Title

High-resolution genetic maps identify multiple type 2 diabetes loci at regulatory hotspots in African Americans and Europeans

Authors:
Winston Lau[1], Toby Andrew[2]*, Nikolas Maniatis[1]*

[1]Department of Genetics, Evolution and Environment, University College London, WC1E 6BT London, UK; [2]Department of Genomics of Common Disease, Imperial College London, London W12 0NN, UK

*Equal authorship

Corresponding author: Nikolas Maniatis

ABSTRACT

Interpretation of results from genome-wide association studies for T2D is challenging. Only very few loci have been replicated in African ancestry populations and the identification of the implicated functional genes remain largely undefined.

We used genetic maps that capture detailed linkage disequilibrium information in European and African-Americans and applied these to large T2D case-control samples in order to estimate locations for putative functional variants in both populations. Replicated T2D locations were tested for evidence of being regulatory hotspots using adipose expression. We validated a sample of our co-location intervals using next generation sequencing and functional annotation, including enhancers, transcription and chromatin modifications.

We identified 111 additional disease-susceptibility locations, 93 of which are cosmopolitan and 18 are European specific. We show that many previously known signals are also risk loci in African-Americans. The majority of the disease locations appear to confer risk of T2D via the regulation of expression levels for a large number (266) of *cis*-regulated genes, the majority of which are not the nearest genes to the disease loci. Sequencing three cosmopolitan locations provided candidate functional variants that precisely co-locate with cell-specific chromatin domains and pancreatic islet enhancers. These variants have large effect sizes and are common across populations.

Results show that disease-associated loci in different populations, gene expression and cell-specific regulatory annotation can be effectively integrated by localizing these effects on high-resolution genetic maps. The *cis*-regulated genes provide insights into the complex molecular pathways involved and can be used as targets for sequencing and functional molecular studies.

INTRODUCTION

No disease with a genetic predisposition has been more intensely investigated than Type 2 diabetes (T2D), the world's most widespread and devastating metabolic disorder. Over the last 10 years, numerous consortia have undertaken to characterize the genetic causes of T2D through a very large number (>30) of genome-wide association studies (GWAS), and large-scale meta-analyses. Initially based on Europeans, the focus has now shifted to the replication of risk loci in additional ethnicities (trans-ethnic studies), motivated in part by the likely wider application of cosmopolitan variants for translational research, but also the desire for increasingly larger research sample sizes in order to try to boost study power[1]. But since T2D is really a group of diseases[2], increased sample size should be met with scepticism unless accompanied by more detailed clinical phenotypes and strategies to minimise disease heterogeneity. Recent trans-ethnic meta-analysis of T2D for four populations (Europeans, East Asians, South Asians and Mexican Americans) has identified seven T2D loci[3], in addition to the previously published list of 69 loci[4]. However, a large proportion of these 76 loci[3] do not show evidence for nominal association for the same 'lead' SNP. Since the lead SNP is unlikely to be the causal variant, this low replication rate is a general problem for trans-ethnic studies[5]. The inability to account for genetic distance between neighbouring SNPs and genetic heterogeneity (e.g. locus and allele heterogeneity and variation in LD between populations) are both potential thwarting factors in the endeavour to identify trans-ethnic disease loci.

On the other hand, the prevalence of T2D in African Americans (19%) at approximately twice that of European Americans (10%), and the existence of more genetic diversity in peoples of African ancestry, partly due to less extensive linkage disequilibrium (LD), also gives rise to a major opportunity for comparative fine mapping studies[3; 6; 7]. This possibility was missed by a recent major trans-ethnic meta-analysis that unfortunately excluded African Americans[3]. Here we seek to take advantage of this ancestry group by using a mapping approach to identify cosmopolitan T2D locations, which avoids the focus on lead SNPs. Instead we use high-resolution genetic maps to identify new cosmopolitan T2D susceptibility loci that are shared by both European and African American populations. Genetic distances from these maps accurately capture the genetic architecture of the relevant population and have been successfully used in gene mapping studies for other common diseases[8; 9]. We constructed genetic maps for each of these two populations and then used disease-associated location estimates on these maps as the basis for the precise co-localization and replication in both populations. We also analysed all 76 previously known T2D loci[3] to obtain refined location estimates on the same genetic maps. Since an estimated 90% of variants with a functional role in complex traits such as T2D are likely to be non-coding and regulatory[10], we assessed this scenario by exploiting publicly available subcutaneous adipose expression data. The hypothesis we tested is that T2D disease loci confer risk of disease by acting as expression quantitative trail loci (eQTL) that regulate the expression of neighbouring (*cis*-) genes. To test this hypothesis, we used the same genetic maps to assess whether the location estimates for eQTL also precisely collocated with those mapped for T2D in this whole-genome analysis, thereby identifying potential *cis*-genes and pathways regulated by the disease

loci. Finally, we performed fine mapping using targeted next generation sequencing (NGS) of the refined location estimates for one previously known locus (*TCFL72*) and two of the additional cosmopolitan loci from this study using independent case/control sample data, with the aim of identifying the candidate functional variants at the causal location estimates. These examples illustrate a way forward for the systematic identification of putative functional variants at these identified disease-associated eQTLs, coupled with the integration of functional annotation such as cell-specific chromatin domain modifications, enhancers and transcription binding sites.

METHODS

STUDY DESIGN

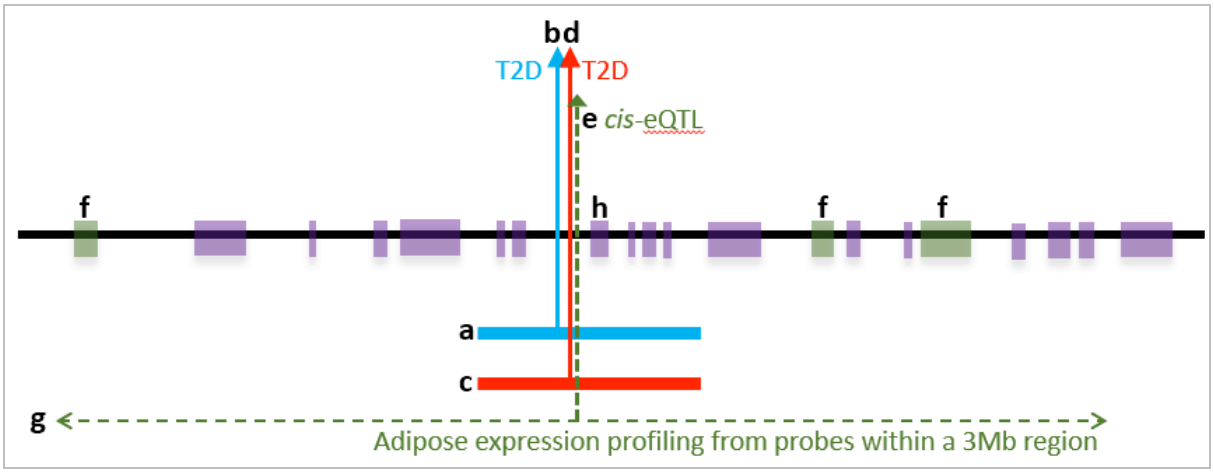
We analysed two European (EUR) and one African American (AA) samples with a total of 5,800 T2D cases and 9,691 controls. The two independent EUR samples (SNP-arrays for GWA and Metabochip) were obtained from the Wellcome Trust Case Control Consortium (WTCCC)[11; 12] with a description of diagnostic criteria and sample matching provided in Supplementary Materials. The AA GWA sample for a population of predominantly African ancestry was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)[13]. Analysing "one-SNP-at-a-time" ignores LD structure when testing for association with disease or gene expression. Here we use population-specific genetic maps, which provide (i) commensurability when making comparisons between different populations and SNP arrays; (ii) the means to implement a multi-marker test of association[14]; (iii) genetic distances between loci when testing for association with disease or adipose expression; and (iv) precise location estimates on the genetic map for potential functional variants, since these estimates are more efficient than using physical maps[15]. We constructed two high-resolution genetic maps based upon HapMap data with genetic distances expressed in LD units (LDU)[16]. The EUR LDU map was used for analyses of the two EUR T2D datasets and the AA LDU map was used for the analysis of the AA T2D dataset. The autosomal genome (sex chromosomes were not included in the analyses) was divided into 4,800 non-overlapping analytical windows with a minimum LDU distance of 10 LDU. In addition to windows being the same

minimum size on the genetic map, each window also had to include a minimum of 30 SNPs. These criteria yielded an average genetic length of 11 LDU. The identical boundaries in kilobases (kb) for all 4,800 analytical windows were used for the AA dataset, but with longer average genetic length (16 LDU), reflecting a population history of greater antiquity with additional historical recombination events. All SNPs in each analytical window were simultaneously used to test for association with disease using a multi-marker LDU model[14]. The analysis returns one estimated location for a causal variant with the strongest signal, along with the association test *P*-value for each window. Utilizing the genetic map in this way, the multi-marker test of association models the degree of regional LD when estimating the location of a putative causal variant on the genetic map. A schematic diagram of the functional genomic strategy used in the current study is provided in Figure 1. Strict criteria were used for the meta-analysis. Location estimates for genome-wide significant meta-analysis loci had to be nominally significant in both ancestry groups for the cosmopolitan loci and in both European samples for the European-specific loci. An interval criterion was used where location estimates from different datasets had to be within <100 kb of one another to qualify as a potential replication. Replicated loci had to pass a Bonferroni corrected meta-analysis *P*-value threshold of $1x10^{-5}$, based on the total number of genomic tests performed (α=0.05/4,800). We refer to the co-location interval (distance between location estimates) as the genomic region that most plausibly include the functional variants that confer risk of T2D.

We conducted *in silico* functional gene expression analyses to assess whether the same T2D loci are also eQTL that regulate the expression of neighbouring *cis*-genes using data generated by the MuTHER consortium[17]. Summary statistics for the probes and SNPs are available from the MuTHER website. Using adipose tissue mRNA expression probes as quantitative traits; we tested for *cis*-association at each disease locus by employing the same multi-locus LDU model, with potential regulated *cis*-genes defined to be within ±1.5 Mb distance either side of each replicated T2D causal locations (Figure 1). We only considered a disease locus to be a potential eQTL, if the estimated eQTL co-located to within 50 kb of the T2D location and passed Bonferroni correction for the total number of probes tested within ±1.5 Mb of each replicated disease locus. All LDU location estimates for both T2D and eQTL on the genetic map were converted back to kb B36 (NCBI36/hg18) for presentational purposes.

Finally, we conducted a NGS targeted re-sequencing experiment for three of the disease loci. Next generation sequencing was conducted using the Agilent SureSelect^{XT2} capture kit following manufacturer protocol guidelines for 100 ng of DNA. Blood DNA samples were sequenced for a total of 94 unrelated European individuals with T2D and 94 unaffected controls 1:1 matched for age, BMI and sex. Cases with a family history of T2D (selection and diagnosis criteria described elsewhere)[18] and controls were selected from families originally collected for an obesity study (without a history of T2D)[19]. Additional method details are provided in the supplementary material.

Figure 1. A schematic presentation of the functional genomic study design



[a]The LDU window of the European (EUR) genetic map; [b]location of the causal variant for T2D estimated on the EUR map using a EUR GWAs; [c]the African-American (AA) LDU genetic map; [d]location of the causal variant for T2D estimated on the AA map using an AA GWAs; [e]location of the *cis*-eQTL for the three [f]associated *cis*-gene that are implicated using adipose expression data[g] from probes for genes within ±1.5 Mb distance either side of the T2D locations[bd]. In this example, the nearest gene[h] is not the implicated regulated gene.

RESULTS

ADDITIONAL LOCI FOR T2D

Tables 1 and 2 present the results for the 111 additional loci associated with T2D. Of the 111 loci, 93 provide evidence of being cosmopolitan (signals 1–93, Table 1), since these loci replicate for both EUR and AA samples, while 18 loci appear to be European-specific (94–111, Table 2), with replication in European samples only. The distances between T2D location estimates for the majority of the 111 loci were narrow (<50kb apart). Estimation of average pairwise D-prime (D') for all HapMap SNPs found within all the identified 111 disease location intervals (ranging from 0 to <100kb) is D'=0.86 in Europeans and D'=0.78 in the AA, which reflect the importance of using a genetic map in LDU distances for localisation and the <100kb interval as a criterion for replication. For the majority of the cosmopolitan loci (signals 6–80, Table 1) the Metabochip array was not informative due to the very low SNP coverage in many regions (symbol '-' in Table 1). Some signals for Metabochip passed the minimal number of SNPs (>30 per window), but did not provide significant evidence of association ('ns' in Table 1) due to the uneven genomic coverage of SNPs on the customized Metabochip design[12]. For this reason, there were only 13 cosmopolitan loci (signals 81–93, Table 1) that provided replicated evidence for AA and Metabochip European samples.

