Genome-wide association study identifies three novel loci for type 2 diabetes

Kazuo Hara[1,2,∗,†], Hayato Fujita[1,†], Todd A. Johnson[3,†], Toshimasa Yamauchi[1,6], Kazuki Yasuda[8], Momoko Horikoshi[1], Chen Peng[9], Cheng Hu[10], Ronald C. W. Ma[11,12,13], Minako Imamura[4], Minoru Iwata[15], Tatsuhiko Tsunoda[3], Takashi Morizono[3], Nobuhiro Shojima[1], Wing Yee So[11,12,13], Ting Fan Leung[14], Patrick Kwan[11], Rong Zhang[10], Jie Wang[10], Weihui Yu[10], Hiroshi Maegawa[16], Hiroshi Hirose[17], DIAGRAM consortium[18], Kohei Kaku[19], Chikako Ito[20], Hirotaka Watada[7], Yasushi Tanaka[21], Kazuyuki Tobe[15], Atsunori Kashiwagi[16], Ryuzo Kawamori[7], Weiping Jia[10], Juliana C. N. Chan[11,12,13], Yik Ying Teo[9,22,23,24,26], Tai E. Shyong[9,25,27], Naoyuki Kamatani[28], Michiaki Kubo[5], Shiro Maeda[4] and Takashi Kadowaki[1]

[1]Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, The University of Tokyo, Tokyo 113-8655, Japan, [2]Department of Integrated Molecular Science on Metabolic Diseases, 22nd Century Medical and Research Center, the University of Tokyo, Tokyo 113-8655, Japan, [3]Laboratory for Medical Science Mathematics, [4]Laboratory for Endocrinology, Metabolism, and Kidney Diseases and [5]Research Group for Genotyping, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan, [6]Sportology Center, Graduate School of Medicine and [7]Department of Medicine, Metabolism and Endocrinology, School of Medicine, Juntendo University, Tokyo 113-8421, Japan, [8]Department of Metabolic Disorder, Diabetes Research Center, National Center for Global Health and Medicine, Research Institute, Tokyo 162-8655, Japan, [9]Saw Swee Hock School of Public Health, National University of Singapore, MD3, 16 Medical Drive, Singapore 117597, Singapore, Singapore, [10]Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai Diabetes Institute, Shanghai Key Laboratory of Diabetes Mellitus, 600 Yishan Road, Shanghai 200233, China, [11]Department of Medicine and Therapeutics, Chinese University of Hong Kong, Prince of Wales Hospital, Hong Kong, China, [12]Li Ka Shing Institute of Health Sciences, [13]Hong Kong Institute of Diabetes and Obesity and [14]Department of Paediatrics, Chinese University of Hong Kong, Hong Kong, China, [15]First Department of Internal Medicine, University of Toyama, Toyama 930-0194, Japan, [16]Department of Medicine, Shiga University of Medical Science, Otsu, Shiga 520-2192, Japan, [17]Health Center, Keio University School of Medicine, Tokyo 160-8582, Japan, [18]For a Full List of Authors, see Supplementary Material and [19]Division of Diabetes, Endocrinology and Metabolism, Department of Internal Medicine, Kawasaki Medical School, Kurashiki, Okayama 701-0192, Japan, [20]Medical Court Life Care Clinic, Hiroshima 730-0012, Japan, [21]Department of Internal Medicine, Division of Metabolism and Endocrinology, St. Marianna University School of Medicine, Kawasaki, Kanagawa 216-8511, Japan, [22]Life Sciences Institute, [23]NUS Graduate School for Integrative Science and Engineering, [24]Department of Statistics and Applied Probability and [25]Department of Medicine, National University of Singapore, Singapore, Singapore, [26]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore, [27]Duke-National University of Singapore Graduate Medical School, Singapore, Singapore and [28]StaGen, Tokyo 111-0051, Japan

---

Although over 60 loci for type 2 diabetes (T2D) have been identified, there still remains a large genetic component to be clarified. To explore unidentified loci for T2D, we performed a genome-wide association study (GWAS) of 6 209 637 single-nucleotide polymorphisms (SNPs), which were directly genotyped or imputed using East Asian

---

*To whom correspondence should be addressed at: Department of Diabetes and Metabolic Diseases, University of Tokyo Hospital, 7-3-1 Hongo, Bunkyo-ku, Tokyo 115-0055, Japan. Tel: 81 358009161; Fax: 81 358009161; Email: thara-tky@umin.ac.jp
†These authors contributed equally to this work.

references from the 1000 Genomes Project (June 2011 release) in 5976 Japanese patients with T2D and 20 829 nondiabetic individuals. Nineteen unreported loci were selected and taken forward to follow-up analyses. Combined discovery and follow-up analyses (30 392 cases and 34 814 controls) identified three new loci with genome-wide significance, which were *MIR129-LEP* [rs791595; risk allele = A; risk allele frequency (RAF) = 0.080; *P* = 2.55 × 10⁻¹³; odds ratio (OR) = 1.17], *GPSM1* [rs11787792; risk allele = A; RAF = 0.874; *P* = 1.74 × 10⁻¹⁰; OR = 1.15] and *SLC16A13* (rs312457; risk allele = G; RAF = 0.078; *P* = 7.69 × 10⁻¹³; OR = 1.20). This study demonstrates that GWASs based on the imputation of genotypes using modern reference haplotypes such as that from the 1000 Genomes Project data can assist in identification of new loci for common diseases.

