Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes

Bonàs-Guarch, Silvia; Guindo-Martínez, Marta; Miguel-Escalada, Irene; Grarup, Niels; Sebastian, David; Rodriguez-Fos, Elias; Sánchez, Friman; Planas-Fèlix, Mercè; Cortes-Sánchez, Paula; González, Santi; Timshel, Pascal; Pers, Tune H; Morgan, Claire C; Moran, Ignasi; Atla, Goutham; González, Juan R; Puiggros, Montserrat; Martí, Jonathan; Andersson, Ehim A; Díaz, Carlos; Badia, Rosa M; Udler, Miriam; Leong, Aaron; Kaur, Varindepal; Flannick, Jason; Jørgensen, Torben; Linneberg, Allan; Jørgensen, Marit E; Witte, Daniel R; Christensen, Cramer; Brandslund, Ivan; Appel, Emil V; Scott, Robert A; Luan, Jian'an; Langenberg, Claudia; Wareham, Nicholas J; Pedersen, Oluf; Zorzano, Antonio; Florez, Jose C; Hansen, Torben; Ferrer, Jorge; Mercader, Josep Maria; Torrents, David

Published in:
Nature Communications

DOI:
10.1038/s41467-017-02380-9

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

Document license
CC BY

Citation for published version (APA):
Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes

Sílvia Bonàs-Guarch et al.

The reanalysis of existing GWAS data represents a powerful and cost-effective opportunity to gain insights into the genetics of complex diseases. By reanalyzing publicly available type 2 diabetes (T2D) genome-wide association studies (GWAS) data for 70,127 subjects, we identify seven novel associated regions, five driven by common variants (LYPLAL1, NEUROG3, CAMKK2, ABO, and GIP genes), one by a low-frequency (EHMT2), and one driven by a rare variant in chromosome Xq23, rs146662057, associated with a twofold increased risk for T2D in males. rs146662057 is located within an active enhancer associated with the expression of Angiotensin II Receptor type 2 gene (AGTR2), a modulator of insulin sensitivity, and exhibits allelic specific activity in muscle cells. Beyond providing insights into the genetics and pathophysiology of T2D, these results also underscore the value of reanalyzing publicly available data using novel genetic resources and analytical approaches.
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. However, the majority of the published GWAS have only reported primary findings, which generally explain a small fraction of the estimated heritability. To examine the missing heritability, most strategies involve generating new genetic and clinical data. Very rarely are new studies based on the revision and reanalysis of existing genetic data by applying more powerful analytic techniques and resources after the primary GWAS findings are published. These cost-effective reanalysis strategies are now possible, given emerging (1) data-sharing initiatives with large amounts of primary genetic data for multiple human genetic diseases, as well as (2) new and improved GWAS methodologies and resources. Notably, genotype imputation with novel sequence-based reference panels can now substantially increase the genetic resolution of GWAS from previously genotyped data sets, reaching good-quality imputation of low frequency (minor allele frequency [MAF] \(0.01 \leq \text{MAF} < 0.05\)) and rare variants (MAF \(\leq 0.01\)), increasing the power to identify novel associations, and fine map the known ones. Moreover, the availability of publicly available primary genetic data allows the homogeneous integration of multiple data sets from different origins providing more accurate meta-analysis results, particularly at the low ranges of allele frequency. Finally, the vast majority of reported GWAS analyses omits the X chromosome, despite representing 5% of the genome and coding for more than 1,500 genes. The reanalysis of publicly available data also enables interrogation of this chromosome.

We hypothesized that a unified reanalysis of multiple publicly available data sets, applying homogeneous standardized quality control (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 hypothesis, we focused this study on type 2 diabetes (T2D), one of the most prevalent complex diseases for which many GWAS have been performed during the past decade. These studies have allowed the identification of more than 100 independent loci, most of them driven by common variants, with a few exceptions. Despite these efforts, there is still a large fraction of genetic heritability hidden in the data, and the role of low-frequency variants, although recently proposed to be minor, has still not been fully explored. The availability of large T2D genetic data sets in combination with larger and more comprehensive genetic variation reference panels, provides the opportunity to impute a significantly increased fraction of low-frequency and rare variants, and to study their contribution to the risk of developing this disease. This strategy also allows us 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 covering the whole genome (including the X chromosome). We then performed imputation using 1,000 Genomes Project (1000G) and UK10K reference panels, followed by 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, allowing us to describe a regulatory mechanism disrupted by a novel rare and large-effect variant identified at the X chromosome.

