Stratified genome-wide association analysis of Type 2 Diabetes reveals subgroups with genetic and environmental heterogeneity

Colette E. Christiansen*[1,5], Ryan Arathimos[2], Oliver Pain[2], Mariam Molokhia[3], Jordana T. Bell[1], Cathryn M. Lewis [2,4]

1 Department of Twin Research and Genetic Epidemiology, King's College London

2 Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK; and NIHR Maudsley Biomedical Research Centre, South London and Maudsley NHS Trust UK.

3 School of Population Health and Environmental Sciences, King's College London

4 Department of Medical and Molecular Genetics, Faculty of Life Sciences & Medicine, King's College London

5 School of Mathematics and Statistics, The Open University

Corresponding authors details:

Email: colette.christiansen@kcl.ac.uk

Address: Department of Twins and Genetic Epidemiology, King's College London, St Thomas' Campus, SE1 7EH, United Kingdom

Abstract

Type 2 diabetes (T2D) is a heterogeneous illness caused by genetic and environmental factors. Previous genome wide association studies (GWAS) have identified many genetic variants associated with T2D and found evidence of differing genetic profiles by age-at-onset. This study seeks to explore further the genetic and environmental drivers of T2D by analysing subgroups based on age-at-onset of diabetes and body mass index (BMI). In UK Biobank, 36,494 T2D cases were stratified into 3 subgroups and GWAS performed for all T2D cases and for each subgroup relative to 421,021 controls. Altogether, 18 SNPs significantly associated genome-wide with T2D in one or more subgroups also showed evidence of heterogeneity between the subgroups, (Cochrane's Q p <0.01) with 2 remaining significant after multiple testing (in *CDKN2B* and *CYTIP*). Combined risk scores, based on genetic profile, BMI and age, resulted in excellent diabetes prediction (AUC=0.92). A modest improvement in prediction (AUC=0.93) was seen when the contribution of genetic and environmental factors was evaluated separately for each subgroup. Increasing sample sizes of genetic studies enables us to stratify disease cases into subgroups which have sufficient power to highlight areas of genetic heterogeneity. Despite some evidence that optimising combined risk scores by subgroup improves prediction, larger sample sizes are likely needed for prediction when using a stratification approach.

Introduction

Diabetes is a metabolic disease characterised by high blood glucose resulting primarily from either insufficient insulin production or insulin resistance. Incidence of diabetes is increasing due both to lifestyle factors, such as increasing levels of obesity, and longer life expectancy[1]. Among minority ethnic communities in the UK, the prevalence is up to 4 times higher than in White populations[2]. Diabetes represents a significant health burden due to the increased rates in individuals with diabetes of physical disability, including blindness and limb amputation, and comorbidities, such as kidney disease, cardiovascular disease and cancer[3].

Glucose homeostasis involves many distinct mechanisms, and genetic susceptibility to diabetes arises from gene variants affecting different gene networks. A complex pattern of genetic susceptibility and environmental exposures by individual leads to significant heterogeneity in the pathogenesis of T2D between individuals. Better understanding of these heterogeneous drivers may aid in predicting both susceptibility to diabetes in individuals, and its downstream complications and thus enable targeted treatments depending on whether, for example, the driver was related to insulin signalling, beta cell function or a combination of both. Accounting clinically for heterogeneity in diabetes is likely to lead to personalised treatment with correspondingly more reliable control of blood sugar levels. Diabetic complications are strongly related to the level of exposure to uncontrolled blood sugar levels, highlighting the importance of treatments which enable individuals to manage their blood sugar levels well.

A clear genetic component to T2D has been identified through population, family and twin-based studies, with heritability estimates ranging from 26% genetic heritability to 50% MZ twin concordance[4]. Many genome-wide association studies (GWAS) of T2D have been carried out, with a large recent meta-analysis including 62,892 diabetic cases and 596,424 controls identifying 139 loci associated with T2D[8]. However, these loci only explain around 20% of T2D heritability[5]. It is becoming increasingly clear that T2D is a disease which has different pathogenic pathways. Most GWAS do not consider the underlying heterogeneity between cases, but studies that stratify cases have found evidence for different genetic profiles between subgroups of T2D cases. Perry et al[6] stratified T2D cases by BMI in a meta-analysis of previous GWAS findings and found a difference in the genetic profile of lean individuals compared with obese. Stratification has also been explored by age-at-onset[7] which showed evidence for different genetic profiles by age. In addition, previous work clustering adult-onset diabetes patients showed five clusters of patients from six variables, each with distinct characteristics and risk of complications[8]. GWAS carried out on these five clusters identified 3 single nucleotide polymorphisms (SNPs) which reached genome-wide significance in at least one cluster[31]. However, the small sample size, with 9,486 individuals divided over 5 clusters, suggests that with greater power further heterogeneity is likely to be detected.

In this study, we aimed to identify differences in the genetic profiles of T2D in European-ancestry individuals, using larger subgroups which broadly captured genetic and environmental drivers of diabetes. We derived three subgroups of T2D cases in UK Biobank, based on BMI and age at onset of diabetes. In GWAS, we identified diabetes-associated SNPs in each subgroup, and tested for heterogeneous SNP-effects across the groups. We explored the genetic correlation between subgroup-derived genetic risk profiles and diabetes traits, including beta cell function and insulin resistance. Finally, we investigated whether this BMI- and age-subgroup approach improved T2D prediction from genetics, or with an integrated risk model across genetic and environmental factors.

Results

In the UK Biobank, 36,494 European-ancestry individuals with T2D were grouped according to BMI and age-at-onset of diabetes into three subgroups. Group 1 consisted of individuals with BMI greater than 30, Group 2 with BMI less than 30 and age-at-onset less than 60 and Group 3 with BMI less than 30 and age-at-onset greater than 60. The characteristics of study participants and subgroups are summarised in Table 1. GWAS were conducted for all individuals and by subgroup using the same 421,021 controls to ascertain whether there were differences in the genetic profiles of these three subgroups.