INTRODUCTION

T2D is a complex disease characterized by hyperglycemia resulting from impaired pancreatic β-cell function and a decreased action of insulin on target tissues (1,2). Familial aggregation and twin studies have shown that a genetic component plays a major role in the onset of T2D. Although there has been a marked increase in the identification of genetic loci for T2D, with over 60 of the discoveries made through genome-wide association studies (GWASs), it is estimated that at best 10% of the genetic component of T2D can be explained by the loci identified so far (3–5). The 1000 Genomes Project was founded with the aim of characterizing over 95% of variants with a minor allele frequency (MAF) >1% to provide a more extensive catalog of the genetic variations in major ethnic groups (6,7). We reasoned that we could take advantage of this recent advance to explore the remaining unidentified part of the genetic component of T2D by extending our search. This idea is supported by a recent article that reported that imputing the Wellcome Trust Case Control Consortium Phase I genotype data by using references from the 1000 Genomes Project detected loci for diabetes that were not identified in the original case–control study (8). In the meantime, genetic studies have failed to explain the reason why the prevalence of T2D differs across different ethnic groups. In light of the high and rising prevalence of T2D in East Asia (9), there is a huge need for exploring genetic variations vigorously in East Asian populations, although GWASs for T2D have been performed not only in European (10–15), but also in Asian (16–22) populations.

RESULTS

GWAS of typed and imputed SNPs using 1000 Genomes Project data

This is a three-stage study comprised of (i) a discovery GWAS, (ii) follow-up analysis and (iii) validation analysis. During the discovery GWAS, we obtained directly genotyped data with a 610K SNP array in 5976 Japanese patients with T2D and 20 829 subjects without diabetes. To increase the genome-wide coverage of genetic variations in this population, we imputed 10 811 164 SNPs of 572 East Asian haplotypes [194 Han Chinese in Beijing (CHB), 200 Chinese in Metropolitan Denver (CHS) and 178 Japanese in Tokyo (JPT)], with 194 generated by the 1000 Genomes Project (June 2011 release). We found that concordance between the directly genotyped data and the imputed genotypes was generally good (94.2%). We then tested the association with T2D for 6 209 637 typed and

imputed SNPs that passed our quality control criteria. An outline of the present study is shown in Supplementary Material, Figure S1, and details of the methods used in each stage are given in Supplementary Material, Table S1. Based on the principal component analysis (PCA), there was no apparent heterogeneity in genetic background among the subjects of the discovery stage, and we observed no indication of a population stratification influencing the result of the discovery analysis; the genomic control inflation factor (λ) was 1.074 (adjusted for 1000 cases and 1000 controls: λ_{1000} = 1.008) (Supplementary Material, Fig. S2). As shown in Figure 1, 9 known loci (in black) reached genome-wide significance ($P < 5 \times 10^{-8}$), and we observed 39 loci with a *P*-value for association with T2D of $<1 \times 10^{-4}$, consisting of 19 previously reported and 20 unreported loci. We selected a top SNP at each of the 20 unreported loci to take forward to follow-up analyses.

Replication and validation of top SNPs selected in the discovery GWAS

A top SNP at each of the two unreported loci was taken forward to follow-up analysis by *in silico* (2799 cases and 3793 controls) and *de novo* (10 319 cases and 6795 controls) genotyping in East Asian populations (Supplementary Material, Table S2). By combining the results obtained in the first discovery GWAS and the second follow-up analyses (in total 19 094 cases and 31 417 controls), four novel SNPs reached genome-wide significance ($P < 5 \times 10^{-8}$), which were near *MIR129-LEP* (rs791595; $P = 5.46 \times 10^{-11}$; OR = 1.19), *GPSM1* (rs11787792; $P = 7.26 \times 10^{-11}$; OR = 1.18), *MRPS35* (rs7316898; $P = 7.36 \times 10^{-9}$, OR = 1.10) and *SLC16A13* (rs312457; $P = 2.15 \times 10^{-8}$; OR = 1.18) (Supplementary Material, Table S2). To validate these associations, the four top SNPs were subsequently genotyped further in another 11 298 cases and 3397 controls. We confirmed that *MIR129-LEP* (rs791595; risk allele = A; RAF = 0.08; $P = 2.55 \times 10^{-13}$; OR = 1.17), *GPSM1* (rs11787792; risk allele = A; RAF = 0.874 ; $P = 1.74 \times 10^{-10}$; OR = 1.15) and *SLC16A13* (rs312457; risk allele = G ; RAF = 0.078; $P = 7.69 \times 10^{-13}$; OR = 1.20) were associated with T2D, reaching a Bonferroni-adjusted *P*-value for significance of 8.05×10^{-9}, which is more stringent than the conventional cut-off of 5×10^{-8} (Table 1) (Fig. 2). In contrast, rs7316898 in *MRPS35* did not reach genome-wide significance ($P = 5.42 \times 10^{-5}$; OR = 0.94) after combining three-stage analyses. Rs147689733 in *TNKS2* also reached genome-wide significance ($P = 2.43 \times 10^{-9}$) (Supplementary Material, Table S2). However, conditioning for rs12219514 in the previously
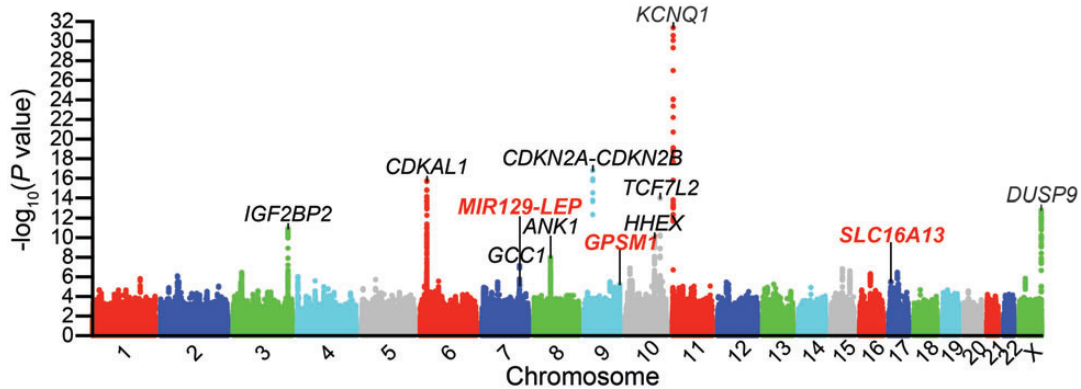
Figure 1. Manhattan plot for the discovery analysis of directly genotyped and imputed SNPs in 5976 T2D cases and 20 829 controls. Known loci that reached genome-wide significance ($P < 5 \times 10^{-8}$) in the discovery analysis (5976 T2D cases and 20 829 controls) are indicated in black and the three loci that reached genome-wide significance in the combined analysis of discovery and follow-up analyses 30 392 cases and 34 814 controls) are indicated in red.