For the 111 additional T2D loci, half of the location estimates are intragenic and half are intergenic. For the latter, we follow the convention of labeling the disease loci using the nearest gene symbol (within 100kb from the T2D location). The *in silico* expression

analyses, however, indicate which *cis*-genes are regulated and therefore functionally implicated by the identified T2D loci. Two-thirds (71/111) of the disease loci also show strong evidence of being eQTLs using our stringent criteria but the remaining one-third may well reflect that these replicated loci could be eQTLs for a T2D-relevant tissue other than subcutaneous adipose. The 71 eQTL signals regulate the expression of a conservatively estimated total of 183 *cis*-genes (Tables 1 and 2), the majority of which are not genes that are the nearest to the disease locus. Indeed, further investigation of the 183 *cis*-genes substantiates quantitatively what has previously been suspected. Namely, that the physical kb distance of the eQTL to the *nearest* gene (Figure 1) is entirely unrelated ($P>0.05$) to the distance between the same eQTL and the actual (*cis*-associated) *functional* gene (see supplementary Figure S1). This result demonstrates that the assumption that the nearest gene is also the most likely candidate functional gene is not justified. Further analysis of the 183 *cis*-genes also showed that the distance between the eQTL and the T2D location estimate is not biased by the distance between the T2D sample location estimates within the <100kb interval (see supplementary Figure S2).

Interestingly, approximately 40% of the eQTL signals observed in this study have at least one *cis*-gene previously identified as one whose expression (in adipose or liver) is also strongly associated with BMI in morbidly obese individuals[20]. We also show that a number of the identified T2D loci also regulate the expression levels of nuclear encoded mitochondrial genes. For example, many *cis*-genes implicate T2D through the regulation of the molecular functions of fatty acid metabolism (*PCCA* [MIM: 232000]), *ACAD11* [MIM: 614288], signals 66, 99, respectively); glycerophospholipid metabolism (*PISD*

[MIM: 612770]), *GPAM* [MIM: 602395])), signals 80 and 117); pyruvate metabolism (*PDHA2* [MIM: 179061]), *HAGH* [MIM: 138760], signals 26/27 and 68); mitochondrial transcription and translation (*MTERFD3* [MIM: 616929]), *LACTB* [MIM: 608440], *TRMT11* [MIM: 609752], signals 61, 90, 105); and mitochondrial protein transport (*GFER* [MIM: 600924], signal 68). The *cis*-genes *PDHA2* and *PCCA* (signals 26/27 and 66, respectively) directly implicate the genetic dysregulation of Krebs cycle function as a risk factor for T2D.

To further charecterise the observed T2D and associated eQTL location estimates in relation to potential functional variants, three of the cosmopolitan loci were targeted for sequencing using an independent sample of Europeans. Figure 2 presents an example of a regulatory intergenic hotspot near *ACTL7B* [MIM: 604304] on chromosome 9q31.3 (signal 46, Table 1). The two genetic maps (Y-axis in LDU) for AA and EUR are plotted against the physical genomic region (X-axis in kb), with each data point representing a HapMap SNP from which the genetic LDU maps were inferred (Figure 2b). Cumulative LDU plots the non-linear relationship between physical distance and the underlying LD, which is typically a "Block-Step" structure. Blocks of LD (SNPs with the same LDU location) represent areas of conserved LD and low haplotype diversity, while Steps (increasing LDU distances) define LD breakdown, primarily caused by recombination, since crossover profiles agree precisely with the corresponding LDU steps[16]. The maps show numerous short blocks across the entire region with total LDU length being greater for AA (greater LD breakdown) than the EUR population. The functional location estimates A and E (13kb apart) indicated by the vertical solid line arrows are associated

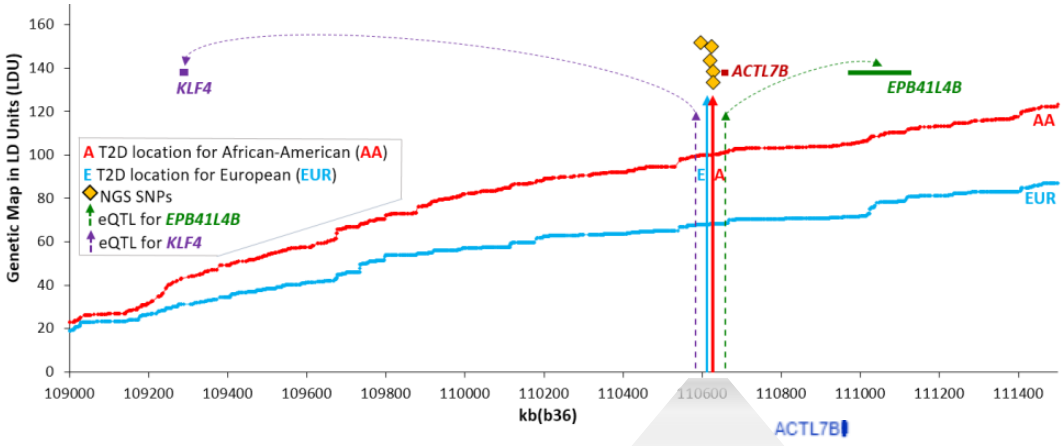
with T2D in AA and EUR samples, respectively. The dotted lines, in close proximity to the T2D locations (<30 kb), represent the location of eQTLs that regulate the expression of *KLF4* [MIM: 602253] and *EPB41L4B* [MIM: 610340] in subcutaneous adipose (nine genes reside between the two illustrated, but for clarity only the *cis*-genes regulated by the T2D-associated eQTL are plotted). *KLF4* and *EPB41L4B* are 1.3Mb downstream and 350kb upstream, respectively. Targeted next-generation re-sequencing of the 39 kb region centred on these A and E locations shows evidence of association between T2D and variants, which coincide with the estimates for the T2D-associated eQTL (summary statistics of the NGS SNPs are provided in Figure 2a). Although only nominally significant (P<0.05) due to small sample size (94 cases and 94 controls), these common variants both confirm the expected location estimate and account for a relatively large risk of disease (odds ratio (OR) of 2.0–2.4), with the risk allele frequencies (RAF) being similar in a number of human populations from the 1000 Genomes Project. Examination of the epigenetic chromatin marks from trimethylation of histone H3 at lysine 4 (H3K4me3) and acetylation of histone H3 at lysine 27 (H3K27ac), which highlight regulatory elements such as active promoters and enhancers[21] has previously been shown to overlap with T2D loci[22], but such marks are often cell type-specific[23]. Figure 2c plots the -log10 *P*-values of the chromatin profiles, demonstrating that T2D causal locations also co-localize with chromatin domains for CD14+ monocytes and adipose nuclei. The most intense chromatin peaks were observed in CD14+ monocytes at the precise eQTL location for *KLF4* and at rs60388922 and rs72756001 SNPs, which reside within the E and A T2D interval. Hence both of these SNPs are good causal candidate variants.

Figure 2. Causal variants at T2D location estimates in the *ACTL7B* region and their regulatory role

a) Novel T2D associated SNPs and risk allele frequency (RAF) in EUR, AA, East Asian (EAS), South Asian (SAS) and Mexican-Americans (AMR)

			Allele		Summary statistics			RAF in other human populations				
NGS SNPs	kb (B36)◆		risk	other	OR	*P*	RAF	EUR	AA	EAS	SAS	AMR
rs13285616	110597		A	G	2.04	2.4E-02	0.908	0.860	0.992	0.997	0.914	0.922
rs60388922	110619		G	T	2.09	3.7E-02	0.136	0.083	0.025	0.174	0.081	0.035
rs72756001	110624		A	G	2.40	3.1E-02	0.109	0.089	0.074	0.163	0.133	0.197
rs10979533	110628		G	T	2.21	4.2E-02	0.946	0.926	0.877	0.997	0.860	0.935
rs12380226	110629		T	C	2.21	4.2E-02	0.946	0.926	0.877	0.997	0.860	0.935

b)



c) -log10 P-values of cell type-specific chromatin profiles are plotted against the kb map. T2D, eQTL and NGS SNP locations overlap with the co-ordinates of the chromatin peaks

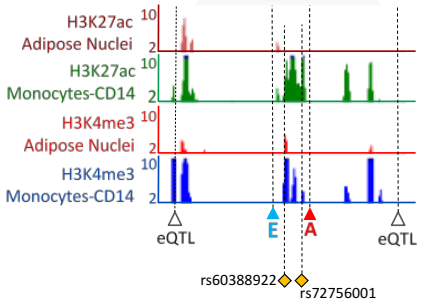


Figure 3 presents the regulatory Potassium channel, subfamily k, member 3 (*KCNK3* [MIM: 603220]) locus (signal 82, Table 1) on chromosome 2p23.3 with identical location estimates for both AA and EUR samples (Fig. 3b). Despite the disease locus residing within *KCNK3*, expression analyses indicate that this locus is not functionally related to this gene, but instead is a *cis*-eQTL that regulates the distant genes *DNMT3A* [MIM: 602769], *ASXL2* [MIM: 612991], *FAM59B* [MIM: none]), *OTOF* [MIM: 603681]), *CENPA* [MIM: 117139]) and *CCDC121* [MIM: none], with *DNMT3A* and *CCDC121* being 1.5Mb and 896kb away from the eQTL, respectively. This is a gene-rich region (57 genes in total), but for clarity only the six identified *cis*-genes have been plotted. Figure 3a presents summary statistics for the variants associated with T2D for a 42kb targeted re-sequence region. These associated variants (*P*-value <0.05) coincide with the promoter region of *KCNK3*, 11kb upstream from the T2D location estimate and account for a high risk of disease (OR=3.5–4.8). The RAF are approximately 0.05 and the results show that these variants are indeed cosmopolitan, since they are common not only in EUR and AA, but also in other human populations (e.g. East and South Asians and Mexican Americans). Using information from the human pancreatic islet regulome[22], where sequences targeted by islet transcription factors highlight active enhancers, we observed that the identified T2D location resides within a cluster of such active enhancers. The transcription factor FOXA2-bound sites are also plotted within the kb boundaries of the active enhancer cluster. Chromatin peaks also overlap with the regulatory T2D and eQTL locations for pancreas and liver cell types (Fig. 3c), suggesting evidence for more than one functional mutation within this region. The full interval for the T2D and eQTL locations were not entirely covered by NGS data, but nevertheless, the associated NGS

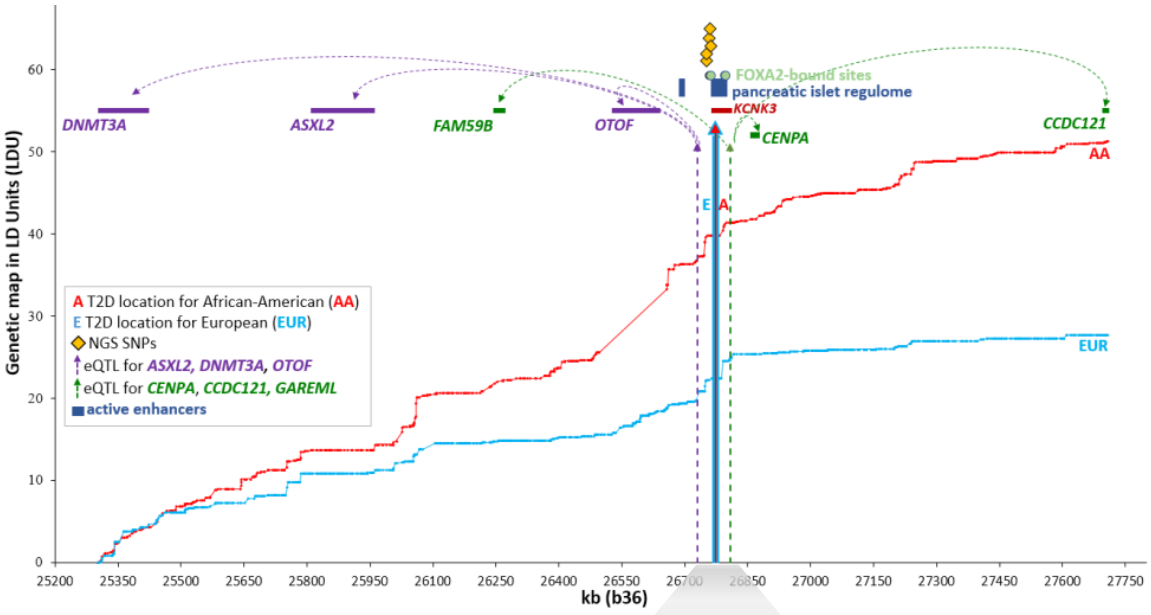
SNPs reside between two active islet enhancers (Fig. 3b). In contrast to imputation methods that use high-resolution "out-of-sample" marker panels to infer missing SNPs to subsequently test one at a time, LDU analysis uses marker panels to infer high-resolution genetic maps. Subsequently, multi-marker tests of association use genetic distances for SNP arrays placed on those maps to infer the location of disease-associated functional variants. Fine mapping is therefore achieved by inferring fine-scale genetic maps, not by imputing SNPs. It is worth noting in relation to this that using the *KCNK3* locus as an example, the NGS SNPs genotyped for the case control data that were associated with T2D status (Fig. 3a) could not be imputed based on the 1000G data as the reference panel.

