Supplementary material

Supplementary Methods

Definition of clinical diagnoses in Genes & Health

We captured all diagnostic codes for type 1 (T1D) and type 2 diabetes (T2D) from primary care (SNOMED) and secondary care (ICD-10), using clinical codesets in routine use in the NHS. These have been previously used [21] and are summarised in Supplementary Table 1. We excluded all individuals with diagnosed secondary diabetes, rare types of diabetes (e.g. monogenic diabetes) or a history of pancreatitis or pancreatic surgery from all analyses using additional codelists (see Supplementary Table 1). We collated additional clinical features, including prescribing data, clinical measurements and laboratory data, including measurements of body mass index (BMI), Haemoglobin A1c (HbA1c), high-density lipoprotein (HDL) and triglycerides (TG) measured within a year of the earliest date of diagnosis. Where C-peptide measures were available (either fasting or random samples), we defined insulin deficiency as C-peptide <200 pmol/l (measured fasting, random or stimulated). We defined diabetes autoantibody positivity as the presence of any NHS laboratory-defined anti-glutamic acid, anti-ZnT8 or anti-islet cell antibody titre above the local laboratory reference range. We calculated time to first use of insulin for patients who had received insulin based on prescription records and dates of diagnosis. We also captured the diagnosis of other organ-specific autoimmune diseases (Supplementary Table 2).

Cleaning of quantitative data from G&H electronic health records

Since there were multiple height measurements per individual, we took the median height measurement recorded over the age of 18, assuming this would be a reasonable proxy for height at diagnosis. BMI at diagnosis was calculated based on the weight nearest to diagnosis date and this median adult height for patients diagnosed over age 18, with clinically improbable results excluded (BMI >75 or <12 kg/m^2). Laboratory results were considered outliers if they fell out with the established reference range for our local NHS laboratory (HbA1c > 160 mmol/l, HDL > 6.24 mmol/l, triglycerides > 50.0 mmol/l). For older HbA1c results, values expressed in Diabetes Control and Complications Trial percentage units were converted to mmol/mol using an established conversion formula.

Preparation of the genotype data from Genes & Health

Quality control

We performed quality control of the Global Screening Array data with Illumina's GenomeStudio and plink v1.9. Using GenomeStudio, we removed variants with h-cluster separation scores <0.57, Gentrain score <0.7, an excess of heterozygotes >0.03, or ChiTest 100 (Hardy-Weinberg test) <0.6. We additionally removed variants that were included on the array in order to tag specific structural variants, and samples with that failed gender checks or had low call rate (<0.995 for male samples and <0.992 for female samples across all 637,829 variants including those on Y chromosome for males). We then removed SNPs with MAF < 0.5%, leaving 44,396 people and 399,911 SNPs. To ensure we did not lose SNPs

that have high quality but that were out of Hardy-Weinberg Equilibrium due to high rates of consanguinity and strong population structure in Pakistanis [38], we removed SNPs that failed Hardy-Weinberg Equilibrium p-value < $1x10^{-6}$ in Bangladeshis alone, as done in [39]. This left 399,047 SNPs.

Inference of genetic ancestry

To infer genetic ancestry, we merged the SNPs with MAF>1% from the imputation dataset (355,136 SNPs) with whole-genome sequence data from unrelated individuals (inferred using KING as described below) from the 1000 Genomes Project and with Central and South Asian individuals from the Human Genome Diversity Project. After first excluding palindromic variants and multiallelic sites from both datasets, we merged the datasets by matching positions and alleles of the common SNPs that passed QC in G&H, and kept variants found in both datasets, leaving 349,632 SNPs. We then excluded 1,285 variants due to allele frequency discrepancies between G&H and South Asian reference individuals (>4 standard deviations from the mean residual of $-\log_{10}$ frequency bins, and Fisher's exact test $P<1x10^{-5}$), leaving in 348,347 SNPs. Before carrying out the principal components analysis (PCA), we pruned the SNPs (window size 1000kb, step size 50, linkage disequilibrium (LD) r^2 cutoff 0.1) and removed long LD regions, leaving 104,552 SNPs. We ran the PCA with plink 1.9, first calculating PCs for the 3,433 reference individuals, then projecting the G&H samples into the reference PC space. We used the umap package in R to perform UMAP on the PCs, and found that the reference individuals separated into known superpopulations optimally when using seven PCs. This allowed us to remove 76 of the G&H individuals who were genetically most similar to non-South Asian samples from the reference datasets leaving 44,320 individuals. We then performed a second PCA on the unrelated G&H individuals, projecting the related G&H individuals on those, and then performing UMAP with four PCs to identify two major clusters. These corresponded well to self-identified Pakistani and Bangladeshi groups, and were thus used to define individuals as genetically Pakistani or Bangladeshi.

