nature research

Peer Review Information

Journal: Nature Geneics
Manuscript Title: Rare variant association analysis in 51,256 type 2 diabetes cases and 370,487 controls informs the spectrum of pathogenicity of monogenic diabetes genes.
Corresponding author name(s): Josep Mercader

Reviewer Comments & Decisions:

Decision Letter, initial version:

16th February 2024

Dear Josep,

Your Article "Rare variant association analysis in 51,256 type 2 diabetes cases and 370,487 controls informs the spectrum of pathogenicity of monogenic diabetes genes" has been seen by two referees. You will see from their comments below that, while they find your work of interest, they have raised several relevant points. We are interested in the possibility of publishing your study in Nature Genetics, but we would like to consider your response to these points in the form of a revised manuscript before we make a final decision on publication.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision, and sometimes overruling referee requests that are deemed beyond the scope of the current study. In this case, we ask that you revise the presentation for clarity where needed and address all technical points related to the association analyses by adjusting significance thresholds where appropriate, incorporating additional data where available and/or modifying claims where required. We hope you will find this prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage, we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already, please begin to revise your manuscript so that it conforms to our Article format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

Please use the link below to submit your revised manuscript and related files:

[redacted]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within 8-12 weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information, please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,
Kyle

Kyle Vogan, PhD
Senior Editor
Nature Genetics
https://orcid.org/0000-0001-9565-9665

Referee expertise:

Referee #1: Genetics, diabetes, monogenic and complex diseases

Referee #2: Genetics, diabetes, monogenic and complex diseases

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:

This paper reports the largest rare-variant GWAS meta-analysis for T2D, using the rare-variant imputation enabled by the TOPMed panel (which it showcases and highlights), in conjunction with the WGS of All of Us. A number of novel variants are found, some of which are low-frequency or rare. More importantly, this paper addresses for the first time, in a systematic way and relying on GWAS methodology, the issue of incomplete penetrance of monogenic diabetes (MD) among individuals clinically diagnosed as T2D. A number of MD variants of intermediate penetrance are identified, which clinicians will find helpful in the interpretation of genetic testing for MD and in pharmacogenetic studies evaluating therapeutic options in carriers. At the same time, a large number of CIPs and VUS ClinVar entries are conclusively shown to be benign, which should go a long way in decluttering the database. Overall, this paper contributes significantly to the genetics of diabetes in multiple ways, but a number of issues need to be resolved.

Points of substance

1. As the paper admits, most of the variants that are both novel and rare are of borderline statistical significance. It is stated that TOPMed imputation discovered 10-fold more rare variants. Correlation of these previously unimputable SNPs with the rest of the GWAS set should be minimal and they should be treated as additional independent hypotheses tested, potentially pushing the threshold by as much as an order of magnitude. The HNF4A variant safely clears any threshold adjustment, and the HbA1c-associate SNPs have additional support. To claim the remaining ones, the paper should make a better effort to estimate the number of hypotheses tested here. Until this is done, it is doubtful that the elaborate epigenomics shown in Fig. 2 are justified without randomly chosen negative controls.

2. The debunking of many ClinVar CIPs is, in principle, an important contribution. However, a look at Supplementary Table 6 shows that most of them should not have been there in the first place. Of the 445 CIPs listed (this number, by the way, should be explicitly stated in line 197 of the main text), 170 should not have been there in the first place, as they can be ruled out on the basis of MAF > 10E-4, incompatible with a penetrant variant for a condition with the low prevalence of MD. Please revise the statement in line 206 to reflect what percentage of the credible CIPs it refers to.

3. The paper correctly points out (lines 222-224) that the pathogenicity of variants rated P and LP in ClinVar could not be confirmed because of lack of statistical power. However, analysis of the mutation burden of these variants as a group for each MD gene (or even in the aggregate) is meaningful and feasible. It will provide information of potential clinical value, on the proportion of MD cases that are misdiagnosed as ordinary T2D. The proportion may be far from negligible.

4. What is the point of constructing a separate rare-variant PRS rather than simply adding to the existing PRS? How many subjects had more than one of these rare variants to make such analysis meaningful? Terminologically, rvPRS may be a contradiction in terms, depending on whether all rare variants were used, or only the statistically significant ones (in which case it should be called rvGRS). This part of the study needs much better documentation.

Suggestions for improvement

5. The Abstract should start with a question, not with a compact summary of complicated methodology. It would read much better if the text in lines 45-47 is preceded by a clear statement of the objectives of this study.

6. In Table 1 and in Supplementary Table 6, the column indicating the chromosomal coordinates is labeled "ID" and "MarkerName" respectively. Genomic coordinate should be labeled as such. It takes more than that to ID a variant —the alternative allele should also be specified.

7. Supplementary Table 6, column heading log10p. Did you mean -log10p?

8. Line 73. Genes "code for" a protein or simply "encode" a protein. "Encodes for" is awkward English.

9. Lines 118-120 repeat verbatim what had just been stated at the end of the introduction. Please eliminate one of the two mentions.

10. In Line 136, the word "performing" is utterly unnecessary. I have been challenging my own trainees to completely eliminate the verb "to perform" from their papers, an action that almost always results in more elegant scientific prose.

Reviewer #2:
Remarks to the Author:

This paper is a clearly described thorough piece of work that moves the field of human genotype-phenotype mapping forwards in several ways. First, it combines deeper imputation of rare variants with rare variants defined from whole genome sequence data. This approach has identified putative alleles that are present in <1 in 500 people but with larger effects than we are used to with standard GWAS. In this reviewer's opinion, this aspect is important because most variants in the genome are rare. Second, it provides a more population-based and therefore likely less biased assessment of true population penetrance of rare variants of clinical and potential clinical significance. The paper focuses on an exemplar common disease, type 2 diabetes.

I have the following comments that hopefully could help:

1. Table 1 describes the novel rare variants of larger effect. Most of these are of borderline significance at 1 – 4 x 10-8, and ideally would benefit from further replication. Have other datasets become available since submission with TOPMed imputation to provide further statistical evidence for or against these alleles? (the exception is the HNF4A variant).

2. The authors acknowledge that P < 5 x 10-8 may not be the equivalent of 0.05 in the context of rare variant testing. They point to orthogonal evidence from some signals, but did they run some "dummy phenotypes" to obtain more empirical estimates of the type 1 error? Running such an analysis in the UK Biobank and All of Us data for example could be informative?

A minor point:

3. The assessment of variants of intermediate significance, etc., and polygenic scores *in the context of rare disease are nice additions. I believe it is well established now that polygenic scores modify penetrance of rare/"monogenic" mutations in many common conditions that also have rare forms – most obviously breast cancer and extreme LDL-cholesterol levels. A sentence or two in the discussion placing in context of this knowledge could be helpful to the reader?

Author Rebuttal to Initial comments

RESPONSE TO REVIEWERS

Response to Reviewers for Manuscript: "Rare variant association analysis in 51,256 type 2 diabetes cases and 370,487 controls informs the spectrum of pathogenicity of monogenic diabetes genes (NG-A63640).

We thank the editors and reviewers for their insightful comments, which have greatly strengthened the manuscript.
Below, we highlight the main improvements of the revised manuscript:

- We performed extensive replication analysis and modified the claims of the novel findings accordingly.
- We conducted luciferase reporter assays to demonstrate reduced enhancer activity of the African/African American risk allele of the variant near *LEP*.
- We have performed burden tests for each of the ClinVar groups to contextualize the effect of VIP variants in monogenic genes.

In the following, please find reviewers' comments in regular font, our point-by-point response in blue font, and changes to the revised manuscript in *italicized blue font*.

REVIEWER'S COMMENTS TO AUTHORS:
Reviewer #1:
Remarks to the Author:

This paper reports the largest rare-variant GWAS meta-analysis for T2D, using the rare-variant imputation enabled by the TOPMed panel (which it showcases and highlights), in conjunction with the WGS of All of Us. A number of novel variants are found, some of which are low-frequency or rare. More importantly, this paper addresses for the first time, in a systematic way and relying on GWAS methodology, the issue of incomplete penetrance of monogenic diabetes (MD) among individuals clinically diagnosed as T2D. A number of MD variants of intermediate penetrance are identified, which clinicians will find helpful in the interpretation of genetic testing for MD and in pharmacogenetic studies evaluating therapeutic options in carriers. At the same time, a large number of CIPs and VUS ClinVar entries are conclusively shown to be benign, which should go a long way in decluttering the database. Overall, this paper contributes significantly to the genetics of diabetes in multiple ways, but a number of issues need to be resolved.

We thank the reviewer for highlighting that this paper contributes significantly to the genetics of diabetes in multiple ways. In the responses below, we describe how we have addressed the points raised by the reviewer.

Points of substance

1. As the paper admits, most of the variants that are both novel and rare are of borderline statistical significance. It is stated that TOPMed imputation discovered 10-fold more rare variants. Correlation of these previously unimputable SNPs with the rest of the GWAS set should be minimal and they should be treated as additional independent hypotheses tested, potentially pushing the threshold by as much as an order of magnitude. The HNF4A variant safely clears any threshold adjustment, and the HbA1c-associate SNPs have additional support. To claim the remaining ones, the paper should make a better effort to estimate the number of hypotheses tested here. Until this is done, it is doubtful that the elaborate epigenomics shown in Fig. 2 are justified without randomly chosen negative controls.