Results

Overall analysis strategy. As shown in Fig. 1, we first gathered all T2D case-control GWAS individual-level data that were available through the EGA and dbGaP databases (i.e., Gene Environment-Association Studies [GENEVA], Wellcome Trust Case Control Consortium [WTCCC], Finland–United States Investigation of NIDDM Genetics [FUSION], Resource for Genetic Epidemiology Research on Aging [GERA], and Northwestern NuGENE project [NuGENE]). We harmonized these cohorts, applied standardized quality control procedures, and filtered out low-quality variants and samples (Methods and Supplementary Notes). After this process, a total of 70,127 subjects (70KforT2D, 12,931 cases, and 57,196 controls, Supplementary Data 1) were retained for downstream analysis. Each of these cohorts was then imputed to the 1000G and UK10K reference panels using an integrative method, which selected the results from the reference panel that provided the highest accuracy for each variant, according to IMPUTE2 info score (Methods). Finally, the results from each of these cohorts were meta-analyzed (Fig. 1), obtaining a total of 15,115,281 variants with good imputation quality (IMPUTE2 info score \(\geq 0.7\), MAF \(\geq 0.001\), and \(F^2\) heterogeneity score \(\leq 0.75\)), across 12,931 T2D cases and 57,196 controls. Of these, 6,845,408 variants were common (MAF \(\geq 0.05\)), 3,100,848 were low-frequency (0.01 \(\leq \text{MAF} < 0.05\)), and 5,169,025 rare variants were \(\leq 0.01\), 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 (Fig. 1, Supplementary Data 1). Second, the remaining variants (13,197,048), which included mainly non-HapMap variants (MAF \(< 0.05\)) or variants not tested above, were meta-analyzed using all five cohorts from the 70KforT2D resource (Supplementary Data 1). Finally, low-frequency variants located in coding regions and with \(p \leq 1 \times 10^{-4}\) were meta-analyzed using the non-overlapping fraction of samples with the data from the T2D Portal through the interrogation of exome array and whole-exome sequence data from ~80,000 and ~17,000 individuals, respectively.

Pathway and functional enrichment analysis. To explore whether our results recapitulate the pathophysiology of T2D, we performed gene-set enrichment analysis with all the variants with \(p \leq 1 \times 10^{-5}\) using DEPICT (Methods). This analysis showed enrichment of genes expressed in pancreas (ranked first in tissue enrichment analysis, \(p = 7.8 \times 10^{-4}\), FDR \(< 0.05\), Supplementary Data 2) and cellular response to insulin stimulus (ranked second in gene-set enrichment analysis, \(p = 3.9 \times 10^{-8}\), FDR = 0.05,
The high coverage of genetic variation ascertained in this study allowed us to fine-map known and novel loci, providing more candidate causal variants for downstream functional interpretations. We constructed 99% credible variant sets for each of the 57 loci (Supplementary Data 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 Data 6). Interestingly, by integrating imputed results from 1,000G and UK10K reference panels, we gained up to 41% of indels, which were only identified by either 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 that we replicated (p ≤ 5.3 × 10^{-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 previously reported locus for T2D, we highlighted the advantage of integrating the results from both reference panels.

Three association strategies allowed us to identify 57 genome-wide significant associated loci (p ≤ 5 × 10^{-8}), of which seven were not previously reported as associated with T2D (Table 1). The remaining 50 loci have been previously reported and included, for example, two low-frequency variants recently discovered in east and south Asian populations (Supplementary Data 4).

### Supplementary Data

- **Supplementary Data 3**: Discovery and replication strategy. Publicly available GWAS datasets representing a total of 12,931 cases and 57,196 controls (70KforT2D) were used to discover 57 genome-wide significant associated loci, of which seven were not previously reported as associated with T2D (Table 1). The remaining 50 loci have been previously reported and included, for example, two low-frequency variants recently discovered in east and south Asian populations (Supplementary Data 4).

- **Supplementary Data 2**: Three association strategies allowed us to identify 57 genome-wide significant associated loci (p ≤ 5 × 10^{-8}), of which seven were not previously reported as associated with T2D (Table 1). The remaining 50 loci have been previously reported and included, for example, two low-frequency variants recently discovered in east and south Asian populations (Supplementary Data 4).

- **Supplementary Data 4**: The high coverage of genetic variation ascertained in this study allowed us to fine-map known and novel loci, providing more candidate causal variants for downstream functional interpretations. We constructed 99% credible variant sets for each of the 57 loci (Supplementary Data 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 Data 6). Interestingly, by integrating imputed results from 1,000G and UK10K reference panels, we gained up to 41% of indels, which were only identified by either 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 that we replicated (p ≤ 5.3 × 10^{-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, we found that 12 of the 57 variants within its 99% credible set correspond to indels with genome-wide significance (5.6 × 10^{-16} < p < 2.4 × 10^{-15}), which collectively represented 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. In addition, we tested the effect of all variants on expression across multiple tissues by interrogating GTEx and RNA-sequencing gene expression data. A protective effect for T2D in our study (rs8176719, \( p = 3.4 \times 10^{-4}, \text{OR} = 0.95 [0.91–0.98] \)). In addition, several variants within this credible set are associated with the expression of the ABO gene in multiple tissues including skeletal muscle, adipose tissue, and pancreatic islets (Supplementary Data 8 and 9).

### Novel T2D-associated loci driven by common variants

Beyond the detailed characterization of the known T2D-associated regions, we also identified seven novel loci, among which, one was driven by common variants with modest effect sizes (1.06 < OR < 1.12; Table 1, Supplementary Fig. 6 and 7).

