medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

# Genome-Wide Association Meta-Analysis Using a Recessive Model Illuminates Genetic Architecture of Type 2 Diabetes

Mark J. O'Connor [1, 2, 3, 4, 5], Alicia Huerta-Chagoya [6], Paula Cortés-Sánchez [7], Silvía Bonàs-Guarch [7], Joanne B. Cole [4, 5, 8, 9], David Torrents [7, 10], Kumar Veerapen [8, 11, 12], Niels Grarup [13], Mitja Kurki [8, 11, 12], Carsten F. Rundsten [13], Oluf Pedersen [13], Ivan Brandslund [14, 15], Allan Linneberg [16, 17], Torben Hansen [13], Aaron Leong [1, 2, 3, 4, 5, 8, 18,#], Jose C. Florez [1, 2, 3, 4, 5, 8,#], Josep M. Mercader [3, 4, 5, 8,#]

1. *Department of Medicine, Massachusetts General Hospital, Boston, MA, USA.*
2. *Endocrine Division, Massachusetts General Hospital, Boston, MA, USA.*
3. *Diabetes Unit, Massachusetts General Hospital, Boston, MA, USA.*
4. *Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA.*
5. *Programs in Metabolism and Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA.*
6. *Consejo Nacional de Ciencia y Tecnología (CONACYT), Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico.*
7. *Barcelona Supercomputing Center (BSC), Barcelona, Spain.*
8. *Department of Medicine, Harvard Medical School, Boston, MA, USA.*
9. *Center for Basic and Translations Obesity Research, Boston Children's Hospital, Boston, MA, USA.*
10. *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.*
11. *Stanley Center for Psychiatric Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA.*
12. *Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA.*
13. *Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2100, Copenhagen, Denmark.*
14. *Department of Clinical Biochemistry, Lillebaelt Hospital, Vejle, Denmark.*
15. *Institute of Regional Health Research, University of Southern Denmark, Odense, Denmark.*
16. *Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark.*
17. *Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.*
18. *Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA.*

# These authors jointly directed this work.

**Correspondence:**
Josep M. Mercader, Ph.D.
The Broad Institute of Harvard and MIT
75 Ames Street, Cambridge, MA 02142.
mercader@broadinstitute.org
Phone: 617-714-7000

**NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

## ABSTRACT

**Objective:** Most genome-wide association studies (GWAS) of complex traits are performed using models with additive allelic effects. Hundreds of loci associated with type 2 diabetes have been identified using this approach. Additive models, however, can miss loci with recessive effects, thereby leaving potentially important genes undiscovered.

**Research Design and Methods:** We conducted the largest GWAS meta-analysis using a recessive model for type 2 diabetes. Our discovery sample included 33,139 cases and 279,507 controls from seven European-ancestry cohorts including the UK Biobank. We then used two additional cohorts, FinnGen and a Danish cohort, for replication. For the most significant recessive signal, we conducted a phenome-wide association study across hundreds of traits to make inferences about the pathophysiology underlying the increased risk seen in homozygous carriers.

**Results:** We identified 51 loci associated with type 2 diabetes, including five variants with recessive effects undetected by prior additive analyses. Two of the five had minor allele frequency less than 5% and were each associated with more than doubled risk. We replicated three of the variants, including one of the low-frequency variants, rs115018790, which had an odds ratio in homozygous carriers of 2.56 (95% CI 2.05-3.19, $P=1\times10^{-16}$) and a stronger effect in men than in women (interaction $P=7\times10^{-7}$). Colocalization analysis linked this signal to reduced expression of the nearby *PELO* gene, and the signal was associated with multiple diabetes-related traits, with

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

homozygous carriers showing a 10% decrease in LDL and a 20% increase in triglycerides.

**Conclusions:** Our results demonstrate that recessive models, when compared to GWAS using the additive approach, can identify novel loci, including large-effect variants with pathophysiological consequences relevant to type 2 diabetes.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

## INTRODUCTION

Type 2 diabetes affects nearly 1 in 12 adults globally (1), but its genetic architecture is still not fully understood. Over the last decade, large genome-wide association studies (GWAS) have used additive models to identify hundreds of associated loci (2-4). Additive models are most powerful when the effect of two copies of a risk allele is twice that of one copy. This model is computationally simple and statistically powerful, but it does not always match the pattern of inheritance of Mendelian disorders, including monogenic forms of diabetes, which can be transmitted in a dominant or recessive fashion (5). Variants with recessive effects, particularly low-frequency variants, can go undetected by additive models (6), suggesting that non-additive models have the potential to generate new biological insights.

To date, a handful of studies have used recessive models to identify genetic associations with type 2 diabetes, but these have been limited by small sample sizes (6-8). Nevertheless, some promising findings have emerged. In a Greenlandic population, homozygous carriers of two copies of a nonsense mutation in the *TBC1D4* gene, which facilitates glucose transfer into skeletal muscle in the setting of insulin stimulation, were found to have a tenfold increase in diabetes risk compared to other individuals in the same population (7). Hyperglycemia due to this variant occurs postprandially, so the diagnosis of type 2 diabetes in homozygous carriers often requires an oral glucose tolerance test, creating an opportunity for precision medicine (9). More recently, members of our group conducted GWAS with non-additive models for several age-related diseases (10) and identified multiple new loci, including one rare variant (rs77704739) associated

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

with type 2 diabetes. This variant was also associated with reduced expression of the *PELO* gene, whose connection to diabetes is not well understood.

We have conducted the largest GWAS meta-analysis using a recessive model reported to date for type 2 diabetes. Over the last few years, GWAS sample sizes have grown exponentially (11), and reference panels for imputation have improved, making it easier to ascertain low-frequency variants accurately (12). To take advantage of these developments, we combined data from seven discovery cohorts and two replication cohorts to conduct the largest recessive-model GWAS yet reported for type 2 diabetes or any other disease. We identified and replicated multiple variants missed by larger additive studies, confirmed and fine-mapped the association near *PELO*, and conducted a phenome-wide association analysis to identify other impacted traits to better understand the pathophysiology underlying this novel association.

## RESEARCH DESIGN AND METHODS

### Study Population and Outcome Definition

We used data from multiple European-ancestry cohorts (Supplemental Table 1) including the UK Biobank (13), five cohorts known collectively as 70K for T2D (4), and the Mass General Brigham (MGB) Biobank (14). The UK Biobank is a sample of approximately half a million people recruited in the United Kingdom between the ages of 40 and 69 years. The 70K for T2D cohort consists of five studies with publicly available data, and the MGB Biobank consists of approximately 50,000 people recruited within a hospital system in the

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

United States. We only considered individuals whose family relatedness was lower than that of third-degree relatives.

Definitions of type 2 diabetes varied according to cohort. In the UK Biobank, for example, we used a validated algorithm designed specifically to identify cases of diabetes in that cohort (15). In the MGB Biobank, type 2 diabetes was defined according to an algorithm developed by the Biobank team (16) to have 99% positive predictive value. In the UK and MGB Biobanks, which both have a relatively low prevalence of type 2 diabetes, we excluded controls younger than 55 years, as the mean age of onset for type 2 diabetes is around 50 years (17).

Recessive Genome-Wide Meta-Analysis

Genotyping, phasing, and imputation as well as sample and variant quality control were done according to cohort-specific protocols (Supplemental Table 1). For the recessive analysis in each cohort, we controlled for age, sex, body mass index (BMI), and principal components. For the UK Biobank, we also controlled for the genotyping platform, as two different genotyping arrays were used. For one of the five cohorts within 70K for T2D (6% of the cases in our discovery sample), age and BMI were not available. In our models, we used the minor allele in Europeans – not necessarily the non-reference allele – as the recessive allele to maximize our chances of identifying variants missed by prior GWAS.