reported locus *HHEX* (13,14) abrogated the signal at rs147689733, showing that it was a proxy for the stronger *HHEX* signal. Therefore, this SNP was not taken forward for validation analysis. To further confirm our imputation-based results, we performed additional direct genotyping of the three novel SNPs in a subset of the discovery GWAS samples that consisted of ~2600 samples and observed almost perfect concordance with the imputed data (Supplementary Material, Table S3). Combined with the direct genotyping performed in the other two analysis stages, the three novel SNPs were directly genotyped in over 34 000 samples.

Similarities and differences between East Asians and Europeans

To briefly assess the similarity of genetic architecture between East Asians and European populations, we compared the reported effect sizes of loci reported first in Europeans and then those obtained in the present study (Supplementary Material, Table S4). A high concordance in the direction of the effects and correlation of ORs between the two populations was observed ($r = 0.49$, $P = 0.0018$), showing that the two populations shared the same susceptibility genes when the data are restricted to common variants identified by previous GWASs. As for the novel loci identified in the present study, there is no evidence for an association between SNPs in and near *MIR129-LEP* (Supplementary Material, Fig. S3A) and *SLC16A13* (Supplementary Material, Fig. S3C) and only modest evidence for association around *GPSM1* (Supplementary Material, Fig. S3B) in European populations (15). Rs4731420, being in perfect linkage disequilibrium (LD) with rs791595 near *MIR129-LEP* ($r^2 = 1.00$), was not associated with T2D in the DIAGRAMv3 ($P = 0.760$). There are no available data on rs312457 in *SLC16A13* and its proxy in European populations. However, rs11652868, which was in modest LD with another SNP that was associated with T2D in the present study (rs312458), was not associated with T2D in the European data. The *GPSM1* locus has been reported to be linked to diabetes-related traits (23,24): rs3829109, located in the adjacent gene *DNLZ*, was previously associated with the fasting

glucose level (23), and rs60980157, a nonsynonymous SNP (Ser391Leu of *GPSM1*), was previously reported to be associated with an insulin secretion measure, the insulinogenic index (24). These two SNPs were in modest LD with rs11787792 ($r^2 = 0.127$ for rs3829109 and 0.344 for rs60980157) and were associated with T2D ($P = 6.10 \times 10^{-6}$ and 9.03×10^{-6}, respectively) in this study, but we did not detect any association between rs11787792 and fasting glucose level or an insulin secretion index, HOMA-β (Supplementary Material, Table S5).

Comparison between association statistics of typed and imputed SNPs

Supplementary Material, Figure S4, demonstrates the usefulness of using imputation to explore previously unknown loci; imputed SNPs (grey circles) were more significantly associated with T2D than typed (red circles) SNPs. The signal at *GPSM1* would not have been taken forward to follow-up analysis, and therefore, it would have been missed (Supplementary Material, Fig. S4). Moreover, rs11787792 in *GPSM1* is included only in 1000 Genomes Project data and not in HapMap2 data. We also sought to define the most relevant SNPs for susceptibility to T2D in previously identified loci. We found rs7656416 in *CTBP1* ($P = 1.29 \times 10^{-8}$) to be more significantly associated with T2D than the previously reported SNP, rs6815464 in *MAEA* ($P = 1.32 \times 10^{-5}$) (20). Rs7656416 and rs6815464 were in LD ($r^2 = 0.58$), and conditioning for rs7656416 in the logistic regression abrogated the signal at rs6815464.

