Genetic risk scores for diabetes diagnosis and precision medicine

Miriam S. Udler, Mark I McCarthy, Jose C. Florez, Anubha Mahajan

Endocrine Reviews
Endocrine Society

Submitted: April 26, 2019
Accepted: July 08, 2019
First Online: July 19, 2019

Endocrine Reviews; Copyright 2019 DOI: 10.1210/er.2019-00088

Genetic risk scores for diabetes diagnosis and precision medicine

Genetic risk scores for diabetes

Miriam S. Udler[1-4], Mark I McCarthy[5-7], Jose C. Florez[1-4], Anubha Mahajan[6]

1. *Diabetes Unit, Massachusetts General Hospital, 50 Staniford St, Boston, MA 02114*

2. *Center for Genomic Medicine, Massachusetts General Hospital, Simches Research Building, 185 Cambridge St, Boston, MA 02114*

3. *Programs in Metabolism and Medical & Population Genetics, Broad Institute of MIT and Harvard, 415 Main St, Cambridge, MA 02142*

4. *Department of Medicine, Harvard Medical School, 25 Shattuck Street, Boston MA 02115*

5. *Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LJ UK*

6. *Wellcome Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK*

7. *Oxford NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, OX3 9DU, UK*

ORCiD numbers:

0000-0003-3824-9162

Udler

Miriam S.

0000-0002-4393-0510

McCarthy

Mark I

0000-0002-1730-9325

Florez

Jose C.

0000-0001-5585-3420

Mahajan

Anubha

Received 26 April 2019. Accepted 08 July 2019.

ORCID identifiers

Mark McCarthy ORCID: 0000-0002-4393-0510

Anubha Mahajan ORCID: 0000-0001-5585-3420

Jose Florez ORCID: 0000-0002-1730-9325

Miriam Udler ORCID: 0000-0003-3824-9162

Over the last decade, there have been substantial advances in the identification and characterization of DNA sequence variants associated with individual predisposition to type 1

and type 2 diabetes. As well as providing insights into the molecular, cellular and physiological mechanisms involved in disease pathogenesis, these risk variants, when combined into a polygenic score, capture information on individual patterns of disease predisposition that have the potential to influence clinical management. In this review, we describe the various opportunities that polygenic scores provide: to predict diabetes risk, to support differential diagnosis, and to understand phenotypic and clinical heterogeneity. We also describe the challenges that will need to be overcome if this potential is to be fully realized.

SEARCH STRATEGY

The literature referenced in this article was selected for inclusion on the basis of the authors' expertise in this area of research, based on a broader set of publications sourced from PubMed and other repositories using relevant search terms, including, but not limited to, "polygenic scores", "risk scores", "precision medicine", "diabetes" and combinations thereof.

ESSENTIAL POINTS

- Over the last decade, there have been major advances in our understanding of the genetic basis of the most common subtypes of type 1 (T1D) and type 2 diabetes (T2D), with over 500 robust associations identified.
- Although individual variants typically have only a modest effect on risk, when combined into a polygenic score, they offer increasing power to capture information on individual patterns of disease predisposition with the potential to influence clinical management.
- The generation of polygenic scores based on overall T2D predisposition can identify individuals with a high future risk of diabetes who may benefit from targeted interventions.
- The generation of polygenic scores based on overall T1D risk can identify individuals who may benefit from early interventions to forestall the risk of T1D, and also supports the identification of those with later-onset diabetes who have an autoimmune etiology, for whom early recourse to insulin therapy may be advantageous.
- The generation of partitioned polygenic scores which capture aspects of the etiological and clinical heterogeneity that contributes to variable clinical outcomes in those with T2D has potential to deliver clinical benefit through enhanced capacity to predict disease progression, complication risk, and response to pharmacological and behavioral interventions.
- Polygenic scores have predominantly been derived from genetic studies performed in European populations and have suboptimal ability to capture risk in individuals of non-European origin.
- Though there are a number of technical and logistical issues to be addressed before the clinical utility of polygenic scores can be fully enumerated, increasing utilization of polygenic scores within diabetes clinical practice is likely to be an important component of efforts to deliver precision medicine for those who have, or are at risk of, diabetes.

1. Introduction

Diabetes is already one of the major contributors to death and ill-health globally, and its prevalence continues to rise. Current projections estimate almost 500M affected by diabetes as of 2017 (and almost 700M by 2045), most of this in the form of type 2 diabetes (T2D) [1]. Escalating rates of T2D speak to the limits of current strategies for prevention, whether they involve lifestyle interventions (for example through dietary modification and increased physical activity) or pharmacotherapy. At the same time, the burden of disease arising from the complications of inadequately controlled diabetes (manifest as renal failure, vision loss, amputation, and accelerated vascular disease) highlights the urgent need for major improvements regarding both the timely diagnosis of diabetes (since much damage is

initiated whilst the disease is subclinical) and the management of those with established disease.

The condition that we currently label as "type 2 diabetes" represents a convenient, but likely suboptimal, construct for the application of 21st century medicine. Though individuals with established T2D have a generalized metabolic derangement (typically associated with hyperlipidemia, adiposity, disturbed hepatic metabolism and the like), formal diagnosis rests entirely upon a single metabolic component (glucose), itself the end-result of multiple metabolic processes. The diagnosis of diabetes depends on numeric thresholds placed within continuous distributions of (fasting, random or postprandial) glucose and/or glycated hemoglobin levels. These thresholds were initially based around the observed relationships between levels of hyperglycemia and the incidence of specific diabetic complications, such as retinopathy, but they may not be equally discriminating for the macrovascular complications [2]. Crucially, T2D remains effectively a diagnosis of exclusion, made after those with hyperglycemia attributable to more defined causation including islet autoimmunity (type 1 diabetes [T1D]), highly penetrant genetic effects (e.g. maturity onset diabetes of the young [MODY]) and certain specified exposures (steroids, pancreatitis, pregnancy) have been excluded. Those left with the diagnosis of T2D demonstrate considerable heterogeneity with respect to presentation, clinical course, and response to available therapies, yet clinical pathways tend to be based around universally-applied algorithms that take little, if any, account of that heterogeneity [3-5].

Human genetics provides a powerful set of approaches for addressing some of these challenges, delivering both an improved understanding of the mechanisms contributing to the development of diabetes, and opportunities for direct translational benefit [6]. Both common major subtypes of diabetes (T1D, T2D) are complex, multifactorial traits: that is, an individual's risk of developing either of these conditions is influenced by the combination of genetic variation at multiple sites across the genome, acting in concert with factors within the external (e.g. nutritional availability, socio-economic status) and internal (e.g. microbiome, metabolic memory) environment [7,8]. Over the past decade, large-scale genetic studies (typically in the form of genome-wide association studies [GWAS]) have identified over 400 distinct genetic signals influencing T2D risk [9] and over 50 with impact on T1D predisposition [10]. Most of these DNA sequence variants are widely shared within and between populations, in contrast to the more private alleles that drive some rarer subtypes of diabetes [11,12]. With the notable exception of the HLA region (which has the major impact on T1D risk), most of these common variants have only modest effects on individual predisposition: the biggest effects for T2D modulate risk by no more than 40% per allele and most have much smaller effects [9,10].

However, in combination, the impact of this variation can be more profound [9,13]. In the most recent GWAS for T2D, the entire set of associated variants so far detected explains around 20% of the overall variation in disease risk [9], in Europeans at least (comparable analyses in non-European populations are limited by the sample sizes available for study). Estimates of the heritability of T2D vary widely [14,15] around a median of 40%, suggesting that around half the genetic contribution to the variation in risk can be quantified for each individual. Estimates of the heritability of T1D are higher [16] and a greater proportion of that genetic risk can be captured using existing approaches. Ongoing efforts to further characterize the genetic basis of both major subtypes of diabetes – through detecting significant associations at variants that have escaped detection because they are too rare, or have small effects – will increase the proportion of individual genetic predisposition that can be directly measured.