For the UK and MGB Biobanks, computations were done using Hail version 0.2 (https://hail.is), and the 70K for T2D cohort was analyzed using the program SNPTEST

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html). After generating summary statistics using a recessive model for each cohort, we used the program METAL to meta-analyze the results (18), weighting cohorts by the inverse of the standard error for each variant. Our threshold for genome-wide significance was $P=5\times10^{-8}$, and we considered signals within 0.5 megabase pairs (Mb) to be part of the same locus. For comparison, we repeated our approach using an additive model. To visually inspect each genome-wide significant locus, we used the program LocusZoom (19). We estimated the power of our recessive and additive models to detect variants acting recessively across a range of allele frequencies and effect sizes using a simulation-based approach, assuming a baseline case prevalence of 10%, similar to our case-control ratio.

Defining Novel Recessive Signals

We compared our results to the largest additive GWAS with available summary statistics (2, 3), and we defined signals as novel if they were not in significant linkage disequilibrium (LD) with a known signal ($r^2 < 0.3$). This analysis was done using R version 3.6 (https://www.R-project.org) and the R package 'LDlinkR' (20, 21). The LD information was calculated using a British reference panel (1000 Genomes Project). For each signal, we used PLINK version 1.9 (22) to calculate a dominance deviation *P* value (23) using UK Biobank data. Signals were deemed to be non-additive if this *P* was less than 0.05. To ensure that signals near the major histocompatibility complex (MHC) region were not due to contamination of our cases with cases of type 1 diabetes, which is known to be heavily associated with haplotypes in the MHC region, we performed conditional analysis in the

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

UK Biobank sample, adjusting for MHC haplotypes relevant to type 1 diabetes (24). We excluded variants that lost significance by more than one order of magnitude.

## Replication

We attempted to replicate our novel findings in two cohorts: FinnGen and a Danish cohort (Supplemental Table 2). FinnGen is a study based in Finland that combines genotyping with digital health data for over 100,000 people, and the Danish cohort consists of over 20,000 individuals (22% cases) from Denmark. The program SNPTEST was used to analyze both cohorts. We meta-analyzed the results from the replication cohorts with our initial results using the R package 'rmeta'.

## Credible Sets

For each novel variant, we identified the set of variants with 99% probability of containing the causal variant. We used a Bayesian refinement approach (25), considering variants in LD with the lead variant ($r^2 > 0.1$). Each credible set is akin to a confidence interval for the true causal variant. Within a locus, each variant is assigned an approximate Bayes factor (ABF) based on the following equation:

$$ABF = \sqrt{1-r}e^{rz^2/2}$$

where $r = 0.04/(SE^2 + 0.04)$ and $z = \beta/SE$. The beta and standard error are the estimated effect size and corresponding standard error from the recessive-model logistic regression. This calculation assumes a Gaussian prior with mean 0 and variance 0.04. The posterior probability for a variant is equal to its ABF divided by the sum of all ABF values for the locus. Variants are ranked by ABF in decreasing order, and the cumulative

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

probability is calculated starting at the top of the list and stopping when the value exceeds 99%.

## Colocalization with Gene Expression

To shed light on variants' functional consequences, we used the Genotype-Tissue Expression (GTEx) project, version 8 (26). This database links genetic variants with tissue-specific gene expression, allowing for the identification of expression quantitative trait loci (eQTLs). For our most significant variant, which was an eQTL for the gene *PELO*, we performed colocalization analysis to confirm that our GWAS signal matched the signal influencing gene expression. Colocalization analysis compares *P* values for two traits across a locus to generate a posterior probability for the hypothesis that both traits are being influenced by the same variant. Due to the rarity of homozygous carriers of our variants, we used additive summary statistics from GTex for this analysis. We used the R package 'coloc' (27) and considered a window of 1 Mb around our leading signal.

## Phenome-Wide Association Study

For the variant located near the *PELO* gene, we performed a phenome-wide association study (PheWAS) in the UK Biobank, which provides detailed information about each participant's health, dietary habits, and lifestyle characteristics. Phenotypes were curated and transformed using the PHEnome Scan ANalysis Tool, or PHESANT (28). As in our GWAS, we used a recessive model. We used logistic regression for binary phenotypes and linear regression for continuous phenotypes. We controlled for age, sex, ten principal components, and the genotyping platform. Limiting our binary phenotypes to those with

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

more than five cases among homozygotes for the risk variant, we analyzed 1,731 binary phenotypes. We also analyzed 30 biomarkers such as cholesterol levels as well as 1,345 other continuous phenotypes. For significant associations, we used colocalization analysis to quantify the probability that the phenotype shared the same causal variant as type 2 diabetes. We also performed a PheWAS in the Danish cohort looking at 16 glycemic traits, using the same covariates as above.

### Sex-Stratified Analysis

To test whether the genetic effects of the variant near the *PELO* gene differed by sex, we performed a sex-stratified analysis within the UK Biobank for the biomarkers in our dataset and also for type 2 diabetes itself. We assessed the significance of the difference between sexes by including an interaction term in our regression model. We then confirmed sex-specific differences for type 2 diabetes in our two replication cohorts.

## RESULTS

### Genome-Wide Meta-Analysis Using a Recessive Model

Our discovery sample consisted of 33,139 cases of type 2 diabetes and 279,507 controls from seven cohorts. We meta-analyzed 11,634,328 variants and fitted additive and recessive models to compare the results. We identified 51 loci (Supplemental Table 3) that reached genome-wide significance in the recessive model, and 121 loci using the additive model (Figure 1). Of the 51 signals identified with the recessive model, 33% deviated from additivity (dominance deviation $P < 0.05$), and of these, five were distinct from the set of previously reported additive signals (Table 1).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The strongest recessive signal (rs115018790) was located within an intron of the *PELO* and *ITGA1* genes on chromosome 5 (Figure 2) and was in complete LD ($r^2 = 1$) with the lead variant (rs77704739) that was previously identified in the GERA cohort (10), one of the discovery cohorts in this study. With minor allele frequency (MAF) 0.04, rs115018790 had an odds ratio (OR) for homozygous carriers of 2.63 (95% CI: 2.03-3.41), much greater than the additive-model OR of 1.07 (1.02-1.12). The *P* value for the recessive model ($P=3\times10^{-13}$) was ten orders of magnitude more significant than the additive one, and the dominance deviation test confirmed the variant's recessive nature ($P=3\times10^{-5}$). This variant was near known additive signals (rs17261179, rs3811978, and rs62357230) associated with type 2 diabetes (2), but it was not in strong LD with any of these previously identified variants (maximum $r^2 = 0.08$).

We identified another non-additive, low-frequency variant (rs140453320) with large effect size on chromosome 5. This variant (MAF=0.01, OR [95% CI] = 6.94 [3.63-13.27], $P=5\times10^{-9}$) lies within an intron of the gene *ADAMTS6*. The additive *P* value was 0.48, leading to highly significant dominance deviation ($P=4\times10^{-9}$). This signal was over 1 Mb away from any previously known signal associated with type 2 diabetes.

The other three novel non-additive signals were significantly more common, each with MAF > 30%. Two of the three were located less than 0.5 Mb from known additive loci, but these signals were in weak LD with previously reported associations, with maximum $r^2$ between 0.1 and 0.3 (Supplemental Table 4). The third (rs755900673) was an indel (OR

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

[95% CI] = 1.13 [1.08-1.17], $P=5\times10^{-9}$) on chromosome 8 located within an intron of the *MYOM2* gene, more than 7 Mb away from any locus additively associated with type 2 diabetes.

We performed power simulations for our top variant (rs115018790, MAF 0.04, OR 2.63) and found that a GWAS with an additive model with our case-control ratio would need approximately 1.8 million participants to have 80% power to detect a genome-wide significant signal whereas a recessive model would only need 160,000 participants. At higher allele frequencies, the benefits of the recessive model become much less pronounced (Supplemental Figure 1).

