



Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, *et al.*

Science **316**, 1331 (2007);

DOI: 10.1126/science.1142358

The following resources related to this article are available online at www.sciencemag.org (this information is current as of June 12, 2007):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/316/5829/1331>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/1142358/DC1>

This article **cites 29 articles**, 7 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/316/5829/1331#otherarticles>

This article has been **cited by 2 articles** hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/316/5829/1331#otherarticles>

This article appears in the following **subject collections**:

Genetics

<http://www.sciencemag.org/cgi/collection/genetics>

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/about/permissions.dtl>

African ape clade, it is the panins and gorillines that innovated and the hominins that were conservative.

References and Notes

- B. G. Richmond, D. S. Strait, *Nature* **404**, 382 (2000).
- J. Kingdon, *Lowly Origins* (Princeton Univ. Press, Princeton, NJ, 2003).
- C. V. Ward, M. G. Leakey, A. Walker, *Evol. Anthropol.* **7**, 197 (1999).
- Y. Haile-Selassie, *Nature* **412**, 178 (2001).
- T. D. White *et al.*, *Nature* **440**, 883 (2006).
- K. Kovarovic, P. Andrews, *J. Hum. Evol.*, in press (available at <http://dx.doi.org/doi:10.1016/j.jhevol.2007.01.001>; doi: 10.1016/j.jhevol.2007.01.001).
- N. Patterson, D. J. Richter, S. Gnerre, E. S. Lander, D. Reich, *Nature* **441**, 1103 (2006).
- K. D. Hunt *et al.*, *Primates* **37**, 363 (1996).
- J. G. Fleagle *et al.*, *Symp. Zool. Soc. London* **48**, 359 (1981).
- R. H. Crompton *et al.*, *Cour. Forsch.-Inst. Senckenb.* **243**, 115 (2003).
- J. T. Stern, *Yrb. Phys. Anthropol.* **19**, 59 (1975).
- S. K. S. Thorpe, R. H. Crompton, *Am. J. Phys. Anthropol.* **131**, 384 (2006).
- K. D. Hunt, *J. Hum. Evol.* **26**, 183 (1994).
- M. Cartmill, K. Milton, *Am. J. Phys. Anthropol.* **47**, 249 (1977).
- E. Larney, S. Larsen, *Am. J. Phys. Anthropol.* **125**, 42 (2004).
- S. K. S. Thorpe, R. H. Crompton, *Am. J. Phys. Anthropol.* **127**, 58 (2005).
- S. K. S. Thorpe, R. H. Crompton, M. M. Gunther, R. F. Ker, R. McN. Alexander, *Am. J. Phys. Anthropol.* **110**, 179 (1999).
- R. McN. Alexander, *Principles of Animal Locomotion* (Princeton Univ. Press, Princeton, NJ, 2003).
- C. V. Ward, *Yrbk. Phys. Anthropol.* **45**, 185 (2002).
- R. W. Wrangham, N. L. Conklin-Brittain, K. D. Hunt, *Int. J. Primatol.* **19**, 949 (1998).
- H. Pontzer, R. W. Wrangham, *J. Hum. Evol.* **46**, 317 (2004).
- R. C. Payne *et al.*, *J. Anat.* **208**, 709 (2006).
- M. Pickford, B. Senut, B. Gommery, in *Late Cenozoic Environments and Hominid Evolution: a Tribute to Bill Bishop*, P. Andrews, P. Banham, Eds. (Geological Society, London, 1999), pp. 27–38.
- N. M. Young, L. MacLachy, *J. Hum. Evol.* **46**, 163 (2004).
- D. Gommery, B. Senu, M. Pickford, E. Musiime, *Ann. Paléontol.* **88**, 167 (2002).
- C. V. Ward, in *Handbook of Paleoanthropology Vol. 2: Primate Evolution and Human Origins*, W. Henke, I. Tattersall, Eds. (Springer, Heidelberg, Germany, 2007), pp. 1011–1030.
- B. Senut, in *Human Origins and Environmental Backgrounds*, H. Ishida, R. H. Tuttle, M. Pickford, N. Ogihara, M. Nakatsukasa, Eds. (Springer, Heidelberg, Germany, 2006), pp. 199–208.
- C. P. E. Zollikofer *et al.*, *Nature* **434**, 755 (2005).
- M. Pickford, *Anthropologie* **69**, 191 (2005).
- We thank the Indonesian Institute of Science, Indonesian Nature Conservation Service, and Leuser Development Programme for granting permission and giving support for research in the Leuser Ecosystem. R. McN. Alexander, T. M. Blackburn, S. Burtles, J. Rees, N. Jeffery, E. E. Verecke, A. Walker, A. Wilson, and B. Wood commented on the manuscript. R. Savage developed the animation (fig. S1). Studies of captive animals were hosted by the North of England Zoological Society. This research was supported by grants from the Leverhulme Trust, the Royal Society, the L.S.B. Leakey Foundation, and the Natural Environment Research Council.

Supporting Online Material

www.sciencemag.org/cgi/content/full/316/5829/1328/DC1

Table S1

Movies S1 to S3

5 February 2007; accepted 18 April 2007

10.1126/science.1140799

Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes for BioMedical Research*†

New strategies for prevention and treatment of type 2 diabetes (T2D) require improved insight into disease etiology. We analyzed 386,731 common single-nucleotide polymorphisms (SNPs) in 1464 patients with T2D and 1467 matched controls, each characterized for measures of glucose metabolism, lipids, obesity, and blood pressure. With collaborators (FUSION and WTCCC/UKT2D), we identified and confirmed three loci associated with T2D—in a noncoding region near *CDKN2A* and *CDKN2B*, in an intron of *IGF2BP2*, and an intron of *CDKAL1*—and replicated associations near *HHEX* and in *SLC30A8* found by a recent whole-genome association study. We identified and confirmed association of a SNP in an intron of glucokinase regulatory protein (GCKR) with serum triglycerides. The discovery of associated variants in unsuspected genes and outside coding regions illustrates the ability of genome-wide association studies to provide potentially important clues to the pathogenesis of common diseases.

Type 2 diabetes, obesity, and cardiovascular risk factors are caused by a combination of genetic susceptibility, environment, behavior, and chance. Whole-genome association studies (WGAS) offer a new approach to gene discovery unbiased with regard to presumed functions or locations of causal variants. This approach is based on Fisher's theory for additive effects at common alleles (1); human heterozygosity being substantially attributable to common ancestral variants (2); and the hypothesis that variants influencing common, late-onset diseases of modernity may not have been subject

to purifying selection, and has been made possible by genomic advances such as the human genome sequence, SNP and HapMap databases, and genotyping arrays (3).