GWAS overall and by subgroup

GWAS with T2D cases in all subgroups (36,494 cases and 421,021 controls) identified 267 lead SNPs as genome-wide significant associations. An additional 10 SNPs showed genome-wide significant associations in subgroup-specific GWAS. Of the 277 lead SNPs, 25 were novel with no previous evidence for association with T2D, at these SNPs or at SNPs in linkage disequilibrium (LD) (Table 2), excluding SNPs from genes with known associations with diabetes. Many of the lead SNPs had been found previously to have suggestive associations with T2D which did not meet genome-wide significance. 13 of the novel SNPs had previously been associated with metabolic measures, and 8 were associated with other non-metabolic traits and the remaining 4 were not previously associated with any traits. Two of the novel SNPs not previously associated with metabolic traits were annotated to genes which have been associated with chronic inflammatory diseases (*PKIG* and *SBNO2*). In addition, one further SNP, rs9934018, was annotated to *CLCN7* encoding a chloride channel protein with previous work indicating a relationship between chloride channels and beta cell health[25]. Three novel SNPs were found to be significant only in subgroup 1, defined by BMI > 30 (1:168960001, rs17153738, rs76798800). Results for all 277 SNPs are given in Supplementary Table 1. There was no evidence of genomic inflation for either the combined or subgroup analysis with $\lambda_{1000} \sim 1$.

Given these indicators of heterogeneity we further explored differences in association across subgroups. 18 of the 277 lead SNPs significantly associated with diabetes also showed significant heterogeneity between the subgroups using a threshold of Cochrane's Q with p <0.01 (Figure 2); 2 SNPs (rs72655474 in *CDKN2B*, rs10166720 in *CYTIP*) remained significant after a Bonferroni correction for multiple testing.

Genetic differences between groups

Manhattan plots illustrating the genetic associations of each subgroup and all T2D cases are shown in Figure 1. The genetic correlations between the groups were high, partly driven by the common set of controls across all GWAS analyses. The correlation between the Group 1 and 2 was the highest at r^2= 0.97 (standard error (SE) 0.05). The correlations between Group 3 and both group 1 and 2 were similar (Group 1- 3, r^2= 0.88, SE 0.05; Groups 2-3, r^2 0.85; SE 0.05). SNP heritability for each subgroup was h^2 =0.02-0.03 (Supplementary Table 3).

In genetic correlation analysis with 15 pre-defined phenotypes related to T2D, 11 phenotypes evaluated had significant non-zero genetic correlations with one or more subgroups (Table S4, Figure 2b). Group 1 and 2 had significant correlations with fasting insulin (Group 1: r^2 =0.49; Group 2: r^2 = 0.31) and insulin resistance (Group 1: r^2 =0.52; Group 2: r^2 =0.38), with lower, non-significant correlations in Group 3. The only group showing a significant correlation with beta cell function was Group 3 (r^2 =-0.31). Longevity only showed a genetic correlation with Group 1 (r^2 =-0.39).

Genetic and environmental contributions

To determine whether the subgroups provided better prediction of diabetes case status, polygenic scores from each subgroup were generated and their predictive ability compared to the full dataset. The polygenic risk scores calculated using the GWAS summary statistics from the full dataset were more predictive of diabetes than those from each subgroup (Figure 3).

As diabetes risk is a combination of genetic and environmental factors including ageing, a combined risk score was optimised for the training set using both an overall and subgroup approach. This combined risk score included the genetic risk score, a BMI score and a smoothed age score (see Figure 4). Combined risk scores based on the optimised parameters were calculated in the test set and diabetes was predicted using the R predict function and an AUC determined.

The AUC prediction including all subgroup components for the full dataset which included 7,300 test cases was 0.92. Also in the full dataset, a model utilising only the PRS component had an AUC of 0.87 compared with an AUC of 0.74 for a model using only BMI and an AUC of 0.78 when both BMI and age were included. Further analysis was then undertaken to assess the impact of using subgroup specific genetic risk scores and subgroup optimisation of the contribution to the risk score of genetics, BMI and Age (Table 3). We found an increase in prediction when using the genetic risk score calculated on the full group but using the weighting of genetic and environmental factors by subgroup. For Group 1 this increased the AUC from 0.92 to 0.93.

Discussion

In this study we explored the different genetic profiles amongst individuals with T2D based on 457,515 European-ancestry participants from UK Biobank (36,494 diabetes cases and 421,021 controls). Individuals with T2D were grouped based on age-at-onset and BMI, where Group1 comprised all individuals with BMI greater or equal to 30 (Group 1), then Group 2 and 3 had BMI <30 with an age-at-onset less than 60 (Group 2) or age-at-onset over 60 (Group 3). Our study builds on Noordam et al[7], which found that the genetic profile of individuals with diabetes varies by age, and previous studies which have found differences between lean and obese individuals, and found clustering adult-onset diabetes produced distinct groups in terms of patient characteristics and risk of complications[8]. Our study found novel SNPs associated with T2D in the overall dataset, differing genetic profiles among the subgroups and SNPs with high heterogeneity between subgroups. The study further found that the contribution to overall risk between genetic and environmental factors varied by subgroup.

