0.2.0 method of Marchini and Howie (www.stats.ox.ac.uk/~marchini/software/gwas/impute.html, accessed 17 December 2008), and haplotypes derived from the CEU samples (Centre d'Etude du Polymorphisme Humain trios originating from northern and western Europe living in UT, USA) and fine-scale recombination maps from HapMap Release 21 (www.hapmap.org/downloads/genotypes/2006-7/, accessed 17 December 2008) including 390,631 measured SNPs. The output of this version gives genotype probabilities for a given SNP based on information from all measured SNPs. Only SNPs with a minor allele frequency ≥1% and with an average maximum-posterior probability score >0.90 were included in the present analyses. Two of the imputed SNPs (rs4430796 for *TCF2* and rs13266634 for *SLC30A8*, with a reported OR for diabetes of 1.10 [1.07–1.14] and 1.15 [1.12–1.19], respectively) did not pass this QC, so that these two genes were removed and 15 genes/SNPs were eventually included in the present analysis. Twelve of these 15 SNPs were used in the report by Lango et al. [18]. Three SNPs (rs7901695 for *TCF7L2*, rs5215 for *KCNJ11* and rs10923931 for *ADAM30*) were used as proxy SNPs in the present study.

Construction of the weighted genetic score The weighted genetic score was generated using the following equation:

$$\text{weighted genetic score} = w_1 \times \text{SNP}_1 + w_2 \times \text{SNP}_2 + \ldots + w_k \times \text{SNP}_k \quad (1)$$

where SNP_i=0, 1 or 2 according to the number of risk alleles for the specific locus, w_i is the appropriate weight to be determined and k is the number of SNPs used (i.e. k=15). The weighted score was derived based on the assumption that the SNPs of interest have independent effects on the disease and contribute to the log risk of the disease in an additive manner [18]. Under these assumptions, we derived the following relationship between the log odds of having disease given the joint value of the SNPs:

$$\log\left(\frac{P(D=1|\,\text{SNP}_1,\ldots,\text{SNP})_k}{P(D=0|\,\text{SNP}_1,\ldots,\text{SNP}_k)}\right) = \text{C} + \sum_{i=1}^{k} \log(\text{OR}_i)\,\text{SNP}_i \quad (2)$$

where $P(\ldots)$ indicates the probability function, C is a constant specific to a dataset, OR_i is the OR per allele at the ith SNP of having the disease, SNP_i is a genotype coded 0/1/2 and D=1 and D=0 indicate an individual having or not having the disease, respectively. Based on Eq. 2, one could have used w_i=log(OR_i) if the true ORs and these estimates were known and were reliable and accurate. However, the ORs used here were based on meta-analyses of various studies, with different sample sizes and designs. Hence, to be conservative and in an attempt to penalise those SNPs with a lower confidence or a less reliable OR estimate, and in an effort to avoid inflation of SNPs with a lower boundary of the 95% CI close to 1, we used $\log_e$ transformation of the lower boundary of 95% CI as the w_i. To make the weighted genetic score more comparable to the unweighted genetic score (cumulative number of alleles), we used the rescaled version of the genetic score using the rescaling factor $k/(w_1+\ldots+w_k)$:

$$\text{(rescaled) weighted genetic score} = k \times (w_1 \times \text{SNP}_1 + w_2 \times \text{SNP}_2 + \ldots + w_k \times \text{SNP}_k)/(w_1 + \ldots w_k). \quad (3)$$

To determine the variables independently associated with type 2 diabetes, we performed a multivariate logistic regression analysis, including the variables listed in Table 1 with the exception of glucose, insulin and type 2 diabetes medication usage. The covariates with p values <0.05 level were included in the final model. The Hosmer–Lemeshow test was used as a calibration statistic to check the goodness of fit of the final models [19, 24]. A χ^2 statistic was calculated to compare the predicted and observed event rates. A model with χ^2 statistic <20 (p>0.01) is usually considered a good calibration. The added contribution of the weighted genetic score to the prediction of prevalent diabetes was evaluated using a ROC analysis and a non-parametric comparison of areas under these curves [25]. We also performed an IDI analysis [23] of the model comparing covariates and weighted genetic score vs covariates only.