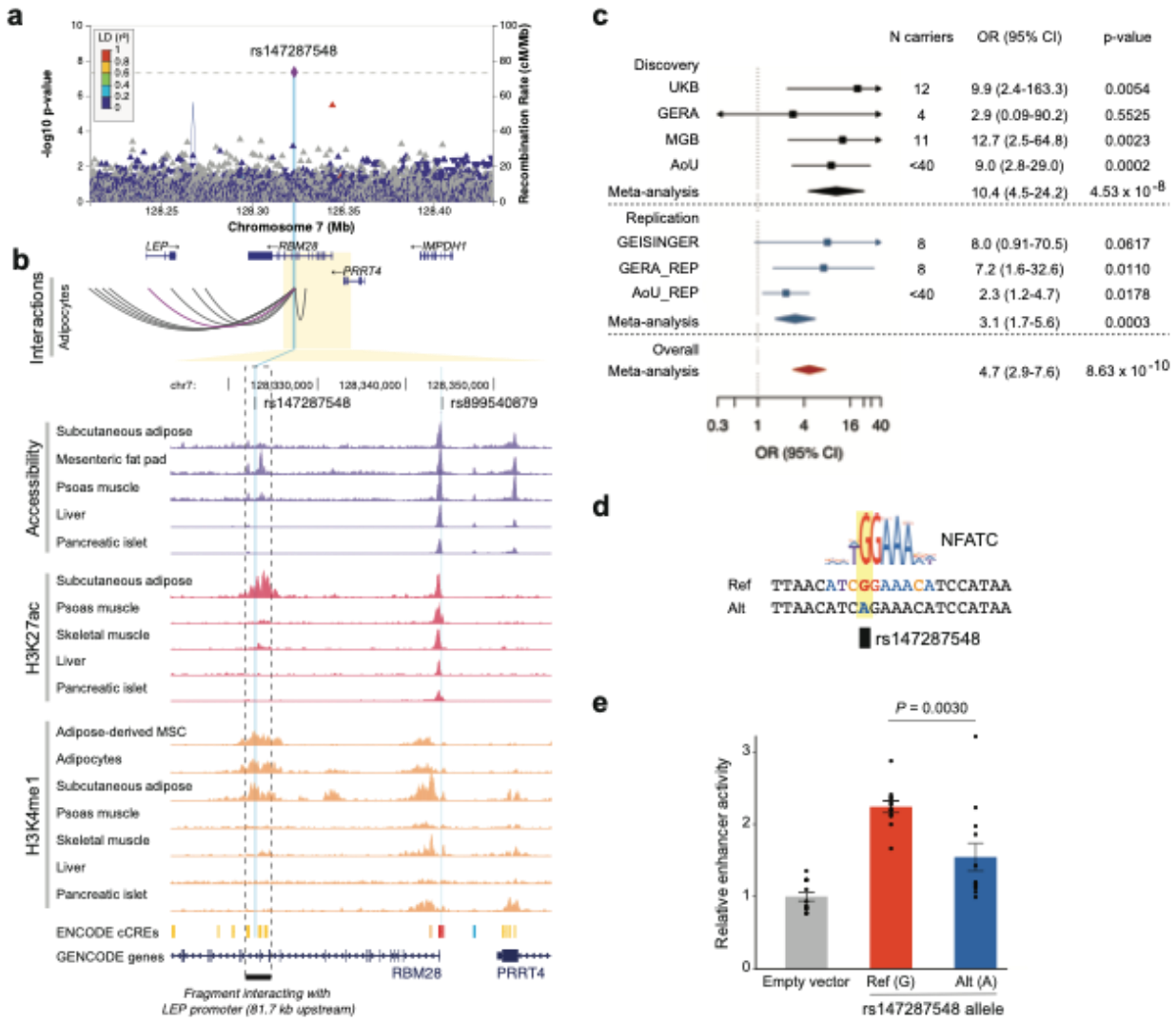
Thank you for raising this important point regarding the multiple testing threshold, given the ability to test a larger number of variants. In order to claim the significance of the novel findings, we pursued replication with several additional data, including (1) a new subset of GERA samples from a non-European subset not included in (and unrelated to) the discovery sample (2,737 T2D cases and 9,270 controls); (2) data from a new release of All of Us (17,693 T2D cases and 28,918 controls), not included in (and unrelated to) the previous discovery meta-analysis; and (3), 52,658 T2D cases and 41,639 controls from the Geisinger Health myCode. After this replication effort, in the revised manuscript, we now only claim the novel variants that achieve a discovery plus replication overall p-value $< 5 \times 10^{-9}$ for the non-coding variants and those that achieve a p-value $< 3.1 \times 10^{-5}$ for coding variants in monogenic diabetes genes, based on Bonferroni correction of 0.05/1,634 variants tested in 22 monogenic diabetes genes.

All the variants showed a consistent direction of effect with the discovery. We were able to successfully replicate the variant rs147287548 near *LEP* (Discovery + replication OR (95% CI) = 4.7 (2.9-7.6), p-value = 8.8×10^{-10}), and three of the coding variants in *HNF4A* (Discovery + replication OR (95% CI) = 7.9 (5.0-12.7), p-value = 3.09×10^{-18}), *GCK* (Discovery + replication OR (95% CI) = 8.0 (3.5-18.34), p-value = 9.38×10^{-7}) and *HNF1A* (Discovery + replication OR (95% CI) = 5.42 (2.9-10.2), p-value = 1.76×10^{-7}). For rs184889639 near *KCNV1* and rs190043928 near *MIDEAS*, we provide orthogonal evidence of their relationship with diabetes risk, given their association with higher fasting glucose and Hba1c levels, respectively. We do not claim as novel all the other variants that we were not able to replicate or for which there is no additional evidence of association with glycemic traits.

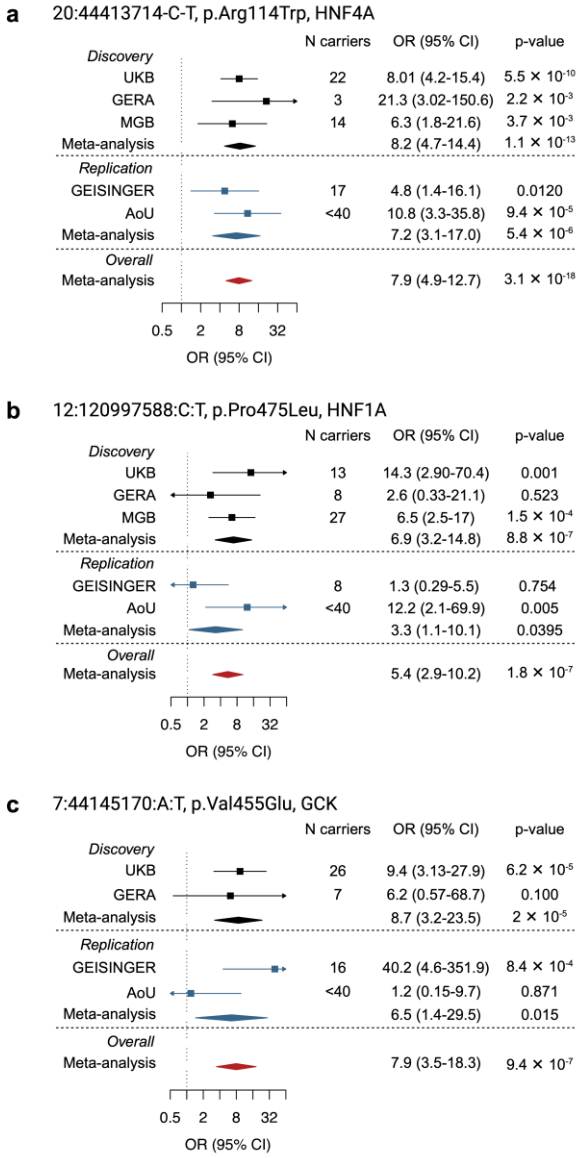
We also updated the list of novel variants based on the latest and largest common variant T2D GWAS meta-analysis (Suzuki et al. 2024) and the tables and figures accordingly.

Furthermore, for the rs147287548 variant in a *LEP* enhancer, we performed luciferase reporter assays and showed that the T2D risk allele reduces the regulatory potential of the enhancer in adipocytes *in vitro*.

We have updated Figure 2, which is now only on the variant near *LEP*, and includes a replication forest plot and results from the luciferase reporter assays (Response Figure 2) and added replication for three of the identified coding VIPs (Response Figure 4).



Response Figure 2. Functional characterization of novel rs147287548 low-frequency variant in *LEP* enhancer associated with T2D.



Response Figure 4. Effect of three identified variants of intermediate penetrance (VIP) on T2D risk.

We described the extended analyses as follows:

Results, Page 5, Lines 160-167: *"We tested these variants for replication in independent TOPMed imputed, whole-exome, and WGS data for a total of 73,088 T2D cases and 79,827 controls from the GEISINGER MyCode cohort and participants from GERA (GERA_REP) and AoU (AoU_REP)*

cohorts not overlapping and unrelated with discovery (Supplementary Table 1). In our replication dataset, all the variants showed a consistent direction of effect with the discovery (binomial p-value = 0.007). We had sufficient statistical power and imputation quality to test 7 of them for replication, of which we replicated two variants and found supporting evidence for two additional variants based on association with other glycemic traits (Supplementary Table 4).”

Results, Page 6, Lines 182-192: *“7:128323039-G-A (rs147287548; MAF_{AFA}=0.002, Discovery: OR=10.4, 95% CI=4.5-24.2, p = 4.53×10^{-8}; Discovery + Replication: OR=4.7, 95% CI=2.8-7.6, p = 8.8×10^{-10}) is an African/African American (AFA) specific variant, located in an enhancer active in adipose tissue and adipose tissue-derived mesenchymal stem cells that interacts with the LEP gene[1] (Figure 2a,b,c, Supplementary Figure 6, Supplementary Table 5). The AFA-specific allele of this variant (A) disrupts a binding motif for NFATc transcription factor, which is implicated in adipogenesis[2]. Using luciferase reporter assays, we demonstrate that the risk allele reduces the regulatory potential of the enhancer in adipocytes in vitro (Figure 2d,e). Moreover, 7:128323039-G-A is associated with lower apolipoprotein A levels (beta = -0.084 g/L, p = 0.003) and lower HDL cholesterol levels in UKB (beta = -0.117 mmol/L, p = 0.002) in participants without diabetes (Supplementary Table 6, Supplementary Figure 7a,b).”*

Results, Page 6, Lines 193-198: *“The second is a non-synonymous variant (20:44413714-C-T, p.Arg114Trp) in HNF4A, a known gene causing Maturity Onset Diabetes of the Young (MODY). During the preparation of this work, it was classified as having “conflicting interpretations of pathogenicity” (CIP) in ClinVar (accessed July 2023)[3] and was associated with ~8-fold increased risk of T2D (rs137853336; MAF=0.0001, Discovery: OR=8.3, 95% CI=4.7-14.14, p = 1.08×10^{-13}; Discovery + Overall: OR=7.9, 95% CI=4.9-12.7, p = 3.1×10^{-18}).”*

Results, Page 7, Lines 209-222: *“Despite the lack of replication for T2D, two additional variants showed evidence of association with glycemic traits. Variant 8:110165438-T-C (rs184889639, MAF_{AFA}=0.02, Discovery: OR=1.9, 95% CI=1.5-2.3, p = 1.06×10^{-9}, Discovery + Replication: OR=1.4, 95% CI=1.2-1.6, p = 9×10^{-6}), near KCNV1, is enriched in AFA ancestry and is associated with higher fasting glucose (p = 0.009) in the latest MAGIC Hispanic subset[4] (Supplementary Table 6, Supplementary Figure 7c). Variant 14:73781721-C-T (rs190043928, MAF_{AFA}=0.01, Discovery: OR = 2.8 95% CI = 2.0-3.9, p = 1.03×10^{-8}; Discovery + Replication: OR=1.7, 95% CI=1.3-2.1, p = 3.5×10^{-5}), near MIDEAS, is also enriched in AFA ancestry and is associated with higher HbA1c (p = 0.003) in the MAGIC European subset[4]. We were not able to*

replicate the three remaining variants, which may be either false positives or have a true lower effect size compared to the discovery."