The steadily expanding list of genetic variants delivered by these successive waves of genetic discovery has delivered novel mechanistic insights into disease pathophysiology.

Some of these have led to an understanding of the major processes contributing to disease risk, such as the role of islet-specific as well as immunological processes with respect to T1D risk [17,18] or the relative impact of defects in insulin secretion and action for T2D [19]. Other studies have attempted to dissect the detailed molecular, genomic and physiological events that mediate risk at individual loci [9,20-23]. These efforts can have direct translational impact, for example through the identification of novel therapeutic targets, or biomarkers that track disease progression.

In this review, however, we focus on a different route from human genetics to translation, one that derives estimates of an individual's predisposition to diabetes and its subtypes (in the form of polygenic scores) from the patterns of individual genetic variation at sites known to influence diabetes predisposition.

2. The concept of polygenic scores

The idea of grouping genetic variants to capture the aggregate genetic risk for a given disease is not new. An early promise of genetic discovery in complex (polygenic) conditions was to predict clinical outcomes. It was recognized that, in contrast to classical Mendelian diseases, where the presence of a specific mutation was *deterministic* and typically heralded the eventual onset of disease (contingent on penetrance), the genetic risk for complex, multifactorial diseases is *probabilistic* and most appropriately used as a predictor that quantified a discrete increment in overall risk [24]. This is because for complex human traits, the overwhelming majority of associated genetic variants exert modest effects, and the ability of any individual variant to influence clinical outcomes is small. The obvious approach is to sum the effects of risk alleles associated with a given condition, to generate an aggregate estimate of genetic risk. This approach was justified by the observation that early genetic associations seemed to work in an additive fashion, with little or no evidence of epistasis. This concept was pioneered for age-related macular degeneration, the first disease for which GWAS proved successful [25] and had also been employed in T2D for the three reproducible genetic associations that had emerged from the pre-GWAS era [26].

This concept can be easily expanded from the disease arena to quantitative traits. Here, rather than expressing "risk" (which connotes the deleterious burden of illness), the aim is to capture the overall variance in a trait conferred by the set of genetic variants grouped into a composite score. Examples include circulating levels of a specific metabolite or the inherited predisposition toward a behavioral pattern, where the deleterious connotations ascribed to the term "risk" no longer apply. In this review, therefore, we favor the use of the term "**polygenic score**" as a more inclusive general descriptor.

The initial uses of polygenic scores deliberately focused on the inclusion of individual genetic variants for which the evidence for association was robust. This occurred as a "route correction" to the historical trend whereby the proliferation of candidate gene studies and the adoption of liberal statistical significance thresholds had led to the publication of many genetic associations which later proved irreproducible, and likely represented false positive findings [27]. In the GWAS era, such high-likelihood variants had to achieve *genome-wide significance*, based on a widely-accepted threshold of $p<5\times10^{-8}$, established to account for the estimated 1,000,000 independent tests that exist among common variants in the European genome) [28]. Thus, polygenic scores began to be constructed through the compilation of genome-wide significant variants emerging from successive and ever larger GWAS, each with increased statistical power [29].

In the literature to date, scores which only incorporate variants that are individually significant (typically weighted to reflect their respective effect sizes on the trait of interest), have often been described as "genetic [risk] scores", sometimes in contrast to use of the term "polygenic scores" to reflect those which build in additional sub-significant variants.

However, these terms have been applied inconsistently and sometimes interchangeably, and in this review we take the opportunity to (re)define these concepts with labels that are easier to interpret. Because the former are composed of variants at the top or extreme of the statistical distribution, we propose the term "**restricted-to-significant polygenic scores**" (rsPS). (**BOX; FIGURE 1**)

In T2D, the use of rsPS was pioneered in a series of publications in 2008, each of which constructed an rsPS from the 16-18 T2D risk variants known at the time [30-32] and compared its predictive performance to that of clinical T2D-risk factors. In Framingham samples, for instance, individuals with a "high" rsPS (score ≥21, ~11% of the cohort) had 2.6 higher odds of developing T2D, than those with a "low" rsPS (score ≤15, ~25% of the cohort) [31]. We discuss these rsPS studies in detail later in this review.

While compiling variants that achieve genome-wide significance ensures that the variants included in the score represent real associations with disease, such a stringent threshold ignores many other variants which, though truly associated with the phenotype, have escaped detection at genome-wide significance due to limited sample sizes. However, an estimate of their likely contribution is available in GWAS datasets, even if they fail to achieve genome-wide significance. Therefore, under the assumption that the effects of variants that have no association with disease will tend to cancel each other by random fluctuations around the null distribution, there is an opportunity to extend the polygenic score beyond the set of individually-significant variants (including potentially, all the variants from the GWAS dataset) in the expectation that the small cumulative effects of many hundreds or thousands of truly associated variants can contribute to the overall score, and improve power. In practice, the scores derived in this way do not typically include all variants. Typically, the full set of variants is pruned to selectively remove highly-correlated variants:this pruning is combined with an optimization step that evaluates the discriminative performance of different sets of variants (defined using a range of progressively more liberal association *p*-value cutoffs), to establish which cutoff maximizes the predictive signal [33]. We propose the term "**global extended polygenic scores**" (gePS) to describe these extended polygenic scores. We prefer "global" over "total" in this context, because current approaches do not capture all aspects of genetic risk: private variants, many structural variants and variants whose effect is modified by environmental factors are not optimally considered in these analysis.

The use of these global scores has been popularized recently with the assembly of large GWAS meta-analyses for multiple traits [13]. Their increase in content allows for a steeper and more granular estimation of risk along the gradient of genetic burden. These scores can include many tens of thousands, even millions, of variants. For example, one such gePS for T2D-risk, comprising 7M variants, was able to demonstrate that, in the UK Biobank, individuals in the top 3.5% of a T2D gePS (generated from and optimized in a subset of independent UK Biobank sample) had an odds ratio ≥3.0 when compared to the mean of the population [13].

The clinical manifestation of disease often reflects the confluence of multiple pathophysiological processes. In T2D, hyperglycemia typically requires the concomitant presence of insulin resistance and inadequate beta-cell function. Each of these may in turn be caused by various mechanisms, such as incretin insufficiency and/or resistance, fatty acid accumulation, glucolipotoxicity, diet, inflammation or the microbiome. To the extent that endophenotypes which reflect these processes can be captured in populations, one can estimate which of these processes is likely to mediate the T2D impact of each T2D-associated variant. Once these associations are established, discrete polygenic scores can be constructed with variants which share mediation of T2D-risk through a specific intermediary process. For example, early efforts to group variants in this fashion revealed that one of the processes contributing to T2D risk involves insulin resistance that is characterized by lower levels of

adiposity [34-36]. This paradoxical combination of phenotypes, which reflects the pattern seen in more extreme form in inherited lipodystrophies [37], likely reflects the consequences of an inherited defect in adipocyte development which limits the storage of excess lipid in "metabolically safe" fat depots.

One systematic approach to group variants in this way involves the use of clustering methods (described in more detail below). Here, investigators use orthogonal lines of evidence (e.g. association with physiological measures of insulin secretion or resistance, pattern of expression of tagged genes, or open chromatin regions) to group genetic variants associated with T2D into specific clusters informed by biology [20,38]. We term these **"partitioned" (or "process-specific") polygenic scores** (pPS), and explain below how these may help to define specific pathways that illuminate disease pathogenesis or highlight opportunities for potential pharmacological modulation. These partitioned scores may also, by capturing the endophenotypic profile driving an individual's progression from health to disease, provide a framework for tailored preventive or therapeutic interventions.