Table 1 Clinical characteristics of the participants in the study

Characteristic	Diabetic individuals	Non-diabetic individuals	*p* value[a]
n	356	5,004	
Sex (% female)	32.6	54.0	<0.0001
Age (years)	60.7±8.5	52.8±10.7	<0.0001
Glucose (mmol/l)	7.66±1.44	5.36±0.56	<0.0001
Insulin (pmol/l)	99.2±74.6	59.1±38.3	<0.0001
BMI (kg/m^2)	30.4±6.2	25.5±4.3	<0.0001
Waist (cm)	103.6±14.8	88.3±12.7	<0.0001
Hip (cm)	108.7±11.6	101.2±8.9	<0.0001
WHR	0.95±0.08	0.87±0.08	<0.0001
Total cholesterol (mmol/l)	5.46±1.19	5.61±1.03	0.003
HDL-cholesterol (mmol/l)	1.37±0.36	1.66±0.43	<0.0001
Triacylglycerol (mmol/l)	2.14±2.08	1.34±1.08	<0.0001
Triacylglycerol/HDL-cholesterol ratio	1.75±1.80	0.95±1.10	<0.0001
LDL-cholesterol (mmol/l)	3.20±1.03	3.36±0.91	0.002
Systolic BP (mmHg)	138.5±17.8	127.8±17.7	<0.0001
Diastolic BP (mmHg)	82.0±11.2	79.2±10.7	<0.0001
Type 2 diabetes medication usage (%)	65.7	0	<0.0001
Statin usage (%)	22.8	7.6	<0.0001
Physical activity (%)	48.9	67.3	<0.0001
Family history of type 2 diabetes (%)	42.1	21.1	<0.0001
Weighted genetic score	15.2±2.9	14.3±2.7	<0.0001
Unweighted genetic score (number of risk alleles)	15.2±2.5	14.6±2.4	<0.0001

Values are means±SD

[a] *p* value for comparison between diabetic and non-diabetic individuals

The IDI was estimated by computing the difference between the integrated difference in sensitivities and the integrated difference in 1−specificities between the covariates and weighted genetic score model and the covariates only model. The integration was performed over all possible cut-offs. Evaluation of model predictive performance using the same dataset used for fitting the model usually leads to a biased assessment. To obtain an unbiased assessment of discriminatory power of the multivariate regression models, a tenfold cross-validation was used in the ROC analysis and in the IDI analysis. Tenfold cross-validation randomly divides the data into ten (roughly) equal subsets and repeatedly uses any nine subsets for model fitting and the remaining subset as validation until each of the ten subsets has been used exactly once as validation data.

Results

The clinical characteristics of the 5,360 participants, including 356 diabetic and 5,004 non-diabetic individuals, are described in Table 1. As expected, diabetic individuals were older than non-diabetic and had higher levels of blood pressure, insulin and triacylglycerol levels, and lower HDL-cholesterol levels. Plasma levels of total- and LDL-cholesterol were lower in diabetic participants than in non-diabetic, presumably because of the broader usage of statins in this former group. The proportion of individuals engaged in regular physical activity was lower among diabetic patients. Twice as many diabetic than non-diabetic individuals had a positive family history of diabetes, defined as having at least one first-degree relative with diabetes.

To determine the variables that were associated with the presence of diabetes in this population, we performed a multivariate logistic regression analysis (Table 2). The variables that were significantly and independently associated with diabetes included age, BMI, family history of diabetes, WHR, triacylglycerol/HDL-cholesterol ratio and lack of regular engagement in physical activity.

Fifteen SNPs were measured or imputed for each of the 5,360 participants (Table 3). Overall, the risk allele frequency was similar in the CoLaus Study and the published meta-analyses. As expected, considering the limited number of diabetic patients in the present study, only three SNPs reached nominally significant *p* values (≤0.01) for association with diabetes in this population (*IGF2BP2*, *CDKAL1* and *TCF7L2*).

