doi:10.1038/nature12828

Supplementary Methods

Study participants

Diabetes in Mexico Study (DMS):
Individuals were enrolled in the study, recruited from two tertiary level institutions (IMSS and ISSSTE) located in Mexico City. The diagnosis of T2D was made based on ADA criteria. 811 unrelated healthy subjects older than 45 years and with fasting glucose levels below 100 mg/dL were classified as controls. 569 unrelated individuals, older than 18 years, with either previous T2D diagnosis or fasting glucose levels above 125 mg/dL were included as T2D cases. Individuals with fasting glycemia between 100-125 mg/dL were excluded. Informed consent was obtained from all participants. The study was conducted with the approval of the Ethics and Research Committees of all institutions involved. Genomic DNA was purified from whole blood samples using a modified salting-out precipitation method (Gentra Puregene, Qiagen Systems, Inc., Valencia, CA, USA).

Mexico City Diabetes Study (MCDS):
The Mexico City Diabetes Study is a population based prospective investigation. All 35-64 years of age men and non-pregnant women residing in the study site (low income neighborhoods equivalent to 6 census tracks with a total population of 15,000 inhabitants) were interviewed and invited to participate in the study. We had a response rate of 67% for the initial exam. Diagnostic criteria for type 2 diabetes were recommended by the ADA. Fasting glucose 126 mg/dL or more or 2 hr post 75 gr of glucose load 200 or more. If a participant was diagnosed as diabetic by a physician and was under pharmacologic therapy for diabetes he was considered as diabetic regardless the blood glucose levels. The study was conducted with the approval of the Ethics and Research Committees of all institutions. Informed consent was obtained from all participants. Genomic DNA was extracted from whole blood using the QIAmp 96 DNA Blood Ki5 (12) (Qiagen, Cat. No. 51162).

Multiethnic Cohort (MEC):
The MEC consists of 215,251 men and women in Hawaii and Los Angeles, and comprises mainly five self-reported racial/ethnic populations: African Americans, Japanese Americans, Latinos, Native Hawaiians and European Americans[30]. Between 1993 and 1996, adults between 45 and 75 years old were enrolled by completing a 26-page, self-administered questionnaire asking detailed information about dietary habits, demographic factors, level of education, personal behaviors, and history of prior medical conditions (e.g., diabetes). Potential cohort members were identified through Department of Motor Vehicles drivers' license files, voter registration files and Health Care Financing Administration data files. In 2001, a short follow-up questionnaire was sent to update information on dietary habits, as well as to obtain information about new diagnoses of medical conditions since recruitment. Between 2003 and 2007, we re-administered a modified version of the baseline questionnaire. All questionnaires inquired about history of diabetes, without specification as to type (1 vs. 2). Between 1995 and 2004, blood specimens were collected from ~67,000 MEC participants at which time a short questionnaire was administered to update certain exposures, and collect current information about medication use.

Cohort members in California are linked each year to the California Office of Statewide Health Planning and Development (OSHPD) hospitalization discharge database which consists of mandatory records of all in-patient hospitalizations at most acute-care facilities in California. Records include information on the principal diagnosis plus up to 24 other diagnoses (coded according to ICD-9), including T1D and T2D. In Hawaii cohort members have been linked with the diabetes care registries for subjects with Hawaii Medical Service Association (HMSA) and Kaiser Permanente Hawaii (KPH) health plans (~90% of the Hawaii population has

one of these two plans). Information from these additional databases has been utilized to assess the percentage of T2D controls (as defined below) with undiagnosed T2D, as well as the percentage of identified diabetes cases with T1D rather than T2D. Based on the OSHPD database <3% of T2D cases had a previous diagnosis of T1D. We did not use these sources to identify T2D cases because they did not include information on diabetes medications, one of our inclusion criteria for cases (see below).

In the MEC, diabetic cases were defined using the following criteria: (a) a self-report of diabetes on the baseline questionnaire, 2nd questionnaire or 3rd questionnaire; and (b) self-report of taking medication for T2D at the time of blood draw; and (c) no diagnosis of T1D in the absence of a T2D diagnosis from the OSHPD (California Residents). Controls were defined as: (a) no self-report of diabetes on any of the questionnaires while having completed a minimum of 2 of the 3 (~80% of controls returned all 3 questionnaires); and (b) no use of medications for T2D at the time of blood draw; and (c) no diabetes diagnosis (type 1 or 2) from the OSHPD, HMSA or KPH registries. To preserve DNA for genetic studies of cancer in the MEC, subjects with an incident cancer diagnosis at time of selection for this study were excluded. Controls were frequency matched to cases on sex, ethnicity and age at entry into the cohort (5-year age groups) and for Latinos, place of birth (U.S. vs. Mexico, South or Central America), oversampling African American, Native Hawaiian and European American controls to increase statistical power. Many of the T2D variants have also been evaluated in studies of cancer in the MEC which allowed for inclusion of additional controls who met the criteria above. Altogether, this study included 2,231 T2D cases and 2,607 controls of Latin American ethnicity. Informed consent was obtained from all participants. The study was conducted with the approval of the Ethics and Research Committees of all institutions. Genomic DNA extraction was done using Qiagen from buffy coat.

UNAM/INCMNSZ Diabetes Study (UIDS):
Cases were recruited at the outpatient diabetes clinic of the Department of Endocrinology and Metabolism of the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ). All Mexican-mestizo individuals were invited to participate in the study. Diagnosis of type 2 diabetes was done following the American Diabetes Association criteria, i.e., fasting plasma glucose values ≥126 mg/dL, current treatment with a hypoglycemic agent, or casual glucose values ≥200 mg/dL.

Control subjects were recruited from a cohort of adults aged 45 years or older among government employees, blue collar workers and subjects seeking for attention in medical units for any condition besides those considered as exclusion criteria (see below). Normoglycemic status was defined as having a fasting plasma glucose concentration < 100 mg/dl and no previous history of hyperglycemia, gestational diabetes or use of metformin.

Patients were interviewed following a standardized questionnaire; it included the medical history, a previously validated, three days food record and a physical activity registry. In addition a blood sample (after 9-12 hours of fasting) was obtained. The questionnaire included demographic, socio-economic and medical history of the patients and their family. Blood pressure, height, waist circumference and weight must be measured in the same visit. For taking blood pressure, systolic and diastolic pressure were recorded using a mercury sphygmomanometer; subjects remained seated and at rest for five minutes before measuring.

Inclusion criteria: Men or women aged 25 years or older, with BMI greater than 20 but lower than 40 kg/m2.

Exclusion criteria: Diabetes, coronary heart disease, stroke, transient ischemic attack, lower limb amputations, alcoholism (more than 10 servings of alcohol per week) or any disease that in opinion of the researcher may limit life expectancy to less than 2 years. Subjects that planned to move out of town permanently during the next three years were also excluded. Pregnant women, individuals with drug addictions, the use of systemic corticosteroids in pharmacologic doses (intravenous, oral or injectable, including injections in the joints) were exclusion criteria also. Replacement dosage of systemic corticosteroids (up 7.5 mg/day of prednisone or 30 mg/day of hydrocortisone or its equivalent; as well as inhaled or topical corticosteroids) was allowed into the study. Other exclusion criteria were: active liver disease (defined as AST (SGOT) or ALT (SGPT) > 2.0 x upper limit of the normal range, alkaline phosphatase (ALK-P) > 1.5 x upper limit of the normal

range or total bilirubin > 1.5 x upper limit of the normal range), significant renal dysfunction (defined as serum creatinine > 1.7 upper limit of the normal range or nephrotic syndrome), any history of malignancy (except for basal cell skin carcinoma) and uncontrolled depression or psychosis.

Informed consent was obtained from all participants. The study was conducted with the approval of the Ethics and Research Committees of all institutions. Genomic DNA was extracted from whole blood using the QIAmp 96 DNA Blood Ki5 (12) (Qiagen, Cat. No. 51162).

SNP genotyping

DNA samples were sent to the Broad Institute and prepared for genetic analysis with two quality control measures. First, DNA quantity was measured by Picogreen, and then all samples with sufficient total DNA and minimum concentrations for downstream activities were genotyped for a set of 24 SNPs using the Sequenom iPLEX Assay. These 24 validated markers include 1 gender assay and 23 SNPs located across the autosomes. The genotypes for these SNPs are used as a quality filter to advance samples, as well as a technical fingerprint validation with the subsequent genome-wide array genotypes. Genotyping was performed at the Broad Institute Genetic Analysis Platform. DNA samples were placed on 96-well plates and genotyped using the Illumina HumanOmni2.5-4v1_B SNP array. Omni genotypes were called using GenomeStudio v2010.3 with the calling algorithm/genotyping module version 1.8.4 using genotype cluster definitions based on study samples. Called genotypes were run through a standard QC pipeline at this stage and only samples passing a call rate threshold of 95%, and passing genetic fingerprint (24 marker panel) and gender concordance were passed on to downstream GWAS QC. SNPs with a GenTrain score <0.6, cluster separation score <0.4 and call rate <97% were considered technical failures at the genotyping laboratory and were automatically deleted before release for further quality control.

Quality control procedure

We included only SNP variants, omitting 543 indel sites, and removed SNPs that had 2% or more missing data in any of the four individual cohorts; subsequently, we removed any samples with 2% or more missing data sites. Gender concordance based on X chromosome heterozygosity rates were performed and discordant samples removed. Sample duplicates were identified and we removed one sample in each pair where the case-control status of the duplicates matched; four duplicate pairs had discordant case-control status and we removed all these samples. We removed SNPs that had < 1% minor allele frequency (MAF) within the full SIGMA dataset. After all these checks, we again removed any samples with 2% or more missing data in the complete SIGMA dataset and 118 SNPs that showed statistically significant difference in frequency between males and females (Z-score ≥ 6). We next checked for structure of missing data among samples by performing principal components analysis[31] (PCA) using missing data status as the value of each site, and we removed outlier samples. We then performed PCA on standard genotypes and, in order to include samples with more uniform ancestry, we removed a small percentage (2.0% n=181) of samples that showed evidence of high African or East Asian ancestry (Supplementary Figure 1). Finally, we detected sample relatives using the PCA-based method *smartrel*[31] and removed one individual in each pair of relatives estimated to have 10% or more relatedness (n=532).

Prior to imputation, we aligned variants to the forward strand as reported by the 1,000 Genomes Project[32]. To do this, we first removing SNPs not present in the 1,000 Genomes Project variant set and SNPs where the allele reported in the genotype data did not match the alleles from the 1,000 Genomes data. We further excluded SNPs in which the frequency difference between the minor allele in the SIGMA data and the AMR population

in 1,000 Genomes is greater than 15%. Finally, we excluded A/T and C/G SNPs with MAF > 35%; sites that were not excluded by these means had strand resolved by consistency between the SIGMA MAF and that of AMR in 1,000 Genomes.

Genotype Imputation

SNP imputation was performed by pre-phasing[33] with HAPI-UR[34] version 1.01 and imputation IMPUTE[35] version 2.2.0 with the 1000 Genomes Phase I integrated variant set[32] (build 37 and haplotype release date in August, 2012) serving as our reference panel. Analysis included all imputed variants with MAF ≥ 1% and info score ≥ .6.

Statistical analyses

GWAS analysis was performed using LTSOFT[36] version 1.0 to convert T2D from a dichotomous trait to a quantitative
liability score using information about disease prevalence stratified by these risk factors. Prevalence of type 2 diabetes (T2D) at different intervals of age and body mass index (BMI) reported by the Mexican National Survey of Health and Nutrition[37] were used as parameters for the model. For each individual, LTSOFT computed the posterior mean of the residual of the liability score, given an individual's disease status, age, and BMI, and the prevalence intervals, and we treated this as a continuous phenotype for performing the association tests via linear regression using PLINK[38] version 1.08p. Sex and the top 2 principal components were also included as fixed covariates in the regression. *P*-values were corrected using genomic control[39], with a λ_{GC} inflation factor of 1.046 . Odds ratios reported throughout are from case-control logistic regression obtained using PLINK with BMI, age, sex, and the top 2 principal components used as covariates. Directional consistency P-value is from a binomial test.

To estimate proportions of Native American ancestry, we used the software ADMIXTURE[40] version 1.22 with *K*=3 clusters. We merged the SIGMA samples with individuals from the Human Genome Diversity Panel (HGDP) dataset[41], including Southern Europeans (Basque, French and Italians), Africans (Mandenka and Yoruba), and Native Americans from Mexico (Pima and Maya); our merged dataset contained 346,490 SNPs. We identified 290 individuals in the SIGMA dataset estimated to have at least 95% Native American ancestry.

Difference in age-of-onset and BMI between carriers of the risk haplotype and non-carriers was performed using SNP rs13342692 to determine carrier status. P-value was calculated using a two-sample t-test analyzing T2D cases that are risk haplotype carriers compared to T2D cases that are non-carriers.

In order to analyze the association between age of onset and SNP rs13342692, we analyzed cases from DMS, UIDS and MCDS cohorts (age of onset was not available in MEC cohort). We classified diabetics who reported being diagnosed before or at age of 45 as early onset (n=752), and all other samples as late onset (n=968). Odds ratios were calculated via logistic regression separately in both age of onset classifications using two randomly selected disjoint sets of controls (both with n=2,164) for each classification. To obtain a p-value, we calculated a Z-score for the difference between the two odds ratios as $z = \frac{\log OR_Y - \log OR_O}{\sqrt{SE_Y^2 + SE_O^2}}$, where OR_Y is the odds ratio for young cases and SE_Y is the standard error for OR_Y, and likewise for OR_O and SE_O.

Local ancestry estimation was performed using LAMP-LD[42] version 1.0. Panels for inference included a diverse collection of 227 Native American samples from Central America and Mexico[41,43], 72 Southern

Europeans from HGDP[41] and 12 Spanish individuals[44], and 109 Yoruba Africans (YRI) from HapMap[45]. Prior to performing local ancestry inference, the panels were merged, yielding an intersected SNP set of 252,402 markers. The panels were then jointly phased using SHAPEIT[46] version 1.532. Next the SNPs from the panels and the SIGMA data were intersected, yielding 235,660 SNPs, and LAMP-LD was run to infer local ancestry.

Replication meta-analyses were performed using inverse standard error weighting of effect sizes in METAL[47] (release date 2011-03-25).

