Supporting Information

Kodama et al. 10.1073/pnas.1114513109

SI Materials and Methods

eGWAS. All T2D-related genome-wide microarray experiments used for this meta-analysis were collected from three public data sources: The NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) was searched by using the keywords: "diabetes," "diabetic," "NIDDM," or "non-insulin-dependent" (from inception until July 2009). From the list of all identified GEO Series (GSE), we selected gene expression microarray studies that met the following criteria: (*i*) investigating about "type 2 diabetes" (studies related to type 1 diabetes, maturity-onset diabetes of the young, drug-induced diabetes, gestational diabetes, diabetic complications, and other specific diseases causing diabetes were excluded from our data analysis); (*ii*) using T2D-relevant tissues, such as adipose, liver, muscle, pancreatic islets, kidney, and hypothalamus; (*iii*) using samples from human and rodent models. We also selected microarray studies involving T2D from the supplied list in the Diabetes Genome Anatomy Project (DGAP; www.diabetesgenome.org) and the Nuclear Receptor Signaling Atlas (NURSA; www.nursa.org) on April 2008. We then reassembled microarray samples in the selected studies into independent data set comparisons (experiments): a curated collection of pathophysiologically comparable microarray samples, always comparing diabetes samples to pathophysiologically matched control samples. This study selection and sample assembly were performed manually by two independent investigators by reviewing the submitter-supplied record and the abstract and full-text of original research articles. Discrepancies in eligibility were discussed between reviewers until agreement was achieved. We subsequently downloaded processed (normalized by the original methods selected by the original submitters) microarray data corresponding to the identified experiments from the three public repositories. Data were stored in a Microsoft Excel 2007 spreadsheet (Microsoft Corporation). Execution of this procedure yielded 130 independent experiments with a total of 1,175 samples (591 diabetic cases and 584 nondiabetic controls) (Tables S1 and S2).

To estimate differences between groups of samples from diabetic subjects and groups representing control, the downloaded microarray data were reanalyzed by using Significance Analysis of Microarrays software (SAM Excel Add-in; www-stat.stanford.edu/~tibs/SAM) with "two class unpaired" design (or with "paired" or "one class" option if necessary) when each group had three or more samples (1). For each gene in every microarray experiment, we calculated a *d* score (d_i), which denotes the standardized change in gene expression:

$$d_i = \frac{\bar{x}_{i-t2d} - \bar{x}_{i-control}}{S_i + S_0},$$

where $\bar{x}_{i-t2d}$ is the mean expression level of gene *i* in group T2D, $\bar{x}_{i-control}$ is the mean expression level of gene *i* in group control, S_i is the SD for the numerator calculation, and S_0 is a small positive constant, and we also calculated fold change as the ratio between the signal averages of control and experimental samples. When each group had only one or two samples, we calculated only a fold-change value for each gene in these microarray experiments.

We considered genes to be significantly dysregulated with either an absolute value of the *d* score ≥2 or a fold change ≥2 between controls and cases (ref; "SAM": Users guide and technical documents; www-stat.stanford.edu/~tibs/SAM/sam.pdf) (2). We then converted all probe identifiers across the various microarray platforms for mouse, rat, and human to the latest human Entrez Gene identifiers by using our published AILUN system (http://ailun.stanford.edu; its Cross-species Mapping uses the NCBI Homologene resource) (3). Gene expression profiles were assigned in our eGWAS database (Excel spreadsheet format) according to the standardized (human) Entrez Gene ID. There were 24,898 genes in the database in total.