BMI-stratifying analysis

It has been reported that the genetic predisposition to T2D is different in lean subjects compared with findings for obese subjects (25,26). We tested the association between the top SNP in each of the eight autosomal known loci with $P < 5 \times 10^{-8}$ (adjusted for sex, age, and the first four principal components from PCA in the present discovery analysis) after dichotomizing the T2D subjects into lean (body mass index, BMI < 25 kg/m²) and overweight (BMI ≥ 25 kg/m²) groups. We found that the top SNPs in the known loci were more significantly associated with T2D

Table 1. Three new T2D loci reaching genome-wide significance from combined analysis

| SNP	Chr	Chromosome position	Nearby gene	Risk allele	Other allele	RAF	First stage OR (95%CI)		Second + third stage OR (95%CI)		1st + 2nd +3rd Stage OR (95%CI)	
							5976 cases and 20 829 controls	*P*-value	24 416 cases and 13 985 controls	*P*-value	30 392 cases and 34 814 controls	*P*-value
rs791595	7	127 862 802	*MIR129-LEP*	A	G	0.080	1.19 (1.11–1.28)	4.69E−06	1.16 (1.06–1.18)	5.64E−07	1.17 (1.12–1.22)	2.55E−13
rs11787792	9	139 252 148	*GPSM1*	A	G	0.874	1.17 (1.09–1.25)	7.12E−06	1.14 (1.06–1.18)	3.94E−06	1.15 (1.10–1.20)	1.74E−10
rs312457	17	6 940 393	*SLC16A13*	G	A	0.078	1.19 (1.10–1.29)	9.40E−06	1.20 (1.13–1.28)	1.04E−08	1.20 (1.14–1.26)	7.69E−13
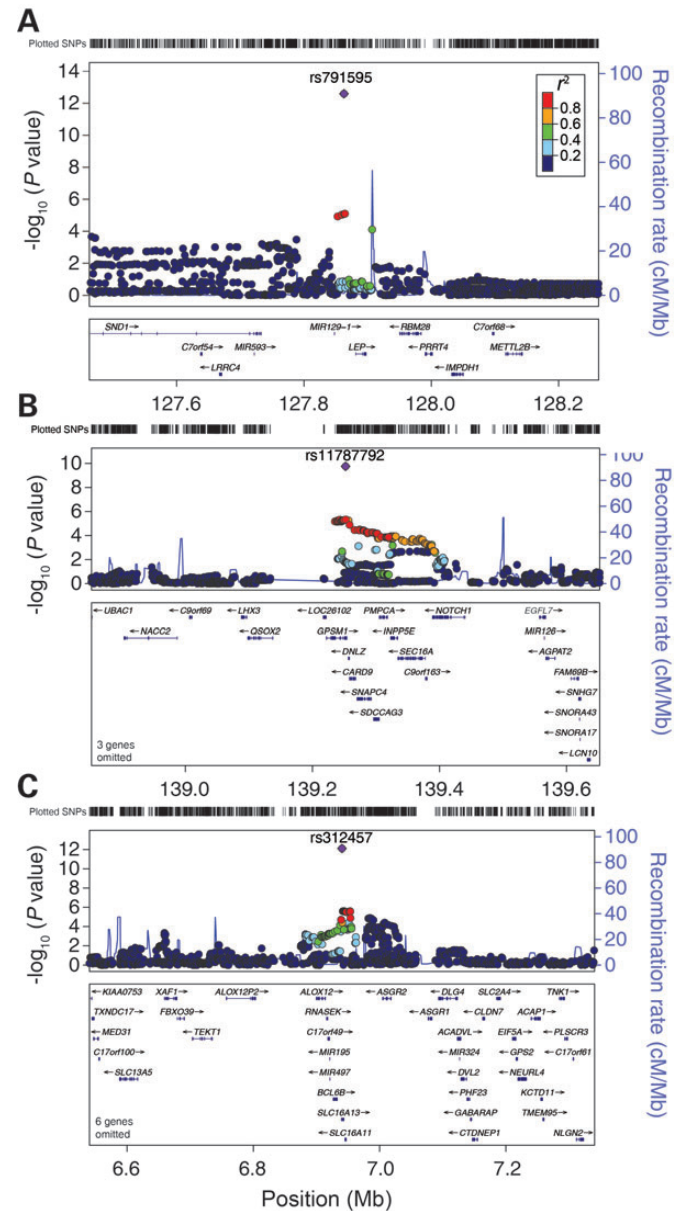
Figure 2. Regional plots of the three newly discovered T2D loci: (**A**) *MIR129-LEP,* (**B**) *GPSM1* and (**C**) *SLC16A13*. Regional associations were plotted for the three novel loci that showed genome-wide significance after combing the results from Stages 1, 2 and 3. Genotyped and imputed SNPs are plotted with the *P*-values (as −log10 values) from discovery analysis versus their physical position (NCBI Build 37). In each panel, the top SNP is represented by a purple diamond and *P*-values derived by combining first + second + third stage results are shown. *P*-values of other SNPs were derived from Stage 1 result alone and are color coded according to their pairwise LD with the top SNP based on 1000 Genomes Project East Asian reference data. Estimated recombination rates are plotted to reflect the local LD structure.

in lean subjects than in overweight subjects (Supplementary Material, Fig. S5). Several loci (*KCNQ1*, *CDC123*, *IGF2BP2* and *CDKAL1*) had large enough heterogeneity *z*-scores to suggest that a substantial difference may exist in the association statistics between lean and overweight groups (Supplementary Material, Table S6).