It is worth emphasizing a critical point that is often neglected in the enthusiastic embrace of the burgeoning power of human genetics. Because, for complex traits such as T1D and T2D, inherited sequence variation is only one component of predisposition, even the best possible distillation of genetic potential will never provide a complete description of individual risk. A fuller assessment of present and future disease state for an individual requires the integration of genetic information with accurate and robust measures of other contributions to individual predisposition (including diet, lifestyle and microbiome), and an assessment of current clinical state (including measurement of biomarkers such as glucose, lipids, islet autoantibodies, and clinical phenotypes such as BMI and WHR). The relative contributions of these various domains of information are likely to shift during life with measures of clinical state becoming ever more impactful in later life as disease becomes overt. However, as we will show, genetic variation has a critical part to play. The long-term stability of genetic variation, which is easily ascertained in peripheral blood, offers the potential for risk stratification throughout the life-course; unlike other risk factors, it is also not subject to the confounding effects of disease or its treatment.

3. Polygenic scores in action

3.1 Predicting T2D onset

The slow onset of T2D, coupled to evidence that the damaging consequences often predate the clinical diagnosis by some years [2], emphasizes the clinical value of early diagnosis. The capacity for drugs and lifestyle interventions to lead to substantial reductions in progression to diabetes [39,40] motivates efforts to identify those at the greatest future risk of developing T2D. As discussed above, genetic predictors have the particular advantage of offering predictive information that is stable throughout life.

Prior to the first GWAS for T2D, three genetic variants had been associated with T2D with high confidence: identified either through candidate gene analyses, or the follow up of linkage signals, these implicated *KCNJ11* p.E23K, *PPARG* p.P12A, and *TCF7L2* rs7903146. In 2006, Weedon *et al.* [26] assessed the combined risk of carrying these variants. As well as observing that the variants influenced T2D risk additively, the authors assessed the predictive value of the genetic tests using a standard approach that uses the trade-off between the sensitivity and specificity of the test to generate a receiver operator characteristics (ROC) curve. The area under this curve (the AUROC, or C-statistic) provides a measure of the proportion of times such a test will correctly assign disease state between a pair of individuals, one who has the disease of interest (or, depending on the study design, will would go on to develop it), and another who is not (or, who remains disease-free on follow-

up). The estimated AUROC was 0.58, exceeding the 0.50 value that indicates no discriminative capacity, but well short of the values seen for most clinically useful tests.

Publication of the first few rounds of T2D GWAS extended the number of significantly associated variants into the teens, enabling better powered studies (involving between 16 and 18 risk alleles) that sought to compare the value of an rsPS to predict incident diabetes to that of clinical factors alone [30-32]. Lyssenko and colleagues examined a 16-SNP rsPS in 16,061 Swedish and 2,770 Finnish subjects followed over a median of 23.5 years [30]. The rsPS alone (adjusted for age and sex) predicted diabetes incidence with an AUROC of 0.62, but this compared poorly to a mix of baseline clinical factors (age, sex, a family history of diabetes, BMI, blood pressure, triglycerides, fasting plasma glucose) that claimed an AUROC of 0.74. Adding the rsPS to these clinical factors had only a modest impact on performance, pushing the AUROC to 0.75. Adding genetic factors to clinical factors reclassified 9% and 20% of subjects from the Swedish and Finnish study subjects, respectively, to a higher risk category.

In a similar study, Meigs *et al.* [31] assessed an 18-SNP rsPS in 2,377 participants of the Framingham Offspring Study over 28 years of follow-up. The AUROC for incident diabetes with the rsPS alone (adjusted for age and sex) was 0.58 whereas an enhanced clinical model incorporating age, sex, family history, BMI, fasting glucose, systolic blood pressure, HDL cholesterol, and triglyceride levels reached 0.90. Adding genetic data to such a well performing clinical model left the AUROC unchanged, and resulted in risk reclassification of, at most, 4% of the subjects. A study of the power of an 18-SNP rsPS to capture T2D case-control status in 4,907 participants from Dundee (Scotland), reached similar conclusions: the AUROC for genetics alone was 0.60, whereas the equivalent metric for age, BMI, and sex was a vastly superior 0.78, with only a slight increment (to 0.80) for the combined analysis [32].

In the decade since, waves of successively larger T2D GWAS efforts have brought the number of significant loci discovered into the hundreds. Concomitant improvements in the performance of rsPS have been more modest. An updated analysis of a 62-SNP rsPS performed in the Framingham Offspring Study [41] generated a much-improved AUROC for T2D prediction (combined with age and sex) of 0.72, but as before, the addition of genetic information provided negligible improvement in performance over the equivalent clinical predictor (AUROC for clinical factors alone, 0.90; for the combined clinical and genetic score 0.91). Predictive performance in a second prospective study (CARDIA) was uniformally worse, particularly in participants of African descent [41].

The studies so far described employed rsPS, restricting the score to variants that demonstrated genome-wide significant associations. In principle, the expansion of the score to accommodate additional information from subthreshold variants should improve performance. Indeed, in a model-based analysis of predictive performance for T2D and other traits that extrapolated from estimates of GWAS effect-size distribution and heritability (as available in 2012), Chatterjee and colleagues deduced that a ten-fold increase in effective GWAS sample size for T2D (to ~220,000) would result in a boost in rsPS performance from 0.57 to 0.74, with a further increment in performance to 0.79 if a more liberal cut-off for variant inclusion was adopted [42].

Sample sizes on that kind of scale are now within reach for T2D, but as yet, those theoretical estimates have not been realized, most likely because some of the assumptions of the model, such as heritability, were overestimated**.** In analyses that update those reported in the original manuscript [9], we include here a gePS generated by Mahajan *et al.* from a T2D GWAS meta-analysis of almost 460,000 European individuals (effective sample size ~158,000) which captured around 20% of the variance in individual predisposition to T2D (about half the total estimated heritability) [9]. An optimized gePS comprising 171,249

variants was constructed using 5,639 cases and 112,307 controls from the UK Biobank and then used to predict T2D case-control status (as a proxy for prospective T2D incidence) in a separate set of 13,480 cases and 311,390 controls, also from the UK Biobank. The AUROC generated was 0.66 without adjustment for age and sex, increasing to 0.73 if age and sex were added (**Table 1; Figure 2**). Khera *et al.* [13] used a similar approach with a deeper gePS of almost 7M variants that, after factoring in age and sex, generated a similar AUROC (0.72). Both studies found that individuals from the UK Biobank (who were aged between 40 and 69 at recruitment, and tended to be relatively healthy) in the top 2.5-5% of the gePS distribution were at approximately 3-fold-increased risk (case-control prevalence of ~11%) compared with the mean of the rest of the sample and at almost 10-fold-increased risk compared with the bottom 2.5% (prevalence~1%) [9,13]. The former odds ratio could be expanded to ≥5.0 in individuals with the very highest gePS, though this high-risk group constituted only the ~150 individuals in the 0.05% extreme of the distribution [13].

One interesting feature to emerge from the re-analysis of the T2D polygenic scores shown in **Figure 2** (based on the data from [9]) is the limited increment in performance seen between the rsPS (which was based on 199 genome-wide significant SNPs) and the gePS (built from ~170K SNPs). A second observation is the reassuring concordance in the estimates of predictive performance obtained in these two studies [9,13], despite differences in the methods, though it is worth noting substantial overlap in the data sets used for training and testing (**Table 1**). Furthermore, these risk estimates are almost identical to those generated by the direct-to-consumer company 23andMe from their data set of ~1M individuals (mean age <50y) [43]. 23andMe have recently started sharing results from their 1244-SNP T2D gePS with their customers, with a recommendation that those deemed at high risk consider lifestyle interventions to mitigate that risk. A T2D-risk score generated by Genomics PLC had similar performance [45].

3.2 Predicting T1D onset

Whilst clinical management strategies exist to prevent the development of T2D in those determined to be at high risk [39,40], there is currently no known effective strategy to prevent T1D. Nevertheless, genetic profiling could have value in defining individuals at the highest future risk of T1D for enhanced surveillance or inclusion in trials of early immunologic interventions, and, in turn, when those trials are successful, could prove instrumental in stratifying those most likely to benefit from those new preventative approaches.