These SNPs were used to construct a weighted genetic score. This score was normally distributed among the non-diabetic individuals in the CoLaus population and was slightly skewed to the right among diabetic individuals

Table 2 Multivariate logistic regression analysis of the risk of prevalent diabetes in the CoLaus Study

Characteristic	Without genetic score			With genetic score		
	OR	95% CI	*p* value	OR	95% CI	*p* value
Age (per 1 year)	1.08	1.06–1.10	1.0×10^{-25}	1.08	1.06–1.09	9.5×10^{-27}
BMI (per 1 kg/m^2)	1.13	1.10–1.17	1.3×10^{-18}	1.13	1.10–1.16	6.9×10^{-19}
Family history of diabetes	2.94	2.27–3.80	2.5×10^{-16}	2.92	2.25–3.77	7.3×10^{-16}
WHR (per 1 SD[a])	1.77	1.47–2.13	1.3×10^{-9}	1.78	1.48–2.14	1.5×10^{-9}
Triacylglycerol/HDL-cholesterol ratio (per 1 SD[b])	1.24	1.14–1.35	3.3×10^{-7}	1.25	1.15–1.36	1.6×10^{-7}
Physical activity	0.64	0.49–0.83	0.0006	0.63	0.49–0.82	0.0004
Weighted genetic score (per unit)	NA	NA	NA	1.15	1.10–1.20	2.9×10^{-9}
Unweighted genetic score (per 1 risk allele)[c]	NA	NA	NA	1.13	1.08–1.19	1.4×10^{-6}

[a] SD=0.08, [b] SD=1.18, as calculated for the entire population
[c] The results for the unweighted genetic score (per allele) were obtained by replacing the weighted genetic score in the model
NA, not applicable

(Fig. 1a), with a correspondingly higher mean score in the latter group (15.2±2.9 vs 14.3±2.7, $p<0.001$, Table 1). After adjustment for the clinical variables independently associated with the disease, the risk of prevalent diabetes rose in proportion to the weighted genetic score, with the 20% of the population with a score within the top quintile having a 2.7 (95% CI 1.8–4.0, $p=0.000006$) higher risk than those within the bottom quintile (Fig. 1b). Figure 2a shows the distribution of the unweighted genetic score, as generated by risk allele counting for the 15 SNPs in these two groups. The relationship between the unweighted genetic score and the weighted genetic score is shown in Fig. 2b. Overall, a direct relationship was observed between the two scores; however, a wide spread was observed in the weighted genetic score for each category of cumulative risk alleles.

We next examined the relationship between BMI, the weighted genetic score and the prevalence of diabetes in the CoLaus sample (Fig. 3). As expected, the disease prevalence increased in proportion to quintiles of BMI. In addition, within each BMI quintile, the prevalence of the

Table 3 Genes and SNPs selected from published meta-analyses to generate the 15 SNP-based weighted score

SNP	CHR	Nearest gene	Non-risk	Risk	Reference			CoLaus Study		
					Risk allele frequency	OR	95% CI[b]	Risk allele frequency	OR	95% CI
rs10923931	1	*NOTCH2*	G	T	0.11	1.13	1.08–1.17	0.10	1.06	0.81–1.39
rs7578597	2	*THADA*	C	T	0.90	1.15	1.10–1.20	0.90	1.31	0.98–1.74
rs1801282[a]	3	*PPARG*	G	C	0.87	1.14	1.08–1.20	0.88	1.25	0.96–1.63
rs4607103	3	*ADAMTS9*	T	C	0.76	1.09	1.06–1.12	0.72	0.84	0.71–1.01
rs4402960[a]	3	*IGF2BP2*	G	T	0.32	1.14	1.11–1.18	0.31	1.29	1.09–1.53
rs10010131[a]	4	*WFS1*	A	G	0.60	1.11	1.08–1.16	0.61	1.05	0.89–1.23
rs10946398[a]	6	*CDKAL1*	A	C	0.32	1.14	1.11–1.17	0.32	1.35	1.14–1.59
rs864745	7	*JAZF1*	C	T	0.50	1.10	1.07–1.13	0.50	1.07	0.91–1.25
rs10811661[a]	9	*CDKN2A–CDKN2B*	C	T	0.83	1.20	1.14–1.25	0.80	0.91	0.75–1.12
rs12779790	10	*CDC123–CAMK1D*	A	G	0.18	1.11	1.07–1.14	0.18	1.12	0.91–1.37
rs1111875[a]	10	*HHEX–IDE*	T	C	0.65	1.15	1.10–1.19	0.59	1.02	0.87–1.20
rs7901695[a]	10	*TCF7L2*	T	C	0.31	1.37	1.31–1.43	0.34	1.59	1.35–1.87
rs5215[a]	11	*KCNJ11*	T	C	0.35	1.14	1.10–1.19	0.37	1.05	0.88–1.24
rs7961581	12	*TSPAN–LGR5*	T	C	0.27	1.09	1.06–1.12	0.30	1.07	0.90–1.28
rs8050136[a]	16	*FTO*	C	A	0.40	1.17	1.12–1.22	0.41	1.12	0.95–1.32