All regional plots were generated using LocusZoom[48].

Testing for heterogeneity of effect by cohort

To examine the potential for heterogeneity of effect by cohort, we performed a retrospective analysis (i.e., using genotype as the target variable) in a logistic regression framework. After including other primary covariates (top two principal components, case-control phenotype, BMI, age), inclusion of any cohort label yielded non-significant p-values at both our novel associated loci ($P \geq 0.50$ at *SLC16A11* and $P \geq 0.77$ at 11p15.5). Thus, cohort-specific effects do not explain genotype differences across cohorts at these loci and there is no evidence for heterogeneity by cohort.

***SLC16A11* haplotype frequencies of SIGMA samples**

To determine haplotype frequencies in the SIGMA samples, we first performed genotyping of three of the associated *SLC16A11* missense variants that were not genotyped on the OMNI 2.5 array (rs117767867, rs75418188, rs75493593) in the MEC individuals. (The fourth missense variant and the associated synonymous SNP (rs13342692, rs13342232) were genotyped on the OMNI 2.5 array.) Separately, we performed imputation in the 17p13.1 region in all SIGMA samples. Correlation between direct genotypes and imputed dosage values for the MEC samples was $r^2 \geq .99$ for all three directly typed variants, and we therefore used imputed dosage values to estimate allele frequencies in the entire SIGMA dataset. As described above, we identified a subset of the SIGMA dataset with ≥ 95% Native American ancestry, and we used the imputed dosage values to estimate allele frequencies in these samples, as reported in the main text and Figure 2a.

Because three of the missense SNPs reside on only one haplotype in the 1,000 Genomes Project data, we utilized their allele frequencies as a surrogate for the "5 SNP" haplotype frequency. Allele frequencies for all three of these SNPs was 29.5% in SIGMA, consistent with their presence on a single haplotype with frequency ~30%. The other two SNPs reside on both the "5 SNP" and "2 SNP" haplotype, and they have frequency 31.5% in the SIGMA samples, consistent with the "2 SNP" haplotype having frequency 31.5% - 29.5% = 2%. In the samples with ≥95% Native American ancestry, all five SNPs have frequency 47.6%, consistent with the "5 SNP" haplotype having frequency ~48% and the "2 SNP" haplotype having frequency 0%.

Genotyping in the MEC samples was conducted by the TaqMan allelic discrimination assay. Blinded duplicates were included to assess genotyping reproducibility; concordance was 100% for all 4 missense SNPs (rs75493593, rs75418188, rs13342692, rs117767867). SNPs rs13342232 and rs13342692 are both on the 2.5M array were also genotyped by TaqMan on 384 Latin American samples; the genotype concordance versus 2.5M data was 100% for one SNP and 99.7% for the other.

Replication data collection and analysis

Multiethnic Cohort (MEC):
The MEC study design and recruitment were described above. Replication utilized the non-Latin American populations in the study: African American (1,084 cases, 1,630 controls), European American (537 cases, 1,827 controls), Japanese American (1,821 cases, 2,736 controls) and Native Hawaiian (625 cases, 1,180 controls)). Informed consent was obtained from all participants. Association testing was performed using logistic

regression with sex, age, BMI, and principal components (PCs) included for all populations except European Americans for which PCs were unavailable. PCs were calculated using AIMs as previously described[49].

The Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium:
The exons of SLC16A11 were sequenced in 10,246 individuals as part of the whole-exome sequencing study of the T2D-GENES consortium. Individuals were selected spanning 5 ethnicities: European (the METSIM study[50] [METSIM] and Ashkenazi individuals recruited from the metropolitan New York region[51] [Ashkenazim]), African-American (the Jackson Heart Study cohort[52] [JHS] and additional individuals recruited from North Carolina, South Carolina, Georgia, Tennessee, or Virginia[53] [WFS]), South Asian (the London Life Sciences Prospective Population Study[54,55] [LOLIPOP] and Singapore Indian Eye Study[56] [Singapore Indians]), East Asian (the Korean Association REsource[57] [KARE] as well as the Singapore Diabetes Cohort Study (SDCS) and Singapore Prospective Study Program[58-61] [Singapore Chinese]), and Hispanic (the San Antonio Family Heart Study[62] (SAFHS), the San Antonio Family Diabetes/Gallbladder Study[63] (SAFDGS), the Veterans Administration Genetic Epidemiology Study[64] (VAGES), the Family Investigation of Nephropathy and Diabetes[65] (FIND), San Antonio component [San Antonio], and individuals from Starr County, Texas[66] [Starr County]). Some individuals in the San Antonio cohort were also genotyped as part of the SAMAFS replication effort; these individuals were excluded from analysis within the T2D-GENES project.

Genotyping was performed via whole-exome sequencing, with target capture via the Agilent SureSelect Human All Exon platform. DNA libraries were barcoded using the Illumina index read strategy and sequenced with an Illumina HiSeq2000. Reads were mapped to the human genome hg19 with the BWA algorithm[67] and processed with the Genome Analysis Toolkit[68] (GATK) to recalibrate base quality-scores and perform local realignment around known indels. Genotypes were called with the Unified Genotyper module of the GATK. Samples with fewer than 76% of targeted bases covered to 20x, with an abnormally high number of non-reference alleles or heterozygosity, or with an abnormally low concordance with prior SNP array genotypes (based on the distribution across all samples) were excluded from analysis. Any sample genotype at a site with fewer than 10x coverage in the sample was ignored (e.g. set as missing).

Singapore Chinese Health Study (SCHS):
The design of the Singapore Chinese Health Study has been previously described[69]. Stage 1 Genotyping was performed at the Genome Institute of Singapore using an Affymetrix ASI (Asian) Axiom array. Genotype calling was performed with the assistance of Affymetrix Corporation. Additional QC was based on sample and SNP call rate, estimation of relatedness, principal components analysis, and comparison of reported and genotyped sex. A total of 4,677 samples (2338 cases and 2339 controls) remained after QC.

Genotype imputation was performed using the Segmented Haplotype Estimation and Imputation Tool (SHAPEIT)[46] version v2.r644 program to phase the main study SNPs. We applied 1000 Genomes Project Phase I data[32] "version 3" as the reference panel, which contained 1,092 individuals of various ethnicities (246 Africans, 181 African Americans, 286 East Asians and 379 Europeans) with 36,648,992 SNPs. IMPUTE2[35] version 2.3.0 was run to perform the imputation. The five SNPs of interest to this study were all imputed using IMPUTE2; estimated imputation r^2 (observed / expected variance) was the range from 0.79 to 0.80 for all five. Association of diabetes with the imputed SNPs was performed using unconditional logistic analysis, adjusting for age, sex, dialect group, and 10 principal components and including imputed SNP minor allele count as an additive effect variable. We also included adjustment for BMI, removing 719 individuals with missing or invalid data for this variable in those analyses.

San Antonio Mexican American Family Studies (SAMAFS):
Genotypes for rs117767867, rs75418188, and rs11564732 was carried out using the MassARRAY system (Sequenom, San Diego, CA). Variant assay primers were designed using Sequenom's online assay design tool in conjunction with their MassARRAY Assay Designer v4.0 software, to amplify ~100bp surrounding the variant for amplification in the MassEXTEND reaction. The MassARRAY Matrix Liquid Handler was used for

automated preparation of reaction products which were then spotted onto 384-sample SpectroCHIP arrays using the MassARRAY Nanodispenser chip spotting station. Spotted arrays were loaded into the MassARRAY Analyzer 4 and sample genotypes determined by measuring the migration times, within a vacuum for each base at a specific locus (MALDI-TOF MS). Analysis of spectra and generation of genotypes was conducted using Sequenom's TyperAnalyzer software v4.0.21. Samples were from three Mexican American family studies from San Antonio, Texas: San Antonio Family Heart Study[62] (SAFHS); San Antonio Family Diabetes/Gallbladder Study[63] (SAFDGS); and the Veterans Administration Genetic Epidemiology Study[64]. In addition, association with rs13342692 was performed using already available GWAS data for SAFHS and SAFDGS[70].

The samples are related through large, multi-generational pedigrees and thus require analysis that accounts for the non-independence of genotypes due to biological relationships. We used a general linear model employing a logit link function and a normal residual to account for this non-independence. The logit link enables estimation of mean effect parameters that are equivalent to those obtained in logistic regression analysis of unrelated samples. We implemented a non-linear link function using the software SOLAR to perform these analyses. The residual error is modeled using a standard pedigree-derived structuring correlation matrix (e.g., kernel) containing the pair-wise coefficients of relationship which allows for genetic correlations between individuals and an identity matrix which yields uncorrelated environmental components across individuals. These two matrices are weighted by the *residual heritability* and 1-*residual heritability*, respectively. Association analysis of the SNP data with T2D is performed using the measured genotype analysis, assuming additivity of allelic effects. The hypothesis of no association is tested by comparing the likelihood of a model in which the effect of measured genotype (i.e., the gene dosage vector) is estimated with a model where the effect of measured genotype is fixed at zero. The test statistic is distributed as a χ^2 distribution with one degree of freedom.

Prior to performing the association analysis, allele frequencies were estimated using maximum likelihood techniques conditional upon all pedigree information. All polymorphisms were tested for Hardy–Weinberg Equilibrium and did not deviate from expectations ($P > 0.05$). All association analyses included age, sex, BMI, and three principal components (PCs) to account for population stratification. For SAFHS and SAFDGS, Illumina genotype data was available for ~1 M SNPs[70], and PCs were computed based on these data. Only unrelated individuals were analyzed to derive the PC axes; PCs for the other subjects were object projection. The VAGES data had available 385 microsatellite markers, and PCs were calculated by first converting these markers to *n* "pseudoSNP" dosages (where *n* is the number of alleles per microsatellite). Principal components analysis was then performed on these pseudoSNP data for unrelated individuals using the R 'prcomp' function, and scores were then projected onto the other VAGES subjects.

Difference in prevalence between Mexican Americans and European Americans accounted for by *SLC16A11*

We model the T2D prevalence accounted for by the *SLC16A11* association separately in both Mexican Americans and European Americans and determine the amount that T2D prevalence would be reduced if this variant were absent from each population. T2D prevalence differs between these groups: 13.3% of Mexican Americans and 14.4% of Mexicans are diagnosed with T2D as compared to 7.1% of U.S. non-Hispanic whites[37,71]. Mexican Americans have a lower prevalence of T2D than Mexicans, so we calculated estimates of reduction relative to the impact of this variant on Mexican American prevalence. We use a standard log-additive effect model (the model our scan tested and found to be associated) and assume an odds ratio of 1.20 for the variant in both populations. We also assume that the odds ratio is a good approximate for R, the relative risk (as is the case for relatively small odds ratios). To obtain a plausible range of values, we report upper and lower bounds by performing this calculation for the lower and upper 95% confidence interval odds ratios for SLC16A11. The overall prevalence in population P is modeled[72] as $K_P = p^2K_{PA} + 2pqRK_{PA} + q^2R^2K_{PA}$, where p is the frequency of the reference allele, q the frequency of the risk variant, and K_{PA} the prevalence if the variant were absent from the population (i.e., if all individuals were homozygous for the reference allele). We then calculate the proportion of difference in prevalence accounted for by *SLC16A11* as

$$[(K_M - K_{MA}) - (K_E - K_{EA})] / (K_M - K_E),$$

where K_M is the overall prevalence in Mexican Americans and K_E is the overall prevalence in European Americans.

Membrane topology prediction

The predicted membrane topology of human SLC16A11 (UniProtKB: Q8NCK7, 471 amino acids) was generated using THMM[73] 2.0 and visualized with TeXtopo[74].

Gene expression analyses

Expression levels of SLC16A family members were determined using the NanoString nCounter system[75]. Fifty genes, including all 14 SLC16A family members, tissue markers, and housekeepers, were measured in 60 independent, commercially-available RNA samples representing 30 different human tissues. The assay was performed using 200 ng total RNA, as per manufacturer's instructions. Data was normalized in two steps. First, variation in sample processing was normalized using the spiked-in positive control probes provided by the nCounter system. Then, variation in input was normalized by median centering. The background level of non-specific binding was determined by calculating the mean + 2 standard deviations of the spiked-in negative control probes. Sample size for each tissue (*n*): pancreas (5), adipose, brain, colon, liver, skeletal muscle, and thyroid (3), adrenal, fetal brain, breast, heart, kidney, lung, placenta, prostate, small intestine, spleen, testes, thymus, and trachea (2), bladder, cervix, fetal liver, oesophagus, ovary, salivary gland, fetal skeletal muscle, skin, umbilical cord, and uterus (1).

For the "55k screen," a microarray-based analysis of gene expression, a dataset was generated from 55,269 samples in the Gene Expression Omnibus (GEO) and Cancer Cell Line Encyclopedia[76] (CCLE) databases that were measured on the Affymetrix U133 Plus 2.0 Array. This array was chosen because it contains 2 probe sets for *SLC16A11*; however, *SLC16A13* is not measured on this array. Each sample in the raw expression data was first linearly transformed using a modified invariant set normalization method on a set of eighty control genes with stable expression on U133 Plus 2.0[77]. The expression data were then log2 transformed to stabilize the variance and expression distribution. Finally, the data were quantile normalized so the expression distribution of each sample matched[78]. Expression values for genes with multiple probe sets were calculated by taking the median value of all probe sets for that gene. Following normalization, a log2 expression value of 4 is considered baseline and log2 expression values greater than 6 are considered expressed. Sample annotations were curated based on GEO descriptions provided by depositors. To account for variation in the number of samples representing each tissue in the dataset, expression of a gene is presented as the fraction of samples of a tissue that exceed a log2 expression value of 6. Sample size for each tissue (*n*): adipose (394), adrenal (69), brain (1990), breast (4104), heart (178), kidney (675), liver (721), lung (1442), pancreas (150), placenta (107), prostate (578), salivary gland (26), skeletal muscle (793), skin (947), testis (102), thyroid (108).