Replication

Our two replication cohorts consisted of 28,336 cases and 62,253 controls. Of the four non-additive signals for which we had sufficient power, three replicated, and one did not (Table 1, Supplemental Figure 2). Variant rs115018790 replicated in both cohorts (meta-analysis OR [95% CI] = 2.56 [2.05-3.19], $P=1\times10^{-16}$).

Two of the other three variants for which we had power also replicated. The indel near *MYOM2*, rs755900673, did not replicate ($P$=0.84) and showed high heterogeneity ($P$=0.008). Our power to replicate the rare variant near *ADAMTS6* was limited because there were only 10 homozygous carriers in our replication sample compared to 74 in our discovery sample. This signal retained genome-wide significance when we meta-analyzed the discovery and replication cohorts ($P=3\times10^{-8}$).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

## Gene Expression Colocalization Analysis

Using GTEx data (26), we found that rs115018790 was associated with reduced *PELO* expression in multiple tissues. Colocalization analysis, which tests the hypothesis that traits are associated and share a single causative variant, confirmed the link between rs115018790 and reduced *PELO* expression in many tissues, including subcutaneous adipose tissue (posterior probability 0.99, *n*=581), skeletal muscle (0.99, *n*=706), and the pancreas (0.96, *n*=305). Colocalization plots (Supplemental Figure 3) comparing the recessive *P* values for the association with type 2 diabetes to additive *P* values from the gene expression dataset showed a high degree of correlation between the two sets of *P* values, visually confirming rs115018790's connection to reduced *PELO* gene expression. The signal's 99% credible set (Supplemental Table 5) contains rs185240714 (posterior probability 0.23), which is located in the 5' untranslated region of the *PELO* transcription start site, further supporting a causal link.

## Phenome-wide Association Study

Using a recessive model, we found that multiple biomarkers (Figure 3) were associated with rs115018790 in the UK Biobank. Homozygotes for the risk allele had significantly higher triglycerides and lower LDL, HDL, and total cholesterol. Effect sizes were large. For example, being a homozygous carrier of the risk allele was associated with a 0.35 mmol/L (14 mg/dL) decrease in LDL (10% change relative to the mean) and a 0.35 mmol/L (31 mg/dL) increase in triglycerides (20%). These associations, particularly for triglycerides, were less significant using an additive model (Supplemental Table 6),

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

suggesting that rs115018790 acts in a recessive manner for these traits as well. Colocalization analysis (Supplemental Figure 3) confirmed that these lipid associations are the result of a single shared variant (posterior probability > 0.99 for each trait). It did not appear that medication use was responsible for the observed effects on lipids because homozygotes for the risk allele were less likely (OR [95% CI] = 0.66 [0.49-0.88], *P*=0.005) to be on LDL-lowering therapy. Other biomarkers associated with rs115018790 included albumin and C-reactive protein. There was also a nominally significant (*P*=0.01) association with estradiol. The other novel variants did not have comparably significant and numerous biomarker associations (Supplemental Table 7).

When we examined non-biomarker phenotypes (Supplemental Table 8) using a recessive model, we found that the variant near *PELO* was associated with a variety of hematologic features (decreased blood cell count, increased reticulocyte count and percent, increased mean corpuscular hemoglobin and volume, and decreased red blood cell distribution width) as well as increased alcohol intake frequency. None of the binary phenotypes reached a strict Bonferroni-corrected significance threshold of $1.6\times10^{-5}$, but the top two phenotypes were metformin use (OR [95% CI] = 2.27 [1.56-3.31], $P=2\times10^{-5}$) and "diabetes diagnosed by a doctor" (1.87 [1.39-2.54], $P=6\times10^{-5}$). We did not detect significant associations with cardiovascular phenotypes such as heart attack or stroke after correcting for multiple testing.

In the Danish cohort, our power to detect recessive associations with glycemic traits was limited due to the low number of homozygous carriers (Supplemental Table 9). None of

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

the traits were recessively associated with the variant. In an additive analysis, the variant was associated with increased insulin and C-peptide levels at the 120-minute timepoint of an oral glucose tolerance test.

Sex-stratified Analysis

Because the variant near *PELO* was nominally associated with estradiol, we performed an analysis stratified by sex and, in the case of women, by menopause status, and we found that the effect of rs115018790 on estradiol was only significant in pre-menopausal women, with homozygotes for the risk allele having higher estradiol levels (174 pmol/L [95% CI: 49-300], $P$=0.006) than other pre-menopausal women. For other biomarkers such as cholesterol and triglycerides, effects were stronger in men than in women (Figure 3, Supplemental Table 6). For example, the recessive association with triglycerides was over twelve orders of magnitude more significant in men ($P=3\times10^{-16}$) than in women ($P$=0.002).

We also performed a sex-stratified analysis for type 2 diabetes itself. In the UK Biobank, we found that the association was limited to men (OR [95% CI] = 3.05 [2.10-4.43], $P=5\times10^{-9}$) as opposed to women (0.89 [0.42-1.87], $P$=0.75), with an interaction $P$ value of $7\times10^{-7}$. We replicated this finding in our replication cohorts (Supplemental Table 10). Meta-analysis across cohorts confirmed a large effect in men (OR=2.99 [2.18-4.10], $P=1\times10^{-11}$) and little to no effect in women (OR=1.41 [0.87-2.28], $P$=0.15).

Comparison with Known Lipid-Associated Variants

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

To put effect sizes into context, we compared rs115018790 to previously described lipid-related variants of large effect size (29) using UK Biobank data. The LDL-lowering effect (0.35 mmol/L or 14 mg/dl) of rs115018790 in homozygotes was comparable to the effect (0.34 mmol/L or 13 mg/dl) of carrying one copy of a well-known protective variant (rs11591147) associated with the *PCSK9* gene. The triglycerides-increasing effect (0.35 mmol/L or 31 mg/dl) of rs115018790 in homozygotes was larger than the change (-0.29 mmol/L or -26 mg/dl) seen in homozygotes for a known variant (rs1569209) linked to the lipoprotein lipase (*LPL*) gene, known to be involved in triglyceride metabolism. For men, the size of rs115018790's effect (0.58 mmol/L or 51 mg/dl) was almost double that of the *LPL* variant.

## DISCUSSION

Type 2 diabetes is a highly polygenic trait, and hundreds of loci associated with the disease have been identified, mostly via large GWAS meta-analyses conducted under additive genetic models (2, 3). This prior work has produced useful results, identifying potential therapeutic targets and also allowing for the creation of polygenic scores capable of quantifying one's genetic risk (30). A sizeable fraction of the heritability of type 2 diabetes, however, remains unexplained by loci identified using additive models. Recessive modeling offers a way to identify new associations, creating opportunities for discovery and improved genetic risk stratification.

Our work takes advantage of the increasing number of genetic datasets now available, and it is currently the largest GWAS using a recessive model yet reported for type 2

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

diabetes or any other complex disease. We were able to identify multiple variants acting recessively, including two low-frequency variants of large effect size. Most of the variants identified via additive analyses have ORs less than 1.1, but the most significant variant we identified had an OR of 2.56 in homozygous carriers. Our minimum sample size to detect this variant was ten times smaller because we used a recessive, not an additive, model.

This variant was located near the *PELO* gene, and one of the six variants in the 99% credible set was in the gene's upstream 5' untranslated region, suggesting a role for this variant in gene expression regulation, a link we confirmed across multiple tissues using colocalization approaches. Members of our group first identified this association in one of our cohorts while conducting recessive-model GWAS for multiple age-related diseases (10). In this study, we confirmed the association with a larger sample size, fine-mapped the region, and used the power of the UK Biobank to demonstrate that the phenotypic effects of this variant are not limited to type 2 diabetes.