[a] SNPs selected from the review by Frayling [9]; other SNPs were chosen from Zeggini et al. [8]
[b] The lower boundary of the 95% CI was logarithmically transformed and used as weighting factor for each SNP
CHR, chromosome number

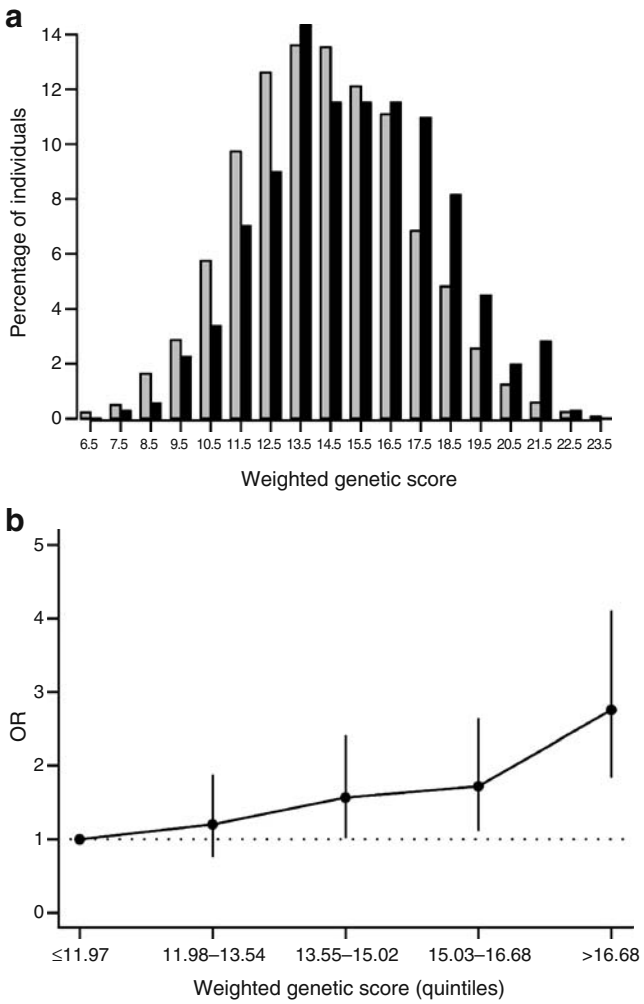

Fig. 1 Distribution of the weighted genetic score among diabetic and non-diabetic individuals and risk of prevalent diabetes in the population. **a** Distribution of the 15 SNP-based weighted genetic score among the 5,004 non-diabetic (grey bars) and 356 diabetic (black bars) participants of the CoLaus Study. This score was calculated for each individual by weighting each risk allele with its reported effect, as described in the Methods. **b** Risk of prevalent diabetes (expressed as OR±95% CI by quintiles) after adjustment for clinical covariates including age, sex, WHR, physical activity, triacylglycerol/HDL-cholesterol ratio and family history of diabetes. The boundaries for each genetic score quintile are shown on the *x*-axis

disease increased for each quintile of the genetic score. The disease prevalence in individuals in the top quintiles for both BMI and the genetic score was 24.7% (n=220), compared with 1.4% among individuals with a genetic score within the bottom quintiles for both variables (n=220). The effect of the weighted genetic score appeared particularly pronounced among the top three quintiles of the distribution. However, the interaction between the weighted genetic score (by quintiles) and BMI (by quintiles) was not significant (p=0.18). No significant differences in BMI were observed between genetic score quintiles. BMI averaged 25.6±4.5 (SD) kg/m^2 in the bottom quintile (n=1,071) and 25.9±4.7 kg/m^2 in the top quintile (n=1,071, p=0.10).