<table>
<thead>
<tr>
<th>Novel Locus</th>
<th>Chr</th>
<th>rsID---Risk Allele</th>
<th>Stage1 Discovery Meta-analysis</th>
<th>Stage2 Replication Meta-analysis</th>
<th>Stage1 + Stage2 Combined Meta-analysis</th>
<th>MAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>LYPAL1/ZC3H11B (1q41)</td>
<td>1</td>
<td>rs2820443-T</td>
<td>1.08 (1.04–1.13)</td>
<td>2.10 × 10^{-5} b</td>
<td>1.07 (1.04–1.09)</td>
<td>0.28</td>
</tr>
<tr>
<td>EHM72 (6p21.33–p21.32)</td>
<td>6</td>
<td>rs115884658-A</td>
<td>1.34 (1.18–1.53)</td>
<td>1.17 (1.09–1.26)</td>
<td>1.21 (1.14–1.29)</td>
<td>0.02</td>
</tr>
<tr>
<td>ABO (9q34.2)</td>
<td>9</td>
<td>rs505922-C</td>
<td>1.07 (1.03–1.11)</td>
<td>1.06 (1.03–1.09)</td>
<td>1.06 (1.04–1.09)</td>
<td>0.34</td>
</tr>
<tr>
<td>NEUROG3 (10q22.1)</td>
<td>10</td>
<td>rs2642587-G</td>
<td>1.12 (1.08–1.16)</td>
<td>1.06 (1.03–1.09)</td>
<td>1.07 (1.04–1.10)</td>
<td>0.32</td>
</tr>
<tr>
<td>CAMKK2 (12q24.31)</td>
<td>12</td>
<td>rs3794205-G</td>
<td>1.09 (1.05–1.14)</td>
<td>1.06 (1.03–1.09)</td>
<td>1.07 (1.04–1.10)</td>
<td>0.47</td>
</tr>
<tr>
<td>CALCOCO2/ATPSG1/UBE2Z/SNF8/GIP (17q21.32)</td>
<td>17</td>
<td>rs12453394-A</td>
<td>1.08 (1.04–1.12)</td>
<td>1.07 (1.03–1.11)</td>
<td>1.07 (1.05–1.10)</td>
<td>0.47</td>
</tr>
<tr>
<td>AGTR2 (Xq23)</td>
<td>X</td>
<td>rs146662075-T</td>
<td>3.09 (2.06–4.60)</td>
<td>1.57 (1.19–2.07)</td>
<td>1.95 (1.56–2.45)</td>
<td>0.008</td>
</tr>
</tbody>
</table>

The lead common variant of the fourth novel locus at chromosome 17q24.31 (rs3794205, OR = 1.07 [1.04–1.10], \( p = 4.1 \times 10^{-8} \)) lies within an intron of the CAMKK2 gene, previously implicated in cytokine-induced beta-cell death. However, other variants within the corresponding credible set could also be causal, such as a missense variant within the P2RX7 gene. The lead variant (rs11065504, \( p = 8.4 \times 10^{-10} \)) includes NEUROG3 (Neurogenin3), which is an essential regulator of pancreatic endocrine cell differentiation. Mutations in this gene have been reported to cause permanent neonatal diabetes, but a role of this gene in T2D has not been yet reported.

A new T2D signal driven by a low-frequency variant. Furthermore, we selected all low-frequency (0.01 < MAF < 0.05) variants with \( p \leq 1 \times 10^{-4} \) in the 70kforT2D meta-analysis that...
were annotated as altering protein sequences, 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 after excluding the overlapping cohorts between the different data sets. This analysis highlighted a novel genome-wide association driven by a low-frequency missense variant (Ser58Phe) within the EHM2 gene at chromosome 6p21.33 (rs115884658, OR = 1.21 [1.14−1.29], p = 3.00 × 10^{-10}; Fig. 2, Supplementary Figures 6 and 7). EHM2 is involved in the mediation of FOXO1 translocation induced by insulin. Since this variant is less than 1 Mb away from HLA-DQA1, a locus reported to be associated with T2D, we performed a series of reciprocal conditional analyses and excluded the possibility that our analysis was capturing previously reported T2D or T1D signals (Supplementary Data 11). Beyond this missense EHM2 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 (multitissue meta-analysis p = 8.9 × 10^{-16}, Supplementary Data 9). In addition, this same variant is associated with the expression of the first and second exon of the CLIC1 mRNA in pancreatic islet donors (p(exon 1) = 1.4 × 10^{-19}, p(exon 2) = 1.9 × 10^{-13}, Supplementary Data 10). Interestingly, CLIC1 has been reported as a direct target of metformin by mediating the antiproliferative effect of this drug in human glioblastoma. All these findings support CLIC1 as an additional possible effector transcript, likely driven by rs115333512.

A novel rare X chromosome variant associated with T2D. Similar to 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. To fill this gap, we tested 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, and then meta-analyzed. 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 liver (rs61503151 [GCCA/G], 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 on 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 a larger effect size (OR = 1.33, p = 0.003), placing this deletion, as a more likely candidate causal variant for this locus (Supplementary Data 14).