Results, Page 8, Lines 250-260: *"Among the 381 variants within the ClinVar category of CIP (64.3% with MAF < 0.0001), we identified five VIPs with OR>5 and 95%CI LB >2 (Figure 3b, Supplementary Figure 8). Including the HNF4A p.Arg114Trp described above, we replicated three of the five CIP VIPs in the full AoU and GEISINGER cohorts (Figure 4, Supplementary Table 8). They are in the known MODY genes, HNF1A (12:120997588:C:T, p.Pro475Leu, enriched in Ashkenazi Jewish, $maxMAF_{AJ}$=0.00086, Discovery: OR=6.9, 95% CI=3.2-14.8, $p = 8.8\times10^{-7}$; Discovery + Replication: OR=5.4, 95% CI=2.9-10.2, $p = 1.8\times10^{-7}$), and GCK (7:44145170:A:T, p.Val455Glu, $maxMAF_{EUR}$=0.00002, Discovery: OR=8.7, 95% CI=3.2-23.6, $p = 2\times10^{-5}$; Discovery + Replication: OR=7.9, 95% CI=3.5-18.3, $p = 9.4\times10^{-7}$). p.Val455Glu is also associated with glucose (beta = 0.781 mmol/l, $p = 8.1\times10^{-5}$) and glycated hemoglobin (beta = 5.11%, $p = 7.4\times10^{-7}$) in the UKB cohort (Supplementary Figure 7o,p)."*

Results, Page 9, Lines 276-279: *"None of these VIPs reached statistical significance after multiple comparisons, and we were not able to replicate them as they were absent or showed a lack of statistical power to detect an association (power < 67% for an OR = 5) in the replication cohorts (Supplementary Table 8)."*

The new analyses are described in the supplementary methods as follows:

Supplementary methods, Lines 731-818: *"We tested the novel rare or population-specific variants for replication in three independent datasets not included in our discovery meta-analysis. Overall, the replication sample included 73,088 T2D cases and 79,827 controls (Supplementary Table 1). After generating the T2D summary statistics for each replication cohort, we used METAL to meta-analyze the results, weighting the cohorts by the inverse of the standard error for each variant. We considered replication when the variant showed evidence of association with T2D ($p < 0.05$) and consistent direction of effect with the T2D discovery GWAS meta-analysis.*

Genetic Epidemiology Research on Adult Health and Aging
The GERA replication cohort (GERA_REP) included individuals from African American, Admixed American, and East Asian ancestry[5] who were not previously included in the discovery GWAS. Because of the multi-ethnic nature of the cohort, different ethnic-specific arrays, based on the Affymetrix Axiom Genotyping System, were employed to enhance genome-wide coverage. The

number of SNPs and SNP content varied by array, with SNP content designed to maximize the coverage of low frequency and more common variants specific to the different race-ethnicity while also maximizing coverage of the whole genome and including new SNPs from sequencing projects and SNPs with established associations with disease phenotypes.
The custom array genotyped data was array-wise QCed and imputed to the TOPMed freeze 8 reference panel. Variant quality checks included the removal of variants with 5% or more missing data, MAF < 0.1%, a Hardy-Weinberg equilibrium p-value < 5×10^{-10}, palindromic single nucleotide variants (AT or CG), and duplicates. We also excluded samples with 2% or more missing data. Clean datasets were phased using SHAPEIT2 and used as input for imputation. We merged the imputed data using IMMERGE[6] to keep all variants present in any of the three imputed chunks with Rsq ≥ 0.3. After merging, we converted the VCF files to BGEN format for analysis with REGENIE.
We applied the same T2D definition as with GERA EUR and tested the association of the novel variants with T2D as a binary outcome and included age, sex, BMI, the top 10 PCs, and the imputation batch. In total, GERA_REP included 2,737 T2D cases and 9,270 controls.

All of Us Research Program
The AoU replication cohort (AoU_REP) included non-overlapping individuals and non-family members (pairwise kinship score > 0.1) to those included in the discovery meta-analysis as they were released in the posterior v7 data freeze. We used the AoU srWGS for a total of 17,693 independent T2D cases and 28,918 controls that were QCed and analyzed to test the association with T2D as previously described.

Geisinger MyCode cohort
The Geisinger MyCode Community Health Initiative (GEISINGER) is a biorepository of blood, serum and DNA samples that are linked to electronic health records for the purpose of broad research use. Participants were recruited from Geisinger, an integrated healthcare system located in central and northeastern Pennsylvania. The study sample consisted of 172,366 individuals with exome sequence and microarray genotype data. Available clinical data includes clinical diagnoses, procedures, medications, and laboratory results, and are updated daily[7]. The cohort is 61% female and consists of 93% genetically inferred EUR, 3% AFA, 1.5% AMR, 0.3% SAS, 0.3% EAS, and 1.7% other ancestries. The mean (SD) age at enrollment is 40.3 (1s8.9) years, range 0 to 99. Participation in MyCode is done through written consent under Geisinger Institutional Review Board Study #2016-0269. The results reported here were determined by the Geisinger IRB

to meet the criteria for "Non-human subject research" as defined in 45CFR46.102(e). All research was performed in accordance with relevant guidelines and regulations.

DNA Samples from 173,585 MyCode participants were exome sequenced using IDT exome capture/Illumina HiSeq. Of those, 172,366 were genotyped using Infinium OmniExpress Exome array (Illumina), and GSA-24v1-0 array (Illumina) through a collaboration between Geisinger and the Regeneron Genetics Center (Dewey FE et al). Rare coding variants for this study have genotype quality ≥ 30, read depth ≥ 15, allelic depth ≥ 20, and strand bias p-value ≥ 10%. For array-based genotyping data, SNPs with minor allele frequency ≤ 1%, significant deviation ($p \leq 1x10^{-15}$) from Hardy-Weinberg Equilibrium, and site-level missingness ≥ 1% were removed. Array-based genotypes were imputed to the TOPMED reference panel (97,256 deeply sequenced genomes) with a GRCh38 build using the TOPMED Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov), which employed Eagle v2.4 and Minimac4 as the phasing and imputation algorithm, respectively. All candidate non-coding rare variants in this study have an imputation info score ≥0.3. A set of LD-pruned (200 variant windows, 50 variant sliding windows, and $r^2 < 0.25$), high quality (info score ≥0.3) common variants (919,227) were applied for PCA (PLINK2.0).

Diabetes was defined using outpatient ICD9/10 codes for diabetes, hbA1c, fasting blood glucose, and antihyperglycemic medications (other than biguanides). Laboratory definition for diabetes used thresholds defined by the American Diabetes Association. In absence of other evidence, individuals with IC9/10 codes for diabetes for at least two encounters more than 3 months apart were included individuals with only an ICD9/10 code for diabetes were excluded from the study. T2D cases were defined 1) having a ratio T1D/T2D ICD9/10 count ≤0.50, and 2) the following exclusion: a. on insulin medication only and has ≥2 insulin medication prescriptions, or b. 1 ICD9/10 code for insulin pump and has ≥2 insulin medication prescriptions, or c. on insulin medication only and diabetes duration ≥3 years and has ≥2 insulin medication prescriptions, or d. has >1 ICD9/10 for insulin pump. Controls were defined using a modified version of the Northwestern T2D algorithm. Controls were individuals ≥45 years of age without an ICD9/10 code for diabetes or antihyperglycemic medications including biguanides, and no HbA1c ≥ 5.6% or random glucose ≥ 200 mg/dl or fasting glucose ≥ 100 mg/dl. The index age for controls was calculated by the subtraction of the last active date (primarily last encounter date) and the birth date. The index age for cases was the age of the first recorded evidence of diabetes.

Genomic-determined sex, index age, and ten major principal components (PCs) were included as covariates in the association study using REGENIE v3.4.1. Only high-quality common variants were used to fit the whole genome regression model for the dichotomous or quantitative traits, and a set of genomic predictions were produced as output of REGENIE Step 1. The candidate rare variants are tested for association in REGENIE Step 2 with Firth logistic regression model to fit

the dichotomous traits. Rank Inverse Normal Transformation was applied to the quantitative phenotypes in both Step 1 and Step 2."

Supplementary methods, Lines 912-949: "*Cell culture and in vitro differentiation of adipocytes.*

3T3-L1 mouse fibroblasts were obtained from the European Collection of Authenticated Cell Cultures (ECACC). Cells were cultured in high glucose DMEM (Sigma) supplemented with 10% fetal bovine serum (FBS) (ThermoFisher), 2 mM glutamine, 100 U/mL penicillin and 100 µg/mL streptomycin (all Sigma). The cells were differentiated to adipocytes by first growing to confluence for 48 hours prior to commencing the differentiation protocol. On differentiation day 0, cells were treated with basal medium supplemented with 0.5 mM 3-isobutyl-1-methylxanthine (IBMX) (Sigma I5879), 1 µM dexamethasone (Sigma D4902), and 1 µg/mL insulin (Sigma I9278). After 48 hours, medium was replaced by medium supplemented with 1 µg/mL insulin only for the next 4 days, refreshing the medium every 48 hours. Mature adipocytes were assessed for accumulation of lipid droplets by visual inspection using a brightfield microscope on day 8 and experiments were performed on day 10.

Luciferase reporter assays.