Homozygous carriers of the *PELO* variant exhibit significantly different circulating triglyceride and cholesterol levels compared to other individuals. These effects were most pronounced in men but were also seen in women, and the effect sizes were clinically relevant and comparable to previously discovered genetic variants that revealed novel therapeutic targets. The reduction in LDL associated with rs115018790 was approximately 10% (given an average LDL of 3.62 mmol/L or 140 mg/dl) whereas statins, the most commonly used LDL-lowering medications, typically lower LDL by 30 to 60%

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(31). As would be expected for carriers of an LDL-lowering variant, homozygotes for the minor allele at rs115018790 were less likely to be on statin medication. For triglycerides, the effect size (20%) was even larger.

The overall consequences of the effect of variant rs115018790 on lipid levels remain unclear. Low LDL is known to protect against cardiovascular events. High triglycerides and low HDL, on the other hand, are associated with cardiovascular disease, although for these two lipid particles, it is not clear whether the relationship is causal (32, 33). For homozygotes at rs115018790, the protective effects of lower LDL may be offset by the high triglycerides and lower HDL, meaning that the net effect on cardiovascular risk could be beneficial, harmful, or neutral. Our PheWAS did not reveal associations with cardiovascular events such as myocardial infarction or stroke. This lack of association, however, must be interpreted with caution in the setting of limited power, automatically curated phenotypes, and “healthy volunteer” selection bias in the UK Biobank (34).

The mechanism by which *PELO* affects diabetes risk is not clear. The gene is ubiquitously expressed, and its genetic deletion in mice leads to embryonic lethality (35). It is evolutionarily conserved and plays a role in rescuing stalled ribosomes, thus affecting the translation of multiple mRNA transcripts (36). It has a known role in sustaining protein synthesis in developing blood cells and platelets (37). A recent CRISPR loss-of-function screen in human pancreatic beta cells suggests that *PELO* may also play a role in insulin secretion (38). The sex-specific effect on diabetes risk in our study was striking, and more investigation is needed to determine what factors underlie the increased risk in men.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

One limitation of our study is the restriction of the analyses to participants of European ancestry. Estimation of recessive effects requires large sample sizes, as homozygous carriers of low-frequency variants are rare. Progress has been made in terms of recruiting diverse participants for genetic studies, but people of European ancestry still make up the bulk of available datasets. In the future, non-additive methods may yield new insights when applied to non-European populations, work that could be particularly fruitful given the increased genetic diversity of these populations (39) and the increasing availability of multi-ethnic cohorts (40).

It is worth noting that most of the associations detected in our recessive analysis had already been uncovered in prior additive GWAS. This observation matches our power simulations comparing additive and recessive models. In these simulations, the benefit of the recessive model was significant at the low end of the allele-frequency spectrum whereas both models had similar power to detect high-frequency variants with recessive effects.

Our work illustrates the value of performing non-additive analyses to uncover low-frequency recessive variants. By conducting what is currently the largest GWAS using a recessive model for type 2 diabetes, we confirmed that a variant linked to reduced *PELO* gene expression appears to have significant effects not just on diabetes but also on lipid metabolism. Recessive models of type 2 diabetes and glycemic traits as part of larger

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

and more diverse genetic discovery efforts are likely to provide additional associations that will in turn provide a better understanding of diabetes pathophysiology.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

## ACKNOWLEDGEMENTS

Funding: M.J.O. was supported by NIH/NIDDK award T32 DK110919. J.M.M. is supported by American Diabetes Association Innovative and Clinical Translational Award 1-19-ICTS-068. S.B.-G. was supported by FI-DGR Fellowship from FI-DGR 2013 from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, Generalitat de Catalunya) and by a 'Juan de la Cierva' post- doctoral fellowship (MINECO;FJCI-2017-32090). J.C.F. is supported by NIDDK K24 DK110550. This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 667191. This work was also supported by grant SEV-2011-00067 of the Severo Ochoa Program, awarded by the Spanish Government. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. This study also makes use of data generated by the UK10K Consortium, derived from samples from UK10K COHORT IMPUTATION (EGAS00001000713). A full list of the investigators who contributed to the generation of the data is available in www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310. Work in the UK Biobank was done under application numbers 27892 and 31063. 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). The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Acknowledgements: We acknowledge Bianca C. Porneala, M.S., for technical assistance in the collection and curation of the genotype and phenotype data from the MGB Biobank. We also acknowledge Josee DuPuis, Ph.D., and Peitao Wu, Ph.D., for their valuable feedback.

Author Contributions: J.M.M. and A.L. planned the analysis. M.J.O. performed the GWAS in the UK Biobank and MGB Biobank. A.H.-C. and J.B.C assisted M.J.O. with these cohorts. P.C.S., S. B.-G., D.T., and J.M.M. analyzed the 70K for T2D cohort. K.V. performed and M.K. supervised replication analysis in FinnGen, and C.F.R. performed the analysis in the Danish cohort supervised by N.G. and T.H. and with assistance from O.P., I.B., and A.L., who all contributed to assembling the data. M.J.O. performed the meta-analysis and the PheWAS. M.J.O. and J.M.M. wrote the manuscript. J.M.M., A.L., and J.C.F. supervised the work. J.M.M. is the guarantor of this work and, as such, had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Duality of Interest: J.C.F. has received consulting honoraria from Goldfinch Bio and AstraZeneca, and speaking honoraria from Novo Nordisk, AstraZeneca and Merck for research presentations over which he had full control of content. None of the other authors have conflicts of interest to declare.

Prior Presentation: Parts of this work were presented in an abstract as part of the American Diabetes Association's 80th Scientific Sessions, June 2020.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

## REFERENCES:

1. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol. 2018;14(2):88-98. Epub 2017/12/09. doi: 10.1038/nrendo.2017.151. PubMed PMID: 29219149.

2. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, Payne AJ, Steinthorsdottir V, Scott RA, Grarup N, Cook JP, Schmidt EM, Wuttke M, Sarnowski C, Magi R, Nano J, Gieger C, Trompet S, Lecoeur C, Preuss MH, Prins BP, Guo X, Bielak LF, Below JE, Bowden DW, Chambers JC, Kim YJ, Ng MCY, Petty LE, Sim X, Zhang W, Bennett AJ, Bork-Jensen J, Brummett CM, Canouil M, Ec Kardt KU, Fischer K, Kardia SLR, Kronenberg F, Lall K, Liu CT, Locke AE, Luan J, Ntalla I, Nylander V, Schonherr S, Schurmann C, Yengo L, Bottinger EP, Brandslund I, Christensen C, Dedoussis G, Florez JC, Ford I, Franco OH, Frayling TM, Giedraitis V, Hackinger S, Hattersley AT, Herder C, Ikram MA, Ingelsson M, Jorgensen ME, Jorgensen T, Kriebel J, Kuusisto J, Ligthart S, Lindgren CM, Linneberg A, Lyssenko V, Mamakou V, Meitinger T, Mohlke KL, Morris AD, Nadkarni G, Pankow JS, Peters A, Sattar N, Stancakova A, Strauch K, Taylor KD, Thorand B, Thorleifsson G, Thorsteinsdottir U, Tuomilehto J, Witte DR, Dupuis J, Peyser PA, Zeggini E, Loos RJF, Froguel P, Ingelsson E, Lind L, Groop L, Laakso M, Collins FS, Jukema JW, Palmer CNA, Grallert H, Metspalu A, Dehghan A, Kottgen A, Abecasis GR, Meigs JB, Rotter JI, Marchini J, Pedersen O, Hansen T, Langenberg C, Wareham NJ, Stefansson K, Gloyn AL, Morris AP, Boehnke M, McCarthy MI. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet. 2018;50(11):1505-13. Epub 2018/10/10. doi: 10.1038/s41588-018-0241-6. PubMed PMID: 30297969; PMCID: PMC6287706.