Sex-stratifying analysis

We performed a sex-differentiated analysis to test for sexual dimorphism of associations between SNPs and T2D, allowing for heterogeneity in allelic effects between males (3938 cases and 9553 controls) and females (1831 cases and 9220 controls) (Supplementary Material, Table S7). We found modest evidence for sexual dimorphism in 23 genes and regions (heterogeneity $P < 1 \times 10^{-4}$) (Supplementary Material, Table S7). We did not find evidence of sex heterogeneity at previously reported loci such as *KCNQ1, DGKB,* and *GRB14*, nor present novel loci.

DISCUSSION

We identified three novel T2D loci in East Asian populations: *MIR129-LEP*, *GPSM1* and *SLC16A13*. *GPSM1* locus could not have been identified as a T2D locus without imputation using the 1000 Genomes Project reference data, as shown in Supplementary Material, Figure S4, This demonstrates the utility of using up-to-date reference haplotype data such as that from the 1000 Genomes Project to perform imputation to identify novel loci for T2D. Given the fact that whole-genome sequencing for thousands of samples is still highly cost intensive, an imputation-based GWAS is still useful to search for novel common disease loci. At *MIR129-LEP* and *SLC16A13* loci, imputed SNPs were more significantly associated with type 2 diabetes (T2D) than directly genotyped SNPs, but imputation was not mandatory for the identification of these loci, because a directly genotyped SNP reached the P-value of 1×10^{-4}, the cut-off for follow-up analysis. Those two loci could have been found because of improved statistical power by increasing the sample size from ∼7000 to 27 000 in the first-stage analysis.

We found no association between the three novel genes identified in this study and T2D in European populations (DIA-GRAMv3) (15). We think there could be three sources for the lack of evidence for associations in DIAGRAM. First, in general, the difference in allele frequencies would cause discrepancy between the association results among different ethnic groups. A lower allele frequency reduces statistical power to detect association. However, we found no apparent differences in the allele frequencies of the top SNPs and those in LD with the top SNP between the East Asian populations and European populations. Taking into account that the DIAGRAM conducted the largest GWAS for T2D in European populations covering ∼2.5 million SNPs, it is unlikely that the DIAGRAM did not find our loci due to the lack of statistical power. Secondly and most likely is dissimilarities in the patterns of LD among populations, which could lead to a substantial difference in the strength of LD between a causal variant and its proxy SNP between our population and European populations. While in *GPSM1* locus, which showed modest association with T2D in DIAGRAMv3 (Supplementary Material, Fig. S3B), CEU and JPT + CHB show quite similar haplotype structure using HapMap Project data (Supplementary Material, Fig. S6B), while the haplotype structure around the other two loci appears to be notably different between the two population samples (Supplementary Material, Fig. S6A and S6C). In particular, the region around *SLC16A13* shows a large block in which many SNPs are in complete LD (D′ = 1). Further examination using the HapMap project's phased haplotype browser shows a

single high-frequency haplotype extending roughly 20 kb in each direction from rs321457 in the JPT or CHB haplotypes that is not present in the CEU haplotypes (data not shown). Lastly and least likely, it is possible that a difference in effect size could lead to a situation where a disease locus would be discovered in a particular group with a strong effect. But no such locus, at least for T2D, has been reported so far, and effect sizes generally appear to be comparable among different ethnic groups. Further study will be needed to clarify whether three loci detected in this study are specific to East Asians.

Of the three newly identified loci, rs791595 is located between *MIR129-1* and *LEP*. The coding product of *LEP*, leptin, plays a critical role in the regulation of body weight by inhibiting food intake and stimulating energy expenditure, and its deficiency in mice and humans causes morbid obesity and diabetes (27,28). Leptin is a hormone produced and secreted by white adipose tissue, and its circulating levels are closely related to body fat mass (29). Thus, *LEP* is one of the most plausible candidates for a susceptibility gene of obesity and T2D, although we could not find any association with BMI in the controls, and adjustment for BMI did not influence the strength of the association with T2D ($P = 9.42 \times 10^{-9}$). Instead, we did find rs791595 to be significantly associated with the homeostasis model assessment of insulin resistance (HOMA-IR) index (30) ($P = 0.005$) (Supplementary Material, Table S5). As for *MIR129−1*, miR129, the mature product of *MIR129−1*, has been reported to be up-regulated in bladder cancers in accordance with the down regulation of its targeted genes *SOX4* and *GALNT1*, which are involved in cell death processes (31). However, whether miR129 has any role in tissues relevant to T2D is unknown. The Genomic Evolutionary Rate Profiling scores for the position of rs791595 near *MIR129-LEP* is 2.96, indicating that this site may be under evolutionary constraint.