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^{-5}; Fig. 3a). Two other variants in LD with the top variant, rs139246371 (chrX:115329804, OR = 1.65, p = 3.5 × 10^{-5}, r^2 =
0.37 with the top variant) and rs6063744 (chrX:115823966, OR = 1.28, \( p = 1.7 \times 10^{-4}, \quad \hat{r}^2 = 0.1 \) with the top variant), comprised the 99% credible set and supported the association. We tested in detail the accuracy of the imputation for the rs146662075 variant by comparing the imputed results from the same individuals genotyped by two different platforms (Methods) and found that the imputation was highly accurate in males only when using UK10K, but not in females, nor when using 1000G \( R^2_{[\text{UK10K males}]} = 0.94, \quad R^2_{[\text{UK10K females}]} = 0.66, \quad R^2_{[\text{1000G males}]} = 0.62, \quad \) and \( R^2_{[\text{1000G females}]} = 0.43; \) Supplementary Fig. 8). Whether this association is specific to men, or whether it also affects female carriers, remains to be clarified with datasets that allow accurate imputation on females, or with direct genotyping or sequencing.

To further validate and replicate this association, we next analyzed four independent data sets (SIGMA\(^6\), INTERACT\(^4\), Partners Biobank\(^4\), and UK Biobank\(^8\)), by performing imputation with the UK10K reference panel. In addition, a fifth cohort was genotyped de novo for the rs146662075 variant in several Danish sample sets. The initial meta-analysis, including the five replication data sets did not reach genome-wide significance (OR = 1.57, \( p = 1.2 \times 10^{-5}; \) Supplementary Fig. 9A), and revealed a strong degree of heterogeneity (heterogeneity \( p_{\text{het}} = 0.004), \) which appeared to be driven by the replication cohorts.

As a complementary replication analysis, within the case-control studies, there was a nested prospective cohort study, the Inter99, which consisted of 1,652 nondoniatric male subjects genotyped for rs146662075, of which 158 developed T2D after 11 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 \( \text{HR} = 3.17 \) [1.3–7.7], \( p = 0.011, \) Fig. 3b). Nearly 30% of carriers of the T risk allele developed incident T2D during 11 years of follow-up, compared to only 10% of noncarriers.

To understand the strong degree of heterogeneity observed after adding the replication datasets, we compared the clinical and demographic characteristics of the discovery and replication cohorts, and found that the majority of the replication datasets contained control subjects that were significantly younger than 55 years, the average age at the onset of T2D reported in this study and in Caucasian populations\(^46\). 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 Fig. 10). Given the supporting results, with the Inter99 prospective cohort, we performed an additional analysis using a stricter definition of controls, to minimize the presence of prediabetics or individuals that may further develop diabetes after reaching the average age at the onset. For this, we applied two additional exclusion criteria: (i) subjects younger than 55 years and (ii), when possible, excluding individuals with measured 2-h plasma glucose values during oral glucose tolerance test (OGTT) above 7.8 mmoll\(^-1\), a threshold employed to identify impaired glucose tolerance (prediabetes)\(^47\), or controls with family history of T2D, both being strong risk factors for developing T2D. While the application of the first filter alone did not yield genome-wide significant results (Supplementary Fig. 9B), upon excluding individuals with prediabetes or a family history of T2D, the replication results were significant and consistent with the initial discovery results (OR = 1.57 [1.19–2.07], \( p = 0.0014). \) The combined analysis of the discovery and replication cohorts resulted in genome-wide significance, confirming the association of rs146662075 with T2D (OR = 1.95 [1.56–2.45], \( p = 7.8 \times 10^{-9}, \) Fig. 3c).

### Allele-specific enhancer activity of the rs146662075 variant

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 Fig. 7, Supplementary Data 5), as well as the highest CADD (scaled C-score = 15.68; Supplementary Data 8), and LINSIGHT score (Supplementary Data 9). 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 150 kb) genes. The closest gene AGTR2, which encodes for the angiotensin II receptor type 2, has been previously associated with insulin secretion and resistance\(^48\–50\).

From the analysis of available epigenomic data sets\(^51\), we found no evidences of H3K27ac or other enhancer regulatory marks in human pancreatic islets; whereas a significant association was observed between the presence of H3K27ac enhancer marks and the expression of AGTR2 across multiple tissues (Fisher test \( p = 4.45 \times 10^{-3}; \) showing the highest signal of both H3K27ac and AGTR2 RNA-seq expression, but not with other genes from the same topologically associated domain (TAD), in fetal muscle (Fig. 4a; Supplementary Figure 11).

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 (non-risk) 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 (Fig. 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 the rare T allele (Fig. 4c). Overall, these data indicate that the risk T allele prevents the binding of a nuclear protein that is associated with decreased activity of an AGTR2-linked enhancer.

### Discussion

Through harmonizing and reanalyzing publicly available T2D GWAS data, and performing genotype imputation with two whole-genome sequence-based reference panels, we are able to perform deeper exploration of 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 1–22 and X.

The reanalysis of these data confirmed a large fraction of already-known T2D loci, and identified novel potential causal variants by fine mapping and functionally annotating each locus.

This reanalysis also allowed us to identify seven novel associations, five driven by common variants in or near LYPLAL1, NEUROG3, CAMKK2, ABO, and GIP; a low-frequency variant in EHM12, and a rare variant in the X chromosome. This rare variant identified in QX23 chromosome was located near the AGTR2 gene, and showed nearly twofold increased risk for T2D in males, which represents, to our knowledge, the largest effect size identified so far in Europeans, and a magnitude similar to other variants with large effects identified in other populations\(^52\,53\).