We studied 1464 patients with T2D and 1467 controls from Finland and Sweden, each characterized for 18 clinical traits: anthropometric measures, glucose tolerance and insulin secretion, lipids and apolipoproteins, and blood pressure. The samples were both population-based [1022 T2D cases and 1075 euglycemic controls, matched on gender, age, body mass index, and region of origin] and family-based (326 sibships discordant for T2D; 442 cases and 392 euglycemic controls; tables S1 and S2).

Genotyping of 500,568 SNPs was attempted in each sample. Overall call rate for passing SNPs was 99.2%. After filtering rare and monomorphic variants ($n = 69,696$ SNPs) and

applying stringent quality-control filters, high-quality genotypes for 386,731 common SNPs were obtained (4). To extend the set of putative causal alleles tested for association, we developed 284,968 additional multimarker (haplotype) tests based on these SNP genotypes (5, 6). The 671,699 allelic tests capture (correlation coefficient $r^2 \geq 0.8$) 78% of common SNPs in HapMap CEU (3).

Each SNP and haplotype test was assessed for association to T2D and each of 18 traits with the software package PLINK (<http://pngu.mgh.harvard.edu/purcell/plink/>). For T2D, a weighted meta-analysis was used to combine results for the population-based and family-based subsamples (4). For quantitative traits, multivariable linear or logistic regression with or without covariates was performed (4). Association results for each SNP, haplotype test, and phenotype are available (www.broad.mit.edu/diabetes/).

In genome-wide analysis involving hundreds of thousands of statistical tests, modest levels of bias imposed on the null distribution can overwhelm a small number of true results. We used three strategies to search for evidence of systematic bias from unrecognized population structure, the analytical approach, and genotyping artifacts (7, 8). First, we examined the distribution of P -values in the population-based sample, observing a close match to that expected for the null distribution (genomic inflation factor $\lambda_{GC} = 1.05$ for T2D). Second, we calculated association statistics using EIGENSTRAT, an independent method based on principal components analysis (9); P -values for T2D derived with the two methods were nearly identical ($r^2 = 0.95$, Fig. 1A). Third, 114 SNPs from the extreme tail of P -values for T2D were genotyped with an independent technology. Genotype concordance was 99.5%, indicating that even the extreme tail of low P -values is not substantially contaminated by genotyping artifacts.

*To whom correspondence should be addressed: David Altshuler, Leif Groop, Thomas E. Hughes. E-mail: altshuler@molbio.mgh.harvard.edu (D.A.); leif.groop@med.lu.se (L.G.); thomase.hughes@novartis.com (T.E.H.)

†All authors with their contributions and affiliations appear at the end of this paper.

Although the observed P -value distribution closely matches expectation over most of its range ($1.0 > P > 0.01$), an excess of low P -values is observed (Fig. 1B and table S3). To evaluate the significance of this excess, we generated 1000 permuted whole-genome analyses in which phenotype data were randomized within matched case-control groups (4). For P -values between 0.01 and 0.001, 6370 SNPs were observed, as compared to an average of 5917 [95% confidence interval (CI): 5714 to 6128] in permuted scans ($Z = 4.3$; $P < 0.001$). For $P < 10^{-4}$, 125 SNPs were observed, compared to 94 [95% CI: 61 to 121] in permuted scans ($Z = 2.4$, $P < 0.02$). These observations support a model in which there are few common variants with large effects, but a substantial number with modest effects of the sort that generate P -values between 0.01 and 10^{-7} in 3000 samples.

Given this distribution, and because WGAS are hypothesis generating, we sought replication in independent samples. An initial set of 107 SNPs (table S4) was selected on the basis of our study ($n = 89$) and by comparison of our results with WGAS of T2D ($n = 18$) by Wellcome Trust Case Control Consortium (WTCCC) (10) and Finland–United States Investigation of NIDDM Genetics (FUSION) (11). Each SNP was genotyped in 10,850 additional subjects (T2D and controls) from Sweden, Poland, and the United States (table S1) and analyzed for association to T2D under the same genetic model as the scan (4).

These results, with those from FUSION (11) and WTCCC/UKT2D (10, 12), identify SNPs at three previously unknown loci as influencing risk of T2D with $P < 10^{-10}$ (Table 1 and tables S4 and S5).

A SNP on chromosome 9p (rs10811661), 125 kb from the nearest annotated genes (*CDKN2A/CDKN2B*), was selected on the basis of strong association to T2D in our WGAS

(rank #51) (Table 1; Fig. 2A). Combined analysis of data from our scan and replication samples provides strong evidence for association: odds ratio (OR) = 1.20, 95% CI 1.07 to 1.36, $P = 5.4 \times 10^{-8}$. Independent evidence of association for the same SNP, phenotype, and genetic model was obtained by WTCCC/UKT2D ($P = 10^{-7}$) and FUSION ($P = 0.001$) (10–12). No association with measured quantitative metabolic traits was observed in our scan or replication samples.

An intriguing aspect of this association is its location far from any annotated gene. The region of association is limited to a 9-kb region flanked by strong recombination hot-spots, in which there are multiple conserved noncoding sequences but no known genes or microRNAs. A member of the nearest gene cluster, cyclin-dependent kinase inhibitor-2A (*CDKN2A*), plays a role in pancreatic islet regenerative capacity (13).

SNPs in the second intron of *IGF2BP2* were selected for replication on the basis of joint analysis of the three scans (10–12) (Fig. 2B). Evidence was weak in our initial scan ($P = 0.034$ for rs4402960), but pronounced in the replication samples ($P = 5.5 \times 10^{-9}$, Table 1). Strong evidence was obtained for the same SNP, phenotype, and genetic model by WTCCC/UKT2D ($P = 10^{-4}$) and FUSION ($P = 10^{-4}$) (10–12). These SNPs showed no association to measured quantitative metabolic traits in our scan or replication samples.

Insulin-like growth factor 2 binding protein 2 (*IGF2BP2*) belongs to a family of three mRNA binding proteins with affinity for leader elements in the untranslated regions of *IGF-2* transcripts. Family members bind with weak sequence specificity and are implicated in transport of RNA targets to enable protein synthesis at specific locations in the cell (14). The IGFBP homolog is necessary for pancreas development in *Xeno-*

pus (15), and *IGF2BP3* transgenic mice exhibit acinar-ductal pancreatic metaplasia (16).