Plasmids and Cell Lines

Plasmids encoding C-terminus, V5-tagged human SLC16A13 and control proteins (BFP, GFP, HcRed, and Luciferase) in the pLX304 lentiviral vector were obtained from the RNAi Consortium at the Broad Institute[79]. The open reading frame of human *SLC16A11* (Consensus CDs, CCDS11086.1) was synthesized and sublconed into pLX304 by Genscript. To generate a *SLC16A1* expression plasmid, a cDNA clone of human *SLC16A1* was purchased from the Dana-Farber/Harvard Cancer Center DNA Resource Core and subcloned into pLX304 through Gateway® recombination. Mycoplasma-free HeLa cells were obtained from within the Broad Institute in 2009. HeLa cells were maintained in culture medium containing DMEM (Cellgro), 10% heat-inactivated fetal bovine serum, 100 U/ml penicillin and 100 ug/ml streptomycin.

Immunocytochemistry

To determine the subcellular localization of SLC16A proteins, HeLa cells plated on chamber slides were transiently transfected with expression plasmids encoding C-terminus, V5-tagged proteins using FuGENE®HD (Promega). Cells were fixed 24-48 h post-transfection with 4% paraformaldehyde, permeabilized with 0.1% TritonX-100, and immunostained using antibodies against the V5 epitope (1:1500, Invitrogen, R960) and Calnexin (1:25, Novus, NBP1-85519) or Golph4 (1:250, Novus, NBP1-91954). Alexa Fluor® 488 goat anti-mouse and Alexa Fluor® 594 goat anti-rabbit secondary antibodies (1:2000) were used for detection. MitoTracker® Red CMXRos staining was performed prior to fixation, as per manufacturer's instructions. Nuclei were stained with Hoechst 33342. Images were captured using a 63x objective on a Zeiss Cell Observer and processed in AxioVision and Adobe Photoshop. Due to heterogeneity in expression levels of overexpressed proteins and endogenous organelle markers, imaging of each protein was optimized for clarity of localization and varied across images; therefore, images are not representative of relative expression levels of each protein.

Metabolic profiling

Three independent experiments for metabolite profiling were performed as follows. HeLa cells were plated in 6 well plates at a density of 150,000 cells per well. The following day, cells were transiently transfected in triplicate with expression plasmids encoding C-terminus, V5-tagged SLC16A11 and control proteins (BFP, GFP, HcRed, and Luciferase) using 1 µg DNA and 4 µl FuGENE®HD (Promega) per well. Four independent plasmids of SLC16A11 were used to compare to the four control proteins. The day after transfection, the media was removed and fresh media was added to the cells. Two days later (after 3 days of gene expression), media and cellular lysates were collected for metabolite profiling, as described below.

Analyses of polar and non-polar lipids in cell extracts and growth media were conducted using a liquid chromatography tandem mass spectrometry (LC-MS) system comprised of an Open Accela 1250 U-HPLC and a Q Exactive hybrid quadrupole orbitrap mass spectrometer (Thermo Fisher Scientific; Waltham, MA). Lipid metabolites were extracted from cells grown in 6 well plates using 800 µL of isopropanol containing 1-dodecanoyl-2-tridecanoyl-sn-glycero-3-phosphocholine (Avanti Polar Lipids; Alabaster, AL) as an internal standard. Prior to analyses, cell extracts (200 μL) were concentrated 2-fold by evaporation under nitrogen gas in a TurboVap LV (Caliper, Hopkinton, MA) and resuspension in 100 μL of isopropanol containing 1-dodecanoyl-2-tridecanoyl-sn-glycero-3-phosphocholine. Media samples (10 μL) were prepared with the addition of 90 μL of isopropanol containing internal standard, followed by centrifugation at 9000 x g for 15 minutes. Samples (10 μL) were injected onto 150 x 3.0 mm Prosphere HP C4 column (Grace, Columbia, MD). The column was eluted isocratically with 80% mobile phase A (95:5:0.1 vol/vol/vol 10mM ammonium acetate/methanol/acetic acid) for 2 minutes followed by a linear gradient to 80% mobile-phase B (99.9:0.1 vol/vol methanol/acetic acid) over 1 minute, a linear gradient to 100% mobile phase B over 12 minutes, then 10 minutes at 100% mobile-phase B. MS data were acquired in the positive ion mode using electrospray ionization and full scan MS over m/z 400-1100. Other MS settings were: spray voltage, 3 kV; capillary temperature, 300°C; sheath gas, 50; auxiliary gas, 15; heater temperature, 300 °C; S-lens level, 60; and resolution, 70 000. Raw data from cell extracts and growth media were integrated and visually inspected using TraceFinder 2.1 software (Thermo Fisher Scientific; Waltham, MA).

Polar metabolites were analyzed using hydrophilic interaction liquid chromatography (HILIC)-MS methods operated in positive and negative ion modes as described previously[80]. Polar metabolites were extracted from cells grown in 6 well plates using 800 μL of 80% methanol. Media samples were extracted using 4 volumes of 80% methanol for negative ion mode polar metabolite analyses and 9 volumes of 75/25 acetonitrile/methanol for positive ion mode polar metabolite analyses.

HeLa Cell Metabolite Analysis

After raw data were integrated and reviewed, cellular lysate values were normalized to the total signal obtained in each sample to reduce any potential skew due to variations in biomass or other extraneous effects. A scaling

factor was computed as the ratio of the sum of the total signal across all measured metabolites for a sample versus the mean of all such total signal values across all samples. Each metabolite value in each sample was then adjusted by multiplying to the scaling factor. For the majority of metabolites and samples, the scaling factor was close to 1 indicating little variation in biomass or other variables that might influence the relative signal obtained.

Calculating the fold change of each metabolite:

In analyzing the metabolite data, non-parametric statistical tests were used since for many metabolites the assumptions of normality were not met and thus t-tests were deemed unsuitable. The fold change, *d*, of each metabolite as measured in the cells expressing reference SLC16A11 versus those expressing control proteins was calculated. This was done by first calculating the fold change within each experiment (d_i for experiment number *i*). This was defined to be the ratio of the median value of the metabolite in the SLC16A11-expressing cells to the median for the control cells within a single experiment. The median value was taken to reduce the sensitivity of the analysis to outliers. Let $x_{ij}^{(V)}$ be the j[th] measurement of the metabolite in experiment *i* for the SLC16A11 variant, and let $x_{ij}^{(C)}$ be the j[th] measurement in experiment *i* for the controls. Then $d_i = median(x_{i1}^{(V)}, \ldots, x_{in}^{(V)}) / median(x_{i1}^{(C)}, \ldots, x_{in}^{(C)})$ where *n* and *m* are the number of measurements in the SLC16A11 variant and in controls respectively. These three values were then averaged over experiments to give the overall fold change, $mean(d_1, d_2, d_3)$.

In order to identify any plasmids that behaved as outliers in their metabolic profiles, we calculated the correlation of these fold change values between each pair of plasmids, for every metabolite. This was done in each of the three experiments. All possible plasmid pairs showed high correlation in each experiment and therefore none were removed in the analysis.

Testing for a change in metabolite values in SLC16A11 versus controls:

To determine if, for each metabolite in turn, its measurements in SLC16A11 cells differed from those in controls we first log-transformed these values (to base 10) as their distributions were often highly skewed. These log-transformed values will be denoted as $\tilde{x}$, hence $\tilde{x}_{ij}^{(V)} = log_{10}(x_{ij}^{(V)})$. To aggregate the data of the three experiments, we subtracted the mean value within each experiment where the mean of experiment *i* was $mean(\tilde{x}_{i1}^{(V)}, \ldots, \tilde{x}_{in}^{(V)}, \tilde{x}_{i1}^{(C)}, \ldots, \tilde{x}_{im}^{(C)})$. Letting $\{z_{ij}^{(V)}\}$ represent the mean-subtracted values of the SLC16A11 variant, we the performed a two-sided Wilcox rank sum test on $\{z_{ij}^{(V)}\}$ versus $\{z_{ij}^{(C)}\}$ for *i* in {1,2,3} and *j* in {1,..,*n*} and {1,...,*m*} respectively.

Comparing metabolite levels of named lipid classes in HeLa cells expressing SLC16A11 reference haplotype to those expressing control genes:

We sought to test if the metabolites within a named lipid class were significantly elevated or reduced in the SLC16A11-expressing cells versus in controls. The mean-subtracted log-transformed values, $\{z_{ij}^{(V)}\}$, were calculated as described above. The median of these values was found for both the variant and the controls, that is $median(\{z_{ij}^{(V)}\})$ and $median(\{z_{ij}^{(C)}\})$. This was done for each metabolite in the lipid class to give a pair of median values for each metabolite. We then performed a two-sided Wilcox signed-rank test on these pairs to determine if the metabolites of the lipid class differ significantly between SLC16A11 and controls. We note that this test assumes that if metabolite levels do change between the variant and controls, then they all do so in the same direction (increase or decrease).

Metabolite Enrichment Analysis

We applied a similar strategy for assessing metabolite pathway enrichment as previously described for gene set enrichment using the GSEA approach[81]. For this analysis, all KEGG pathways from the human reference set and eight additional classes of metabolites covering lipid sub-types and carnitines were used to assess enrichment or depletion. Pathways and classes were pruned to reflect only those metabolites measured by our metabolite profiling platform, and a pathway had to have at least six measurable metabolites in this dataset in order to be scored. For each pathway or metabolite class, enrichment was computed using the unweighted Kolmogorov-Smirnov statistic similar to that described in GSEA. Metabolites were rank ordered by fold change relative to controls, and the enrichment score computed starting from rank 1. Each metabolite belonging to the pathway being scored would increment the enrichment score by 1, and any metabolites that were non-members led to a decrement in the score of 1/n, where n was the total number of metabolites assessed or 339. After scoring all pathways, the rank ordering of metabolites was randomly shuffled and a null enrichment score re-computed for each of the pathways. This was repeated 10,000 times to achieve a null distribution, which was then the basis for computing a p-value for the true enrichment score of each pathway. The false discovery rate associated with each pathway and nominal p-value was computed as described for GSEA[82].

Supplementary Tables

SNP	Chr	Position	Gene	Risk Allele	P-value				OR		OR directionally consistent?	
					Liability w/BMI	Logistic w/o BMI	Logistic w/BMI	Liability w/o BMI	w/BMI	w/o BMI	w/BMI	w/o BMI
rs10923931	1	120517959	NOTCH2	T	0.08	0.06	0.07	0.11	1.11 (0.99-1.25)	1.10 (0.98-1.23)	+	+
rs340874	1	214159256	PROX1	C	7.78E-03	6.60E-03	3.89E-03	5.51E-03	1.10 (1.03-1.18)	1.10 (1.03-1.17)	+	+
rs780094	2	27741237	GCKR	C	0.28	0.16	0.31	0.19	1.04 (0.97-1.11)	1.05 (0.98-1.12)	+	+
rs7578597	2	43732823	THADA	T	0.05	0.08	0.14	0.15	1.12 (0.97-1.30)	1.11 (0.97-1.29)	+	+
rs243021	2	60584819	BCL11A	A	0.02	0.03	0.03	0.02	1.08 (1.01-1.15)	1.08 (1.02-1.16)	+	+
rs2925757	2	161101169	CAPN10	G	0.28	0.25	0.36	0.37	0.95 (0.86-1.05)	0.96 (0.87-1.05)	-	-
rs13389219	2	165528876	GRB14	C	0.05	0.07	0.06	0.08	1.08 (1.00-1.17)	1.08 (0.99-1.16)	+	+
rs2943641	2	227093745	IRS1	C	0.07	0.08	0.12	0.18	1.07 (0.98-1.17)	1.06 (0.98-1.15)	+	+
rs1801282	3	12393125	PPARG	C	0.02	0.07	0.01	0.04	1.14 (1.03-1.25)	1.10 (1.00-1.21)	+	+
rs6780569	3	23198484	UBE2E2	A	0.37	0.4	0.37	0.4	0.94 (0.84-1.07)	0.95 (0.84-1.07)	-	-
rs831571	3	64048297	PSMD6	C	0.47	0.43	0.44	0.38	0.95 (0.85-1.07)	0.95 (0.85-1.06)	-	-
rs4607103	3	64711904	ADAMTS9	C	1	0.82	0.88	0.82	0.99 (0.93-1.07)	0.99 (0.93-1.06)	-	-
rs11708067	3	123065778	ADCY5	A	1.68E-04	4.99E-04	1.57E-04	2.65E-04	1.15 (1.07-1.23)	1.14 (1.06-1.22)	+	+
rs1470579	3	185529080	IGF2BP2	C	6.14E-04	5.12E-03	4.18E-04	2.32E-03	1.14 (1.06-1.23)	1.12 (1.04-1.20)	+	+
rs16861329	3	186666461	ST6GAL1	C	0.24	0.43	0.2	0.41	1.05 (0.98-1.12)	1.03 (0.96-1.10)	+	+
rs6815464	4	1309901	MAEA	C	9.89E-06	1.21E-05	1.62E-05	9.36E-06	1.17 (1.09-1.26)	1.17 (1.09-1.25)	+	+
rs1801214	4	6303022	WFS1	T	9.23E-04	6.96E-04	2.26E-03	1.23E-03	1.12 (1.04-1.21)	1.13 (1.05-1.21)	+	+
rs459193	5	55806751	ANKRD55	G	5.89E-03	9.74E-03	1.06E-02	2.40E-02	1.10 (1.02-1.19)	1.09 (1.01-1.17)	+	+
rs4457053	5	76424949	ZBED3	G	0.54	0.78	0.42	0.79	1.03 (0.96-1.10)	1.01 (0.95-1.08)	+	+
rs10440833	6	20688121	CDKAL1	A	0.03	0.12	0.05	0.17	1.08 (1.00-1.16)	1.05 (0.98-1.13)	+	+
rs9470794	6	38106844	ZFAND3	C	0.08	0.05	0.09	0.07	0.89 (0.78-1.01)	0.88 (0.78-1.01)	-	-
rs1535500	6	39284050	KCNK16/ KCNK17	T	0.03	0.02	0.04	0.03	1.07 (1.00-1.14)	1.07 (1.01-1.14)	+	+
rs17168486	7	14898282	DGKB	T	0.04	0.11	0.06	0.12	1.07 (1.00-1.14)	1.05 (0.99-1.12)	+	+
rs864745	7	28180556	JAZF1	T	7.86E-06	7.66E-05	3.50E-06	2.29E-05	1.18 (1.10-1.26)	1.16 (1.08-1.24)	+	+
rs4607517	7	44235668	GCK	A	0.23	0.22	0.19	0.26	1.06 (0.97-1.14)	1.05 (0.97-1.13)	+	+
rs6467136	7	127164958	GCC1/PAX4	G	0.25	0.25	0.28	0.29	0.96 (0.90-1.03)	0.97 (0.91-1.03)	-	-
rs972283	7	130466854	KLF14	G	0.02	0.05	0.03	0.07	1.08 (1.01-1.15)	1.06 (1.00-1.13)	+	+
rs516946	8	41519248	ANK1	C	4.03E-04	9.02E-04	5.56E-04	1.29E-03	1.16 (1.07-1.26)	1.14 (1.06-1.24)	+	+
rs896854	8	95960511	TP53INP1	T	0.27	0.23	0.25	0.18	1.04 (0.97-1.11)	1.04 (0.98-1.11)	+	+
rs3802177	8	118185025	SLC30A8	G	2.38E-04	1.15E-03	3.61E-04	1.79E-03	1.15 (1.07-1.23)	1.12 (1.05-1.21)	+	+
rs7041847	9	4287466	GLIS3	A	0.01	0.01	0.02	0.03	1.08 (1.01-1.15)	1.08 (1.01-1.15)	+	+