The coding product of G-protein signaling modulator 1 isoform *(GPSM1)* influences the basal activity of G-protein signaling systems through interaction with G-protein subunits (32). *Gpsm1* null mice have a lean phenotype with reduced fat mass and increased nocturnal energy expenditure (33), suggesting that *GPSM1* is a biologically plausible obesity gene. We could not find any association with BMI ($P = 0.11$) and adjustment for BMI did not influence the strength of the association with T2D.

SLC16A13 encodes solute carrier family 16, member 13, which is one of the monocarboxylate transporters (MCTs) (34). The first four MCTs (MCT1-4) catalyze the transport of monocarboxylates, such as lactate and pyruvate, but the functions of MCT13 (encoded by *SLC16A13*) and MCT11 (encoded by *SLC16A11,* which is located adjacent to *SLC16A13*) are unknown, except for one report that showed that intestinal expression of *SLC16A13* was up-regulated by peroxisome proliferator-activated receptor-α agonists (35). Rs312457 in *SLC16A13* is in perfect LD with rs17203120 (RAF = 0.079; $P = 6.71 \times 10^{-6}$, OR = 1.19), which is located between *SLC16A13* and *SLC16A11*. There is evidence that rs17203120 is associated with the expression level of *SLC16A11* in lymphoblastoid cell lines (36).

There was no low-frequency variant with MAF ranging from 1 to 5% that reached genome-wide significance among 967 419 variants tested for association with T2D in this study. The present study yielded ∼80% power to detect variants with an

MAF of 1% and OR of 1.6 and retained 80% power to detect association for variants with an MAF of 1% and OR of 2.0 even when imputed with a minimum quality metric r^2 of 0.3, which was sufficient for detecting low-frequency variants with a relatively large effect. Further studies that use newer data from the 1000 Genomes Project or large-scale meta-analyses across populations will be required to build on these results and to further elucidate the global genetic architecture of T2D.

MATERIALS AND METHODS

Study design

The present study was comprised of a three-stage analysis: a discovery stage (first stage), *de novo* genotyping and *in silico* follow-up (second stage), and *de novo* genotyping (third stage) analysis. The discovery stage was a GWAS that tested directly genotyped and imputed variants for association with T2D. We phased SNPs from the Illumina 610K SNP array and performed imputation using the 1000 Genomes East Asian reference panel. We selected 39 loci that had a top SNP with a $P < 1 \times 10^{-4}$ (adjusted for age, sex, BMI and the first four principal components from PCA), and filtered out 19 loci that were located in or near genes that were found to be associated with T2D in previous GWAS. The remaining 20 top SNPs were subsequently investigated for association with T2D by *de novo* genotyping analysis or *in silico* imputation analysis. We combined the results of the discovery and follow-up analyses to identify new T2D loci. An outline of the study design is shown in Supplementary Material, Figure S1.

Subjects

Subjects for the discovery analysis [BioBank Japan (BBJ) 1; 5976 cases with T2D and 20 829 controls] were recruited at several medical institutions in Japan ([37,38]). The follow-up analysis consisted of five studies, which were BBJ2, Shanghai Jiao Tong University (SJTU), Singapore Diabetes Cohort Study (SDCS)/Singapore Prospective Study Program (SP2), Singapore Chinese Eye Study (SCES) and Chinese University of Hong Kong (CUHK). We obtained a total of 2799 cases and 3793 controls for *in silico* follow-up analysis and a total of 10 319 cases and 6795 controls for the follow-up analysis by *de novo* genotyping. Samples in the validation analysis included an independent case–control sample (11 298 cases and 3397 controls) from subjects enrolled in the BioBank Japan (BBJ3). T2D cases were from individuals registered as having T2D. Diabetes was diagnosed according to the WHO criteria. The exclusion criteria for cases were individuals positive for antibody to glutamic acid decarboxylase (GAD) or those with diabetes due to liver dysfunction, steroids and other drugs that might raise glucose levels, malignancy or monogenic disorder known to cause diabetes. Control subjects were healthy volunteers confirmed at annual health check-up or individuals not having T2D but with diseases other than T2D: bronchial asthma, myocardial infarction, breast cancer, Basedow's disease, cerebral infarction, cerebral aneurism, osteoporosis, heart failure, unstable angina, pollinosis, arteriosclerosis obliterance, emphysema or atopic dermatitis. Note that BBJ1, BBJ2 and BBJ3 did not overlap for samples. Ethnicity was self-reported by the enrolled

individuals. For each study, approval was obtained from the appropriate institutional review boards of the participating institutions, and a written informed consent was obtained from all participants.

Genotyping and imputation: in the discovery (BBJ1) and *in silico* follow-up analyses (SDCS/SP2, SCES, and CUHK), genotyping was done with genome-wide SNP arrays. In the *de novo* follow-up analysis, genotyping was carried out by using a multiplex polymerase chain reaction invader assay (BBJ2) and Mass ARRAY (SJTU). The typing platforms and quality control methods for each study are described in Supplementary Material, Table S1. We included SNPs from the SNP array for imputation and the association analysis with a call rate of ≥ 0.99 and a Hardy–Weinberg equilibrium (HWE) $P \geq 1 \times 10^{-6}$, selecting 6 209 637 imputed SNPs with an MAF of ≥ 0.01 and r^2 higher than a set of MAF specific thresholds as described in the previous study: MAF $0-0.1 = 0.75$, MAF $0.1-0.2 = 0.70$, MAF $0.2-0.3 = 0.66$, MAF $0.3-0.4 = 0.60$, MAF $0.4-0.5 = 0.55$ for the discovery analysis. In the discovery and *in silico* analyses, SNP imputation was done using 572 East Asian haplotypes (194 CHB, 200 CHS and 178 JPT) from the 1000 Genomes Project data (June 2011 release).

eQTL

Potential candidates for association with T2D were pursued with eQTL studies in available datasets in lymphoblastoid cell lines ([36]) and a liver tissue gene expression database ([39]).