When the weighted genetic score was included in the multivariate logistic regression analysis, this score was significantly and independently associated with the risk of prevalent diabetes (Table 2). The fact that, in this analysis, the genetic score and family history of diabetes were both independent predictors of diabetes suggested that these variables each added additional information to the risk prediction, i.e. that the genetic score did not capture the entirety of the information contained in family history. The

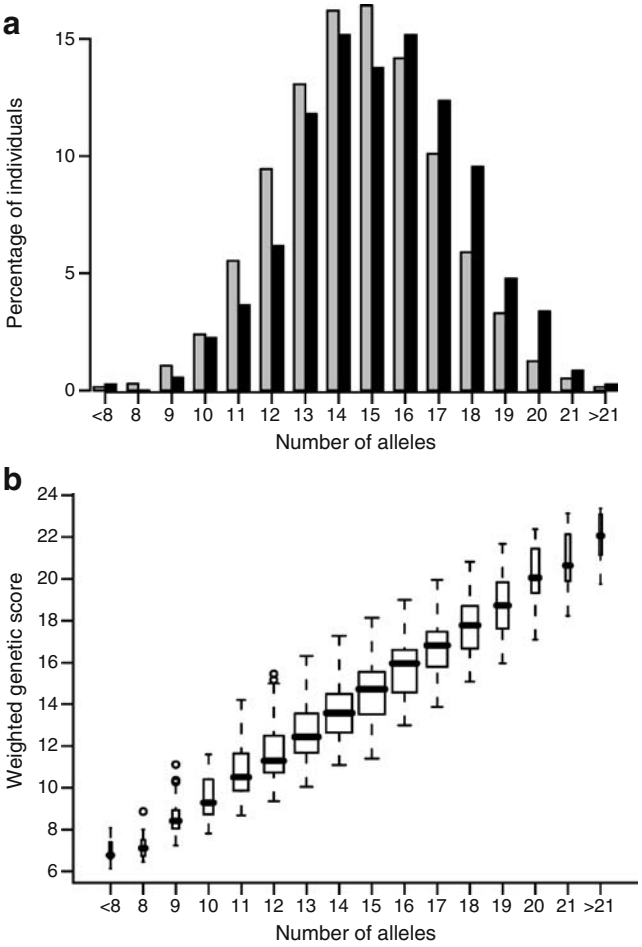

Fig. 2 Distribution of the unweighted genetic score and relationship between weighted and unweighted genetic score. **a** Distribution of the unweighted genetic score (corresponding to the cumulative number of risk alleles) among 356 diabetic (black bars) and 5,004 non-diabetic (grey bars) participants in the CoLaus Study. **b** Box plot of weighted genetic score by the number of risk alleles (unweighted genetic score). The horizontal line within each box represents the median value of weighted genetic score for a particular number of alleles, with upper and lower limits of the box corresponding to 75th (Q3) and 25th (Q1) percentile of the distribution, respectively. Horizontal lines at the top and bottom of the dashed lines correspond to 1.5×interquartile range (IQR; Q3–Q1) from each end of the box. Individuals with weighted genetic score 1.5×IQR higher than Q3 or 1.5×IQR lower than Q1 are represented by small circles. The width of the box is proportional to the square root of the number of observations