rs17584499	9	8879118	PTPRD	T	0.93	0.91	0.82	0.67	0.99 (0.92-1.07)	0.98 (0.91-1.06)	-	-
rs10965250	9	22133284	CDKN2A/B	G	0.09	0.1	0.11	0.13	1.09 (0.98-1.21)	1.08 (0.98-1.20)	+	+
rs824248	9	28772700	LINGO2	T	2.11E-06	7.45E-06	1.21E-06	3.58E-06	1.19 (1.11-1.27)	1.18 (1.10-1.26)	+	+
rs13292136	9	81952128	CHCHD9	C	0.57	0.84	0.78	0.98	1.01 (0.93-1.10)	1.00 (0.92-1.09)	+	+
rs2796441	9	84308948	TLE1	G	0.43	0.52	0.69	0.68	1.01 (0.95-1.08)	1.01 (0.95-1.08)	+	+
rs12779790	10	12328010	CDC123/ CAMK1D	G	0.12	0.14	0.08	0.1	1.08 (0.99-1.17)	1.07 (0.99-1.16)	+	+
rs1802295	10	70931474	VPS26A	T	0.92	0.97	0.92	0.98	1.00 (0.92-1.07)	1.00 (0.93-1.07)	-	-
rs12571751	10	80942631	ZMIZ1	A	0.02	0.03	0.02	0.05	1.08 (1.01-1.15)	1.07 (1.00-1.13)	+	+
rs1111875	10	94462882	HHEX	C	0.11	0.11	0.18	0.09	1.05 (0.98-1.12)	1.06 (0.99-1.13)	+	+
rs7903146	10	114758349	TCF7L2	T	2.47E-17	1.02E-14	6.20E-17	3.55E-15	1.41 (1.30-1.53)	1.37 (1.27-1.48)	+	+
rs2334499	11	1696849	HCCA2	T	0.07	0.14	0.06	0.1	0.94 (0.88-1.00)	0.95 (0.89-1.01)	-	-
rs231362	11	2691471	KCNQ1	G	0.18	0.29	0.19	0.39	1.05 (0.98-1.13)	1.03 (0.96-1.11)	+	+
rs2237897	11	2858546	KCNQ1	C	4.94E-16	8.50E-15	2.23E-14	7.97E-13	1.35 (1.25-1.45)	1.31 (1.22-1.41)	+	+
rs5219	11	17409572	KCNJ11	T	6.36E-03	1.95E-02	6.98E-03	2.31E-02	1.10 (1.03-1.17)	1.08 (1.01-1.15)	+	+
rs1552224	11	72433098	CENTD2	A	0.02	0.03	0.02	0.02	1.16 (1.02-1.33)	1.16 (1.02-1.32)	+	+
rs1387153	11	92673828	MTNR1B	T	0.4	0.39	0.5	0.39	1.03 (0.95-1.11)	1.03 (0.96-1.11)	+	+
rs11063069	12	4374373	CCND2	G	0.49	0.51	0.39	0.4	1.05 (0.95-1.15)	1.04 (0.95-1.15)	+	+
rs10842994	12	27965150	KLHDC5	C	0.39	0.45	0.35	0.48	1.04 (0.96-1.14)	1.03 (0.95-1.12)	+	+
rs1531343	12	66174894	HMGA2	C	0.06	0.12	0.08	0.14	1.11(0.99-1.25)	1.09 (0.97-1.22)	+	+
rs7961581	12	71663102	TSPAN8/LGR5	C	9.17E-04	2.50E-03	1.34E-03	3.21E-03	1.16 (1.06-1.27)	1.14 (1.05-1.24)	+	+
rs7957197	12	121460686	HNF1A	T	0.06	0.1	0.04	0.09	1.12 (1.01-1.24)	1.10 (0.99-1.22)	+	+
rs1359790	13	80717156	SPRY2	G	1.10E-04	5.30E-04	8.49E-05	3.59E-04	1.15 (1.07-1.22)	1.13 (1.06-1.20)	+	+
rs7172432	15	62396389	C2CD4A/B	A	0.03	0.04	0.02	0.02	1.08 (1.02-1.15)	1.08 (1.01-1.15)	+	+
rs7178572	15	77747190	HMG20A	G	0.02	0.04	0.01	0.03	1.09 (1.02-1.16)	1.08 (1.01-1.15)	+	+
rs11634397	15	80432222	ZFAND6	G	0.12	0.11	0.09	0.08	1.06 (0.99-1.13)	1.06 (1.00-1.13)	+	+
rs2028299	15	90374257	AP3S2/ C15orf38/ AP3S2	C	0.61	0.51	0.45	0.57	0.97 (0.89-1.05)	0.98 (0.90-1.06)	-	-
rs8042680	15	91521337	PRC1	A	0.94	0.95	0.84	0.93	1.01 (0.94-1.09)	1.00 (0.93-1.08)	+	+
rs11642841	16	53845487	FTO	A	4.97E-03	2.96E-04	4.56E-03	1.11E-04	1.13 (1.04-1.22)	1.17 (1.08-1.27)	+	+
rs7202877	16	75247245	BCAR1	T	0.07	0.1	0.06	0.15	1.13 (1.00-1.29)	1.10 (0.97-1.24)	+	+
rs4523957	17	2208899	SMG6/SRR	T	0.85	0.99	0.84	0.94	1.01 (0.94-1.08)	1.00 (0.94-1.06)	+	-
rs757210	17	36096515	HNF1B	T	0.61	0.49	0.44	0.48	0.97 (0.91-1.04)	0.98 (0.91-1.04)	-	-
rs12970134	18	57884750	MC4R	A	0.43	0.29	0.41	0.37	1.04 (0.95-1.14)	1.04 (0.95-1.14)	+	+
rs10401969	19	19407718	CILP2	C	0.14	0.18	0.13	0.18	1.12 (0.97-1.30)	1.11 (0.96-1.28)	+	+

rs3786897	19	33893008	PEPD	A	0.25	0.26	0.24	0.21	0.96 (0.89-1.03)	0.95 (0.89-1.03)	-	-
rs8108269	19	46158513	GIPR	G	0.19	0.22	0.22	0.28	1.04 (0.98-1.11)	1.04 (0.97-1.10)	+	+
rs6017317	20	42946966	HNF4A	G	0.29	0.38	0.23	0.23	1.04 (0.98-1.11)	1.04 (0.98-1.11)	+	+
rs4812829	20	42989267	HNF4A	A	0.04	0.06	0.03	0.03	1.08 (1.01-1.15)	1.07 (1.01-1.15)	+	+

Supplementary Table 1: 68 previously associated variants, with SNP id, chromosome, physical position, nearby gene, and risk allele. Listed are liability p-values from LTSOFT with and without BMI correction, logistic regression p-values from PLINK with and without BMI correction, odds ratios (OR) from logistic regression with and without BMI correction, and directional consistency of these odds ratios with previous studies.

Cohort	PLINK		LTSOFT	
	OR (CI)	P	Beta (CI)	P
UIDS (*n*=1,953)	0.72 (0.60-0.87)	4.8×10^{-4}	-0.15 (-0.22 to -0.08)	6.2×10^{-5}
DMS (*n*=1,162)	0.72 (0.57-0.90)	4.2×10^{-3}	-0.12 (-0.21 to -0.03)	7.0×10^{-3}
MCDS (*n*=900)	0.92 (0.70-1.20)	0.54	-0.03 (-0.12 to 0.07)	0.56
MEC (*n*=4,199)	0.76 (0.65-0.88)	3.3×10^{-4}	-0.10 (-0.16 to -0.05)	3.4×10^{-4}

Supplementary Table 2: Individual cohort association results for rs11564732, the top SNP association in 11p15.5. Results are given for both logistic regression from PLINK and a liability threshold model from LTSOFT (*P*-values are not corrected for genomic control). Table includes estimated odds ratio (OR), 95% confidence interval on the odds ratio (CI), and p-value for association.

Cohort	Age (years) ± SD	PLINK		LTSOFT	
		OR (CI)	P	Beta (CI)	P
UIDS (*n*=1,953)	55.7 ± 10.8	1.43 (1.24-1.65)	6.1×10^{-7}	0.16 (0.11-0.22)	1.7×10^{-8}
DMS (*n*=1,162)	54.6 ± 10.1	1.38 (1.15-1.65)	6.2×10^{-4}	0.13 (0.06-0.20)	2.6×10^{-4}
MCDS (*n*=900)	62.9 ± 7.7	1.00 (0.81-1.24)	0.98	0.00 (-0.07 to 0.08)	0.91
MEC (*n*=4,199)	59.3 ± 7.0	1.22 (1.09-1.35)	3.0×10^{-4}	0.07 (0.03-0.11)	3.3×10^{-4}

Supplementary Table 3: Individual cohort association results for rs13342232, the top SNP association in 17p13.1. Results are given for both logistic regression from PLINK and a liability threshold model from LTSOFT. Table includes average age in years for each cohort ± standard deviation, estimated odds ratio (OR), 95% confidence interval on the odds ratio (CI), and P-value for association. Elsewhere, we report heterogeneity in odds ratio based on age, with older individuals having a lower odds ratio than younger; cohort-specific odds ratios and ages presented here are consistent with this finding.

rs13342692 genotype	SIGMA			UIDS			DMS			MCDS			MEC		
	N=0	N=1	N=2	N=0	N=1	N=2	N=0	N=1	N=2	N=0	N=1	N=2	N=0	N=1	N=2
g=0	1673	1732	556	294	383	161	90	205	130	96	196	98	1193	948	167
% of g=0	42%	44%	14%	35%	46%	19%	21%	48%	31%	25%	50%	25%	52%	41%	7%
g=1	194	1865	1274	31	407	387	18	245	288	6	170	212	139	1043	387
% of g=1	6%	56%	38%	4%	49%	47%	3%	44%	52%	2%	44%	55%	9%	66%	25%
g=2	8	124	785	2	20	267	0	19	167	0	13	108	6	72	243
% of g=2	1%	14%	86%	1%	7%	92%	0%	10%	90%	0%	11%	89%	2%	22%	76%

Supplementary Table 4: Native American local ancestry counts (N) stratified by genotype count at rs13342692. Also shown are within-cohort (or SIGMA study-wide) percentages of samples within a given genotype class that have the a given Native American local ancestry count. The risk haplotype at *SLC16A11* (tagged by rs13342692) has higher frequency in Native Americans and thus genotype count correlates with Native American local ancestry.

rs13342692 genotype	SIGMA					
	Cases			Controls		
	N=0	N=1	N=2	N=0	N=1	N=2
g=0	709	701	250	964	1031	306
% of g=0	43%	42%	15%	42%	45%	13%
g=1	88	922	668	106	943	606
% of g=1	5%	55%	40%	6%	57%	37%
g=2	5	59	445	3	65	340
% of g=2	1%	12%	87%	1%	16%	83%

Supplementary Table 5: Native American local ancestry counts (N) stratified by genotype count and case-control status at rs13342692. Note that population stratification (which is captured by global ancestry) is a confounder for association, but significant deviations in local ancestry between cases and controls at a given locus is evidence for association. Thus, while it is important to include covariates representing global ancestry, such as principal components, including adjustment for local ancestry is incorrect and in general can lead to type II error.

SNP	Substitution	Region	SNP Type	Prediction	SIFT Score
rs13342232	L187L	Exon	Synonymous	N/A	N/A
rs13342692	D127G	Exon	Nonsynonymous	Damaging	0.01
rs117767867	V113I	Exon	Nonsynonymous	Tolerated	0.41
rs75418188	G340S	Exon	Nonsynonymous	Tolerated	0.59
rs75493593	P443T	Exon	Nonsynonymous	Tolerated	0.44

Supplementary Table 6: Associated SNPs in chromosome 17p13.1 that are in the protein coding region of *SLC16A11* with SIFT scores for the four nonsynonymous changes.

Trait	P-Value (with BMI correction)	N	Samples Used
Total Cholesterol	0.83 (0.74)	3,855	Cases and Controls; subsets of UIDS and DMS cohorts with trait data available
Triglycerides	0.84 (0.76)	3,855	
LDL Cholesterol	0.25 (0.65)	2,756	
HDL Cholesterol	0.54 (0.44)	2,957	
Fasting Insulin	0.80 (0.59)	1,505	Controls only; subsets of UIDS, DMS, and MCDS cohorts with trait data available
HOMA2-β	0.83 (0.67)		
HOMA2-IR	0.74 (0.72)		
Fasting Glucose	0.53 (0.69)		
HOMA-β	0.78 (0.82)		
HOMA-IR	0.67 (0.80)		

Supplementary Table 7: Quantitative trait association results. P-values from association with and without correcting for BMI in linear regression models; in all cases, we adjusted for age, gender, T2D status, top two principal components, and cohort membership. All phenotypes were log-transformed. HOMA-β, homeostasis model assessment estimates of beta-cell function; HOMA-IR, homeostasis model assessment estimates of insulin resistance; HOMA2-β and HOMA2-IR, updated computer models.