Statistical analysis

Associations between SNPs and T2D were tested by logistic regression analysis using an additive model with or without adjustment for age, sex, BMI and the first four principal components from PCA. A quantile–quantile plot was constructed by plotting the distribution of the observed *P*-values for the SNPs against the theoretical distribution of the expected *P*-values for T2D. In the discovery analysis, the genomic control inflation factor (λ) was calculated as the median χ^2 statistic divided by 0.456. Meta-analysis was performed by an inverse variance method assuming fixed effects using *R* software. Quantitative trait analyses were done for BMI, FPG, HbA1c, log-transformed HOMA-β, log-transformed HOMA-IR, total cholesterol, HDL cholesterol and triglycerides by multiple linear regression analysis, employing an additive association model with or without adjustment for the relevant covariates. The power of detecting previously reported loci in the present study was estimated by using QUANTO, employing the RAF and the sample sizes in the discovery stage, the reported ORs, an assumed T2D prevalence of 10%, and $\alpha = 0.05$. For the analysis of lean versus obese subjects, samples were dichotomized into lean (BMI < 25) and obese (BMI ≥ 25) groups and SNPs analyzed for each group using a logistic regression base model adjusted for gender, age and four principal components. Heterogeneity between lean and obese samples was calculated as a *z*-score using the beta and standard error (s.e.) from the logistic regression estimates as: $z = (\text{beta.lean} - \text{beta.obese})/\text{sqrt (s.e.lean}^2 + \text{s.e.obese}^2)$. To test whether the reduced sample size of the obese sample group could affect its observed decrease in significance, we performed a re-sampling-based analysis, whereby samples with

sizes matching that of the obese case and control sizes were drawn without replacement 10 000 times from the complete set of samples. We estimated the probability of observing the obese group's *P*-value due to random sample size differences using R's empirical distribution function.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

DIAGRAMv3 study were downloaded via the website at http:// diagram-consortium.org/index.html.

REFERENCES

1. Ashcroft, F.M. and Rorsman, P. (2012) Diabetes mellitus and the β cell: the last ten years. *Cell*, **148**, 1160–1171.
2. Samuel, V.T. and Shulman, G.I. (2012) Mechanisms for insulin resistance: common threads and missing links. *Cell*, **148**, 852–871.
3. McCarthy, M.I. (2010) Genomics, type 2 diabetes, and obesity. *N. Engl. J. Med.*, **363**, 2339–2350.
4. Prokopenko, I., McCarthy, M.I. and Lindgren, C.M. (2008) Type 2 diabetes: new genes, new understanding. *Trends Genet.*, **24**, 613–621.
5. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
6. 1000 Genomes Project ConsortiumAbecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. and McVean, G.A. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
7. 1000 Genomes Project ConsortiumAbecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. and McVean, G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
8. Huang, J., Ellinghaus, D., Franke, A., Howie, B. and Li, Y. (2012) 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 data. *Eur. J. Hum. Genet.*, **20**, 801–805.
9. Yang, W., Lu, J., Weng, J., Jia, W., Ji, L., Xiao, J., Shan, Z., Liu, J., Tian, H., Ji, Q. *et al.* (2010) Prevalence of diabetes among men and women in China. *N. Engl. J. Med.*, **362**, 1090–1101.
10. Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S. *et al.* (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, **445**, 881–885.
11. Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U. *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**, 1341–1345.
12. Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
13. Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G. *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.
14. Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G. *et al.* (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.*, **42**, 579–589.
15. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A. *et al.* (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.*, **44**, 981–990.
16. Yasuda, K., Miyake, K., Horikawa, Y., Hara, K., Osawa, H., Furuta, H., Hirota, Y., Mori, H., Jonsson, A., Sato, Y. *et al.* (2008) Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat. Genet.*, **40**, 1092–1097.
17. Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., Ng, D.P., Holmkvist, J., Borch-Johnsen, K., Jørgensen, T. *et al.* (2008) SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asians and European populations. *Nat. Genet.*, **40**, 1098–1102.
18. Yamauchi, T., Hara, K., Maeda, S., Yasuda, K., Takahashi, A., Horikoshi, M., Nakamura, M., Fujita, H., Grarup, N., Cauchi, S. *et al.* (2010) A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B. *Nat. Genet.*, **42**, 864–868.
19. Kooner, J.S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., Been, L.F., Chia, K.S., Dimas, A.S., Hassanali, N. *et al.* (2011) Genome-wide

association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.*, **43**, 984–989.

20. Cho, Y.S., Chen, C.H., Hu, C., Long, J., Ong, R.T., Sim, X., Takeuchi, F., Wu, Y., Go, M.J., Yamauchi, T. *et al.* (2011) Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in East Asians. *Nat. Genet.*, **44**, 67–72.