For each of the 24,898 genes, we counted the observed number of microarray experiments in which each gene was significantly dysregulated, using the Excel "pivot table" function. We then aligned the number of positive/negative experiments for every one of genes, summed the number of positive/negative experiments for all other genes and calculated χ^2 formula by using these data in the one row of Excel spreadsheet, and ran the χ^2 test to calculate χ^2 value and its corresponding *P* value for all 24,898 genes (rows). As an alternative methodology, we also conducted a Fisher's exact test by using the same data of contingency tables for all of the genes, using the R statistical package (http://www.r-project.org) (4). We subsequently ranked all of the genes according to their *P* values ($-\log_{10}(P)$). We also conducted a weighted Z-method (5) using the same experiments in our eGWAS database and confirmed that *CD44* is still the top gene in this alternative method based on a different concept.

Animal Experiments. Mice for breeding, C57BL/6J wild-type (diabetes-prone) and *CD44*-deficient mice backcrossed to C57BL/6J for at least 10 generations (B6.Cg-*Cd44*^{tm1Hbg}/J), were obtained from The Jackson Laboratory. To examine the role of CD44 for the development of insulin resistance, male *CD44*$^{-/-}$ and littermate wild-type *CD44*$^{+/+}$ mice (8 wk of age) were fed diets containing either 12% kcal fat (normal-fat diet; NFD) (CE-2; CLEA Japan) or 60% kcal fat (high-fat diet; HFD) (D12492; Research Diets) for 12 wk. To assess CD44 and SPP1 expression in obese adipose tissue, C57BL/6J male mice (8 wk of age) were fed either a NFD or a HFD for 16 wk. To investigate the therapeutic effect of Anti-CD44 antibody on HFD-induced diabetes, C57BL/6J male mice (8 wk of age) were fed a HFD for 18 wk. Weight gain was monitored by weighing mice weekly. Mice had free access to autoclaved water. Mice were housed in a barrier facility under specific pathogen-free conditions. The Animal Care and Use Committee of Kitasato University approved all animal experiments.

Metabolic measurements were conducted on male *CD44*$^{-/-}$ and littermate wild-type *CD44*$^{+/+}$ mice after feeding a NFD or HFD for 12 wk. Fasting blood glucose levels were measured after a 14-h overnight fast. Glucose tolerance tests (GTT) were performed by giving glucose (2 g/kg of body weight) intraperitoneally after a 14-h overnight fast. Venous blood was obtained for measurement of blood glucose and serum insulin levels at 0, 15, 30, 60, 90, and 120 min after the injection. Insulin tolerance tests (ITT) were performed by giving insulin (1.0 unit/kg; Actrapid, Novo Nordisk) intraperitoneally after a 4-h fast. Venous blood was obtained for measurement of blood glucose at 0, 30, and 45 min after the injection. Blood glucose concentration was determined with a glucose meter (Medisafe-Mini; Terumo). Serum insulin levels were measured with an ultrasensitive mouse insulin ELISA kit (Morinaga Institute of Biological Science).

CD44 and *SPP1* mRNA expression in visceral adipose tissue was measured by quantitative real-time RT-PCR. At 24 wk of age, epididymal white adipose tissue (EWAT) was removed from wild-type C57BL/6J mice fed either a NFD or HFD, immediately frozen by liquid nitrogen and stored in −80 °C freezer. Total RNA of EWAT was isolated by using the TRIzol RNA isolation

method (Invitrogen) and purified with the RNeasy Mini Kit spin columns (QIAGEN) according to the manufacturer's instructions. Quantity and quality of isolated RNA was determined by spectrophotometric measurements at optical density (OD) 260 and OD280. The integrity of isolated RNA was checked by agarose gel electrophoresis. Two micrograms of RNA was reverse-transcribed to cDNA by using a first-strand cDNA Synthesis Kit for RT-PCR (AMV) (Roche Diagnostics). PCRs were performed with the LightCycler FastStart DNA master SYBR Green I system (Roche Diagnostics). Each sample was analyzed in triplicate and normalized to the values for glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) mRNA expression. Mouse primer sequences used were as follows: CD44, CCA GGC TTT CAA CAG TAC CTT ACC (forward), CTG AGG CAT TGA AGC AAT ATG TGT C (reverse); SPP1, TGA TAG CTT GGC TTA TGG ACT GA (forward), CCA CTG AAC TGA GAA ATG AGC AG (reverse); GAPDH, TGA ACG GGA AGC TCA CTG G (forward), TCC ACC ACC CTG TTG CTG TA (reverse).