The LEP enhancer region containing the variant 7:128323039-G-A was amplified from human genomic DNA using the primers Fw ctggcctaactggccggtacTCACCCCTGTTCCTCATC and Rv tatataccctctagtgtctaGCCCAGAACTTAGCCCTG (amplicon coordinates: chr7:128322865-128323766). The PCR amplicon was then cloned into the pGL4.23 vector (Promega) at KpnI and HindIII, upstream of a minimal promoter and the Firefly luciferase coding sequence (CDS), using Gibson Assembly[8] (New England Biolabs). The cloned enhancer sequence contained the non-risk allele (7:128323039-G); therefore, the enhancer sequence containing the risk allele (7:128323039-A) was produced via site directed mutagenesis, which was performed using Q5 Site-Directed Mutagenesis Kit (New England Biolabs) with primers GGTTAACATCaGAAACATCCATAAAATG and GATAACTCATATTCCTTTGC. Correct cloning was confirmed by Sanger sequencing (Genewiz). Plasmids were amplified in DH5α and purified using PureYield™ Plasmid Midiprep System (Promega).

Enhancer reporter construct transfections were carried out on 40,000 3T3-L1-derived adipocytes in 24-well plates with 1 µg of tested construct and 0.1 µg of the internal control Renilla-expressing pGL7.74 vector, by reverse transfection using 1 µL of Lipofectamine 2000 (Thermo Fisher) per 1 µg of construct. Luciferase activity was measured 24 hours post transfection with Dual-Luciferase

Reporter Assay kit (Promega) on a GloMax-Multi Microplate Multimode Reader (Promega). The empty pGL4.23 vector was used as an internal control and showed no significant signal above background. The construct pGL4.13 (Promega), containing the SV40 early enhancer/promoter for high expression, was used as internal positive control in all experiments (not shown). Firefly luciferase measurements were normalized to Renilla luciferase. Three experiments with four independent transfections were performed per construct. Two-sided Student's t-test was used to calculate significance."

2. The debunking of many ClinVar CIPs is, in principle, an important contribution. However, a look at Supplementary Table 6 shows that most of them should not have been there in the first place. Of the 445 CIPs listed (this number, by the way, should be explicitly stated in line 197 of the main text), 170 should not have been there in the first place, as they can be ruled out on the basis of MAF > 10E-4, incompatible with a penetrant variant for a condition with the low prevalence of MD. Please revise the statement in line 206 to reflect what percentage of the credible CIPs it refers to.

We are glad that the reviewer finds the analysis of ClinVar CIPs an important contribution.

We appreciate that a number of variants could be filtered out based on MAF > 1E-4. While this cutoff is used to assess pathogenic variants, we are interested in using our approach also to assess whether any of these variants are of intermediate penetrance, for which a MAF threshold has not been established yet. In addition, MAF can still be not fully accurate for populations that are underrepresented in genomic databases, such as gnomAD. Therefore, while some of the variants can be filtered out to discard pathogenicity, we decided to still analyze all the VUS and CIP variants present in ClinVar, with MAF < 0.001, and assess their association with T2D as a metric to inform the likelihood of each variant being benign, intermediate penetrance, or pathogenic.

Nevertheless, we have also clarified in the manuscript the total number of CIPs and how many of them still have MAF < 1E-4 to improve clarity as follows:

Results, page 8, lines 241-244: *"In a meta-analysis that included UKB, MGB, and GERA, we evaluated the effect of 1,634 well-imputed variants (MAF < 0.001) in 22 MD genes reported in ClinVar (Supplementary Table 7). Of those, 1,007 of the variants (61.6%) had a MAF < 0.0001."*

Results, page 8, lines 250-253: "*Among the 381 variants within the ClinVar category of CIP (64.3% with MAF < 0.0001), we identified five VIPs with OR>5 and 95%CI LB >2 (Figure 3b, Supplementary Figure 8).*"

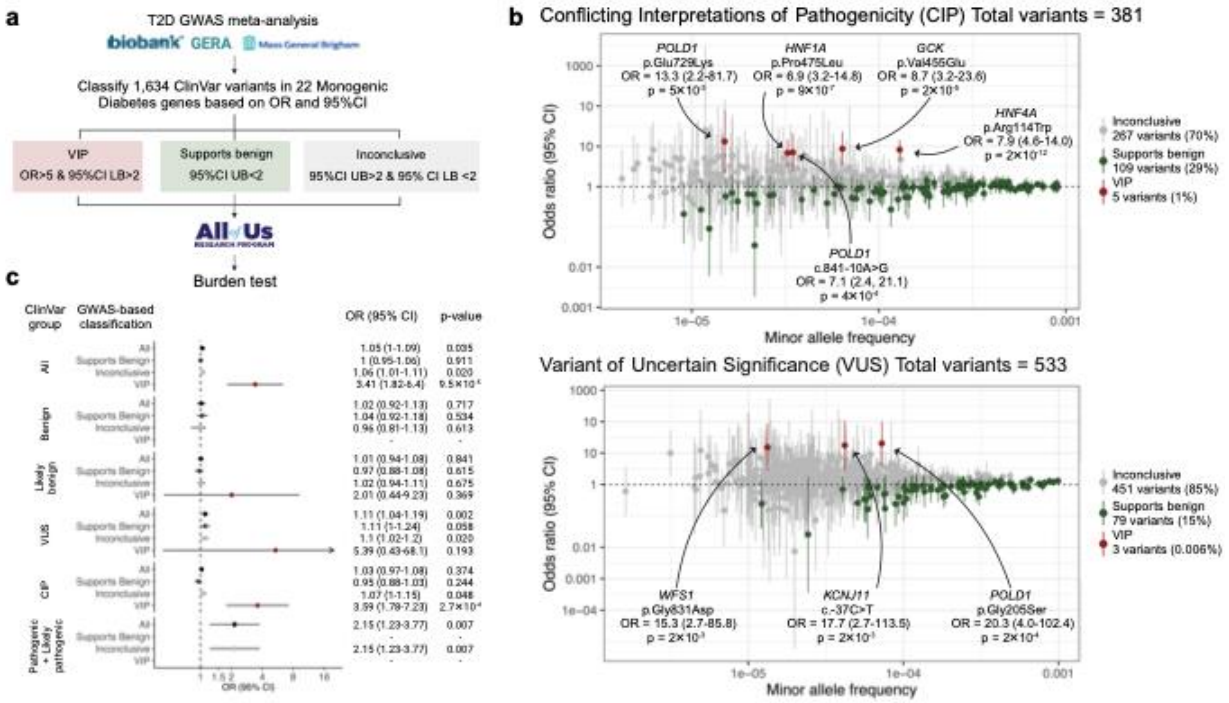
Results, page 9, lines 270-271: "*Within the 533 variants in the ClinVar category of "uncertain significance" (75.6% with MAF < 0.0001), we identified three additional VIPs…*"

3. The paper correctly points out (lines 222-224) that the pathogenicity of variants rated P and LP in ClinVar could not be confirmed because of lack of statistical power. However, analysis of the mutation burden of these variants as a group for each MD gene (or even in the aggregate) is meaningful and feasible. It will provide information of potential clinical value, on the proportion of MD cases that are misdiagnosed as ordinary T2D. The proportion may be far from negligible.

We thank the reviewer for this great suggestion. Following this idea, we tested the association of all P and LP in aggregate, using a burden test approach, and did the same for all the other ClinVar categories. We observed that, as a burden, carriers of P and LP show a higher risk of T2D (OR (95% CI) = 2.15 (1.23-3.8), p = 0.007, proportion of burden carriers in T2D cases = 1.3 x10^{-3} (35 out of 26,271) and controls = 6.9x10^{-4} (30 out of 43,174)). We also found an association of VUS group with T2D (OR (95% CI) = 1.1 (1.04,1.19), p = 0.002), probably reflecting that the VUS category is enriched with risk variants.

In contrast, carriers of VIPs showed an odds ratio of 3.41 (95% CI = 1.82-6.4, p = 9.5x10^{-5}), including all VIP variants from all ClinVar categories, and consistently increased risk when stratifying by any of the ClinVar categories, showing that our approach to identifying VIPs can provide added value in addition to the ClinVar categories.

We have created a revised version of Figure 3 where we now show, for each category, the burden association results, including all the variants, and then stratified according to our T2D GWAS meta-analysis results. Please see below the updated version of Figure 3 (Response Figure 3).



Response Figure 3. Classification of variants in MD genes from ClinVar and assessment of their effect when collapsing them in single burden variables according to the GWAS-based classification.

We have updated the description of these results as follows:

Results, pages 10, lines 296-312 *"To further validate the effects of the VIPs, we tested the association of carriers of VIPs in aggregate with T2D in the AoU cohort, using a burden test approach (see supplementary methods). We compared the association results using the whole set of ClinVar variants and stratified by the ClinVar categories (Figure 1d, Figure 3c). Except for the VUS POLD1 p.Gly205Ser, for which no carriers were present in the AoU cohort, carriers of the remaining 8 VIP variants exhibited a 3.4-fold increased risk for T2D (OR=3.4, 95% CI = 1.82-6.4, p = 9.5×10^{-5}), in contrast to variants identified as "supporting benign" (OR=1.0, 95% CI = 0.95-1.06, p = 0.911) and inconclusive variants (OR=1.06, 95% CI = 1.01-1.11, p = 0.02), which, in aggregate, showed no significant effect on T2D risk (Figure 3c).*

When aggregating the variants according to the ClinVar criteria, only the VUS (OR=1.1, 95% CI=1.04-1.19, p = 0.002) and the "likely pathogenic" and "pathogenic" groups (OR=2.15, 95% CI=1.23-3.8, p = 0.007) showed association with T2D. However, VIPs showed consistent direction

when stratifying by ClinVar categories, with VIP in CIP ClinVar category showing significantly increased risk (OR=3.59, 95% CI=1.78-7.23, p = 2.7×10^{-4}), and "likely benign-VIP" (OR=2.01, 95% CI=0.44-9.23, p = 0.369), "VUS-VIP" (OR=5.39, 95% CI=0.43-68.1, p = 0.193) groups not reaching statistical significance (Figure 3c)."

We also added a description of this analysis in supplementary methods:

Supplementary methods, Lines 841-853: "*Testing a rare variant burden in an independent sample in All of Us. We aggregated the variants from each of the three classes described above ("VIP", "supports benign", and "inconclusive") either as a whole or stratified by ClinVar groups into single burden variables and then regressed the T2D phenotype in the burden variable to test for the cumulative effects of all possible combinations of ClinVar and GWAS-based classified variants, using the AoU cohort as a validation dataset (Figure 3). We ran burden tests using T2D as a binary outcome and included age, sex, body mass index (BMI), and 10 PCs as covariates with REGENIE version 2. We used a block size of 1,000 for step 1 and 400 for step 2."*

4. What is the point of constructing a separate rare-variant PRS rather than simply adding to the existing PRS? How many subjects had more than one of these rare variants to make such analysis meaningful? Terminologically, rvPRS may be a contradiction in terms, depending on whether all rare variants were used, or only the statistically significant ones (in which case it should be called rvGRS). This part of the study needs much better documentation.