This study analysed a larger sample size than previous UK Biobank Diabetes GWAS, by including related individuals using a mixed model implemented in regenie. Case control studies with a small number of cases relative to the number of controls can suffer from bias which regenie addresses using the Firth correction method. This improved power enabled the identification of 25 novel SNPs. Four of the novel SNPs were annotated to genes with no annotated SNPs previously associated with diabetes including genes associated with chronic inflammatory diseases (*PKIG* and *SBNO2)* and *CLCN7* encoding a chloride channel protein. *PKIG* is a protein kinase inhibitor, blocking protein kinases from phosphorylating proteins which affects the level of activity and function. Specifically, it inhibits cAMP-dependent protein kinase (PKA) and the cAMP/PKA signalling pathway is important for regulating glucose homeostasis in a wide range of processes including both insulin and glucagon secretion and glucose uptake[26]. *SBNO2* regulates inflammatory processes[27] and has been found to be differentially methylated with BMI[28]. Finally, *CLCN7* is part of the family of chloride channel proteins. Beta cells contain chloride channels which respond to glucose concentration and in turn lead to insulin secretion[24].

The three subgroups of T2D cases differed in size, which makes comparisons of the number of significant SNPs detected difficult. However, Group 3, despite being around half the size of Group 1, showed a greater predictive genetic risk score (AUC=0.83) than Group 1 (AUC=0.74). This may indicate that the pathogenesis of diabetes in the obese group has a greater environmental component than the non-obese group. A high degree of genetic correlation was seen between groups but with statistically significant differences for individual SNPs. This observation is in line with the results obtained by stratifying by age[7] which found different genetic profiles between older age-at-onset and younger age-at-onset. *TCF7L2* which contained SNPs more strongly associated with different age of diagnosis in the previous study also contained SNPs heterogeneous between the three groups in the current study. Aly et al[31] also identified a variant in *TCF7L2* (rs7903146) as being significantly associated with only three of their five clusters, finding it not to be associated with severe insulin resistant patients (characterised by late onset and obesity) or severe autoimmune diabetes. In this study, rs4917644 was not even nominally significantly associated with T2D in Group 3, but it was with Group 1 and 2. In addition, 17 further SNPs had significant heterogeneity between the groups with 4 of these significant only in a subgroup and not genome-wide significant overall. Two SNPs met a multiple testing threshold for significance for heterogeneity. A SNP in *CDKN2B*, a gene which previous work has suggested plays a role in beta cell physiology and diabetes risk[34] , and a SNP in *CYTIP*, a gene which has been found in animal models to have significantly different expression in mice deficient in insulin receptor substrate-2[35]. These links to physiological processes suggest the underlying genetic heterogeneity may be reflected in the variation in disease pathogenesis by individual.

There were differences between the groups in their genetic correlation with other traits. Only Group 1 and 2 had significant genetic correlations with insulin resistance and fasting insulin with the older age-at-onset group showing no significant correlation. This group instead showed a significant negative genetic correlation with beta cell function, which neither Group 1 or 2 did. This suggests that older age-at-onset diabetes for those who are not obese has a different pathogenesis from those whose diabetes develops at a younger age or who are obese. Previous work by Udler et al[32] clustered individuals by previously identified GWAS variants and diabetic traits identifying five clusters (Beta cell, Proinsulin, Obesity, Lipodystrophy and Liver). In the current study, we found differences between the subgroups in loci identified by Udler et al in their Proinsulin and Lipodystrophy clusters there were also differences with *ARAP*, *CCND2*, *HNF4A*, *PPARG* and *FAF1* only significant in Group 1 and *ARAP*, *HNF4A* and *CMP* only significant in Group 2.

T2D is a complex disease driven by both genetic and environmental factors. The study sought to assess the contribution by subgroup of genetic and environmental factors to the overall risk by computing combined risk scores for a training data set and assessing their predictive ability in a test data set. Genetic risk was determined using polygenetic risk scores and the study found that the most predictive genetic risk scores for each subgroup were those that were calculated on all T2D cases, rather than by subgroup. Polygenic risk scores were calculated using all nominally significant SNPs, but the power may be too low within subgroups given the modest sample sizes. The combined risk score using all subgroups had an excellent level of prediction with an AUC 0.92, but this was improved for Group 1 by using only Group 1 to optimise the balance between genetic and environmental risk factors. Group 1 consists of obese individuals and the genetic risk score prediction for this group is not strong (AUC 0.74) suggesting a greater environmental component to the pathogenesis. However, by utilising the power of the significantly larger overall group the improvement in genetic risk score prediction combined with the weighting for BMI leads to an improvement in the prediction for this group.

Whilst the study had a relatively large discovery sample, the clustering process meant that each subgroup contained around 10,000 individuals with diabetes. This results in a lower statistical power for analyses within subgroups. The process used to group individuals also resulted in different sized groups leading to differing statistical power, making it harder to compare the resulting genetic profiles. Due to the low numbers of diverse ancestry participants in the UK Biobank, analyses were restricted to individuals of European-ancestry. Diabetes incidence varies substantially by ethnicity both in terms of the level and age of incidence and results may therefore not be generalisable to other ancestries.

In summary, by stratifying T2D cases by age-at-onset and BMI, we found subgroup specific genetic variation and furthermore differing contributions to disease pathogenesis from genetic and environmental risk by subgroup. However, larger sample sizes than those currently available are likely needed to optimise prediction of T2D in a stratification approach.

Materials and Methods

Study Participants

This study included participants from the UK Biobank which includes over 500,000 individuals aged 40-69 at the time of recruitment. Participants were recruited from across the UK between 2006 and 2010[9] and genotype data is available for all individuals[10]. Data up until 1 January 2021 was included in the study. Ethical approval was provided by the Research Ethics Committee (REC reference 11/NW/0382).