Inference of relatedness

We used the PropIBD metric from KING to estimate relatedness between the 9,675 genetically-inferred Pakistani or Bangladeshi individuals in the following groups, listed here in priority order: 1. T1D cases, 2. Ambiguous cases, 3. T2D cases and 4. non-diabetic controls. We then removed one from each pair of related individuals inferred to be third-degree relatives or closer, preferentially retaining individuals in the higher priority groups. To maximise the sample size, we first removed the relatives within the group of interest (starting with T1D cases, followed by the other groups in priority order). We then ranked individuals by their number of inferred relatives, then sequentially removed the individual with the highest number of relatives from the groups of lower priority, until no relative pairs remained.

Quality control of genotype chip data in the Indian cohort

Genotype data on the Indian subjects were generated using the Global Screening Array v3.0. Genome studio 2.0 was used to convert microarray intensity files (.idat) to genotypes. We applied a sample call rate of 95% and Gentrain > 0.5, cluster separation > 0.4 and AA/AB/BB Tdev < 0.06 for the SNPs. Further poorly clustering SNPs were manually inspected and removed as per recommendations [40]. This retained 649 people and 594,542

SNPs. Using plink v1.9, we applied MAF < 0.5% and Hardy-Weinberg p-value < $1x10^{-6}$, leaving 341,343 SNPs. The data were merged with QCed data from the G&H cohort before imputation, as described in the main Methods.

Principal component analysis and imputation of G&H and Indian samples

We first merged the post-QC genotype data from the G&H and Indian cohorts, then pruned the SNPs (window size 1000kb, step size 50, LD r^2 cutoff 0.1) and removed regions of long-range linkage disequilibrium except those overlapping the HLA region (chr6:25391792-33424245), leaving 113,928 SNPs. We then carried out PCA using plink v1.9.

Imputation was carried out using the Michigan TOPMED Imputation Server on the common subset of genotyped variants between the G&H and Indian samples, excluding palindromic SNPs and chrY SNPs. Non-ACGT variants were removed using --snp-only just-acgt in plink2 at a MAF cut-off of 0.005. TOPMed r2 was used as the imputation reference panel. We removed variants with imputation accuracy r^2<0.3, leaving 45,946,210 SNPs.

Supplementary Tables

Supplementary Table 1: Clinical codelists used to define the set of type 1 and type 2 diabetes cases. Codes included in the NHS primary care Quality and Outcomes Framework are denoted in blue.