3. Vujkovic M, Keaton JM, Lynch JA, Miller DR, Zhou J, Tcheandjieu C, Huffman JE, Assimes TL, Lorenz K, Zhu X, Hilliard AT, Judy RL, Huang J, Lee KM, Klarin D, Pyarajan S, Danesh J, Melander O, Rasheed A, Mallick NH, Hameed S, Qureshi IH, Afzal MN, Malik U, Jalal A, Abbas S, Sheng X, Gao L, Kaestner KH, Susztak K, Sun YV, DuVall SL, Cho K, Lee JS, Gaziano JM, Phillips LS, Meigs JB, Reaven PD, Wilson PW, Edwards TL, Rader DJ, Damrauer SM, O'Donnell CJ, Tsao PS, Consortium H, Regeneron Genetics Center, V. A. Million Veteran Program, Chang KM, Voight BF, Saleheen D. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. Nat Genet. 2020;52(7):680-91. Epub 2020/06/17. doi: 10.1038/s41588-020-0637-y. PubMed PMID: 32541925; PMCID: PMC7343592.

4. Bonas-Guarch S, Guindo-Martinez M, Miguel-Escalada I, Grarup N, Sebastian D, Rodriguez-Fos E, Sanchez F, Planas-Felix M, Cortes-Sanchez P, Gonzalez S, Timshel P, Pers TH, Morgan CC, Moran I, Atla G, Gonzalez JR, Puiggros M, Marti J, Andersson EA, Diaz C, Badia RM, Udler M, Leong A, Kaur V, Flannick J, Jorgensen T, Linneberg A, Jorgensen ME, Witte DR, Christensen C, Brandslund I, Appel EV, Scott RA, Luan J, Langenberg C, Wareham NJ, Pedersen O, Zorzano A, Florez JC, Hansen T, Ferrer J, Mercader JM, Torrents D. Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. Nat Commun. 2018;9(1):321. doi: 10.1038/s41467-017-02380-9. PubMed PMID: 29358691; PMCID: PMC5778074.

5. Riddle MC, Philipson LH, Rich SS, Carlsson A, Franks PW, Greeley SAW, Nolan JJ, Pearson ER, Zeitler PS, Hattersley AT. Monogenic Diabetes: From Genetic Insights to Population-Based Precision in Care. Reflections From a Diabetes Care Editors' Expert Forum. Diabetes Care. 2020;43(12):3117-28. Epub 2021/02/10. doi: 10.2337/dci20-0065. PubMed PMID: 33560999.

6. Grarup N, Moltke I, Andersen MK, Bjerregaard P, Larsen CVL, Dahl-Petersen IK, Jorsboe E, Tiwari HK, Hopkins SE, Wiener HW, Boyer BB, Linneberg A, Pedersen O, Jorgensen ME, Albrechtsen A, Hansen T. Identification of novel high-impact recessively inherited type 2 diabetes risk variants in the Greenlandic

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

population. Diabetologia. 2018;61(9):2005-15. Epub 2018/06/22. doi: 10.1007/s00125-018-4659-2. PubMed PMID: 29926116; PMCID: PMC6096637.

7. Moltke I, Grarup N, Jorgensen ME, Bjerregaard P, Treebak JT, Fumagalli M, Korneliussen TS, Andersen MA, Nielsen TS, Krarup NT, Gjesing AP, Zierath JR, Linneberg A, Wu X, Sun G, Jin X, Al-Aama J, Wang J, Borch-Johnsen K, Pedersen O, Nielsen R, Albrechtsen A, Hansen T. A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. Nature. 2014;512(7513):190-3. doi: 10.1038/nature13425. PubMed PMID: 25043022.

8. Wood AR, Tyrrell J, Beaumont R, Jones SE, Tuke MA, Ruth KS, consortium G, Yaghootkar H, Freathy RM, Murray A, Frayling TM, Weedon MN. Variants in the FTO and CDKAL1 loci have recessive effects on risk of obesity and type 2 diabetes, respectively. Diabetologia. 2016;59(6):1214-21. Epub 2016/03/11. doi: 10.1007/s00125-016-3908-5. PubMed PMID: 26961502; PMCID: PMC4869698.

9. Manousaki D, Kent JW, Jr., Haack K, Zhou S, Xie P, Greenwood CM, Brassard P, Newman DE, Cole S, Umans JG, Rouleau G, Comuzzie AG, Richards JB. Toward Precision Medicine: TBC1D4 Disruption Is Common Among the Inuit and Leads to Underdiagnosis of Type 2 Diabetes. Diabetes Care. 2016;39(11):1889-95. Epub 2016/08/27. doi: 10.2337/dc16-0769. PubMed PMID: 27561922.

10. Guindo-Martinez M, Amela R, Bonas-Guarch S, Puiggros M, Salvoro C, Miguel-Escalada I, Carey CE, Cole JB, Rueger S, Atkinson E, Leong A, Sanchez F, Ramon-Cortes C, Ejarque J, Palmer DS, Kurki M, FinnGen Consortium, Aragam K, Florez JC, Badia RM, Mercader JM, Torrents D. The impact of non-additive genetic associations on age-related complex diseases. Nat Commun. 2021;12(1):2436. Epub 2021/04/25. doi: 10.1038/s41467-021-21952-4. PubMed PMID: 33893285.

11. Mills MC, Rahal C. A scientometric review of genome-wide association studies. Commun Biol. 2019;2:9. Epub 2019/01/10. doi: 10.1038/s42003-018-0261-x. PubMed PMID: 30623105; PMCID: PMC6323052.

12. Kowalski MH, Qian H, Hou Z, Rosen JD, Tapia AL, Shan Y, Jain D, Argos M, Arnett DK, Avery C, Barnes KC, Becker LC, Bien SA, Bis JC, Blangero J, Boerwinkle E, Bowden DW, Buyske S, Cai J, Cho MH, Choi SH, Choquet H, Cupples LA, Cushman M, Daya M, de Vries PS, Ellinor PT, Faraday N, Fornage M, Gabriel S, Ganesh SK, Graff M, Gupta N, He J, Heckbert SR, Hidalgo B, Hodonsky CJ, Irvin MR, Johnson AD, Jorgenson E, Kaplan R, Kardia SLR, Kelly TN, Kooperberg C, Lasky-Su JA, Loos RJF, Lubitz SA, Mathias RA, McHugh CP, Montgomery C, Moon JY, Morrison AC, Palmer ND, Pankratz N, Papanicolaou GJ, Peralta JM, Peyser PA, Rich SS, Rotter JI, Silverman EK, Smith JA, Smith NL, Taylor KD, Thornton TA, Tiwari HK, Tracy RP, Wang T, Weiss ST, Weng LC, Wiggins KL, Wilson JG, Yanek LR, Zollner S, North KE, Auer PL, Consortium NT-OfPM, TOPMed Hematology, Hemostasis Working Group, Raffield LM, Reiner AP, Li Y. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet. 2019;15(12):e1008500. Epub 2019/12/24. doi: 10.1371/journal.pgen.1008500. PubMed PMID: 31869403; PMCID: PMC6953885 received consulting fees from Genentech. Scott T. Weiss and Kathleen C. Barnes received royalties from UpToDate. Patrick T. Ellinor is supported by a grant from Bayer AG to the Broad Institute focused on the genetics and therapeutics of cardiovascular diseases, and has also served on advisory boards or consulted for Bayer AG, Quest Diagnostics and Novartis. Steven A Lubitz receives sponsored research support from Bristol Myers Squibb / Pfizer, Bayer HealthCare, and Boehringer Ingelheim, and has consulted for Abbott, Quest Diagnostics, Bristol Myers Squibb / Pfizer. Other authors declared no conflicts of interest.

13. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779. Epub 2015/04/01. doi: 10.1371/journal.pmed.1001779. PubMed PMID: 25826379; PMCID: PMC4380465.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

14. Karlson EW, Boutin NT, Hoffnagle AG, Allen NL. Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. J Pers Med. 2016;6(1). Epub 2016/01/20. doi: 10.3390/jpm6010002. PubMed PMID: 26784234; PMCID: PMC4810381.