Data on participants' diabetes status was based primarily on hospital admission data but also included primary care data and self-reported status. A combination of data fields was used to determine diabetes status including self-reported "Diabetes diagnosed by a doctor" (data field #2443) and the existence of the first reported date in data fields #130706 (insulin-dependent diabetes mellitus), #130708 (non-insulin-dependent diabetes mellitus), #130710 (malnutrition-related diabetes mellitus) and #130712 (other specified diabetes mellitus). For age-at-onset, self-reported responses to "Age diabetes diagnosed" (data field #2976) were used along with the first occurrence of diabetes reporting in data fields #130706, #130708, #130710 and #130710. Individuals with missing BMI were not included in the study. Two further exclusions were made to remove Type 1 diabetes (T1D) cases and individuals of non-European ancestry. The exclusions made for T1D were, all cases with an age-at-onset of 18 or younger in recognition that cases arising at these ages are predominantly T1D[11] and those cases where the individual progressed to insulin treatment within 1 year identified in data field #2986. Previous work by Thomas et al[33] indicates that 90% over T1D cases can be identified by this indicator. Individuals with no records of diabetes were used as controls. The same group of controls was used for all subgroup analyses for consistency to ensure results arising were due only to differences in cases. Individuals with non-European ancestry were identified through 4-means clustering of the first two genetic principal components (PCs) as supplied by UK Biobank and excluded from analysis. Of the remaining individuals, 96% identified as white in data field #21000.

Subgroups

Individuals were put into groups based on BMI and age-at-onset. Given the number of overweight T2D cases the BMI threshold for grouping was chosen as obesity (BMI > $30kg/m^2$). Group 1 included all individuals who had a BMI greater than or equal to 30. To establish the differences between onset as age related disease and onset at earlier ages the age-at-onset was set to be 60. Group 2 and 3 consisted of the remaining individuals divided by age-at-onset with Group 2 individuals having an age-at-onset of less than 60 and Group 3 with an age-at-onset greater than or equal to 60.

Genetic data

UK Biobank profiled the genotypes using Affymetrix UK BiLEVE Axiom and Affymetrix UK Biobank Axiom arrays (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=263). UK Biobank also carried out genotype imputation and preliminary QC on the resulting genetic data (see Supplementary Note). The first 20 principal components (PCs) were recalculated for the individuals to be included in the genome wide association study (following exclusions for ancestry, juvenile diabetes and QC) using FlashPCA v2.0[12].

Genome-wide association analyses

GWAS were conducted using logistic regression models using regenie, C++ program for whole genome regression modelling of large genome-wide association studies[13]. Covariates included the first 20 PCs to account for population structure, sex and batch. In addition, further covariates were included to account for risk factors of diabetes including age, BMI and smoking. Analysis was restricted to SNPs on the autosomes, with MAF > 1% and an imputation information score >0.6. FUMA, a platform to annotate, prioritize, visualize and interpret GWAS results[14] was used to identify independent lead SNPs based on a p-value threshold of 5 x 10^{-8}, r2 <0.6 and LD <0.1. LDlink software programme[15] was used to check for novel sites that were not in LD with sites previously associated with diabetes or measures of blood glucose. For SNPs not in LDlink a manual check of the GWAS catalog[16] was carried out for all sites with LD <0.2 within a 0.5MB window. Lead SNPs were determined to be significant only in one subgroup if the p-values for that SNP in the GWAS of the other subgroups and combined analysis were all greater than 5 x 10^{-8}.

Analysing genetic differences in the subgroups

To assess the heterogeneity of each SNP across the GWAS results for the subgroups we used a fixed effects meta- analysis implemented through GWAMA (Genome-Wide Association Meta-Analysis) software [17] to calculate heterogeneity statistics. Heterogeneity was determined based on Cochrane's Q p-value at a threshold of 1% with I^2 >50% and further with adjustment for multiple testing at a Bonferroni threshold 1.9 x10^{-4}. Genetic correlations and SNP heritability were calculated using the LDSC software[18], based on LD score regression. This is in line with previously reported subgroup GWAS[7] and LDSC has been reported to be unbiased with sample overlap[18]. Genetic correlations between the subgroups were calculated to determine the extent to which genetic profiles for each group overlapped. Correlations were also calculated between GWAS results from each subgroup and 15 other phenotypes (Table S2), including five related to diabetes (2hr glucose, fasting insulin, HbA1c, insulin resistance and beta cell function), four metabolic phenotypes (waist-hip ratio adjusted for BMI, Visceral Fat, Body Fat percentage), a longevity measure (longest 10% survival), four psychiatric phenotypes (anxiety, depression, Alzheimer's disease, autism) and two inflammatory phenotypes (inflammatory bowel disease and rheumatoid arthritis). GWAS summary statistics for these phenotypes were accessed from LD hub[29]. A 95% confidence interval was constructed to test the significance between the correlations observed.

Polygenic risk scores

Polygenic risk scores for the combined groups and each individual diabetes subgroup were calculated based on GWAS results after first rerunning the GWAS using a split training/test (80%/20%) set approach. GWAS were carried out as above in the training set. Polygenic risk scores (PRS) were calculated overall and for each diabetes subgroup using Polygenic Risk Score software, PRSice v2[19,20] using the GWAS summary statistics from the training set analysis with clumping (r^2< 0.1 and 500kb window) and a p-value threshold 0.05 (based on a previous study in UK Biobank[30]).

Combined risk scores

Combined risk scores were calculated using methodology described in Moldovan et al[21]. based on the three risk factors of PRS, BMI and age. Each of these risk factors were transformed to account for the non-linear relationship between diabetes risk across risk factor percentiles. After transformation of each risk factor, each individual then had an assigned genetic risk score (GR), a BMI risk score (BR) and an age risk score (AR). Combined risk scores (CRS) were then calculated as in the formula below with regression model parameters estimated as described in Moldovan et al[21].

$$CRS = \alpha\, GR + \beta\, BR + \gamma\, AR$$

The ability of the combined risk scores to predict diabetes was then assessed in the test set using the area under the ROC curve (AUC). The AUC was calculated using the pROC package in R[22]. The combined risk scores were calculated twice for each subgroup. Once using genetic risk scores based on the subgroup PRS with optimisation by subgroup for α, β, and γ and once using the genetic risk scores based on the overall data set also with optimisation by subgroup. The AUC was determined using PRS alone, BMI alone and BMI and age to assess the impact of including genetics in the score.

