A comprehensive reanalysis of publicly available GWAS datasets reveals an X chromosome rare regulatory variant associated with high risk for type 2 diabetes.

Sílvia Bonàs-Guarch[1], Marta Guindo-Martínez[1], Irene Miguel-Escalada[2,3,4], Niels Grarup[5], David Sebastian[3,6,7], Elias Rodriguez-Fos[1], Friman Sánchez[1,8], Mercè Planas-Fèlix[1], Paula Cortes-Sánchez[1], Santi González[1], Pascal Timshel[5,9], Tune H Pers[5,9,10,11], Claire C. Morgan[4], Ignasi Moran[4], Juan R González[12,13,14], Ehm A. Andersson[5], Carlos Díaz[8], Rosa M. Badia[8,15], Miriam Udler[16,17], Jason Flannick[16,17,18], Torben Jørgensen[19,20,21], Allan Linneberg[19,22,23], Marit E. Jørgensen[24,25], Daniel R. Witte[26,27], Cramer Christensen[28], Ivan Brandslund[29,30], Emil V. Appel[5], Robert A. Scott[31], Jian'an Luan[31], Claudia Langenberg[31], Nicholas J. Wareham[31], InterAct Consortium[32], The SIGMA T2D consortium[32], Oluf Pedersen[5], Antonio Zorzano[3,6,7], Jose C Florez[16,17,33], Torben Hansen[5,34], Jorge Ferrer[2,3,4], Josep Maria Mercader[1,16,17,]*, David Torrents[1,35,]*.

*Both authors jointly directed this work.

1. Barcelona Supercomputing Center (BSC). Joint BSC-CRG-IRB Research Program in Computational Biology, 08034 Barcelona, Spain.
2. Genomic Programming of Beta-cells Laboratory, Institut d'Investigacions August Pi i Sunyer (IDIBAPS), 08036 Barcelona, Spain.
3. Instituto de Salud Carlos III, Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Madrid, Spain.
4. Department of Medicine, Imperial College London, London W12 0NN, United Kingdom
5. The Novo Nordisk Foundation Center for Basic Metabolic Research, Section for Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, 2100, Denmark.
6. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10-12, 08028 Barcelona, Spain.
7. Departament de Bioquímica i Biomedicina Molecular, Facultat de Biologia, Universitat de Barcelona, 08028 Barcelona, Spain.
8. Computer Sciences Department, Barcelona Supercomputing Center (BSC-CNS), 08034 Barcelona, Spain.
9. Department of Epidemiology Research, Statens Serum Institut, Copenhagen, 2300 Denmark.
10. Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, 02116, USA.
11. Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, 0214, USA.
12. ISGlobal, Centre for Research in Environmental Epidemiology (CREAL).
13. CIBER Epidemiología y Salud Pública (CIBERESP).
14. Universitat Pompeu Fabra (UPF).
15. Artificial Intelligence Research Institute (IIIA), Spanish Council for Scientific Research (CSIC).

16. Programs in Metabolism and Medical & Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.
17. Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA.
18. Department of Molecular Biology, Harvard Medical School, Boston, Massachusetts, USA.
19. Research Centre for Prevention and Health, Capital Region of Denmark, Copenhagen, Denmark.
20. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
21. Faculty of Medicine, University of Aalborg, Aalborg, Denmark.
22. Department of Clinical Experimental Research, Rigshospitalet, Glostrup, Denmark.
23. Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
24. Steno Diabetes Center, Gentofte, Denmark.
25. National Institute of Public Health, Southern Denmark University, Denmark.
26. Department of Public Health, Aarhus University, Aarhus, Denmark.
27. Danish Diabetes Academy, Odense, Denmark.
28. Medical department, Lillebaelt Hospital, Vejle, Denmark.
29. Department of Clinical Biochemistry, Lillebaelt Hospital, Vejle, Denmark.
30. Institute of Regional Health Research, University of Southern Denmark, Odense, Denmark.
31. MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK.
32. Members of the consortia are provided in Appendix S1.
33. Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA.
34. Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark.
35. Institució Catalana de Recerca i Estudis Avançats (ICREA).

Corresponding authors:

Josep M Mercader
Life Science Department
Barcelona Supercomputing Center (BSC)
C/Jordi Girona 29, Edifici Nexus II
08003 Barcelona, CATALUNYA, Spain
Phone: 3493 413 40 74
E-mail: Josep.mercader@bsc.es

and

David Torrents
Life Science Department
Barcelona Supercomputing Center (BSC)
Institució Catalana de Recerca i Estudis Avançats (ICREA)
C/Jordi Girona 29, Edifici Nexus II
08003 Barcelona, CATALUNYA, Spain
Phone: 3493 413 40 74
E-mail: david.torrents@bsc.es

Abstract

The reanalysis of publicly available GWAS data represents a powerful and cost-effective opportunity to gain insights into the genetics and pathophysiology of complex diseases. We demonstrate this by gathering and reanalyzing public type 2 diabetes (T2D) GWAS data for 70,127 subjects, using an innovative imputation and association strategy based on multiple reference panels (1000G and UK10K). This approach led us replicate and fine map 50 known T2D *loci*, and identify seven novel associated regions: five driven by common variants in or near *LYPLAL1*, *NEUROG3*, *CAMKK2*, *ABO* and *GIP* genes; one by a low frequency variant near *EHMT2*; and one driven by a rare variant in chromosome Xq23, associated with a 2.7-fold increased risk for T2D in males, and located within an active enhancer associated with the expression of Angiotensin II Receptor type 2 gene (*AGTR2*), a known modulator of insulin sensitivity. We further show that the risk T allele reduces binding of a nuclear protein, resulting in increased enhancer activity in muscle cells. Beyond providing novel insights into the genetics and pathophysiology of T2D, these results also underscore the value of reanalyzing publicly available data using novel analytical approaches.

Keywords: dbGaP, publicly available GWAS data, data sharing, genotype imputation, type 2 diabetes, rare variants, 1000 Genomes, UK10K, data sharing, enhancer, AGTR2, angiotensin receptor 2.

During the last decade, hundreds of genome-wide association studies (GWAS) have been performed with the aim of providing a better understanding of the biology of complex diseases, improving their risk prediction, and ultimately discovering novel therapeutic targets[1]. However, the majority of the published GWAS just report the primary findings, which generally explain a small fraction of the estimated heritability. In order to better uncover the missing heritability, most strategies usually involve the generation of new genetic and clinical data. Very rarely, new studies are based on the revision and reanalysis of existing genetic data by applying more powerful analytic techniques and resources generated after the primary GWAS findings are published. These cost-effective reanalysis strategies are now possible, given the existence of (1) different data-sharing initiatives that gather large amounts of primary genotype and sequencing data for multiple human genetic diseases, as well as (2) new and improved GWAS methodologies and resources. Notably, genotype imputation with novel and denser sequence-based reference panels can now substantially increase the genetic resolution of GWASs from previously genotyped datasets[2], reaching good quality imputation of low-frequency (minor allele frequency [MAF]: $0.01 \leq MAF < 0.05$) and rare variants ($MAF < 0.01$). This increases the power to identify novel associations, as well as to refine known associated *loci* through fine-mapping. Moreover, the availability of public primary genetic data further allows the homogeneous integration of multiple datasets with different origins providing more accurate meta-analysis results, particularly at the low ranges of allele frequency. In contrast, common meta-analytic strategies usually involve the aggregation of summary statistics generated independently by different centers, each applying their own internal quality-control, data cleaning, genotype imputation, and association testing methodologies. Each of these steps represents a potential source of additional heterogeneity, which may ultimately result in loss of statistical power. Finally, the vast majority of reported GWAS analyses omits the X chromosome, even though it

represents 5% of the genome and encodes for more than 1,500 genes[3,4]. The reanalysis of publicly available data also enables interrogation of this chromosome.

We hypothesized that a unified reanalysis of multiple publicly available datasets, applying homogeneous standardized QC, genotype imputation and association methods, as well as novel and denser sequence-based reference panels for imputation would provide new insights into the genetics and the pathophysiology of complex diseases. To test this we focused on type 2 diabetes (T2D), one of the most prevalent complex diseases for which many GWAS have been performed during the past decade[5,6,7,8,9,10,11]. These studies have allowed the identification of more than 100 independent *loci*, most of them driven by common variants, with a few exceptions[11,12,13]. Despite all these efforts, only a small fraction of the genetic heritability can be explained by established T2D *loci*, and the role of low-frequency variants, although recently proposed to be minor[14], has still not been fully explored. The availability of large T2D genetic datasets in combination with larger and more comprehensive genetic variation reference panels[15,16,17,18], gives us the opportunity to impute a significant increased fraction of low-frequency and rare variants, and to study their contribution to the risk of developing this disease. At the same time, this strategy also allows to fine map known associated *loci*, increasing the chances of finding causal variants and understanding their functional impact. We therefore gathered publicly available T2D GWAS cohorts with European ancestry, comprising a total of 13,857 T2D cases and 62,126 controls, to which we first applied harmonization and quality control protocols across the whole genome (including the X chromosome) and across samples, to later carry out imputation using 1,000 Genomes Project (1000G)[15] and UK10K[16,17,18] reference panels, as well as association testing. By using this strategy, we identified novel associated regions driven by common, low-frequency, and rare variants, fine-mapped and functionally annotated the existing and novel ones, and

confirmed experimentally a regulatory mechanism disrupted by a novel rare and large effect variant identified at the X-chromosome.