SNP	Cohort	Ethnicity	MAF	# Controls	# Cases	OR (95% CI)	P	Meta OR	Meta P
rs75493593 (Specific to 5-SNP haplotype)	T2D-GENES	Mexican American	0.21	832	896	1.21 (0.98-1.49)	0.08	1.20 (1.09-1.31)	1.1×10^{-4}
		East Asian	0.10	1,133	990	1.22 (0.98-1.53)	0.08		
		South Asian	0.003	1,105	1,082	0.90 (0.29-2.73)	0.85		
		European	0.02	801	990	1.02 (0.59-1.76)	0.94		
		African American	0.006	987	1,008	1.01 (0.41-2.49)	0.97		
	MEC	Japanese	0.08	1,747	1,778	1.11 (0.93-1.33)	0.23		
		European American	0.008	532	884	2.42(1.13-5.19)	0.02		
		African American	0.008	1,116	1,119	0.93 (0.44-1.94)	0.84		
		Native Hawaiian	0.03	579	698	1.24 (0.75-2.05)	0.40		
	SCHS	Singaporean	0.10	1,959	2,009	1.27 (1.05-1.52)	0.008		
	SAMAFS	Mexican American	NA	NA	NA	NA	NA		
rs13342692 (Present on 2-SNP and 5-SNP haplotypes)	T2D-GENES	Mexican American	0.23	709	802	1.14 (0.90-1.44)	0.28	1.13 (1.06-1.20)	1.5×10^{-4}
		East Asian	0.10	1,133	989	1.20 (0.95-1.51)	0.12		
		South Asian	0.004	1,102	1,081	0.70 (0.26-1.83)	0.46		
		European	0.03	799	982	1.33 (0.86-2.06)	0.20		
		African American	0.17	984	1,004	1.04 (0.90-1.19)	0.60		
	MEC	Japanese	0.08	1,758	1,788	1.11 (0.92-1.34)	0.26		
		European American	0.01	532	885	1.91 (1.03-3.55)	0.04		
		African American	0.31	1,117	1,120	1.09 (0.95-1.24)	0.24		
		Native Hawaiian	0.04	578	698	1.18 (0.74-1.87)	0.49		
	SCHS	Singaporean	0.10	1,959	2,009	1.26 (1.06-1.51)	0.01		
	SAMAFS	Mexican American	0.28	1,491	594	1.13 (0.96-1.32)	0.15		

Supplementary Table 8: Replication results from two missense SNPs present on the risk haplotype of *SLC16A11*. SNP rs75493593 is specific to the 5 SNP risk haplotype identified by this study. SNP rs13342692 is present on both the 5 SNP haplotype and the 2 SNP haplotype that is common in Africa. The missense SNP rs75418188 and rs117767867 were excluded due to low call rate in T2D-GENES. The T2D-GENES, MEC, and SCHS analyses were all performed using logistic regression corrected for age, BMI, sex, and principal components (one exception is the MEC European American cohort for which principal components were unavailable and therefore not included). The SAMAFS cohort used a logit model (analogous to a logistic regression model) that corrects for relatedness among samples and also included age, BMI, sex, and principal components as covariates.

	T – % of Human-Macaque TMRCA	T – k Years
Risk to European haplotype	3.20 (2.04 – 4.78)	799 (509 – 1,194)

Supplementary Table 9: Estimates of divergence time of the two haplotypes of 1,000 Genomes Project individual NA11930, who is heterozygous for the risk haplotype. Shown are divergence in percent of human-macaque TMRCA at *SLC16A11* and thousands of years (k Years) assuming human-macaque TMRCA of 25 million years. Table includes maximum likelihood estimates and 95% confidence intervals.

Population	ID	T – % of Human-Macaque TMRCA	T – k Years
ASW	NA20314	0.007 (0.222 – 1.596)	171 (55 – 399)
		0.008 (0.301 – 1.786)	205 (75 – 446)
CLM	HG01250	0.007 (0.222 – 1.596)	171 (55 – 399)
		0.008 (0.301 – 1.786)	205 (75 – 446)
	HG01272	0.010 (0.385 – 1.972)	239 (96 – 493)
		0.008 (0.301 – 1.786)	205 (75 – 446)
MXL	NA19654	0.008 (0.301 – 1.786)	205 (75 – 446)
		0.008 (0.301 – 1.786)	205 (75 – 446)
	NA19728	0.007 (0.222 – 1.596)	171 (55 – 399)
		0.008 (0.301 – 1.786)	205 (75 – 446)
	NA19752	0.012 (0.563 – 2.336)	308 (141 – 584)
		0.007 (0.222 – 1.596)	171 (55 – 399)
	NA19753	0.007 (0.222 – 1.596)	171 (55 – 399)
		0.007 (0.222 – 1.596)	171 (55 – 399)
	NA19783	0.007 (0.222 – 1.596)	171 (55 – 399)
		0.008 (0.301 – 1.786)	205 (75 – 446)
CHB	NA18539	0.008 (0.301 – 1.786)	205 (75 – 446)
		0.007 (0.222 – 1.596)	171 (55 – 399)
	NA18547	0.007 (0.222 – 1.596)	171 (55 – 399)
		0.005 (0.149 – 1.400)	137 (37 – 350)
	NA18574	0.026 (1.564 – 4.057)	650 (391 – 1,014)
		0.030 (1.885 – 4.555)	752 (471 – 1,139)
	NA18609	0.023 (1.354 – 3.722)	581 (339 – 930)
		0.007 (0.222 – 1.596)	171 (55 – 399)
	NA18621	0.005 (0.149 – 1.400)	137 (37 – 350)
		0.008 (0.301 – 1.786)	205 (75 – 446)
	NA18745	0.015 (0.751 – 2.691)	376 (188 – 673)
		0.011 (0.472 – 2.155)	273 (118 – 539)
FIN	HG00171	0.011 (0.472 – 2.155)	273 (118 – 539)
		0.008 (0.301 – 1.786)	205 (75 – 446)

Supplementary Table 10: Estimates of divergence times to the Neandertal sequence of the risk haplotype in the 1000 Genomes populations. We selected all individuals that are homozygous for the risk haplotype. Shown are divergence in percent of human-macaque TMRCA at *SLC16A11* and years assuming human-macaque TMRCA of 25 million years. Table includes the maximum likelihood estimates and 95% confidence intervals.

Population	ID	T – % of Human-Macaque TMRCA	T – k Years
CLM	HG01359	0.025 (0.015 – 0.039)	615 (365 – 972)
	HG01359	0.019 (0.010 – 0.032)	479 (262 – 803)
	HG01374	0.027 (0.017 – 0.042)	684 (418 – 1,056)
	HG01374	0.025 (0.015 – 0.039)	615 (365 – 972)
MXL	NA19679	0.030 (0.019 – 0.046)	752 (471 – 1,139)
	NA19679	0.019 (0.010 – 0.032)	479 (262 – 803)
	NA19756	0.023 (0.014 – 0.037)	581 (339 – 930)
	NA19756	0.033 (0.021 – 0.049)	820 (526 – 1,221)
CHB	NA18616	0.033 (0.021 – 0.049)	820 (526 – 1,221)
	NA18616	0.033 (0.021 – 0.049)	820 (526 – 1,221)
	NA18747	0.025 (0.015 – 0.039)	615 (365 – 972)
	NA18747	0.034 (0.022 – 0.050)	855 (553 – 1,262)
FIN	HG00312	0.019 (0.010 – 0.032)	479 (262 – 803)
	HG00312	0.023 (0.014 – 0.037)	581 (339 – 930)
	HG00331	0.034 (0.022 – 0.050)	855 (553 – 1,262)
	HG00331	0.031 (0.020 – 0.047)	786 (498 – 1,180)

Supplementary Table 11: Estimates of divergence times to the Neandertal sequence of non-risk haplotypes in the 1000 Genomes populations. We randomly selected two individuals that do not carry the risk haplotype from each of the populations with individuals that are homozygous risk haplotype carriers (Supplementary Table 10 lists these homozygous carriers; note that the individual in ASW is of Mexican descent). Shown are divergence in percent of human-macaque TMRCA at *SLC16A11* and years assuming human-macaque TMRCA of 25 million years. Table includes the maximum likelihood estimates and 95% confidence intervals.

								1000 Genomes		Americas		Asia		Europe		Africa	
						Variant Effect Predictor		N=1092		N=181		N=286		N=379		N=316	
Gene	SNP	Base Pair Position	Number of Transcripts	Position in protein	Amino acid change	Prediction	Weighted Average of PolyPhen and SIFT scores	MAF	R^2	MAF	R^2	MAF	R^2	MAF	R^2	MAF	R^2
ALOX12	rs138589208	6900166	1	53	V/I	neutral	0.059	0.0005	0.0014	0	-	0	-	0	-	0	-
	rs199823896	6900216	1	69	H/Q	neutral	0.141	0.0005	0.0003	0	-	0	-	0.001	0.0001	0	-
	rs148560839	6901831	1	114	R/L	neutral	0.253	0.0009	0.0006	0	-	0	-	0.003	0.0003	0	-
	rs145526271	6901839	1	117	G/R	neutral	0.019	0.0046	0.0025	0	-	0	-	0	-	0.028	0.003
	rs148602792	6901889	1	133	K/N	neutral	0.410	0.0009	0.0028	0	-	0	-	0	-	0.006	0.002
	rs114985038	6901890	1	134	D/H	neutral	0.321	0.0041	0.0028	0.003	0.003	0	-	0	-	0.016	0.004
	rs199758224	6902071	1	153	A/T	neutral	0.293	0.0005	0.0014	0	-	0	-	0	-	0	-
	rs115276151	6902295	1	189	R/H	deleterious	0.494	0.0032	0.0014	0	-	0	-	0	-	0.009	9.71E-05
	rs1126667	6902760	1	261	Q/R	neutral	0.029	0.3878	0.0432	0.343	0.092	0.463	0.026	0.397	0.022	0.317	0.016
	rs149957595	6903716	1	290	R/Q	neutral	0.034	0.0005	0.0003	0	-	0	-	0	-	0.003	0.007
	rs434473	6904934	1	322	N/S	neutral	0.056	0.3429	0.0626	0.296	0.088	0.458	0.023	0.398	0.022	0.1614	7.14E-05
	rs183466632	6904979	1	337	P/L	deleterious	0.873	0.0005	0.0085	0	-	0	-	0	-	0	-
	rs143493293	6905060	1	364	T/I	deleterious	0.761	0.0005	0.0014	0	-	0	-	0	-	0	-
	rs202168295	6908624	1	404	R/W	deleterious	0.875	0.0005	0.0003	0	-	0	-	0	-	0.003	0.007
	rs147158964	6908625	1	404	R/Q	deleterious	0.945	0.0009	0.0006	0.003	0.003	0	-	0.001	0.0001	0	-
	rs11571342	6909217	1	430	R/H	deleterious	0.641	0.0046	0.0010	0.003	0.005	0	-	0	-	0.022	0.008
	rs200546604	6909219	1	431	R/W	deleterious	0.856	0.0005	0.0085	0	-	0	-	0	-	0	-
	rs146737781	6909821	1	479	V/I	neutral	0.450	0.0005	0.0014	0	-	0	-	0	-	0.003	0.001
	rs200671464	6909840	1	485	R/K	neutral	0.250	0.0009	0.0006	0.006	0.006	0	-	0	-	0	-
	rs199517856	6909880	1	498	W/C	deleterious	0.945	0.0005	0.0003	0	-	0	-	0.001	0.0001	0	-
	rs184982217	6913119	2	532	C/S	deleterious	0.833	0.0005	0.0003	0	-	0	-	0.001	0.0001	0	-
	rs41359946	6913336	2	568	T/N	deleterious	0.507	0.0005	0.0003	0	-	0	-	0.001	0.0001	0	-
C17orf49	rs201691412	6920576	4	213	D/G	deleterious	-	0.0005	0.0014	0	-	0	-	0	-	0.003	0.001
BCL6B	rs202178963	6927450	1	76	D/E	neutral	0.392	0.0014	0.0042	0	-	0	-	0	-	0.003	0.001
	rs200898852	6927535	1	105	P/A	deleterious	0.484	0.0005	0.0003	0	-	0.002	0.0009	0	-	0	-
	rs200272841	6927587	1	122	H/L	deleterious	0.586	0.0005	0.0003	0	-	0	-	0	-	0.003	0.007
	rs202166674	6927859	1	181	P/A	neutral	0.061	0.0005	0.0003	0.003	0.003	0	-	0	-	0	-
	rs200779770	6927958	1	214	G/R	neutral	0.006	0.0009	0.0028	0	-	0	-	0	-	0.006	0.002
	rs147572675	6929827	1	314	S/L	deleterious	0.544	0.0005	0.0014	0	-	0	-	0	-	0.003	0.001
	rs181137320	6930137	1	390	G/S	deleterious	0.655	0.0005	0.0003	0	-	0	-	0	-	0.003	0.007
	rs200507334	6930290	1	403	V/M	deleterious	0.858	0.0005	0.0014	0	-	0	-	0	-	0	-
SLC16A13	rs200399927	6939760	1	20	A/V	neutral	0.328	0.0005	0.0003	0	-	0.002	0.0009	0	-	0	-
	rs181076938	6939802	1	34	F/S	neutral	0.452	0.0009	0.0006	0	-	0	-	0	-	0.006	0.014
	rs200577398	6940052	1	69	V/G	deleterious	0.714	0.0005	0.0003	0	-	0.002	0.0009	0	-	0	-
	rs182203916	6940084	1	80	R/G	deleterious	0.945	0.0005	0.0003	0	-	0.002	0.0009	0	-	0	-
	rs186831581	6940120	1	92	L/V	-	-	0.0005	0.0003	0	-	0.002	0.0009	0	-	0	-
	rs200099105	6940130	1	95	L/P	deleterious	0.758	0.0005	0.0003	0	-	0.002	0.0009	0	-	0	-
	rs201673210	6941498	1	124	P/L	deleterious	0.945	0.0005	0.0085	0	-	0	-	0	-	0.003	0.022
	rs61747374	6941726	1	200	V/E	-	-	0.0023	0.0118	0	-	0	-	0	-	0.016	0.016
	rs116931082	6941749	1	208	T/S	neutral	0.013	0.0037	0.0009	0	-	0.014	0.0015	0	-	0	-
	rs200392931	6941768	1	214	G/V	neutral	0.035	0.0005	0.0014	0	-	0	-	0.001	0.045	0	-
	rs201941350	6941803	1	226	I/V	neutral	0.011	0.0005	0.0003	0	-	0.002	0.0009	0	-	0	-
	rs142971810	6941924	1	266	R/H	deleterious	0.945	0.0005	0.0003	0	-	0	-	0.001	0.0001	0	-
	rs184174774	6941960	1	278	G/V	neutral	0.004	0.0380	0.0039	0.025	0.008	0.030	0.001	0.033	0.007	0.089	0.005
	rs151102974	6942035	1	303	A/V	neutral	0.013	0.0027	0.0188	0	-	0	-	0	-	0.019	0.032
	rs143183384	6943090	1	364	R/W	deleterious	0.530	0.0055	0.0013	0	-	0	-	0	-	0.028	0.0003
	rs201413089	6943093	1	365	D/Y	deleterious	0.945	0.0005	0.0003	0	-	0	-	0	-	0.003	0.007
	rs200850489	6943123	1	375	V/M	neutral	0.034	0.0005	0.0003	0	-	0.002	0.0009	0	-	0	-
	rs148247138	6943258	1	420	K/E	neutral	0.013	0.0014	0.0009	0	-	0	-	0.004	0.0004	0	-