Acknowledgements

This research was funded/supported by the National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) at Guy's and St Thomas' NHS Foundation Trust and King's College London, and the NIHR Maudsley BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. This research has been conducted using the UK Biobank Resource under Application Number 65912. Data are available from UK Biobank subject to standard access procedures (www.ukbiobank.ac.uk). The code for carrying out GWAS using regenie has been added to Github (https://github.com/applyfun/gwas_methods).

The authors would like to thank the research volunteers who participated in the study.

Conflict of Interest Statement

None

References

1. Zhu R, Zhou S, Xia L, Bao X (2022) “Incidence, Morbidity and years Lived With Disability due to Type 2 Diabetes Mellitus in 204 Countries and Territories: Trends From 1990 to 2019” *Front. Endocrinol (Lausanne)* **11**
2. Whicher C, O'Neill S, Holt R (2020) “Diabetes in the UK: 2019” *Diab. Med.*, 2020, **37** , 242-247
3. World Health Organisation, (2016) “Global Report on Type 2 Diabetes”
4. Ali O (2013) Genetics of Type 2 diabetes. *World. J. Diab.s*. **4**, 114–123
5. Xue A, Wu Y, Zhu Z, Zhang F, Kemper, K, Zheng Z, Yengo L, Lloyd-Jones L, Sidorenko J, Wu Y, *et al.* (2018) ”Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes” *Nat. Comm..* **9**
6. Perry J, Voight B, Yengo L, Amin N, Dupuis J, Ganser M, Grallert H, Navarro P, Li M, Qi Li, *et al.* (2012) “Stratifying Type 2 Diabetes Cases by BMI Identifies Genetic Risk Variants in *LAMA1* and Enrichment for Risk Variants in Lean Compared to Obese Cases” *PLOS Gen.* **8**
7. Noordam R, Läll K, Smit R, Laisk T, Metspalu A, Esko T, Milani L, Loos R, Mägi R, van Dijk K, *et al* (2021) “Stratification of Type 2 Diabetes by Age of Diagnosis in the UK Biobank Reveals Subgroup-Specific Genetic Associations and Causal Risk Profiles” *Diab.*; **70**, 1816–1825
8. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, Vikman P, Prasad R, Aly D, Almgren P, *et al*. (2018) “Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables” *Lan. Diab. Endocrinol.* **6**, 361-369
9. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, *et al.* (2015) “UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age” *PLoS Med.* **12**
10. Bycroft C, Freeman C, Petkova D, Band G, Elliott L T, Sharp K, Motyer A, Vukcevic D, Delaneau O, O’Connell J, *et al.* (2018) “The UK Biobank resource with deep phenotyping and genomic data” *Nature* **562**

11. Maahs D, West N, Lawrence J, Mayer-Davis E, (2010) “Epidemiology of Type 1 Diabetes” *Endocrin. and Metab. Clinics of N. America* **39**, 481-497

12. Abraham G and Inouye M, (2016) “Fast Principal Component Analysis of Large-Scale Genome-Wide Data", *PLOS ONE* **9**, e93766

13. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki J, Ziyatdinov A, Benner C, O’Dushlaine C, Barber M, Boutkov B, *et al.* (2021) “Computationally efficient whole-genome regression for quantitative and binary traits” *Nat. Genet.* **53**, 1097–1103

14. Watanabe K,Taskesen E, van Bochoven A and Posthuma D (2017) “Functional mapping and annotation of genetic associations with FUMA” *Nat. Commun.* **8**

15. Machiela, M, & Chanock, S (2015) “LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants” *Bioinf.* **31**, 3555–3557

16. Buniello A, MacArthur J, Cerezo M, Harris L, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al (2019) “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019” *Nuc. Acid. Res.*, **47**

17. Mägi R, Morris A, (2016) “GWAMA: software for genome-wide association meta-analysis” *BMC. Bioinf.* **11**

18. Bulik-Sullivan B, Finucane H, Anttila V, Gusev A, Day F, Loh P, ReproGen Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium, Duncan L et al. (2015) “An Atlas of Genetic Correlations across Human Diseases and Traits” *Nat. Gen.***47,** 1236-1241

19. Euesden J, Lewis C, O'Reilly P. (2015) “PRSice: Polygenic Risk Score software” *Bioinf.*. **31**,1466-1468

20. Choi S, O’Reilly P (2019) “PRSice-2: Polygenic Risk Score software for biobank-scale data” *GigaSci.*. **8**

21. Moldovan A, Waldman Y, Brandes N, Linial M (2021) “Body Mass Index and Birth Weight

Improve Polygenic Risk Score for Type 2 Diabetes" *J. Pers. Med.* **11**

22. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves." *BMC Bioinf.* **12**
23. Verdi S, Abbasian G, Bowyer R, Lachance G, Yarand D, Christofidou P, Mangino M, Menni C, Bell J.T, Falchi M, (2019) "TwinsUK: The adult twin registry update" *Twin Res. and Hum. Gen.* **22**: 523–529
24. Tsai P, Glastonbury C, Eliot M, Bollepalli S, Yet I, Castillo-Fernandez J, Carnero-Montero E, Hardiman H, Martin T, Vickers A, (2018) "Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health" *Clin. Epigen.* **10**
25. Di Fulvio M, and Aguilar-Bryan L (2019) "Chloride transporters and channels in β-cell physiology: revisiting a 40-year-old model" *Biochem. Soc. Trans*, **47** 1843–1855
26. Yang H, Yang Y (2016) "Targeting cAMP/PKA pathway for glycemic control and type 2 diabetes therapy" *J. Mol. Endocrin.* **57**, 93-108
27. El Kasmi K, Smith A, Williams L, Neale G, Panopoulos A, Watowich S, Häcker H, Foxwell B, Murray P *et al* (2007) "Cutting edge: A transcriptional repressor and corepressor induced by the STAT3-regulated anti-inflammatory signaling pathway" *J. Immunol.* **179**, 7215-7219
28. Orozco L, Farrell C, Hale C, Rubbi L, Rinaldi A, Civelek M, Pan C, Lam L, Montoya D, Edillor C, *et al.* (2018). "Epigenome-wide association in adipose tissue from the METSIM cohort" *Hum. Mol. Gen.*, **27**, 1830–1846
29. Zheng J, Erzurumluoglu A, Elsworth A, Kemp J, Howe L, Haycock P, Hemani G, Tansey K, Laurin C, "Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis" *Bioinf. (Oxford, England)*, **33**, 272–279
30. Liu W, Zhuang Z, Wang W, Huang T, Liu Z. (2021) "An Improved Genome-Wide Polygenic