Results

Overall analysis strategy

As shown in Figure 1, we first gathered all T2D case-control GWAS individual-level data that was available through the EGA and dbGaP databases (i.e. Gene Environment Association Studies [GENEVA], Welcome Trust Case Control Consortium [WTCCC], Finland-United States Investigation of NIDDM Genetics, Resource for Genetic Epidemiology Research on Aging [GERA] and Northwester NuGENE project [NuGENE]). We harmonized these cohorts, standardized quality control and the filtering of low-quality variants and samples (see Online Methods). After this process, a total of 70,127 subjects (70KforT2D; 12,931 cases and 57,196 controls; Supplementary Table 1) were retained for downstream analysis. Each of these cohorts was then imputed to the 1000G and UK10K reference panels using an integrative method, that selects for each variant the results from the reference panel that provides the highest accuracy, according to IMPUTE2 info score (see Online Methods). Finally, the results from each of these cohorts were meta-analyzed (Figure 1), obtaining a total of 15,115,281 variants with good imputation quality (IMPUTE2 info score ≥ 0.7, MAF ≥ 0.001, and I^2 heterogeneity score < 0.75), across 12,931 T2D cases and 57,196 controls. Of these, 6,845,408 variants were common (MAF ≥ 0.05), 3,100,848 were low-frequency ($0.01 \leq$ MAF < 0.05) and 5,169,025 were rare ($0.001 \leq$ MAF < 0.01). Interestingly, the strategy of merging the imputation results derived from the UK10K and 1000G reference panels substantially improved the number of good quality imputed single nucleotide variants (SNVs) and insertion/deletions (*indels*), particularly within the low-frequency and rare spectrum, when compared with the imputation results obtained with each of the reference panels separately. For example, a set of 5,169,025 rare variants with good

quality was obtained after integrating 1000G and UK10K results, while only 2,878,263 rare variants were imputed with 1000G and 4,066,210 with UK10K (Supplementary Figure 1 A). This strategy also allowed us to impute of 1,357,753 *indels* with good quality (Supplementary Figure 1 B).

In order to take full advantage of publicly available genetic data, we used three main meta-analytic approaches to adapt to three of the most common strategies for genetic data sharing: individual-level genotypes, summary statistics, as well as single-case queries through web services, like the Type 2 Diabetes Knowledge Portal (http://www.type2diabetesgenetics.org/). These three approaches allowed us to maximize the power of detection, by prioritizing largest sample sizes and best imputation quality, in particular for low-frequency, rare variants, and coding variants. We first meta-analyzed all summary statistics results from the DIAGRAM trans-ancestry meta-analysis[6] (26,488 cases and 83,964 controls), selecting 1,918,233 common variants (MAF $\geq$ 0.05), mostly imputed from HapMap, with the corresponding fraction of non-overlapping samples in our 70KforT2D set, i.e. the GERA and the NuGENE cohorts, comprising a total of 7,522 cases and 50,446 controls (Figure 1, Supplementary Table 1). Second, the rest of variants (13,197,048), consisting of non-HapMap variants (mostly with MAF < 0.05) or not tested above, were meta-analyzed using all five cohorts that constitute the 70KforT2D resource (Supplementary Table 1). Finally, low-frequency variants located in coding regions and with $p \leq 1\times10^{-4}$ were meta-analyzed using the non-overlapping fraction of samples with the data from the Type 2 Diabetes Knowledge Portal (http://www.type2diabetesgenetics.org) through the interrogation of exome array and whole-exome sequence data from ~80,000 and ~17,000 individuals, respectively[19,20,21].

T2D associated variants show enrichment of pathways involved in insulin response and pancreatic islet enhancers.

As a first exploration of how our association results recapitulate the pathophysiology of T2D, we performed gene-set enrichment analysis with all the variants with p-value $\leq 1\times10^{-5}$ using DEPICT[22] (see Online Methods). This analysis showed enrichment of genes expressed in pancreas (ranked 1st in tissue enrichment analysis, $p=7.8\times10^{-4}$, FDR<0.05, Supplementary Table 2) and cellular response to insulin stimulus (ranked 2nd in gene set enrichment analysis, $p=3.9\times10^{-8}$, FDR=0.05, Supplementary Table 3, Supplementary Figure 2, Supplementary Figure 3), in concordance with the current knowledge of the molecular basis of T2D.

In addition, variant set enrichment analysis of the T2D-associated credible sets across regulatory elements defined in isolated human pancreatic islets showed a significant enrichment for active regulatory enhancers (Supplementary Figure 4), suggesting that causal SNPs within associated region have a regulatory function, as previously reported[23].

Identification, fine-mapping and functional characterization of novel and previously known loci

The three association strategies allowed us to finally identify 57 genome-wide significant associated *loci* ($p\leq5\times10^{-8}$), of which seven were not previously reported as associated with T2D (Table 1). The remaining 50 *loci* were already known, and included, for example, two low-frequency variants recently discovered in Europeans, one located within one of the *CCND2* introns (rs76895963) and a missense variant within the *PAM*[13] gene. Furthermore, as a quality control of these results, we confirmed that the magnitude and direction of the effect of all the associated variants ($p\leq0.001$) were highly consistent with those reported previously (Rho=0.92, $p=1\times10^{-248}$, Supplementary Figure 5). In fact, the direction of effect was consistent with all 139 previously reported variants except three that were discovered in east and south Asian populations (Supplementary Table 4).

The high coverage of genetic variation achieved in this study allowed us to fine-map known and novel *loci*, providing more candidate causal variants for downstream functional interpretations. For this, we constructed 99% credible variant sets[24], -i. e. the subset of variants that have, in aggregate, 99% probability of containing the true causal variant- for all 57 *loci* (Supplementary Table 5). As an important improvement over previous T2D genetic studies, we identified small structural variants within the credible sets, consisting mostly of insertions and deletions between 1 and 1,975 nucleotides. In fact, out of the 8,348 variants included within the credible sets for these *loci*, 927 (11,1%) were *indels*, of which 105 were genome-wide significant (Supplementary Table 6). Interestingly, by integrating imputed results from 1000G and UK10K reference panels we gain up to 41% of *indels*, which were only identified by one of the two reference panels, confirming the advantage of integrating the results from both reference panels. Interestingly, 15 of the 71 previously reported *loci* we replicated ($p \leq 5.3\times10^{-4}$ after correcting for multiple testing), have an *indel* as the top variant, highlighting the potential role of this type of variation in the susceptibility for T2D. For example, within the *IGF2BP2* intron, a well-established and functionally validated *locus* for T2D[25,26], we found that 12 of the 57 variants within its 99% credible set correspond to *indels* with genome-wide significance ($5.6\times10^{-16} < p < 2.4\times10^{-15}$), which collectively represent 18.4% posterior probability of being causal.

To prioritize causal variants within all the identified associated *loci*, we annotated their corresponding credible sets using the variant effector predictor (VEP) for coding variants[27] (Supplementary Table 7), and the Combined Annotation Dependent Depletion (CADD) tool for non-coding variation (Supplementary Table 8)[28]. Additionally, we tested the effect of all variants on expression across multiple tissues by interrogating GTEx[29] and RNA-sequencing gene expression data from pancreatic islets[30].

Identification of new signals driven by common variants

Beyond the detailed characterization of the known T2D associated regions, we also identified novel *loci*, among which, five were driven by common variants linked to modest effect sizes (1.06 < OR <1.12; Table 1, Figure 2, Supplementary Figure 6 and 7).

Within the first novel T2D-associated *locus* in chromosome 1q41 (*LYPLAL1-ZC3H11B*, rs2820443, OR=1.07 [1.04-1.09], $p=2.6\times10^{-8}$), several variants have been previously associated with waist-to-hip ratio in females, height, visceral adipose fat in females, adiponectin levels, fasting insulin, and non-alcoholic fatty liver disease[31,32,33,34,35,36]. Among the genes of this *locus*, *LYPLAL1*, which encodes for lysophospholypase-like 1, appears as the most likely effector gene, as it has been found to be downregulated in mouse models of diet-induced obesity and upregulated during adipogenesis[37].

Second, a novel *locus* at chromosome 9q34.2 region (*ABO*, rs505922, OR=1.06 [1.04-1.09], $p=4.9\times10^{-8}$) includes several variants that have been previously associated with other metabolic traits. For example, the variant rs651007, in linkage disequilibrium (LD) with rs505922 ($r^2=0.507$), has been recently associated with fasting glucose[38]; and rs514659 (r^2 with top=1) is associated with an increased risk for cardio-metabolic disorders[39]. Interestingly one of the variants within the credible set is the one base-pair frame-shift deletion defining the blood group O[40]. In concordance with previous results that linked O blood type with a lower risk of developing T2D[41], the frameshift deletion determining the blood group type O was associated with a protective effect for T2D in our study (rs8176719, $p=3.4\times10^{-4}$, OR=0.95 [0.91-0.98]). In addition, several variants within this credible set are associated with the expression of the *ABO* gene in multiple tissues including skeletal muscle and adipose tissue, and pancreatic islets (Supplementary Table 9, Supplementary Table 10).

Third, a novel *locus* at chromosome 10q22.1 *locus* (*NEUROG3/COL13A1/RPL5P26*, rs2642587, OR=1.12 [1.08-1.16], $p=8.4\times10^{-9}$), includes *NEUROG3* (Neurogenin3), which is

an essential regulator of pancreatic endocrine cell differentiation[42,43]. Mutations in this gene have thus been reported to cause permanent neonatal diabetes[44], but a role of this gene in T2D has not been previously reported[45].

