**Cohort Description of the UK Biobank population**

The UK Biobank cohort is a large prospective general population cohort. Baseline assessments took place between 2006 and 2010 in 22 different assessment centers across the UK (1). A total of 502,628 participants between the age of 40 and 70 years were recruited from the general population. Invitation letters were sent to eligible adults registered with the National Health Service and living within a 25 miles distances from one of the study assessment centers across the country. At the study assessment center, participants completed a questionnaire through touchscreen that included topics such as sociodemographic characteristics, physical and mental health, lifestyle and habitual food intake. The UK Biobank study was approved by the North-West Multi-center Research Ethics Committee (MREC). Access for information to invite participants was approved by the Patient Information Advisory Group (PIAG) from England and Wales. All participants in the UK Biobank study provided written informed consent.

For the present study, we restricted the analyses to the UK Biobank participants who were in the full release imputed genomics datasets, and who had a self-reported European ancestry and excluded individuals with outlying principal components.

*Data and Resource Availability*

The present study, completed under project number 22474, was mainly conducted in the UK Biobank, an open database where researchers can use the data after acceptance of a research project by UK Biobank Resources (<https://www.ukbiobank.ac.uk/>).

*Genotyping and genetic imputations*

UK Biobank genotyping was conducted by Affymetrix using a bespoke BiLEVE Axium array for approximately 50,000 participants; the remaining participants were genotyped using the Affymetrix UK Biobank Axiom array. All genetic data were quality controlled centrally by UK Biobank resources. More information on the genotyping processes can be found online (<https://www.ukbiobank.ac.uk>).

Based on the genotyped SNPs, UK Biobank resources performed centralized imputations on the autosomal SNPs using the UK10K haplotype (2), 1000 Genomes Phase 3 (3), and Haplotype Reference Consortium (HRC) reference panels (4). Autosomal SNPs were pre-phased using SHAPEIT3 and imputed using IMPUTE4. In total, ~96 million SNPs were imputed. Related individuals were identified by estimating kinship coefficients for all pairs of samples using only markers weakly informative of ancestral background.

**Cohort description of the Estonian Biobank**

The Estonian Biobank is a population-based biobank with over 200,000 participants, which is volunteer-based sample of the Estonian resident adult population (aged ≥18 years). The 150K data freeze was used for the analyses described in this paper. All biobank participants have signed a broad informed consent form.  Information on ICD codes is obtained via regular linking with the national Health Insurance Fund and other relevant databases (5). Estonian Biobank samples were genotyped in Core Genotyping Lab of Institute of Genomics, University of Tartu using Illumina GSAv1.0, GSAv2.0, and GSAv2.0\_EST arrays. Altogether 155,772 samples were genotyped and PLINK format files were created using Illumina GenomeStudio v2.0.4. Individuals were excluded from the analysis if their call-rate was < 95% or sex defined based on heterozygosity of X chromosome did not match sex in phenotype data. Variants were filtered by call-rate < 95% and HWE p-value < 1e-4 (autosomal variants only). Variant positions were updated to b37 and all variants were changed to be from TOP strand using tools and reference files provided in <https://www.well.ox.ac.uk/~wrayner/strand/> webpage. Before imputation variants with MAF<1% and indels were removed. QC was made using R markdown script "gsa\_array\_analysis\_pipeline.Rmd", which uses PLINK v1.9 (6) and several in-house R and PERL scripts.

After QC the dataset contained 154,201 samples for imputation. Prephasing was done using Eagle v2.3 (7) software (number of conditioning haplotypes Eagle2 uses when phasing each sample was set to: --K pbwt=20000) and imputation was done using Beagle v.28Sep18.793 (8) with effective population size ne=20,000. Estonian population specific imputation reference of 2297 WGS samples were used (9).