Anti-CD44 antibody treatment was performed on C57BL/6J male mice fed a HFD for 18 wk. Mice were injected intraperitoneally with purified rat anti-mouse CD44 (IM7; 553131, BD Pharmingen) or purified rat IgG2b, κ isotype control (A95-1; 559478, BD Pharmingen) for 8 d (100 μg at day 0 and 50 μg at day 1–7). Morning blood glucose was measured at day 0, 1, 3, 5, and 7 during the treatment.

Human Studies. Venous peripheral blood samples were collected from human subjects who went through a 75g oral glucose tolerance test after an overnight fast [$n = 55$: age (yr), 60.3 ± 15; sex (M/F), 36/19; BMI (kg/m^2), 23.2 ± 4.3]. HbA1c (glycosylated hemoglobin) was measured in Japan Diabetes Society (JDS)-HbA1c units by using an ion-exchange HPLCy method. HbA1c was converted to National Glycohemoglobin Standardization Program (NGSP) levels by the formula HbA1c (%) (NGSP) = HbA1c (JDS) (%) + 0.4%, considering the relational expression of HbA1c (JDS) (%) measured by the previous Japanese standard substance and measurement methods and HbA1c (NGSP) (6), and these converted values were used throughout the study. We then calculated homeostasis model assessment as an index of insulin resistance [HOMA-IR = fasting plasma insulin (μU/mL) × fasting plasma glucose (mg/dL)/405] as described (7). Serum sCD44std (standard soluble CD44) and SPP1 concentrations were determined by using a quantitative ELISA technique (sCD44std ELISA, Bender MedSystems; Human Osteopontin Quantikine ELISA, R&D Systems).

Informed consent was obtained from all of the subjects enrolled in this study, and the protocol was approved by the ethics committee of the University of Tokyo.

Immunohistochemistry. For histological analysis of CD44 expression in adipose tissue, EWAT was removed from mouse models, and omental adipose tissue obtained from consented donors undergoing elective gastric bypass surgery (lot number. OM020304B) was purchased from Zen-Bio. Formalin-fixed paraffin-embedded sections (5 μm thick) were stained for CD44 by using the DAKO CSA II signal amplification system. After a peroxidase and protein block, the slides were incubated overnight (4 °C) with mouse monoclonal antibody against CD44 at 1:50 dilution (DF1485/sc-7297; Santa Cruz Biotechnology), followed by reactions with anti-mouse immunoglobulins-HRP, an amplification reagent, and anti-fluorescein-HRP. 3,3′-Diaminobenzidine (DAB) was used as a chromogen, and the sections were counterstained with hematoxylin.

Analysis of inflammatory cell (macrophage) content in EWAT was performed on tissue pads isolated from model mice. Formalin-fixed paraffin-embedded sections were incubated overnight (4 °C) with primary antibody: Purified Anti Mouse MAC-2 (macrophage marker) Monoclonal Antibody (CL8942AP, 1:100; Cedarlane Laboratories), and stained using Histofine Simple Stain Mouse MAX-PO secondary antibody (Nichirei Biosciences) with a DAB solution and counterstained with hematoxylin.

In anti-CD44 antibody-treated and control mice, adipose inflammation was quantified as the density of crown-like structures (CLSs). The total number of CLSs was counted in five random fields (magnification: 100×) of each mouse in a blinded manner, and the average number of CLSs was calculated in all animals in each group, by creating a digitized image with a BIOQUANT Image Analysis System (BIOQUANT Image Analysis).

1. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98:5116–5121.
2. Orlova MO, et al. (2006) Constitutive differences in gene expression profiles parallel genetic patterns of susceptibility to tuberculosis in mice. *Infect Immun* 74:3668–3672.
3. Chen R, Li L, Butte AJ (2007) AILUN: Reannotating gene expression data automatically. *Nat Methods* 4:879.
4. R Development Core Team (2004) R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna), Version 1.9.1.
5. Whitlock MC (2005) Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *J Evol Biol* 18:1368–1373.
6. Seino Y, et al. (2010) Report of the Committee on the Classification and Diagnostic Criteria of Diabetes Mellitus. *J Diabetes Invest* 1:212–228.
7. Matthews DR, et al. (1985) Homeostasis model assessment: Insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* 28:412–419.

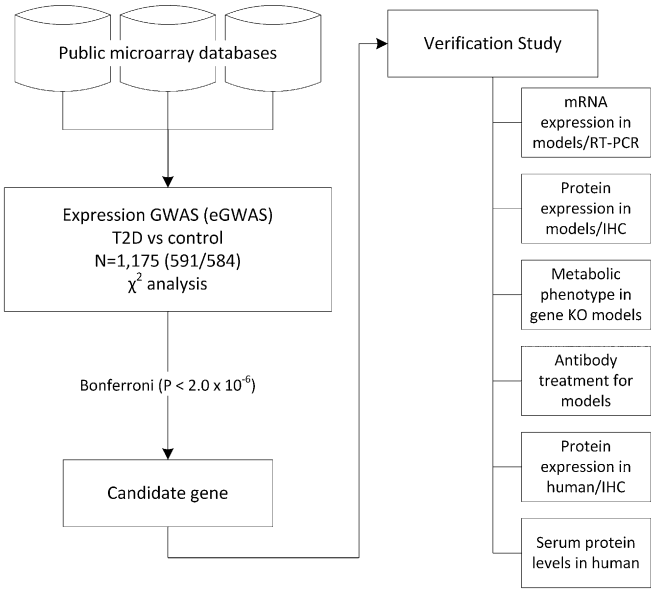

Fig. S1. Study design. eGWAS for type 2 diabetes (T2D) was carried out in 1,175 microarray samples collected from public databases. *P* values ($-\log_{10}(P)$) were calculated by comparing dysregulation distribution of genes between T2D and control microarrays. Our T2D candidate gene extracted from eGWAS was verified by functional studies in mouse models and human subjects.

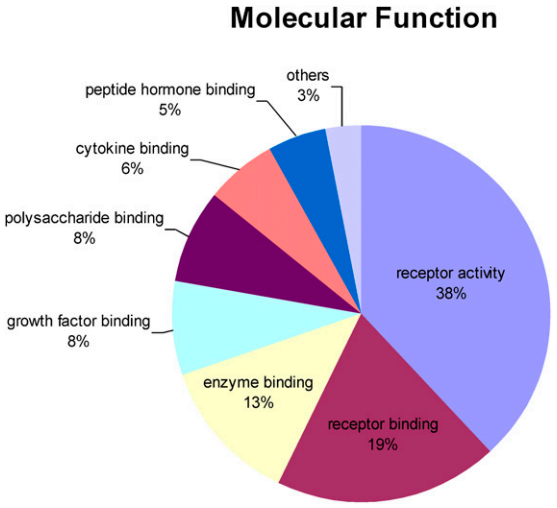

Fig. S2. Distribution of Gene Ontology (GO) molecular function annotations for the 127 top-ranked genes in our T2D eGWAS (Bonferroni threshold, $P < 2.0 \times 10^{-6}$). The graph shows the percentage of GO terms on the total number of annotated gene products with significantly enriched GO terms.

