



# The contribution of functional *HNF1A* variants and polygenic susceptibility to risk of type 2 diabetes in ancestrally diverse populations

Lauren A. Stalbow<sup>1,2,3</sup> · Michael H. Preuss<sup>1,2,3</sup> · Roelof A. J. Smit<sup>1,2,3,4</sup> · Nathalie Chami<sup>1,2,3</sup> · Lise Bjørkhaug<sup>5</sup> · Ingvild Aukrust<sup>6,7</sup> · Anna L. Gloyn<sup>8,9</sup> · Ruth J. F. Loos<sup>1,2,3,10</sup>

Received: 6 June 2022 / Accepted: 8 August 2022 / Published online: 11 October 2022  
© The Author(s) 2022

## Abstract

**Aims/hypothesis** We examined the contribution of rare *HNF1A* variants to type 2 diabetes risk and age of diagnosis, and the extent to which their impact is affected by overall genetic susceptibility, across three ancestry groups.

**Methods** Using exome sequencing data of 160,615 individuals of the UK Biobank and 18,797 individuals of the BioMe Biobank, we identified 746 carriers of rare functional *HNF1A* variants (minor allele frequency  $\leq 1\%$ ), of which 507 carry variants in the functional domains. We calculated polygenic risk scores (PRSs) based on genome-wide association study summary statistics for type 2 diabetes, and examined the association of *HNF1A* variants and PRS with risk of type 2 diabetes and age of diagnosis. We also tested whether the PRS affects the association between *HNF1A* variants and type 2 diabetes risk by including an interaction term.

**Results** Rare *HNF1A* variants that are predicted to impair protein function are associated with increased risk of type 2 diabetes in individuals of European ancestry (OR 1.46,  $p=0.049$ ), particularly when the variants are located in the functional domains (OR 1.89,  $p=0.002$ ). No association was observed for individuals of African ancestry (OR 1.10,  $p=0.60$ ) or Hispanic-Latino ancestry (OR 1.00,  $p=1.00$ ). Rare functional *HNF1A* variants were associated with an earlier age at diagnosis in the Hispanic-Latino population ( $\beta=-5.0$  years,  $p=0.03$ ), and this association was marginally more pronounced for variants in the functional domains ( $\beta=-5.59$  years,  $p=0.03$ ). No associations were observed for other ancestries (African ancestry  $\beta=-2.7$  years,  $p=0.13$ ; European ancestry  $\beta=-3.5$  years,  $p=0.20$ ). A higher PRS was associated with increased odds of type 2 diabetes in all ancestries (OR 1.61–2.11,  $p<10^{-5}$ ) and an earlier age at diagnosis in individuals of African ancestry ( $\beta=-1.4$  years,  $p=3.7 \times 10^{-6}$ ) and Hispanic-Latino ancestry ( $\beta=-2.4$  years,  $p<2 \times 10^{-16}$ ). Furthermore, a higher PRS exacerbated the effect of the functional *HNF1A* variants on type 2 diabetes in the European ancestry population ( $p_{\text{interaction}}=0.037$ ).

**Conclusions/interpretation** We show that rare functional *HNF1A* variants, in particular those located in the functional domains, increase the risk of type 2 diabetes, at least among individuals of European ancestry. Their effect is even more pronounced in individuals with a high polygenic susceptibility. Our analyses highlight the importance of the location of functional variants within a gene and an individual's overall polygenic susceptibility, and emphasise the need for more genetic data in non-European populations.

✉ Ruth J. F. Loos  
ruth.loos@sund.ku.dk

<sup>1</sup> The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

<sup>2</sup> The Genetics of Obesity and Related Metabolic Traits Program, Icahn School of Medicine at Mount Sinai, New York, NY, USA

<sup>3</sup> The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

<sup>4</sup> Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, the Netherlands

<sup>5</sup> Department of Safety, Chemistry and Biomedical Laboratory Sciences, Western Norway University of Applied Sciences, Bergen, Norway

<sup>6</sup> Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway

<sup>7</sup> Department of Medical Genetics, Haukeland University Hospital, Bergen, Norway

<sup>8</sup> Department of Pediatrics, Division of Endocrinology, Stanford University School of Medicine, Stanford, CA, USA

<sup>9</sup> Stanford Diabetes Research Center, Stanford University School of Medicine, Stanford, CA, USA

<sup>10</sup> The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medicine, University of Copenhagen, Copenhagen, Denmark

## Research in context

### What is already known about this subject?

- The role of rare functional *HNF1A* variants in type 2 diabetes risk remains inconclusive
- There is some evidence that the level of functional impairment of the variants as well as the location within the gene determine whether or not a variant affects risk of type 2 diabetes
- For other complex diseases, it has been shown that an individual's polygenic susceptibility affects the association of rare variants with disease risk

### What is the key question?

- What is the effect of rare functional *HNF1A* variants on the risk of type 2 diabetes, and does their location in the gene and the overall polygenic susceptibility affect that relationship?

### What are the new findings?

- Individuals of European ancestry who carry functional *HNF1A* variants are at higher risk of type 2 diabetes, particularly if they carry variants located in the functional domains
- A high polygenic susceptibility, assessed using a polygenic risk score, exacerbates the effect of *HNF1A* variants on type 2 diabetes risk
- There is a need for more genetic association studies in populations of non-European ancestry, with a focus on rare variants, as these are more likely to be population specific

### How might this impact on clinical practice in the foreseeable future?

- The improved characterisation of functional *HNF1A* variants may to help identify individuals who are at increased risk of diabetes before diagnosis, allowing time for implementation of preventive strategies

**Keywords** Age at diagnosis · Functional domain · *HNF1A* · Interaction effects · Polygenic risk scores · Risk stratification · Type 2 diabetes

### Abbreviations

|       |  |
|-------|--|
| HNF1A | Hepatocyte nuclear factor 1 homeobox A |
| PRS   | Polygenic risk score                   |
| GWAS  | Genome-wide association study          |
| PC    | Principal component                    |

## Introduction

Over the past 15 years, genome-wide association studies (GWASs) have identified more than 700 independent genetic loci that are associated with type 2 diabetes [1–3], of which the vast majority were first identified in individuals of European ancestry [1]. Summary statistics from GWASs have been used to quantify an individual's overall genetic susceptibility to type 2 diabetes, by aggregating the risk-increasing alleles, weighted by the effect size, into a polygenic risk score (PRS). People with a high PRS (top 3.5%) have been shown

to have a more than threefold increased risk of type 2 diabetes, which is comparable to the risk of diabetes in individuals who carry a monogenic variant [4].