Score Model for Predicting the Risk of Type 2 Diabetes” *Front. Gen.* **12**

31. Mansour Aly, D., Dwivedi, O.P., Prasad, R.B, Käräjämäki A, Hjort , Thangam M, Åkerlund M, MahajanA, Udler M, Florez J *et al.* (2021) “Genome-wide association analyses highlight etiological differences underlying newly defined subtypes of diabetes” *Nat. Genet.* **53**, 1534–1542
32. Udler M, Kim J, von Grotthuss M, Bonàs-Guarch S, Cole JB, Chiou J, Christopher D. Anderson on behalf of METASTROKE and the ISGC, Boehnke M, Laakso M, Atzmon G, et al (2018) “Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis” *PLoS Med.* **15**
33. Thomas N, Jones S, Weedon M, Shields B, Oram R, Hattersley A (2017) “Frequency and phenotype of type 1 diabetes in the first six decades of life: a cross-sectional, genetically stratified survival analysis from UK Biobank” *Lanc. Diab. Endocrinol* **6**, 122-129.
34. Kong Y, Sharma R, Nwosu B, Alonso L (2016) “Islet biology, the CDKN2A/B locus and type 2 diabetes risk” *Diabetologia* **59**, 1579-1593.
35. Oliveira J, Rebuffat S, Gasa R, Burks D, Garcia A, Kalko S, Zafra D, Guinovart J, Gomi R (2014) “Tungstate promotes β-cell survival in Irs2−/− mice” *Am. J. Physiol. Endocrin. Metab.***306**

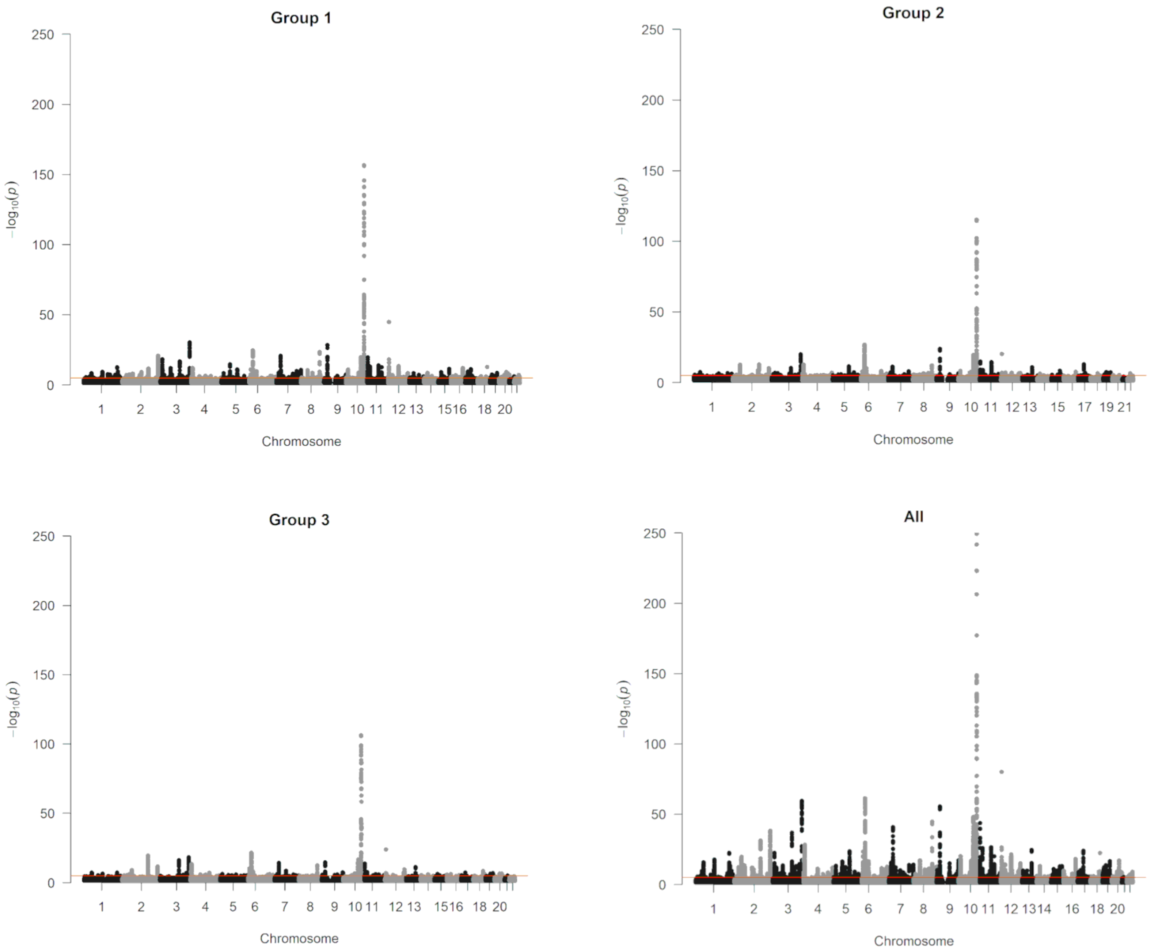

Figure 1: Genome wide association study results by subgroup. Manhattan plots GWAS results from T2D cases and controls from (a) Group1, (b) Group 2, (c) Group 3, (d) all groups combined**.**