Figure 3. Causal variants at T2D locations estimates in the *KCNK3* region and their regulatory role

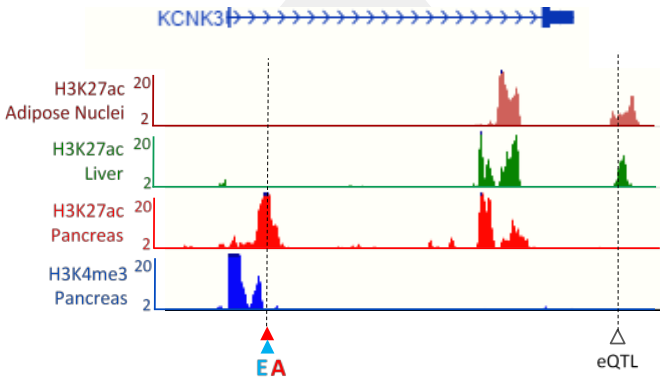
a) Novel T2D associated SNPs and risk allele frequency (RAF) in EUR, AA, East Asian (EAS), South Asian (SAS) and Mexican-Americans (AMR)

		Allele		Summary statistics			RAF in other human populations				
NGS SNPs ◆	kb (B36)	risk	other	OR	*P*	RAF	EUR	AA	EAS	SAS	AMR
rs78489206	26751	A	G	4.73	3.1E-02	0.049	0.047	0.025	0.053	0.112	0.013
rs77786658	26753	A	G	4.85	2.8E-02	0.051	0.047	0.025	0.053	0.111	0.013
rs6707973	26755	G	A	3.51	4.6E-02	0.054	0.050	0.443	0.057	0.117	0.059
rs59284336	26760	T	C	4.73	3.1E-02	0.049	0.047	0.033	0.058	0.099	0.013
rs113076024	26763	A	G	4.79	2.9E-02	0.050	0.047	0.033	0.057	0.087	0.013

b)



c) -log10 P-values of cell type-specific chromatin profiles are plotted against the kb map. T2D and eQTL locations overlap with the co-ordinates of the chromatin peaks



FINDINGS AT PREVIOUSLY KNOWN T2D LOCI

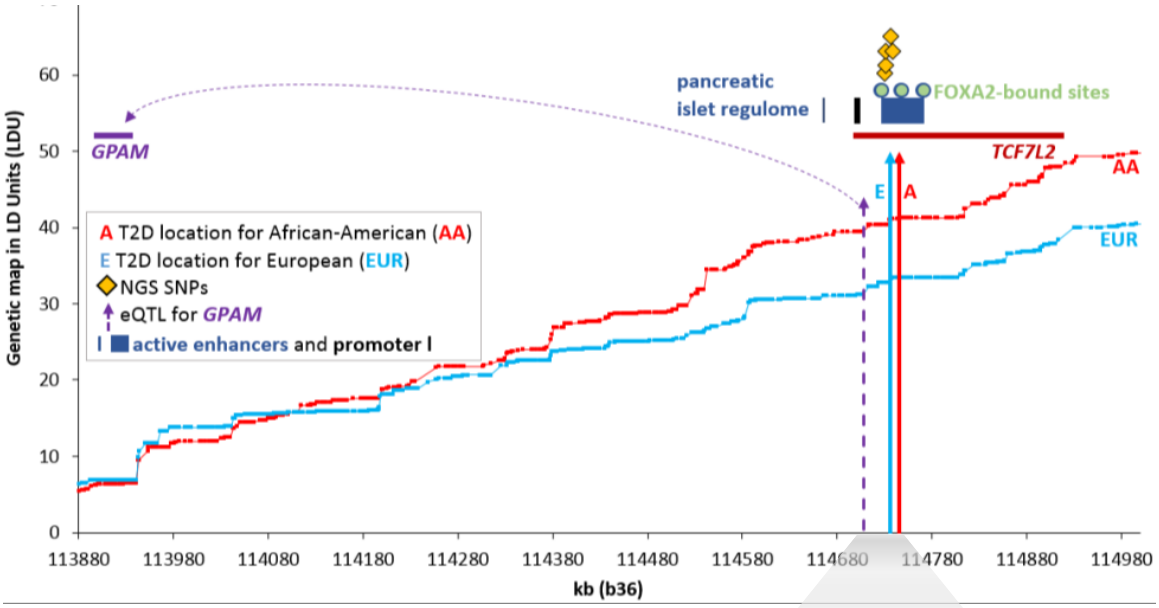
Using the same data and analytical methods, we have confirmed disease location estimates for 62 out of 76 previously known loci[3] (supplementary Table S1). Of these, about half (33/62) show evidence of being eQTL with the majority regulating *cis*-genes over 1Mb away. In addition, over one third (22/62) of the loci replicate for AA samples, which were previously excluded from trans-ethnic meta-analysis[3]. We investigated the Transcription factor 7-like 2 (*TCF7L2* [MIM: 602228]) locus (signal 117, supplementary Table S1). Figure 4b plots the T2D locations for EUR and AA and shows that this signal harbours a *cis*-eQTL for the distant *GPAM*. The T2D re-sequence variants we identified, which co-locate to the <30 kb interval between the T2D and eQTL locations, account for a large risk of disease (OR=1.7–2.6). The summary statistics in Figure 4a show that these risk variants are the major allele in all other human populations. The T2D locations for both populations reside within an active enhancer cluster that is targeted by transcription factor FOXA2 (kb locations of the regulatory elements are plotted on the X-axis). Inferring the likely transcriptional activity by observing the chromatin state, we show that the eQTL, T2D co-locations and NGS SNPs all map precisely to highly significant H3K4me3 and H3K27ac peaks, in particular in adipose cells (Fig. 4c). This illustrates the importance of co-locating within an interval on the genetic map, since it allows for potential allelic heterogeneity.

Figure 4. Causal variants at T2D location estimates in the *TCF7L2* region and their regulatory role

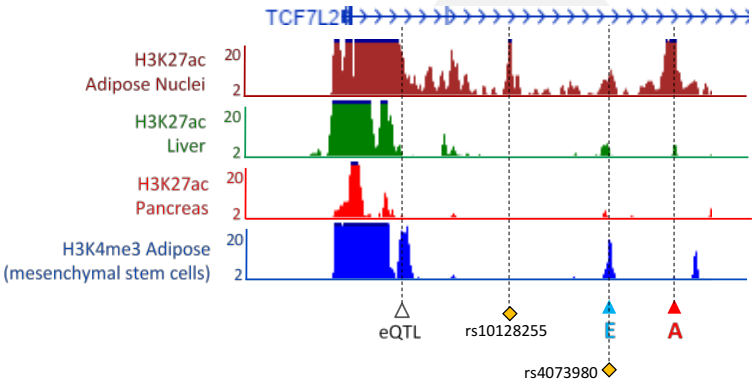
a) T2D associated SNPs and risk allele frequency (RAF) in EUR, AA, East Asian (EAS), South Asian (SAS) and Mexican-Americans (AMR)

		Allele		Summary statistics			RAF in other human populations				
NGS SNPs ◆	kb (B36)	risk	other	OR	*P*	RAF	EUR	AA	EAS	SAS	AMR
rs7080591	114731	T	C	2.61	2.8E-04	0.746	0.615	0.549	0.678	0.591	0.503
rs7081062	114731	A	G	2.04	1.6E-03	0.755	0.643	0.639	0.694	0.640	0.530
rs10128255	114733	A	G	2.00	2.3E-03	0.755	0.643	0.639	0.695	0.640	0.530
rs4073980	114737	G	C	1.73	8.9E-03	0.609	0.498	0.779	0.049	0.376	0.339
rs4074718	114739	A	G	1.73	8.9E-03	0.609	0.499	0.779	0.050	0.377	0.337

b)



c) -log10 P-values of cell type-specific chromatin profiles are plotted against the kb map. T2D, eQTL and NGS SNP locations overlap with the co-ordinates of the chromatin peaks



DISCUSSION

This study provides a comprehensive genomic catalogue of susceptibility loci for T2D in European and African ancestry populations and evidence that the majority of the additional 111 and 62 previously known disease loci are eQTLs for 183 and 83 *cis*-genes, respectively. This implies that these disease loci confer risk of T2D, via the *cis*-regulation of the expression levels in tissue relevant to T2D for a large number (266 in total) of neighbouring genes. This study identifies a large number of disease loci at regulatory hotspots and replicates them in both European and African American populations, with 84% (93/111) of the additional loci being cosmopolitan. This replication was made possible by analyses that make use of, rather than being confounded by, the fine-scale differences in LD between these two populations, where causal locations and also eQTLs are estimated on an LDU map, avoiding the ambivalence of interpreting individual GWAS SNPs. Interestingly, recent results in the literature support our conclusion that cosmopolitan loci are more widespread than previously thought. For example, it was not until recently that the most established bona-fide *TCF7L2* locus for T2D in Europeans was also confirmed in African ancestry groups in an extensive study that included 17 African American GWA samples[7]. This is a locus that was identified in the present study, which included only one African American sample.

The T2D associated common genetic variation in *TCF7L2* is well established, but the mechanism by which risk of disease is conferred remains elusive. Our refined localization

of this locus reveals that this is a regulatory site for the distant nuclear-encoded mitochondrial gene, *GPAM,* and identifies additional candidate causal variants *GPAM* is interesting as it is a key rate-limiting enzyme in lipogenesis and highly expressed in liver and adipose tissue. Nuclear-encoded mitochondrial genes are of particular interest in relation to T2D, since mitochondrial function has been demonstrated to impact upon a myriad of molecular and cellular functional processes implicated in T2D[24-26]. However, to date human genetic association studies have identified few, if any, nuclear-encoded mitochondrial genes that directly confer risk of T2D in its common form. In this study we have identified at least 11 further nuclear-encoded mitochondria genes, which are regulated by eQTLs that also appear to confer risk of T2D. This indicates that one of the molecular mechanisms that contribute to inherited risk of T2D is mitochondrial dysfunction in relation to energy metabolism. These findings are also supported by an independent study design that utilises transcriptome and proteome data to reconstruct metabolic pathways in myocytes and identify the same mitochondrial pathways implicated in our study for adipose tissue (namely, fatty acid beta-oxidation, Krebs cycle, pyruvate metabolism and branched-chain amino acid [valine, leucine, and isoleucine] metabolism)[27].

Recent large scale whole genome and exome association studies[28] have empirically questioned any major role for coding variants in the aetiology of T2D and therefore make regulatory loci such as the ones highlighted in this study, all the more important to investigate. We believe that the targeted re-sequencing of informative refined regions using case/control data has the power to dissect the genetic epidemiology of T2D. While

imputation methods can successfully infer missing genotypes for most genomic regions using population data as the reference data, ironically it appears that imputation methods are likely to fail to impute or to make correct inferences where it matters most (e.g. at disease loci with allelic spectra that differ from the general population). This important point requires further investigation.

Chromatin analyses and targeted re-sequencing of these refined regions can be used to identify potential causal variant locations. The two identified signals (*ACTL7B* and *KCNK3*), which we have investigated lends support to this approach. A cosmopolitan disease location interval that includes *KCNK3* was observed to be a hotspot for the regulation of six neighboring genes, with *OTOF* of particular interest because of its known association with hearing loss (autosomal recessive forms of deafness[29]). According to the NIDDK, the prevalence of low- or mid-frequency hearing impairment among diabetics is three times that of non-diabetics (28% compared with 9%)[30] with the potential mechanism operating via microvascular and neural damage due to long-term hyperglycaemia[31]. Interestingly, adipose gene expression for *OTOF* has also been previously shown to be associated with BMI in the morbidly obese[29] providing further evidence of a functional role through a regulatory mechanism. Targeted re-sequencing within the *KCNK3* location interval identified candidate causal variants with large effect sizes. As with the *TCF7L2* locus, the T2D and eQTL locations were found to reside in pancreatic islet enhancer and FOXA2 transcription factor-binding sites. Studies have shown that dysregulation of islet enhancers is relevant to the underlying mechanisms of

T2D[22] and a recent examination of some of the previously found T2D loci have been found to overlap with FOXA2-bound sites[32].