The lead common variant of the fourth novel *locus* at chromosome 12q24.31 (rs3794205, OR=1.07 [1.04-1.10], p=4.1×10^{-8}), lies within an intron of the *CAMKK2* gene, previously implicated in cytokine-induced beta cell death[46]. However other variants within the corresponding credible set could be, in fact, responsible for the molecular action behind this association, like a missense variant within the *P2RX7*, a gene previously associated with glucose homeostasis in humans and mice[47]; or other variant (rs11065504, r^2 with lead variant=0.81) that is associated with the regulation of the *P2RX4* gene in tibial artery and in whole blood according to GTEx (Supplementary Table 9).

The fifth novel *locus* driven by common variants is located within 17q21.32 (rs12453394, OR=1.07 [1.05-1.10], p=3.23×10^{-8}). It involves three missense variants located within the *CALCOCO2*, *SNF8* and *GIP* genes. *GIP* encodes for glucose-dependent insulinotropic peptide, a hormonal mediator of enteral regulation of insulin secretion[48]. Variants in the GIP receptor (GIPR) have previously been associated with insulin response to oral glucose challenge and beta-cell function[49]. *GIP* is thus a plausible candidate effector gene of this *locus*[50].

Identification of a new signal driven by a low-frequency variant

We selected all low-frequency (0.01≤MAF<0.05) variants with p≤1×10^{-4} in the 70KforT2D meta-analysis that were annotated as altering protein-coding variants according to VEP. This resulted in 15 coding variants that were meta-analyzed with exome array and whole-exome sequencing data from a total of ~97,000 individuals[19,20,21] after excluding the overlapping cohorts between the different datasets. This analysis highlighted a novel genome-wide association driven by a low-frequency missense variant (Ser58Phe) within the *EHMT2* gene

at chromosome 6p21.33 (rs115884658, OR=1.21 [1.14-1.29], p=3.00×10^{-10}; Figure 2, Supplementary Figures 6 and 7). *EHMT2* is involved in the mediation of FOXO1 translocation induced by insulin[51]. Since this variant was less than 1 Mb away from *HLA-DQA1*, which was recently reported to be associated with T2D[52], we performed a series of reciprocal conditional analyses and excluded the possibility that our analysis was capturing previously reported T2D[52] or T1D[10,53,54] signals (Supplementary Table 11). Beyond this missense *EHMT2* variant, other low-frequency variants within the corresponding credible set may also be causal. For example, rs115333512 (r^2 with lead variant=0.28) is associated with the expression of *CLIC1* in several tissues according to GTEx (multi-tissue Meta-Analysis $p=8.9\times10^{-16}$, Supplementary Table 9). In addition, this same variant is also associated with the expression of the first and second exon of the *CLIC1* mRNA in pancreatic islet donors[30] (p(exon 1)=$1.4\text{x}10^{-19}$; p(exon 2)=1.9×10^{-13}; Supplementary Table 10). Interestingly, *CLIC1* has been reported as a direct target of metformin by mediating the anti-proliferative effect of this drug in human glioblastoma[55]. All these findings support *CLIC1*, as an additional possible effector transcript, likely driven by rs115333512.

Identification of a novel rare variant in the X chromosome associated with 2.7-fold increased risk for T2D

As for many other complex diseases, the majority of published large-scale T2D GWAS studies have omitted the analysis of the X chromosome, with the notable exception of the identification of a T2D associated region near the *DUSP9* gene in 2010[56]. To fill this gap, we tested the X chromosome genetic variation for association with T2D. To account for heterogeneity of the effects and for the differences in imputation performance between males and females, the association was stratified by sex and tested separately, as well as together by meta-analyzing the results from males and females. This analysis was able to replicate the *DUSP9 locus*, not only through the known rs5945326 variant (OR=1.15, p=0.049), but also

through a three-nucleotide deletion located within a region with several promoter marks in the liver (rs61503151 [G/GCCA], OR=1.25, p=3.5×10^{-4}), and in high LD with the first reported variant (R^2=0.62). Conditional analyses showed that the originally reported variant was no longer significant (OR=1.01, p=0.94) when conditioning by the newly identified variant, rs61503151. On the other hand, when conditioning on the previously reported variant, rs5945326, the effect of the newly identified *indel* remained significant and with higher effect size (OR=1.33, p=0.003), placing this deletion, as a more likely candidate causal variant for this *locus*.

In addition, we identified a novel genome-wide significant signal in males at the Xq23 *locus* driven by a rare variant (rs146662075, MAF=0.008, OR=2.94 [2.00-4.31], p=3.5×10^{-8}; Figure 3A). We tested the accuracy of the imputation of this variant by comparing the imputed results from the same individuals genotyped by two different platforms (see Online Methods) and found that the imputation for this variant was highly accurate in males, and when using UK10K, but not in females, or when using 1000G ($R^2_{[UK10K,males]}$=0.94, $R^2_{[UK10K,females]}$=0.66, $R^2_{[1000G,males]}$=0.62, $R^2_{[1000G,females]}$=0.43, Supplementary Figure 8). Whether this association is specific to men or whether it also affects female carriers remains to be clarified, since the imputation for this variant in females was not accurate enough.

To further validate this association and to discard potential imputation artifacts, we next analyzed two independent cohorts by performing imputation with the UK10K reference panel (SIGMA[57], INTERACT[58]), and a third cohort by *de-novo* genotyping the rs146662075 variant in several Danish sample sets. The initial meta-analysis, including the three replication datasets did not reach genome-wide significance (OR=1.72, p=1.6×10^{-4}) (Figure 3B), and revealed a strong degree of heterogeneity (heterogeneity p_{het}=0.002), which appeared to be driven by the replication cohorts. Within one of the case-control studies there was a nested cohort study; this study, the Inter99 prospective cohort, consisted of 1,652 non-

diabetic male subjects, of whom 158 developed T2D after eleven years of follow-up. Analysis of incident diabetes in this cohort, confirmed the association with the same allele as previously seen in the case-control studies, with carriers of the rare T allele having increased risk of developing incident diabetes, compared to the C carriers (Cox-proportional Hazards Ratio (HR)=3.17 [1.3-7.7], p=0.011, Figure 3C): in fact, nearly 30% carriers of the T risk allele developed incident T2D during 11 years of follow-up, compared to only ~10% of non-carriers.

In order to gain further support for this association, we thoroughly compared the clinical and demographic characteristics of the discovery and replication datasets in an attempt to understand the strong degree of heterogeneity observed when adding the replication datasets. We found that some of the replication datasets contained control subjects that were significantly younger than the average age at onset of T2D reported in this study and in Caucasian populations in general[59]. This was particularly clear for the Danish cohort (age controls [95%CI] = 46.9 [46.6-47.2] vs age cases [95%CI] = 60.7 [60.4-61.0]) and for Interact (age controls [95%CI] = 51.7 [51.4-52.1] vs age cases [95%CI] = 54.8 [54.6-55.1]) (Supplementary Figure 9). Since in the Inter99 prospective cohort we observed that 30% of subjects developed incident T2D over 11 years of follow-up, we performed an additional analysis using a stricter definition of controls, in order to avoid the presence of pre-diabetes or individuals that may further develop diabetes after reaching the average age at onset. We therefore applied two additional exclusion criteria in order to obtain a stricter set of controls: (i) excluding subjects younger than 55 years old, and (ii) restricting the analysis, when possible, to individuals with measured 2 hours plasma glucose values during oral glucose tolerance test (OGTT) below 7.8 mmol/l, which is employed to diagnose impaired glucose tolerance (pre-diabetes), a strong risk factor for developing T2D[60]. We repeated the meta-analysis using the first filtering criterion for controls for both the discovery and replication

datasets, including only controls older than 55. While the application of the first filter did not yet provide genome-wide significant results (Supplementary Figure 10), when we added the second filter (only possible in the Danish cohort), we obtained replication results consistent with the initial discovery results. Moreover, we also integrated the Cox-proportional hazards results into the meta-analysis by using a method that accounts for overlapping subjects (MAOS)[61] allowing us to integrate the longitudinal (follow-up) and the case-control study. This final meta-analysis confirmed the association at rs146662075 resulting in genome-wide significance and without significant heterogeneity (OR=2.7 (1.91, 3.81), p=1.7×10^{-08}, p_{het}=0.51, Figure 3D). These results therefore support the existence of a genetic association with T2D is driven by a rare variant at X chromosome.

The rs146662075 T risk allele is associated with 5-fold greater enhancer activity and disruption of allele specific nuclear protein binding.

We next explored the possible molecular mechanism behind this association by using different genomic resources and experimental approaches. The credible set of this region contained three variants, with the leading SNP alone (rs146662075), showing 78% posterior probability of being causal (Supplementary Figure 7, Supplementary Table 5), as well as the highest CADD score (scaled C-score=15.68, Supplementary Table 8). rs146662075 lies within a chromosomal region enriched in regulatory (DNAse I) and active enhancer (H3K27ac) marks, between the *AGTR2* (at 103 kb) and the *SLC6A14* (at 150Kb) genes. Interestingly, the closest gene, *AGTR2*, which encodes for the angiotensin II receptor type 2, has been previously associated with insulin secretion and resistance[62,63,64]. Analysis of publicly available epigenomic datasets[65] showed that this region lacks H3K27ac enrichment in human islet chromatin, whereas there is a positive correlation between the H3K27ac enhancer marks and the expression of *AGTR2* across multiple tissues, showing the highest

signal of both H3K27ac and *AGTR2* RNA-seq (but not *SLC6A14*) expression in fetal muscle (Figure 4A).