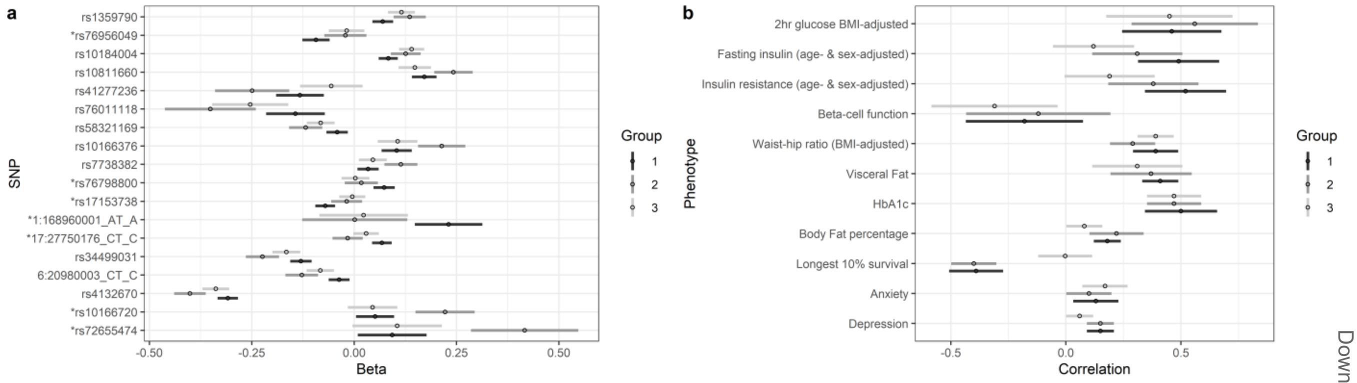

Figure 2 Genetic differences between subgroups a. Heterogeneity between SNPs showing SNPs with Cochrane's Q p <0.01. SNPs which are genome-wide significant in a subgroup but not genome-wide significant overall marked with an asterisk **b.** Genetic correlations between subgroups and genetic profiles of other phenotypes.

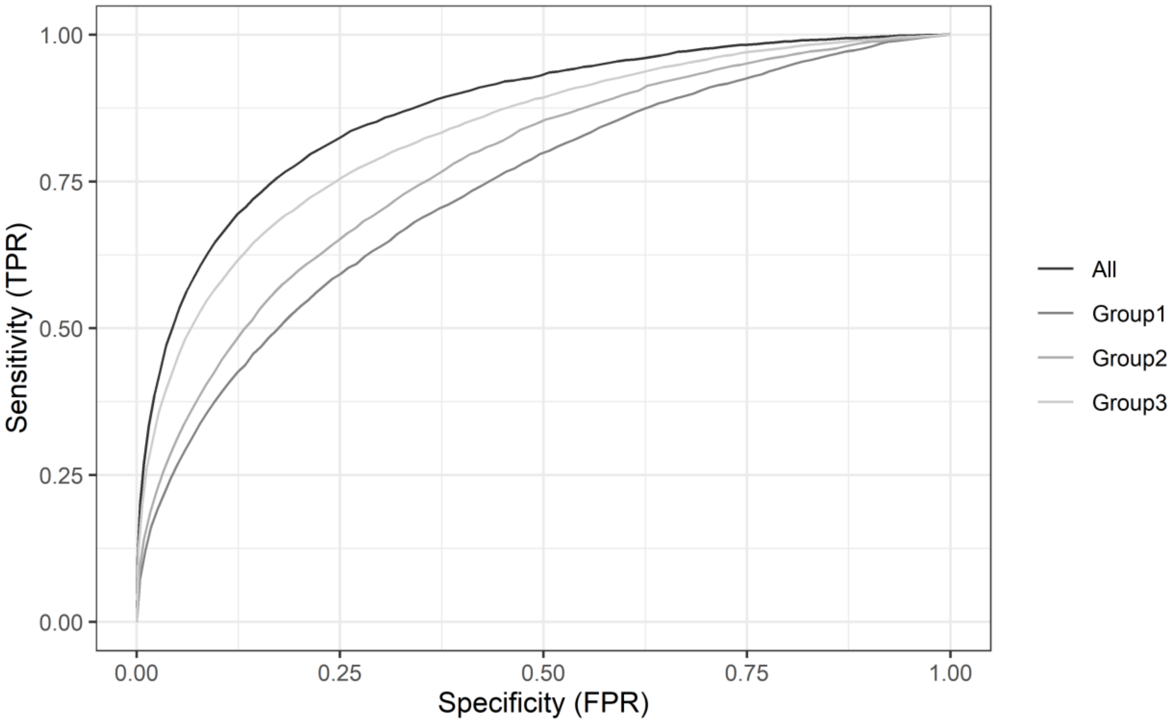

Figure 3 ROC curves for polygenic risk scores by subgroup. All group analysis AUC = 0.87, Group 1 AUC = 0.74, Group 2 AUC = 0.78, Group 3 AUC = 0.83.