Like T2D, T1D has a substantial heritable component, estimated to be between 65 to 88% [46,47,48]. Genetic variation in the HLA region on chromosome 6p21 accounts for ~50% of that heritability [49]. The DR and DQ loci confer the strongest association with odds ratios as high as 16 for the DR4-DQ8/DR3-DQ2 genotype [50]. Subsequent GWAS have identified over 50 non-HLA genetic loci contributing to T1D risk, including SNPs near the *INS*, *PTPN22* and *CTLA4* genes with substantial impact on T1D risk [10,51-53].

Over the past 15 years, genetic prediction for T1D has evolved from the use of HLA alleles alone [54], to incorporation of over 40 non-HLA variants [55-58]. Two rsPS for T1D developed independently by Winkler and Oram and colleagues [56,57] were recently merged into a single rsPS including 41 HLA and non-HLA SNPs [59]. This 41-SNP rsPS was deployed within the TEDDY (T1D in the Environmental Determinants of Diabetes in the Young) study which followed several thousand children with high T1D-risk HLA genotypes from birth, using the development of islet autoantibodies and diabetes as outcomes indicating disease progression. The 41-SNP T1D rsPS successfully stratified risk: children with a score >14.4 had 11.0% risk of developing multiple islet autoantibodies by age 6 and 7.6% risk of diabetes by age 10, compared with those with scores below this who had rates of 4.1% and 2.7% respectively [59].

Leveraging advances in density of SNP arrays as well as larger reference panels, the most recently updated T1D rsPS includes 67 SNPs and accounts for interactions between 18 HLA DR-DQ combinations [60]. When applied to the UK Biobank, this enhanced T1D rsPS significantly outperformed previous scores, identifying individuals with T1D with AUROC of 0.92. These figures are close to the maximum performance figures predicted for T1D, based on the modelling analyses described earlier in the context of T2D [42].

3.3 Refining the diagnosis of major diabetes subtypes

As well as predicting future disease risk, polygenic scores are emerging as powerful tools to support diagnosis of major diabetes subtypes. Determining whether a particular patient has T1D, T2D, or one of the other specified forms of diabetes is not always straightforward. A clinical diagnosis of T1D can often, but not always, be substantiated by the presence of one or more islet autoantibodies (GAD, IA2, IAA, ZnT8), as these are found in >90% of newly diagnosed patients [61]. However, these antibodies are not always measured in clinical practice, and do not provide perfect determination of T1D diagnosis due to a combination of (i) background presence in some individuals without T1D, (ii) lower rates of positivity for T1D individuals diagnosed in adulthood, and (iii) waning titers over time from initial diagnosis [57]. The measurement of C-peptide levels in plasma or urine can also help distinguish T1D from other forms of diabetes, but use of this test is not routine, not least because it has reduced value at the time of diagnosis (where it can be suppressed even in T2D or monogenic diabetes) or during the “honeymoon period” of T1D, given residual beta-cell function in the early years following presentation [62]. The consequence is relatively high rates of both under- and over-diagnosis of T1D when trying to differentiate it from both T2D and less common forms of diabetes, such as MODY [63]. The stable nature of a polygenic score, unchanged throughout life, offers a useful tool to aid in diagnostic characterization of individuals with established diabetes.

An early application of the initial T1D rsPS developed by Oram and colleagues was in discriminating between T1D and T2D. The authors applied both a 69-SNP T2D rsPS and a 30-SNP T1D rsPS to a sample of well-defined cases of T1D and T2D from the Wellcome Trust Case Control Consortium GWAS [64]. They found the T1D rsPS was highly discriminative (AUROC 0.88), whereas the T2D EPS was less so (AUROC 0.64), and that combining the two offered little improvement beyond the T1D score alone (AUROC 0.89) [57].

Application of the 30-SNP T1D rsPS alone to a cohort of 223 adults, aged between 20 and 40 diagnosed with diabetes at least 3 years previously, predicted progression to insulin deficiency (AUROC 0.87) and offered information additional to that provided by antibody status [57]. In 8,608 individuals with a clinical diagnosis of T2D after 35 years of age, treated without insulin for at least 6 months following diagnosis, the same T1D rsPS predicted progression to insulin use at five years, but only in the small subset of GAD antibody-positive participants: the probability of insulin use ranged from 17.6% in those in the lowest tertile of T1D-risk to 47.9% in the highest [65].

T1D polygenic scores have also provided a clearer sense of the extent of T1D prevalence across the age spectrum. Using a 29-SNP T1D rsPS, Thomas and colleagues demonstrated that, amongst individuals participating in the UK Biobank, 42% of genetically-defined T1D was observed in those diagnosed with diabetes between 31 and 60 years, pointing to a far higher proportion of overall T1D presenting in adulthood than is commonly appreciated [66]. It can be challenging to detect these individuals clinically since, in this age range, they represent only a small minority (~4%) of patients with any form of diabetes. Compared to those with T2D, individuals with T1D defined on the basis of a high T1D rsPS had lower BMI, were more likely to use insulin in the first year of diagnosis, and were at higher risk of diabetic ketoacidosis [66].

T1D polygenic scores have also shown utility in discriminating early-onset T1D from monogenic forms of diabetes including MODY [67], neonatal diabetes [67], and monogenic autoimmune diabetes [68], that typically present during childhood. In these settings, a T1D score can prioritize patients who are most likely to benefit from sequence-based testing for rare causal variants, and support correct interpretation of novel variants of uncertain functional significance that emerge from such sequencing. Prioritization of patients in this way is important both for providing a cost-effective strategy to increase diagnosis rates for known forms of monogenic diabetes and for facilitating new gene discovery by reducing study subject heterogeneity. A related application of polygenic scores may be to explain some of the variable presentation of monogenic forms of diabetes, with respect to age of diagnosis for example [69]. The same variant within the *HNF1A* gene may segregate with early onset diabetes in some pedigrees, but also be observed in individuals who retain normal glucose tolerance into late adulthood and beyond [70]. Studying 410 individuals from 203 *HNF1A*-MODY families, Lango Allen and colleagues found that a 15-SNP T2D rsPS was significantly associated with earlier age of diabetes diagnosis, with each additional risk allele accelerating diagnosis by around four months [71].

3.4 Clinical application of predictive scores

These data provide a sound basis for the use of polygenic scores to support discrimination of major diabetes subtypes and lend credence to their wider clinical value. Given analogous applications of the polygenic score approach for other multifactorial disease traits [13,72,73], these findings have collectively bolstered excitement about their potential to deliver clinical benefit across a wide range of common diseases.

One major focus of current research activity lies in exploring the value of polygenic scores to predict individuals at the highest risk of T2D so as to enable early targeting of intervention strategies. If the estimates of relative risk seen in UK Biobank participants in recent studies generalize to the population level (and the current data indicate that performance seems to be sustained throughout the age ranges studied [9,43]), then there are likely to be in excess of one million individuals in the UK, who, on the basis of their polygenic score alone, have a ~50% lifetime risk of T2D [9]. With the price of whole-genome sequencing falling, and the potential to achieve near-perfect imputation by harnessing the combination of large-scale whole-genome sequencing (in a subset of the population) and dense GWAS arrays (in the rest), several countries are starting to plan for a future of universal genetic screening. The rationale is that “one-time” measurement of genome-wide genetic variation (achievable for the cost similar to that of a single outpatient appointment or a chest X-ray), would support a wide range of clinical applications throughout a person’s lifetime, including, but not limited to, the optimization of therapies (based on pharmacogenetic insights) and the prediction of future illness using polygenic scores for a range of diseases.

However, there are clearly multiple obstacles to be overcome before this becomes the standard of care.