Our study complemented other efforts that also aim at unraveling the genetics behind T2D through the generation of new
**Fig. 3** Discovery and replication of rs146662075 association signal. **a** Forest plot of the discovery of rs146662075 variant. Cohort-specific odds ratios are denoted by boxes proportional to the size of the cohort and 95% CI error bars. The combined OR estimated for all the data sets is represented by a 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. **b** Kaplan–Meier plot showing the cumulative incidence of T2D for a 11 years follow-up. The red line represents the T carriers and in light blue, C carriers are represented (n = 1,652, cases = 158). **c** Forest plot after excluding controls younger than 55 years, OGTT >7.8 mmol l⁻¹, and controls with family history of T2D in both the discovery and replication cohorts when available.
genetic data. For example, we provided for the first time a comprehensive coverage of structural variants, which point to previously unobserved candidate causal variants in known and novel loci, as well as a comprehensive coverage of the X chromosome through sequence-based imputation.

This study also highlights the importance of a strict classification of both cases and controls, in order to identify rare variants associated with disease. Our initial discovery of the Xq23 locus was only replicated when the control group was restricted to T2D-free individuals who were older than 55 years (average age...
at the onset of T2D), had normal glucose tolerance, and no family history of T2D. 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 OGGT as criteria for classifying cases and controls, but not when using HbA1c. Our observation that 30% of the rs146662075 risk allele carriers developed T2D over 11 years of follow-up, compared to 10% of noncarriers, further supports the association of this variant and suggests that an early identification of these subjects through genotyping may be useful to tailor pharmacological or lifestyle intervention to prevent or delay the onset of T2D.

Using binding and gene-reporter analyses, we demonstrated a functional role of this variant and proposed a possible mechanism behind the pathophysiology of T2D in risk allele carriers, in which this rare variant could favor a gain of function of AGTR2, previously associated with insulin resistance. Therefore, as a potential therapeutic target for this disease, which would be in line with previous studies showing that the blockade of the renin-angiotensin system is effective in mice and in humans presenting typical features of T2D, and in scores.

Overall, beyond our significant contribution toward expanding the number of genetic associations with T2D, our study also highlights the potential of the reanalysis of public data, as a complement to large studies that use newly generated data. This study informs the open debate in favor of data sharing and democratization initiatives, for investigating the genetics and pathophysiology of complex diseases, which may lead to new preventive and therapeutic applications.

Methods

Quality filtering for imputed variants. In order to assess genotype imputation quality and to determine an accurate post-imputation quality filter, we made use of the Wellcome Trust Case Control Consortium (WTCCC) data available through the European Genotype Archive (EGA, https://www.ebi.ac.uk/ega/studies/EGAS000000000028). The genotyping data and the subjects included in the following tests were filtered according to the guidelines provided by the WTCCC, whose criteria of exclusion are in line with standard quality filters for GWAS. We used the 1958 British Birth cohort (≈3,000 samples, SBC) that was genotyped by Affymetrix v6.0 and Illumina 1M chips. After applying the quality filtering criteria, 2,706 and 2,699 subjects from the Affymetrix and Illumina data, respectively, were available for the SBC samples, leaving an intersection of 2,509 individuals genotyped by both platforms. After variant quality filtering and excluding all the variants with minor allele frequency (MAF) below 0.01, 717,556, and 892,516 variants remained for SBC Affymetrix and Illumina platforms, respectively.

We used a two-step genotype imputation approach based on prephasing the study genotypes into full haplotypes with SHAPEIT2 to ameliorate the computational burden required for genotype imputation through IMPUTE. We included variants with IMPUTE2 info score ≥ 0.5. The correlation between genotype dosages estimated when imputing using Affy or IL as the backbone. The quality of the imputed variants was evaluated using the allelic dosage R2 correlation coefficient, between the genotype dosages estimated when imputing using Affy or IL as the backbone. The Affy GT and IL GT SNPs were used to evaluate the correspondence between the allelic dosage R2 scores and the IMPUTE2 info scores for the imputed genotypes.

The 70KforT2D resource. We collected genetic individual-level data for T2D case/control studies from five independent datasets, Gene Environment Association Studies initiative [GENEVA], Wellcome Trust Case Control Consortium [WTCCC], Finland–United States Investigation of NIDDM Genetics [FUSION], Resource for Genetic Epidemiology Research on Aging [GERA], and the Northwestern NUgene project [NuGENE] publicly available in the dbGaP (http://www.ncbi.nlm.nih.gov/gap) and EGA (https://www.ega-archive.org) public repositories, comprising a total of 13,201 cases and 59,656 controls (for the description of each cohort, see Supplemental Note 1 and Supplementary Data 1).

Each dataset was independently harmonized and quality controlled with a three-step protocol, including two stages of SNP removal and an intermediate stage of sample exclusion. The exclusion criteria for variants were i) missing call rate ≥ 0.05, ii) significant deviation from Hardy-Weinberg equilibrium (HWE) p ≤ 1×10−6 for controls and p ≤ 1×10−7 for the entire cohort, iii) significant differences in the proportion of missingness between cases and controls p ≤ 1×10−4, and iv) MAF < 0.01 (for the GERA cohort, we considered a MAF of 0.001). The exclusion criteria for samples were i) genotypes for samples between the two genotyping platforms were not genetically predicted sex, ii) subject relatedness (pairs with π ≥ 0.125 from which we removed the individual with the highest proportion of missingness), iii) missing call rates per sample ≥ 0.02, and iv) population structure showing more than four standard deviations within the distribution of the study population according to the first four principal components.