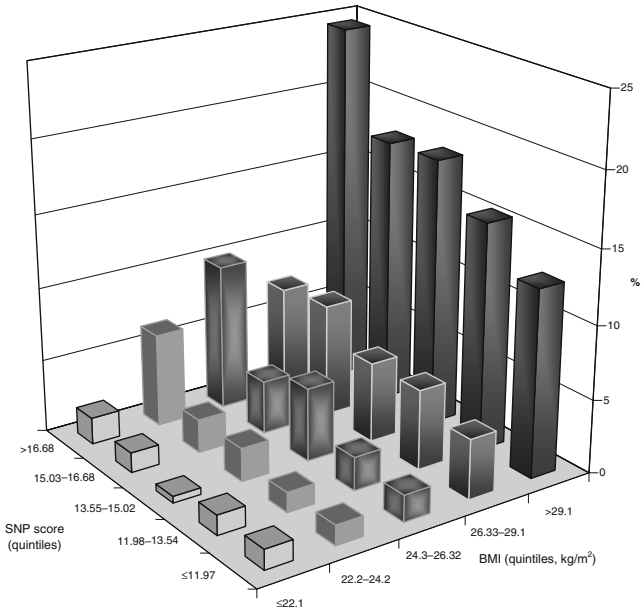

Fig. 3 Prevalence of type 2 diabetes in the CoLaus Study according to BMI and the weighted genetic score. The 5,360 participants of the Lausanne Study were partitioned into 25 groups, each containing an average of 220 individuals, by quintiles of BMI and quintiles of the weighted genetic score

calibration statistic (Hosmer–Lemeshow χ^2 statistic) for the model with the clinical predictors was 5.55 (p=0.70) and adding the weighted score yielded 12.55 (p=0.13), which indicated that both models represented a good fit.

To explore further the discriminatory power of the weighted genetic score, we performed a ROC curve analysis (Fig. 4). The values for the area under the ROC curve for the weighted and unweighted genetic scores only were 59% and 57%, respectively (p=0.008 for comparing the area under ROC curves for the weighted and unweighted scores), whereas the value was 86% for the clinical covariates listed above only. Adding the weighted genetic score to the clinical covariates led to a limited yet significant improvement in the area under the ROC curve to 87% (p=0.002). In contrast, adding the unweighted genetic score to the clinical covariates did not lead to a statistically significant improvement in the area under the ROC curve (p=0.07, compared with the area under the ROC curve of the model with the covariates only). To further characterise whether or not the weighted genetic score improved the prediction of prevalent diabetes, we also performed an IDI analysis. In this analysis, adding the weighted genetic score to the covariates led to a statistically significant improvement in the prediction (IDI=1.2%, p=0.0003). This indicated that by adding the weighted genetic score the improvement of average sensitivity offset by the potential increase in average 'one minus specificity' was about 1.2%. The IDI for adding the unweighted genetic score led to a smaller and less significant improvement (IDI=0.7%, p=0.002) than the weighted score. This analysis confirmed the ROC curve analysis results and reinforced the concept that the weighted genetic score added some predictive ability to the clinical covariates.

Discussion

In this study, we show that, at a population level, accumulation of several susceptibility genes for diabetes is accompanied by a substantial increase in the risk of having the disease. This was particularly apparent, in terms of prevalence, among obese individuals. We also show that the weighted genetic score added some information that was not captured by clinical variables, including family history of diabetes. The present data also show that weighting the genetic score with the reported effect of risk alleles provided more predictive value than an unweighted genetic score generated by counting the number of risk alleles. The clinical usefulness of the score, however, remains to be demonstrated.

The present population-based cross-sectional study is in line with two very recently published prospective studies [19, 20]. In both of these studies, a high unweighted genetic score was associated with a marked increase in the

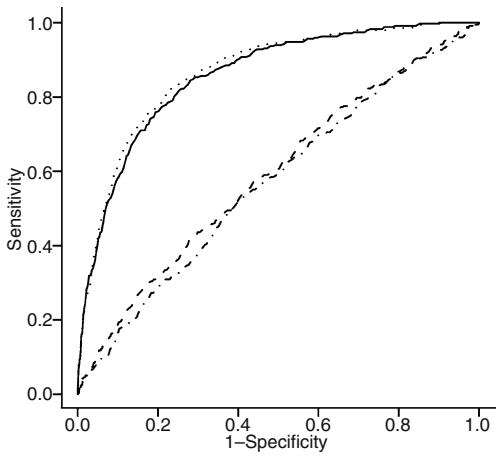

Fig. 4 ROC curve analysis of the discriminatory power of the genetic score in predicting the presence of diabetes in the CoLaus Study. A ROC curve analysis was performed using the 15 SNP-based weighted score only (dashed line, area under ROC curve=59%), the 15 SNP-based unweighted genetic score only (dotted-dashed line, area under ROC curve=57%, p=0.008 vs weighted genetic score), the clinical covariates only (solid line, area under ROC curve=86%) and the combination of the weighted genetic score and the clinical covariates (dotted line, area under ROC curve=87%, p=0.002 vs covariates only) to predict the presence of diabetes in the CoLaus Study. Clinical covariates included age, sex, family history of diabetes, physical activity, triacylglycerol/HDL-cholesterol ratio and WHR. The ROC curves were based on cross-validation prediction for the multivariate logistic regression models

incidence of diabetes. However, the predictive value of this score beyond clinical variables was modest.