						Variant Effect Predictor		1000 Genomes N=1092		Americas N=181		Asia N=286		Europe N=379		Africa N=316	
Gene	SNP	Base Pair Position	Number of Transcripts	Position in protein	Amino acid change	Prediction	Weighted Average of PolyPhen and SIFT scores	MAF	R^2	MAF	R^2	MAF	R^2	MAF	R^2	MAF	R^2
SLC16A11	rs201074878	6945080	2	445	E/A	deleterious	0.750	0.0009	0.0006	0	-	0.003	0.002	0	-	0	-
	rs75493593	**6945087**	**2**	**443**	**P/T**	**deleterious**	**0.665**	**0.0728**	**0.3549**	**0.196**	**0.851**	**0.121**	**0.954**	**0.018**	**0.754**	**0.016**	**0.058**
	rs35712788	6945201	2	405	F/L	deleterious	0.750	0.0041	0.0283	0	-	0	-	0.001	0.0001	0.019	0.076
	rs75418188	**6945483**	**2**	**340**	**G/S**	**neutral**	**0.083**	**0.0714**	**0.3644**	**0.196**	**0.851**	**0.1154**	**1**	**0.0185**	**0.754**	**0.016**	**0.058**
	rs187584131	6945657	2	282	V/M	neutral	0.406	0.0014	0.0094	0.003	0.005	0	-	0	-	0.006	0.017
	rs191656427	6945674	2	276	G/V	deleterious	0.532	0.0018	0.0004	0	-	0	-	0	-	0.013	0.003
	rs200366816	6945797	2	235	G/D	deleterious	0.492	0.0055	0.0013	0	-	0	-	0	-	0.028	0.0003
	rs199749576	6946252	2	139	A/T	neutral	0.442	0.0005	0.0003	0	-	0.002	0.0009	0	-	0	-
	rs13342692	**6946287**	**2**	**127**	**D/G**	**neutral**	**0.383**	**0.1571**	**1**	**0.218**	**1**	**0.115**	**1**	**0.025**	**1**	**0.358**	**1**
	rs117767867	**6946330**	**2**	**113**	**V/I**	**neutral**	**0.064**	**0.0710**	**0.3676**	**0.193**	**0.868**	**0.115**	**1**	**0.018**	**0.754**	**0.016**	**0.058**
	rs75636181	6946357	2	104	A/S	neutral	0.034	0.0325	0.0118	0.047	0.030	0	-	0.070	0.0001	0.003	0.001
	rs77302172	6946392	2	92	S/N	deleterious	0.847	0.0165	0.1158	0.008	0.015	0	-	0	-	0.057	0.146
	rs199869009	6946662	2	81	S/R	deleterious	0.744	0.0005	0.0003	0	-	0.002	0.0009	0	-	0	-
	rs201214748	6946826	2	27	P/A	deleterious	0.533	0.0005	0.0003	0	-	0.002	0.0009	0	-	0	-
CLEC10A	rs115347328	6978501	2	275	G/R	deleterious	0.541	0.0046	0.0245	0	-	0	-	0	-	0.009	0.003
	rs146480775	6978504	2	274	H/Y	deleterious	0.945	0.0027	0.0004	0.003	0.005	0	-	0.007	0.0006	0	-
	rs200496384	6978516	2	270	D/N	neutral	0.029	0.0005	0.0003	0.003	0.003	0	-	0	-	0	-
	rs35101468	6979117	2	203	A/G	neutral	0.322	0.0032	0.0034	0.003	0.003	0	-	0	-	0.013	6.70E-05
	rs200740525	6979189	2	179	T/S	neutral	0.351	0.0005	0.0014	0	-	0.002	0.009	0	-	0	-
	rs112729653	6979331	2	165	C/R	deleterious	0.883	0.0018	0.0057	0	-	0	-	0	-	0.006	0.001
	rs201647706	6979363	2	154	Q/R	deleterious	0.734	0.0005	0.0003	0	-	0	-	0	-	0.003	0.007
	rs36097216	6980061	2	115	R/W	deleterious	0.858	0.0005	0.0003	0	-	0	-	0.001	0.0001	0	-
	rs35318160	6980105	2	100	T/M	neutral	0.059	0.0188	0.0017	0.017	0.007	0.016	1.08E-05	0.025	0.002	0.013	0.003
	rs200378911	6980258	2	78	N/T	deleterious	0.945	0.0009	0.0002	0	-	0	-	0	-	0.006	0.001
	rs16956478	6980273	2	73	R/K	neutral	0.150	0.0444	0.0236	0.022	0.013	0.016	1.08E-05	0.02507	0.002	0.136	0.016
	rs147504959	6981331	2	57	V/M	deleterious	0.945	0.0037	0.0009	0	-	0.0140	0.0015	0	-	0	-
	rs90951	6981397	2	35	C/R	neutral	0.345	0.4895	0.0290	0.329	0.027	0.4948	0.0091	0.356	6.19E-05	0.0981	0.0007
	rs78714016	6981403	2	33	R/C	deleterious	0.882	0.0023	0.0002	0.003	0.005	0	-	0.005	0.0005	0	-
	rs142535416	6982125	2	3	R/G	deleterious	0.574	0.0005	0.0003	0.003	0.003	0	-	0	-	0	-
ASGR2	rs201316903	7004913	4	306	A/V	neutral	0.012	0.0005	0.0003	0	-	0	-	0.001	0.0001	0	-
	rs35381090	7005058	4	258	N/Y	deleterious	0.504	0.0037	0.0076	0.003	0.003	0	-	0	-	0.013	0.016
	rs150471603	7005508	4	224	I/T	deleterious	0.656	0.0009	0.0006	0	-	0.003	0.002	0	-	0	-
	rs199839502	7010380	4	201	A/V	deleterious	0.945	0.0005	0.0003	0	-	0.002	0.001	0	-	0	-
	rs2304978	7012079	3	85	G/R	neutral	0.001	0.2376	0.0074	0.345	0.046	0.271	7.25E-06	0.276	0.0007	0.060	0.009
	rs200704590	7017483	2	26	P/L	neutral	0.242	0.0005	0.0003	0.003	0.003	0	-	0	-	0	-
	rs200104102	7017559	5	1	M/V	deleterious	0.919	0.0005	0.0003	0	-	0	-	0.001	0.0001	0	-

Supplementary Table 12: Coding SNPs in genes near *SLC16A11* with ancestry-specific minor allele frequency and correlation with rs13342692. Data was accessed on November 19, 2012 from the 1000 Genomes site (www.1000genomes.org). For each SNP, the gene, base pair position, and number of gene transcripts observed with a SNP at the position are listed. The Variant Effect Predictor was applied to all SNPs from each gene and the non-synonymous coding SNPs were extracted. Weighted averages of the Polyphen and SIFT scores were used to predict if each variant is neutral or deleterious. Minor allele frequencies and correlation with rs13342692, calculated using PLINK v. 1.07, are listed for the entire 1000 Genomes Phase 1 sample (N=1092) and by ancestry groups: Americas (people with Mexican ancestry in Los Angeles, California; Puerto Ricans in Puerto Rico; Colombians in Medellin, Colombia), Asia (Han Chinese in Beijing, China; Han Chinese South, China; Japanese in Tokyo, Japan), Europe (Utah residents with ancestry from Northern and Western Europe; Finnish in Finland; British from England and Scotland, UK; Iberian populations in Spain; Toscani in Italia) and Africa (people with African ancestry in Southwest United States; Luhya in Webuye, Kenya).

	SLC16A11 Reference Haplotype	
Lipid Classes	**Mean SLC16A11:Control**	**P-Value**
Lysophosphatidylcholines	0.905	1.95×10^{-3}
Lysophosphatidylethanolamines	0.92	3.13×10^{-2}
Phosphatidylcholines	0.982	8.94×10^{-2}
Sphingomyelins	0.942	3.91×10^{-3}
Cholesterol Esters	0.855	9.77×10^{-4}
Diacylglycerols	1.13	7.81×10^{-3}
Triacylglycerols	1.34	7.60×10^{-12}

Supplementary Table 13: For each lipid class, the Wilcoxon signed rank test was performed using the median value of each metabolite, calculated for SLC16A11 and that of the control samples. The p-value of a two-sided test is reported along with the mean ratio of the change for SLC16A11 over the controls. More details are available in Online Methods.

	SLC16A11 Reference Haplotype		
Metabolic Pathway or Class	**P-Value**	**FDR**	**Effect**
Alanine, Asparate, and Glutamate Metabolism	0.0001	0.0016	Depleted
Aminoacyl-tRNA Biosynthesis	0.0011	0.0061	Depleted
Citric Acid Cycle	0.0002	0.0015	Depleted
Cyanoamino Acid Metabolism	0.0101	0.0414	Depleted
Glyoxylate and Dicarboxylate Metabolism	0.0001	0.0011	Depleted
Nitrogen Metabolism	0.0001	0.0014	Depleted
Proximal Tubule Bicarbonate Reclamation	0.0252	0.0015	Depleted
Purine Metabolism	0.0002	0.0015	Depleted
Lysophophatidylcholines	0.0027	0.0105	Depleted
Phosphatidylcholines	0.0128	0.0452	Depleted
Cholesterol Esters	0.0001	0	Depleted
Sphingomyelins	0.0256	0.0755	Depleted
Diacylglycerols	0.0003	0.0007	Enriched
Triacylglycerols	0.0001	0	Enriched

Supplementary Table 14: Metabolite pathway enrichment in HeLa cells expressing SLC16A11 as compared to controls. For this analysis, all KEGG pathways from the human reference set and eight additional classes of metabolites covering lipid sub-types and carnitines were used to assess enrichment or depletion. Pathways and classes were pruned to reflect only those metabolites measured by our metabolite profiling platform, and a pathway had to have at least six measurable metabolites in this dataset in order to be scored. For each pathway or metabolite class, enrichment was computed using the unweighted Kolmogorov-Smirnov statistic. P-values for enrichment scores were computed from an empirically determined null distribution and then corrected for false discovery rate (FDR). Only those pathways with $P \leq 0.05$ and FDR ≤ 0.25 are shown.

Supplementary Note

Association at *KCNQ1*

The top associated SNP at *KCNQ1* within SIGMA is a known associated variant originally identified in East Asians[83]. *KCNQ1* also harbors a second associated site previously identified in Europeans[84] that is independent of the East Asian SNP. Conditional analysis based on the top SIGMA SNP identified a potential third independent association (rs139647931; P=5.3×10^{-8}; OR=0.78 [0.70-0.86]; Supplementary Figure 3b) that appears independent of both the East Asian (r^2=0.056 to rs2237897) and European (r^2=0.028 to rs231362) SNPs. Further work is needed to determine the relationship of these associations to one another; ideally, one should conduct parent-of-origin and cross ethnic analyses at all these sites.

Examination and test of Neandertal introgression as source for *SLC16A11* risk haplotype

We examined a recently generated genome sequence of a Neandertal (or a close Neandertal relative) obtained from a foot phalanx bone from Denisova Cave[85] (Prüfer, Shunkov, Derevianko, and Pääbo, unpublished) for affinity to the *SLC16A11* and chromosome 11p15.5 associations. For the top two associated SNPs in chromosome 11p15.5 (rs11564732 and rs192912194), this sequence contains the ancestral chimpanzee alleles, and thus gives no evidence of introgression at that locus. In contrast, the associated synonymous and missense SNPs in *SLC16A11* (rs13342232, rs13342692, rs117767867, rs75418188, and rs75493593) are all present in homozygous form in this Neandertal sequence.

To determine whether the risk haplotype in modern humans at *SLC16A11* is introgressed from Neandertals, we examined 1,000 Genomes[86] samples that are homozygous for the synonymous and missense variants associated to T2D. We identified a region spanning 90 kb that showed low divergence between the Neandertal sequence and the homozygous 1,000 Genomes samples. To conservatively estimate the putatively introgressed haplotype boundaries for confidently Neandertal-derived ancestry, we identified variants within this 90 kb region that are absent in 1,000 Genomes African YRI and LWK populations. These variants are likely Neandertal-derived since Neandertal admixture occurred in the ancestors of non-Africans[87]. This conservative haplotype spans 73 kb and a genetic distance of .1196 cM.

We examined whether this haplotype was recently introgressed via admixture with Neandertals or whether it may have been segregating in the ancestors of modern humans and Neandertals. To do this, we tested a null model of a haplotype with genetic length .1196 cM segregating in modern humans without recombination since the split with Neandertals ~9,000 generations ago. We conservatively ignore the possibility of recombination in the last 15,000 years ≈ 517 generations[88] and of recombination within the Neandertal lineage. This null model gives P = exp(-.1196 / 100 × (9,000 – 517)) = 3.9×10^{-5}, suggesting that the haplotype is very likely to have entered the modern human population via relatively recent admixture with Neandertals[89].