Firstly, there are technical issues. The most critical amongst these involves ensuring that polygenic scores are appropriately calibrated to the ethnicity of the individual being tested. An rsPS or gePS generated using data solely derived from Europeans will have suboptimal ability to capture risk in individuals of non-European origin. The T2D gePS recently released by 23andMe demonstrates a marked fall-off in predictive performance in individuals of Asian and African-American origin [43]. In some settings, these issues with the transethnic portability of polygenic scores go beyond a simple dilution of performance: unpredictable biases and the consequences of genetic drift can result in entirely misleading results [74,75]. Recent studies have also emphasized the impact of residual population stratification effects

on the performance of these scores [76,77]. RsPS are likely to be more robust to these biases than gePS.

The second question to be addressed concerns whether a given polygenic score adds clinical value to the predictions that are possible using existing risk factors. In the case of coronary artery disease, there is evidence that a substantial proportion of those at highest polygenic risk would not have been detected using classical risk factors [13]. In contrast, and as described earlier, the incremental benefit of a polygenic score over easily-accessible clinical parameters seems more limited for T2D, at least when applied at older ages. In fact, the non-genetic risk factors we already collect in clinic (family history, ethnicity, BMI, fat distribution) perform quite well in predicting T2D, particularly in the near term, especially when supported by direct biomarkers of the underlying disease process such as measures of glycemia [30-32,41]. There is an intrinsic limitation to the added value of a polygenic score arising from the fact that trait heritability provides a ceiling for the performance of any purely-genetic measure.

Third, there is the issue as to whether early diagnosis can be shown to result in beneficial outcomes, for example by motivating improvements in lifestyle or treatment that reduce the risk of disease. In the case of T2D, the potential for lifestyle modification and/or pharmaceutical intervention (for example with metformin) to reduce diabetes progression is clear [39,40], and these benefits seem to accrue irrespective of genetic risk. In the Diabetes Prevention Program, for example, lifestyle intervention was effective at reducing diabetes incidence compared to placebo even among those with the highest quartile of T2D rsPS [78]. However, there is limited evidence to date that the communication of genetic risk is sufficient to motivate most individuals to undertake the kind of long-term behavioral modification required for sustained benefit [79-81]. There is also some (at least theoretical) risk of harm if the communication of risk information is mishandled. This could arise through failure to use ethnically appropriate scores, or to incorporate other relevant health information. For example, an overweight person with a low T2D polygenic score may be at far greater risk of disease than the polygenic score alone would suggest. Some individuals may be liable to interpret high genetic risk in a deterministic and fatalistic way, failing to appreciate that remediation of risk through lifestyle modification is no less likely to be effective in their case.

Finally, there are questions related to implementation. Several countries (Finland, Estonia, UK, Taiwan, amongst others) are expanding the clinical roll-out of genome-wide genetic data, with plans to deliver genetic profiling to the population scale through a combination of sequence- and array-based strategies. Such universal availability of genomic data would open up much wider use of polygenic scores: the costs of acquiring such data (which only needs to be done once in the life of the individual) could be amortized across multiple applications (rather than needing to be justified based on any single indication) and the marginal costs of any specific use of those data would be minimal. Having said that, any valid assessment of clinical utility needs to consider the full costs of any given application: if the consequence of the unregulated use of genetic information is to identify a large proportion of the population as at high risk, there may be substantial financial and health costs to be incurred in follow-up screening, unnecessary treatment, patient stress, and the unproductive use of medical resources. A rigorous pipeline for the interpretation of these findings and their translation into evidence-based clinical interventions at the point of care will need to be created and deployed for multiple phenotypes across health care systems.

4) Partitioned polygenic risk scores

So far, in this review, we have focused on the use of restricted (rsPS) and expanded (gePS) polygenic scores, both of which aim to capture the genetic contribution to predisposition for the major disease phenotypes conventionally used to define morbid states – such as T1D and

T2D. These scores are designed to enable prediction of an individual's risk of developing of one of these forms of diabetes, or, as described above, to support differential diagnosis in those who have recently been diagnosed with diabetes. For these indications, it makes sense to combine as many risk variants as possible, irrespective of the mechanisms through which they influence that risk.

However, these are not the only clinical questions that polygenic scores are equipped to address. Many of the most difficult problems in the clinical management of T2D, in particular, arise out of the clinical and phenotypic heterogeneity that is an obvious feature of this condition. Clinical management of someone with a diagnosis of T2D would be substantially improved if it were possible to sense how fast their diabetes is likely to progress, their propensity for developing macrovascular and microvascular complications, and their likely response to the range of treatments (therapeutic, surgical, and behavioral) that could be deployed to improve outcomes. Since these are questions that relate to clinical and etiological heterogeneity in those with established T2D, polygenic scores based on overall disease risk are unlikely to offer discriminatory value.

As discussed earlier, one promising route to capture elements of this clinical heterogeneity is through the use of "partitioned risk scores" (pPS). These seek to "deconstruct" the overall (restricted or extended) polygenic score along biological axes that represent contributory etiological pathways, and thereby provide a framework upon which to map the variable response to clinical outcomes.

One way of conceptualizing these pPS is in terms of the "palette" model of diabetes predisposition, which seeks to focus attention not on T2D itself, but on the various intermediary processes that collectively contribute to T2D-risk [3,38]. These include well-studied processes such as obesity, fat distribution, islet development and function, and insulin sensitivity, though there are likely to be others that are, as yet, less clearly described. Each of these processes is itself under multifactorial (genetic and non-genetic) control, and a given individual may be positioned at any point on the spectrum from "low-T2D-risk" to "high-T2D-risk" for each of these. Whilst the overall load of T2D-risk across the set of processes is likely to be a useful measure of the overall T2D-risk of an individual, the disposition of that risk across the various axes is likely to be more informative regarding disease presentation and clinical course. In accordance with the "palette" analogy, each of these processes can be considered to be represented by a particular base color (red, blue, yellow etc): for any given individual, risk along each axis would be captured by the saturation of the relevant base color, and their overall profile of T2D-predisposition visualized in terms of the mix of those colors which results when they are combined.

This "palette" model is consistent with current understanding of the pathogenesis and the genetic architecture of T2D. Over the past decade, T2D-associated variants have been shown to modulate T2D risk through diverse mechanisms: some increase T2D risk through an impact on obesity (e.g. *FTO*), others reduce insulin sensitivity (e.g. *PPARG*, *IRS1*) whilst others compromise insulin secretion, either through direct effects on islet function (e.g. *KCNJ11)* or development (e.g. *HNF1A*), or indirectly through impact on incretin signalling (e.g. *GLP1R)* [82]. The various classes of T2D therapeutics operate through the same range of mechanisms to reverse the diabetic phenotype or control its glycemic consequences. The weight of evidence indicating that the genetic contribution to T2D predisposition mostly arises from common variants of limited individual effect [11,12] emphasizes the need to think in terms of a gradation of polygenic risk across individuals, rather than a classification based around rigid, discrete subtypes [3]. As well as providing a framework for capturing the mechanistic basis of T2D heterogeneity, this model also offers an approach to understanding how an individual's particular genetic profile contributes to their progression from normal metabolic health towards the diabetic state.

In 2010, Voight *et al.* were first to demonstrate that patterns of genetic association across diabetes-related quantitative traits could be utilized to annotate T2D-risk loci with respect to their physiological impact, analyses which highlighted the predominant role played by variants influencing insulin secretion [19]. This approach was further developed by Dimas and colleagues [83] to perform a systematic analysis of the relationships between 36 T2D-risk alleles and a range of glycemic measures including indices of insulin secretion and insulin resistance gathered in nondiabetic individuals. Scott and colleagues extended this approach to a larger set of 93 T2D-risk alleles and included BMI and lipid measures in their clustering in addition to glycemic traits [44]. Three main patterns of multi-trait association emerged from this analysis, two of them reflecting defects in insulin secretion and insulin action respectively, and a third characterized by obesity and dyslipidemia. One major limitation of the unsupervised hierarchical "hard" clustering approach used in these papers [44,83] is that it requires each variant to be assigned to a single cluster, based on the questionable assumption that each variant can only be involved in one pathophysiological pathway.