We agree with the reviewer that constructing a rare-variant PRS is the same as performing a burden analysis because there are no subjects that carry more than one of these VIP rare variants. To avoid confusion, we have now described this analysis as a burden testing, where we compare carriers of any of the identified rare variants in aggregate (See response to previous request). We have updated this along with the manuscript text and figures.

Suggestions for improvement

5. The Abstract should start with a question, not with a compact summary of complicated methodology. It would read much better if the text in lines 45-47 is preceded by a clear statement of the objectives of this study.

We thank the reviewer for this suggestion, which will add clarity to the manuscript. The abstract now starts as follows:

Abstract, lines 51-57: *"Previous type 2 diabetes (T2D) Genome-Wide Association Studies (GWAS) have not investigated rare variation at scale, mostly due to the inability of previous reference panels to impute rare variation. To address the contribution of coding and non-coding rare genetic variants to T2D susceptibility, we meta-analyzed array data imputed with the TOPMed panel and whole-genome sequence (WGS) datasets and performed the largest, rare variant (minor allele frequency as low as 5×10^{-5}) GWAS meta-analysis of T2D, comprising 51,256 cases and 370,487 controls.*

6. In Table 1 and in Supplementary Table 6, the column indicating the chromosomal coordinates is labeled "ID" and "MarkerName" respectively. Genomic coordinate should be labeled as such. It takes more than that to ID a variant —the alternative allele should also be specified.

We have corrected the column labels accordingly.

7. Supplementary Table 6, column heading log10p. Did you mean -log10p?

Following the reviewer's suggestion, we have changed the heading to p-value.

8. Line 73. Genes "code for" a protein or simply "encode" a protein. "Encodes for" is awkward English.

We have corrected this in the manuscript.

9. Lines 118-120 repeat verbatim what had just been stated at the end of the introduction. Please eliminate one of the two mentions.

We thank the reviewer for flagging this repetition. We have eliminated this from the introduction (lines 125-126).

10. In Line 136, the word "performing" is utterly unnecessary. I have been challenging my own trainees to completely eliminate the verb "to perform" from their papers, an action that almost always results in more elegant scientific prose.

We have eliminated this word from this sentence.

Reviewer #2:
Remarks to the Author:

This paper is a clearly described thorough piece of work that moves the field of human genotype-phenotype mapping forwards in several ways. First, it combines deeper imputation of rare variants with rare variants defined from whole genome sequence data. This approach has identified putative alleles that are present in <1 in 500 people but with larger effects than we are used to with standard GWAS. In this reviewer's opinion, this aspect is important because most variants in the genome are rare. Second, it provides a more population-based and therefore likely less biased assessment of true population penetrance of rare variants of clinical and potential clinical significance. The paper focuses on an exemplar common disease, type 2 diabetes.

I have the following comments that hopefully could help:

1. Table 1 describes the novel rare variants of larger effect. Most of these are of borderline significance at 1 – 4 x 10-8, and ideally would benefit from further replication. Have other datasets become available since submission with TOPMed imputation to provide further statistical evidence for or against these alleles? (the exception is the HNF4A variant).

We agree with the reviewer that replication would be ideal to confirm these findings. In order to claim the significance of the novel findings, we pursued replication with several additional data, including (1) a new subset of GERA samples from a non-European subset not included in (and unrelated to) the discovery sample (2,737 T2D cases and 9,270 controls); (2) data from a new release of All of Us (17,693 T2D cases and 28,918 controls), not included in (and unrelated to) the previous discovery meta-analysis; and (3), 52,658 T2D cases and 41,639 controls from the Geisinger Health myCode. After this replication effort, in the revised manuscript, we now only claim the novel variants that achieve a discovery plus replication overall p-value $< 5 \times 10^{-9}$ for the non-coding variants and those that achieve a p-value $< 3.1 \times 10^{-5}$ for coding variants in monogenic diabetes genes, based on Bonferroni correction of 0.05/1,634 variants tested in 22 monogenic diabetes genes.

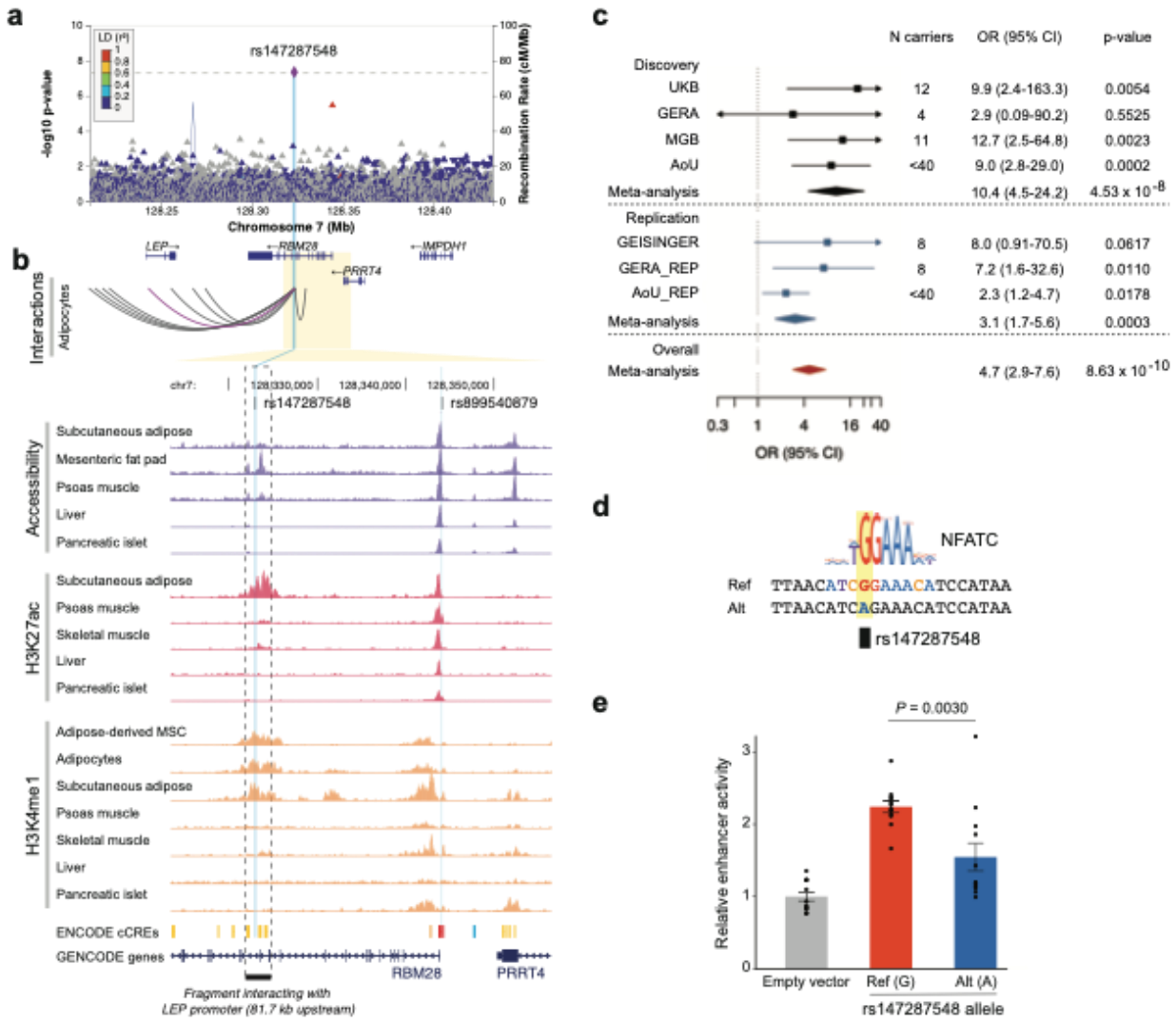
All the variants showed a consistent direction of effect with the discovery. We were able to successfully replicate the variant rs147287548 near *LEP* (Discovery + replication OR (95% CI) =

4.7 (2.9-7.6), p-value = 8.8×10^{-10}), and three of the coding variants in *HNF4A* (Discovery + replication OR (95% CI) = 7.9 (5.0-12.7), p-value = 3.09×10^{-18}), *GCK* (Discovery + replication OR (95% CI) = 8.0 (3.5-18.34), p-value = 9.38×10^{-7}) and *HNF1A* (Discovery + replication OR (95% CI) = 5.42 (2.9-10.2), p-value = 1.76×10^{-7}). For rs184889639 near *KCNV1* and rs190043928 near *MIDEAS*, we provide orthogonal evidence of their relationship with diabetes risk, given their association with higher fasting glucose and Hba1c levels, respectively. We do not claim as novel all the other variants that we were not able to replicate or for which there is no additional evidence of association with glycemic traits.

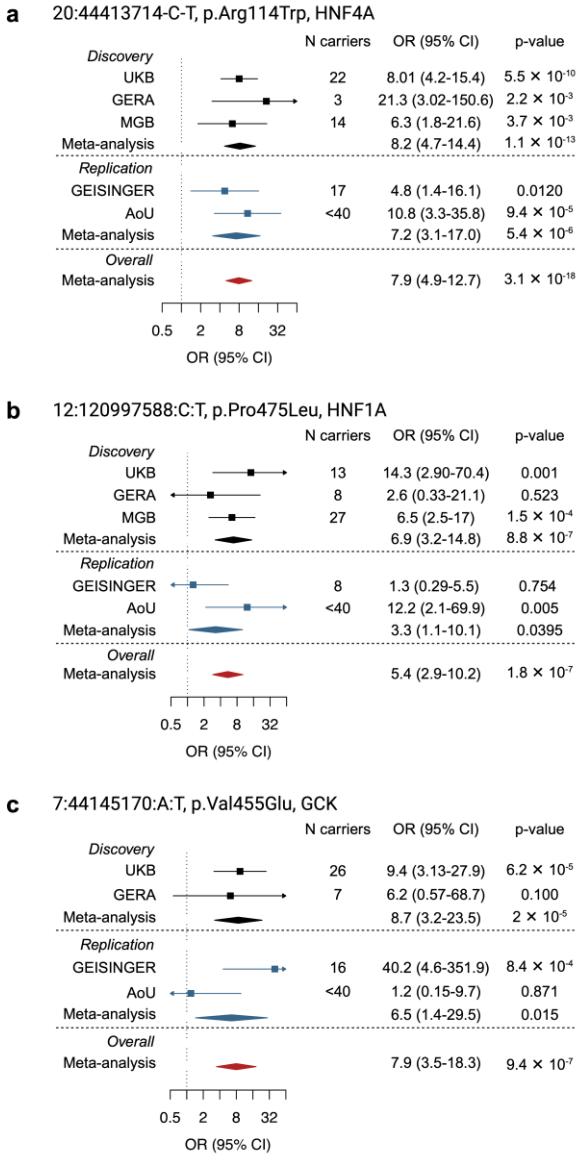
We also updated the list of novel variants based on the latest and largest common variant T2D GWAS meta-analysis (Suzuki et al. 2024) and the tables and figures accordingly.