We performed genotype imputation independently for each cohort by prephasing the genotypes to whole haplotypes with SHAPEIT2 and then, we performed genotype imputation with IMPUTE2. We tested for association with additive logistic regression using SNPRTEST, seven derived principal components set as covariates, body-mass-index and sex, except for WTCCC, where BMI was available and was not applied to the final X chromosome data set, for meta-analyzing the results from males and females separately (which were already filtered by R2 < 0.75).

For the meta-analysis with the DIAGRAM trans-ethnic study, we excluded from the whole 70KforT2D datasets those cohorts that overlapped with the DIAGRAM data. Therefore, we meta-analyzed the GERA and NuGENE cohorts (7,884 cases and 50,446 controls) and the 70KforT2D dataset.

The 70KforT2D and inclusion of previous summary statistics data. We meta-analyzed the different sets from the 70KforT2D dataset set with METALL, using the inverse variance-weighted fixed effect model. We included variants with I2 heterogeneity < 75. This filter was not applied to the final X chromosome data set, after meta-analyzing the results from males and females separately (which were already filtered by R2 < 75).

Fig. 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 is shown. The association between RNA-seq signals and H3K27ac marks suggests 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.
DIAGRAM trans-ethnic meta-analysis, we performed a sample size based meta-analysis, which converts the direction of the effect and the p-value into a Z-score. In addition, we also performed an inverse variance-weighted fixed effect meta-analysis to estimate the final effect sizes. This approach required the estimation of the beta and standard errors from the summary statistics (p-value and odds ratio). For the meta-analysis of coding low-frequency variants with the Type 2 Diabetes Knowledge Portal (T2D Portal), we included the 70kForT2D data set and the NuGENE and GERA cohorts (7,522 cases and 50,446 controls), to avoid overlapping samples. Like in the previous scenario, standard errors were not provided for the T2D Portal data and we used a sample size based meta-analysis with METAL. However, to estimate the effect sizes, we also calculated the standard errors from the p-values and odds ratios, and we performed an inverse variance-weighted fixed effect meta-analysis.

See further details about the cohorts in Supplementary Note 1.

Pathway and enrichment analysis. Summary statistics that resulted from the 70kForT2D meta-analysis were analyzed by Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT) to prioritize likely causal genes, to highlight enriched pathways, and to identify the most relevant tissues/cell types; DEPICT relies on publicly available gene sets (including molecular pathways) and leverages gene expression data from 77,840 gene expression arrays, to perform gene prioritization and gene-set enrichment based on predicted gene function and the so-called reconstituted gene sets. A reconstituted gene set contains a membership probability for each gene and conversely, each gene is functionally characterized by its membership in multiple gene sets across 14,461 reconstituted gene sets. As an input to DEPICT, we used all summary statistics from autosomal variants with \( p < 1 \times 10^{-5} \) in the 70kForT2D meta-analysis. We used an updated version of DEPICT, which handled 100G Phase3-integrated haplotypes (June 2014, www.broadinstitute.org/depic).

DEPICT was run using 3,412 associated SNPs (\( p < 1 \times 10^{-5} \)), from which we identified independent SNPs using PLINK and the following parameters: \(-c\)lump-p1 5e-8, \(-c\)lump-p2 1e-5, \(-c\)lump-r2 0.6, and \(-c\)lump-kB 250. We used LD \( r^2 > 0.5 \) distance to define locus limits yielding 70 autosomal loci comprising 119 genes (note that this is not the same locus definition which we used elsewhere in the text). We ran DEPICT with default settings, i.e., using 500 permutations to adjust for bias to estimate false discovery rate (FDR). We used normalized expression data from 77,840 Affymetrix microarrays to reconstitute gene sets. The resulting 14,461 reconstituted gene sets were tested for enrichment analysis. A total of 209 tissue or cell type expression data sets assembled from 37,427 Affymetrix U133 Plus 2.0 Arrays were used for enrichment in tissue/cell-type expression. DEPICT identified 103 reconstituted genes sets significantly enriched (FDR < 5%) for genes found among the 70 loci associated to T2D. We did not consider reconstituted sets in which genes of the original gene set were not nominally enriched (Wilcoxon rank-sum test), as these are expected to be enriched in the reconstituted gene set by design. The lack of enrichment makes the interpretation of the reconstituted gene set challenging because the label of the reconstituted gene set will not be accurate. Hence, the following reconstituted gene sets were removed from the results (Wilcoxon rank sum and \( P \)-values in parentheses): MP:0004247 gene set (\( p = 0.73 \)), GO:00070491 gene set (\( p = 0.04 \)), MP:00040731 gene set (\( p = 0.17 \)), MP:0005491 gene set (\( p = 0.54 \)), GO:0005159 gene set (\( p = 0.04 \)), MP:00012661 gene set (\( p = 0.02 \)), MP:0006344 gene set (\( p = 0.42 \)), MP:0004188 gene set (\( p = 0.22 \)), MP:0002189 gene set (\( p = 0.02 \)), MP:00000033 gene set (\( p = 0.08 \)), ENSG00000116604 gene set (\( p = 0.13 \)), GO:0005138 gene set (\( p = 0.07 \)), and MP:0001715 gene set (\( p = 0.01 \)). After applying the filters described above, there were 89 significantly enriched reconstituted gene sets. We used the affinity projection tool to cluster related reconstituted gene sets (network diagram script available from https://github.com/perslab/DEPICT).