Access to an expanded range of large-scale quantitative trait association data (from large-scale GWAS efforts within global consortia such as GIANT [anthropometric traits], MAGIC [continuous glycemic traits] and GLGC [lipids]) plus advancements in clustering algorithms have enabled a new wave of variant clustering analyses [20,38]. These described efforts to aggregate GWAS data from more diverse sets of T2D-related quantitative traits and employed more sophisticated "soft" clustering techniques [84,85] to pick out clusters of T2D-associated variants with similar patterns of impact across the suite of phenotypes. These soft clustering approaches explicitly allow for the possibility that a variant influences more than one process. Mahajan *et al.* [20] deployed a C-means clustering approach across GWAS data from 10 T2D-related quantitative traits for a set of 94 T2D association signals that emerged from a T2D-GWAS of ~450K individuals, identifying 6 variant clusters (based on a threshold of 80% for cluster membership). Udler *et al.* [38] employed a complementary soft clustering approach - Bayesian nonnegative matrix factorization - to a partly overlapping set of 94 T2D-risk variants, gathering GWAS data from 47 diabetes-related traits, and identifying five clusters. Reassuringly, despite these differences, the clusters identified by both were broadly similar (**Table 2**).

The variants within each of the genetic clusters can be used to generate "partitioned" polygenic scores that capture the genetic contribution to each intermediary process. Each of these clusters (and the pPS generated therefrom) can be assigned mechanistic labels based on the observed patterns of GWAS effects: for example, a cluster which features T2D risk alleles most clearly associated with decreased fasting insulin, can, on the basis of known pathophysiological relationships, be attributed to reduced insulin secretion. On this basis, two of the clusters were associated with an adverse impact on beta-cell function, three were characterized by insulin sensitivity (differing with respect to their relationship to obesity, fat distribution, and lipid metabolism), and a sixth cluster (designated only in the Mahajan *et al.* paper) had less clearcut phenotypic features [20,38] (**Table 2**).

The T2D-risk variants assigned to the three insulin sensitivity clusters displayed the most obvious overlap across the two approaches. Variants near *FTO*, *MC4R*, and *NRXN3*, all loci known to have substantial impact on variation in BMI, mapped to a cluster of T2D-risk variants thereby assumed to be driven primarily by obesity. Variants at *IRS1*, *PPARG*, and *KLF14* implicated in effects on adipocyte differentiation and body fat distribution, were co-located to a cluster of T2D-risk variants featuring lipodystrophy-like effects on insulin sensitivity, partly overlapping with the set of "favorable adiposity" loci identified by others [34-36]. Finally, variants at *GCKR* and *TM6SF2*, known for their profound impact on ectopic

fat accumulation in liver and altered circulating lipid levels [86,87] were members of a cluster which seems to be driven by alterations in hepatic metabolism.

Though there was broad agreement concerning the variants deemed to influence beta-cell function, disposition across the pair of beta-cell clusters was less consistent, particularly for variants with less dramatic effects on the continuous glycemic traits that distinguished them. T2D-risk variants at *SLC30A8*, *TCF7L2*, *ADCY5*, *HNF1A*, and *MTNR1B* consistently mapped to a cluster characterized by an association between T2D-risk, reduced insulin levels but elevated proinsulin levels, whilst those at *ARAP1*, *IGFBP2*, *DGKB*, and *CCND2*, combined T2D-risk and reduced beta-cell function with reduced proinsulin levels. Some of the variation in the assignment of other variants across these two clusters reflects differences in the traits included in the respective analyses, compounded by substantial differences in the size of the GWAS data sets available across traits (which has an impact on discriminatory power). Nevertheless, the replicated subdivision of beta-cell function variants into two clusters distinguished by the direction of the association to proinsulin speaks to two distinctive mechanisms whereby T2D-associated variation results in beta-cell dysfunction [88].

Despite some of the differences in the assignment of individual variants across clusters, the mechanistic basis of these clusters appears robust, mapping as it does to current understanding concerning the major pathophysiological processes influencing T2D development. Allocation of variants to these physiologically-defined clusters is also broadly supported by orthogonal analyses of tissue-specific patterns of chromatin accessibility, histone modification, and transcriptional regulation. The various subsets of T2D-risk variants identified by clustering of GWAS data demonstrate clear evidence of genome-wide enrichment with respect to tissue-specific active enhancers and promoters [9,38,44,89-91], *cis*-eQTL signals [90,92], and enhanced connectivity in tissue-specific protein-protein interaction networks [93]. As anticipated, these link variants in the insulin secretion clusters to altered transcriptional regulation in the islet, and those in insulin action clusters to events in liver, fat and muscle.

Beyond the ability of these efforts to identify disease pathways, a critical question in terms of clinical translation is whether or not the pPS generated from these clusters show associations with clinically relevant outcomes: early results are encouraging. For example, differential cluster associations have been observed for coronary artery disease, stroke, and the renal complications of diabetes [38,94,95], each emphasizing enhanced risk associated with T2D predisposition mediated through insulin resistance. In the case of macrovascular disease, of course, this is likely to reflect the pleiotropic impact of these variants on non-glycemic risk factors such as lipids. A specific role for pPS-captured defects in insulin secretion and altered gut microbiome has also been reported: those microbiome changes include an effect on butyrate-producing pathways shown to play a causal role with respect to diabetic and obesity phenotypes [8].

These findings support the notion that whilst, by definition, all cluster-defined pPS associate with T2D risk, differential effects can be detected with respect to aspects of mechanism, phenotype, and clinical outcomes. However, further effort is needed to validate and extend these findings, and to define the contribution that these can make to the delivery of more personalized management in diabetes. So far, clustering analyses have been restricted to a subset of the most robust genome-wide significant T2D-associated variants, primarily those discovered in Europeans, and for which association statistics are available across multiple related traits. More complete analyses (delving deeper into the list of T2D-associated variants, and embracing a wider range of traits) capable of generating more powerful pPS will become possible as GWAS efforts for those other traits scale up. Inclusion of additional phenotypes should provide more granular clustering, attributing mechanism to

variants which currently show only weak phenotypic features, and bringing to light new pathways involved in T2D development. Integration with tissue- and cell-type-specific regulatory annotation maps will continue to support mechanistic inference [38,44]. Greater access to association data on T2D and other traits from non-European ethnicities will enable broader exploration of ethnic-specific variants and the heterogeneity of clinical presentation and course across major ethnic groups. As confidence grows in the mechanistic basis of these variant clusters, it will become possible to use trait-specific GWAS data to "build out" cognate pPSs and generate more powerful genetic instruments. For example, the pPS formed from the handful of genome-wide significant T2D variants in the "obesity" cluster could be superseded by using a polygenic score constructed from the BMI GWAS efforts themselves, and a pPS capturing islet autoimmunity generated from existing polygenic scores for T1D.

For diseases such as T2D, the characterization of clinical phenotype using genetic measures alone is constrained by the fact that individual variation within each of the endophenotypic axes is also influenced by non-genetic factors. Diagnostic and predictive accuracy would be much improved, and the ability to track an individual's journey from health to disease much enhanced, if the genetic contribution to phenotypic variation (as captured by the pPS) can be integrated with robust longitudinal measures of relevant features of the external environment (e.g. related to diet and physical activity) and internal milieu (e.g. metabolic memory and microbiome). Integration of this "predictive" information with evolving measures of the individual's clinical state would add another dimension. In the context of T2D, the latter would involve capturing anthropometric data, and glycemic and metabolic state, forming an integrated profile of that individual that can be tracked over time.