15. Eastwood SV, Mathur R, Atkinson M, Brophy S, Sudlow C, Flaig R, de Lusignan S, Allen N, Chaturvedi N. Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. PLoS One. 2016;11(9):e0162388. doi: 10.1371/journal.pone.0162388. PubMed PMID: 27631769; PMCID: PMC5025160.

16. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, Cai T. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. J Am Med Inform Assoc. 2015;22(5):993-1000. Epub 2015/05/02. doi: 10.1093/jamia/ocv034. PubMed PMID: 25929596; PMCID: PMC4986664.

17. Koopman RJ, Mainous AG, 3rd, Diaz VA, Geesey ME. Changes in age at diagnosis of type 2 diabetes mellitus in the United States, 1988 to 2000. Ann Fam Med. 2005;3(1):60-3. Epub 2005/01/27. doi: 10.1370/afm.214. PubMed PMID: 15671192; PMCID: PMC1466782.

18. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1. Epub 2010/07/10. doi: 10.1093/bioinformatics/btq340. PubMed PMID: 20616382; PMCID: PMC2922887.

19. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26(18):2336-7. Epub 2010/07/17. doi: 10.1093/bioinformatics/btq419. PubMed PMID: 20634204; PMCID: PMC2935401.

20. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31(21):3555-7. Epub 2015/07/04. doi: 10.1093/bioinformatics/btv402. PubMed PMID: 26139635; PMCID: PMC4626747.

21. Myers TA, Chanock SJ, Machiela MJ. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. Front Genet. 2020;11:157. Epub 2020/03/18. doi: 10.3389/fgene.2020.00157. PubMed PMID: 32180801; PMCID: PMC7059597.

22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75. Epub 2007/08/19. doi: 10.1086/519795. PubMed PMID: 17701901; PMCID: PMC1950838.

23. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. Epub 2015/02/28. doi: 10.1186/s13742-015-0047-8. PubMed PMID: 25722852; PMCID: PMC4342193.

24. Nguyen C, Varney MD, Harrison LC, Morahan G. Definition of high-risk type 1 diabetes HLA-DR and HLA-DQ types using only three single nucleotide polymorphisms. Diabetes. 2013;62(6):2135-40. Epub 2013/02/05. doi: 10.2337/db12-1398. PubMed PMID: 23378606; PMCID: PMC3661605.

25. Wellcome Trust Case Control Consortium, Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, Howson JM, Auton A, Myers S, Morris A, Pirinen M, Brown MA, Burton PR, Caulfield MJ, Compston A, Farrall M, Hall AS, Hattersley AT, Hill AV, Mathew CG, Pembrey M, Satsangi J, Stratton MR, Worthington J, Craddock N, Hurles M, Ouwehand W, Parkes M, Rahman N, Duncanson A, Todd JA, Kwiatkowski DP, Samani NJ, Gough SC, McCarthy MI, Deloukas P, Donnelly P. Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet. 2012;44(12):1294-301. Epub 2012/10/30. doi: 10.1038/ng.2435. PubMed PMID: 23104008; PMCID: PMC3791416.

26. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, DeLuca DS, Peter-Demchok J, Gelfand ET, Guan P, Korzeniewski GE, Lockhart NC, Rabiner CA, Rao AK, Robinson KL, Roche

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

NV, Sawyer SJ, Segre AV, Shive CE, Smith AM, Sobin LH, Undale AH, Valentino KM, Vaught J, Young TR, Moore HM, GTEx Consortium. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreserv Biobank. 2015;13(5):311-9. Epub 2015/10/21. doi: 10.1089/bio.2015.0032. PubMed PMID: 26484571; PMCID: PMC4675181.

27. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10(5):e1004383. Epub 2014/05/17. doi: 10.1371/journal.pgen.1004383. PubMed PMID: 24830394; PMCID: PMC4022491.

28. Millard LAC, Davies NM, Gaunt TR, Davey Smith G, Tilling K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. Int J Epidemiol. 2018;47(1):29-35. Epub 2017/10/19. doi: 10.1093/ije/dyx204. PubMed PMID: 29040602; PMCID: PMC5837456.

29. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, Gagnon DR, DuVall SL, Li J, Peloso GM, Chaffin M, Small AM, Huang J, Tang H, Lynch JA, Ho YL, Liu DJ, Emdin CA, Li AH, Huffman JE, Lee JS, Natarajan P, Chowdhury R, Saleheen D, Vujkovic M, Baras A, Pyarajan S, Di Angelantonio E, Neale BM, Naheed A, Khera AV, Danesh J, Chang KM, Abecasis G, Willer C, Dewey FE, Carey DJ, Global Lipids Genetics Consortium, Myocardial Infarction Genetics Consortium, Geisinger-Regeneron Discov E. H. R. Collaboration, V. A. Million Veteran Program, Concato J, Gaziano JM, O'Donnell CJ, Tsao PS, Kathiresan S, Rader DJ, Wilson PWF, Assimes TL. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet. 2018;50(11):1514-23. Epub 2018/10/03. doi: 10.1038/s41588-018-0222-9. PubMed PMID: 30275531; PMCID: PMC6521726.

30. Udler MS, McCarthy MI, Florez JC, Mahajan A. Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine. Endocr Rev. 2019;40(6):1500-20. Epub 2019/07/20. doi: 10.1210/er.2019-00088. PubMed PMID: 31322649; PMCID: PMC6760294.

31. Jones P, Kafonek S, Laurora I, Hunninghake D. Comparative dose efficacy study of atorvastatin versus simvastatin, pravastatin, lovastatin, and fluvastatin in patients with hypercholesterolemia (the CURVES study). Am J Cardiol. 1998;81(5):582-7. Epub 1998/03/26. doi: 10.1016/s0002-9149(97)00965-x. PubMed PMID: 9514454.

32. Sarwar N, Danesh J, Eiriksdottir G, Sigurdsson G, Wareham N, Bingham S, Boekholdt SM, Khaw KT, Gudnason V. Triglycerides and the risk of coronary heart disease: 10,158 incident cases among 262,525 participants in 29 Western prospective studies. Circulation. 2007;115(4):450-8. Epub 2006/12/28. doi: 10.1161/CIRCULATIONAHA.106.637793. PubMed PMID: 17190864.

33. Rosenson RS. The High-Density Lipoprotein Puzzle: Why Classic Epidemiology, Genetic Epidemiology, and Clinical Trials Conflict? Arterioscler Thromb Vasc Biol. 2016;36(5):777-82. Epub 2016/03/12. doi: 10.1161/ATVBAHA.116.307024. PubMed PMID: 26966281.

34. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R, Allen NE. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol. 2017;186(9):1026-34. Epub 2017/06/24. doi: 10.1093/aje/kwx246. PubMed PMID: 28641372; PMCID: PMC5860371.

35. Nyamsuren G, Kata A, Xu X, Raju P, Dressel R, Engel W, Pantakani DV, Adham IM. Pelota regulates the development of extraembryonic endoderm through activation of bone morphogenetic protein (BMP) signaling. Stem Cell Res. 2014;13(1):61-74. Epub 2014/05/20. doi: 10.1016/j.scr.2014.04.011. PubMed PMID: 24835669.

36. Liakath-Ali K, Mills EW, Sequeira I, Lichtenberger BM, Pisco AO, Sipila KH, Mishra A, Yoshikawa H, Wu CC, Ly T, Lamond AI, Adham IM, Green R, Watt FM. An evolutionarily conserved ribosome-rescue pathway maintains epidermal homeostasis. Nature. 2018;556(7701):376-80. Epub 2018/04/13. doi: 10.1038/s41586-018-0032-3. PubMed PMID: 29643507.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

37. Mills EW, Wangen J, Green R, Ingolia NT. Dynamic Regulation of a Ribosome Rescue Pathway in Erythroid Cells and Platelets. Cell Rep. 2016;17(1):1-10. Epub 2016/09/30. doi: 10.1016/j.celrep.2016.08.088. PubMed PMID: 27681415; PMCID: PMC5111367.