The cosmopolitan T2D location interval near *ACTL7B* overlapped with chromatin peaks for CD14+ monocytes and included an eQTL for *KLF4* and *EPB41L4B*. *KLF4* is highly expressed in CD14+ monocytes and belongs to the Krüppel-like factor (KLF) family that consists of transcription factors that can activate or repress different genes involved in processes such as differentiation, development and cell cycle progression, with several of these proteins implicated in glucose homeostasis[33]. *KLF4* is also used experimentally to induce pluripotent cells that can differentiate into insulin-producing cells[34]. *EPB41L4B* codes for an erythrocyte membrane protein (EMP) with greatly increased EMP glycosylation observed in T2D subjects[35] with likely clinical implications[36]. It is recognized clinically that both obesity and T2D are associated with a state of abnormal inflammatory response. Here we show that the T2D variants and eQTL locations for *KLF4* and *EPB41L4B* reside in regions with chromatin modifications mainly observed in CD14+ monocytes. Monocytes play a pivotal role in innate immunity and are involved in metabolic regulation[37]. It has been shown that unbalanced proinflammatory/anti-inflammatory markers of CD14+ cells is associated with metabolic disorder in obese T2D patients[38]. *KLF4* is a critical regulator of monocyte differentiation[39] and *EPB41L4B* expression in subcutaneous and omental adipose is strongly associated with BMI in morbidly obese individuals[20]. Therefore, these T2D intragenic variants and the regulated *cis*-genes (*KLF4* and *EPB41L4B*) are likely to be involved in an inflammatory pathway for obesity and T2D.

The complex causal chain between a gene and its effect on susceptibility cannot be unravelled until we have a full understanding of the regulatory genetic architecture that underpins T2D, and until the causal changes have been localized in the DNA sequence[40]. Our results show that disease-associated loci in different populations, gene expression and cell-specific regulatory annotation can be effectively integrated by localizing these effects on high-resolution linkage disequilibrium maps. By exploiting these maps to refine causal location estimates, we have identified a genomic catalogue of cosmopolitan and European disease loci with correspondingly important clinical implications that provides important molecular insights and opportunities to understand the molecular basis of this devastating common disease.

Acknowledgements

We would like to thank the WTCCC, UK, for making the WTCCC T2D genomic data available. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. We are grateful to the NIDDK, USA, for making the AA T2D phenotype and genomic data available to us. The NIDDK whole genome association search for T2D genes in African Americans was conducted by Donald Bowden, Center for Human Genomics, Center for Diabetes Research, Wake Forest University School of Medicine, Winston-Salem, USA, with support from the NIDDK. The datasets used were obtained from the database of Genotypes and Phenotypes (dbGaP) at accession number phs000140. This manuscript was not prepared in collaboration with the labs of any of the investigators responsible for generating the data, and does not necessarily reflect the views or opinions of these investigators. TA would like to acknowledge the Medical Research Council UK [Investigator Award 91993] for supporting his work. All authors are grateful to Professor Dallas Swallow (UCL) for her valuable comments on the manuscript, Professor Philippe Froguel (Imperial College, CNRS 8199, EGID 59045) for generously supplying us with European DNA samples for the NGS pilot work and Aminah Ali (UCL) for her valuable contributions to processing the NGS data. NM would like to acknowledge Newton E. Morton (Southampton) for the previous body of work on LDU maps.

Conflicts of interest

No conflict of interest to declare.

Table 1. Identified cosmopolitan T2D susceptibility loci and their regulatory role of neighbouring gene expression

Table 2. Identified European-specific T2D susceptibility loci and their regulatory role of neighbouring gene expression

Figure 1. A schematic presentation of the functional genomic study design

Figure 2. Candidate causal variants at T2D location estimates on the LDU genetic maps in the *ACTL7B* region and their regulatory role

Figure 3. Candidate causal variants at T2D locations estimates on the LDU genetic maps in the *KCNK3* region and their regulatory role

Figure 4. Candidate causal variants at T2D location estimates on the LDU genetic maps in the *TCF7L2* region and their regulatory role

Table 1. Identified cosmopolitan T2D susceptibility loci and their regulatory role of neighbouring gene expression

[a]T2D associated intervals in kb (<100) that harbour T2D susceptibility loci in both populations, the minimum distance is provided for signals 1-5; [b]Location estimates for the European (E) GWAS; [c]Location estimates for the African-American (A) GWAS; [d]Location estimates for the Metabochip European (E) samples, signals with low SNP coverage '-' were not meta-analysed; [e]Genes in bold denote the intragenic localization and genes with '+' for self-regulatory; [f]Number of *cis*-genes regulated by the eQTL; [g]List of *cis*-genes associated with eQTLs that co-located within <50kb of the T2D locations, cis-genes with '*' have previously shown evidence of association between Body Mass Index in morbidly obese and adipose/liver expression profiles[20]; [h]Distance in kb (<50) between eQTL and T2D locations, the minimum is given when more than one *cis*-gene is implicated.

Signal	Chr.	Meta *P*-value	Distance between locations[a]	T2D location GWAS-E[b]	T2D location GWAS-A[c]	T2D location metabo-E[d]	Nearest gene to T2D locations[e]	no. of *cis*-genes[f]	eQTL associated *cis*-genes[g]	eQTL distance from T2D[h]
1	4p	1.73E-56	1	44797	44857	44858	-	0	-	-
2	6q	5.28E-10	4	72479	72509	72505	-	1	*FAM135A*	1
3	13q	2.14E-35	28	109848	109805	109833	***COL4A2***	1	*ANKRD10*	20
4	17q	2.80E-12	7	65769	65762	65769	***KCNJ2***	1	*MAP2K6*	37
5	20q	1.83E-06	77	44104	44181	44104	***SLC12A5, CD40***$^{+}$	2	*CD40, CDH22*	17
6	1p	2.48E-06	41	82826	82867	-	-	1	*LPHN2**	8
7	1p	5.20E-07	73	84451	84524	-	***PRKACB***$^{+}$, ***SAMD13***$^{+}$	3	*PRKACB, SAMD13, C1orf52*	6
8	1p	1.79E-11	94	106441	106535	-	-	1	*PRMT6*	3
9	1q	5.03E-14	8	207662	207670	-	***MIR205HG***	0	-	-
10	1q	3.40E-09	3	232337	232340	-	***SLC35F3***	0	-	-
11	1q	1.07E-07	52	234896	234948	-	***ACTN2***	3	*TBCE*, B3GALNT2, ARID4B*	23
12	1q	6.33E-07	49	243993	244042	-	***SMYD3***	0	-	-
13	2p	4.57E-07	95	12139	12044	-	***MIR3681HG***	0	-	-
14	2p	2.62E-15	89	45025	44936	-	***SIX3***, *CAMKMT*	3	*PRKCE*, DYNC2LI1, ABCG8*	27
15	2q	2.61E-09	89	139530	139527	-	-	0	-	-
16	3p	4.05E-07	42	7503	7461	-	***GRM7***	1	*RAD18*	17
17	3p	1.20E-10	0	38370	38369	ns	***XYLB***	0	-	-
18	3p	5.10E-11	95	41078	41173	ns	*CTNNB1*	2	*CTNNB1*, ZNF621*	1
19	3p	4.22E-23	13	47556	47569	ns	*CSPG5*	4	*IP6K2 (IHPK2)*, KIF9, TESSP5, P4HTM*	3
20	3p	8.39E-06	20	54986	54965	-	***CACNA2D3***	0	-	-
21	3p	1.13E-13	59	67827	67768	-	*SUCLG2*	0	-	-
22	3p	6.67E-10	9	87317	87327	ns	*CHMP2B*	0	-	-
23	3q	7.80E-06	18	122056	122038	-	*BC032918*	2	*CD86*, ILDR1*	4
24	3q	5.39E-08	49	184743	184694	-	***KLHL6***	3	*AP2M1, ABCF3, MAGEF1*	0
25	4q	1.87E-11	24	53132	53108	ns	***USP46***$^{+}$	1	*USP46*	2
26	4q	2.68E-07	6	92162	92168	-	***CCSER1***	2	*HSD17B13, PDHA2*	0
27	4q	5.24E-13	56	96319	96376	-	***UNC5C***	1	*PDHA2*	2
28	4q	4.02E-13	3	102095	102098	-	*PPP3CA*	0	-	-
29	4q	1.50E-06	74	148449	148375	>100kb	-	1	*EDNRA*	1
30	5q	8.84E-06	44	52127	52083	-	***PELO***	1	*ITGA1*	20
31	5q	1.56E-08	14	59263	59277	-	***PDE4D***$^{+}$	1	*PDE4D**	38
32	5q	2.81E-06	77	97282	97359	-	-	0	-	-
33	6p	1.00E-07	86	40514	40429	-	***LRFN2, BC132805***	2	*KCNK5, UNC5CL**	0
34	6p	1.30E-15	72	44790	44718	-	*BX647715*	1	*PTK7*	4
35	6q	3.04E-07	5	168842	168837	-	*SMOC2*	3	*MLLT4*, CCR6, C6orf122*	1
36	7p	6.93E-08	85	24390	24305	-	*NPY*	2	*CCDC126, OSBPL3*	35
37	7p	1.49E-08	4	35510	35514	-	*HERPUD2*	1	*AAA1*	34
38	7p	5.97E-06	75	37471	37396	-	***ELMO1***$^{+}$	2	*ELMO1, GPR141*	30
39	7q	4.00E-06	77	82225	82149	-	***PCLO***	1	*HGF**	26
40	7q	4.78E-12	62	132516	132454	-	*CHCHD3*	0	-	-
41	7q	1.30E-09	73	133716	133789	-	*SLC35B4*, ***AKR1B1***	0	-	-
42	7q	4.44E-10	14	134336	134322	-	***AGBL3***	1	*TMEM140**	0

43	7q	6.98E-14	75	141873	141798	-	***TCRBV20S1*, *TCRB***	4	*FAM131B*, ZYX, OR2A25, OR2F1*	0
44	8p	4.81E-08	93	14339	14246	-	***SGCZ***	2	*C8orf79, TUSC3*	11
45	8q	7.00E-08	87	70416	70503	-	*SULF1*	1	*PRDM14*	46
46	9q	6.46E-08	13	110613	110626	ns	*ACTL7B*	2	*EPB41L4B*, KLF4*	29
47	9q	8.64E-07	19	132799	132781	-	***FIBCD1***	4	*ABL1*, AIF1L, RAPGEF1* PRDM12*	23
48	10q	5.13E-10	11	44467	44456	>100kb	-	1	*ZNF22**	32
49	10q	1.28E-06	8	57261	57268	-	-	1	*CDC2*	0
50	10q	9.29E-09	0	65408	65408	ns	*CR622643*	0	-	-
51	10q	8.37E-11	66	77817	77883	-	***C10orf11***$^{+}$	2	*C10orf11, DUPD1*	1
52	11p	4.29E-09	5	22257	22262	-	***ANO5***	0	-	-
53	11p	5.53E-13	88	32453	32364	-	***WT1***$^{+}$	4	*PRRG4*, FBXO3*, ELP4, WT1*	14
54	11q	4.12E-08	0	57558	57558	ns	***OR9Q1***	1	*MPEG1*	33
55	11q	1.11E-09	21	59295	59315	-	***STX3***	0	-	-
56	12p	1.40E-07	31	25404	25373	-	*KRAS*	1	*SSPN*	0
57	12p	8.75E-07	6	29568	29574	-	***TMTC1***	0	-	-
58	12q	3.45E-06	91	57035	57127	ns	-	2	*SLC26A10, MBD6*	1
59	12q	8.78E-06	34	61602	61636	-	***PPM1H***	0	-	-
60	12q	2.45E-06	62	98425	98363	-	***ANKS1B***	0	-	-
61	12q	1.45E-06	22	104931	104909	-	*NUAK1*	1	*MTERFD3*	0
62	13q	2.09E-07	85	26479	26564	-	***USP12***	0	-	-
63	13q	8.19E-06	40	47133	47173	-	-	0	-	-
64	13q	7.16E-14	13	65393	65380	-	*MIR548X2*	0	-	-
65	13q	1.10E-14	82	66885	66803	-	-	0	-	-
66	13q	5.74E-07	44	101207	101251	-	***FGF14***$^{+}$	3	*TMTC4, PCCA, FGF14*	0
67	15q	1.13E-06	16	45938	45954	-	*LINC01494*	0	-	-
68	16p	2.90E-10	80	1700	1781	-	*MAPK8IP3, IGFALS*	7	*HAGH*, PGP*, TPSAB1*, FAHD1, GFER, GNPTG, PRR25*	0
69	16p	2.24E-06	32	12605	12636	-	*SNX29*	3	*CLEC16A*, SOCS1, LITAF*	10
70	16q	1.75E-26	4	70999	70995	>100kb	*AK055364*	1	*HYDIN*	2
71	16q	1.03E-11	20	77036	77016	-	***WWOX***	0	-	-
72	16q	5.52E-19	29	78435	78406	-	*LOC101928248*	2	*MAF*, WWOX*	11
73	17q	4.15E-07	47	34640	34593	>100kb	*STAC2*, ***CACNB1***	4	*PLXDC1, CRKRS, ZPBP2, KRT222*	6
74	18q	9.58E-10	0	36270	36271	-	-	0	-	-
75	18q	6.84E-06	72	74880	74809	-	*SALL3*	0	-	-
76	21q	3.09E-09	30	20521	20490	-	*DYRK1A*	0	-	-
77	21q	9.56E-08	61	27738	27677	-	*MIR5009*	0	-	-
78	21q	1.35E-10	39	34148	34109	-	***ITSN***	2	*C21orf66, TCP10L*	0
79	21q	1.53E-06	1	37635	37633	-	*DYRK1A*	1	*KCNJ6*	0
80	22q	7.29E-09	0	31376	31376	-	***SYN3***	1	*PISD*	12
81	2p	6.46E-06	7	ns	21294	21301	-	1	*SDC1*	33
82	2p	5.02E-09	0	>100kb	26774	26774	*KCNK3*	6	*OTOF*, FAM59B, CCDC121, ASXL2, DNMT3A, CENPA*	3
83	2q	1.02E-11	1	>100kb	135313	135313	***ACMSD***	0	-	-
84	2q	2.25E-11	0	ns	203732	203732	*NBEAL1*	0	-	-
85	4q	2.77E-09	8	>100kb	103404	103395	***SLC39A8***$^{+}$	1	*SLC39A8*	0
86	6p	2.09E-13	60	ns	28525	28586	*ZSCAN23, PX6*	8	*ZKSCAN3, PGBD1*, MAS1L, OR2W1, OR11A1, HLA-F, IFITM4P, ZNF184*	6
87	8p	1.54E-18	44	>100kb	18194	18238	*NAT1, NAT2*	3	*PDGFRL*, CSGALNACT1, PCM1*	1
88	11p	1.86E-20	86	>100kb	8508	8594	***STK33, TRIM66***	0	-	-
89	12q	4.10E-16	44	>100kb	48574	48618	***FAIM2***$^{+}$	1	*FAIM2*	10
90	15q	1.93E-07	77	ns	61133	61210	***TPM1***$^{+}$, ***LACTB***$^{+}$	9	*TPM1, RPS27L*, DAPK2*, SNX1, LACTB, APH1B, HERC1, FAM96A, VPS13C*	1
91	17q	2.02E-06	86	>100kb	44633	44548	***GNGT2***$^{+}$, *B4GALNT2*	3	*GNGT2, SAMD14*, HOXB3**	19
92	17q	3.95E-08	85	>100kb	60951	60866	***AXIN2***$^{+}$	1	*AXIN2*	38
93	20p	1.90E-19	31	>100kb	25718	25687	***FAM182B***	0	-	-