We selected a SNP in a 90-kb intron within *CDKAL1* (rs7754840) for replication on the basis of nominal association in our scan, WTCCC (10), and FUSION (11) (Table 1). Analysis of the scan and replication samples (Fig. 2C) supports association under the same phenotype and genetic model (OR = 1.08, 95% CI 1.03 to 1.14, combined $P = 0.0024$; Table 2), as does evidence from WTCCC/UKT2D (10, 12) ($P = 10^{-8}$) and FUSION (11) ($P = 0.01$) (Table 1). The risk allele was nominally associated with reduced insulin secretion in controls from our scan ($P = 0.01$ for insulinogenic index).

CDKAL1 is homologous to CDK5RAP1, an inhibitor of cyclin-dependent kinase CDK5; CDK5 transduces glucotoxicity signals in pancreatic beta cells (17). As with the other variants, how SNPs in CDKAL1 might influence risk of T2D awaits further investigation.

Common variation in an intron of *TCF7L2* has been reproducibly associated with T2D (18). In our WGAS, *TCF7L2* was the third-ranked association (Fig. 2D, $P < 3 \times 10^{-6}$) and was among the top results in each of the three other well-powered whole-genome scans of T2D (10, 11, 19) (Table 1). The consistency of these findings suggests that *TCF7L2* is the single largest effect of a common SNP on T2D risk in European populations. Associations in *KCNJ11* (20) and *PPARG* (21) were not strongly observed in any single scan, but across the three scans provided $P < 10^{-10}$ and $P < 10^{-6}$, respectively (Table 1).

In 2007, Sladek *et al.* (19) reported four previously unknown associations to T2D in a WGAS, two with particularly strong evidence of replication (*HHEX* and *SLC30A8*). We confirm association at *HHEX* (Table 1; Fig. 2E) in our scan (OR = 1.15, $P = 0.01$) and in replication genotyping ($P = 10^{-3}$), as do WTCCC/UKT2D ($P = 10^{-6}$) (10, 12) and FUSION ($P = 0.03$) (11). At the zinc transporter *SLC30A8*, our data were less compelling ($P = 0.90$ in our scan and $P = 0.01$ in replication samples), but convincing evidence was obtained by WTCCC/UKT2D (10, 12) ($P = 10^{-3}$) and FUSION ($P = 10^{-5}$) (11). We observed no evidence for association at *LOC387761* ($n = 7401$, OR = 1.00, $P = 0.93$ for rs7480010) and *EXT2-ALX4* ($n = 7401$, OR = 1.06, $P = 0.12$ for rs3740878), nor was evidence obtained by WTCCC (10) or FUSION (11).

We observed intriguing replication signals at additional loci (10, 11). For example, rs17044137 in *FLJ393370* was associated with T2D in our scan (OR = 1.27, $P = 3.7 \times 10^{-4}$) and replication (OR = 1.09, $P = 3.1 \times 10^{-3}$), but not in WTCCC (10) or FUSION (11) (Table 1). Similarly, rs6698181 in *PKN2* demonstrated evidence in our scan and replication samples ($P = 5.3 \times 10^{-5}$), and in FUSION (11) ($P = 10^{-3}$), but not WTCCC (10) ($P = 0.93$). Genotyping in more samples is needed to resolve these and other hypotheses.

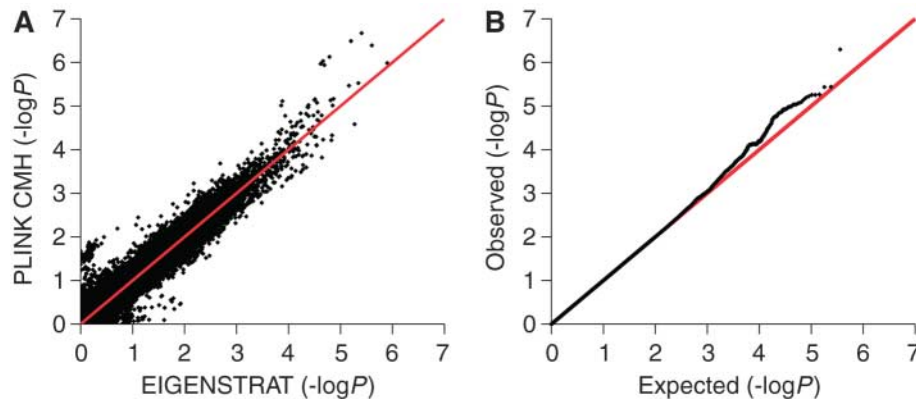


Fig. 1. P -value distribution for the association with type 2 diabetes. (A) P -values obtained from the Cochran-Mantel-Haenszel stratified test implemented in PLINK are plotted (as $-\log_{10}$ values) as a function of the corresponding P -value computed by EIGENSTRAT in the population-based case/control sample ($n = 2097$). These distributions are strongly correlated ($r^2 = 0.95$). (B) P - P plot for the combined (Z score) association analysis of type 2 diabetes in the population-based case/control sample and the discordant sibships ($n = 2,931$). The P -values for the corresponding Z scores are plotted (as $-\log_{10}$ values) as a function of P -values from the expected (uniform) null distribution. The observed distribution matches the expected distribution closely and shows an excess in the tail at $P < 10^{-3}$.

Table 1. Association results for type 2 diabetes. Odds ratios (OR), 95% confidence intervals (CI), and *P*-values are given for SNPs from Diabetes Genetics Initiative (DGI) scan, replication samples, and data from the WTCCC/UKT2D (10,12) and FUSION (11) studies. Proxies include: for rs7754840, rs10946398 ($r^2 = 1$, WTCCC/UKT2D); for rs7903146, rs7901695 ($r^2 = 0.92$, WTCCC/UKT2D); and for rs5219, rs5215 ($r^2 = 0.99$, WTCCC/UKT2D).