38. Grotz AK, Navarro-Guerrero E, Bevacqua RJ, Baronio R, Thomsen SK, Nawaz S, Rajesh V, Wesolowska-Andersen A, Kim SK, Ebner D, Gloyn AL. 2021. A genome-wide CRISPR screen identifies regulators of beta cell function involved in type 2 diabetes risk. medRxiv doi: 10.1101/2021.05.28.445984.

39. Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526(7571):68-74. Epub 2015/10/04. doi: 10.1038/nature15393. PubMed PMID: 26432245; PMCID: PMC4750478.

40. All of Us Research Program Investigators, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, Jenkins G, Dishman E. The "All of Us" Research Program. N Engl J Med. 2019;381(7):668-76. Epub 2019/08/15. doi: 10.1056/NEJMsr1809937. PubMed PMID: 31412182.

| | | | | | OR (95% CI), *P* | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Discovery | | Replication | Combined | |
| **Variant (Position)** | **Major Allele** | **Minor Allele** | **MAF** | **Nearest Gene** | **Additive** | **Recessive** | | | **Dominance Deviation P** |
| rs115018790 (5:52088271) | G | A | 0.04 | *PELO* | 1.07 (1.02-1.12), 0.004 | 2.63 (2.03-3.41), $3\times10^{-13}$ | 2.37 (1.55-3.62), $6\times10^{-5}$ | 2.56 (2.05-3.19), $1\times10^{-16}$ | $3\times10^{-5}$ |
| rs140453320 (5:64485239) | T | C | 0.01 | *ADAMTS6* | 1.03 (0.95-1.11), 0.49 | 6.94 (3.63-13.27), $5\times10^{-9}$ | 1.18 (0.23-6.07), 0.84 | 5.46 (2.99-9.98), $3\times10^{-8}$ | $4\times10^{-9}$ |
| rs2714337 (6:7240577) | A | T | 0.35 | *RREB1* | 1.06 (1.04-1.08), $1\times10^{-10}$ | 1.12 (1.08-1.16), $3\times10^{-9}$ | 1.10 (1.06-1.15), $5\times10^{-6}$ | 1.11 (1.08-1.14), $7\times10^{-14}$ | 0.02 |
| rs755900673 (8:2008956) | TC | T | 0.36 | *MYOM2* | 1.04 (1.02-1.06), $3\times10^{-5}$ | 1.13 (1.08-1.17), $5\times10^{-9}$ | 1.01 (0.96-1.06), 0.64 | 1.08 (1.05-1.12), $2\times10^{-6}$ | 0.001 |
| rs33932777 (10:12311465) | G | A | 0.50 | *CDC123* | 1.06 (1.04-1.08), $8\times10^{-11}$ | 1.11 (1.07-1.14), $9\times10^{-12}$ | 1.07 (1.03-1.11), $4\times10^{-5}$ | 1.09 (1.07-1.12), $4\times10^{-15}$ | 0.04 |

**Table 1: Novel recessively acting variants.** Position is from genome assembly GRCh37 (hg19). For replication, variant rs140453320 was only assessed in FinnGen as there were only three homozygotes in the Danish cohort. Dominance deviation *P* values were calculated in the UK Biobank.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint
**(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

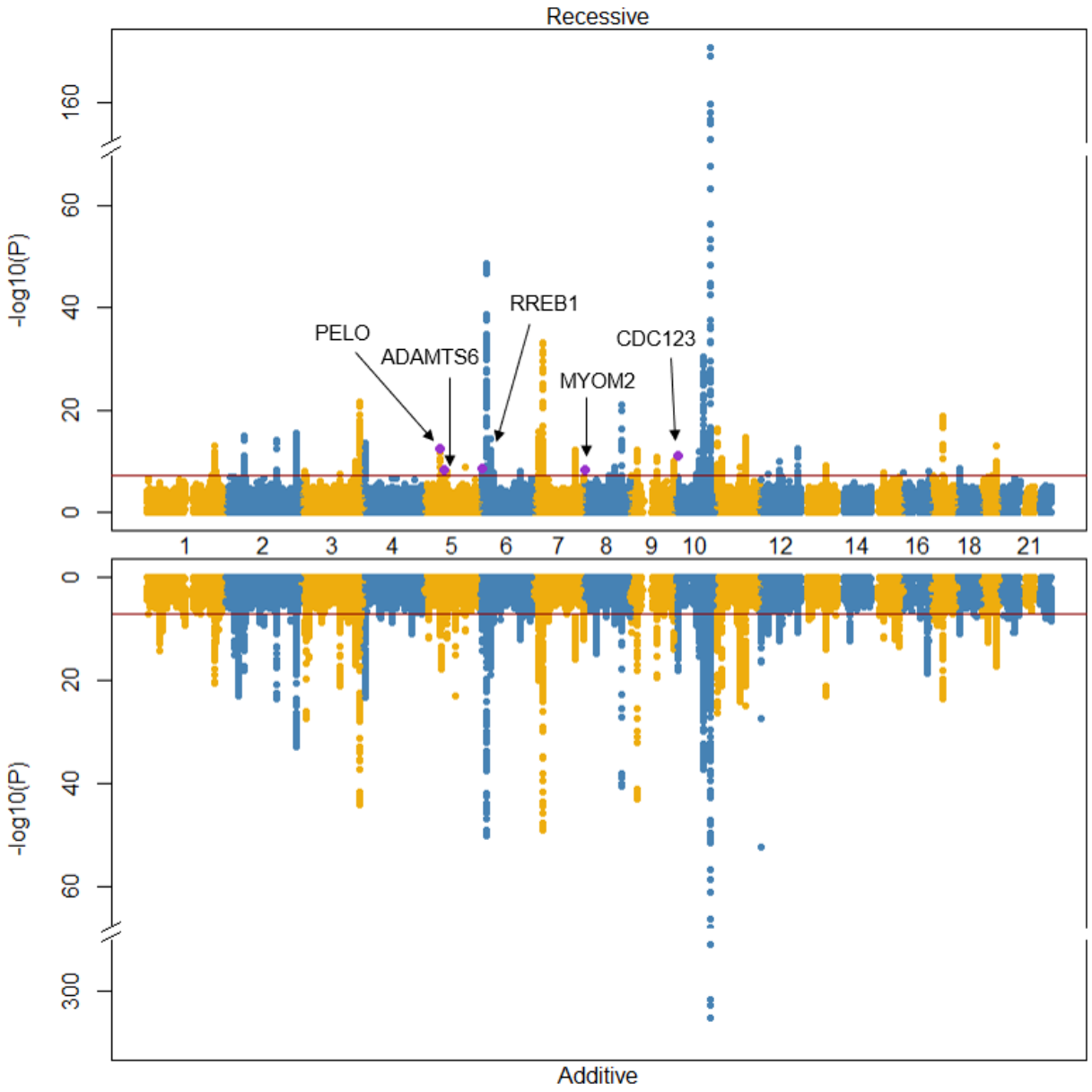


**Figure 1: Miami plot comparing recessive and additive results.** Non-additive signals are purple and labeled. The dark red line is the threshold for genome-wide significance.