While the majority of variants identified so far are common and mostly located in untranslated/non-coding regions of the genome, some variants are rare and protein-encoding, and located in genes that are directly involved in diabetes-related pathways [5]. Some of these genes have been previously implicated in development of monogenic diabetes. For example, rare variants in the hepatocyte nuclear factor 1 homeobox A gene (*HNF1A*), which encodes a transcription factor involved in pancreatic beta cell development and function, are responsible for 30–65% of all diagnoses of maturity-onset diabetes of the young [6]. The role of *HNF1A* variants in type 2 diabetes in the general population remains unclear. In an exome-wide analysis in 3756 Mexican and US Latino individuals, p.E508K in *HNF1A* was the only variant significantly associated with risk of type 2 diabetes (OR 5.48) [7, 8]. The p.G319S variant, which is unique to the Oji-Cree people of Northern Ontario (Canada), was found to increase the risk of

type 2 diabetes in 451 individuals (OR 1.97 for heterozygote carriers and 4.00 for homozygote carriers) [7, 8]. However, a large-scale study of almost 75,000 individuals of European ancestry that examined pathogenic variants across *HNF1A* found no evidence of association with type 2 diabetes. Another study that combined data from three cohorts ( $n=4115$ ) found no association between any of the 27 *HNF1A* variants and type 2 diabetes, except when analyses were restricted to the 11 variants that reduced transcriptional activity to <60% of normal activity (OR 5.04) [9].

Whether the effect of these rare *HNF1A* variants on type 2 diabetes is exacerbated or attenuated by an individual's overall genetic susceptibility to type 2 diabetes, as shown for other conditions [10], has so far not been reported.

Here, we examine the association between rare functionally damaging *HNF1A* variants and type 2 diabetes in the ancestrally diverse Mount Sinai BioMe Biobank and the less diverse population-based UK Biobank, together comprising almost 180,000 individuals. We compare the effect of all functionally damaging variants identified in *HNF1A* with that of variants located within a functional domain. We also examine how an individual's polygenic susceptibility, assessed using a PRS, affects the impact of rare *HNF1A* variants on type 2 diabetes.

## Methods

### Study participants

**The Mount Sinai BioMe Biobank** The Mount Sinai BioMe Biobank, founded in 2007, is an ongoing electronic health record-linked biorepository that enrolls participants non-selectively from the Mount Sinai Health System, comprising approximately 60,000 participants. The Mount Sinai Health System serves a diverse group of communities in the greater New York City area. Participants are between 18 and 89 years of age, with a broad racial and ethnic diversity (24% African, 32% European, 35% Hispanic-Latino, 9% other ancestries). At enrolment, participants consent to linkage of their DNA and plasma samples to their de-identified electronic health records. The clinical and electronic health record information is complemented by a questionnaire that gathers demographic and lifestyle information. The median number of clinical encounters for BioMe participants is 21. In the present study, data from 5244 participants of African ancestry, 6107 participants of European ancestry and 7446 participants of Hispanic-Latino ancestry were used (total 18,797), after restricting participation to those individuals with both exome sequencing and array data available, were of the three largest ancestries (African, European and Hispanic-Latino ancestries; self-reported), were considered a case or control for type 2 diabetes, had BMI data available, and were not first- or second-degree relatives of each other, as determined using

KING software (<https://www.kingrelatedness.com>) [11] (see electronic supplementary material [ESM] Fig. 1).

**The UK Biobank** The UK Biobank is a large population-based prospective cohort study from the UK comprising genotypic and phenotypic data on approximately 500,000 individuals, aged 40–69 years at enrolment [12]. The participants are predominantly of European ancestry (90.3%), and the remainder are of Asian (2.5%), African (1.8%) or other ancestries (5.4%), as determined by genetic ancestry analysis. At enrolment, participants provided baseline information and biological samples, and answered questionnaires that collected health and lifestyle information. In the present study, exome sequencing data from 3433 participants of African ancestry and 157,182 participants of European ancestry were used (total 160,615), after restricting participation to: those individuals who had both exome sequencing and array data available; were of African or European ancestries; were considered a case or control for type 2 diabetes; had BMI data available; and were not first- or second-degree relatives of each other, as determined using KING software [11] (ESM Fig. 1).

The North West–Haydock Research Ethics Committee approved the UK Biobank study (REC reference 11/NW/0382), and the current analysis was carried out under UK Biobank application 1251.

### Genotyping and quality control

**The Mount Sinai BioMe Biobank** BioMe participants ( $n=32,595$ ) were genotyped using the Illumina Global Screening Array (GSA, USA) version 1.0 platform. Individuals were removed if the sample call rate was <95%, or if the heterozygosity rate was outside 6 SDs of the mean ( $p < 1 \times 10^{-5}$  in those of African and European ancestry, or  $p < 1 \times 10^{-13}$  in those of Hispanic-Latino ancestry) ( $n=684$ ). We removed data for individuals with discordant or missing data on sex ( $n=88$ ) and any duplicates ( $n=102$ ). Missing genotypes were imputed on the Michigan Imputation server pipeline using the TOPMed freeze 5 variants as the reference panel [13].

The Regeneron Genetics Center (Tarrytown, NY, USA) generated exome sequencing files from 31,591 BioMe participants, using the Illumina NovaSeq 6000 platform. Samples that had low coverage, were genotype–exome-discordant or sex-discordant or were duplicates were removed. After quality control measures, 30,813 samples were available for analysis.

**The UK Biobank** A total of 487,409 participants (97%) were genotyped using genome-wide genotyping arrays. Among these, a subset of approximately 50,000 participants were genotyped using the Applied Biosystems KBB Lung Exome

Variant Evaluation (UK BiLEVE) Axiom Array (Affymetrix). The remaining participants were genotyped using the related Applied Biosystems UK Biobank Axiom Array.

Exome sequencing was performed for 200,633 participants (initially sequencing data from 49,960 participants followed by the remainder) using the Illumina NovaSeq 6000 platform [14]. The data that were used were the Original Quality Functional Equivalent (OQFE), and SNPs were restricted to those that met published criteria [15]. We further restricted to SNPs that had a read depth >10, genotype quality >20 and Phred-scaled likelihoods >20.

## Phenotypes

**The Mount Sinai BioMe Biobank** We identified individuals with a type 2 diabetes diagnosis using an electronic phenotyping algorithm developed by the Electronic Medical Records and GENomics (eMERGE) consortium [16, 17]. In brief, the presence in the patient's record of a type 2 diabetes-related ICD-9 (<http://www.icd9data.com/2007/Volume1/default.htm>) or ICD-10 (<https://icd.who.int/browse10/2016/en>) code in combination with either (1) prescription of insulin or other glucose-lowering medications or (2) an  $HbA_{1c} > 6.5\%$  (48 mmol/mol), was necessary to qualify as a case. In the BioMe Biobank, the age of diabetes diagnosis was defined as the age at which the participant met the criteria for the type 2 diabetes algorithm. For this analysis, we calculated the median BMI ( $\text{kg/m}^2$ ) across all outpatient encounters where BMI was measured, filtering out outlying and pregnancy-related measurements [18]. Age, ancestry and sex were self-reported. A total of 5244 (28%) individuals of African ancestry, 6107 (32%) individuals of European ancestry and 7446 (40%) individuals of Hispanic-Latino ancestry were included in the analyses, of whom 1720 (33%), 672 (11%) and 2596 (35%), respectively, had been diagnosed with type 2 diabetes.