SNP	Chr.	Risk allele*	Nearest gene	DGI					WTCCC/UKT2D		FUSION		All studies		Sample size for 80% power [¶]		
				Genome scan		Replication [‡]		Combined		OR	<i>P</i> -value	OR	<i>P</i> -value	OR (95%CI)		<i>P</i> -value	
				Freq. [†]	OR	<i>P</i> -value	OR	<i>P</i> -value [§]	OR	<i>P</i> -value	OR	<i>P</i> -value	OR (95%CI)	<i>P</i> -value			
				<i>n</i> = 2931	<i>n</i> = 10,850			<i>n</i> = 13,781	<i>n</i> ≤ 13,965		<i>n</i> ≤ 4808		<i>n</i> ≤ 32,554				
Novel loci																	
rs10811661	9	T	<i>CDKN2B</i>	0.83	1.37	3.6×10 ⁻⁵	1.16	2.2×10 ⁻⁵	1.20	5.4×10 ⁻⁸	1.19	4.9×10 ⁻⁷	1.20	2.2×10 ⁻³	1.20	7.8×10 ⁻¹⁵	~3400
					(1.18–1.59)		(1.09–1.27)		(1.12–1.28)		(1.11–1.28)		(1.07–1.36)		(1.14–1.25)		
rs4402960	3	T	<i>IGF2BP2</i>	0.29	1.14	0.034	1.18	5.5×10 ⁻⁹	1.17	1.7×10 ⁻⁹	1.11	1.6×10 ⁻⁴	1.18	2.4×10 ⁻⁴	1.14	8.9×10 ⁻¹⁶	~4300
					(1.01–1.28)		(1.12–1.25)		(1.11–1.23)		(1.05–1.16)		(1.08–1.28)		(1.11–1.18)		
rs1470579	3	C	<i>IGF2BP2</i>	0.30	1.12	0.052	1.19	2.1×10 ⁻⁹	1.17	1.3×10 ⁻⁹	–	–	–	–	–	–	–
					(1.00–1.26)		(1.12–1.26)		(1.11–1.23)								
rs7754840	6	C	<i>CDKAL1</i>	0.31	1.17	7.5×10 ⁻³	1.06	0.024	1.08	2.4×10 ⁻³	1.16	1.3×10 ⁻⁸	1.12	9.5×10 ⁻³	1.12	4.1×10 ⁻¹¹	~4800
					(1.04–1.31)		(1.00–1.12)		(1.03–1.14)		(1.10–1.22)		(1.03–1.22)		(1.08–1.16)		
Previously published loci																	
rs1111875	10	C	<i>HHEX</i>	0.53	1.15	0.01	1.13	6.0×10 ⁻³	1.14	1.7×10 ⁻⁴	1.13	4.6×10 ⁻⁶	1.10	0.025	1.13	5.7×10 ⁻¹⁰	~4200
					(1.04–1.29)		(1.04–1.23)		(1.06–1.22)		(1.07–1.19)		(1.01–1.19)		(1.08–1.17)		
rs13266634	8	C	<i>SLC30A8</i>	0.65	1.01	0.92	1.12	0.014	1.07	0.047	1.12	7.0×10 ⁻⁵	1.18	6.8×10 ⁻⁵	1.12	5.3×10 ⁻⁸	~5400
					(0.90–1.31)		(1.02–1.23)		(1.00–1.16)		(1.05–1.18)		(1.09–1.29)		(1.07–1.16)		
rs7903146	10	T	<i>TCF7L2</i>	0.26	1.33	5.4×10 ⁻⁶	1.40	3.9×10 ⁻²⁸	1.38	2.3×10 ⁻³¹	1.37	6.7×10 ⁻¹³	1.34	1.4×10 ⁻⁸	1.37	1.0×10 ⁻⁴⁸	~800
					(1.17–1.50)		(1.32–1.49)		(1.31–1.46)		(1.25–1.49)		(1.21–1.49)		(1.31–1.43)		
rs5219	11	T	<i>KCNJ11</i>	0.47	1.14	0.012	1.14	2.6×10 ⁻⁶	1.15	1.0×10 ⁻⁷	1.15	1.3×10 ⁻³	1.11	0.014	1.14	6.7×10 ⁻¹¹	~3700
					(1.03–1.28)		(1.08–1.21)		(1.09–1.21)		(1.05–1.25)		(1.02–1.21)		(1.10–1.19)		
rs1801282	3	C	<i>PPARG</i>	0.86	1.02	0.83	1.11	6.2×10 ⁻³	1.09	0.019	1.23	1.3×10 ⁻³	1.20	1.4×10 ⁻³	1.14	1.7×10 ⁻⁶	~7900
					(0.87–1.19)		(1.02–1.19)		(1.01–1.16)		(1.09–1.41)		(1.07–1.33)		(1.08–1.20)		
Interesting for follow-up																	
rs17044137	4	A	<i>FLJ39370</i>	0.23	1.27	3.7×10 ⁻⁴	1.09	3.1×10 ⁻³	1.13	4.1×10 ⁻⁵	1.01	0.90	1.02	0.79	–	–	–
					(1.11–1.44)		(1.03–1.17)		(1.06–1.19)		(0.92–1.11)		(0.89–1.17)				
rs6698181	1	T	<i>PKN2</i>	0.29	1.17	7.6×10 ⁻³	1.09	9.7×10 ⁻⁴	1.11	5.3×10 ⁻⁵	1.00	0.93	1.21	4.1×10 ⁻³	–	–	–
					(1.04–1.32)		(1.03–1.16)		(1.05–1.16)		(0.92–1.09)		(1.06–1.37)				

*Alleles are indexed to the forward strand of NCBI Build 35. †Frequency of risk allele based on controls in DGI genome scan. ‡rs1111875 and rs13266634 genotyped in U.S. and Polish replication samples (*n* = 4470). §Replication *P*-value is one-tailed for additive model; all other *P*-values reported are additive and two-tailed. ||Results reported for same SNP, or proxy with $r^2 > 0.95$ (in HapMap-CEU) for FUSION and WTCCC except for *TCF7L2*; different SNPs selected based on genotyping platform (FUSION) or lowest *P*-value (FUSION, WTCCC/UKT2D). ¶Sample size required for 80% power (nominal $\alpha = 0.05$) calculated with the pooled OR estimate from DGI, FUSION, WTCCC studies (*n* > 32,000), assuming an equal number of cases and controls, allele frequency in controls (DGI genome scan), and T2D prevalence of 10% (consistent with estimates in adults from the population-based Botnia PPP study in Finland).

For validated loci, variability in significance across studies may appear surprising, and suggestive of heterogeneity. However, formal tests for heterogeneity in effect size were not significant ($P > 0.05$). Moreover, in a simulated association study of 1500 cases and 1500 controls, allele frequency of 20%, and an OR of 1.20 per copy, the median P -value was 10^{-4} , but ~5% of simulations P -values were >0.025 , and 5% of P -values were $<10^{-7}$. Thus, substantial variability in rank and significance is expected where power is modest, particularly if a SNP is selected based on the study with an extreme P -value.