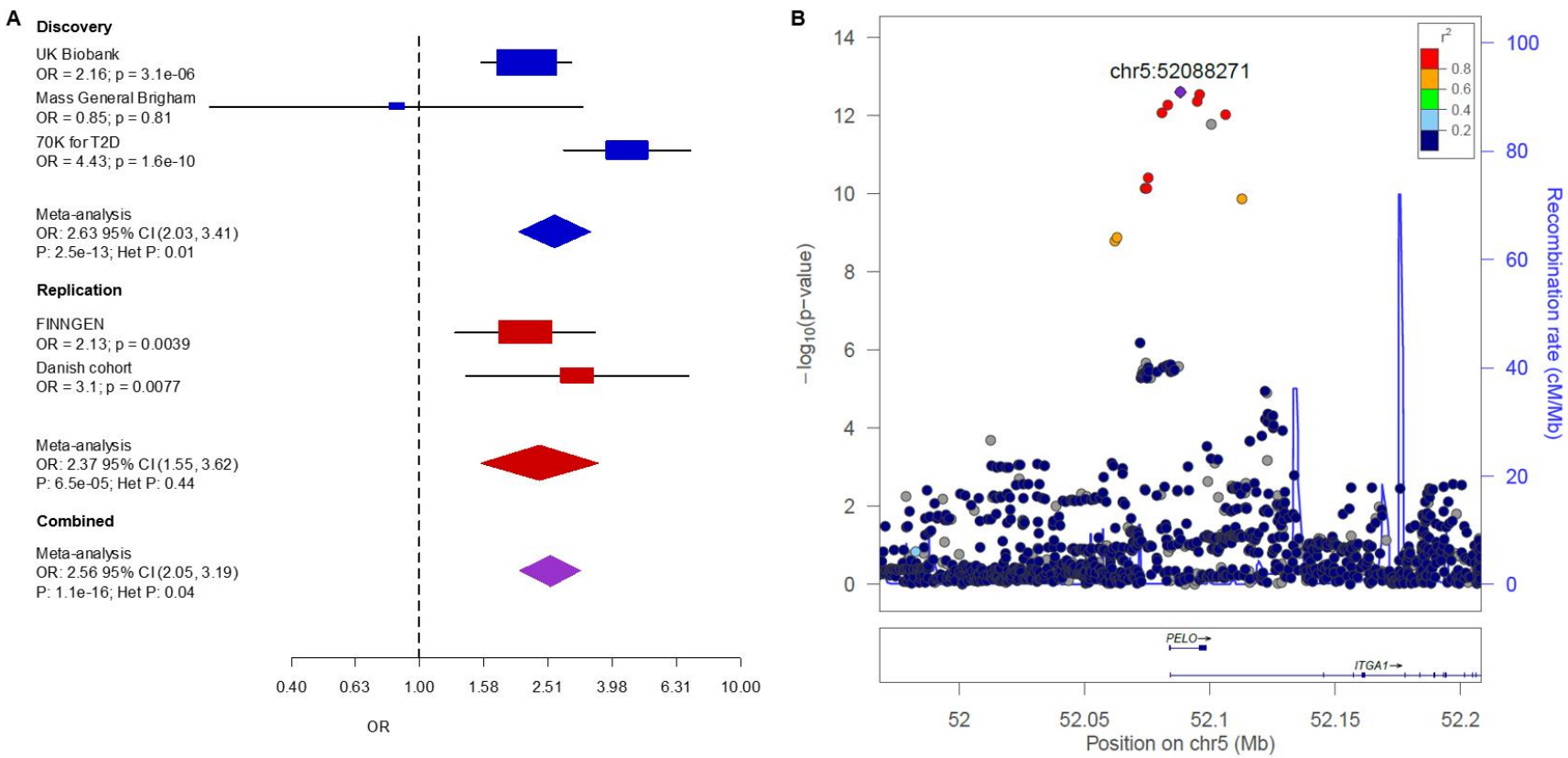


**Figure 2. Replication of variant rs115018790.** Panel A shows a forest plot of the discovery and replication cohorts. Cohort-specific odds ratios are denoted by boxes proportional to the size of the cohort, and error bars represent the 95% confidence interval. Panel B shows discovery GWAS *P* values at the *PELO* locus. Each dot represents a variant, with its genomic position (hg19) on the x axis and its *P* value (-log10) on the y-axis. Nearby genes are shown at the bottom of the plot.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.08.21258700; this version posted July 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

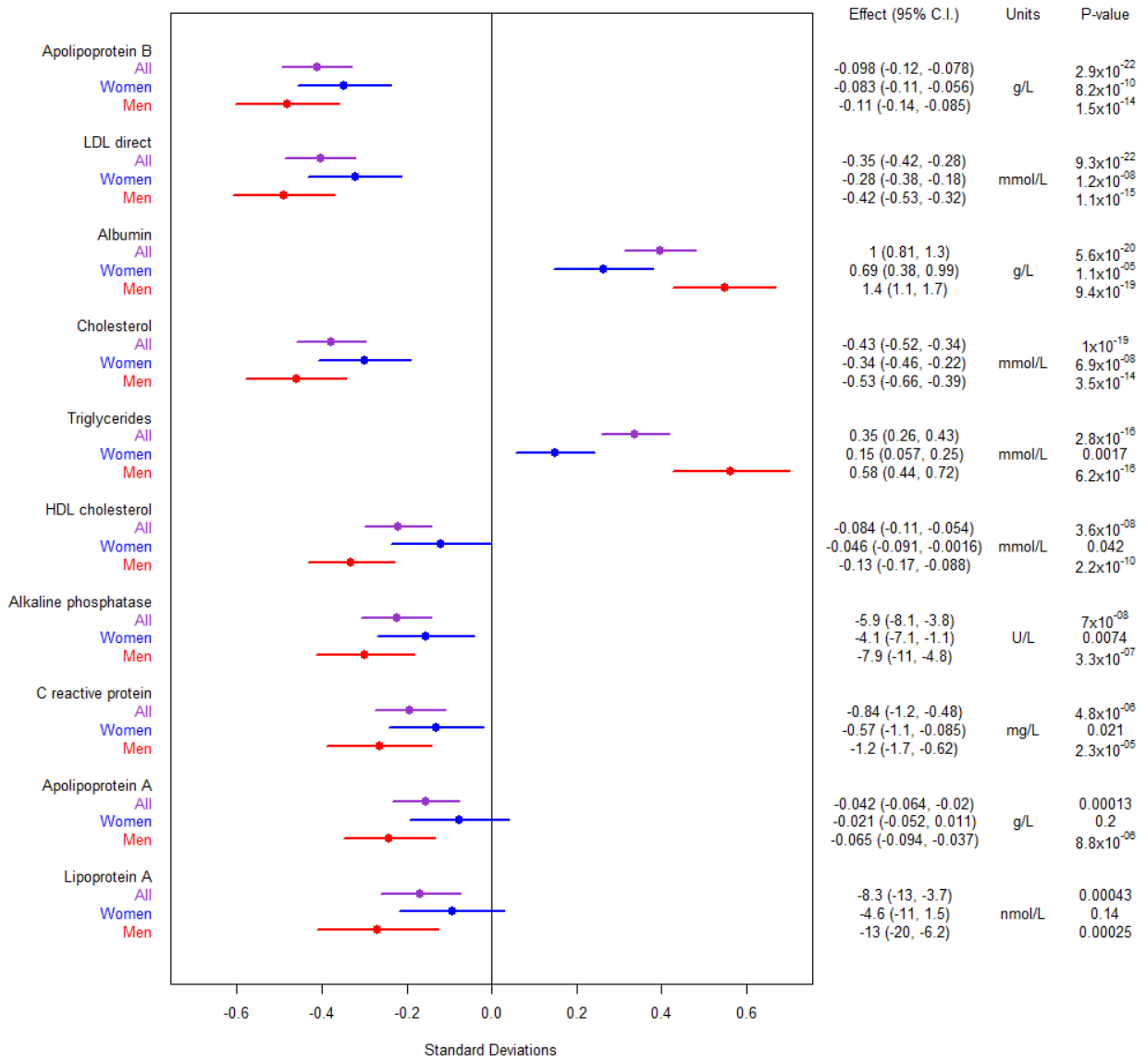


**Figure 3: Biomarker associations for variant rs115018790.** The figure to the left shows effect sizes normalized by each trait's standard deviation. The error bars in the figure to the left represent 95% confidence intervals.