Table 2. Identified European-specific T2D susceptibility loci and their regulatory role of neighbouring gene expression

[a]T2D associated intervals in kb (<100) that harbour T2D susceptibility loci in two European populations; [b]T2D location estimates for the European (E) GWAS; [c]The African-American (A) GWAS yielded either significant but distant locations from the European T2D location (>100kb) or not significant 'ns' estimates; [d]Location estimates for the Metaboship European (E) samples that were within <100kb of the GWAS-E location; [e]Genes in bold denote the intragenic localization and genes with '+' for self-regulatory genes; [f]Number of cis-genes regulated by the eQTLs; [g]List of *cis*-genes associated with eQTLs that co-located within <50kb of the T2D locus, *cis*-genes with '*' have previously shown evidence of association between Body Mass Index for morbidly obese and adipose/liver expression profiles[20]; [h]Distance in kb (<50) between eQTL and T2D locations, the minimum is given when more than one *cis*-gene is implicated.

Signal	Chr.	Meta *P*-value	Distance of T2D locations[a]	T2D location GWAS-E[b]	T2D location GWAS-A[c]	T2D location metabo-E[d]	Nearest gene to T2D locations[e]	no. of *cis*-genes[f]	eQTL associated *cis*-genes[g]	eQTL distance from T2D[h]
94	1p	2.02E-07	3	25879	ns	25876	***MAN1C1***	2	*DHDDS, C1orf172*	0
95	1q	6.94E-12	21	228322	ns	228344	***GALNT2***	0	-	-
96	2p	4.11E-34	34	605	ns	639	*TMEM18*	0	-	-
97	2p	1.18E-07	46	40320	>100kb	40274	***SLC8A1-AS1***	4	*THUMPD2, TMEM178, MORN2, DHX57*	0
98	3p	1.03E-09	5	53518	ns	53513	***CACNA1D***	0	-	-
99	3q	9.35E-07	7	133919	ns	133912	***ACAD11***[+]	8	*ACAD11, SLC02A1, RYK, NPHP3, ACKR4, SRPRB, CDV3, RAB6B*	0
100	4q	1.86E-22	66	104224	>100kb	104157	*BDH2, NHEDC1*	0	-	-
101	5q	2.62E-09	12	76449	ns	76461	*ZBED3*	1	*PDE8B**	15
102	6p	2.64E-24	13	29662	ns	29674	*GABBR1*	7	*GNL1*, TRIM10, TRIM27, DDR1, TRIM40, TRIM15, OR10C1*	9
103	6p	1.46E-28	6	31709	>100kb	31704	*BAT2, AIF1*	14	*AIF1*, TRIM39*, AGER, BAT4, CCHCR1, DOM3Z, EHMT2, HLA-DMB, HLA-DPA1, HSPA1B, LST1, TRIM10, NOTCH4, HLA-DRA*	0
104	6q	6.29E-56	0	118801	>100kb	118801	*SLC35F1*	0	-	
105	6q	4.23E-17	37	127581	ns	127544	***RSPO3***[+]	2	*RSPO3, TRMT11*	1
106	10q	3.39E-24	40	104790	>100kb	104830	***CNNM2***	0	-	
107	11p	2.50E-06	1	43836	ns	43836	*HSD17B12*	1	*HSD17B12*	1
108	11q	6.39E-06	21	65337	>100kb	65357	*OVOL1,* ***SNX32***[+]	4	*SNX32, MAJIN*, SNX15, EFEMP2*	9
109	12q	6.16E-06	4	54905	>100kb	54909	***OBFC2B***[+]	4	*OBFC2B, RGL1*, RPS26, STAT2*	0
110	12q	4.66E-08	3	111476	>100kb	111473	*PTPN11*	4	*ERP29, TMEM116, MAPKAPK5, NAP1*	20
111	17q	3.05E-09	51	25003	>100kb	25054	***SSH2***	1	*CORO6*	9

Web Resources

GTEx Portal

http://www.gtexportal.org/home/

HapMap:

ftp://ftp.ncbi.nlm.nih.gov/hapmap/

Islet Regulome Browser

http://gattaca.imppc.org/isletregulome/home

MuTHER

http://www.muther.ac.uk

Online Mendelian Inheritance in Man:

http://www.omim.org

Roadmap Epigenetics Project

http://www.roadmapepigenomics.org

The 1000 Genomes Project

http://www.internationalgenome.org

References

1. Altshuler, D., and Daly, M. (2007). Guilt beyond a reasonable doubt. Nat Genet 39, 813-815.
2. Lebovitz, H.E. (1999). Type 2 diabetes: an overview. Clinical chemistry 45, 1339-1345.
3. Replication, D.I.G., Meta-analysis, C., Asian Genetic Epidemiology Network Type 2 Diabetes, C., South Asian Type 2 Diabetes, C., Mexican American Type 2 Diabetes, C., Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples, C., Mahajan, A., Go, M.J., Zhang, W., Below, J.E., et al. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet 46, 234-244.
4. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segre, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet.
5. Li, Y.R., and Keating, B.J. (2014). Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome medicine 6, 91.
6. Asimit, J.L., Hatzikotoulas, K., McCarthy, M., Morris, A.P., and Zeggini, E. (2016). Trans-ethnic study design approaches for fine-mapping. European journal of human genetics : EJHG.
7. Ng, M.C., Shriner, D., Chen, B.H., Li, J., Chen, W.M., Guo, X., Liu, J., Bielinski, S.J., Yanek, L.R., Nalls, M.A., et al. (2014). Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. PLoS Genet 10, e1004517.
8. Direk, K., Lau, W., Small, K.S., Maniatis, N., and Andrew, T. (2014). ABCC5 transporter is a novel type 2 diabetes susceptibility gene in European and African American populations. Ann Hum Genet 78, 333-344.
9. Elding, H., Lau, W., Swallow, D.M., and Maniatis, N. (2013). Refinement in localization and identification of gene regions associated with Crohn disease. Am J Hum Genet 92, 107-113.
10. Zhang, X., Bailey, S.D., and Lupien, M. (2014). Laying a solid foundation for Manhattan--'setting the functional basis for the post-GWAS era'. Trends Genet 30, 140-149.
11. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661-678.
12. Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet 8, e1002793.
13. Palmer, N.D., McDonough, C.W., Hicks, P.J., Roh, B.H., Wing, M.R., An, S.S., Hester, J.M., Cooke, J.N., Bostrom, M.A., Rudock, M.E., et al. (2012). A genome-wide association search for type 2 diabetes genes in African Americans. PLoS One 7, e29202.
14. Maniatis, N., Collins, A., and Morton, N.E. (2007). Effects of single SNPs, haplotypes, and whole-genome LD maps on accuracy of association mapping. Genet Epidemiol 31, 179-188.
15. Maniatis, N., Collins, A., Gibson, J., Zhang, W., Tapper, W., and Morton, N.E. (2004). Positional cloning by linkage disequilibrium. Am J Hum Genet 74, 846-855.
16. Maniatis, N., Collins, A., Xu, C.F., McCarthy, L.C., Hewett, D.R., Tapper, W., Ennis, S., Ke, X., and Morton, N.E. (2002). The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci U S A 99, 2228-2233.

17. Grundberg, E., Small, K.S., Hedman, A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.P., Meduri, E., Barrett, A., et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet 44, 1084-1089.
18. Vionnet, N., Hani, E.H., Dupont, S., Gallina, S., Francke, S., Dotte, S., De Matos, F., Durand, E., Lepretre, F., Lecoeur, C., et al. (2000). Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. Am J Hum Genet 67, 1470-1480.
19. Meyre, D., Lecoeur, C., Delplanque, J., Francke, S., Vatin, V., Durand, E., Weill, J., Dina, C., and Froguel, P. (2004). A genome-wide scan for childhood obesity-associated traits in French families shows significant linkage on chromosome 6q22.31-q23.2. Diabetes 53, 803-811.
20. Greenawalt, D.M., Dobrin, R., Chudin, E., Hatoum, I.J., Suver, C., Beaulaurier, J., Zhang, B., Castro, V., Zhu, J., Sieberts, S.K., et al. (2011). A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. Genome research 21, 1008-1016.
21. Consortium, E.P. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS biology 9, e1001046.
22. Pasquali, L., Gaulton, K.J., Rodriguez-Segui, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J.J., Moran, I., Gomez-Marin, C., van de Bunt, M., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat Genet 46, 136-143.
23. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet 45, 124-130.
24. Lowell, B.B., and Shulman, G.I. (2005). Mitochondrial dysfunction and type 2 diabetes. Science 307, 384-387.
25. Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34, 267-273.
26. Patti, M.E., and Corvera, S. (2010). The role of mitochondria in the pathogenesis of type 2 diabetes. Endocr Rev 31, 364-395.
27. Varemo, L., Scheele, C., Broholm, C., Mardinoglu, A., Kampf, C., Asplund, A., Nookaew, I., Uhlen, M., Pedersen, B.K., and Nielsen, J. (2015). Proteome- and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes. Cell reports 11, 921-933.
28. Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D.J., et al. (2016). The genetic architecture of type 2 diabetes. Nature 536, 41-47.
29. Duman, D., and Tekin, M. (2012). Autosomal recessive nonsyndromic deafness genes: a review. Frontiers in bioscience 17, 2213-2236.
30. Bainbridge, K.E., Hoffman, H.J., and Cowie, C.C. (2008). Diabetes and hearing impairment in the United States: audiometric evidence from the National Health and Nutrition Examination Survey, 1999 to 2004. Annals of internal medicine 149, 1-10.
31. Bainbridge, K.E., Cheng, Y.J., and Cowie, C.C. (2010). Potential mediators of diabetes-related hearing impairment in the U.S. population: National Health and Nutrition Examination Survey 1999-2004. Diabetes Care 33, 811-816.
32. Gaulton, K.J., Ferreira, T., Lee, Y., Raimondo, A., Magi, R., Reschen, M.E., Mahajan, A., Locke, A., William Rayner, N., Robertson, N., et al. (2015). Genetic fine mapping and

genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nat Genet 47, 1415-1425.
33. Gray, S., Feinberg, M.W., Hull, S., Kuo, C.T., Watanabe, M., Sen-Banerjee, S., DePina, A., Haspel, R., and Jain, M.K. (2002). The Kruppel-like factor KLF15 regulates the insulin-sensitive glucose transporter GLUT4. The Journal of biological chemistry 277, 34322-34328.
34. Noguchi, H. (2009). Recent advances in stem cell research for the treatment of diabetes. World journal of stem cells 1, 36-42.
35. Yamaguchi, M., Nakamura, N., Nakano, K., Kitagawa, Y., Shigeta, H., Hasegawa, G., Ienaga, K., Nakamura, K., Nakazawa, Y., Fukui, I., et al. (1998). Immunochemical quantification of crossline as a fluorescent advanced glycation endproduct in erythrocyte membrane proteins from diabetic patients with or without retinopathy. Diabetic medicine : a journal of the British Diabetic Association 15, 458-462.
36. Adewoye, E.O., Akinlade, K.S., and Olorunsogo, O.O. (2001). Erythrocyte membrane protein alteration in diabetics. East African medical journal 78, 438-440.
37. Fernandez-Real, J.M., and Pickup, J.C. (2008). Innate immunity, insulin resistance and type 2 diabetes. Trends in endocrinology and metabolism: TEM 19, 10-16.
38. Satoh, N., Shimatsu, A., Himeno, A., Sasaki, Y., Yamakage, H., Yamada, K., Suganami, T., and Ogawa, Y. (2010). Unbalanced M1/M2 phenotype of peripheral blood monocytes in obese diabetic patients: effect of pioglitazone. Diabetes Care 33, e7.
39. Feinberg, M.W., Wara, A.K., Cao, Z., Lebedeva, M.A., Rosenbauer, F., Iwasaki, H., Hirai, H., Katz, J.P., Haspel, R.L., Gray, S., et al. (2007). The Kruppel-like factor KLF4 is a critical regulator of monocyte differentiation. The EMBO journal 26, 4138-4148.
40. Morton, N.E. (2005). Linkage disequilibrium maps and association mapping. The Journal of clinical investigation 115, 1425-1430.