**UK Biobank** The presence of type 2 diabetes at the time of enrolment was defined using the algorithm described by Eastwood et al [19], involving diabetes diagnosis, type, medications and complications, as well as age at diagnosis. Age of diabetes diagnosis and sex were self-reported. BMI was calculated at the time of enrolment. Ancestry was defined using the first four genetic principal components (PCs) of the genotyped dataset with *k*-means clustering (*k*=4). A total of 3433 (2%) individuals of African ancestry and 157,182 (98%) individuals of European ancestry were included, of whom 374 (11%) and 6621 (4%), respectively, were diagnosed with type 2 diabetes.

## *HNFI1A* variant classification

We used Variant Effect Predictor version 96.0 [20] to identify all rare non-synonymous variants in *HNFI1A* (transcript

NM\_000545.8) that had a high or moderate impact on protein function (transcript ablation, splice acceptor variant, splice donor variant, stop gained, frameshift variant, stop lost, start lost, transcript amplification, in-frame insertion, in-frame deletion, missense variant and protein altering variant). Rare variants were defined as those that had a minor allele frequency  $\leq 1\%$  in any ancestry in the Genome Aggregation Database (gnomAD). Importantly, we retained variants that have previously been shown to impair luciferase-based transactivation or nuclear localisation to  $\leq 60\%$  of wild-type function [9, 21–25] (ESM Fig. 2 and ESM Table 1). We restricted the data on the luciferase assays to those performed in the HeLa cell line, which lacks endogenous *HNFI1A* expression [26]. We considered participants who carried one of these variants as a carrier. No individuals carried more than one variant. We further classified the functional variants into those that were or were not located in one of the domains required for function: the  $\text{NH}_2$ -terminal dimerisation domain (amino acids 1–32), the DNA-binding domain (amino acids 91–281) and the COOH-terminal transactivation domain (282–631) [27]. We use the terms 'functional variant' or 'any variant' to describe rare non-synonymous variants in *HNFI1A* with reduced activity, and 'functional domain variants' to describe those variants that fall within one of the *HNFI1A* functional domains.

## Polygenic risk scores

An individual's overall genetic susceptibility to type 2 diabetes was assessed using PRSs. We used the PRS-CS software [28] to calculate trans-ancestry and European-ancestry PRSs. Summary statistics from the GWAS by Scott et al [29] ( $n=159,208$ ; DIAGRAM Consortium) were used to calculate a PRS for the UK Biobank European ancestry population. Summary statistics from the trans-ancestry GWAS by Vujkovic et al [2] ( $n=1,407,282$ ) were used to calculate a trans-ancestry PRS for all BioMe ancestries, and for the African ancestry group of the UK Biobank. Summary statistics were downloaded from dbGaP (Vujkovic) [30] (study accession: phs001672.v1.p1) and the DIAGRAM Consortium website (Scott) [31]. The trans-ancestry GWAS summary statistics could not be used for the European ancestry group of the UK Biobank, because the Vujkovic et al GWAS already included the UK Biobank European ancestry data, which would potentially lead to inflation of the results [32]. We removed BioMe summary statistics from the Scott et al GWAS, to avoid overlap, using MetaSubtract [33].

For the trans-ancestry PRS, default parameters were used as defined by the PRS-CS software [27], which does not pre-specify the global shrinkage parameter, allowing the software to specify it ('auto'). For the European-ancestry PRS, the global shrinkage parameter,  $\phi$ , was set to  $1e-2$ , as recommended by the PRS-CS developers for highly polygenic traits. SNPs were restricted to those with an imputation quality greater than 0.4 and a minor allele frequency above 0.1%.



Summary statistics from SNPs in the *HNF1A* region were excluded before calculating the PRS. We summed the scores using the score command from PLINK2 [34] (<https://www.cog-genomics.org/plink/2.0/>).

For the European-ancestry PRS, we used the publicly available European Linkage Disequilibrium reference panel, developed by the PRS-CS developers, based on the 1000 Genomes Project phase 3 [28]. For the trans-ancestry PRS, we used a trans-ancestry reference panel [35] that was developed based on the 1000 Genomes Project phase 3 [35]. A total of 1,223,016 (BioMe) and 1,097,294 (UK Biobank) variants were included in the trans-ancestral PRSs, and 1,118,835 variants in the European-ancestry PRS.