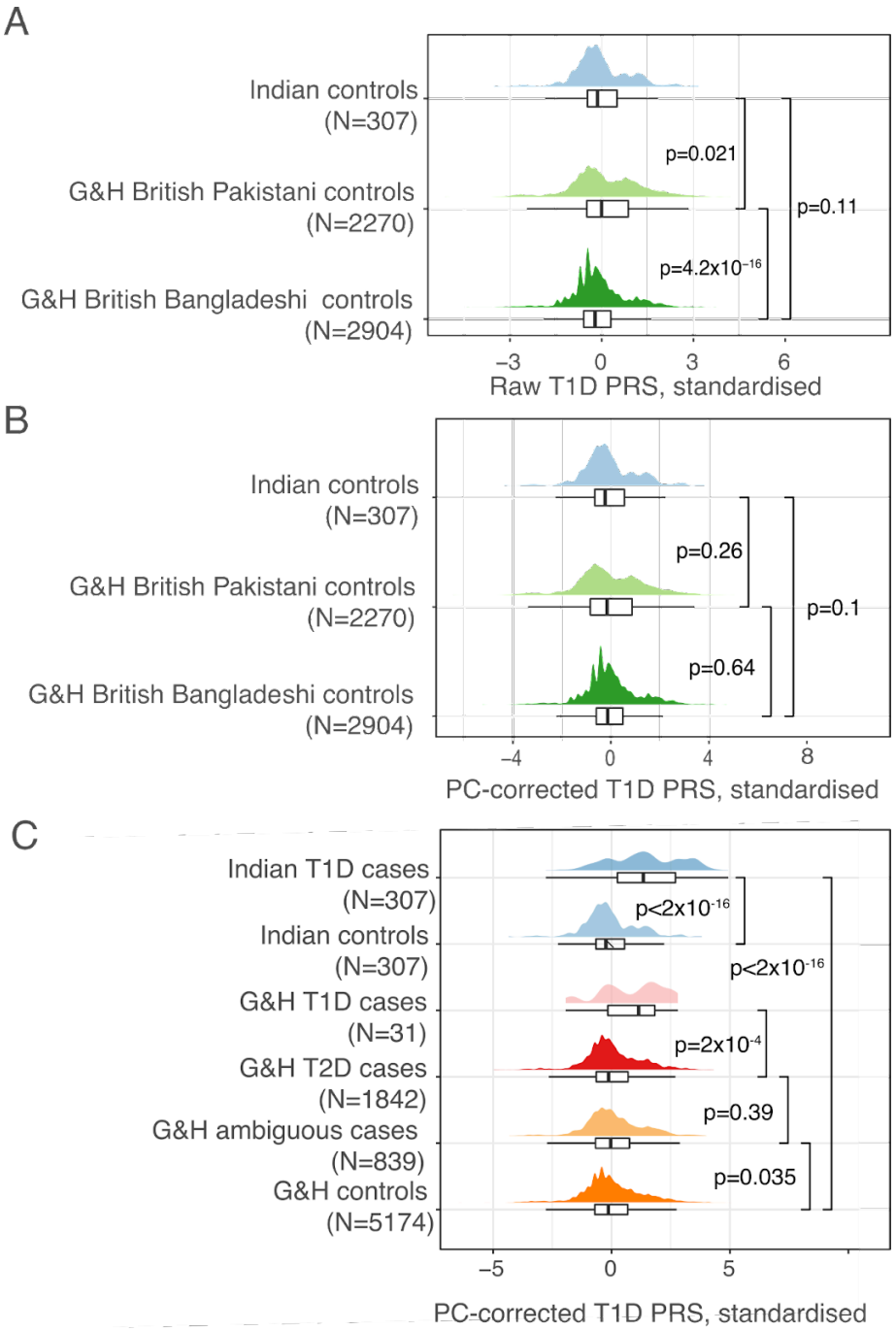
Supplementary Table 2: Clinical code lists used to define autoimmune diseases.

Supplementary Table 3: Clinical codes denoting people at risk of diabetes, which were used to exclude people from the control group. Codes included in the NHS Primary Care Quality and Outcomes Framework are denoted in blue.