The best way to construct a genetic score for individuals who carry risk alleles for several susceptibility genes is not yet known. The present analysis provides evidence that, at least in the CoLaus dataset, weighting the score with the individual effect of the genes provides more predictive information for the presence of prevalent diabetes than simply counting the number of risk alleles. Considering the fact that most susceptibility genes for type 2 diabetes have a relatively similar effect, one may anticipate that weighting the genetic score may be even more advantageous for diseases where risk alleles have a much more divergent effect, like type 1 diabetes [26].

BMI was similar between quintiles of genetic score. This observation, which is in line with the data reported from the combined prospective Malmö and Botnia studies [20], suggests that the cumulative genetic susceptibility to diabetes, as mediated by the 15 genes under investigation, could be accounted for by a dysfunction of the insulin-secreting beta cell or other unknown mechanisms, rather than by merely a propensity to obesity. This observation is supported by the data reported by Lango et al. [18], in which, among diabetic cases, individuals with a high genetic score had an earlier age of diagnosis, and had 1.6 kg/m^2 less in BMI than those with a low genetic score. In contrast, we observed that the genetic score and increasing BMI together could identify individuals at high risk of diabetes, an effect that was also observed in Lyssenko et al. [20]. This suggests that even though overall the genetic score alone does not add a great deal to our ability to predict who will develop diabetes, in combination with other risk factors there may be usefulness in identifying high-risk individuals who could benefit from early preventative interventions.

This study has some limitations. First, it is cross-sectional, allowing the predictive potential to be inferred from the clinical state only at the time of observation. The number of diabetic patients in the present study was relatively modest, and documentation of the value of the score in larger population-based collections is warranted. Another limitation of this study is that two SNPs/genes were not included in the present analysis because of technical reasons. If anything, we expect that incorporation of these additional SNPs in the weighted genetic score should improve its predictive value. In addition, the weighting scheme we used to construct the genetic score is not necessarily an optimal solution, but an attempt to account for the variability in allelic contribution and estimate variation. Finally, the CoLaus Study includes only whites, so extrapolation to other ethnic groups should be considered with great caution.

One may anticipate that future genetic scores will have a better capacity to predict diabetes. First, the genes under investigation only account for part of the genetic susceptibility to diabetes [18]. Additional diabetes-susceptibility genes are likely to be discovered when larger collections are assembled for meta-analyses or when technological advances increase SNP coverage in areas of the genome that are currently poorly represented. Incorporation of these new genes into the genetic score may make it more predictive, although the gain could be relatively modest unless a large number of additional new genes are identified [17]. Most of the SNPs selected here are not causal, and are presumably in linkage disequilibrium with causative SNP. Once causal SNPs will be identified, the score may further improve. Similarly, it is conceivable that incorporation of several SNPs within each gene, by capturing more variability within these genes, could also ameliorate the predictive value of the score. Finally, it is conceivable that larger studies will improve the risk prediction associated with risk alleles, so that the OR (rather than its lower 95% CI boundary) could be used in the risk assessment.

Despite these limitations, this study shows that, at a population level, accumulation of risk alleles within multiple susceptibility genes for diabetes increases significantly the disease risk and that a simple weighted genetic score has the ability to predict the presence of the disease beyond the clinical predictors. At this stage, however, the weighted genetic score does seem to have limited ability to predict the presence of diabetes, and additional studies are required to demonstrate its clinical usefulness.