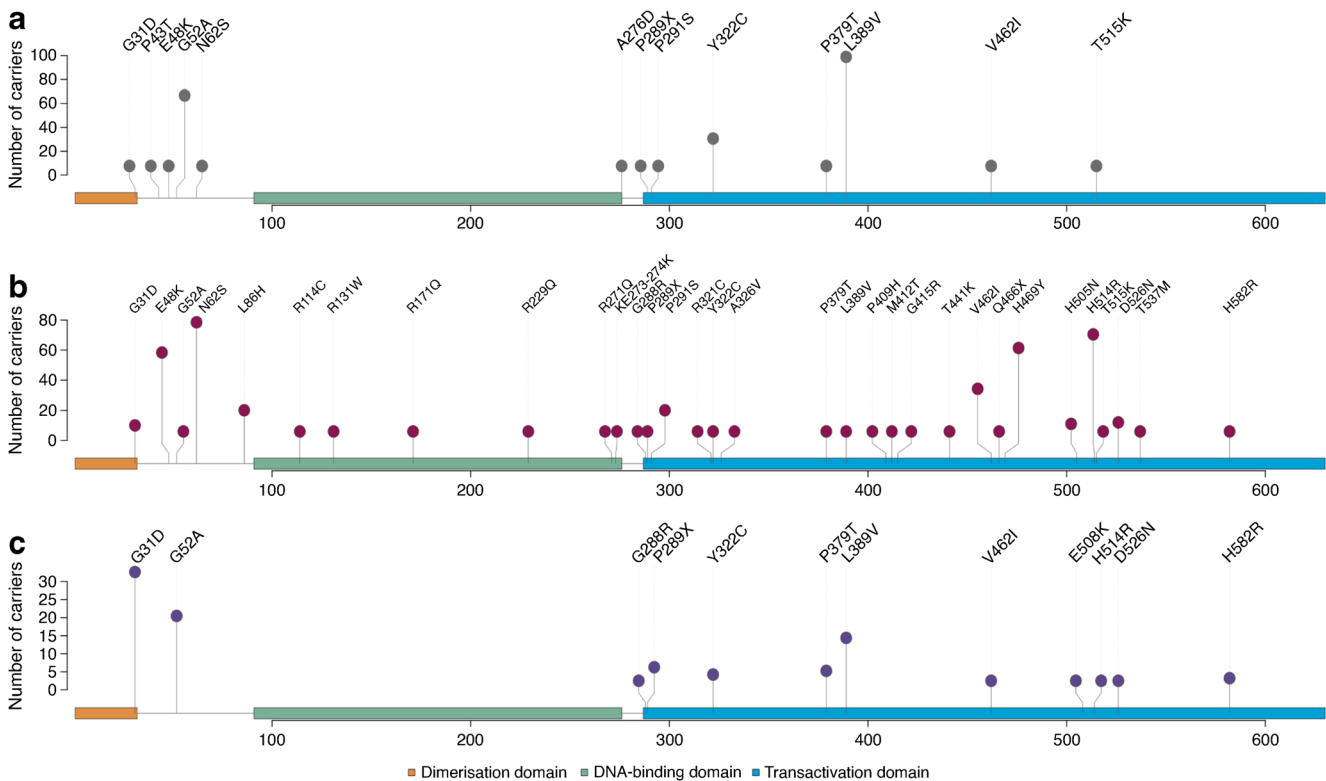
## Statistical methods

Before conducting any analyses, we standardised each of the PRSs to mean 0 and SD 1 in each ancestry group separately. To calculate the goodness-of-fit of the PRS, we generated a Nagelkerke  $R^2$  [36] and a liability threshold score [37]. We used both statistical measures because the Nagelkerke  $R^2$  is affected by disease prevalence, whereas liability threshold scores rely on accurate population prevalence, which can be difficult to ascertain. We ascertained a population prevalence from the 2017–2018 National Health and Nutrition Examination Survey (NHANES, 2017–2018) for BioMe data using the code

DIQ010 (Doctor told you have diabetes) [38]. We then performed age adjustment using the WHO age distribution. For the UK Biobank data, we used prevalence estimates reported for the UK [39].

We performed logistic regression analyses to assess the association of *HNF1A* variant carrier status and PRSs (standardised-by-ancestry) with risk of type 2 diabetes, adjusting for age, sex, BMI and the first ten genetic PCs (plus assessment centre and genotyping chip in UK Biobank), for each ancestry group separately. The reference group comprises individuals who do not carry any functionally damaging *HNF1A* variant. We then tested whether the PRS affects the association between *HNF1A* carrier status and type 2 diabetes by including an interaction term (carrier status  $\times$  PRS). We grouped the PRS into quintiles (1–20%, 20–40%, 40–60%, 60–80%, 80–100%) for visualisation. We performed all regression analyses using either carrier status of (1) any rare *HNF1A* functionally damaging variant or (2) only rare functionally damaging variants within the functional domains of *HNF1A*. We repeated all analyses for age of diabetes diagnosis, using linear regression models.

All results were meta-analysed by ancestry using the Meta package in R [40]. Random-effects statistics are reported for all the African and European ancestry analyses where the biobank data is combined. A  $p$  value of  $<0.05$  was considered statistically significant.



**Fig. 1** Position of functionally damaging *HNF1A* variants in the *HNF1A* protein sequence identified in the three ancestry groups: (a) African, (b) European and (c) Hispanic-Latino. The illustrations show the number of

carriers of each variant and where in the protein the variant is located. The three functional domains of *HNF1A* are indicated in orange, green and blue

## Results

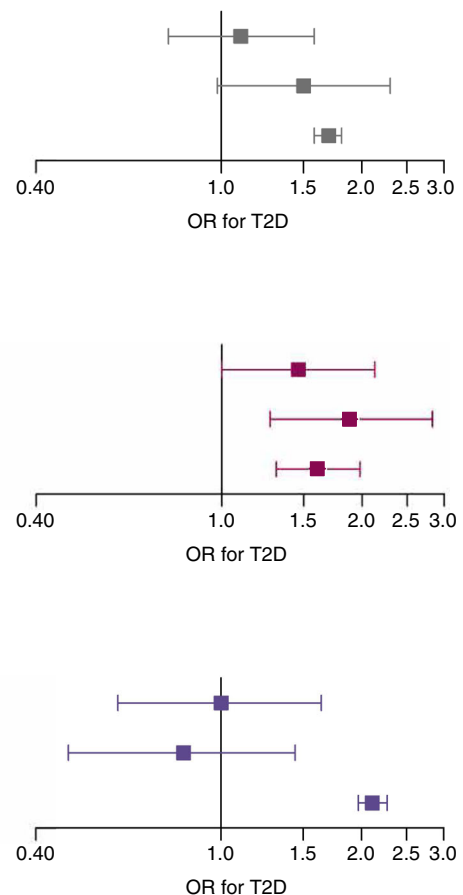
### Identification of variants that affect HNF1A function

A total of 269 coding variants in *HNF1A* across BioMe and UK Biobank participants ( $n=746$ ) had a minor allele frequency  $\leq 1\%$ . Of these, 35 variants reduced the functional activity of HNF1A to  $\leq 60\%$  compared with the wild-type genotype (BioMe  $N_{\text{Unique}}=10$ ; UK Biobank  $N_{\text{Unique}}=9$ ), of which 16 variants were observed in both biobanks ( $n=507$ ) (ESM Fig. 2, ESM Table 1). The majority of functional variants are within one of the three domains (dimerisation, DNA binding and transactivation domains) ( $n=30$ ) (Fig. 1). Most of the functional variants (22 out of 35, or 63%) are located in the transactivation domain, which is known to be the most tolerant to missense variants when diagnosing maturity-onset diabetes of the young [27, 41].

### *HNF1A* variants associate with type 2 diabetes and age of diabetes diagnosis in an ancestry-specific manner

We observed a significant association between *HNF1A* variants and type 2 diabetes in the population of European ancestry (OR 1.46 [95% CI 1.00, 2.13],  $p=0.049$ ), but not in those of African ancestry (OR 1.10 [95% CI 0.77, 1.58],  $p=0.60$ ) or Hispanic-Latino ancestry (OR 1.00 [95% CI 0.60, 1.64],  $p=1.0$ ) (Fig. 2, ESM Table 2). When we restricted our analyses to variants located in one of the three functional domains of *HNF1A*, the association in individuals of European ancestry was more pronounced (OR 1.89 [95% CI 1.27, 2.83],  $p=0.002$ ), but no such effect was seen for other ancestries (African ancestry OR 1.50 [95% CI 0.98, 2.30],  $p=0.064$ ; Hispanic-Latino ancestry OR 0.83 [95% CI 0.47, 1.44],  $p=0.52$ ).

| African         |                            |         |  |
|-----------------|----------------------------|---------|--|
| a               | Associations               | $n$     | OR (95% CI) $p$                          |
|                 | Any variant                | 201     | 1.10 (0.77, 1.58) 0.600                  |
|                 | Functional domain variants | 137     | 1.50 (0.98, 2.30) 0.064                  |
|                 | PRS                        | 8677    | 1.70 (1.58, 1.81) $<2.0 \times 10^{-16}$ |
| European        |                            |         |  |
| b               | Associations               | $n$     | OR (95% CI) $p$                          |
|                 | Any variant                | 458     | 1.46 (1.00, 2.13) 0.049                  |
|                 | Functional domain variants | 298     | 1.89 (1.27, 2.83) 0.002                  |
|                 | PRS                        | 163,289 | 1.61 (1.31, 1.98) $4.8 \times 10^{-6}$   |
| Hispanic-Latino |                            |         |  |
| c               | Associations               | $n$     | OR (95% CI) $p$                          |
|                 | Any variant                | 87      | 1.00 (0.60, 1.64) 1.000                  |
|                 | Functional domain variants | 72      | 0.83 (0.47, 1.44) 0.520                  |
|                 | PRS                        | 7446    | 2.11 (1.97, 2.27) $<2.0 \times 10^{-16}$ |