We also used the VSE R package to compute the enrichment or depletion of genetic variants comprised in the 57 credible sets listed in Supplementary Data5 across regulatory genomic annotations, as described in14. Each GWAS lead variant from the final meta-analysis was considered as a tag SNP and variants from the corresponding 99% credible set (Supplementary Data 5) in LD with the tag SNP (\( R^2 \geq 0.4 \)), as a cluster or associated variant set (AVS). In order to account for the size and structure of the AVS, a null distribution was built based on random permutations of the AVS. Each permuted AVS was matched to the original AVS, cluster by cluster using HapMap data by size and structure. This Matched Random Variant Set (MRVS) was calculated using 500 permutations. Significant enrichments or depletions were considered when the Bonferroni-adjusted \( p \)-value was < 0.01. Human islet regulatory elements (C1–C5) were obtained from15.

Definition of 99% credible sets of GWAS-significant loci. For each genome-wide significant region, we identified the fraction of variants that have, in aggregate, 99% probability of containing the causal T2D-associated variant. We used our 70kForT2D meta-analysis based on imputed data (NuGENE, GERA, FUSION, GENEVA, and WTCCC data sets, comprising 12,231 cases and 57,196 controls), we defined the 99% credible set of variants for each locus with a Bayesian posterior probability of being the causal one11. We estimated the approximate Bayes factor (ABF) for each variant as

\[
ABF = \sqrt{1 - e^{r^2/2}},
\]

where

\[
0 \leq r = \frac{\text{SE}^2 + 0.04}{\beta^2} \leq 1
\]

and the \( \text{SE} \) and the \( \beta \) are the estimated effect size and the corresponding standard error resulting from testing for association under a logistic regression model. The posterior probability for each variant was obtained as

\[
\text{Posterior Probability}_i = \frac{ABF_i}{\sum_i ABF_i}
\]

where \( ABF \), corresponds to the approximate Bayes factor for the marker \( t \) and \( \beta \) represents the sum of all the \( ABF \) values from the candidate variants enclosed in the interval being evaluated. This calculation assumes that the prior of the \( \beta \) corresponds to a Gaussian with mean 0 and variance 0.04, which is also the same prior commonly used by SNPTEST, the program being used for calculating single variant associations.

Finally, we ranked variants according to the \( ABF \) (in decreasing order) and from this ordered list, we calculated the cumulative posterior probability. We included variants in the 99% credible set of each region until the SNP that pushed the cumulative posterior probability of association over 0.99.

The 99% credible sets of variants for each of the 57 GWAS-significant regions are summarized in Supplementary Data 5.

Characterization of indels. We examined whether indels from the 99% credible sets were present or absent in the 1000G Phase1 or UK10K reference panels, and also checked whether they were present or not in the 1000G Phase3 reference panel. All the information has been summarized in Supplementary Data 6. We also visually inspected the aligned sequences of the most relevant indels from both projects to discard that they could be alignment artifacts.

Functional annotation of the 99% credible set variants. To determine the effect of 99% credible set variants on genes, transcripts, and protein sequence, we used the variant effect predictor (VEP, GRCh37.p13 assembly)12. The VEP application determines the effect of variants (SNPs, insertions, deletions, CNVs, or structural variant(s) on genes, transcripts, proteins, and regulatory regions. We used as input the coordinates of variants within 99% credible sets and the corresponding alleles, to find out the affected genes and RefSeq transcripts and the consequence on the protein sequence by using the GRCh37.p13 assembly. We also manually checked all these annotations with the Exome Aggregation Consortium data set (ExAC, http://exac.broadinstitute.org) and the most updated VEP server based on the GRCh38.p7 assembly. All these annotations are provided in Supplementary Data 7.

We used combined annotation-dependent depletion (CADD) scoring function to prioritize functional, deleterious, and disease causal variants. We obtained the scaled C-score (PHRED-like scaled C-score ranking each variant with respect to all possible substitutions of the human genome) metric for each 99% credible set variant, as it highly ranks causal variants within individual genome sequences14 (Supplementary Data 8). We also used the LINSIGHT score to prioritize functional variants, which measures the probability of negative selection on noncoding sites by combining a generalized linear model for functional genomic data with a probabilistic model of molecular evolution15. For each credible set variant, we retrieved the precomputed LINSIGHT score at that particular nucleotide site, as well as the mean LINSIGHT precomputed score for a region of 20 bp centered on each credible set variant, respectively (https://github.com/CdbiSiepelLab/LINSIGHT/). These metrics are summarized in Supplementary Data 9.