It would be particularly valuable in this regard to develop process-specific biomarkers that provide clinical readouts for each of the endophenotypic axes that corresponds to a particular pPS. The best illustration of this concept is the use of LDL-cholesterol as an integrated biomarker for that component of cardiovascular risk attributable to genetic and environmental influences on lipoprotein metabolism. The growing availability of large, publicly-available metabolomic and proteomic datasets makes it possible to use pPS as instruments to identify biomarkers correlated to pPS-defined risk as candidates for further prospective testing [96,97].

A key focus of ongoing research relates to understanding how these pPS might be deployed in clinical practice. One interesting possibility is that pPS profiling will allow identification of individuals whose diabetes is mostly attributable to defects in a single process. In the analysis by Udler *et al*. [38], one third of individuals fell within the top decile of T2D-risk for at least one cluster and, of these, 75% were not placed at the top decile of any other cluster. These individuals would be obvious recruits for the testing of targeted interventions. An alternative, possibly complementary, approach would make use of the full range of scores for a given individual to assign risk, and optimize management. In either case, much will depend on the extent to which these various ways of representing etiological heterogeneity (with or without additional environmental and clinical state information) can be shown to optimize clinical management (for example, the selection of therapeutic agents).

One important corollary is that, by conceptualizing a disease such as T2D as arising from the coming together of diverse, largely-orthogonal underlying processes, these models question some of the tacit concepts underlying precision medicine. One of these is the notion that characterization of the specific defect contributing to an individual's disease invites therapeutic approaches that are designed to specifically correct it. This model has proven effective in monogenic diabetes – where one molecular defect is largely responsible for the phenotype – but it is less clear this can be implemented in polygenic disease. In people in whom the disease is caused by multiple processes, it will be unlikely that modulating a single pathway will be sufficient to correct metabolic derangements; whereas in those in whom the

contributions of specific genetic defects are modest, equivalent reductions in disease risk and progression may be possible through interventions that boost the performance of other processes contributing to overall T2D risk, even those that are already performing at healthy levels. Indeed, because the effects of common variants on the hyperglycemic phenotype are modest, current T2D drugs that target specific pathways (e.g. sulfonylureas and thiazolidinediones) appear to be effective in both carriers and non-carriers of T2D-associated alleles in the respective target-encoding genes [98,99]. Nevertheless, it is possible that some individuals will be identified whose pathophysiology is predominantly driven by one process, and in whom the monogenic paradigm of a drug targeting that very process could be applied effectively. Whether, and in whom, such approaches may prove successful will require the conduct of appropriately designed precision clinical trials.

These pPS approaches to analysing phenotypic heterogeneity, which build out from genetic risk, offer a complementary perspective to the results emerging from the analysis of real world data [100, 101]. These real-world methods have focused on efforts to classify T2D into distinct subtypes, analogous to the categorization of monogenic forms of disease. Such an objective, if successful, would offer clinical expediency.

However, these efforts to sift individuals into discrete subtypes of disease would appear to run counter to the evidence that points to a complex, graded, architecture of risk, one that is consistent with a multifactorial etiology, composed of genetic predisposition dominated by multiple common variants of modest effect, and pervasive exposures contributing to risk. In one recent study, Ahlqvist *et al.* used basic clinical information from patients with newly-diagnosed adult-onset diabetes, to define five subtypes of T2D: an autoimmune form (covering T1D and other related clinical entities), two severe forms (one dominated by insulin deficiency, the other by insulin resistance), and two milder forms (termed “obesity-related“ and “age-related” diabetes) [101]. Whereas the genetic clusters that form the basis of pPS are defined at the level of the variants, these clinical subtypes are defined at the level of the individual, and based on biomarkers and clinical data gathered at a specific point in the progression of an individual from health disease. The latter is likely to limit their relevance to those who have not yet developed disease, and/or those who are on treatment.

It is worth emphasizing the different, but complementary, nature of these two approaches: the partitioned risk approach involves first clustering genetic signals by mechanism to derive pPS, and then exploring how the quantitative pPS scores perform across individuals. In contrast, the phenotypic clustering approach attempts to hard cluster individuals on the basis of their physiology. Further work is required to understand how these two approaches to capturing clinical heterogeneity relate to each other, and to objective measures of clinical utility. One of the fundamental issues – which pervades diverse aspects of precision medicine – relates to the relative merits of retaining as much quantitative information on an individual as possible until the point when a substantive (typically binary) clinical decision needs to be made, as opposed to early diagnostic categorization of the individual in a way that bases subsequent clinical decision-making on the optimized outcomes of the group to which they have been assigned. While further investigation is needed, a recent analysis by Dennis *et al.* in the ADOPT and RECORD clinical trials indicated that the former approach – considering phenotypic traits as continuous measures – provided better predictive value of treatment response, than an approach that binned individuals using the phenotypic clustering approach of Ahlqvist *et al* [101, 102].

5) Summary and further discussion

After many years of frustration at the slow progress that had been made in the translation of recent discoveries in human genetics – notably the many risk variants for common, multifactorial forms of diabetes identified through GWAS and sequencing – there is now

growing optimism that the use of polygenic scores will offer substantial clinical benefit and contribute to efforts to forestall the growing morbidity and mortality associated with these conditions. Some early clinical applications have emerged – mostly related to positive identification of those who have developed, or at highest imminent risk of developing, T1D [57,65-68].

It is inevitable that clinical applications of the polygenic score approach will roll out at a different pace across disease conditions, with a focus on different clinical questions, dictated by the additional clinical benefit that they provide, and the extent of the unmet clinical need. One size certainly does not fit all, and the relative merits of the different types of polygenic score described in this review (gePS, rsPS, pPS) will differ according to the specific clinical situation. It also remains to be determined how or whether pPS and phenotypic trait clustering will impact clinical care and be deployed in practice.

Recent developments in relation to the potential clinical use of polygenic scores have led to heated debate between those who are enthusiastic about the potential, and those who are of the view that the clinical value of human genetics discovery has been consistently hyped, and who feel that polygenic scores represent just the latest chapter in that story of scientific over-selling [103,104]. As in other similar situations, the outcome of this debate will become clearer as theoretical and basic knowledge develops and the collection of real-world data expands. However, it is already possible to identify a series of obstacles that need to be overcome before the full potential of this approach can be realized.

The most critical is the need to ensure that the benefits of accurate, robust polygenic score determination are equally available to all. As others have pointed out, most GWAS and sequence data have been derived from the European-descent individuals who live in the developed nations of Europe and North America, and the polygenic scores generated from these data perform best when applied to the same populations [74,75]. There is a critical need to generate equivalent data and polygenic scores in other populations, to explore and characterize the extent to which transethnic portability of polygenic scores can be tolerated, and to define strategies for their deployment in special situations such as recently-admixed and isolate populations. Concerns about the impact of population stratification and the limits of transethnic portability provide arguments for the use of rsPS over gePS [74-77]. This may be particularly true for T1D and T2D given the limited increment in performance available with more extended scores.

Wider recognition needs to be given that, for multifactorial traits with an appreciable non-genetic component, a wholly genetic explanation of disease prediction and state will never provide a perfect clinical instrument. In some settings, the information from genetics may simply recapitulate measures already available from other risk factors. The clinical use of cholesterol measures as a biomarker for CAD risk provides a counterexample, reflecting the benefits it offers as an integrator of both genetic and environmental risk. At the same time, some of those who are less enthusiastic about the clinical value of polygenic scores often fail to acknowledge that many established clinical tools (for example the use of BMI to predict T2D risk, or the use of islet cell antibodies for the differential diagnosis of T1D in late-onset diabetes) are likely to have performance metrics that limit their discriminative power. As the costs associated with the generation and interpretation of individual genomic information decline, there will be a growing roster of clinical applications where polygenic scores can add value.

There is clearly a need to develop novel approaches to establish the clinical validity and utility of polygenic scores in medical practice that take account not just of the marginal cost of acquiring the data, but the full costs of implementation. Randomized clinical trials are unlikely to be the answer here, not least because the dynamic nature of the underpinning genetic databases means that polygenic scores are likely to evolve, rapidly rendering

redundant any precise quantification of cost and benefit based around a historical set of scores [106]. There will need to be concomitant efforts to document the provenance, content and performance of polygenic scores using standardized metrics and conventions which do not currently exist.