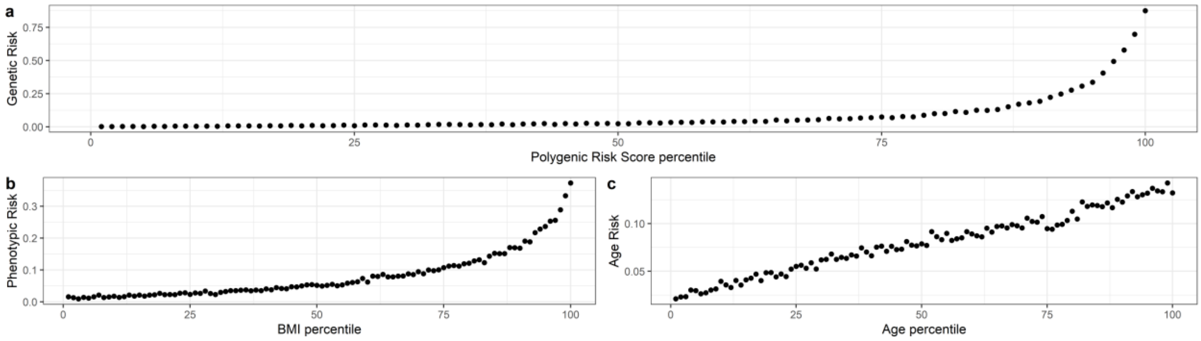

Figure 4: Combined risk score components a. Genetic risk at each percentile category for polygenic risk calculated for the combined group **b** Diabetes risk due to BMI calculated at each percentile category **c.** Diabetes risk due to age calculated at each percentile category.

Tables

	Diabetes Cases	**Controls**	**Group 1**	**Group 2**	**Group 3**
Number	38,481	421,021	20,588	6,328	9,344
BMI (mean, SD)	31.6 (5.7)	27.0 (4.5)	35.2 (4.7)	26.8 (2.7)	26.8 (2.3)
Age-at-onset (mean, SD)	58.6 (11.1)	n/a	59.0 (10.0)	51.6 (7.2)	68.1 (5.7)
Age (at 1/1/21)	71.5 (7.0)	68.0 (8.1)	70.7 (7.1)	75.1 (4.4)	68.5 (7.4)
Sex					
Female	39%	56%	41%	31%	40%
Male	61%	44%	59%	69%	60%
Smoking Status					
Never	42%	54%	42%	43%	41%
Ex-smoker	45%	35%	46%	41%	45%
Current	13%	10%	12%	16%	14%

Table 1: Participant characteristics of T2D cases and controls and characteristics of the three subgroups

ID	Chrom	Position	Gene	Effect Allele	Effect Allele Freq	Beta	se	P-value	Previous Associations
1:168960001	1	168960001		AT	0.978861	0.118722	0.030362	9.22E-05	unrelated
rs17153738	7	106658309		T	0.654342	-0.03904	0.008896	1.14E-05	unrelated
rs76798800	1	154994978	DCST2	G	0.733766	0.037472	0.009505	8.08E-05	BMI/WHR
rs371649660	7	74353884	GTF2I	C	0.890097	0.095554	0.014636	6.63E-11	BMI/WHR
rs2821226	1	203517292		A	0.472495	0.048463	0.008491	1.15E-08	BMI/WHR
rs199679345	6	34234953		C	0.951861	-0.13681	0.019876	5.85E-12	related
20:43246633	20	43246633	PKIG	CACAA	0.952468	0.126113	0.02029	5.11E-10	None
rs201458438	12	123801250	SBNO1	C	0.320434	-0.06178	0.00955	9.84E-11	BMI/WHR
rs38169	7	15893300		C	0.244283	0.061601	0.009802	3.28E-10	related
rs10823909	10	73989184	ANAPC16	T	0.911994	-0.08372	0.014924	2.03E-08	BMI/WHR
rs72752197	5	44627323		C	0.780923	0.057415	0.010266	2.23E-08	unrelated
rs11057368	12	124309574	DNAH10	G	0.633135	0.048748	0.008813	3.18E-08	BMI/WHR
rs11187152	10	94500111		G	0.923309	0.098068	0.016017	9.21E-10	unrelated
rs1916334	12	122484294	BCL7A	G	0.804831	-0.06304	0.010707	3.93E-09	BMI/WHR
rs56218834	2	25520857	DNMT3A	G	0.572974	0.062063	0.008558	4.12E-13	BMI/WHR
rs556132116	17	17931884	ATPAF2	C	0.689899	-0.05928	0.0108	4.04E-08	BMI/WHR
6:19809493	6	19809493	RP1-167F1.2	TA	0.689373	0.050394	0.009156	3.71E-08	None
10:114673015	10	114673015		GGT	0.856163	0.071575	0.012742	1.94E-08	None
rs142201902	3	185153047	MAP3K13	C	0.92574	-0.09285	0.016139	8.76E-09	BMI/WHR
rs34636896	10	114647936		G	0.183938	-0.10405	0.011741	7.84E-19	unrelated
19:1149092	19	1149092	SBNO2	GC	0.638361	0.053763	0.009113	3.65E-09	unrelated
rs876475	9	81545012		G	0.599162	-0.04786	0.008595	2.57E-08	unrelated
rs59521405	2	112264989		T	0.752715	0.060811	0.009975	1.09E-09	BMI/WHR
rs78535155	13	49435399		A	0.985318	0.198544	0.036376	4.81E-08	None
rs9934018	16	1504934	CLCN7	T	0.593146	0.046978	0.00857	4.22E-08	unrelated

Table 2: Novel SNPs which were either significant in the analysis including all groups or significant in a subgroup analysis. P-values are shown for the all group analysis.

Training and Test set	Genetic Risk population	AUC (95% CI)
All groups	All groups	0.921 (0.918-0.924)
Group 1	All groups	0.932 (0.929-0.934)-)
Group 1	Group 1	0.883 (0.880-0.882)
Group 2	All groups	0.915(0.912-0.918)
Group 2	Group 2	0.825 (0.820-0.829)
Group 3	All groups	0.916 (0.913-0.919)
Group 3	Group 3	0.894 (0.890-0.898)

Table 3: Predicting diabetes using combined risk scores within groups and across groups