We also performed genome-wide analyses for 18 clinical traits (table S2). The distribution of P -values was similar to that observed for T2D, with close match to expectation under the null hypothesis and a modest excess of signals in the tail (table S3). We observed strong evidence ($P < 10^{-4}$, rank in the top 100) for six previously reported common variants that influence lipid levels (table S3).

A previously unknown association with triglycerides was observed for rs780094 ($P = 3.7 \times 10^{-8}$), explaining 1% of residual variance in triglyceride levels (Fig. 2F; fig. S1A). This single SNP was tested in 5217 individuals from

the Malmö Diet and Cancer Study, Cardiovascular Arm (MDC-CVA); the association replicated ($P = 8.7 \times 10^{-8}$) (Table 2, fig. S1B). The association was observed by FUSION ($P = 10^{-4}$ controls; $P = 10^{-3}$ cases) (11).

SNP (Rs780094) is in a large block of LD, spanning 416 kb and 16 genes. The SNP resides, however, within a highly plausible biological candidate gene: glucokinase regulatory protein (*GCKR*). *GCKR* regulates glucokinase (*GCK*), the first glycolytic enzyme. Adenoviral-mediated overexpression of *GCKR* in mouse liver increased *GCK* activity and lowered fasting blood glu-

Fig. 2. Regional plots of six confirmed associations. For each of the (A) *CDKN2A/CDKN2B*, (B) *IGF2BP2*, (C) *CDKAL1*, (D) *TCF7L2*, (E) *HHEX* regions associated with T2D, and (F) *GCKR* region associated with triglyceride levels, all genotyped SNPs in the DGI genome scan are plotted with their P -values (as $-\log_{10}$ values) as a function of genomic position (with NCBI Build 35). In each panel, the SNP with the most significant association in the DGI combined analysis is listed (blue diamond) and its initial P -value in the genome scan (red diamond). Estimated recombination rates (taken from HapMap) are plotted to reflect the local LD structure around the associated SNPs and their correlated proxies (red: $r^2 \geq 0.8$; orange: $0.5 \leq r^2 < 0.8$; gray: $0.2 \leq r^2 < 0.5$; white: $r^2 < 0.2$). Gene annotations were taken from the University of California–Santa Cruz genome browser.

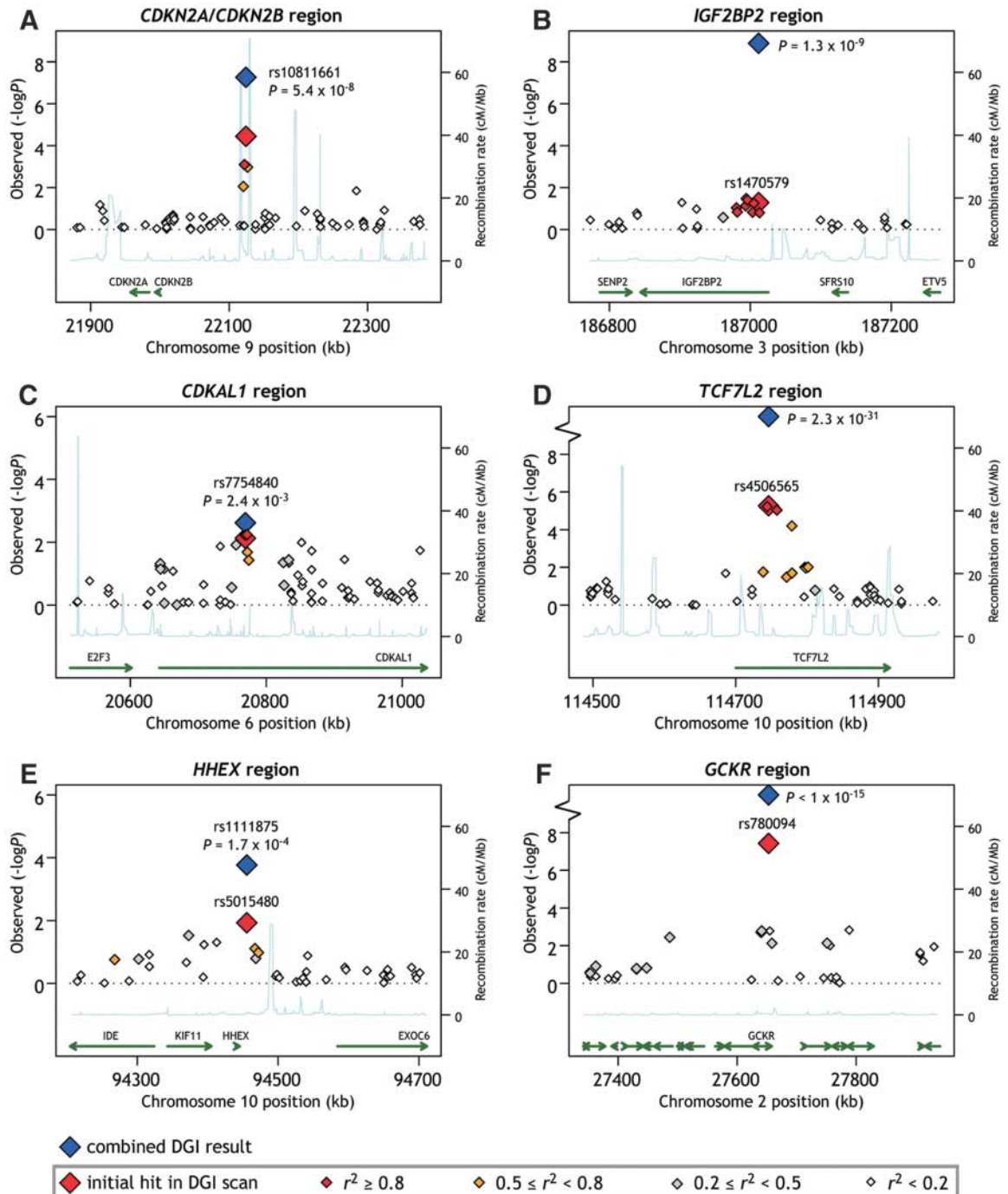


Table 2. Associations results for lipid/apolipoprotein traits in the top 100 of the DGI genome scan.