In order to prioritize functional regulatory variants, we used the V6 release from the GTEx data that provides gene-level expression quantifications and eQTL results based on the annotation with GENCODE v19. This release included 450 genotyped donors, 8,555 RNA-seq samples across 51 tissues, and two cell lines, which led to the identification of eQTLs across 44 tissues16. Moreover, RNA-seq data from human pancreatic islets from 89 deceased donors cataloged as eQTLs and exon use (sQTL) were also integrated with the GWAS data to prioritize candidate regulatory variants17 but in pancreatic islets, which is a target tissue for T2D. Both analyses are summarized in Supplementary Data 10 and Supplementary Data 11, respectively.

Conditional analysis. To confirm the independence between novel loci and previously known T2D signals, we performed reciprocal conditional analyses (Supplementary Data 5, Supplementary Data 12, Supplementary Data 13, and Supplementary Data 14). We included the conditioning SNP as a covariate in the
In silico functional characterization of rs146662075. This variant is located in an intergenic region, flanked by AGTR2 and SLC6A14 genes, and within several DNase I hypersensitive sites. We searched for regulatory markers (i.e., H3K4me1 and H3K27ac marks) through the HaploReg web server (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php), in order to assess which type of regulatory element was associated with the rs146662075 variant.

replication

Replication of the rare variant association at Xq23. To replicate the association of the rs146662075 variant, we performed genotype imputation with the UK10K reference panel in four independent data sets: the InterAct case-control study44, the Slim Initiative in Genomic Medicine for the Americas (SIGMA) consortium GWAS45, the VenaHealth Care HealthCare Partners Biobank cohort46, and the UK Biobank cohort47. We used Cox-proportional hazard regression models to assess the association of the rs146662075 variant for association with type 2 diabetes using SNPTEST v2.5.1 and exclusion of relatives reported to be third degree or higher according to UK Biobank, we performed a luciferase assay. A region of 969 bp surrounding rs146662075 was labeled using dCTP [γ-32P] (Perkin Elmer). Oligonucleotide sequences are as follows (SNP location is underlined): probe-C: 5′-gatcTTTGAACACGAGGGGAAAAT-3′ and R: 5′-gatcATTTTCCCCCTCTGTGGT-3′ pre-bp 5′-gatcTTTGAACACGAGGGGAAAAT-3′ and R: 5′-gatcATTTTCCCCCTCTGTGGT-3′. Assay specificity was assessed by preincubation of nuclear extracts with 50- and 100-fold excess of unlabelled wild-type or mutant probes, followed by electrophoresis on a 5% nondenaturing polyacrylamide gel. Findings were confirmed by repeating binding assays on separate days.

Allelic-specific enhancer activity at rs146662075. The mouse C2C12 cell line (ATCC CRL-1772) was grown in DMEM medium supplemented with 10% FBS and was induced to differentiate in DMEM with 10% horse serum for 4 days.

The human myofibblast cell line was established by Prof. Giulio Cossu (Institute of Inflammation and Repair, University of Manchester)48. The authors played no role in the procurement of the tissue. Cells were cultured in DMEM medium supplemented with 10% fetal calf serum and was induced to differentiate in DMEM with 2% horse serum for 4 days. To perform an electrophoretic mobility shift assay, nuclear extracts from mouse myoblasts (C2C12 cells) and the human myoblast cell line (ATCC CRL-1772) were obtained as described before49. Double-stranded oligonucleotides containing either the common or rare variants of rs146662075 were labeled using dCTP [γ-32P] (Perkin Elmer). Oligonucleotide sequences are as follows (SNP location is underlined): probe-C: 5′-gatcTTTGAACACGAGGGGAAAAT-3′ and R: 5′-gatcATTTTCCCCCTCTGTGGT-3′. Assay specificity was assessed by preincubation of nuclear extracts with 50- and 100-fold excess of unlabelled wild-type or mutant probes, followed by electrophoresis on a 5% nondenaturing polyacrylamide gel. Findings were confirmed by repeating binding assays on separate days.

In silico functional characterization of rs146662075. This variant is located in an intergenic region, flanked by AGTR2 and SLC6A14 genes, and within several DNase I hypersensitive sites. We searched for regulatory markers (i.e., H3K4me1 and H3K27ac marks) through the HaploReg web server (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php), in order to assess which type of regulatory element was associated with the rs146662075 variant.

Received: 13 April 2017 Accepted: 24 November 2017

Published online: 22 January 2018
References


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 and Beatriu de Pinós fellowship from the Agency for Management of University and Research Grants (AGAUR). Silvia Bonás was FI-DGR fellowship from the Generalitat de Catalunya. This study makes use of data generated by the WTCCC. A full list of the investigators of the SIGMA Type 2 Diabetes Consortium for access to the data to replicate the rs146662075 variant. A full list of the investigators of the SIGMA Type 2 Diabetes and the InterAct consortia is provided in Supplementary Notes 3 and 4. 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). This research has been conducted using the UK Biobank Resource (application number 16803). We also acknowledge Bianca C. Porneala, MS for his technical assistance in the collection and curation of the genotype and phenotype data from Partners Biobank. 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 RSC for their helpful discussions and valuable comments on the manuscript.

Author contributions

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-017-02380-9.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2018