Divergence time estimate for *SLC16A11* risk haplotype and Neandertal sequence

To estimate the divergence of the risk haplotype at *SLC16A11* to the sequenced Neandertal genome (Prüfer, Shunkov, Derevianko, and Pääbo, unpublished), we used the computationally phased 1000 Genomes phase I data. We examined an individual in the CEU population (NA11930) that is heterozygous for the synonymous and four missense variants in *SLC16A11* that we identified as associated to T2D risk. We also identified and performed analysis of 15 individuals that are homozygous for these variants (and thus, homozygous for the risk haplotype). One of these 15 individuals is listed as part of the African American ASW population, however principal components analysis shows that this individual is likely that to be Mexican-descent (Giulio Genovese, unpublished).

We restricted our analysis to the conservatively estimated haplotype boundaries spanning 73 kb. In this 73kb region, we attempted to estimate the divergence of the risk haplotype to the Neandertal sequence by:

- Examining all sites at which genotypes were called in this region in the 1000 Genomes data. At sites that had no genotype calls, we used the base found in the human genome reference *hg19*.
- Requiring sites to have a confident ancestral allele determined according to the 6-primate EPO alignment[90].
- Further restricting our analysis to sites at which the Neandertal genotype passes quality filters and is determined to be homozygous for the ancestral allele. Our filters consisted of a Map20 field of 1, genotype quality ≥ 30 and read depth between 31 and 79.

At these sites, we count the number of sites at which the risk haplotype contains the derived allele. Because we restrict to sites that are homozygous ancestral in the Neandertal sequence, these sites correspond to mutations that arose on the risk haplotype after it diverged from both the Neandertal haplotype. In total, we examined L=29110 sites. At these sites, for each risk haplotype, we count (d) the number of sites at which the haplotype carries the derived state.

We compute a point estimate of the time (as a fraction of the human-Rhesus Macaque TMRCA) since the risk haplotype diverged from the lineage leading to the Neandertal haplotype as

$$T_f = \frac{2d}{d_m L}$$

Here, d_m is the expected fraction of pairwise differences between the human reference genome and the Macaque genome at this locus. Specifically, d_m is the product of the per-base-mutation rate and the human-macaque TMRCA at this locus multiplied by two (doubling is necessary because we compare two lines of descent from the human-macaque ancestor). We estimated d_m from the number of observed human-macaque differences at this locus using a Jukes-Cantor model[91]. We restricted our analysis to sites that have confident assignment of the base in human and macaque.

We converted this scaled time T_f to years assuming a human-macaque genetic divergence of 25 million years[92] and obtained confidence intervals on this point estimate using the relationship between the Poisson and chi-squared distributions[93]. Supplementary Table 10 shows the divergence times for each of the risk haplotypes we examined. The mean divergence time (obtained by averaging the point estimates for each risk haplotype) is roughly 250k years before present (kY BP). Three of the CHB haplotypes have considerably higher divergence times than the others haplotypes, and these appear to contain crossovers that artificially inflate their divergences.

Recent estimates for Neandertal and modern human population split times are between 170-700 kY BP[94]; because genetic divergence time is always greater than population split times, the risk haplotype is most consistent with a scenario in which recent gene flow from Neandertals introduced the risk haplotype into modern human populations outside Africa.

As a control, we also computed divergence time estimates in the same region for individuals that do not carry the risk haplotype. For this analysis, we randomly selected two individuals from each of the populations that contain samples that are homozygous for the risk haplotype (i.e., populations from Supplementary Table 10). We obtained a mean divergence time estimate of roughly 677 kY BP —nearly three times older than for the risk haplotype (Supplementary Table 11), consistent with the risk haplotype being closely related to Neandertals.

The above analysis assumes that there are no phase switch errors in the 1000 Genomes haplotypes. Another feature of this analysis is that, by restricting to sites where the Neandertal genome is homozygous for the ancestral allele (thus only counting mutations on the modern human side of the tree), our analysis does not require knowledge of the Neandertal haplotype in this region. It also does not depend on the topology of the

tree that relates the two Neandertal haplotypes to the introgressed Neandertal haplotype. And finally, by considering only mutations that arose on the modern human side of the tree, the analysis avoids the complexity associated with the fact that the Neandertal bone is old and therefore has had less time to accumulate mutations than present-day modern human individuals.

Subconsortia Authors

Broad Genomics Platform: Adal Abebe[1], Justin Abreu[1], Kristin Anderka[1], Scott Anderson[1], Sarah Babchuck[1], Maria Baco[1], Samira Bahl[1], Danielle Bain[1], Kylee Bergin[1], Amy Biasella[1], Bill Biggs[1], Brendan Blumenstiel[1], Harry Bochner[1], Claude Bonnet[1], Wendy Brodeur[1], Joseph BuAbbud[1], Emily C. Davis[1], Jody Camarata[1], Jason Carey[1], Mauricio Carneiro[1], Brynne Cassidy[1], Clinton Chalk[1], Sheridon Channer[1], Andrew Cheney[1], Michelle Cipicchio[1], Kristen Connolly[1], Matthew Coole[1], Maura Costello[1], Miguel Covarrubias[1], Cassandra Crawford[1], Lindsay Croshier[1], Michael Dasilva[1], Matthew Defelice[1], Tim Desmet[1], Alexandra Dimitriou[1], Katerina Dimitriou[1], Michael Dinsmore[1], Danielle Dionne[1], Sheli Dookran[1], Teni Dowdell[1], Phil Dunlea[1], Cassandra Elie[1], M. Erik Husby[1], Emelia Failing[1], Yossi Farjoun[1], Timothy Fennell[1], Damien Fenske-Corbiere[1], Steven Ferriera[1], Sheila Fisher[1], Jennifer Franklin[1], Paul Frere[1], Shemifhar Freytes[1], Dennis Friedrich[1], Stacey Gabriel[1], Diane Gage[1], Christina Gearin[1], Jeff Gentry[1], Lizz Gottardi[1], Alexander Graff[1], George Grant[1], Lisa Green[1], Jonna Grimsby[1], Namrata Gupta[1], Kunsang Gyaltsen[1], Bertrand Haas[1], Susanna Hamilton[1], Maegan Harden[1], Ryan Hegarty[1], Desiree Hernandez[1], Andrew Hollinger[1], Laurie Holmes[1], Tracey Honan[1], Tom Howd[1], Maria Jenkins[1], Ryan Johnson[1], Andrew Johnson[1], Kevin Joseph[1], Fontina Kelley[1], Edward Kelliher[1], Cristyn Kells[1], Amanda Kennedy[1], Sharon Kim[1], Kevinson Kim[1], Samuel Kim[1], Catherine King[1], Charles Kivolowitz[1], Jessica Klopp[1], Anna Koutoulas[1], Massami Laird[1], Katie Larkin[1], Katie Larsson[1], Zach Leber[1], Matthew Lee[1], James Lee[1], Niall Lennon[1], Frances Letendre[1], Tsamla Lhanyitsang[1], Shuqiang Li[1], Kenneth Livak[1], Hayley Lyon[1], Alyssa Macbeth[1], Vasilia Magnisalis[1], Tsheko Makuwa[1], Lauren Margolin[1], Tamara Mason[1], Scott Matthews[1], Michael McCowan[1], Susan McDonough[1], Kaitlyn McGrath[1], James Meldrim[1], Atanas Mihalev[1], Mariela Mihaleva[1], Tyler Miselis[1], Ruchi Munshi[1], Gregory Nakashian[1], Jillian Nolan[1], Nyima Norbu[1], Deborah Norman Farlow[1], Sam Novod[1], Robert Onofrio[1], Veronika Osherov[1], Melissa Parkin[1], Danielle Perrin[1], Caroline Petersen[1], Prapti Pokharel[1], Eliot Polk[1], Samuel Pollock[1], Shannon Power[1], Katelin Pratt[1], Mark Puppo[1], Anthony Rachupka[1], Howard Rafal[1], Ashley Ray[1], Brian Reilly[1], Scott Rich[1], Dana Robbins[1], Joseph Rose[1], Carsten Russ[1], Dennis Ryan[1], Surayya Sana[1], Ahmed Sandakli[1], Michael Saylor[1], Sampath Settipalli[1], Philip Shapiro[1], Kara Slowik[1], Cherylyn Smith[1], Brian Sogoloff[1], Carrie Sougnez[1], Sharon Stavropoulos[1], Gregory Stoneham[1], Jordan Sullivan[1], Katherine Sullivan[1], Danielle Sutherby[1], Frederick Ta[1], Alvin Tam[1], Bradley Taylor[1], Jon Thompson[1], Kathleen Tibbetts[1], Charlotte Tolonen[1], Kristina Tracy[1], Austin Tzou[1], Gina Vicente[1], Fernando Viloria[1], Andy Vo[1], Louisa Walker[1], John Walsh[1], Cole Walsh[1], Kendra West[1], Emily Wheeler[1], Jane Wilkinson[1], Michael Wilson[1], Ellen Winchester[1], Jennifer Wineski[1], Betty Woolf[1], Chin-Lee Wu[1], Alec Wysoker[1], Qing Yu[1], David Zdeb[1], Andrew Zimmer[1]

The T2D-GENES Consortium: Gonçalo Abecasis[2], Marcio Almeida[3], David Altshuler[4,5,6,7,8,9,10], Jennifer L. Asimit[11], Gil Atzmon[12], Mathew Barber[13], Nicola L. Beer[14], Graeme I. Bell[13,15], Jennifer Below[16], Tom Blackwell[2], John Blangero[3], Michael Boehnke[2], Donald W. Bowden[17,18,19,20], Noël Burtt[4], John Chambers[21,22,23], Han Chen[24], Peng Chen[25], Peter S.Chines[26], Sungkyoung Choi[27], Claire Churchhouse[4], Pablo Cingolani[28], Belinda K. Cornes[29], Nancy Cox[13,15], Aaron G. Day-Williams[11], Ravindranath Duggirala[3], Josée Dupuis[24], Thomas Dyer[3], Shuang Feng[2], Juan Fernandez-Tajes[30], Teresa Ferreira[30], Tasha E. Fingerlin[31], Jason Flannick[4,6], Jose Florez[4,6,7], Pierre Fontanillas[4], Timothy M. Frayling[32], Christian Fuchsberger[2], Eric R. Gamazon[15], Kyle Gaulton[30], Saurabh Ghosh, Anna Gloyn[14], Robert L. Grossman[15,33], Jason Grundstad[33], Craig Hanis[16], Allison Heath[33], Heather Highland[16], Momoko Hirokoshi[30], Ik-Soo Huh[27], Jeroen R. Huyghe[2], Kamran Ikram[34,29,35,36], Kathleen A. Jablonski[37], Young Jin Kim[38], Goo Jun[2], Norihiro Kato[39], Jayoun Kim[27], C. Ryan King[40], Jaspal Kooner[22,23,41], Min-Seok Kwon[27], Hae Kyung Im[40], Markku Laakso[42] , Kevin Koi-Yau Lam[25], Jaehoon Lee[27], Selyeong Lee[27], Sungyoung Lee[27], Donna M. Lehman[43], Heng Li[4], Cecilia M. Lindgren[30], Xuanyao Liu[25,44], Oren E. Livne[13], Adam E. Locke[2], Anubha Mahajan[30], Julian B. Maller[30,45], Alisa K. Manning[4]. Taylor J. Maxwell[16], Alexander Mazoure[46],

Mark I. McCarthy[30,14,47], James B. Meigs[7,48], Byungju Min[27], Karen L. Mohlke[49], Andrew Morris[50], Solomon Musani[51], Yoshihiko Nagai[46], Maggie C.Y. Ng[17,18], Dan Nicolae[13,15,52], Sohee Oh[27], Nicholette Palmer[17,18,19], Taesung Park[27], Toni I. Pollin[53], Inga Prokopenko[30,54], David Reich[4,5], Manuel A. Rivas[30], Laura J. Scott[2], Mark Seielstad[55], Yoon Shin Cho[56], E-Shyong Tai[34,25,57], Xueling Sim[2], Robert Sladek[46,58], Philip Smith[59], Ioanna Tachmazidou[11], Tanya M. Teslovich[2], Jason Torres[13,15], Vasily Trubetskoy[13,15], Sara M. Willems[60], Amy L. Williams[4,5], James G. Wilson[61], Steven Wiltshire[62], Sungho Won[63], Andrew R. Wood[32], Wang Xu[57], Yik Ying Teo[64,65,66,67,68], Joon Yoon[27], Jong-Young Lee[69], Matthew Zawistowski[2], Eleftheria Zeggini[11], Weihua Zhang[21], Sebastian Zöllner[2,70]

[1]The Genomics Platform, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA.

[2]Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA.

[3]Department of Genetics, Texas Biomedical Research Institute, San Antonio, Texas 78227, USA.

[4]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA.

[5]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

[6]Center for Human Genetic Research and Diabetes Research Center (Diabetes Unit), Massachusetts General Hospital, Boston 02114, Massachusetts, USA.

[7]Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA.

[8]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.

[9]Department of Molecular Biology, Harvard Medical School, Boston, Massachusetts 02114, USA.

[10]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

[11]Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, UK.

[12]Department of Medicine, Department of Genetics, Albert Einstein College of Medicine, Bronx, New York 10461, USA.

[13]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA.

[14]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, OX3 7LJ, UK.

[15]Department of Medicine, University of Chicago, Chicago, Illinois 60637, USA.

[16]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas 77030, USA.

[17]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA.

[18]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA.

[19]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA.

[20]Internal Medicine-Endocrinology, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA.

[21]Department of Epidemiology and Biostatistics, Imperial College London, London SW7 2AZ, UK.

[22]Imperial College Healthcare NHS Trust, London W2 1NY, UK.

[23]Ealing Hospital National Health Service (NHS) Trust, Middlesex UB1 3HW, UK.

[24]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts 02115, USA.

[25]Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117597, Singapore.

[26]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA.

[27]Seoul National University, Seoul 110-799, South Korea.

[28]McGill Centre for Bioinformatics, McGill University, Montréal, Quebec, H3G 0B1, Canada.

[29]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore 168751, Singapore.

[30]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.

[31]Department of Epidemiology, Colorado School of Public Health, Aurora, Colorado 80045, USA.

[32]Genetics of Complex Traits, University of Exeter Medical School, Exeter, EX4 4SB UK.

[33]Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois 60637, USA.

[34]Duke National University of Singapore Graduate Medical School, Singapore 169857, Singapore.

[35]Department of Ophthalmology, National University of Singapore and National University Health System, Singapore 119228, Singapore.