21. Imamura, M., Maeda, S., Yamauchi, T., Hara, K., Yasuda, K., Morizono, T., Takahashi, A., Horikoshi, M., Nakamura, M., Fujita, H. *et al.* (2012) A single-nucleotide polymorphism in ANK1 is associated with susceptibility to type 2 diabetes in Japanese populations. *Hum. Mol. Genet.*, **21**, 3042–3049.

22. Saxena, R., Saleheen, D., Been, L.F., Garavito, M.L., Braun, T., Bjonnes, A., Young, R., Ho, W.K., Rasheed, A., Frossard, P. *et al.* (2013) Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in sikhs of punjabi origin from India. *Diabetes*, **62**, 1746–1755.

23. Scott, R.A., Lagou, V., Welch, R.P., Wheeler, E., Montasser, M.E., Luan, J., Mägi, R., Strawbridge, R.J., Rehnberg, E., Gustafsson, S. *et al.* (2012) Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.*, **44**, 991–1005.

24. Huyghe, J.R., Jackson, A.U., Fogarty, M.P., Buchkovich, M.L., Stančáková, A., Stringham, H.M., Sim, X., Yang, L., Fuchsberger, C., Cederberg, H. *et al.* (2013) Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.*, **45**, 197–201.

25. Perry, J.R., Voight, B.F., Yengo, L., Amin, N., Dupuis, J., Ganser, M., Grallert, H., Navarro, P., Li, M., Qi, L. *et al.* (2012) Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases. *PLoS Genet.*, **8**, e1002741.

26. Timpson, N.J., Lindgren, C.M., Weedon, M.N., Randall, J., Ouwehand, W.H., Strachan, D.P., Rayner, N.W., Walker, M., Hitman, G.A., Doney, A.S. *et al.* (2009) Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data. *Diabetes*, **58**, 505–510.

27. Zhang, Y., Proenca, R., Maffei, M., Barone, M., Leopold, L. and Friedman, J.M. (1994) Positional cloning of the mouse obese gene and its human homologue. *Nature*, **372**, 425–432.

28. Montague, C.T., Farooqi, I.S., Whitehead, J.P., Soos, M.A., Rau, H., Wareham, N.J., Sewter, C.P., Digby, J.E., Mohammed, S.N., Hurst, J.A. *et al.* (1997) Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature*, **387**, 903–908.

29. Maffei, M., Halaas, J., Ravussin, E., Pratley, R.E., Lee, G.H., Zhang, Y., Fei, H., Kim, S., Lallone, R., Ranganathan, S. *et al.* (1995) Leptin levels in human and rodent: measurement of plasma leptin and ob RNA in obese and weight-reduced subjects. *Nat. Med.*, **1**, 1155–1161.

30. Matthews, D.R., Hosker, J.P., Rudenski, A.S., Naylor, B.A., Treacher, D.F. and Turner, R.C. (1985) Homeostasis model assessment: insulin resistance and B cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*, **28**, 412–419.

31. Dyrskjøt, L., Ostenfeld, M.S., Bramsen, J.B., Silahtaroglu, A.N., Lamy, P., Ramanathan, R., Fristrup, N., Jensen, J.L., Andersen, C.L., Zieger, K. *et al.* (2009) Genomic profiling of microRNAs in bladder cancer: miR-129 is associated with poor outcome and promotes cell death in vitro. *Cancer Res.*, **69**, 4851–4860.

32. Pizzinat, N., Takesono, A. and Lanier, S.M. (2001) Identification of a truncated form of the G-protein regulator AGS3 in heart that lacks the tetratricopeptide repeat domains. *J. Biol. Chem.*, **276**, 16601–16610.

33. Blumer, J.B., Lord, K., Saunders, T.L., Pacchioni, A., Black, C., Lazartigues, E., Varner, K.J., Gettys, T.W. and Lanier, S.M. (2008) Activator of G protein signaling 3 null mice: unexpected alterations in metabolic and cardiovascular function. *Endocrinology*, **149**, 3842–3849.

34. Halestrap, A.P. and Meredith, D. (2004) The SLC16 gene family-from monocarboxylate transporters (MCTs) to aromatic amino acid transporters and beyond. *Pflugers Arch.*, **447**, 619–628.

35. Hirai, T., Fukui, Y. and Motojima, K. (2007) PPAR Alpha agonists positively and negatively regulate the expression of several nutrient/drug transporters in mouse small intestine. *Biol. Pharm. Bull.*, **30**, 2185–2190.

36. Veyrieras, J.B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M. and Pritchard, J.K. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.*, **4**, e1000214.

37. Okada, Y., Sim, X., Go, M.J., Wu, J.Y., Gu, D., Takeuchi, F., Takahashi, A., Maeda, S., Tsunoda, T., Chen, P. *et al.* (2012) Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. *Nat. Genet.*, **44**, 904–909.

38. Okada, Y., Kubo, M., Ohmiya, H., Takahashi, A., Kumasaka, N., Hosono, N., Maeda, S., Wen, W., Dorajoo, R., Go, M.J. *et al.* (2012) Common variants at CDKAL1 and KLF9 are associated with body mass index in east Asian populations. *Nat. Genet.*, **44**, 302–306.

39. Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C. *et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, **6**, e107.