Furthermore, for the rs147287548 variant in a *LEP* enhancer, we performed luciferase reporter assays and showed that the T2D risk allele reduces the regulatory potential of the enhancer in adipocytes *in vitro*.

We have updated Figure 2, which is now only on the variant near *LEP*, and includes a replication forest plot and results from the luciferase reporter assays (Response Figure 2) and added replication for three of the identified coding VIPs (Response Figure 4).



Response Figure 2. Functional characterization of novel rs147287548 low-frequency variant in *LEP* enhancer associated with T2D.



Response Figure 4. Effect of three identified variants of intermediate penetrance (VIP) on T2D risk.

We described the extended analyses as follows:

Results, Page 5, Lines 160-167: *"We tested these variants for replication in independent TOPMed imputed, whole-exome, and WGS data for a total of 73,088 T2D cases and 79,827 controls from the GEISINGER MyCode cohort and participants from GERA (GERA_REP) and AoU (AoU_REP)*

cohorts not overlapping and unrelated with discovery (Supplementary Table 1). In our replication dataset, all the variants showed a consistent direction of effect with the discovery (binomial p-value = 0.007). We had sufficient statistical power and imputation quality to test 7 of them for replication, of which we replicated two variants and found supporting evidence for two additional variants based on association with other glycemic traits (Supplementary Table 4)."

Results, Page 6, Lines 182-192: *"7:128323039-G-A (rs147287548; MAF_{AFA}=0.002, Discovery: OR=10.4, 95% CI=4.5-24.2, p = 4.53×10^{-8}; Discovery + Replication: OR=4.7, 95% CI=2.8-7.6, p = 8.8×10^{-10}) is an African/African American (AFA) specific variant, located in an enhancer active in adipose tissue and adipose tissue-derived mesenchymal stem cells that interacts with the LEP gene[1] (Figure 2a,b,c, Supplementary Figure 6, Supplementary Table 5). The AFA-specific allele of this variant (A) disrupts a binding motif for NFATc transcription factor, which is implicated in adipogenesis[2]. Using luciferase reporter assays, we demonstrate that the risk allele reduces the regulatory potential of the enhancer in adipocytes in vitro (Figure 2d,e). Moreover, 7:128323039-G-A is associated with lower apolipoprotein A levels (beta = -0.084 g/L, p = 0.003) and lower HDL cholesterol levels in UKB (beta = -0.117 mmol/L, p = 0.002) in participants without diabetes (Supplementary Table 6, Supplementary Figure 7a,b)."*

Results, Page 6, Lines 193-198: *"The second is a non-synonymous variant (20:44413714-C-T, p.Arg114Trp) in HNF4A, a known gene causing Maturity Onset Diabetes of the Young (MODY). During the preparation of this work, it was classified as having "conflicting interpretations of pathogenicity" (CIP) in ClinVar (accessed July 2023)[3] and was associated with ~8-fold increased risk of T2D (rs137853336; MAF=0.0001, Discovery: OR=8.3, 95% CI=4.7-14.14, p = 1.08×10^{-13}; Discovery + Overall: OR=7.9, 95% CI=4.9-12.7, p = 3.1×10^{-18})."*

Results, Page 7, Lines 210-222: *"Despite the lack of replication for T2D, two additional variants showed evidence of association with glycemic traits. Variant 8:110165438-T-C (rs184889639, MAF_{AFA}=0.02, Discovery: OR=1.9, 95% CI=1.5-2.3, p = 1.06×10^{-9}, Discovery + Replication: OR=1.4, 95% CI=1.2-1.6, p = 9×10^{-6}), near KCNV1, is enriched in AFA ancestry and is associated with higher fasting glucose (p = 0.009) in the latest MAGIC Hispanic subset[4] (Supplementary Table 6, Supplementary Figure 7c). Variant 14:73781721-C-T (rs190043928, MAF_{AFA}=0.01, Discovery: OR = 2.8 95% CI = 2.0-3.9, p = 1.03×10^{-8}; Discovery + Replication: OR=1.7, 95% CI=1.3-2.1, p = 3.5×10^{-5}), near MIDEAS, is also enriched in AFA ancestry and is associated with higher HbA1c (p = 0.003) in the MAGIC European subset[4]. We were not able to*

replicate the three remaining variants, which may be either false positives or have a true lower effect size compared to the discovery."

Results, Page 8, Lines 259-260: *"Among the 381 variants within the ClinVar category of CIP (64.3% with MAF < 0.0001), we identified five VIPs with OR>5 and 95%CI LB >2 (Figure 3b, Supplementary Figure 8). Including the HNF4A p.Arg114Trp described above, we replicated three of the five CIP VIPs in the full AoU and GEISINGER cohorts (Figure 4, Supplementary Table 8). They are in the known MODY genes, HNF1A (12:120997588:C:T, p.Pro475Leu, enriched in Ashkenazi Jewish, $maxMAF_{AJ}$=0.00086, Discovery: OR=6.9, 95% CI=3.2-14.8, $p = 8.8\times10^{-7}$; Discovery + Replication: OR=5.4, 95% CI=2.9-10.2, $p = 1.8\times10^{-7}$), and GCK (7:44145170:A:T, p.Val455Glu, $maxMAF_{EUR}$=0.00002, Discovery: OR=8.7, 95% CI=3.2-23.6, $p = 2\times10^{-5}$; Discovery + Replication: OR=7.9, 95% CI=3.5-18.3, $p = 9.4\times10^{-7}$). p.Val455Glu is also associated with glucose (beta = 0.781 mmol/l, $p = 8.1\times10^{-5}$) and glycated hemoglobin (beta = 5.11%, $p = 7.4\times10^{-7}$) in the UKB cohort (Supplementary Figure 7o,p)."*

Results, Page 9, Lines 276-279: *"None of these VIPs reached statistical significance after multiple comparisons, and we were not able to replicate them as they were absent or showed a lack of statistical power to detect an association (power < 67% for an OR = 5) in the replication cohorts (Supplementary Table 8)."*

The new analyses are described in the supplementary methods as follows:

Supplementary methods, Lines 731-818: *"We tested the novel rare or population-specific variants for replication in three independent datasets not included in our discovery meta-analysis. Overall, the replication sample included 73,088 T2D cases and 79,827 controls (Supplementary Table 1). After generating the T2D summary statistics for each replication cohort, we used METAL to meta-analyze the results, weighting the cohorts by the inverse of the standard error for each variant. We considered replication when the variant showed evidence of association with T2D ($p < 0.05$) and consistent direction of effect with the T2D discovery GWAS meta-analysis.*

Genetic Epidemiology Research on Adult Health and Aging
The GERA replication cohort (GERA_REP) included individuals from African American, Admixed American, and East Asian ancestry[5] who were not previously included in the discovery GWAS. Because of the multi-ethnic nature of the cohort, different ethnic-specific arrays, based on the Affymetrix Axiom Genotyping System, were employed to enhance genome-wide coverage. The

number of SNPs and SNP content varied by array, with SNP content designed to maximize the coverage of low frequency and more common variants specific to the different race-ethnicity while also maximizing coverage of the whole genome and including new SNPs from sequencing projects and SNPs with established associations with disease phenotypes.
The custom array genotyped data was array-wise QCed and imputed to the TOPMed freeze 8 reference panel. Variant quality checks included the removal of variants with 5% or more missing data, MAF < 0.1%, a Hardy-Weinberg equilibrium p-value < 5×10^{-10}, palindromic single nucleotide variants (AT or CG), and duplicates. We also excluded samples with 2% or more missing data. Clean datasets were phased using SHAPEIT2 and used as input for imputation. We merged the imputed data using IMMERGE[6] to keep all variants present in any of the three imputed chunks with Rsq ≥ 0.3. After merging, we converted the VCF files to BGEN format for analysis with REGENIE.
We applied the same T2D definition as with GERA EUR and tested the association of the novel variants with T2D as a binary outcome and included age, sex, BMI, the top 10 PCs, and the imputation batch. In total, GERA_REP included 2,737 T2D cases and 9,270 controls.

All of Us Research Program
The AoU replication cohort (AoU_REP) included non-overlapping individuals and non-family members (pairwise kinship score > 0.1) to those included in the discovery meta-analysis as they were released in the posterior v7 data freeze. We used the AoU srWGS for a total of 17,693 independent T2D cases and 28,918 controls that were QCed and analyzed to test the association with T2D as previously described.

Geisinger MyCode cohort
The Geisinger MyCode Community Health Initiative (GEISINGER) is a biorepository of blood, serum and DNA samples that are linked to electronic health records for the purpose of broad research use. Participants were recruited from Geisinger, an integrated healthcare system located in central and northeastern Pennsylvania. The study sample consisted of 172,366 individuals with exome sequence and microarray genotype data. Available clinical data includes clinical diagnoses, procedures, medications, and laboratory results, and are updated daily[7]. The cohort is 61% female and consists of 93% genetically inferred EUR, 3% AFA, 1.5% AMR, 0.3% SAS, 0.3% EAS, and 1.7% other ancestries. The mean (SD) age at enrollment is 40.3 (1s8.9) years, range 0 to 99. Participation in MyCode is done through written consent under Geisinger Institutional Review Board Study #2016-0269. The results reported here were determined by the Geisinger IRB

to meet the criteria for "Non-human subject research" as defined in 45CFR46.102(e). All research was performed in accordance with relevant guidelines and regulations.