SUPPLEMENTARY MATERIAL

High resolution genetic maps identify novel T2D loci at regulatory hotspots in African Americans and Europeans

Winston Lau[1], Toby Andrew[2*], Nikolas Maniatis[1*]

[1]Department of Genetics, Evolution and Environment, University College London, London, UK;

[2]Department of Genomics of Common Disease, Imperial College, London, UK;

[*]Equal authorship

Corresponding author: Nikolas Maniatis

Supplementary material includes, Methods, Figure S1 and Table S1

SUPPLEMENTARY METHODS

WELLCOME TRUST CASE CONTROL CONSORTIUM AND AFRICAN AMERICAN SAMPLE SELECTION

The Wellcome Trust Case Control Consortium (WTCCC phase I) described the diagnosis and selection of T2D cases for the original study as "based on either current prescribed treatment with sulphonylureas, biguanides, other oral agents and/or insulin or, in the case of individuals treated with diet alone, historical or contemporary laboratory evidence of hyperglycaemia (as defined by the World Health Organization)"[1]. The two pooled control groups from the 1958 Birth Cohort (aged 44-45) and the UK Blood Service (aged 18-69) are used as shared controls for all seven disease cases (including T2D) in the original study[1], which implies that the controls cannot be assumed to be group matched by BMI, age or sex. The WTCCC2 (phase II) indirectly describes case selection in relation to the Metabochip array design[2], which targets T2D genomic disease loci identified by the DIAGRAM consortium[3]. Although the T2D diagnostic criteria used by the >20 participating research groups varied, most used ADA[4; 5] and WHO[6] guidelines and/or based on treatment with oral anti-diabetic medication or insulin[3]. The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)[7] recruited patients with T2DM and End Stage Renal Disease (ESRD) from dialysis facilities, with the stipulation that cases had to meet at least one of the following three criteria to be included: i) T2DM diagnosed at least 5 years before initiating renal replacement therapy; ii) diabetic retinopathy and/or c) diabetic nephropathy (T2D-ESRD cases).

Given the WTCCC groups did not specify if case and controls were matched by BMI, it cannot be ruled out that gene-mapping studies based upon these European samples might identify genetic loci that are confounded by BMI/ adiposity rather than being associated with T2D alone.

By contrast, this possibility is to some extent mitigated for this study, where the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)[7] samples are BMI matched. WTCCC1 used population controls, which (as noted in the publication[1]) reduces power due the control group including an expected proportion of T2D cases equal to the population prevalence. No published documentation for the selection of WTCCC2 controls is recorded[8]. For the NIDDK, unrelated African-American controls screened for no diagnosis of diabetes or renal disease were recruited from the community and internal medicine clinics (controls)[7]. This suggests that where cosmopolitan T2D disease loci (i.e. the co-location of disease loci for European and African American samples) are identified in this study, we can be more confident that these are not confounded by BMI/ adiposity.

TARGETED RE-SEQUENCING - EUROPEAN CASE/ CONTROL SAMPLES

For the purposes of targeted re-sequencing at the loci *ACTL7B*, *KCNK3* and *TCF7L2*, we used French samples with cases selected from multiplex families from linkage studies with a history of T2D[9; 10] and unrelated controls selected from families with obese individuals, but no history of T2D[11; 12] and 1:1 matched for age, sex and body mass index (BMI, see Table S2).

CONSTRUCTION OF THE GENETIC LDU MAPS

HapMap (Release 28) was used to construct genetic maps for the European samples using 56 unrelated European (EUR) individuals genotyped for 2,270,218 SNPs (screened for quality control) and a second genetic map constructed for the African-American samples using 57 unrelated individuals of African ancestry from South West USA (ASW) genotyped for 1,333,297

quality-control SNPs. The LD maps are based upon HapMap data with genetic distance provided in additive LD units (LDU). The power of the multi-marker approach compared with conventional GWA analysis (single-SNP tests) is primarily provided by the additional information contained in the high resolution LDU maps[13-15] and the reduced number of genomic tests that reduces the multiple-testing burden. The construction of the LDU maps is based on the Malécot-Morton model, which describes the observed decline of pair-wise LD between SNPs as measured by rho, ρ, as an exponential function of physical distance in kilobases (d). The expected decline in pairwise LD is modelled as:

$\hat{\rho} = (1-L)Me^{-\varepsilon d} + L$, with M being the intercept, reflecting the maximum value of LD prior to LD breakdown (~1 for monophyletic origin, i.e. one ancestral haplotype) and L being the asymptote, reflecting spurious LD at large distance, not due to linkage. The parameter ε is the exponential decline of LD and, together with distance d, in kb, an estimate of LDU = $\Sigma\varepsilon_i d_i$ is provided for every ith interval. In this way, all SNPs in the T2D datasets have a genetic location measured in additive LDU[16]. The parameters M, L and ε are the iterative maximum likelihood estimations. The autosomal genome was divided into 4,800 non-overlapping analytical windows of approximately equal size on the genetic map.

ASSOCIATION MAPPING USING LDU MAPS

We carried out association analyses for all autosomal chromosomes using three T2D datasets with a total of 5,800 cases and 9,691 controls. The first genome-wide association (GWA) dataset was obtained from the WTCCC (phase I) and included T2D cases (n=1,925) and controls

(n=2,938) of North European ancestry with available genotypes (Affymetrix, ~500K SNPs)[1; 2; 7]. The second independent dataset was also from the WTCCC (phase II), and included T2D cases (n=2,910) and controls (n=5,724) of North European ancestry (UK)[2], who were genotyped using the Metabochip array (Illumina, ~200K SNPs). The third dataset was obtained from a GWA study for a population of predominantly African ancestry conducted by the NIDDK[7] and included African American (AA) T2D cases (965) and AA controls (1029). The AA NIDDK T2D cases and controls were genotyped at a much higher SNP resolution array (Affymetrix, ~1M SNPs)[7]. All three datasets were screened using standard quality control filters described in previous publications[1; 2; 7] and online data sources. For the eQTLs analysis we used data generated by the MuTHER consortium[17]. Subcutaneous adipose mRNA levels were measured in 825 European twins (TwinsUK) by the MuTHER consortium with data generation and normalization methods described in their initial report and online data sources[17].

The multi-marker association test[18] is based on composite likelihood (Λ)[16; 19], in which all observed genotyped SNPs within each window are simultaneously tested. We therefore do not use imputation and conditional analysis, because the aim of LDU analysis is to estimate the location of functional variants in any given genomic region that provides the strongest evidence of association with disease. For this approach observed (not imputed missing) genotype data are required for reasons explained in the main manuscript. Application of this method to each analytical window returns one estimated location ($\hat{S}$) for the causal variant (± standard error) at the strongest signal, along with the association test *P*-value. The association test is based upon the same Malécot model used to construct the LDU maps described above, although in this case the T2D-by-SNP association (z)[14] is included in the model instead of SNP-by-SNP association

(ρ), along with an additional parameter of causal variant location ($\hat{S}$), with all distances measured in LDU. Therefore the Malécot model prediction of association between disease and markers is estimated by the equation $\hat{z}_i = (1-L)Me^{-\varepsilon|(Si - \hat{S})|} + L$, where S_i the ith SNP LDU location and $\hat{S}$ the estimated location of the putative functional variant on the genetic LDU map. The genetic distance standard errors of $\hat{S}$ were used to obtain the 95% confidence intervals (CI) of the putative causal variants[14], but in this study we only present the co-location intervals (distance between the $\hat{S}$ location estimates), which are the genomic regions within the CIs that most plausibly include the functional variants that confer risk of T2D. For the three gene regions used as examples (Figures 2, 3 and 4 in the manuscript), we constructed LDU maps from the 1000 Genomes Project data, but no differences on the T2D locations estimates were observed based upon the 1000G and HapMap LDU maps.

The regression coefficient *b* was used instead of *z* for the adipose expression quantitative phenotype (eQTL analysis). All the regression coefficients, standard errors and *P*-values for expression probes regressed upon SNPs and probe corresponding gene names were obtained from the MuTHER website (http://www.muther.ac.uk). Our eQTL analysis targeted 174 replicated T2D signals (111 novel and 63 from the previously found list[20]). The Malécot model was then applied after assigning the EUR LDU locations to the SNPs used from the MuTHER data for these 174 signals. The MuTHER probe gene names were updated based on common nomenclature as provide by the UCSC website and to be consistent with the publications that we have referenced in the manuscript.

For convenience, all the functional location estimates ($\hat{S}$) for T2D and eQTLs were converted back to the physical map Build 36 (B36, NCBI36/hg18) in kb by linear interpolation of the two flanking SNPs on the HapMap LDU map. When the $\hat{S}$ was located in an LDU block (horizontal LDU line) then all markers within that block have the same LDU location. In such cases, we took the midpoint of that block as an estimate of $\hat{S}$ in kb. All eQTL locations ($\hat{S}$) had to co-locate within 50kb from the T2D $\hat{S}$ estimates. A detailed description of the Malécot multi-marker test of association is provided by Maniatis et al.[18] and the construction of the LDU maps for this study using the HapMap phase II data are described in more detail elsewhere[14; 16].

Analytical window *P*-values were meta-analysed using Fisher's method to provide overall evidence of association. We did not use other types of meta-analysis (e.g. fixed or random effects), because the multi-marker test of association estimates the causal variant location, but not the association effect size. In order to account for multiple testing, analytical windows were filtered for having a meta *P*-value less than the Bonferroni corrected, genomic *P*-value threshold of $1x10^{-5}$, based on the total number of tests performed (n=4,800; $\alpha = 0.05/4,800$). Loci were only considered biologically plausible if the significant $\hat{S}$ location estimates from different datasets were within a <100 kb interval.

For the eQTL analysis, adipose tissue expression probes were tested for *cis*-association and co-localization, with *cis* defined in this study to be within ±1.5 Mb distance either side of the replicated T2D causal location estimate. This approach provided eQTL location estimates on the LDU maps after Bonferroni correction for the total number of probes tested per 3Mb window. If the eQTL location estimate was within 50kb of the disease susceptibility location, this locus was

considered to be a disease eQTL (i.e. associated with both T2D and *cis*-gene expression) and only these eQTL are presented in the result tables. The results table includes a column for the list of *cis*-genes regulated by identified T2D disease loci.

Here we make an important distinction between an eQTL and an eSNP, which relates to ability to make functional inferences about disease loci. In this study an eQTL is defined by a location estimate for a putative functional variant(s) that regulates gene expression levels for one or more neighbouring genes in a relevant tissue *and* is associated with T2D. In other words, a potential molecular mechanism is immediately suggested for how risk may be conferred by a disease locus, which previously was unknown. By contrast, an eSNP study is defined only by the location of a SNP that is most strongly associated with neighbouring gene expression levels (and may or may not be associated with disease). For eSNP studies, the problems of inconsistency between different lead SNPs associated with disease and expression, between different arrays and across different populations can only be indirectly addressed using genotype imputation methods[21; 22]. For this study, it has been established that the majority of the 111 novel susceptibility loci are also eQTLs. This implies that these disease loci may confer risk of T2D, at least in part, via the *cis*-regulation of expression levels for a large number of neighbouring genes (conservatively, a total of 174 genomic disease loci, both novel and previously known, that regulate the expression levels of a further 267 cis-genes).

Our final set of *cis*-genes (Tables 1 and 2) were then further investigated in order to identify which adipose and liver gene expression profiles have previously also shown evidence of association with body mass index (BMI), a well-established co-morbidity of T2D. We used the

results generated by an independent gene expression study[23] which was based upon 701 subcutaneous adipose and liver samples collected at Massachusetts General Hospital (MGH study) from morbidly obese individuals (BMI >30) who underwent Roux-en-Y gastric bypass surgery.