**Cohort description of the BioMe Biobank**

The BioMe Biobank, founded in 2007, is an ongoing electronic health record (EHR)-linked biorepository that enrolls participants non-selectively from across the Mountain Sinai Health System (MSHS). As of January 2020, >50K participants comprising diverse ancestries have been recruited from >26 outpatients sites located in Manhattan and Queens. The median number of outpatient encounters is ~20 per participant, reflecting predominant enrollment of participants with common chronic conditions from primary care practices. Demographic and lifestyle information is collected by interview-based questionnaire at enrollment, and blood is drawn for plasma and DNA extraction. Participant data is regularly updated through linkage to the EHR, and phenotype algorithms are continuously developed and implemented by a multidisciplinary team at Mount Sinai, as well as through collaboration within the Electronic Medical Records and Genomics (eMERGE) consortium. Cases of diabetes mellitus were defined on the basis of an algorithm developed by the eMERGE consortium [6], which previously demonstrated a 98% positive predictive value for the identification of T2D patients compared to physician review in Northwestern University’s NUgene biobank [7], and which has been implemented within the BioMe biobank. To limit cases to those with non-insulin dependent T2D at diagnosis, we excluded individuals who were prescribed insulin (in any form) twice or more in an outpatient setting within the first half year after diagnosis. Age of diagnosis was calculated on the basis of the earliest date at which a T2D medication prescription, T2D-specific ICD code, or HbA1C measurement >6.5% was registered (medications/ICD codes as defined by algorithm). For controls, we included individuals who had not been assigned case status by the algorithm, nor had a Phecode related to T1D (codes: 250.1, 250.11, 250.12, 250.13, 250.14, 250.15), and for whom evidence of contact with the MSHS past the age of 70 years could be found in the EHR. Related individuals (2nd degree or stronger) were excluded.

**Reference List**

1. 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 medicine 2015;12:e1001779

2. UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, Lawson D, Iotchkova V, Schiffels S, Hendricks AE, Danecek P, Li R, Floyd J, Wain LV, Barroso I, Humphries SE, Hurles ME, Zeggini E, Barrett JC, Plagnol V, Richards JB, Greenwood CM, Timpson NJ, Durbin R, Soranzo N: The UK10K project identifies rare variants in health and disease. Nature 2015;526:82-90

3. 1000 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;526:68-74

4. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, Luo Y, Sidore C, Kwong A, Timpson N, Koskinen S, Vrieze S, Scott LJ, Zhang H, Mahajan A, Veldink J, Peters U, Pato C, van Duijn CM, Gillies CE, Gandin I, Mezzavilla M, Gilly A, Cocca M, Traglia M, Angius A, Barrett JC, Boomsma D, Branham K, Breen G, Brummett CM, Busonero F, Campbell H, Chan A, Chen S, Chew E, Collins FS, Corbin LJ, Smith GD, Dedoussis G, Dorr M, Farmaki AE, Ferrucci L, Forer L, Fraser RM, Gabriel S, Levy S, Groop L, Harrison T, Hattersley A, Holmen OL, Hveem K, Kretzler M, Lee JC, McGue M, Meitinger T, Melzer D, Min JL, Mohlke KL, Vincent JB, Nauck M, Nickerson D, Palotie A, Pato M, Pirastu N, McInnis M, Richards JB, Sala C, Salomaa V, Schlessinger D, Schoenherr S, Slagboom PE, Small K, Spector T, Stambolian D, Tuke M, Tuomilehto J, Van den Berg LH, Van Rheenen W, Volker U, Wijmenga C, Toniolo D, Zeggini E, Gasparini P, Sampson MG, Wilson JF, Frayling T, de Bakker PI, Swertz MA, McCarroll S, Kooperberg C, Dekker A, Altshuler D, Willer C, Iacono W, Ripatti S, Soranzo N, Walter K, Swaroop A, Cucca F, Anderson CA, Myers RM, Boehnke M, McCarthy MI, Durbin R, Haplotype Reference C: A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016;48:1279-1283

5. Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, Perola M, Ng PC, Magi R, Milani L, Fischer K, Metspalu A: Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol 2015;44:1137-1147

6. 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

7. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, Schoenherr S, Forer L, McCarthy S, Abecasis GR, Durbin R, A LP: Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 2016;48:1443-1448

8. Browning SR, Browning BL: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;81:1084-1097

9. Mitt M, Kals M, Parn K, Gabriel SB, Lander ES, Palotie A, Ripatti S, Morris AP, Metspalu A, Esko T, Magi R, Palta P: Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet 2017;25:869-876