DNA Samples from 173,585 MyCode participants were exome sequenced using IDT exome capture/Illumina HiSeq. Of those, 172,366 were genotyped using Infinium OmniExpress Exome array (Illumina), and GSA-24v1-0 array (Illumina) through a collaboration between Geisinger and the Regeneron Genetics Center (Dewey FE et al). Rare coding variants for this study have genotype quality ≥ 30, read depth ≥ 15, allelic depth ≥ 20, and strand bias p-value ≥ 10%. For array-based genotyping data, SNPs with minor allele frequency ≤ 1%, significant deviation ($p \leq 1x10^{-15}$) from Hardy-Weinberg Equilibrium, and site-level missingness ≥ 1% were removed. Array-based genotypes were imputed to the TOPMED reference panel (97,256 deeply sequenced genomes) with a GRCh38 build using the TOPMED Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov), which employed Eagle v2.4 and Minimac4 as the phasing and imputation algorithm, respectively. All candidate non-coding rare variants in this study have an imputation info score ≥0.3. A set of LD-pruned (200 variant windows, 50 variant sliding windows, and $r^2 < 0.25$), high quality (info score ≥0.3) common variants (919,227) were applied for PCA (PLINK2.0).

Diabetes was defined using outpatient ICD9/10 codes for diabetes, hbA1c, fasting blood glucose, and antihyperglycemic medications (other than biguanides). Laboratory definition for diabetes used thresholds defined by the American Diabetes Association. In absence of other evidence, individuals with IC9/10 codes for diabetes for at least two encounters more than 3 months apart were included individuals with only an ICD9/10 code for diabetes were excluded from the study. T2D cases were defined 1) having a ratio T1D/T2D ICD9/10 count ≤0.50, and 2) the following exclusion: a. on insulin medication only and has ≥2 insulin medication prescriptions, or b. 1 ICD9/10 code for insulin pump and has ≥2 insulin medication prescriptions, or c. on insulin medication only and diabetes duration ≥3 years and has ≥2 insulin medication prescriptions, or d. has >1 ICD9/10 for insulin pump. Controls were defined using a modified version of the Northwestern T2D algorithm. Controls were individuals ≥45 years of age without an ICD9/10 code for diabetes or antihyperglycemic medications including biguanides, and no HbA1c ≥ 5.6% or random glucose ≥ 200 mg/dl or fasting glucose ≥ 100 mg/dl. The index age for controls was calculated by the subtraction of the last active date (primarily last encounter date) and the birth date. The index age for cases was the age of the first recorded evidence of diabetes.

Genomic-determined sex, index age, and ten major principal components (PCs) were included as covariates in the association study using REGENIE v3.4.1. Only high-quality common variants were used to fit the whole genome regression model for the dichotomous or quantitative traits, and a set of genomic predictions were produced as output of REGENIE Step 1. The candidate rare variants are tested for association in REGENIE Step 2 with Firth logistic regression model to fit

the dichotomous traits. Rank Inverse Normal Transformation was applied to the quantitative phenotypes in both Step 1 and Step 2."

Supplementary methods, Lines 912-949: "*Cell culture and in vitro differentiation of adipocytes.*

3T3-L1 mouse fibroblasts were obtained from the European Collection of Authenticated Cell Cultures (ECACC). Cells were cultured in high glucose DMEM (Sigma) supplemented with 10% fetal bovine serum (FBS) (ThermoFisher), 2 mM glutamine, 100 U/mL penicillin and 100 µg/mL streptomycin (all Sigma). The cells were differentiated to adipocytes by first growing to confluence for 48 hours prior to commencing the differentiation protocol. On differentiation day 0, cells were treated with basal medium supplemented with 0.5 mM 3-isobutyl-1-methylxanthine (IBMX) (Sigma I5879), 1 µM dexamethasone (Sigma D4902), and 1 µg/mL insulin (Sigma I9278). After 48 hours, medium was replaced by medium supplemented with 1 µg/mL insulin only for the next 4 days, refreshing the medium every 48 hours. Mature adipocytes were assessed for accumulation of lipid droplets by visual inspection using a brightfield microscope on day 8 and experiments were performed on day 10.

Luciferase reporter assays.

The LEP enhancer region containing the variant 7:128323039-G-A was amplified from human genomic DNA using the primers Fw ctggcctaactggccggtacTCACCCCTGTTCCTCATC and Rv tatataccctctagtgtctaGCCCAGAACTTAGCCCTG (amplicon coordinates: chr7:128322865-128323766). The PCR amplicon was then cloned into the pGL4.23 vector (Promega) at KpnI and HindIII, upstream of a minimal promoter and the Firefly luciferase coding sequence (CDS), using Gibson Assembly[8] (New England Biolabs). The cloned enhancer sequence contained the non-risk allele (7:128323039-G); therefore, the enhancer sequence containing the risk allele (7:128323039-A) was produced via site directed mutagenesis, which was performed using Q5 Site-Directed Mutagenesis Kit (New England Biolabs) with primers GGTTAACATCaGAAACATCCATAAAATG and GATAACTCATATTCCTTTGC. Correct cloning was confirmed by Sanger sequencing (Genewiz). Plasmids were amplified in DH5α and purified using PureYield™ Plasmid Midiprep System (Promega).

Enhancer reporter construct transfections were carried out on 40,000 3T3-L1-derived adipocytes in 24-well plates with 1 µg of tested construct and 0.1 µg of the internal control Renilla-expressing pGL7.74 vector, by reverse transfection using 1 µL of Lipofectamine 2000 (Thermo Fisher) per 1 µg of construct. Luciferase activity was measured 24 hours post transfection with Dual-Luciferase

Reporter Assay kit (Promega) on a GloMax-Multi Microplate Multimode Reader (Promega). The empty pGL4.23 vector was used as an internal control and showed no significant signal above background. The construct pGL4.13 (Promega), containing the SV40 early enhancer/promoter for high expression, was used as internal positive control in all experiments (not shown). Firefly luciferase measurements were normalized to Renilla luciferase. Three experiments with four independent transfections were performed per construct. Two-sided Student's t-test was used to calculate significance."

2. The authors acknowledge that P < 5 x 10-8 may not be the equivalent of 0.05 in the context of rare variant testing. They point to orthogonal evidence from some signals, but did they run some "dummy phenotypes" to obtain more empirical estimates of the type 1 error? Running such an analysis in the UK Biobank and All of Us data for example could be informative?

We agree with the reviewer that additional work should be done to establish a genome-wide significance threshold for rare variant testing. However, we believe that this analysis may be beyond the scope of this paper and would require a large amount of computing time. Instead, we opted to focus our efforts on replicating our findings, including the association with T2D and related metabolic traits analyses, as well as functional analysis of the non-coding variants, to claim the validity of the novel findings (see above). We are therefore confident of the results for which we have a robust replication and have toned down and do not discuss further those that we did not successfully replicate.

A minor point:

3. The assessment of variants of intermediate significance, etc., and polygenic scores *in the context of rare disease are nice additions. I believe it is well established now that polygenic scores modify penetrance of rare/"monogenic" mutations in many common conditions that also have rare forms – most obviously breast cancer and extreme LDL-cholesterol levels. A sentence or two in the discussion placing in context of this knowledge could be helpful to the reader?

We agree with the reviewer that it is pertinent to mention in the discussion that this is one new example of how PRS can modify the penetrance of large effect variants. Following the reviewer's suggestion, we have added this idea to the discussion.

Discussion, page 13, lines 426-29: "*This is in line with previous studies that have shown how polygenic background can influence the penetrance of pathogenic mutations related to obesity, cancer susceptibility, lipid disorders, and coronary artery disease*[9,10,11,12-15]*.*"

References:

1. Madsen, J.G.S. *et al.* Highly interconnected enhancer communities control lineage-determining genes in human mesenchymal stem cells. *Nat Genet* **52**, 1227-1238 (2020).
2. Yang, T.T. *et al.* Role of transcription factor NFAT in glucose and insulin homeostasis. *Mol Cell Biol* **26**, 7372-87 (2006).
3. Landrum, M.J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067 (2018).
4. Chen, J. *et al.* The trans-ancestral genomic architecture of glycemic traits. *Nat Genet* **53**, 840-860 (2021).
5. Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1285-95 (2015).
6. Zhu, W. *et al.* IMMerge: merging imputation data at scale. *Bioinformatics* **39**(2023).
7. Mirshahi, U.L. *et al.* Reduced penetrance of MODY-associated HNF1A/HNF4A variants but not GCK variants in clinically unselected cohorts. *Am J Hum Genet* **109**, 2018-2028 (2022).
8. Gibson, D.G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343-5 (2009).
9. Chami, N., Preuss, M., Walker, R.W., Moscati, A. & Loos, R.J.F. The role of polygenic susceptibility to obesity among carriers of pathogenic mutations in MC4R in the UK Biobank population. *PLoS Med* **17**, e1003196 (2020).
10. Barnes, D.R. *et al.* Polygenic risk scores and breast and epithelial ovarian cancer risks for carriers of BRCA1 and BRCA2 pathogenic variants. *Genet Med* **22**, 1653-1666 (2020).
11. Lecarpentier, J. *et al.* Prediction of Breast and Prostate Cancer Risks in Male BRCA1 and BRCA2 Mutation Carriers Using Polygenic Risk Scores. *J Clin Oncol* **35**, 2240-2250 (2017).
12. Fahed, A.C. *et al.* Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat Commun* **11**, 3635 (2020).
13. Goodrich, J.K. *et al.* Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes. *Nat Commun* **12**, 3505 (2021).
14. Oetjens, M.T., Kelly, M.A., Sturm, A.C., Martin, C.L. & Ledbetter, D.H. Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. *Nat Commun* **10**, 4897 (2019).
15. Mars, N. *et al.* The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat Commun* **11**, 6383 (2020).

Decision Letter, first revision:

13th June 2024

Dear Josep,

Your revised manuscript "Rare variant association analysis in 51,256 type 2 diabetes cases and 370,487 controls informs the spectrum of pathogenicity of monogenic diabetes genes" (NG-A63640R) has been seen by the original referees. As you will see from their comments below, they find that the paper has improved in revision, and therefore we will be happy in principle to publish it in Nature Genetics as an Article pending final revisions to satisfy Reviewer #2's remaining requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper, and we will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials or make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics. Please do not hesitate to contact me if you have any questions.

Sincerely,
Kyle

Kyle Vogan, PhD
Senior Editor
Nature Genetics
https://orcid.org/0000-0001-9565-9665

Reviewer #1 (Remarks to the Author):

This revision corrects important flaws in the original version and reads much better. I have no further comments.