We next studied whether the region encompassing the rs146662075 variant could act as a transcriptional enhancer and whether its activity was allele-specific. For this, we linked the DNA region, with either the T (risk) or the C (common) allele, to a minimal promoter and performed luciferase assays in a mouse myoblast cell line. The luciferase analysis showed an average 4.4 fold increased activity for the disease-associated T allele, compared to the expression measured with the common C allele, suggesting an activating function of the T allele, or a repressive function of the C allele (Figure 4B). Consistent with these findings, electrophoretic mobility shift assays using nuclear protein extracts from mouse myoblast cell lines, differentiated myotubes, and human fetal muscle cell line, revealed sequence-specific binding activity of the C allele, but not of the rare T allele (Figure 4C). Overall, these data indicate that the risk T allele prevents binding by a nuclear protein that is associated with decreased activity of an *AGTR2*-linked enhancer.

Discussion

To date, a large fraction of GWAS datasets have been made publicly available through repositories such as dbGaP and EGA. However, how these resources are useful to increase the knowledge of the genetic factors contributing to disease susceptibility has not been fully explored. In this study, we demonstrate the utility of these publicly available resources. By harmonizing and reanalyzing existing and publicly available T2D GWAS data, and by performing genotype imputation with two whole-genome sequence-based reference panels, we have been able to explore deeply the genetic architecture of T2D. This strategy allowed us to impute and test for association with T2D more than 15 million of high quality imputed variants, including low-frequency, rare, and small insertions and deletions across chromosomes 1 to 22 and X.

The reanalysis of these data confirmed a large fraction of already known T2D *loci*, for which we identify and propose novel potential causal variants by fine-mapping and functionally annotating each *locus*. Our in-depth characterization of each of these *loci* reached an improved accuracy over prior efforts, providing for the first time a comprehensive coverage of structural variants, which point to previously unobserved candidate causal variants in known and novel *loci.*

This reanalysis also allowed us to identify seven novel associations driven by common variants in or near *LYPLAL1*, *NEUROG3*, *CAMKK2*, *ABO* and *GIP;* a low-frequency variant in *EHMT2*, and one rare variant in the X chromosome. This rare variant identified in Xq23 chromosome, was located near the *AGTR2* gene, and showed a nearly three-fold increased risk for T2D in males, which represents, to our knowledge, the largest effect size identified so far in Europeans, and complements other large effect size variants identified in other specific populations[11,12].

This study also highlights the importance of a strict classification of both cases and controls in order to identify rare variants associated with disease, as our initial discovery for the Xq23 *locus* was not replicated when using the standard classification of cases and controls (based on fasting plasma glucose or glycated hemoglobin, HbA1c), but only when reclassifying the control group by restricting it to non-diabetic individuals older than 55 years (average age at onset of T2D), and with confirmed normal glucose tolerance using the stricter OGTT measure. This is in line with previous results obtained for a T2D population specific variant found in Inuit within the *TBC1D4* gene, which was only significant when using OGTT as criteria for classifying cases and controls, but not when using HbA1c[11]. The use of this strict criterion for T2D diagnosis was in fact very clinically relevant, as 32% carriers of the *TB1CD4* variant would remain undiagnosed T2D or prediabetes patients when OGTT was not assessed[66]. Despite the need for a stricter definition of controls in our replication case-

control study, our results were further confirmed by the observation that 30% of the rs146662075 risk allele carriers developed T2D over 11 years of follow-up, compared to 10% of non-carriers, and suggest that early identification of these subjects by genotyping this variant may be useful to tailor pharmacological or lifestyle intervention in order to prevent or delay their onset of T2D.

Using binding and gene-reporter analyses we demonstrate a functional role of this variant and propose a possible mechanism behind the pathophysiology of T2D in T risk allele carriers, where this rare variant could favor a gain of function of *AGTR2*, which has been previously associated with insulin resistance[62]. *AGTR2* appears, therefore, as a potential therapeutic target for this disease, which would be in line with previous studies that showed that the blockade of the renin-angiotensin system in mice[67] and in humans[68] prevents the onset of type 2 diabetes and restores normoglycemia[69,70].

Overall, beyond our significant contribution to the understanding of the molecular basis of T2D, our study also highlights the potential of the reanalysis of public data, as a complement to large studies using newly-generated data. This study will inform the open debate in favor of pushing data-sharing and democratization initiatives[71,72], highlighting their importance for the study of the genetics and pathophysiology of complex diseases, which may lead to new preventive and therapeutic applications.

Methods

Online Content Methods, along with any additional Extended Data display items, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Data availability: The complete summary statistics are deposited at the Type 2 Diabetes Knowledge portal (www.type2diabetesgenetics.org/).

Acknowledgements:

This work has been sponsored by the grant SEV-2011-00067 of Severo Ochoa Program, awarded by the Spanish Government. This work was supported by an EFSD/Lilly research fellowship. Josep M. Mercader was supported by Sara Borrell Fellowship from the Instituto Carlos III. Sílvia Bonàs was FI-DGR Fellowship from FI-DGR 2013 from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, Generalitat de Catalunya). This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. This study also makes use of data generated by the UK10K Consortium, derived from samples from UK10K COHORT IMPUTATION (EGAS00001000713). A full list of the investigators who contributed to the generation of the data is available in www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310. We acknowledge PRACE for awarding us access to MareNostrum supercomputer, based in Spain at Barcelona. The technical support group from the Barcelona Supercomputing Center is gratefully acknowledged. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 667191. Mercè Planas-Fèlix is funded by the Obra Social Fundación la Caixa fellowship under the Severo Ochoa 2013 program. Irene Miguel-Escalada has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 658145. The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (www.metabol.ku.dk). We also thank Marcin von Grotthuss for their support for uploading the summary statistics data to the Type 2 Diabetes Genetic Portal (AMP-T2D portal). Finally,

we thank all the Computational Genomics group at the BSC for their helpful discussions and valuable comments on the manuscript.

References:

1. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-6 (2014).
2. Zheng, H.F. *et al.* Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* (2015).
3. Wise, A.L., Gyi, L. & Manolio, T.A. eXclusion: toward integrating the X chromosome in genome-wide association analyses. *Am J Hum Genet* **92**, 643-7 (2013).
4. Tukiainen, T. *et al.* Chromosome X-wide association study identifies Loci for fasting insulin and height and evidence for incomplete dosage compensation. *PLoS Genet* **10**, e1004127 (2014).
5. Replication, D.I.G. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**, 234-44 (2014).
6. SIGMA Type 2 Diabetes Consortium *et al.* Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97-101 (2014).
7. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**, 981-990 (2012).
8. Ng, M.C. *et al.* Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet* **10**, e1004517 (2014).
9. Kim, Y.J. *et al.* Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nat Genet* **43**, 990-5 (2011).
10. Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
11. Moltke, I. *et al.* A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190-3 (2014).
12. Sigma Type 2 Diabetes Consortium *et al.* Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA* **311**, 2305-14 (2014).
13. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* **46**, 294-8 (2014).
14. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41-7 (2016).
15. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
16. Stevens, E. *et al.* Mutations in B3GALNT2 cause congenital muscular dystrophy and hypoglycosylation of alpha-dystroglycan. *Am J Hum Genet* **92**, 354-65 (2013).

17. Consortium, U.K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* (2015).
18. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* **6**, 8111 (2015).
19. Mahajan, A. *et al.* Identification and functional characterization of G6PC2 coding variants influencing glycemic traits define an effector transcript at the G6PC2-ABCB11 locus. *PLoS Genet* **11**, e1004876 (2015).
20. Flannick, J. *et al.* Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat Genet* **46**, 357-63 (2014).
21. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* (2016).
22. Pers, T.H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **6**, 5890 (2015).
23. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* **46**, 136-43 (2014).
24. Wellcome Trust Case Control, C. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294-301 (2012).
25. Diabetes Genetics Initiative of Broad Institute of, H. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331-6 (2007).
26. Dai, N. *et al.* IGF2BP2/IMP2-Deficient mice resist obesity through enhanced translation of Ucp1 mRNA and Other mRNAs encoding mitochondrial proteins. *Cell Metab* **21**, 609-21 (2015).
27. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).
28. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
29. Mele, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660-5 (2015).
30. Fadista, J. *et al.* Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc Natl Acad Sci U S A* **111**, 13924-9 (2014).
31. Manning, A.K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* **44**, 659-69 (2012).
32. Randall, J.C. *et al.* Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet* **9**, e1003500 (2013).
33. Berndt, S.I. *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet* **45**, 501-12 (2013).
34. Dastani, Z. *et al.* Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet* **8**, e1002607 (2012).
35. Fox, C.S. *et al.* Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS Genet* **8**, e1002695 (2012).
36. Speliotes, E.K. *et al.* Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet* **7**, e1001324 (2011).