**Fig. 2** Risk of type 2 diabetes in each ancestry by *HNF1A* variant (any functionally damaging variant identified, and those within a functional domain of the protein) and PRS in (a) African, (b) European and (c) Hispanic-Latino ancestry groups. The OR was calculated using a logistic regression model, with age, sex, BMI and the first ten ancestry PCs as

covariates. In the UK Biobank, centre and chip were included as covariates. The estimates obtained for the specific biobanks were meta-analysed together, and the random-effect ORs are shown. Boxes represent OR; horizontal lines represent 95% CI. T2D, type 2 diabetes

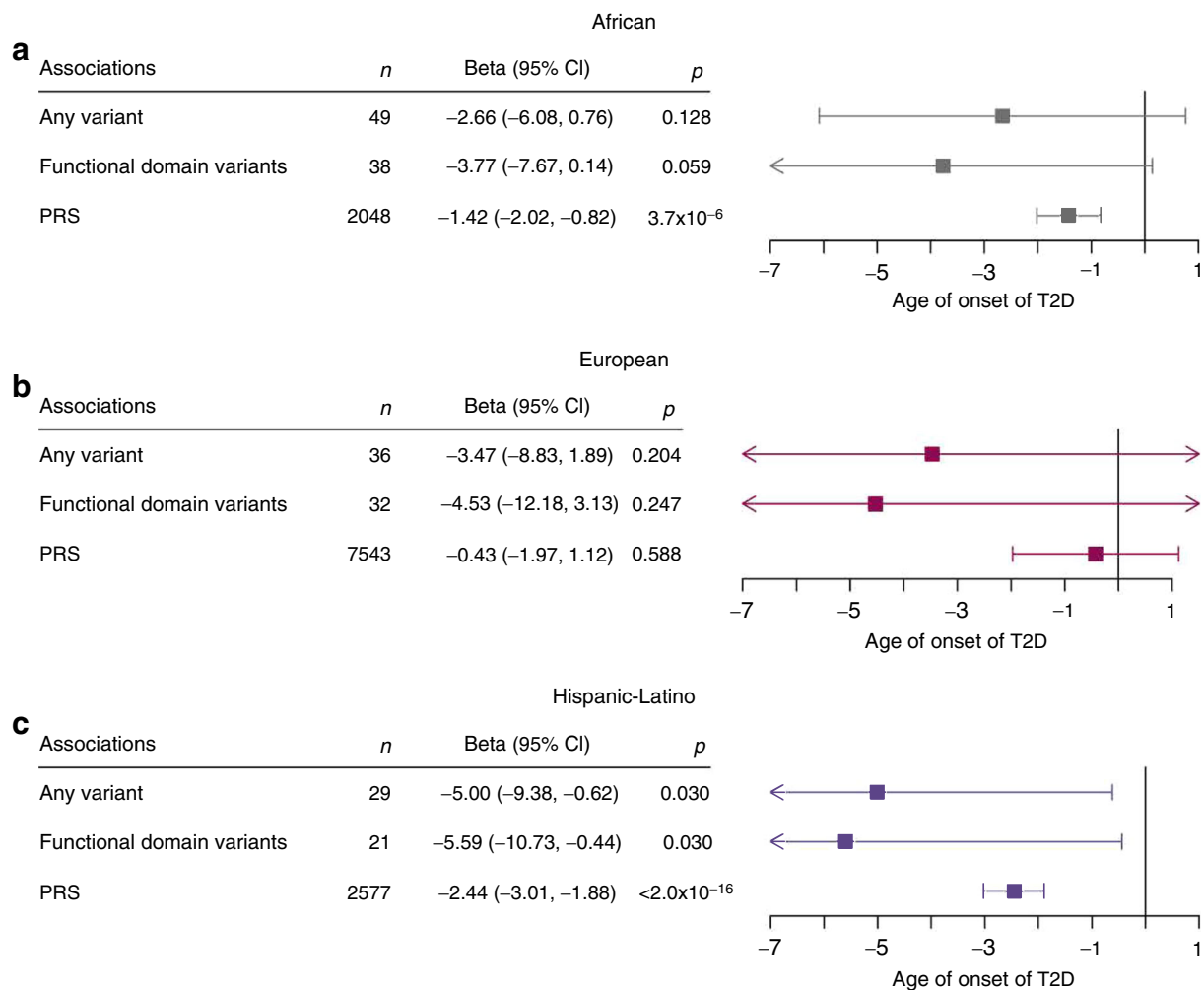
In the Hispanic-Latino population, the age at diagnosis for individuals carrying a functional variant was 5.00 years earlier ([95% CI  $-9.38, -0.62$ ],  $p=0.03$ ). Restricting analyses to variants in one of the three functional *HNF1A* domains only marginally affected the association ( $-5.59$  years [95% CI  $-10.73, -0.44$ ],  $p=0.03$ ) (Fig. 3, ESM Table 2). In participants of African or European ancestry, no association with age of diagnosis was observed (Fig. 3, ESM Table 2).

### Polygenic risk is associated with type 2 diabetes and an earlier age of diabetes diagnosis

A higher PRS was associated with an increased odds of type 2 diabetes (Fig. 2, ESM Table 2). Specifically, per SD increase in the PRS, the odds of type 2 diabetes increase by 1.70 (95%

CI 1.58, 1.81,  $p < 2 \times 10^{-16}$ ) in the population of African ancestry, 1.61 (95% CI 1.31, 1.98,  $p=4.8 \times 10^{-6}$ ) in the population of European ancestry, and 2.11 (95% CI 1.97, 2.27,  $p < 2 \times 10^{-16}$ ) in the population of Hispanic-Latino ancestry.

A higher PRS is also associated with an earlier age of diagnosis in the participants of non-European ancestry, such that per SD increase in the PRS, the diagnosis of diabetes occurred 1.42 years earlier (95% CI  $-2.02, -0.82$ ],  $p=3.71 \times 10^{-6}$ ) in individuals of African ancestry and 2.44 years earlier ([95% CI  $-3.01, -1.88$ ],  $p < 2 \times 10^{-16}$ ) earlier in those of Hispanic-Latino ancestry. However, no association was observed in the population of European ancestry (Fig. 3, ESM Table 2). The explained variance of the trans-ancestry PRS in relation to type 2 diabetes susceptibility is higher for individuals of European ancestry than for individuals of African or Hispanic-Latino ancestry (ESM Table 3).