PREVIOUSLY IDENTIFIED T2D LOCI

We also analysed 76 previously identified T2D loci[20] to obtain refined location estimates on the same genetic maps. For these loci, we undertook commensurable association analyses by centralising the analytical window on the reported lead SNP. These 76 windows were then examined using the same procedures described above to identify T2D locations and assess whether these are eQTL or not. We confirmed 62 out of 76 loci and provide T2D location estimates along with associated *cis*-regulated genes in the supplementary Table S1. Results from the further investigation of the *TCF7L2* locus (signal 117) are provided in the main manuscript. Other notable examples from the supplementary Table S1 are the *HHEX* (signal 149) and *FTO* (signal 152). *HHEX* is observed to regulate *MARCH5* expression levels, which codes for a mitochondrial E3 ubiquitin-protein ligase that plays a crucial role in the control of mitochondrial morphology by acting as a positive regulator of mitochondrial fission[24]. This is the first time a mitochondrial fission gene has been implicated as a risk factor for metabolic disease. Despite testing T2D and not obesity, we observed *FTO* to also be a European T2D disease susceptibility locus with a co-located eQTL that regulates *IRX3*[25]. It is possible this observed association may reflect that the Wellcome Trust T2D cases for this study are overweight and/or poorly matched for BMI with the controls[1; 2]. However, we do not present this result in the table, because while nominally significant (P=0.03) and similar to previous studies[25], the eQTL location for *IRX3* did

not pass Bonferroni correction for the total number of probes tested for this window. We also observed an eQTL within the promoter of *IRX3* that regulates *IRX5*, but we did not further investigate the regulatory landscape of *IRX3*, since the focus of this study was to identify T2D loci that are also eQTL. The *IRX3* was not observed to be associated with T2D either for this or in other studies.

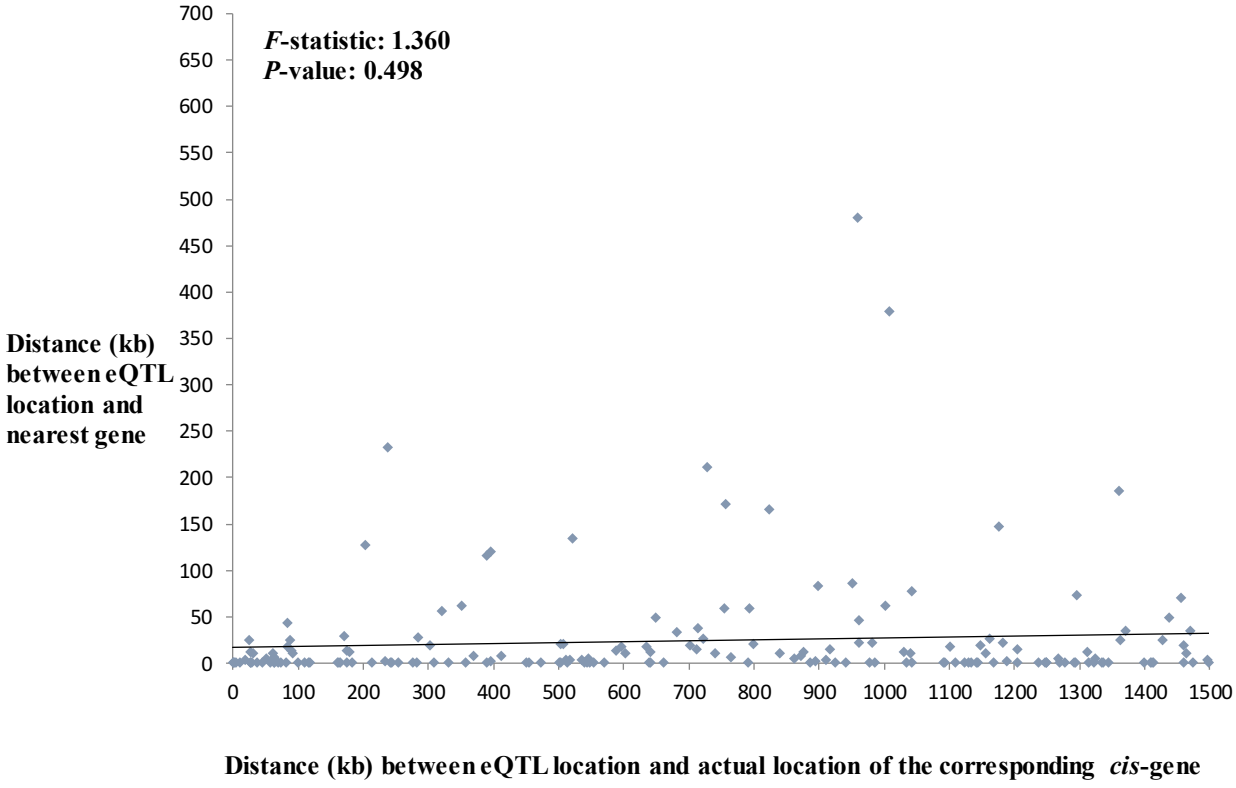
References

1. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661-678.
2. Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet 8, e1002793.
3. Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G., et al. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42, 579-589.
4. (1997). Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Diabetes Care 20, 1183-1197.
5. Expert Committee on the, D., and Classification of Diabetes, M. (2003). Report of the expert committee on the diagnosis and classification of diabetes mellitus. Diabetes Care 26 Suppl 1, S5-20.
6. (1999). Definition, diagnosis and classification of diabetes mellitus, report of a WHO consultation, part 1: diagnosis and classification of diabetes mellitus. In, W.H. Organization, ed. (Geneva.
7. Palmer, N.D., McDonough, C.W., Hicks, P.J., Roh, B.H., Wing, M.R., An, S.S., Hester, J.M., Cooke, J.N., Bostrom, M.A., Rudock, M.E., et al. (2012). A genome-wide association search for type 2 diabetes genes in African Americans. PLoS One 7, e29202.
8. (2008). Wellcome Trust Case Control Consortium 2 (WTCCC2). In, W.T.S. Institute, ed. (Wellcome Trust Sanger Institute.
9. Martin, L.J., Comuzzie, A.G., Dupont, S., Vionnet, N., Dina, C., Gallina, S., Houari, M., Blangero, J., and Froguel, P. (2002). A quantitative trait locus influencing type 2 diabetes susceptibility maps to a region on 5q in an extended French family. Diabetes 51, 3568-3572.
10. Vionnet, N., Hani, E.H., Dupont, S., Gallina, S., Francke, S., Dotte, S., De Matos, F., Durand, E., Lepretre, F., Lecoeur, C., et al. (2000). Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. Am J Hum Genet 67, 1470-1480.
11. Hager, J., Dina, C., Francke, S., Dubois, S., Houari, M., Vatin, V., Vaillant, E., Lorentz, N., Basdevant, A., Clement, K., et al. (1998). A genome-wide scan for human obesity genes reveals a major susceptibility locus on chromosome 10. Nat Genet 20, 304-308.
12. Meyre, D., Lecoeur, C., Delplanque, J., Francke, S., Vatin, V., Durand, E., Weill, J., Dina, C., and Froguel, P. (2004). A genome-wide scan for childhood obesity-associated traits in French families shows significant linkage on chromosome 6q22.31-q23.2. Diabetes 53, 803-811.
13. Lau, W., Kuo, T.Y., Tapper, W., Cox, S., and Collins, A. (2007). Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. Bioinformatics 23, 517-519.
14. Maniatis, N. (2007). Linkage disequilibrium maps and disease-association mapping. Methods Mol Biol 376, 109-121.
15. Maniatis, N., Collins, A., Xu, C.F., McCarthy, L.C., Hewett, D.R., Tapper, W., Ennis, S., Ke, X., and Morton, N.E. (2002). The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci U S A 99, 2228-2233.
16. Maniatis, N., Collins, A., Gibson, J., Zhang, W., Tapper, W., and Morton, N.E. (2004). Positional cloning by linkage disequilibrium. Am J Hum Genet 74, 846-855.

17. Grundberg, E., Small, K.S., Hedman, A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.P., Meduri, E., Barrett, A., et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet 44, 1084-1089.
18. Maniatis, N., Collins, A., and Morton, N.E. (2007). Effects of single SNPs, haplotypes, and whole-genome LD maps on accuracy of association mapping. Genet Epidemiol 31, 179-188.
19. Morton, N., Maniatis, N., Zhang, W., Ennis, S., and Collins, A. (2007). Genome scanning by composite likelihood. Am J Hum Genet 80, 19-28.
20. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segre, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet.
21. Morris, A.P. (2014). Fine mapping of type 2 diabetes susceptibility loci. Curr Diab Rep 14, 549.
22. Price, A.L., Spencer, C.C., and Donnelly, P. (2015). Progress and promise in understanding the genetic basis of common diseases. Proceedings Biological sciences / The Royal Society 282.
23. Greenawalt, D.M., Dobrin, R., Chudin, E., Hatoum, I.J., Suver, C., Beaulaurier, J., Zhang, B., Castro, V., Zhu, J., Sieberts, S.K., et al. (2011). A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. Genome research 21, 1008-1016.
24. Chan, D.C. (2006). Mitochondria: dynamic organelles in disease, aging, and development. Cell 125, 1241-1252.
25. Claussnitzer, M., Hui, C.C., and Kellis, M. (2016). FTO Obesity Variant and Adipocyte Browning in Humans. The New England journal of medicine 374, 192-193.
26. Direk, K., Lau, W., Small, K.S., Maniatis, N., and Andrew, T. (2014). ABCC5 transporter is a novel type 2 diabetes susceptibility gene in European and African American populations. Ann Hum Genet 78, 333-344.

Supplementary Figure S1: No relationship between the distance of the eQTL to the nearest gene (Y-axis) and the distance of the eQTL from the corresponding *cis*-regulated gene (X-axis).

This regression analysis plot demonstrates that the practice of naming disease location estimates after the nearest neighbouring gene annotation is misleading, since the implicated functional genes the eQTLs regulate are just as likely to be distant or near to the eQTL. The same analysis of the Y and X variables, but only including signals where the distance between T2D sample location estimates (Tables 1 and 2) were < 5kb yielded the same result (P>0.05).



Supplementary Figure S2: No relationship between the estimated distance of the eQTL to the T2D location (Y-axis) and the distance between T2D sample location estimates (X-axis; i.e. between EUR and AA in Table 1 and between the two EUR samples in Table 2).

This regression analysis plot shows no relationship between eQTL co-location and disease co-location and demonstrates that the threshold of <100kb used as the criterion for considering estimated disease loci to be co-located and replicated, does not introduce any bias compared to the more conservative threshold of <50kb for the co-location of disease and eQTL.

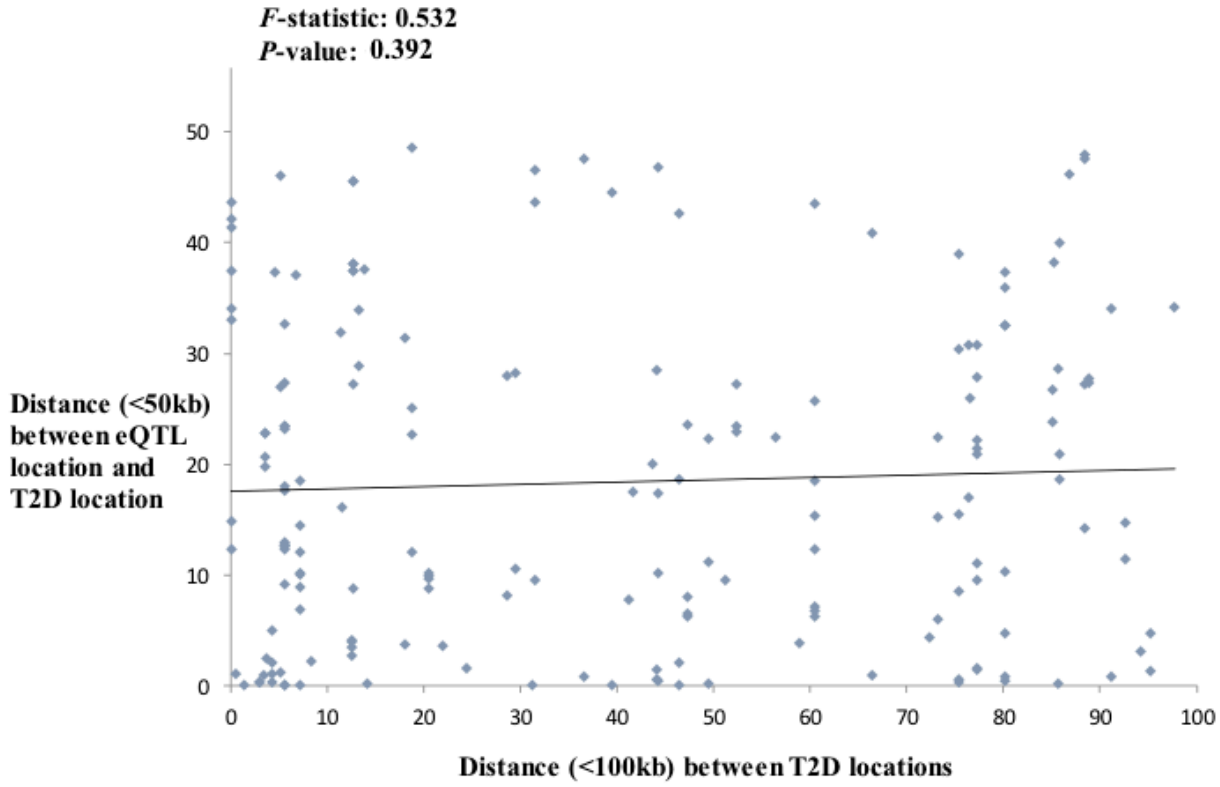


Table S1. Refined information on the previously known T2D loci and their regulatory role of neighbouring gene expression

All locations and distances are given in build 36; ¶Replication with the WTCCC (W), NIDDK AA (A), Metabochip (M) datasets; [a]T2D associated intervals in kb (<100) that harbour T2D locations between datasets; [b]Location estimates for the European (E) GWAS; [c]Location estimates for the African-American (A) GWAS; [d]Location estimates for the Metabochip European (E) samples, signals with low SNP coverage '-' were not meta-analysed; [e]Genes in bold denote the intragenic localization and genes with '+' for self-regulatory; [f]Number of *cis*-genes regulated by the eQTL; [g]List of *cis*-genes associated with eQTLs that co-located within <50kb of the T2D locations on the genetic maps; cis-genes with '*' have previously shown evidence of association between Body Mass Index for morbidly obese and adipose/liver expression profiles[8]; [h]Distance in kb (<50) between eQTL and T2D locations, the minimum is given when more than one *cis*-gene is implicated. Previously observed loci for signals 112-173 are derived from[20] and signal 174 from[26].