37. Lei, X., Callaway, M., Zhou, H., Yang, Y. & Chen, W. Obesity associated Lyplal1 gene is regulated in diet induced obesity but not required for adipocyte differentiation. *Mol Cell Endocrinol* **411**, 207-13 (2015).
38. Wessel, J. *et al.* Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat Commun* **6**, 5897 (2015).
39. Consortium, C.A.D. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* (2015).
40. Yamamoto, F., Clausen, H., White, T., Marken, J. & Hakomori, S. Molecular genetic basis of the histo-blood group ABO system. *Nature* **345**, 229-33 (1990).
41. Fagherazzi, G., Gusto, G., Clavel-Chapelon, F., Balkau, B. & Bonnet, F. ABO and Rhesus blood groups and risk of type 2 diabetes: evidence from the large E3N cohort study. *Diabetologia* **58**, 519-22 (2015).
42. Gradwohl, G., Dierich, A., LeMeur, M. & Guillemot, F. neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc Natl Acad Sci U S A* **97**, 1607-11 (2000).
43. Wang, J. *et al.* Mutant neurogenin-3 in congenital malabsorptive diarrhea. *N Engl J Med* **355**, 270-80 (2006).
44. Rubio-Cabezas, O. *et al.* Permanent Neonatal Diabetes and Enteric Anendocrinosis Associated With Biallelic Mutations in NEUROG3. *Diabetes* **60**, 1349-53 (2011).
45. del Bosque-Plata, L. *et al.* Mutations in the coding region of the neurogenin 3 gene (NEUROG3) are not a common cause of maturity-onset diabetes of the young in Japanese subjects. *Diabetes* **50**, 694-6 (2001).
46. Beck, A. *et al.* An siRNA screen identifies transmembrane 7 superfamily member 3 (TM7SF3), a seven transmembrane orphan receptor, as an inhibitor of cytokine-induced death of pancreatic beta cells. *Diabetologia* **54**, 2845-55 (2011).
47. Todd, J.N. *et al.* Variation in glucose homeostasis traits associated with P2RX7 polymorphisms in mice and humans. *J Clin Endocrinol Metab* **100**, E688-96 (2015).
48. Hinke, S.A., Hellemans, K. & Schuit, F.C. Plasticity of the beta cell insulin secretory competence: preparing the pancreatic beta cell for the next meal. *J Physiol* **558**, 369-80 (2004).
49. Saxena, R. *et al.* Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet* **42**, 142-8 (2010).
50. Lyssenko, V. *et al.* Pleiotropic effects of GIP on islet function involve osteopontin. *Diabetes* **60**, 2424-33 (2011).
51. Arai, T., Kano, F. & Murata, M. Translocation of forkhead box O1 to the nuclear periphery induces histone modifications that regulate transcriptional repression of PCK1 in HepG2 cells. *Genes Cells* **20**, 340-57 (2015).
52. Cook, J.P. & Morris, A.P. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *Eur J Hum Genet* (2016).
53. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* **41**, 703-7 (2009).
54. Hakonarson, H. *et al.* A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* **448**, 591-4 (2007).
55. Gritti, M. *et al.* Metformin repositioning as antitumoral agent: selective antiproliferative effects in human glioblastoma stem cells, via inhibition of CLIC1-mediated ion current. *Oncotarget* **5**, 11252-68 (2014).

56. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* **42**, 579-89 (2010).
57. Consortium, S.T.D. *et al.* Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97-101 (2014).
58. Langenberg, C. *et al.* Gene-lifestyle interaction and type 2 diabetes: the EPIC interact case-cohort study. *PLoS Med* **11**, e1001647 (2014).
59. (CDC), C.f.D.C.a.P. & U.S. Department of Health and Human Services, A., GA. Centers for Disease Control and Prevention National diabetes statistics report: estimates of diabetes and its burden in the United States.
60. Bartoli, E., Fra, G.P. & Carnevale Schianca, G.P. The oral glucose tolerance test (OGTT) revisited. *Eur J Intern Med* **22**, 8-12 (2011).
61. Lin, D.Y. & Sullivan, P.F. Meta-analysis of genome-wide association studies with overlapping subjects. *Am J Hum Genet* **85**, 862-72 (2009).
62. Shao, C., Zucker, I.H. & Gao, L. Angiotensin type 2 receptor in pancreatic islets of adult rats: a novel insulinotropic mediator. *Am J Physiol Endocrinol Metab* **305**, E1281-91 (2013).
63. Yvan-Charvet, L. *et al.* Deletion of the angiotensin type 2 receptor (AT2R) reduces adipose cell size and protects from diet-induced obesity and insulin resistance. *Diabetes* **54**, 991-9 (2005).
64. Liu, M., Jing, D., Wang, Y., Liu, Y. & Yin, S. Overexpression of angiotensin II type 2 receptor promotes apoptosis and impairs insulin secretion in rat insulinoma cells. *Mol Cell Biochem* **400**, 233-44 (2015).
65. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
66. Manousaki, D. *et al.* Toward Precision Medicine: TBC1D4 Disruption Is Common Among the Inuit and Leads to Underdiagnosis of Type 2 Diabetes. *Diabetes Care* **39**, 1889-1895 (2016).
67. Frantz, E.D., Crespo-Mascarenhas, C., Barreto-Vianna, A.R., Aguila, M.B. & Mandarim-de-Lacerda, C.A. Renin-angiotensin system blockers protect pancreatic islets against diet-induced obesity and insulin resistance in mice. *PLoS One* **8**, e67192 (2013).
68. Leung, P.S. Mechanisms of protective effects induced by blockade of the renin-angiotensin system: novel role of the pancreatic islet angiotensin-generating system in Type 2 diabetes. *Diabet Med* **24**, 110-6 (2007).
69. Geng, D.F., Jin, D.M., Wu, W., Liang, Y.D. & Wang, J.F. Angiotensin converting enzyme inhibitors for prevention of new-onset type 2 diabetes mellitus: a meta-analysis of 72,128 patients. *Int J Cardiol* **167**, 2605-10 (2013).
70. Investigators, D.T. *et al.* Effect of ramipril on the incidence of diabetes. *N Engl J Med* **355**, 1551-62 (2006).
71. Flannick, J. & Florez, J.C. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat Rev Genet* **17**, 535-49 (2016).
72. The ups and downs of data sharing in science. *Nature* **534**, 435-6 (2016).

Table 1. Novel T2D associated *loci*.

Novel *Locus*	Chr	rsID – Risk Allele	OR (95% CI) P-value			MAF
			Stage1 Discovery meta-analysis	Stage2 Replication meta-analysis	Stage1 + Stage2 Combined meta-analysis	
LYPLAL1/ZC3H11B (1q41)	1	rs2820443-T	1.08 (1.04-1.13) 2.94×10^{-4} †	1.06 (1.03-1.09) 2.10×10^{-5} §	1.07 (1.04-1.09) 2.56×10^{-8}*	0.28
EHMT2 (6p21.33-p21.32)	6	rs115884658-A	1.34 (1.18-1.53) 1.00×10^{-5} †	1.17 (1.09-1.26) 2.90×10^{-6} ‡*	1.21 (1.14-1.29) 3.00×10^{-10}*	0.02
ABO (9q34.2)	9	rs505922-C	1.07 (1.03-1.11) 6.93×10^{-4} †	1.06 (1.03-1.09) 1.90×10^{-5} §	1.06 (1.04-1.09) 4.94×10^{-8}*	0.34
NEUROG3 (10q22.1)	10	rs2642587-G	1.12 (1.08-1.16) 8.45×10^{-9} ¢	-	-	0.22
CAMKK2 (12q24.31)	12	rs3794205-G	1.09 (1.05-1.14) 4.18×10^{-5} †	1.06 (1.03-1.09) 1.60×10^{-4} §	1.07 (1.04-1.10) 4.11×10^{-8}*	0.32
CALCOCO2/ATP5G1/ UBE2Z/SNF8/GIP (17q21.32)	17	rs12453394-A	1.08 (1.04-1.12) 7.86×10^{-5} †	1.07 (1.03-1.11) 9.60×10^{-5} §	1.07 (1.05-1.10) 3.23×10^{-8}*	0.47
AGTR2 (Xq23)	X	rs146662075-T	3.09 (2.06-4.60) 3.24×10^{-8} ∆	1.82 (0.92-3.61) 0.09 Ø	2.70 (1.91-3.81) 1.73×10^{-8}	0.008

Abbreviations are as follows: Chr, Chromosome; OR, odds ratio; MAF, minor allele frequency.

† Imputed based public GWAS discovery meta-analysis (NuGENE + GERA cohort, 7,522 cases and 50,446 controls)
¢ Full imputed based public GWAS meta-analysis (NuGENE + GERA cohort + GENEVA+FUSION+WTCCC, 12,931 cases and 57,196 controls)
∆ 70KforT2D Men Cohort (GERA cohort + GENEVA + FUSION, 5,277 cases and 15,702 controls older than 55 years old)
‡ T2D Diabetes Genetic Portal (Exome-Chip+Exome Sequencing, 35,789 cases and 56,738 controls)
§ Transancestry DIAGRAM Consortium (26,488 cases and 83,964 controls)
Ø Replication Men Cohort SIGMA UK10K imputation + InterAct + Danish Cohort (case-control and follow-up) (9,259 cases and 4,664 controls older than 55 years old and OGTT>7.8 mmol/l, when available).
*Meta P-value Estimated using a weighted Z-score method due to unavailable SE information from the Stage 2 replication cohorts

Figure Legends:

Figure 1: Discovery and replication strategy. Publicly available GWAS datasets representing a total of 12,931 cases and 57,196 controls (70KforT2D) were first quality controlled, phased and imputed using 1000G and UK10K separately. For those variants that were present in the DIAGRAM trans-ethnic meta-analysis we used the summary statistics to meta-analyze our results with the cohorts that had no overlap with any of the cohorts included in the DIAGRAM trans-ethnic meta-analysis. With this first meta-analysis we discovered four novel *loci* (within magenta panels). For the rest of the variants we meta-analyzed all the 70KforT2D datasets, which resulted in two novel *loci* (in blue panels). All the variants that were coding and that showed a p-value $\leq 1x10^{-4}$ were tested for replication by interrogating the summary statistics in the Type 2 Diabetes Knowledge Portal (T2D Portal) (http://www.type2diabetesgenetics.org/). This resulted in a novel low-frequency variant in the *EHMT2* gene (highlighted with a green panel).