Trait	Previously reported SNP	Locus	SNP in genome scan*	Sample size	Minor allele†	Frequency	Residual variance explained (%)‡	P-value	Rank for trait
Novel locus									
TG	—	GCKR	rs780094	2659	T	0.35	1.2	3.7×10 ⁻⁸	1
TG	—	GCKR	rs780094	5217 [§]	T	0.35	0.5	8.7×10 ⁻⁸	—
Previously published loci									
LDL-C	rs4420638	APOE cluster	rs4420638	2601	G	0.22	2.1	3.4×10 ⁻¹³	1
LDL-C	rs693 (C7673T)	APOB	rs693	2600	A	0.49	1.0	7.1×10 ⁻⁷	4
HDL-C	rs1800775 (C-629A)	CETP	rs1800775	2623	C	0.47	2.1	2.5×10 ⁻¹³	1
HDL-C	rs328 (Ser447×)	LPL	rs17482753 (r ² = 1)	2624	T	0.10	0.9	3.6×10 ⁻⁵	24
HDL-C	rs1800588 (C-514T)	LIPC	rs261332 (r ² = 0.92)	2628	A	0.18	0.7	3.4×10 ⁻⁵	23
TG	rs328 (Ser447×)	LPL	rs17482753 (r ² = 1)	2652	T	0.10	1.0	4.9×10 ⁻⁷	2
TG	rs2266788 (T1259C)	APOA5	rs481843 (r ² = 0.52)	2649	T	0.11	0.7	3.3×10 ⁻⁵	28
apoB	rs693 (C7673T)	APOB	rs693	1448	A	0.45	1.1	9.4×10 ⁻⁵	40
apoA-I	rs1800775 (C-629A)	CETP	rs1800775	1451	C	0.44	1.6	2.7×10 ⁻⁶	1

*r² between proxy and previously reported SNP, based on HapMap-CEU. †Minor allele is indexed to the forward strand of NCBI Build 35. ‡Residual after adjustment of lipid/apolipoprotein levels for age, age², sex, diabetes status, and geographic region. §Replication in Malmö Diet and Cancer Study—Cardiovascular Arm.

cose (22); overexpression of GCK in liver led to lowered blood glucose and increased triglyceride levels (23, 24).

On the basis of these findings, we examined measures of glucose homeostasis. In both our scan and replication samples, the T allele of rs780094 trended toward association to lower glucose ($P < 0.10$, $P < 0.02$ respectively), less insulin resistance (HOMA-IR $P < 0.05$ and $P < 0.01$), and lower risk of T2D ($P < 0.20$, $P < 0.03$). The association of higher triglycerides with lower blood glucose reverses the correlation normally seen in humans, but is consistent with overexpression studies of GCKR and GCK in mouse models.

In summary, we carried out WGAS for T2D and 18 clinical traits. With collaborators we provide compelling evidence for associations at three previously unknown loci with risk of T2D, the first replications of two additional T2D loci, and a previously unknown association to triglyceride levels. Including long-recognized associations, our data provide strong support for 15 common variants as influencing T2D and lipid levels in European populations. The annotations of the new T2D genes suggest a primary role of the pancreatic beta cell, but much additional work will be required to develop and test this hypothesis.

Our results have general implications for genome-wide association studies of common diseases. The modest effect of each SNP demonstrates that large sample sizes will be required to discover and validate genetic risk factors for common disease. Although the eight T2D variants discussed in this report each conveys a substantial population attributable risk (5 to 27% at each locus), each contributes very modestly to overall variance in diabetes risk (0.04 to 0.5%, ~2.3% combined across the eight SNPs). Thus, many more variants remain to be found as risk factors for T2D, and many questions remain about the balance between common and rare variants, SNPs and copy-number alterations, main effects and epistasis. Additional associated

variants may be found in or near these loci, as has been the case for other examples (25–31).

The most notable aspect of this and other such studies may be the generation of new hypotheses. Before this work, few would have argued that these genes and noncoding genomic regions were a high priority for T2D research. Now, on the basis of their validated relationship to disease, it is evident that they should be explored and understood. The ability to discover etiological factors that fall outside previous biological hypotheses is a major motivation for unbiased genome-wide approaches and is well supported by these and other emerging data from genome-wide association studies.

References and Notes

- R. A. Fisher, *Trans. R. Soc. Edinb.* **52**, 399 (1918).
- M. Kimura, T. Ota, *Genetics* **75**, 199 (1973).
- The International Haplotype Map Consortium, *Nature* **437**, 1299 (2005).
- Materials and methods are available as supporting material on Science Online.
- I. Pe'er et al., *Nat. Genet.* **38**, 663 (2006).
- P. I. W. de Bakker et al., *Nat. Genet.* **37**, 1217 (2005).
- C. D. Campbell et al., *Nat. Genet.* **37**, 868 (2005).
- D. G. Clayton et al., *Nat. Genet.* **37**, 1243 (2005).
- A. L. Price et al., *Nat. Genet.* **38**, 904 (2006).
- The Wellcome Trust Case Control Consortium, *Nature*, in press.
- L. J. Scott et al., *Science* **316**, 1341 (2007).
- E. Zeggini et al., *Science* **316**, 1336 (2007).
- J. Krishnamurthy et al., *Nature* **443**, 453 (2006).
- F. C. Nielsen, J. Nielsen, J. Christiansen, *Scand. J. Clin. Lab. Invest. Suppl.* **234**, 93 (2001).
- F. M. Spagnoli, A. H. Brivanlou, *Dev. Biol.* **292**, 442 (2006).
- M. Wagner et al., *Gastroenterology* **124**, 1901 (2003).
- M. Ubeda, J. M. Rukstalis, J. F. Habener, *J. Biol. Chem.* **281**, 28858 (2006).
- S. F. Grant et al., *Nat. Genet.* **38**, 320 (2006).
- R. Sladek et al., *Nature* **445**, 881 (2007).
- E. H. Hani et al., *Diabetologia* **41**, 1511 (1998).
- S. S. Deeb et al., *Nat. Genet.* **20**, 284 (1998).
- E. D. Slosberg et al., *Diabetes* **50**, 1813 (2001).
- R. M. O'Doherty, D. L. Lehman, S. Telegaue-Potts, C. B. Newgard, *Diabetes* **48**, 2022 (1999).
- T. Ferre, E. Riu, F. Bosch, A. Valera, *FASEB J.* **10**, 1213 (1996).
- R. R. Graham et al., *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6758 (2007).
- J. Maller et al., *Nat. Genet.* **38**, 1055 (2006).