**Fig. 3** Age of type 2 diabetes diagnosis in each ancestry by *HNF1A* variant (any functionally damaging variant identified, and those within a functional domain of the protein) and PRS in (a) African, (b) European and (c) Hispanic-Latino ancestry groups. The estimates were calculated using a linear regression model, with sex, BMI and the first ten ancestry

PCs as covariates. In the UK Biobank, centre and chip were included as covariates. The estimates for the specific biobanks were meta-analysed together, and the random-effect estimates ( $\beta$ ) are shown. Boxes represent estimates; horizontal lines represent 95% CI. T2D, type 2 diabetes

## Impact of PRS on the association between *HNF1A* variants and type 2 diabetes risk

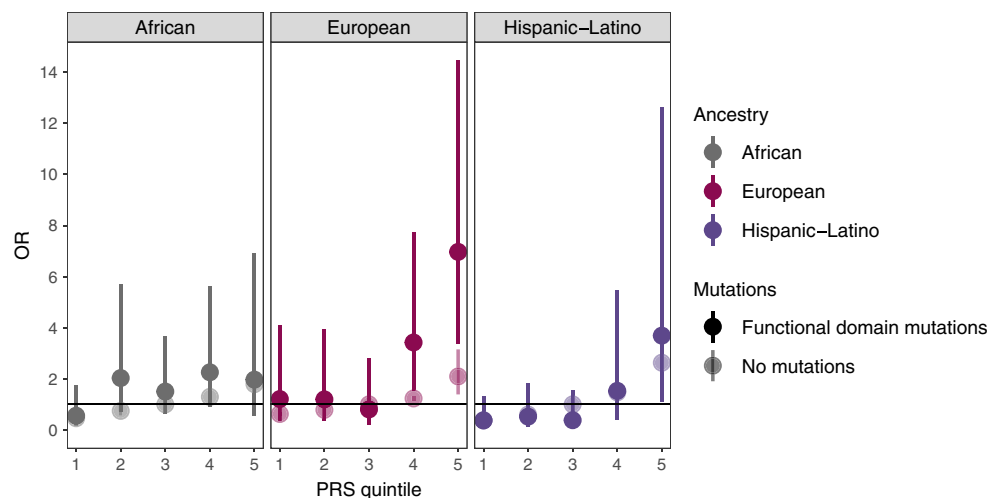
The association between rare *HNF1A* functional domain variants and type 2 diabetes was more pronounced in individuals with a higher PRS compared to those with a lower PRS among those of European ancestry ( $OR_{\text{interaction}}=1.60$ ,  $p_{\text{interaction}}=0.037$ ), but not other ancestries ( $p_{\text{interaction}}=0.32$  for the African population,  $p_{\text{interaction}}=0.45$  for the Hispanic-Latino population) (Fig. 4). For example, individuals of European ancestry who carried functional domain variants and had a high polygenic susceptibility (PRS in the top quintile;  $n=58$ ) had 6.97 higher odds (95% CI 3.36, 14.46,  $p=1.85 \times 10^{-7}$ ) of type 2 diabetes compared with individuals with an average polygenic susceptibility (non-carriers in the middle PRS quintile;  $n=32,717$ ). We observed a similar increased susceptibility in the Hispanic-Latino population (OR 3.69 [95% CI 1.11, 12.65],  $p=0.03$ ) but not in individuals of African ancestry (OR 1.97 [95% CI 0.56, 3.28],  $p=0.29$ ) (Fig. 4, ESM Table 4).

## Discussion

Using data from more than 180,000 individuals across three ancestry groups, from two biobanks, we observed that functional *HNF1A* variants are associated with an increased risk of type 2 diabetes, but only in individuals of European ancestry. This association was even more pronounced when analyses were restricted to functional variants located in the functional domains. *HNF1A* variants were associated with an earlier age

of diabetes diagnosis in the Hispanic-Latino population, but not in the other populations. We did not observe any significant associations between *HNF1A* and type 2 diabetes or age of diagnosis in the population of African ancestry. While these results may reflect true ancestry-specific differences, they may be due to the growing need to functionally validate variants that are exclusively seen in non-European ancestry populations. A higher PRS was associated with an increased risk of type 2 diabetes across all ancestries, and was also associated with an earlier onset of diabetes in the populations of non-European ancestry. Of further interest is that, in the European ancestry population, the association between functional domain *HNF1A* variants and type 2 diabetes is more pronounced in individuals with a high polygenic burden. We observed a similar association in the Hispanic-Latino population, but not in the population of African ancestry. Our study highlights the importance of the location of functional variants in the *HNF1A* gene, the need for more and larger studies in populations of non-European ancestry, and the role of polygenic burden on the impact of rare *HNF1A* variants on type 2 diabetes risk and age of diagnosis.

Few studies have examined the association between rare *HNF1A* variants and type 2 diabetes at a population level [9, 42, 43]. A large study in approximately 75,000 individuals of European ancestry examined the role of pathogenic or likely pathogenic *HNF1A* variants as defined by the American College of Medical Genetics and Genomics, and found no association with type 2 diabetes ( $p=0.4$ ) [42]. In a study that pooled data from three cohorts ( $n=4115$ ), 27 rare variants in *HNF1A* were identified, and no significant association with



**Fig. 4** Risk of type 2 diabetes contributed by rare *HNF1A* variants in the functional domains and common type 2 diabetes risk strata in the three ancestry groups. Individuals were divided into groups based on their type 2 diabetes PRS quintile (0–20%, 20–40%, 40–60%, 60–80% and 80–100%) and their *HNF1A* carrier status. The OR was calculated using a logistic regression model, with age, sex, BMI and the first ten ancestry PCs as covariates. In the UK Biobank, centre and chip were included as

covariates. The estimates obtained for the specific biobanks were meta-analysed together, and the random-effect ORs are shown. The circles represent the OR in each group. The solid colour represents the *HNF1A* rare variant carriers, and the shaded circles represent the non-carriers. Non-carriers in the middle quintile (40–60%) served as the reference group for each ancestry



type 2 diabetes was found. However, when analyses were restricted to 11 damaging functional variants (as assessed by nuclear localisation or transcriptional assays), they found that carriers ( $N_{\text{Participants}}=59$ ) had a fivefold increased risk of type 2 diabetes ( $p=0.0007$ ) [9]. Even though the association that we observed for variants located in the functional domains of *HNF1A* was much lower (OR 1.89,  $p=0.002$ ), possibly due to different variants being included in our analyses, both studies highlight the importance of the location of variants in genes and the need for large sample sizes, consistent with previous observations for maturity-onset diabetes of the young [44].

The associations between the PRS and functional domain carrier status prompted us to investigate the interplay between the two. Our findings are consistent with the results of a recent nested case–control study of type 2 diabetes showing that rare variants across 27 monogenic diabetes genes are associated with increased risk of early-onset type 2 diabetes (age  $\leq 35$  years), particularly among those with a high overall genetic susceptibility [45]. Furthermore, similar findings have been reported regarding other genes and diseases showing that a high polygenic susceptibility exacerbates the effect of rare variants on coronary artery disease, breast cancer and colon cancer [10], prostate cancer [46] and obesity [47]. Here, we show that polygenic burden also affects the impact of rare *HNF1A* variants in a functional domain that have functional implications but are not known to cause monogenic forms of diabetes, at least among individuals of European ancestry.