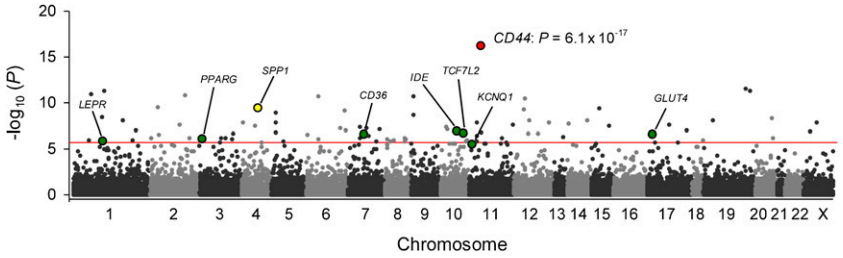

Fig. S3. eGWAS for T2D using a Fisher's exact test. Plot of $-\log_{10}$ (*P* value) (*y* axis) by chromosomal position (*x* axis). *P* values for each gene were calculated by using a Fisher's exact test. *CD44* is the top gene in both the Fisher's exact test and the χ^2 analysis (Fig. 1). The red line indicates the Bonferroni threshold ($P = 2.0 \times 10^{-6}$). The green dots indicate the several well known T2D-susceptibility genes significant in the χ^2 analysis (Fig. 1). All of the well-known genes from Fig. 1 except *KCNQ1* were still significant in the Fisher's exact test ($P < 2.0 \times 10^{-6}$).

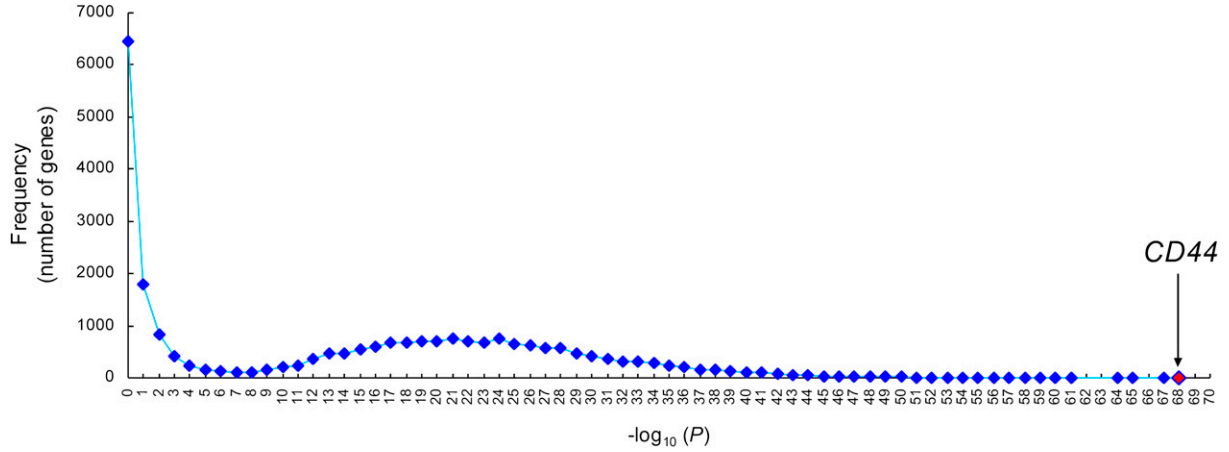

Fig. S4. Liptak–Stouffer's weighted Z-method for T2D case-control microarray experiments. The frequency distribution of -$\log_{10}$ (combined *P* value) is shown. The combined *P* values for each gene were calculated by using a weighted Z-method; *P* values were computed by using a one-tailed *t* test for each gene in each of the 110 experiments with four or more samples, and then the *P* values were converted to Z-scores. Subsequently, the combined Z-scores (Zw) across all of the experiments were calculated for each gene, using a weighted Z-method, by weighting each experiment by its sample size (degrees of freedom; d.f.). Then the combined *P* values for each gene were obtained by converting the weighted Z-scores (Zw) into two-tailed *P* values. The -$\log_{10}$ (combined *P* values) were rounded into the nearest integer, and then the frequency distribution was determined. We confirmed that *CD44* is still the top gene in this alternative method (Zw = 17.48, $P = 2.0 \times 10^{-68}$).