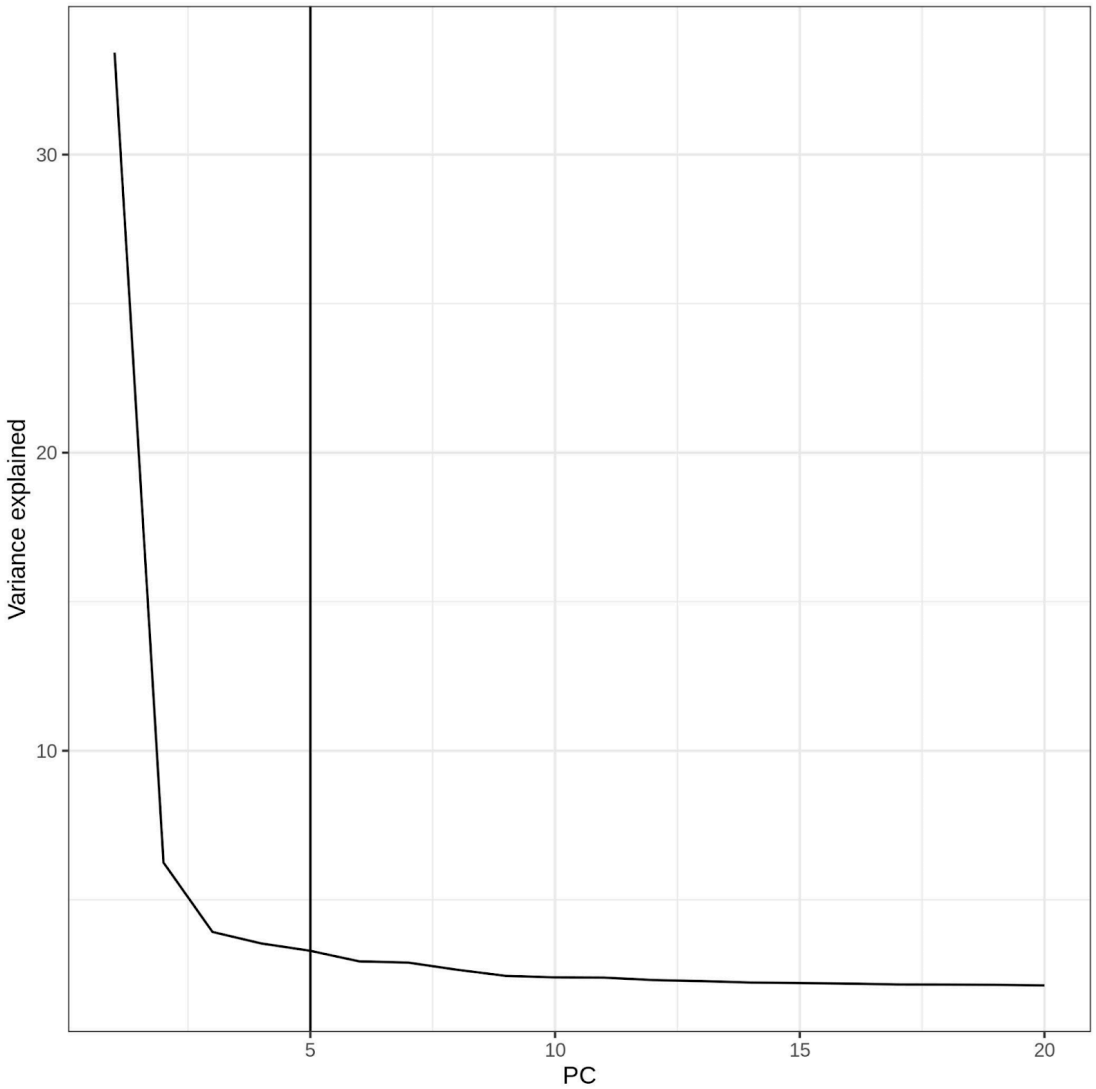
Supplementary Table 4. Numbers of samples remaining after each filter used to define T1D, T2D and ambiguous cases and the controls in G&H.

Supplementary Table 5: Exact p-values for post-hoc comparisons of clinical features between the groups in G&H. The group with which the comparison was made is indicated before the p-value. If the post-hoc comparison was not made (e.g. there was no statistical significance at ANOVA or if only two groups were compared for this feature), NA is given.