Because our focus is on functionally validated variants, we include fewer than 20% of all known exonic variants in our study. As such, we may have excluded variants that have not yet been functionally validated, but that indeed may turn out to affect the function of *HNF1A*. Furthermore, we identified fewer variants in the populations of non-European ancestry, probably due to the much smaller sample sizes. We identified one BioMe participant of Hispanic-Latino ancestry who was a heterozygous carrier of the p.E508K variant, which has been shown to increase the odds of type 2 diabetes 5.5-fold [8]. This variant was found to be polymorphic in Hispanic populations, but was hardly present in populations of other ancestry. We note that the allele frequency observed for the BioMe Hispanic-Latino ancestry population (0.007%) is much lower than previously reported for Hispanic populations (0.36%), probably due to differences in the ancestral diversity. Specifically, in the BioMe Biobank, the population of Hispanic-Latino ancestry is predominantly of Puerto Rican and Dominican origin [48], whereas the Hispanic population in which the variant was observed were predominantly from Mexico [8]. Because non-European ancestry populations are under-represented in genetic association studies, including monogenic diabetes studies, not all variants observed may

have been accurately annotated in terms of their functional implications.

In conclusion, we have shown that biologically important functional variants, specifically those within the functional domains of *HNF1A*, are associated with type 2 diabetes in populations of European ancestry, but not in those of African or Hispanic-Latino ancestry. While only few individuals carry these rare variants in *HNF1A*, those who do have a substantially increased risk of type 2 diabetes. Thus, screening a population for rare *HNF1A* variants will have a large impact for those who do indeed carry a risk allele. Moreover, we demonstrate that overall polygenic susceptibility to type 2 diabetes affects the association between rare *HNF1A* variants in functional domains and type 2 diabetes risk in European-ancestry populations. We highlight the importance of including large-scale biobanks when studying rare variants, and specifically the need to include populations of non-European ancestry in the design.

**Supplementary Information** The online version of this article (<https://doi.org/10.1007/s00125-022-05806-2>) contains peer-reviewed but unedited supplementary material.

**Acknowledgements** We are grateful to the UK Biobank participants. The research was conducted under the application code 1251. We also thank the staff of the Mount Sinai BioMe Biobank and all our recruiters who have assisted and continue to assist in data collection and management, and all the participants and their families for supporting the research. We are grateful for the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai, New York.

**Data availability** The summary-level BioMe data that support these findings are available upon request from the corresponding author. The data are not publicly available because they contain participants' private healthcare information. Data from the UK Biobank are available from <https://www.ukbiobank.ac.uk/>

**Funding** LAS was supported by a grant from the National Institutes of Health/National Institute of Diabetes and Digestive and Kidney Diseases (DK130576-02). RJFL is supported by grants from the National Institutes of Health (R01DK110113, R01DK107786 and R01DK124097), the Novo Nordisk Foundation (Laureate Research Grant 2020-0059313) and the Danish National Fund for Research (DNRF Chair Grant). ALG is funded by Wellcome (200837) and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (U01-DK105535 and UM1DK126185) and the Stanford Diabetes Research Center (NIDDK award P30DK116074).

**Authors' relationship and activities** ALG discloses that her spouse is an employee of Genentech and holds stock options in Roche. The remaining authors declare that there are no relationships or activities that might bias, or be perceived to bias, their work.

**Contribution statement** LAS and RJFL conceived the study. LAS and MHP conducted the analyses. LAS, ALG, LB and IA contributed to the variant threshold selection. LAS, MHP, RAJS, NC and RJFL were responsible for interpretation of the data. LAS and RJFL wrote the first draft of the manuscript and all authors made substantial contributions to iterations and approved the final version. RJFL is responsible for the integrity of this work as a whole.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Flannick J, Florez JC (2016) Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat Rev Genet* 17(9):535–549. <https://doi.org/10.1038/nrg.2016.56>
- Vujkovic M, Keaton JM, Lynch JA et al (2020) 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* 52:680–691. <https://doi.org/10.1038/s41588-020-0637-y>
- Mahajan A, Spracklen CN, Zhang W et al (2022) Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat Genet* 54:560–572. <https://doi.org/10.1038/s41588-022-01058-3>
- Khera AV, Chaffin M, Aragam KG et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 50:1219–1224. <https://doi.org/10.1038/s41588-018-0183-z>
- Fuchsberger C, Flannick J, Teslovich TM et al (2016) The genetic architecture of type 2 diabetes. *Nature* 536(7614):41–47. <https://doi.org/10.1038/nature18642>
- Naylor R, Knight Johnson A, del Gaudio D (2018) Maturity-onset diabetes of the young overview. In: Adam MP, Ardinger HH, Pagon RA et al (eds) *GeneReviews*®. University of Washington, Seattle, USA
- Hegele RA, Cao H, Harris SB, Hanley AJG, Zinman B (1999) The hepatic nuclear factor-1 $\alpha$  G319S variant is associated with early-onset type 2 diabetes in Canadian Oji-Cree. *J Clin Endocrinol Metab* 84(3):1077–1082. <https://doi.org/10.1210/jcem.84.3.5528>
- SIGMA Type 2 Diabetes Consortium, Estrada K, Aukrust I et al (2014) Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA* 311(22):2305–2314. <https://doi.org/10.1001/jama.2014.6511>
- Najmi LA, Aukrust I, Flannick J et al (2017) Functional investigations of HNF1A identify rare variants as risk factors for type 2 diabetes in the general population. *Diabetes* 66(2):335–346. <https://doi.org/10.2337/db16-0460>
- Fahed AC, Wang M, Homburger JR et al (2020) Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat Commun* 11(1):3635. <https://doi.org/10.1038/s41467-020-17374-3>
- KING: Relationship Inference Software. <https://www.kingrelatedness.com/>. Accessed 10 Sep 2022
- Bycroft C, Freeman C, Petkova D et al (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562(7726):203–209. <https://doi.org/10.1038/s41586-018-0579-z>
- Taliun D, Harris DN, Kessler MD et al (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590(7845):290–299. <https://doi.org/10.1038/s41586-021-03205-y>
- Van Hout CV, Tachmazidou I, Backman JD et al (2020) Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* 586(7831):749–756. <https://doi.org/10.1038/s41586-020-2853-0>
- Szustakowski JD, Balasubramanian S, Sasson A et al (2020) Advancing human genetics research and drug discovery through exome sequencing of the UK biobank. *Nat Genet* 53:942–948. <https://doi.org/10.1038/s41588-021-00885-0>
- Kho AN, Hayes MG, Rasmussen-Torvik L et al (2012) Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 19(2):212–218. <https://doi.org/10.1136/amiainl-2011-000439>
- Pacheco J, Thompson W (2012) Type 2 diabetes mellitus. Available from <https://phekb.org/phenotype/type-2-diabetes-mellitus>. Accessed 1 Mar 2021
- Chen S, Banks WA, Sheffrin M, Bryson W, Black M, Thielke SM (2018) Identifying and categorizing spurious weight data in electronic medical records. *Am J Clin Nutr* 107(3):420–426. <https://doi.org/10.1093/ajcn/nqx056>
- Eastwood SV, Mathur R, Atkinson M et al (2016) Algorithms for the capture and adjudication of prevalent and incident diabetes in UK biobank. *PLoS One* 11(9):e0162388. <https://doi.org/10.1371/journal.pone.0162388>
- McLaren W, Gil L, Hunt SE et al (2016) The ensembl variant effect predictor. *Genome Biol* 17(1):122. <https://doi.org/10.1186/s13059-016-0974-4>
- Althari S, Najmi LA, Bennett AJ et al (2020) Unsupervised clustering of missense variants in the HNF1A gene using multidimensional functional data aids clinical interpretation. *Am J Hum Genet* 107(4):670–682. <https://doi.org/10.1016/j.ajhg.2020.08.016>
- Juszczak A, Pavić T, Vučković F et al (2019) Plasma fucosylated glycans and C-reactive protein as biomarkers of HNF1A-MODY in young adult-onset nonautoimmune diabetes. *Diabetes Care* 42(1):17–26. <https://doi.org/10.2337/dc18-0422>
- Bjørkhaug L, Sagen JV, Thorsby P, Sovik O, Molven A, Njølstad PR (2003) Hepatocyte nuclear factor-1 $\alpha$  gene mutations and diabetes in Norway. *J Clin Endocrinol Metab* 88(2):920–931. <https://doi.org/10.1210/jc.2002-020945>
- Yoshiuchi I, Yamagata K, Yang Q et al (1999) Three new mutations in the hepatocyte nuclear factor-1 $\alpha$  gene in Japanese subjects with diabetes mellitus: clinical features and functional characterization. *Diabetologia* 42(5):621–626. <https://doi.org/10.1007/s001250051204>
- Nocera D, Menniti M, Belviso S et al (2019) Functional characterization of p.Pro409His variant in HNF1A, a hypomorphic mutation involved in pancreatic  $\beta$ -cell dysfunction. *Acta Diabetol* 56(8):883–888. <https://doi.org/10.1007/s00592-019-01298-6>
- Balamurugan K, Bjørkhaug L, Mahajan S et al (2016) Structure-function studies of HNF1A (MODY3) gene mutations in South Indian patients with monogenic diabetes. *Clin Genet* 90(6):486–495. <https://doi.org/10.1111/cge.12757>
- Bellanné-Chantelot C, Carette C, Riveline J-P et al (2008) The type and the position of HNF1A mutation modulate age at diagnosis of diabetes in patients with maturity-onset diabetes of the young (MODY)-3. *Diabetes* 57(2):503–508. <https://doi.org/10.2337/db07-0859>
- Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW (2019) Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun* 10(1):1776. <https://doi.org/10.1038/s41467-019-09718-5>
- Scott RA, Scott LJ, Mägi R et al (2017) An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* 66(11):2888–2902. <https://doi.org/10.2337/db16-1253>
- National Institutes of Health. dbGaP. Available from <https://www.ncbi.nlm.nih.gov/gap/>. Accessed 19 Apr 2021
- DIAGRAM Consortium. <http://diagram-consortium.org/downloads.html>. Accessed 19 Apr 2021

32. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM (2013) Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet* 14(7):507–515. <https://doi.org/10.1038/nrg3457>
33. Nolte IM (2020) Metasubtract: an R-package to analytically produce leave-one-out meta-analysis GWAS summary statistics. *Bioinformatics* 36(16):4521–4522. <https://doi.org/10.1093/bioinformatics/btaa570>
34. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7. <https://doi.org/10.1186/s13742-015-0047-8>
35. Graham SE, Clarke SL, Wu K-HH et al (2021) The power of genetic diversity in genome-wide association studies of lipids. *Nature* 600(7890):675–679. <https://doi.org/10.1038/s41586-021-04064-3>
36. Nagelkerke NJD (1991) A note on a general definition of the coefficient of determination. *Biometrika* 78(3):691–692. <https://doi.org/10.1093/biomet/78.3.691>
37. Lee SH, Goddard ME, Wray NR, Visscher PM (2012) A better coefficient of determination for genetic profile analysis. *Genet Epidemiol* 36(3):214–224. <https://doi.org/10.1002/gepi.21614>
38. National Health and Nutrition Examination Survey (2020) NHANES 2017–2018 Questionnaire Data. Available from <https://www.cdc.gov/nchs/nhanes/Search/DataPage.aspx?Component=Questionnaire&CycleBeginYear=2017>. Accessed 29 Apr 2021.
39. Pham TM, Carpenter JR, Morris TP, Sharma M, Petersen I (2019) Ethnic differences in the prevalence of type 2 diabetes diagnoses in the UK: cross-sectional analysis of the Health Improvement Network primary care database. *Clin Epidemiol* 11:1081–1088. <https://doi.org/10.2147/CLEP.S227621>
40. Balduzzi S, Rücker G, Schwarzer G (2019) How to perform a meta-analysis with R: a practical tutorial. *Evid Based Ment Health* 22(4): 153–160. <https://doi.org/10.1136/ebmental-2019-300117>
41. Frayling TM, Evans JC, Bulman MP et al (2001) beta-cell genes and diabetes: molecular and clinical characterization of mutations in transcription factors. *Diabetes* 50(Suppl 1):S94–S100. <https://doi.org/10.2337/diabetes.50.2007.s94>
42. Bonnefond A, Boissel M, Bolze A et al (2020) Pathogenic variants in actionable MODY genes are associated with type 2 diabetes. *Nat Metab* 2(10):1126–1134. <https://doi.org/10.1038/s42255-020-00294-3>
43. Flannick J, Mercader JM, Fuchsberger C et al (2019) Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* 570(7759):71–76. <https://doi.org/10.1038/s41586-019-1231-2>
44. Harries LW (2006) Alternate mRNA processing of the hepatocyte nuclear factor genes and its role in monogenic diabetes. *Expert Rev Endocrinol Metab* 1:715–726. <https://doi.org/10.1586/17446651.1.6.715>
45. Pezzilli S, Tohidirad M, Biagini T et al (2022) Contribution of rare variants in monogenic diabetes-genes to early-onset type 2 diabetes. *Diabetes Metab* 48(5):101353. <https://doi.org/10.1016/j.diabet.2022.101353>
46. Darst BF, Sheng X, Eeles RA, Kote-Jarai Z, Conti DV, Haiman CA (2021) Combined effect of a polygenic risk score and rare genetic variants on prostate cancer risk. *Eur Urol* 80(2):134–138. <https://doi.org/10.1016/j.eururo.2021.04.013>
47. Chami N, Preuss M, Walker RW, Moscati A, Loos RJF (2020) The role of polygenic susceptibility to obesity among carriers of pathogenic mutations in MC4R in the UK Biobank population. *PLoS Med* 17(7):e1003196. <https://doi.org/10.1371/journal.pmed.1003196>
48. Belbin GM, Cullina S, Wenric S et al (2021) Toward a fine-scale population health monitoring system. *Cell* 184(8):2068–2083.e11. <https://doi.org/10.1016/j.cell.2021.03.034>

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