Reviewer #2 (Remarks to the Author):

The authors should be congratulated for an extensive replication effort and clear revision. However, I think it would be helpful for the field to just be very clear in the abstract and results and main tables, that this approach has resulted in one new signal , not 35 or 8 or other number. See below my rationale:

1. I suggest the abstract needs another revision. It starts by saying that 35 novel variants have been discovered, but 20 of these are common and 7 low frequency and not discussed further in the paper after the first few lines of the relevant results section. There is no replication of these 27 variants and so it maybe better to focus on the novel rare signals in the abstract, in keeping with the results

sections.

2. There are 8 variants in Table 1 and on line 159 in the results the authors talk about novel rare variants, but only 1, that in the LEP region could really be described as novel and rare, because the others do not replicate, or , in the case of the R114W HNF4A variant , are well known. I think it is too much to claim the two variants with glycaemic evidence as novel findings. The p values for T2D drop from -8 and -9 to -5 and -5 on adding in the large replication samples and the p values for glucose are around 0.01 and 0.003. there are several tests going on here – several variants, glucose, HbA1c, presumably fasting insulin was also tested. I think these could be real, but equally could be false positives. I suggest some further toning down, or just be more upfront and say that this approach has identified one really clear rare signal, and helped clarify roles of ClinVar annotated variants.

Minor point

1. For readability I suggest the authors do not start sentences with numbers.

Author Rebuttal, first revision:

RESPONSE TO REVIEWERS

Response to Reviewers for Manuscript: "Rare variant association analysis in 51,256 type 2 diabetes cases and 370,487 controls informs the spectrum of pathogenicity of monogenic diabetes genes (NG-A63640R).

We thank the editors and the reviewers for their interest in our study and for giving us the opportunity to address their concerns.

Below, we highlight the improvements of the revised manuscript:

- We now report the replication analysis performed for the low-frequency and common variants we had previously omitted, given our focus only on the rare variants.
- We only claim the novelty of the variants that we could replicate in external cohorts.
- We have removed any statement about the variants we could not replicate from the manuscript.

In the following, please find reviewers' comments in regular font, our point-by-point response in blue font, and changes to the revised manuscript in *italicized blue font*.

Reviewer #1:
Remarks to the Author:
This revision corrects important flaws in the original version and reads much better. I have no further comments.

We thank the reviewer for the insightful comments that helped substantially improve the manuscript's quality.

Reviewer #2:
Remarks to the Author:
The authors should be congratulated for an extensive replication effort and clear revision. However, I think it would be helpful for the field to just be very clear in the abstract and results and main tables that this approach has resulted in one new signal, not 35 or 8 or other number. See below my rationale:

1. I suggest the abstract needs another revision. It starts by saying that 35 novel variants have been discovered, but 20 of these are common and 7 low frequency and not discussed further in the paper after the first few lines of the relevant results section. There is no replication of these 27 variants and so it maybe better to focus on the novel rare signals in the abstract, in keeping with the results sections.

We thank the reviewer for thoroughly examining our manuscript. In the previous versions of the manuscript, we did not include the replication results of the common and low-frequency variants since we wanted to stress the ability of our meta-analysis to identify and assess the association of rare variants with diabetes. However, we acknowledge that the manner in which we presented the data may have been unclear.

We now provide replication data for the full set variants we found to be associated with type 2 diabetes at $p < 5 \times 10^{-8}$, irrespective of their minor allele frequencies, and then focus on the rare variants, for which we undertook additional replication efforts. We considered a significant replication *p*-value of 0.0015, from 0.05/34 tested variants. We replicated 11 out of 26 (42%) common and low-frequency variants and 1 out of 8 rare variants. We clearly declare that *HNF4A* p.Arg114Trp had been previously related to MODY-like with reduced penetrance.

The abstract (lines 48-62) now reads as:
"Type 2 diabetes (T2D) Genome-Wide Association Studies (GWAS) often overlook rare variants due to previous imputation panels' limitations and scarce whole-genome sequencing (WGS) data. We used TOPMed-imputation and WGS to conduct the largest T2D GWAS meta-analysis involving 51,256 T2D cases and 370,487 controls, targeting variants with minor allele frequency as low as 5×10^{-5}. We identified 12 novel variants, including a rare African/African American-enriched enhancer variant near the LEP gene (rs147287548), associated with 4-fold T2D risk. We also identified a rare missense variant in HNF4A (p.Arg114Trp), associated with 8-fold T2D risk, previously reported in Maturity Onset Diabetes of the Young with reduced penetrance, but observed for the first time in a T2D GWAS. We further leveraged this data to analyze 1,634 ClinVar variants in 22 genes related to monogenic diabetes, identifying 2 additional rare variants in HNF1A and GCK, associated with 5-fold and 8-fold T2D risk, respectively, whose effects were modified by the individual's polygenic risk score. For 21% of the variants with conflicting interpretations or uncertain significance in ClinVar, we provide support of being benign based on their lack of association with T2D. Our work provides a framework for using rare variant GWAS to identify large-effect variants and assess variant pathogenicity in monogenic diabetes genes."

We have revised the manuscript to thoroughly explain the number of variants we identified and how many of them we could replicate.

About the discovery meta-analysis, we state in lines 138-142:
"*Through approximate conditional and joint association analysis (COJO)[28], we identified 284 distinct signals in 214 loci. Of those, 239 were common (MAF > 0.05), 37 were low-frequency (MAF 0.05-0.001), and 8 were rare (MAF < 0.001). 34 out of the 284 total signals were novel, of which 8% (19 out of 239 variants), 19% (7 out of 37 variants), and 100% (8 variants) were common, low-frequency, and rare, respectively.*"

Regarding the replication, we report in lines 146-156 and lines 169-176:
"11 out of the 26 common and low-frequency novel variants replicated in two previous T2D GWAS meta-analyses[1,68], including independent samples (p < 0.0015 from 0.05/34 tested novel variants) (Supplementary Table 4). Since rare variants are not considered in previous meta-analyses, we looked for replication of the 8 novel rare variants in independent TOPMed imputed, whole-exome, and WGS data for a total of 73,088 T2D cases and 79,827 controls from the GEISINGER MyCode cohort and participants from GERA (GERA_REP) and AoU (AoU_REP) cohorts not overlapping and unrelated with discovery (Supplementary Table 1). In our replication dataset, all the rare variants showed a consistent direction of effect with the discovery (binomial p-value = 0.007). We had sufficient statistical power and imputation quality to test 7 of them, of which we replicated two variants (Supplementary Table 5), one non-coding near LEP gene, and a missense variant in HNF4A."

"The second is a non-synonymous variant (20:44413714-C-T, p.Arg114Trp) in HNF4A, a known gene causing Maturity Onset Diabetes of the Young (MODY). This variant has previously been reported as a mutation causing a distinct clinical subtype of MD, with reduced penetrance, reduced sensitivity to sulfonylurea treatment, and no effect on birth weight[33]."

Discussion, lines 301-305:
"*This study has enabled several discoveries. First, we identified and replicated 12 novel T2D-associated variants, of which 4 were low-frequency (MAF < 0.05), and 1 was rare and population-specific (rs147287548), near the LEP gene, which encodes for leptin, a hormone/adipokine primarily secreted by white adipose tissue and whose deficiency causes insulin resistance[37]."*

Instead of only providing details about the replication analysis for the identified rare variants, we have also added the replication results for the common and low-frequency variants (Supplementary Table 4). We have revised the entire manuscript, tables, figures, and supplementary information to reflect the updates clearly.

2. There are 8 variants in Table 1 and on line 159 in the results the authors talk about novel rare variants, but only 1, that in the LEP region, could really be described as novel and rare, because the others do not replicate, or, in the case of the R114W HNF4A variant, are well known. I think it is too much to claim the two variants with glycaemic evidence as novel findings. The p values for T2D drop from -8 and -9 to -5 and -5 on adding in the large replication samples and the p values for glucose are around 0.01 and 0.003. There are several tests going on here – several variants, glucose, HbA1c, presumably fasting insulin was also tested. I think these could be real, but equally could be false positives. I suggest some further toning down,

or just be more upfront and say that this approach has identified one really clear rare signal, and helped clarify roles of clinvar annotated variants.

We agree with the reviewer that the orthogonal evidence provided is insufficient to claim the novelty of those variants. Accordingly, we only claim novelty for the findings we could replicate in independent cohorts at adjusted for multiple comparisons significance. In addition, we have removed from the manuscript and extended data any statement concerning the variants without robust replication evidence.

Minor point

1. For readability, I suggest the authors do not start sentences with numbers.

We have corrected the manuscript accordingly.

Final Decision Letter:

10th September 2024

Dear Josep,

I am delighted to say that your manuscript "Rare variant association analysis in 51,256 type 2 diabetes cases and 370,487 controls informs the spectrum of pathogenicity of monogenic diabetes genes" has been accepted for publication in an upcoming issue of Nature Genetics.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Genetics style. Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

You will not receive your proofs until the publishing agreement has been received through our system.

Due to the importance of these deadlines, we ask that you please let us know now whether you will be difficult to contact over the next month. If this is the case, we ask you provide us with the contact information (email, phone and fax) of someone who will be able to check the proofs on your behalf, and who will be available to address any last-minute problems.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections.

You may wish to make your media relations office aware of your accepted publication, in case they consider it appropriate to organize some internal or external publicity. Once your paper has been scheduled you will receive an email confirming the publication details. This is normally 3-4 working days in advance of publication. If you need additional notice of the date and time of publication, please let the production team know when you receive the proof of your article to ensure there is sufficient time to coordinate. Further information on our embargo policies can be found here: https://www.nature.com/authors/policies/embargo.html

Before your paper is published online, we will be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that Nature Genetics is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles), then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including <a href="https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we strongly recommend that you upload the step-by-step protocols used in this manuscript to protocols.io. protocols.io is an open online resource that allows researchers to share their detailed experimental know-how. All uploaded protocols are made freely available and are assigned DOIs for ease of citation. Protocols can be linked to any publications in which they are used and will be linked to from your article. You can also establish a dedicated workspace to collect all your lab Protocols. By uploading your Protocols to protocols.io, you are enabling researchers to more readily reproduce or adapt the methodology you use, as well as increasing the visibility of your protocols and papers. Upload your Protocols at https://protocols.io. Further information can be found at https://www.protocols.io/help/publish-articles.

Sincerely,
Kyle

Kyle Vogan, PhD
Senior Editor
Nature Genetics
https://orcid.org/0000-0001-9565-9665