- M. Li et al., *Nat. Genet.* **38**, 1049 (2006).
- R. H. Duerr et al., *Science* **314**, 1461 (2006).
- J. P. Hugot et al., *Nature* **411**, 599 (2001).
- C. A. Haiman et al., *Nat. Genet.* **39**, 638 (2007).
- I. K. Kotowski et al., *Am. J. Hum. Genet.* **78**, 410 (2006).
- We are indebted to all participants in the studies for their support and contributions. This work was supported by Novartis Institutes for BioMedical Research (D.A.), with additional support from The Richard and Susan Smith Family Foundation/American Diabetes Association Pinnacle Program Project Award (D.A., J.N.H., M.J.D.). We thank our colleagues from the UKT2D study and the FUSION study for sharing data before publication and for close and enjoyable collaboration. We thank P. Donnelly for sharing results before publication on behalf of the WTCCC. The following are supported by NIH Research Career Awards: J.C.F. (K23 DK65978-04), H.N.L. [National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) K23 DK067288], NIH Research Service Award (R.S.), and a NIH Training grant award (E.K.S.). G.L. is supported by March of Dimes research grant (6-FY04-61). S.K.'s work is supported by grants from the Doris Duke Charitable Foundation, the Fannie E. Rippel Foundation, and NIH (grant K23-HL083102). M.J.D., D.A., S.P., and P.I.W.d.B. acknowledge support from NIH/National Heart, Lung, and Blood Institute (grant U01 HG004171). D.A. is a Burroughs Wellcome Fund Clinical Scholar in Translational Research, and a Distinguished Clinical Scholar of the Doris Duke Charitable Foundation. L.G., T.T., B.I., M.R.T., and the Botnia Study are principally supported by the Sigrid Juselius Foundation, the Finnish Diabetes Research Foundation, The Folkhalsan Research Foundation, and Clinical Research Institute HUCH Ltd; work in Malmö, Sweden, was also funded by a Linné grant from the Swedish Research Council (349-2006-237). We thank H. Hilden, H. Perttunen-Nio, and S. Mahan for technical assistance. We thank the Botnia and Skara research teams for clinical contributions and colleagues at Massachusetts General Hospital, Harvard, Broad, and Novartis for helpful discussions. We thank J. Avruch, M. Fishman, E. Lander, R. Lifton, and E. Scolnick for support and helpful comments on the manuscript.

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research:

Writing team: Richa Saxena¹⁻⁶ (Team Leader), Benjamin F. Voight,^{1-3,5} Valeriya Lyssenko,⁷ Noël P. Burtt,¹ Paul I. W. de Bakker,¹⁻⁶ Hong Chen,⁸ Jeffrey J. Roix,⁸ Sekar Kathiresan,^{1,3,5} Joel N. Hirschhorn,^{1,6,9-11} Mark J. Daly,^{1-3,5} Thomas E. Hughes,⁸ Leif Groop,^{7,12} David Altshuler¹⁻⁶ (Chair)
Project manager: Noël P. Burtt,¹ Leif Groop,^{7,12} Thomas E. Hughes,⁸ David Altshuler¹⁻⁶

Study design: Richa Saxena¹⁻⁶ and Valeriya Lyssenko⁷ (Team Leaders), Peter Almgren,⁷ Paul I. W. de Bakker,¹⁻⁶ Noël P. Burtt,¹ Jose C. Florez,¹⁻⁶ Hong Chen,⁸ Joanne Meyer,⁸ Mark N. Hirschhorn,^{1,6,9-11} Mark J. Daly,^{1-3,5} Thomas E. Hughes,⁸ Leif Groop,^{7,12} David Altshuler¹⁻⁶ (Chair)

Clinical characterization and phenotypes: Valeriya Lyssenko⁷ and Richa Saxena¹⁻⁶ (Team Leaders), Peter Almgren,⁷ Kristin Ardlie,¹ Kristina Bengtsson Boström,¹³ Noël P. Burtt,¹ Hong Chen,⁸ Jose C. Florez,¹⁻⁶ Bo Isomaa,^{14,15} Sekar Kathiresan,^{1,3,5} Guillaume Lettre,^{1,6,9-11} Ulf Lindblad,¹⁶ Helen N. Lyon,^{1,6,9-11} Olle Melander,⁷ Christopher Newton-Cheh,^{1-3,5} Peter Nilsson,¹⁷ Marju Orho-Melander,⁷ Lennart Råstam,¹⁶ Elizabeth K. Speliotes,^{1,3,6,9-11} Marja-Riitta Taskinen,¹² Tiinamaija Tuomi,^{12,15} Benjamin F. Voight,^{1-3,5} David Altshuler,¹⁻⁶ Joel N. Hirschhorn,^{1,6,9-11} Thomas E. Hughes,⁸ Leif Groop,^{7,12} (Chair)

DNA sample QC and diabetes replication genotyping: Candace Guiducci¹ and Valeriya Lyssenko⁷ (Team Leaders), Anna Berglund,⁷ Joyce Carlson,¹⁸ Lauren Gianniny,¹ Rachel Hackett,¹ Liselotte Hall,¹⁸ Johan Holmkvist,⁷ Esa Laurila,⁷ Marju Orho-Melander,⁷ Marketa Sjögren,⁷ Maria Sterner,¹⁸ Aarti Surti¹ Margareta Svensson,⁷ Malin Svensson,⁷ Ryan Tewhey,¹ Noël P. Burtt¹ (Chair)

Whole genome scan genotyping: Brendan Blumenstiel¹ (Team Leader), Melissa Parkin,¹ Matthew DeFelice,¹ Candace Guiducci,¹ Ryan Tewhey,¹ Rachel Barry,¹ Wendy Brodeur,¹ Noël P. Burtt,¹ Jody Camarata,¹ Nancy Chia,¹ Mary Fava,¹ John Gibbons,¹ Bob Handsaker,¹ Claire Healy,¹ Kieu Nguyen,¹ Casey

Gates,¹ Carrie Sougnez,¹ Diane Gage,¹ Marcia Nizzari,¹ David Altshuler,¹⁻⁶ Stacey B. Gabriel¹ (Chair)

GCKR replication genotyping and analysis (Malmö Diet and Cancer Study): Sekar Kathiresan^{1,3,5} (Team Leader), Candace Guiducci,¹ Aarti Surti,¹ Noël P. Burtt,¹ Olle Melander,⁷ Marju Orho-Melander⁷ (Chair)

Statistical analysis: Benjamin F. Voight^{1-3,5} and Paul I. W. de Bakker¹⁻⁶ (Team Leaders), Richa Saxena,¹⁻⁶ Valeriya Lyssenko,⁷ Peter Almgren,⁷ Noël P. Burtt,¹ Hong Chen,⁸ Gung-Wei Chirn,⁸ Qicheng Ma,⁸ Hemang Parikh,⁷ Delwood Richardson,⁸ Darrell Ricke,⁸ Jeffrey J. Roix,⁸ Leif Groop,^{7,12} Shaun Purcell,^{1,2} David Altshuler,¹⁻⁶ Mark J. Daly^{1-3,5} (Chair)