Figure 2: Manhattan and Quantile-Quantile plot (QQ-plot) of the discovery and replication genome-wide meta-analysis. The upper corner represents the QQ-plot. Expected $-\log_{10}$ p-values under the null hypothesis are represented in the x-axis while observed $-\log_{10}$ p-values are represented in the y-axis. Observed p-values were obtained according to the suitable replication dataset used (as shown in Figure 1) and were depicted using different colors. HapMap variants were meta-analyzed using the Trans-Ethnic summary statistics from the DIAGRAM study and our meta-analysis based on the Genetic Epidemiology Research on Aging (GERA) cohort and the Northwestern NuGENE Project and that resulted in novel associations depicted in magenta. The rest of non-HapMap variants meta-analyzed using the full 70KforT2D cohort are represented in grey, highlighting in light blue the fraction of novel GWAS-significant variants. Coding low-frequency variants meta-analyzed using the 70KforT2D and the T2D Portal data that resulted in novel GWAS-significant associations are depicted in green. The shaded area of the QQ-plot indicates the 95% confidence interval under the null and a density function of the distribution of the p-values was plotted using a dashed line. The λ is a measure of the genomic inflation and corresponds to the observed median $\chi 2$ test statistic divided by the median expected χ^2 test statistic under the null hypothesis. The Manhattan plot, representing the $-\log_{10}$ p-values were colored as explained in the QQ-plot. All known GWAS-significant associated variants within known T2D genes are also depicted in red. X-chromosome results for females (F), males (M) and all individuals (A) are also included.

Figure 3: Discovery and replication of rs14666075 association signal. Forest plot for the rs146662075 variant using (A) the discovery cohorts and when including also (B) the replication datasets. Cohort-specific odds ratios are denoted by boxes proportional to the size of the cohort and 95% CI error bars. The combined OR estimate for all the datasets is represented by a green diamond, where the diamond width corresponds to 95% CI bounds. The p-value for the meta-analysis (Meta P) and for the heterogeneity (Het P) of odds ratio is shown. C) Plot showing the cumulative incidence of type 2 diabetes for an 11 years follow-up. The red line represents the T carriers and in light blue, C carriers are represented (n=1,652, cases=158.). D) Forest plot after excluding controls younger than 55 years old and OGTT>7.8 mmol/l in both the discovery and replication cohorts when available.

Figure 4: Functional characterization of rs146662075 association signal. A) Signal plot for X chromosome region surrounding rs146662075. Each point represents a variant, with its *p*-value (on a –log10 scale, y-axis) derived from the meta-analysis results from association

testing in males. The x-axis represents the genomic position (Hg19). Below, representation of H3K27ac and RNA-seq in a subset of cell-types for which RNA-seq and H3K27ac was available is also shown. The association between RNA-seq signals and H3K27ac marks suggest that *AGTR2* is the most likely regulated gene by the enhancer that harbors rs146662075. B) The presence of the common allelic variant rs146662075-C reduces enhancer activity in luciferase assays performed in a mouse myoblast cell line. C) Electrophoretic mobility shift assay in C2C12 myoblast cell lines, C2C12 differentiated myotubes and human fetal myoblasts showed allele-specific binding of a ubiquitous nuclear complex. The arrows indicate the allele-specific binding event. Competition was carried out using 50- and 100- fold excess of corresponding unlabeled probe.

Figure 1

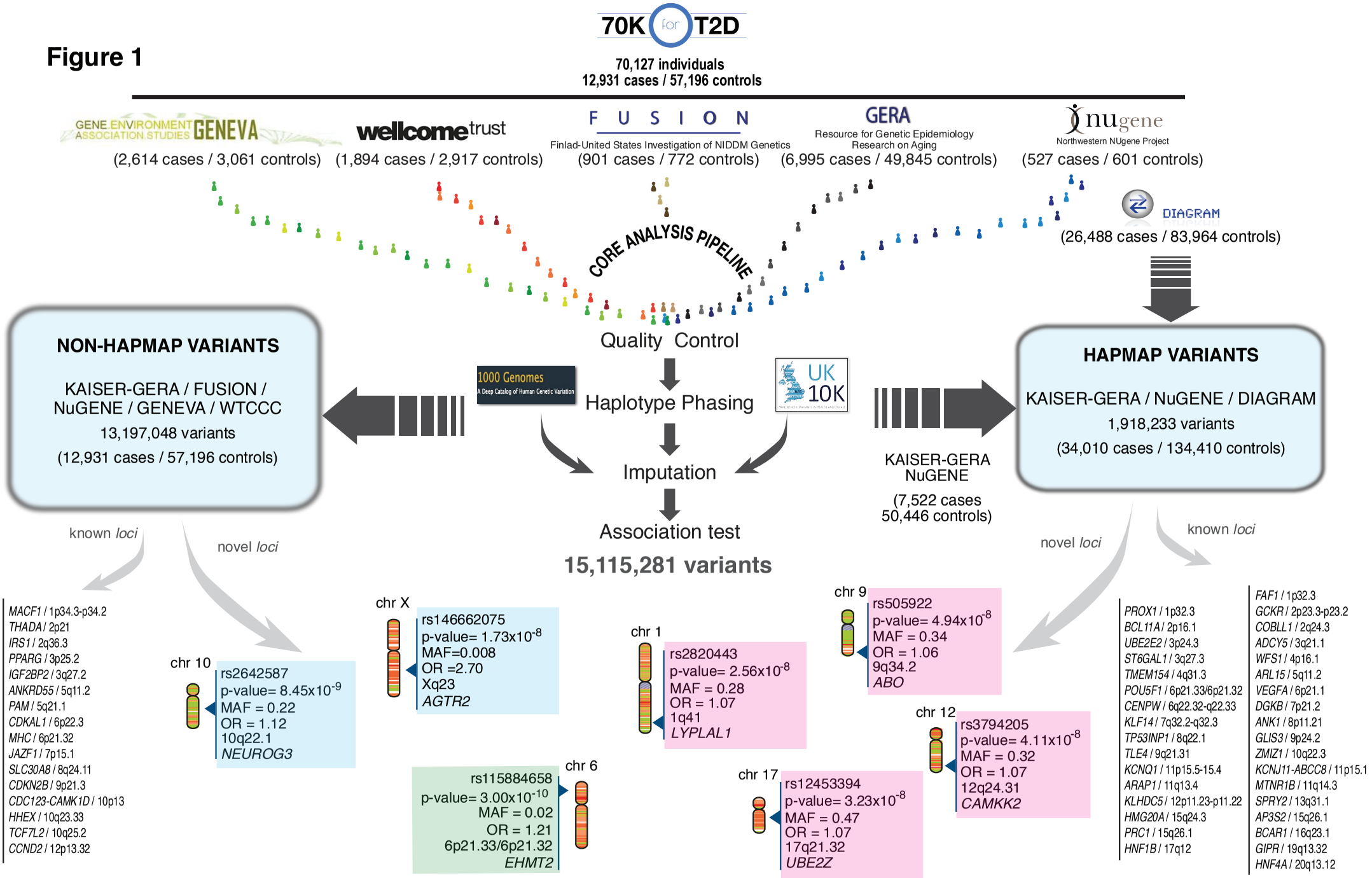

70K for T2D
70,127 individuals
12,931 cases / 57,196 controls
GENE ENVIRONMENT ASSOCIATION STUDIES GENEVA
(2,614 cases / 3,061 controls)
wellcome trust
(1,894 cases / 2,917 controls)
FUSION
Finlad-United States Investigation of NIDDM Genetics
(901 cases / 772 controls)
GERA
Resource for Genetic Epidemiology Research on Aging
(6,995 cases / 49,845 controls)
nugene
Northwestern NUgene Project
(527 cases / 601 controls)
DIAGRAM
(26,488 cases / 83,964 controls)
CORE ANALYSIS PIPELINE
Quality Control
Haplotype Phasing
Imputation
Association test
15,115,281 variants
1000 Genomes
A Deep Catalog of Human Genetic Variation
UK 10K
NON-HAPMAP VARIANTS
KAISER-GERA / FUSION / NuGENE / GENEVA / WTCCC
13,197,048 variants
(12,931 cases / 57,196 controls)
HAPMAP VARIANTS
KAISER-GERA / NuGENE / DIAGRAM
1,918,233 variants
(34,010 cases / 134,410 controls)
KAISER-GERA
NuGENE
(7,522 cases
50,446 controls)
known loci
novel loci
novel loci
known loci
MACF1 / 1p34.3-p34.2
THADA / 2p21
IRS1 / 2q36.3
PPARG / 3p25.2
IGF2BP2 / 3q27.2
ANKRD55 / 5q11.2
PAM / 5q21.1
CDKAL1 / 6p22.3
MHC / 6p21.32
JAZF1 / 7p15.1
SLC30A8 / 8q24.11
CDKN2B / 9p21.3
CDC123-CAMK1D / 10p13
HHEX / 10q23.33
TCF7L2 / 10q25.2
CCND2 / 12p13.32
chr 10
rs2642587
p-value= 8.45x10-9
MAF = 0.22
OR = 1.12
10q22.1
NEUROG3
chr X
rs146662075
p-value= 1.73x10-8
MAF=0.008
OR =2.70
Xq23
AGTR2
rs115884658
p-value= 3.00x10-10
MAF = 0.02
OR = 1.21
6p21.33/6p21.32
EHMT2
chr 6
chr 1
rs2820443
p-value= 2.56x10-8
MAF = 0.28
OR = 1.07
1q41
LYPLAL1
chr 17
rs12453394
p-value= 3.23x10-8
MAF = 0.47
OR = 1.07
17q21.32
UBE2Z
chr 9
rs505922
p-value= 4.94x10-8
MAF = 0.34
OR = 1.06
9q34.2
ABO
chr 12
rs3794205
p-value= 4.11x10-8
MAF = 0.32
OR = 1.07
12q24.31
CAMKK2
PROX1 / 1p32.3
BCL11A / 2p16.1
UBE2E2 / 3p24.3
ST6GAL1 / 3q27.3
TMEM154 / 4q31.3
POU5F1 / 6p21.33/6p21.32
CENPW / 6q22.32-q22.33
KLF14 / 7q32.2-q32.3
TP53INP1 / 8q22.1
TLE4 / 9q21.31
KCNQ1 / 11p15.5-15.4
ARAP1 / 11q13.4
KLHDC5 / 12p11.23-p11.22
HMG20A / 15q24.3
PRC1 / 15q26.1
HNF1B / 17q12
FAF1 / 1p32.3
GCKR / 2p23.3-p23.2
COBLL1 / 2q24.3
ADCY5 / 3q21.1
WFS1 / 4p16.1
ARL15 / 5q11.2
VEGFA / 6p21.1
DGKB / 7p21.2
ANK1 / 8p11.21
GLIS3 / 9p24.2
ZMIZ1 / 10q22.3
KCNJ11-ABCC8 / 11p15.1
MTNR1B / 11q14.3
SPRY2 / 13q31.1
AP3S2 / 15q26.1
BCAR1 / 16q23.1
GIPR / 19q13.32
HNF4A / 20q13.12