[36]Department of Ophthalmology, Erasmus Medical Center, Rotterdam 3000 CA, the Netherlands.

[37]The Biostatistics Center, George Washington University, Rockville, Maryland 20852, USA.

[38]Department of Neurology, Konkuk University School of Medicine, Seoul 143-701, South Korea.

[39]Department of Gene Diagnostics and Therapeutics, Research Institute, National Center for Global Health and Medicine, Tokyo 162-8655, Japan.

[40]Department of Health Studies, University of Chicago, Chicago, Illinois 60637, USA.

[41]National Heart and Lung Institute (NHLI), Imperial College London, Hammersmith Hospital, London W12 0HS, UK.

[42] Department of Medicine, University of Eastern Finland, Kuopio Campus and Kuopio University Hospital, FI-70211 Kuopio, Finland.

[43]Division of Clinical Epidemiology, Department of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78229, USA.

[44]Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore 117456, Singapore.

[45]Department of Statistics, University of Oxford, Oxford, OX1 3TG UK.

[46]McGill University, Montréal, Québec H3A 0G4, Canada.

[47]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Headington OX3 7LE, UK.

[48]General Medicine Division, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.

[49]Department of Genetics, University of North Carolina-Chapel Hill, Chapel Hill, North Carolina 27599, USA.

[50]Department of Genetic Medicine, Manchester Academic Health Sciences Centre, Manchester M13 9NT, UK.

[51]Department of Medicine, University of Mississippi Medical Center, Jackson, Mississippi 39126, USA.

[52]Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA.

[53]Department of Medicine, Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA.

[54]Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, 751 05 Uppsala, Sweden.

[55]University of California San Francisco, San Francisco, California 94143, USA.

[56]Department of Biomedical Science, Hallym University, Chuncheon, Gangwon-do, 200-702 South Korea.

[57]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore.

[58]Department of Medicine, Royal Victoria Hospital, Montréal, Québec H3A 1A1, Canada.

[59]National Institute of Diabetes and Digestive and Kidney Disease, National Institutes of Health, Bethesda, MD 20817, USA.

[60]Department of Genetic Epidemiology, Erasmus Medical Center, Rotterdam 3000 CA, the Netherlands.

[61]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi 39216, USA.

[62]Centre for Medical Research, Western Australian Institute for Medical Research, The University of Western Australia, Nedlands WA 6008, Australia.

[63]Chung-Ang University, Seoul 156-756, South Korea.

[64]Department of Epidemiology and Public Health, National University of Singapore, Singapore 117597, Singapore.

[65]Centre for Molecular Epidemiology, National University of Singapore, Singapore 117456, Singapore.

[66]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore.

[67]Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore 117456, Singapore.

[68]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore.

[69]Center for Genome Science, Korea National Institute of Health, Osong Health Technology Administration Complex, Chungcheongbuk-do, 363-951, South Korea.

[70]Department of Psychiatry, University of Michigan, Ann Arbor, Michigan 48109, USA.

References

30 Kolonel, L. N. *et al*. A Multiethnic Cohort in Hawaii and Los Angeles: Baseline Characteristics. *American Journal of Epidemiology* **151**, 346-357 (2000).

31 Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLoS genetics* **2**, e190, doi:10.1371/journal.pgen.0020190 (2006).

32 An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:http://www.nature.com/nature/journal/v491/n7422/abs/nature11632.html#supplementary-information (2012).

33 Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics* **44**, 955-959, doi:http://www.nature.com/ng/journal/v44/n8/abs/ng.2354.html#supplementary-information (2012).

34 Williams, Amy L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of Many Thousands of Genotyped Samples. *American journal of human genetics* **91**, 238-251 (2012).

35 Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS genetics* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

36 Zaitlen, N. *et al*. Informed Conditioning on Clinical Covariates Increases Power in Case-Control Association Studies. *PLoS genetics* **8**, e1003032, doi:10.1371/journal.pgen.1003032 (2012).

37 Villalpando, S. *et al*. Prevalence and distribution of type 2 diabetes mellitus in Mexican adult population: a probabilistic survey. *Salud Pública de México* **52**, S19-S26 (2010).

38 Purcell, S. *et al*. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559-575, doi:10.1086/519795 (2007).

39 Devlin, B. & Roeder, K. Genomic Control for Association Studies. *Biometrics* **55**, 997-1004, doi:10.1111/j.0006-341X.1999.00997.x (1999).

40 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, doi:10.1101/gr.094052.109 (2009).

41 Li, J. Z. *et al*. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* **319**, 1100-1104, doi:10.1126/science.1153717 (2008).

42 Baran, Y. *et al*. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **28**, 1359-1367, doi:10.1093/bioinformatics/bts144 (2012).

43 Reich, D. *et al*. Reconstructing Native American population history. *Nature* **488**, 370-374, doi:http://www.nature.com/nature/journal/v488/n7411/abs/nature11258.html#supplementary-information (2012).

44 Behar, D. M. *et al*. The genome-wide structure of the Jewish people. *Nature* **466**, 238-242, doi:http://www.nature.com/nature/journal/v466/n7303/abs/nature09103.html#supplementary-information (2010).

45 Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58, doi:http://www.nature.com/nature/journal/v467/n7311/abs/nature09298.html#supplementary-information (2010).

46 Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2012).

47 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191, doi:10.1093/bioinformatics/btq340 (2010).

48 Pruim, R. J. *et al*. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-2337, doi:10.1093/bioinformatics/btq419 (2010).

49 Haiman, C. A. *et al*. Consistent Directions of Effect for Established Type 2 Diabetes Risk Variants Across Populations: The Population Architecture using Genomics and Epidemiology (PAGE) Consortium. *Diabetes* **61**, 1642-1647, doi:10.2337/db11-1296 (2012).

50 Stančáková, A. *et al*. Changes in Insulin Sensitivity and Insulin Release in Relation to Glycemia and Glucose Tolerance in 6,414 Finnish Men. *Diabetes* **58**, 1212-1221, doi:10.2337/db08-1607 (2009).

51 Atzmon, G. *et al*. Abraham's Children in the Genome Era: Major Jewish Diaspora Populations Comprise Distinct Genetic Clusters with Shared Middle Eastern Ancestry. *The American Journal of Human Genetics* **86**, 850-859, doi:http://dx.doi.org/10.1016/j.ajhg.2010.04.015 (2010).

52 Taylor, H. A. *et al*. Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. *Ethnicity & disease* **15**, S6-4-17 (2005).

53 Yu, H., Bowden, D. W., Spray, B. J., Rich, S. S. & Freedman, B. I. Linkage analysis between loci in the renin-angiotensin axis and end-stage renal disease in African Americans. *Journal of the American Society of Nephrology* **7**, 2559-2564 (1996).

54 Chahal, N. S. *et al*. Ethnicity-related differences in left ventricular function, structure and geometry: a population study of UK Indian Asian and European white subjects. *Heart* **96**, 466-471, doi:10.1136/hrt.2009.173153 (2010).

55 Chahal, N. S. *et al*. Does subclinical atherosclerosis burden identify the increased risk of cardiovascular disease mortality among United Kingdom Indian Asians? A population study. *American heart journal* **162**, 460-466, doi:http://dx.doi.org/10.1016/j.ahj.2011.06.018 (2011).

56 Methodology of the Singapore Indian Chinese Cohort (SICC) Eye Study: Quantifying ethnic variations in the epidemiology of eye diseases in Asians. *Ophthalmic Epidemiology* **16**, 325-336, doi:doi:10.3109/09286580903144738 (2009).

57 Cho, Y. S. *et al*. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nature genetics* **41**, 527-534, doi:http://www.nature.com/ng/journal/v41/n5/suppinfo/ng.357_S1.html (2009).

58 Hughes, K. *et al*. Cardiovascular diseases in Chinese, Malays, and Indians in Singapore. II. Differences in risk factor levels. *Journal of Epidemiology and Community Health* **44**, 29-35, doi:10.1136/jech.44.1.29 (1990).

59 Tan, C. E., Emmanuel, S. C., Tan, B. Y. & Jacob, E. Prevalence of diabetes and ethnic differences in cardiovascular risk factors. The 1992 Singapore National Health Survey. *Diabetes Care* **22**, 241-247, doi:10.2337/diacare.22.2.241 (1999).

60 Hughes, K., Aw, T. C., Kuperan, P. & Choo, M. Central obesity, insulin resistance, syndrome X, lipoprotein(a), and cardiovascular risk in Indians, Malays, and Chinese in Singapore. *Journal of Epidemiology and Community Health* **51**, 394-399, doi:10.1136/jech.51.4.394 (1997).

61 Cutter, J., Tan, B. Y. & Chew, S. K. Levels of cardiovascular disease risk factors in Singapore following a national intervention programme *Bulletin of the World Health Organization : the International Journal of Public Health 2001* **79**, 908-915 (2001).

62 Mitchell, B. D. *et al*. Genetic and Environmental Contributions to Cardiovascular Risk Factors in Mexican Americans: The San Antonio Family Heart Study. *Circulation* **94**, 2159-2170, doi:10.1161/01.cir.94.9.2159 (1996).

63 Hunt, K. J. *et al*. Genome-Wide Linkage Analyses of Type 2 Diabetes in Mexican Americans: The San Antonio Family Diabetes/Gallbladder Study. *Diabetes* **54**, 2655-2662, doi:10.2337/diabetes.54.9.2655 (2005).

64 Coletta, D. K. *et al*. Genome-Wide Linkage Scan for Genes Influencing Plasma Triglyceride Levels in the Veterans Administration Genetic Epidemiology Study. *Diabetes* **58**, 279-284, doi:10.2337/db08-0491 (2009).

65 Knowler, W. C. *et al*. The Family Investigation of Nephropathy and Diabetes (FIND): Design and methods. *Journal of diabetes and its complications* **19**, 1-9 (2005).

66 Hanis, C. L. *et al*. Diabetes among Mexican Americans in Starr County, Texas. *American Journal of Epidemiology* **118**, 659-672 (1983).

67 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).

68 DePristo, M. A. *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:http://www.nature.com/ng/journal/v43/n5/abs/ng.806.html#supplementary-information (2011).

69 Hankin, J. H. *et al*. Singapore Chinese Health Study: Development, Validation, and Calibration of the Quantitative Food Frequency Questionnaire. *Nutrition and cancer* **39**, 187-195, doi:10.1207/S15327914nc392_5 (2001).

70 Rubicz, R. *et al*. A Genome-Wide Integrative Genomic Study Localizes Genetic Factors Influencing Antibodies against Epstein-Barr Virus Nuclear Antigen 1 (EBNA-1). *PLoS genetics* **9**, e1003147, doi:10.1371/journal.pgen.1003147 (2013).

71 (U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA, 2011).

72 Risch, N. & Merikangas, K. The Future of Genetic Studies of Complex Human Diseases. *Science* **273**, 1516-1517, doi:10.2307/2891043 (1996).

73 Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology* **305**, 567-580, doi:10.1006/jmbi.2000.4315 (2001).

74 Beitz, E. TEXtopo: shaded membrane protein topology plots in LaTEX2ε. *Bioinformatics* **16**, 1050-1051, doi:10.1093/bioinformatics/16.11.1050 (2000).

75 Geiss, G. K. *et al*. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotech* **26**, 317-325, doi:http://www.nature.com/nbt/journal/v26/n3/suppinfo/nbt1385_S1.html (2008).

76 Barretina, J. *et al*. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-307, doi:http://www.nature.com/nature/journal/v483/n7391/abs/nature11003.html#supplementary-information (2012).

77 Li, C. & Hung Wong, W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology* **2**, research0032.0031 - research0032.0011 (2001).

78 Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193, doi:10.1093/bioinformatics/19.2.185 (2003).

79 Yang, X. *et al*. A public genome-scale lentiviral expression library of human ORFs. *Nat Meth* **8**, 659-661, doi:http://www.nature.com/nmeth/journal/v8/n8/abs/nmeth.1638.html#supplementary-information (2011).

80 Jiang, D., Zhao, L., Clish, C. B. & Clapham, D. E. Letm1, the mitochondrial Ca2+/H+ antiporter, is essential for normal glucose metabolism and alters brain function in Wolf–Hirschhorn syndrome. *Proceedings of the National Academy of Sciences*, doi:10.1073/pnas.1308558110 (2013).

81 Mootha, V. K. *et al*. PGC-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* **34**, 267-273, doi:http://www.nature.com/ng/journal/v34/n3/suppinfo/ng1180_S1.html (2003).

82 Subramanian, A. *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

83 Unoki, H. *et al*. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nature genetics* **40**, 1098-1102, doi:http://www.nature.com/ng/journal/v40/n9/suppinfo/ng.208_S1.html (2008).

84 Voight, B. F. *et al*. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics* **42**, 579-589, doi:http://www.nature.com/ng/journal/v42/n7/suppinfo/ng.609_S1.html (2010).

85 Mednikova, M. B. A proximal pedal phalanx of a Paleolithic hominin from denisova cave, Altai. *Archaeology, Ethnology and Anthropology of Eurasia* **39**, 129-138, doi:10.1016/j.aeae.2011.06.017 (2011).

86 A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:http://www.nature.com/nature/journal/v467/n7319/abs/10.1038-nature09534-unlocked.html#supplementary-information (2010).

87 Green, R. E. *et al*. A Draft Sequence of the Neandertal Genome. *Science* **328**, 710-722, doi:10.1126/science.1188021 (2010).
88 Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology* **128**, 415-423, doi:10.1002/ajpa.20188 (2005).
89 Sankararaman, S., Patterson, N., Li, H., Pääbo, S. & Reich, D. The Date of Interbreeding between Neandertals and Modern Humans. *PLoS genetics* **8**, e1002947, doi:10.1371/journal.pgen.1002947 (2012).
90 Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome research* **18**, 1814-1828, doi:10.1101/gr.076554.108 (2008).
91 Jukes, T. H. & Cantor, C. R. *Evolution of Protein Molecules*. (Academy Press, 1969).
92 Sequencing, R. M. G. *et al*. Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science* **316**, 222-234, doi:10.1126/science.1139247 (2007).
93 Garwood, F. Fiducial Limits for the Poisson Distribution. *Biometrika* **28**, 437-442 (1936).
94 Meyer, M. *et al*. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, doi:10.1126/science.1224344 (2012).