¹Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Cambridge, MA 02142, USA. ²Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA. ³Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA. ⁴Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA. ⁵Department of Medicine, Harvard Medical School, Boston, MA 02115, USA. ⁶Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. ⁷Department of Clinical Sciences, Diabetes and Endocrinology Research Unit, University Hospital Malmö, Lund University, Malmö, Sweden. ⁸Diabetes and Metabolism Disease Area, Novartis Institutes for BioMedical Research, 100 Technology Square, Cambridge, MA 02139, USA. ⁹Depart-

ment of Pediatrics, Harvard Medical School, Boston, MA 02115, USA. ¹⁰Division of Endocrinology, Children's Hospital, Boston, MA 02115, USA. ¹¹Division of Genetics, Children's Hospital, Boston, MA 02115, USA. ¹²Department of Medicine, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. ¹³Skaraborg Institute, Skövde, Sweden. ¹⁴Malmåska Municipal Health Center and Hospital, Jakobstad, Finland. ¹⁵Folkhälsan Research Center, Helsinki, Finland. ¹⁶Department of Clinical Sciences, Community Medicine Research Unit, University Hospital Malmö, Lund University, Malmö, Sweden. ¹⁷Department of Clinical Sciences, Medicine Research Unit, University Hospital Malmö, Lund University, Malmö, Sweden. ¹⁸Clinical Chemistry, University Hospital Malmö, Lund University, Malmö, Sweden. ¹⁹Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02115, USA.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1142358/DC1

Materials and Methods

Figs. S1 and S2

Tables S1 to S6

References

9 March 2007; accepted 20 April 2007

Published online 26 April 2007;

10.1126/science.1142358

Include this information when citing this paper.

Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes

Eleftheria Zeggini,^{1,2*} Michael N. Weedon,^{3,4*} Cecilia M. Lindgren,^{1,2*} Timothy M. Frayling,^{3,4*} Katherine S. Elliott,² Hanaango, ^{3,4} Nicholas J. Timpson,^{2,5} John R. B. Perry,^{3,4} Nigel W. Rayner,^{1,2} Rachel M. Freathy,^{3,4} Jeffrey C. Barrett,² Beverley Shields,⁴ Andrew P. Morris,² Sian Ellard,^{4,6} Christopher J. Groves,¹ Lorna W. Harries,⁴ Jonathan L. Marchini,⁷ Katharine R. Owen,¹ Beatrice Knight,⁴ Lon R. Cardon,² Mark Walker,⁸ Graham A. Hitman,⁹ Andrew D. Morris,¹⁰ Alex S. F. Doney,¹⁰ The Wellcome Trust Case Control Consortium (WTCCC),† Mark I. McCarthy,^{1,2,‡§} Andrew T. Hattersley^{3,4,‡}

The molecular mechanisms involved in the development of type 2 diabetes are poorly understood. Starting from genome-wide genotype data for 1924 diabetic cases and 2938 population controls generated by the Wellcome Trust Case Control Consortium, we set out to detect replicated diabetes association signals through analysis of 3757 additional cases and 5346 controls and by integration of our findings with equivalent data from other international consortia. We detected diabetes susceptibility loci in and around the genes *CDKAL1*, *CDKN2A/CDKN2B*, and *IGF2BP2* and confirmed the recently described associations at *HHEX/IDE* and *SLC30A8*. Our findings provide insight into the genetic architecture of type 2 diabetes, emphasizing the contribution of multiple variants of modest effect. The regions identified underscore the importance of pathways influencing pancreatic beta cell development and function in the etiology of type 2 diabetes.

The pathophysiological basis of type 2 diabetes (T2D) remains unclear despite its growing global importance (1). Candidate gene and positional cloning efforts have suggested many putative susceptibility variants, but unequivocal replications are so far limited to variants in just three genes: *PPARG*, *KCNJ11*, and *TCF7L2* (2–4).

Improved understanding of the correlation between genetic variants [linkage disequilibrium (LD)], allied to advances in genotyping technology, have enabled systematic searches for disease-associated common variants on a genome-wide

scale. The Wellcome Trust Case Control Consortium (WTCCC) recently completed such a genome-wide association (GWA) scan in 1924 T2D cases and 2938 population controls from the United Kingdom, using the Affymetrix GeneChip Human Mapping 500 k Array Set (5). The strongest association signals genome-wide were observed for single-nucleotide polymorphisms (SNPs) in *TCF7L2*. [For example, for rs7901695, odds ratio (OR) = 1.37, 95% confidence interval (CI) = 1.25–1.49, and $P = 6.7 \times 10^{-13}$.] The other known T2D susceptibility variants were detected with effect sizes consistent with previous reports (2, 3).

Here, we describe how integration of data from the WTCCC scan and our own replication studies with similar information generated by the Diabetes Genetics Initiative (DGI) (6) and the Finland–United States Investigation of NIDDM Genetics (FUSION) (7) has identified several additional susceptibility variants for T2D.

In the WTCCC study, analysis of 490,032 autosomal SNPs in 16,179 samples yielded 459,448 SNPs that passed initial quality control (5). We considered only the 393,453 autosomal SNPs with minor allele frequency (MAF) exceeding 1% in both cases and controls and no extreme departure from Hardy-Weinberg equilibrium ($P < 10^{-4}$ in cases or controls) (8). This T2D-specific data set shows no evidence of substantial confounding from population substructure and genotyping biases (8).

To distinguish true associations from those reflecting fluctuations under the null or residual errors arising from aberrant allele calling, we first submitted putative signals from the WTCCC study to additional quality control, including cluster-plot visualization and validation genotyping on a second platform (8). Next, we attempted replication of selected signals in up to 3757 additional cases and 5346 controls (replication sets RS1 to RS3). RS1 comprised 2022 cases and 2037 controls from the U.K. Type 2 Diabetes Genetics Consortium collection (UKT2DGC) (all from Tayside, Scotland). RS2 included 632 additional T2D cases and 1750 population controls from the Exeter Family Study of Child Health (EFSOCH). A subset of SNPs were typed in RS3, comprising a further 1103 cases and 1559 controls from the UKT2DGC (table S1).

The first wave of validated SNPs sent for replication was selected from the 30 SNPs, in nine distinct chromosomal regions (excluding *TCF7L2*), which had, in the WTCCC scan alone,