Signal	Known locus	Lead SNP	Lead SNP b36	chr	Data¶	Meta *P*-value	Distance between locations[a]	T2D location GWAS-E[b]	T2D location GWAS-A[c]	T2D location metabo-E[d]	Nearest gene to T2D locations[e]	no. of *cis*-genes[f]	eQTL associated cis-genes[g]	eQTL distance from T2D[h]
112	*BCL11A*	rs243088	60422	2p	WAM	4.22E-34	0	60441	60427	60441	*MIR4432*	1	*PEX13*	31
113	*TMEM154*	rs6813195	153740	4q	WAM	1.38E-08	1	153747	153739	153740	*TMEM154*	0	-	-
114	*ANKRD55*	rs459193	55843	5q	WAM	2.15E-14	2	55834	55926	55832	***LOC101928448***	0	-	-
115	*CDKAL1*	rs7756992	20788	6p	WAM	5.99E-180	0	20787	20750	20787	***CDKAL1***	0	-	-
116	*CDKN2A/B*	rs944801	22042	9p	WAM	8.16E-41	1	21987	21986	22022	***CDKN2A/B***	2	*KIAA1797, MTAP*	26
117	*TCF7L2*	rs7903146	114748	10q	WAM	3.55E-86	9	114736	114745	114737	***TCF7L2***	1	*GPAM*	28
118	*RBMS1*	rs7569522	161055	2q	WA	4.42E-21	95	160935	160840	>100kb	*RBMS1*	1	*RBMS1**	4
119	*KCNK16*	rs1535500	39392	6p	WA	1.04E-07	86	39505	39419	-	***KIF6***	0	-	-
120	*ZFAND6*	rs11634397	78219	15q	WA	1.95E-03	67	78193	78126	-	***ZFAND6***	0	-	-
121	*TMEM163*	rs6723108	135196	2q	AM	3.54E-12	1	>100kb	135313	135312	***ACMSD***	0	-	-
122	*KCNQ1*	rs231361	2648	11p	AM	1.96E-14	15	ns	2648	2663	***KCNQ1***	0	-	-
123	*KCNJ11*	rs5215	17365	11p	AM	5.32E-26	2	ns	17384	17382	***ABCC8***	2	*MYOD1, UEVLD*	0
124	*HNF1B*	rs4430796	33172	17q	AM	1.21E-11	30	ns	33135	33165	***HNF1B***	0	-	-
125	*PSMD6*	rs12497268	64065	3p	A	1.87E-05	-	>100kb	63759	-	*C3orf49*	0	-	-
126	*HMGA2*	rs2261181	64499	12q	A	5.53E-03	-	>100kb	64404	ns	*RPSAP52*	0		-
127	*MAEA*	rs6815464	1300	4p	A	1.84E-03	-	ns	1267	ns	*MAEA*	3	*CTBP1, KIAA1530, CRIPAK**	8
128	*ANK1*	rs516946	41638	8p	A	7.57E-07	-	ns	41608	-	*AGPAT6*	1	*ANK1*	14
129	*TLE4*	rs13292136	81142	9q	A	2.97E-02	-	ns	81146	-	*CHCHD9*	0	-	-
130	*FAF1*	rs17106184	50683	1p	A	5.38E-02	-	ns	50894	-	***FAF1***	2	*EPS15, TXNDC12**	4
131	*BCAR1*	rs7202877	73805	16q	A	2.92E-03	-	ns	73490	-	***WDR59***	1	*FA2H*	16
132	*SRR*	rs2447090	2246	17p	A	2.28E-07	-	ns	2039	-	***SMG6***	7	*SRR, RPA1,CAMKK1, ZZEF1, TSR1, SMG6*, TMEM93*	0
133	*PEPD*	rs8182584	38602	19q	A	7.27E-03	-	ns	38543	-	*CEBPG*	0	-	-
134	*ADCY5*	rs11717195	124565	3q	WM	3.23E-23	7	124531	>100kb	124538	*ADCY5*	2	*SEC22A, CCDC14**	15
135	*POU5F1*	rs3130501	31244	6p	WM	3.21E-35	4	31773	>100kb	31777	*LINC00243*	13	*LST1, LY6G6C, C6ORF25, MSH5, SLC44A4*, VARS2, DDR1, FLOT1, ABCF1, HLA-DQB2, TAP2*, TRIM15, TRIM40*	0
136	*DGKB*	rs6960043	15019	7p	WM	3.90E-47	2	15034	>100kb	15032	***DGKB***	0	-	-
137	*TSPAN8*	rs7955901	69720	12q	WM	4.01E-37	11	69867	>100kb	69878	*TSPAN8*	2	*LRRC10, FRS2*	0
138	*MPHOSPH9*	rs4275659	122014	12q	WM	4.46E-05	61	121953	>100kb	122014	*VPS37B,* ***ABCB9***[+]	9	*DNAH10, PITPNM2, ABCB9, VPS37B, TMED2, RSRC2, ZCCHC8, NCOR2, DIABLO*	0
139	*HMG20A*	rs7177055	75620	15q	WM	2.25E-04	40	75058	>100kb	75098	*PSTPIP1*	2	*HMG20A, TSPAN3*	0
140	*IRS1*	rs7578326	226729	2q	WM	3.08E-41	59	226788	ns	226729	***LOC646736***	0	-	-
141	*PPARG*	rs13081389	12265	3p	WM	4.33E-17	18	12311	ns	12292	***PPARG***	1	*WNT7A*	6
142	*ADAMTS9*	rs6795735	64680	3p	WM	1.01E-13	1	64707	ns	64706	***ADAMTS9***	0	-	-
143	*IGF2BP2*	rs4402960	186994	3q	WM	1.89E-23	1	187031	ns	187032	***IGF2BP2***	0	-	-
144	*ARL15*	rs702634	53307	5q	WM	2.85E-04	99	53347	ns	53248	***ARL15***	1	*FST**	19
145	*ZBED3*	rs6878122	76463	5q	WM	1.81E-11	0	76457	ns	76457	***ZBED3***	1	*PDE8B**	16

146	*JAZF1*	rs849135	28163	7p	WM	2.51E-67	90	28226	ns	28136	***JAZF1***	0	-	-
147	*KLF14*	rs13233731	130088	7q	WM	2.59E-06	46	130074	ns	130120	*KLF14*	0	-	-
148	*TP53INP1*	rs7845219	96007	8q	WM	1.28E-11	97	96132	ns	96035	*NDUFAF6, TP53INP1*	6	*GDF6, GEM, MTERFD1, FAM92A1, C8orf37, KIAA1429*	0
149	*HHEX/IDE*	rs1111875	94453	10q	WM	2.27E-61	21	94490	ns	94469	*HHEX*	1	*MARCH5*	1
150	*HNF1A*	rs12427353	119911	12q	WM	3.79E-31	74	119794	ns	119720	***SPPL3***	1	*MSI1*	48
151	*PRC1*	rs8042680	89322	15q	WM	3.68E-56	59	89245	ns	89304	*MAN2A2,* ***RCCD1***$^{+}$	4	*RCCD1, UNC45A, IQGAP1*, FAM174B**	1
152	*FTO*	rs9936385	52377	16q	WM	2.52E-176	11	52368	ns	52357	***FTO***	0	-	-
153	*MC4R*	rs12970134	56036	18q	WM	5.19E-21	1	55879	ns	55880	*RPS3A*	0	-	-
154	*GCKR*	rs780094	27595	2p	M	2.16E-17	-	>100kb	>100kb	27228	***TCF23***	2	*PLB1, KHK*	6
155	*GCC1*	rs17867832	126784	7q	M	1.01E-02	-	>100kb	ns	126964	*GCC1*	1	*IMPDH1*	23
156	*NOTCH2*	rs10923931	120319	1p	M	1.42E-24	-	ns	ns	120238	***ADAM30***	0	-	-
157	*PROX1*	rs2075423	212221	1q	M	1.71E-04	-	ns	ns	212226	***PROX1***	0	-	-
158	*THADA*	rs10203174	43544	2p	M	4.35E-15	-	ns	ns	43555	***THADA***	0	-	-
159	*GRB14*	rs13389219	165237	2q	M	7.31E-03	-	ns	ns	165209	*GRB14*	1	*SCN2A*	32
160	*WFS1*	rs4458523	6341	4p	M	2.61E-46	-	ns	ns	6359	*WFS1*	3	*GRPEL1, STK32B, KIAA0232*	2
161	*SLC30A8*	rs3802177	118254	8q	M	6.10E-05	-	ns	ns	118251	***SLC30A8***	1	*SAMD12*	0
162	*GLIS3*	rs10758593	4282	9p	M	2.45E-10	-	ns	ns	4273	***GLIS3***	0	-	-
163	*CDC123*	rs11257655	12348	10p	M	4.54E-06	-	ns	ns	12189	***DHTKD1***	0	-	-
164	*ARAP1*	rs1552224	72111	11q	M	1.67E-03	-	ns	ns	72534	*FCHSD2*	1	*POLD3*	8
165	*CILP2*	rs10401969	19269	19p	M	1.35E-39	-	ns	ns	19188	***NCAN***	3	*ATP13A1, KIAA0892, TM6SF2**	2
166	*GIPR*	rs8108269	50850	19q	M	1.49E-04	-	ns	ns	51124	*NOVA2*	0	-	-
167	*RND3*	rs7560163	151346	2q	W	4.53E-02	-	151248	ns	-	*LOC101929282*	0	-	-
168	*SSR1*	rs9505118	7235	6p	W	4.31E-02	-	7229	>100kb	-	***SSR1***	1	*BMP6*	46
169	*ZMIZ1*	rs12571751	80613	10q	W	4.06E-04	-	80700	ns	-	***ZMIZ1***	1	*DYDC2*	9
170	*GRK5*	rs10886471	121139	10q	W	3.28E-02	-	121233	ns	-	*RGS10*	0	-	-
171	*CCND2*	rs11063069	4245	12p	W	1.04E-02	-	4170	ns	-	*CCND2*	0	-	-
172	*VPS26A*	rs1802295	70601	10q	W	1.03E-02	-	70421	ns	ns	***KIAA1279***	1	*HERC4*	49
173	*C2CD4A*	rs4502156	60170	15q	W	2.02E-02	-	59905	ns	ns	*VPS13C*	4	*APH1B, RORA, VPS13C, TPM1*	14
174	*ABCC5*	-	-	3q	WA	1.00E-07	0	185136	185136	-	***ABCC5***$^{+}$	1	*ABCC5*	0

Table S2. Demographic characteristics for targeted re-sequence T2D case/ control European samples.

Cases	Variable	Obs	Mean	Std Dev.	Min	Max
Female	Age	57	47.4	7.0	26.0	72.0
	BMI	57	27.1	4.6	17.6	34.7
Male	Age	49	43.5	7.5	20.0	53.0
	BMI	49	25.9	3.5	17.6	34.5
Controls						
Female	Age	57	47.8	7.3	26.0	72.0
	BMI	57	27.7	4.0	21.1	34.8
Male	Age	49	40.7	7.1	20.0	53.0
	BMI	49	27.0	3.6	18.7	34.5