Figure 2

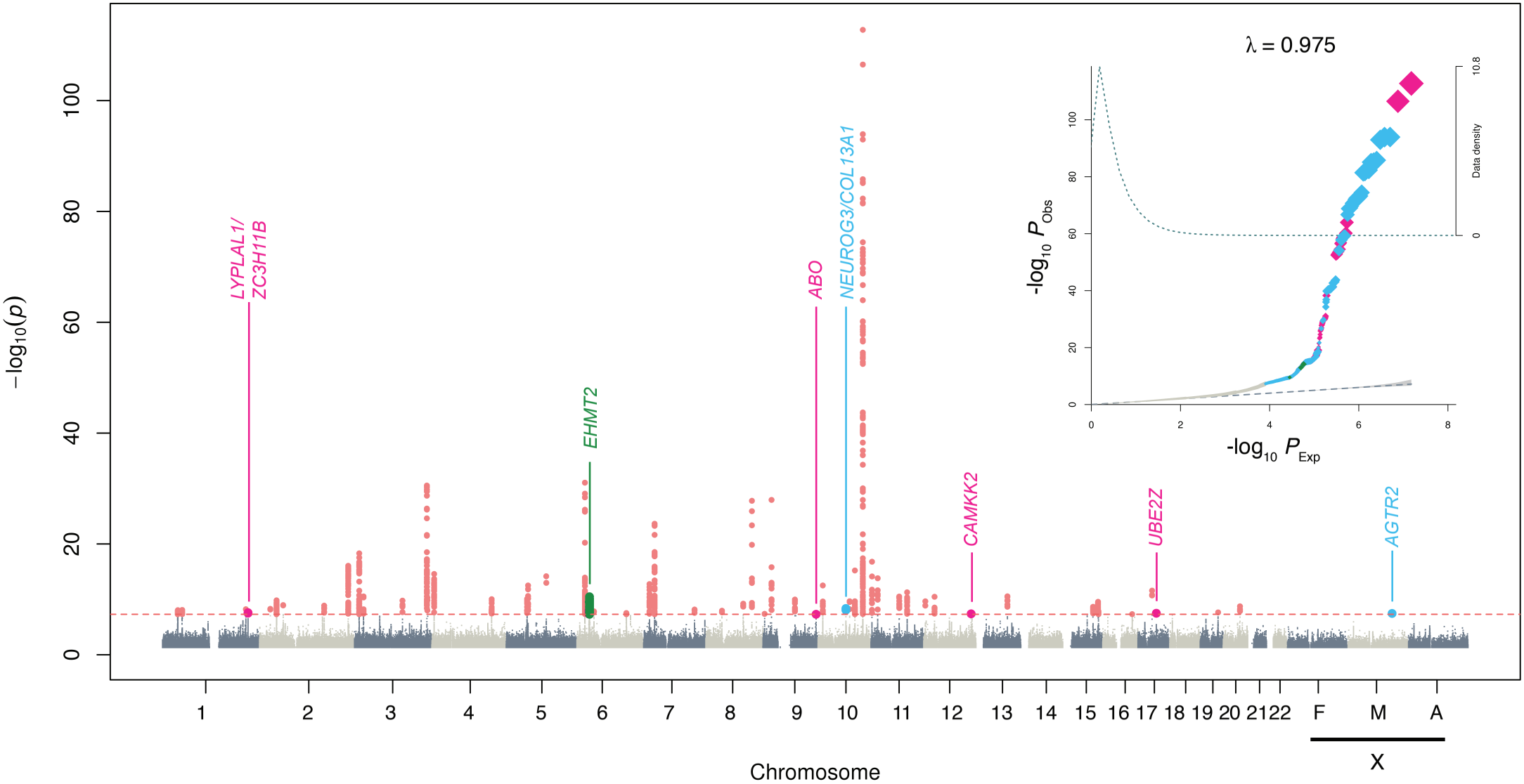

λ = 0.975
LYPLAL1/
ZC3H11B
EHMT2
ABO
NEUROG3/COL13A1
CAMKK2
UBE2Z
AGTR2
−log10(p)
Chromosome
X
-log10 PObs
-log10 PExp
Data density

Figure 3

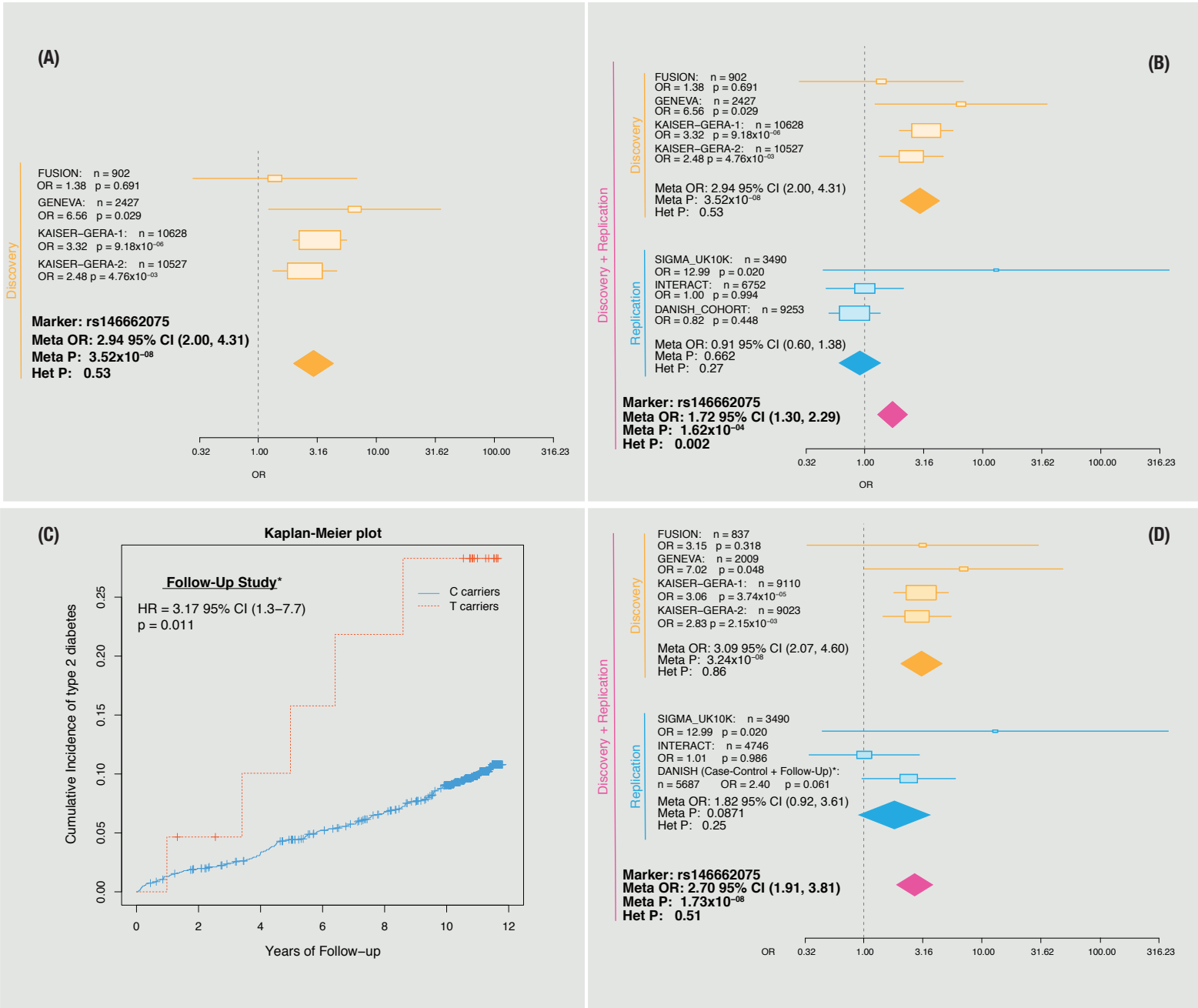
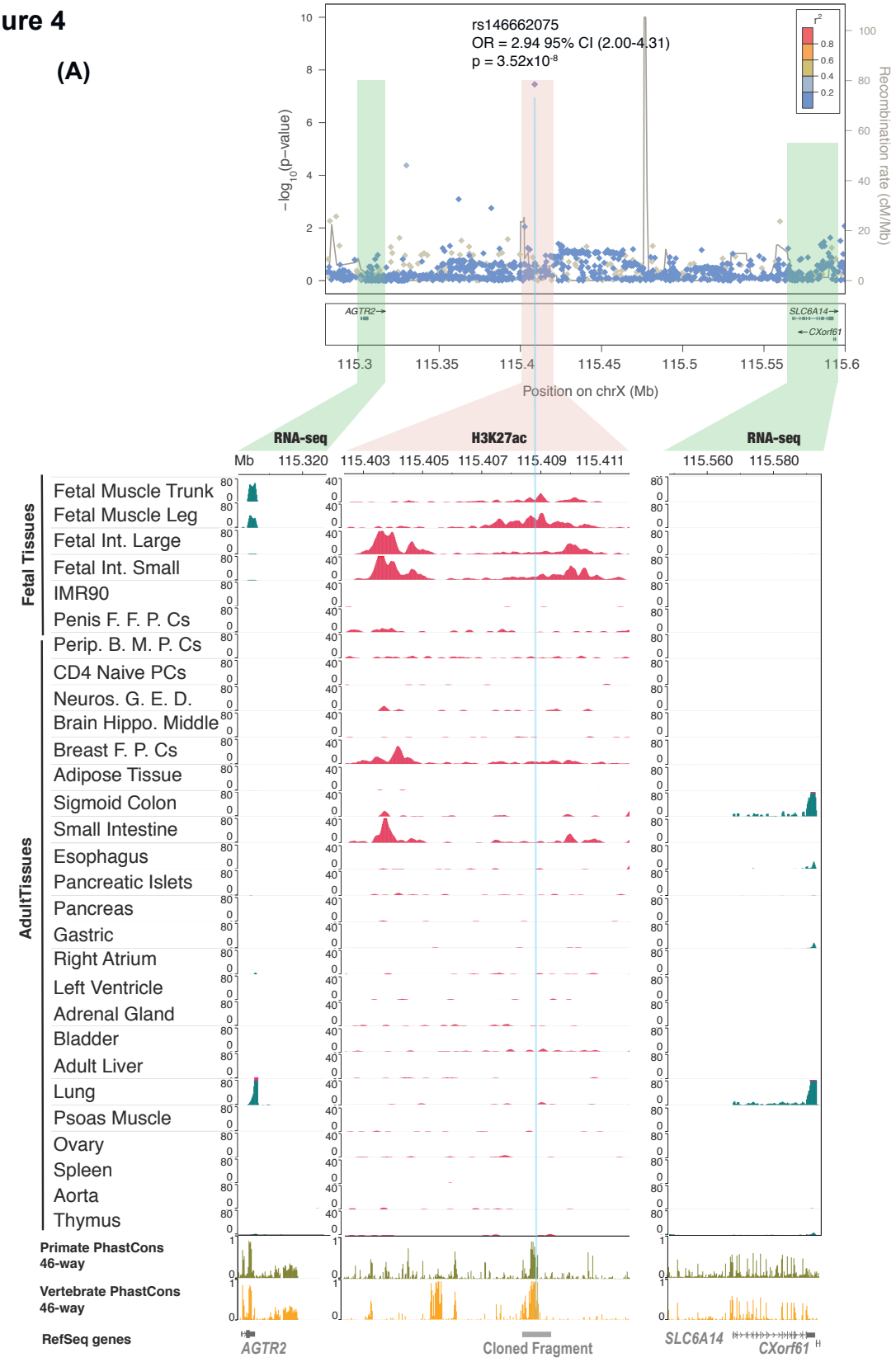
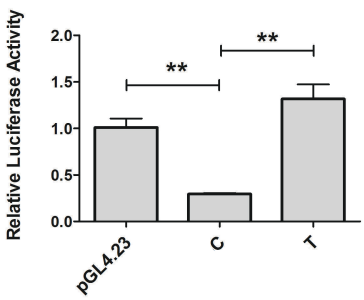