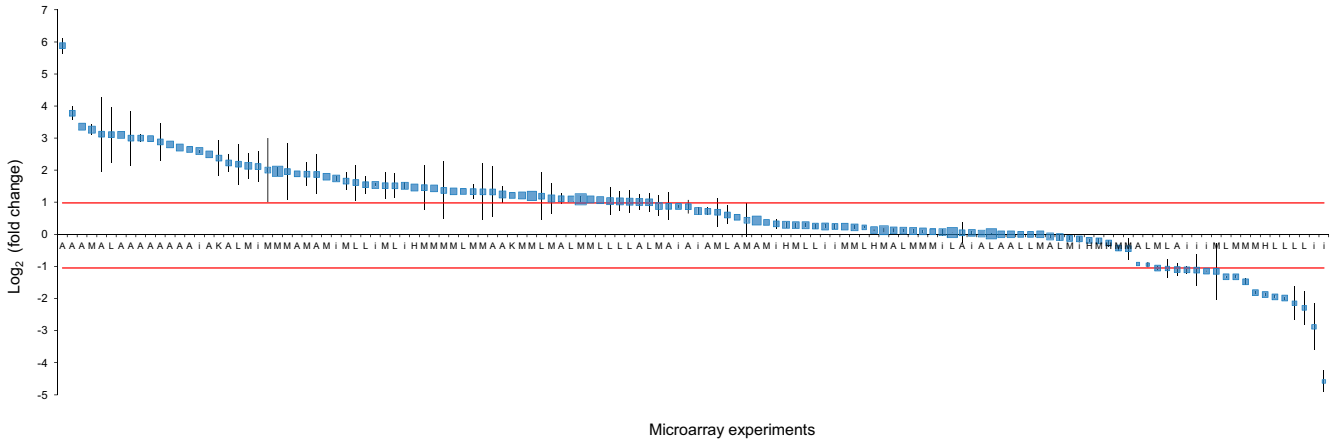

Fig. S5. The distribution of fold change for *CD44* across all of the experiments in T2D eGWAS. Plot of $\log_2$ (fold change) for *CD44* (*y* axis; error bar = 95% confidence intervals) in microarray experiments (*x* axis; A, adipose tissue; H, hypothalamus; i, pancreatic islets; K, kidney; L, liver; M, muscle). The area of the box is proportional to the sample size of each experiment. The red line indicates the significance threshold (fold change = 2). *CD44* mRNA was the most highly differentially expressed gene across experiments and was also the top-most highly up-regulated mRNA in diabetes. *CD44* mRNA was more highly differentially expressed in adipose tissue than other tissues.

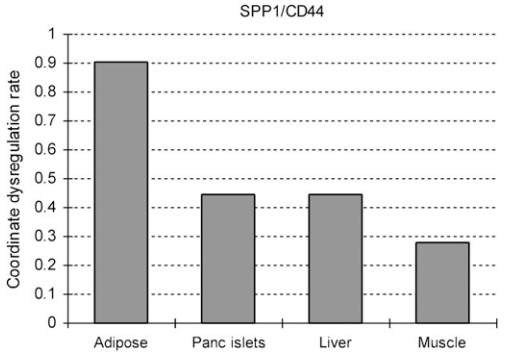

Fig. S6. Coordinate dysregulation rate between *CD44* and *SPP1*. The coordinate dysregulation rate (CDR) is calculated as the probability of *SPP1* dysregulation, given the occurrence of *CD44* dysregulation:

$$CDR = P(SPP1|CD44) = \frac{n(SPP1 \cap CD44)}{n(CD44)}.$$

Table S1. List of microarray experiments in our T2D eGWAS

Tables S1

Table S2. List of microarray platforms in the 130 experiments

Tables S2

Table S3. List of top 127 genes in our T2D eGWAS (Bonferroni threshold, $P < 2.0 \times 10^{-6}$)

Tables S3