Acknowledgements We thank the individuals who volunteered to participate in the CoLaus Study, as well as the staff at Lausanne University Hospital and within GlaxoSmithKline who made this collection possible. Genotype imputation computation was performed at the Vital-IT Center (http://www.vital-it.ch) for high-performance computing of the Swiss Institute of Bioinformatics. We are grateful to N. Wareham (MRC, Cambridge, UK) for helpful discussions.

Duality of interest The CoLaus Study was sponsored in part by GlaxoSmithKline. X. Lin, K. Song, N. Lim, X. Yuan, H. Stirnadel, S. S. Sundseth, E. Lai, D. K. Burns, L. T. Middleton, A. D. Roses, P. M. Matthews, L. Cardon, D. M. Waterworth and V. Mooser are, or were (L. T. Middleton and A. D. Roses), full-time employees of GlaxoSmithKline. The CoLaus Study was sponsored in part by GlaxoSmithKline and by a grant from the Swiss National Foundation (to G. Waeber). The CoLaus Study was designed jointly by scientists within GlaxoSmithKline and the Investigators at Lausanne CHUV University Hospital in Switzerland. Scientists from GlaxoSmithKline and Lausanne University Hospital jointly analysed and interpreted the results of the present study and agreed to submit this manuscript to *Diabetologia*.

References

1. Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB Sr (2007) Prediction of incident diabetes mellitus

in middle-aged adults: the Framingham Offspring Study. Arch Intern Med 167:1068–1074
2. Simmons RK, Harding AH, Wareham NJ, Griffin SJ (2007) Do simple questions about diet and physical activity help to identify those at risk of type 2 diabetes? Diabet Med 24:830–835
3. Katzmarzyk PT, Craig CL, Gauvin L (2007) Adiposity, physical fitness and incident diabetes: the physical activity longitudinal study. Diabetologia 50:538–544
4. Macchia A, Levantesi G, Borrelli G et al (2006) A clinically practicable diagnostic score for metabolic syndrome improves its predictivity of diabetes mellitus: the Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto miocardico (GISSI)—Prevenzione scoring. Am Heart J 151:754.e7–754.e17
5. Rahman M, Simmons RK, Harding AH, Wareham NJ, Griffin SJ (2008) A simple risk score identifies individuals at high risk of developing type 2 diabetes: a prospective cohort study. Fam Pract 25:191–196
6. 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
7. 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
8. 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
9. Frayling TM (2007) Genome-wide association studies provide new insights into type 2 diabetes aetiology. Nat Rev Genet 8:657–662
10. Saxena R, Voight BF, Lyssenko V et al (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316:1331–1336
11. Bottinger EP (2007) Foundations, promises and uncertainties of personalized medicine. Mt Sinai J Med 74:15–21
12. Scheuner MT, Sieverding P, Shekelle PG (2008) Delivery of genomic medicine for common chronic adult diseases: a systematic review. JAMA 299:1320–1334
13. Couzin J (2008) Genetic risk. With new disease genes, a bounty of questions. Science 319:1754–1755
14. Lyssenko V, Almgren P, Anevski D et al (2005) Genetic prediction of future type 2 diabetes. PLoS Med 2:e345
15. Weedon MN, McCarthy MI, Hitman G et al (2006) Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. PLoS Med 3:e374
16. Vaxillaire M, Veslot J, Dina C et al (2008) Impact of common type 2 diabetes risk polymorphisms in the DESIR prospective study. Diabetes 57:244–254
17. Lu Q, Elston RC (2008) Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. Am J Hum Genet 82:641–651
18. Lango H, Palmer CN, Morris AD et al (2008) Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. Diabetes 57:3129–3135
19. Meigs JB, Shrader P, Sullivan LM et al (2008) Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 359:2208–2219
20. Lyssenko V, Jonsson A, Almgren P et al (2008) Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med 359:2220–2232
21. Firmann M, Mayor V, Marques VP et al (2008) The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC Cardiovasc Disord 8:6
22. Sandhu MS, Waterworth DM, Debenham SL et al (2008) LDL-cholesterol concentrations: a genome-wide association study. Lancet 371:483–491
23. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27:157–172
24. D'Agostino RB Sr, Grundy S, Sullivan LM, Wilson P (2001) Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. JAMA 286:180–187
25. Delong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845
26. Todd JA, Walker NM, Cooper JD et al (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39:857–864