(A)
Discovery
FUSION: n = 902
OR = 1.38 p = 0.691
GENEVA: n = 2427
OR = 6.56 p = 0.029
KAISER–GERA-1: n = 10628
OR = 3.32 p = 9.18x10-06
KAISER–GERA-2: n = 10527
OR = 2.48 p = 4.76x10-03
Marker: rs146662075
Meta OR: 2.94 95% CI (2.00, 4.31)
Meta P: 3.52x10-08
Het P: 0.53
0.32
1.00
3.16
10.00
31.62
100.00
316.23
OR
(B)
Discovery + Replication
Discovery
FUSION: n = 902
OR = 1.38 p = 0.691
GENEVA: n = 2427
OR = 6.56 p = 0.029
KAISER–GERA-1: n = 10628
OR = 3.32 p = 9.18x10-06
KAISER–GERA-2: n = 10527
OR = 2.48 p = 4.76x10-03
Meta OR: 2.94 95% CI (2.00, 4.31)
Meta P: 3.52x10-08
Het P: 0.53
Replication
SIGMA_UK10K: n = 3490
OR = 12.99 p = 0.020
INTERACT: n = 6752
OR = 1.00 p = 0.994
DANISH_COHORT: n = 9253
OR = 0.82 p = 0.448
Meta OR: 0.91 95% CI (0.60, 1.38)
Meta P: 0.662
Het P: 0.27
Marker: rs146662075
Meta OR: 1.72 95% CI (1.30, 2.29)
Meta P: 1.62x10-04
Het P: 0.002
0.32
1.00
3.16
10.00
31.62
100.00
316.23
OR
(C)
Kaplan-Meier plot
Follow-Up Study*
HR = 3.17 95% CI (1.3–7.7)
p = 0.011
C carriers
T carriers
Cumulative Incidence of type 2 diabetes
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
12
Years of Follow–up
(D)
Discovery + Replication
Discovery
FUSION: n = 837
OR = 3.15 p = 0.318
GENEVA: n = 2009
OR = 7.02 p = 0.048
KAISER–GERA-1: n = 9110
OR = 3.06 p = 3.74x10-05
KAISER–GERA-2: n = 9023
OR = 2.83 p = 2.15x10-03
Meta OR: 3.09 95% CI (2.07, 4.60)
Meta P: 3.24x10-08
Het P: 0.86
Replication
SIGMA_UK10K: n = 3490
OR = 12.99 p = 0.020
INTERACT: n = 4746
OR = 1.01 p = 0.986
DANISH (Case-Control + Follow-Up)*:
n = 5687 OR = 2.40 p = 0.061
Meta OR: 1.82 95% CI (0.92, 3.61)
Meta P: 0.0871
Het P: 0.25
Marker: rs146662075
Meta OR: 2.70 95% CI (1.91, 3.81)
Meta P: 1.73x10-08
Het P: 0.51
OR
0.32
1.00
3.16
10.00
31.62
100.00
316.23

Figure 4

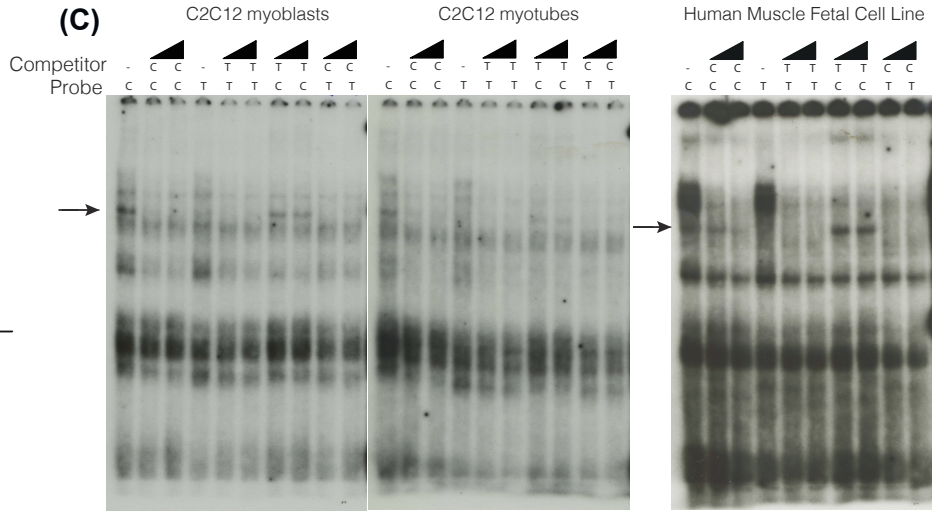
(A)


rs146662075
OR = 2.94 95% CI (2.00-4.31)
p = 3.52x10-8
−log10(p−value)
Recombination rate (cM/Mb)
r2
AGTR2
SLC6A14
CXorf61
Position on chrX (Mb)
RNA-seq
H3K27ac
RNA-seq
Mb
Fetal Tissues
Fetal Muscle Trunk
Fetal Muscle Leg
Fetal Int. Large
Fetal Int. Small
IMR90
Penis F. F. P. Cs
AdultTissues
Perip. B. M. P. Cs
CD4 Naive PCs
Neuros. G. E. D.
Brain Hippo. Middle
Breast F. P. Cs
Adipose Tissue
Sigmoid Colon
Small Intestine
Esophagus
Pancreatic Islets
Pancreas
Gastric
Right Atrium
Left Ventricle
Adrenal Gland
Bladder
Adult Liver
Lung
Psoas Muscle
Ovary
Spleen
Aorta
Thymus
Primate PhastCons 46-way
Vertebrate PhastCons 46-way
RefSeq genes
AGTR2
Cloned Fragment
SLC6A14
CXorf61

(B)


Relative Luciferase Activity
pGL4.23
C
T
**
**

(C)


C2C12 myoblasts
C2C12 myotubes
Human Muscle Fetal Cell Line
Competitor
Probe