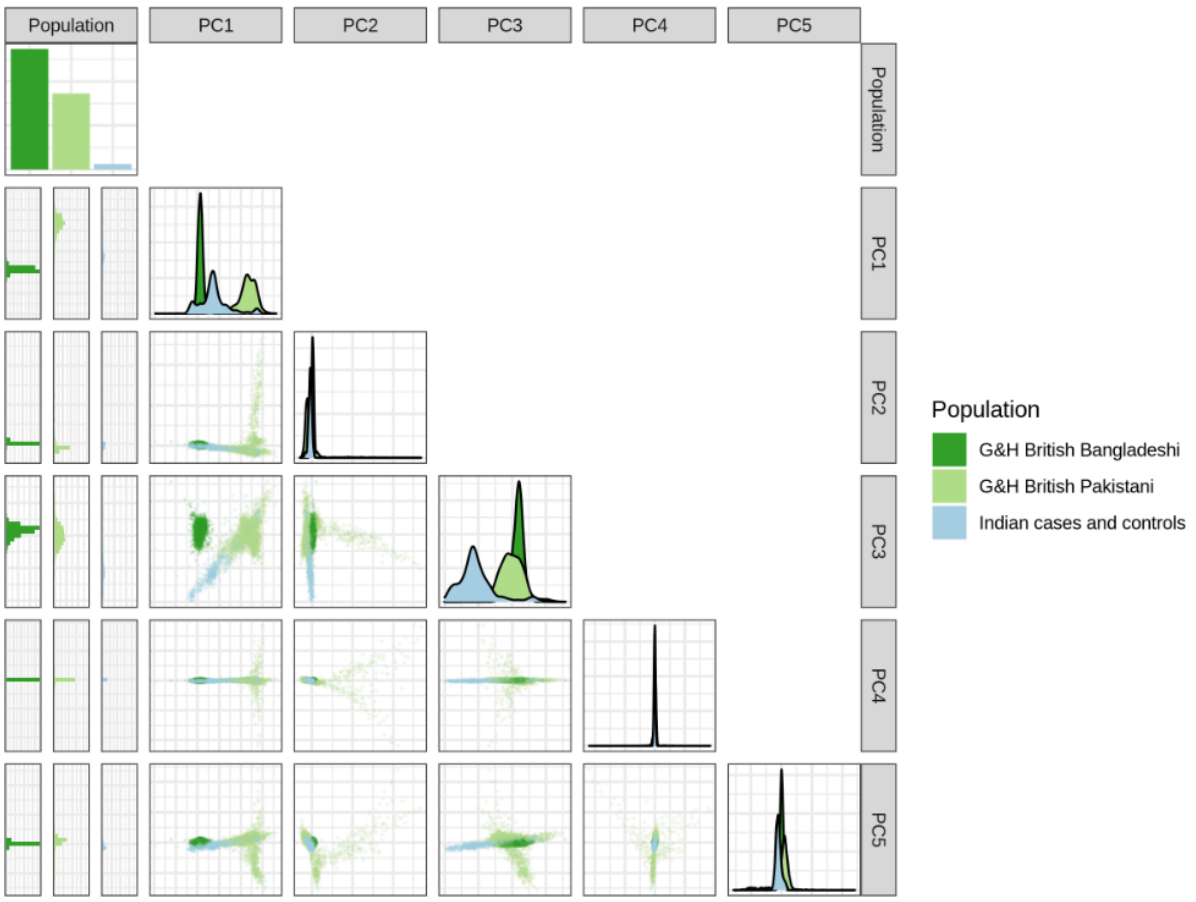
Supplementary Figures



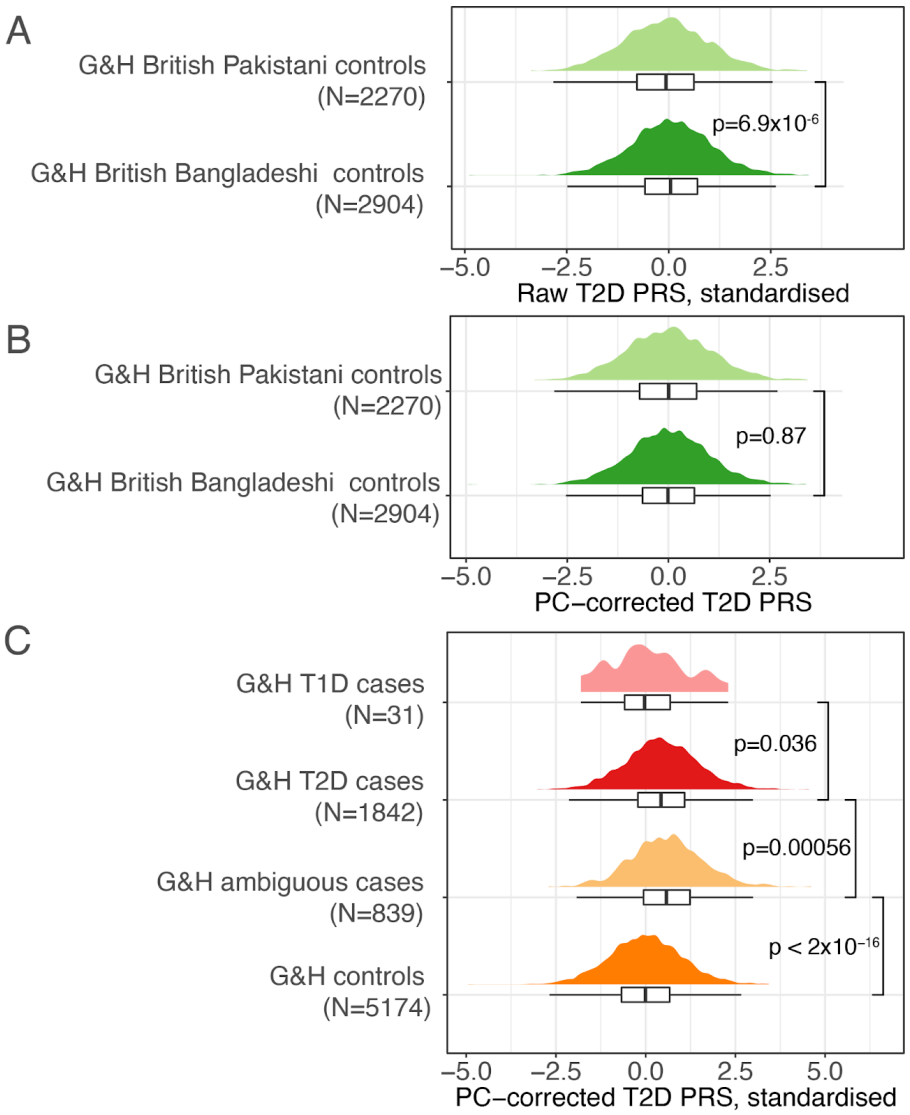
Supplementary Figure 1. Distribution of T1D polygenic risk scores (PRS) in the Indian and Genes and Health (G&H) clinical groups. A) and B) show the distributions in the control groups split by ancestry, before and after PC correction respectively. C) shows the distributions in the various groups of cases and controls used in the analyses, after PC-correction. P-values are from Wilcoxon signed-rank tests comparing the two indicated groups. Supplementary Figure 4 shows the equivalent distributions for the T2D PRS.



Supplementary Figure 2. Scree plot indicating the variance explained by principal components 1-20 from the PCA of G&H and Indian samples. Based on this, we decided to regress out five principal components from the PRSs.



Supplementary Figure 3. Principal component analysis of G&H and Indian samples (cases and controls combined).



Supplementary Figure 4. Distribution of T2D polygenic risk scores (PRS) in the G&H clinical groups. A) and B) show the distributions in the control groups split by ancestry, before and after PC correction respectively. C) shows the distributions in the various groups of cases and controls used in the analyses. P-values are from Wilcoxon signed-rank tests comparing the two indicated groups.