There will need to be education of citizens and professionals to appreciate the benefits and limitations of polygenic scores [105]. It should be clear that genetics represents only one contributor to individual disease risk and profile, that genetically-defined risk should not, for multifactorial traits at least, be considered deterministic, and that most of the evidence indicates that behavioral modifications are just as likely to succeed (and in fact to be even more beneficial) in those at highest genetic risk [78]. The ease with which polygenic score information can be integrated with conventional approaches to risk profiling that are already widely used in clinical practice (e.g. to estimate future risk of CAD) should facilitate widespread introduction, and minimize the need for the health care professionals involved to develop an intimate knowledge of human genetics. It goes without saying that any clinical application of genetic data will need to fully address issues related to privacy and informed consent [107].

At the heart of precision medicine is the notion that an improved specification of disease risk or subtype will allow better targeted interventions to prevent or treat disease. Such efforts must compete for resources with population-based interventions that seek to achieve the same ends through non-targeted means [108]. In many existing clinical settings (e.g. related to reducing rates of cardiovascular disease, melanoma or breast cancer), these two strategies are seen to be complementary and are pursued in parallel. The development of polygenic score-based approaches to support targeting of high-risk individuals will not alter these assessments. As now, the balance of effort between targeted and non-targeted approaches to the reduction of disease and disability will, for any clinical indication, continue to be dependent on the relative impact, cost, acceptability and sustainability of these complementary strategies.

Box: Polygenic Score Terminology used in this article

BOX: Polygenic score terminology

1. Restricted-to-significant Polygenic Scores (rsPS): scores composed of variants at the extreme of a statistical distribution, most usually those that pass the genome-wide significant threshold for the trait concerned.
2. Global extended Polygenic Scores (gePS): scores generated from a deeper set of variants generated from genome-wide analyses, typically involving large numbers of sub-threshold significant variants.
3. Partitioned or Process-specific Polygenic Scores (pPS): scores composed of variants grouped according to some common biological process (e.g. association with a related endophenotype, tissue expression of related genes, chromatin state)

ACKNOWLEDGEMENTS

We acknowledge the timely assistance from Amit Khera (Cambridge, MA) and Michael Multhaup and colleagues at 23andMe (Mountain View, CA) who provided additional details concerning the basis of the T2D polygenic scores that allowed us to complete the comparisons reported in this paper.

GRANTS and FELLOWSHIPS

MMcC is a Wellcome Investigator and an NIHR Senior Investigator. Relevant funding support for this work comes from Wellcome (090532, 106130, 098381, 203141, 212259), NIDDK (U01-DK105535), and NIHR (NF-SI-0617-10090). JCF is supported by NIH grants U01 DK105554, R01 GM117163, R01 DK105154, K24 DK110550 and U54 DK118612.

MSU is supported by NIH/NIDDK K23 1K23DK114551.

Wellcome Trust, 090532, Mark I McCarthy; Wellcome, 106130, Mark I McCarthy; Wellcome, 098381, Mark I McCarthy; Wellcome, 203141, Mark I McCarthy; Wellcome, 212259, Mark I McCarthy; NIDDK, u01-DK105535, Mark I McCarthy; NIHR, NF-SI-0617-10090, Mark I McCarthy; NIDDK, U01 DK105554, Jose C. Florez; NIH, R01 GM117163, Jose C. Florez; NIDDK, R01 DK105154, Jose C. Florez; NIDDK, K24 DK110550, Jose C. Florez; NIDDK, U54 DK118612, Jose C. Florez; NIDDK, K23 1K23DK114551, Miriam S. Udler

CURRENT ADDRESS AND ADDRESS FOR CONTACT: **Mark McCarthy** Genentech, 1 DNA Way, South San Francisco, CA 94080, Tel: (1) 650 467 3970; Email: mccarthy.mark@gene.com, Requests for reprints should be made to the above address

DISCLOSURE SUMMARY

MMcC: The views expressed in this article are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health. He has served on advisory panels for Pfizer, NovoNordisk, Zoe Global; has received honoraria from Merck, Pfizer, NovoNordisk and Eli Lilly; has stock options in Zoe Global and has received research funding from Abbvie, Astra Zeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Merck, NovoNordisk, Pfizer, Roche, Sanofi Aventis, Servier & Takeda. As of June 2019, MMcC is an employee of Genentech, and holds stock in Roche. JCF has received a consulting honorarium from Janssen.

DATA AVAILABILITY

All data generated or analyzed during this study are included in this published article or in the data repositories listed in References.

REFERENCES

1. Cho NH, Shaw JE, Karuranga S, Huang Y, da Rocha Fernandes JD, Ohlrogge AW, Malanda B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract.* 2018;138:271-281. PMID: 29496507.

2. Haffner SM, Stern MP, Hazuda HP, Mitchell BD, Patterson JK. Cardiovascular risk factors in confirmed prediabetic individuals. Does the clock for coronary heart disease start ticking before the onset of clinical diabetes? JAMA. 1990;263(21):2893-8.

3. McCarthy MI. Painting a new picture of personalised medicine for diabetes. *Diabetologia.* 2017;60(5):793-799. PMID: 28175964.

4. Schwartz SS, Epstein S, Corkey BE, Grant SF, Gavin JR 3rd, Aguilar RB. The Time Is Right for a New Classification System for Diabetes: Rationale and Implications of the β-Cell-Centric Classification Schema. *Diabetes Care.* 2016;39(2):179-86. PMID: 26798148.

5. Florez JC. Precision medicine in diabetes: Is it time? *Diabetes Care.* 2016;39(7):1085-8. PMID: 27289125.

6. Barroso I, McCarthy MI. The Genetic Basis of Metabolic Disease. *Cell.* 2019;177(1):146-161. PMID: 30901536.

7. Kolb H, Martin S. Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. BMC Med. 2017;15:131. PMID: 28720102;

8. Sanna S, van Zuydam NR, Mahajan A, Kurilshikov A, Vich Vila A, Võsa U, Mujagic Z, Masclee AAM, Jonkers DMAE, Oosting M, Joosten LAB, Netea MG, Franke L, Zhernakova A, Fu J, Wijmenga C, McCarthy MI. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat Genet.* 2019;51(4):600-605. PMID: 30778224.

9. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, Payne AJ, Steinthorsdottir V, Scott RA, Grarup N, Cook JP, Schmidt EM, Wuttke M, Sarnowski C, Mägi R, Nano J, Gieger C, Trompet S, Lecoeur C, Preuss MH, Prins BP, Guo X, Bielak LF, Below JE, Bowden DW, Chambers JC, Kim YJ, Ng MCY, Petty LE, Sim X, Zhang W, Bennett AJ, Bork-Jensen J, Brummett CM, Canouil M, Ec Kardt KU, Fischer K, Kardia SLR, Kronenberg F, Läll K, Liu CT, Locke AE, Luan J, Ntalla I, Nylander V, Schönherr S, Schurmann C, Yengo L, Bottinger EP, Brandslund I, Christensen C, Dedoussis G, Florez JC, Ford I, Franco OH, Frayling TM, Giedraitis V, Hackinger S, Hattersley AT, Herder C, Ikram MA, Ingelsson M, Jørgensen ME, Jørgensen T, Kriebel J, Kuusisto J, Ligthart S, Lindgren CM, Linneberg A, Lyssenko V, Mamakou V, Meitinger T, Mohlke KL, Morris AD, Nadkarni G, Pankow JS, Peters A, Sattar N, Stančáková A, Strauch K, Taylor KD, Thorand B, Thorleifsson G, Thorsteinsdottir U, Tuomilehto J, Witte DR, Dupuis J, Peyser PA, Zeggini E, Loos RJF, Froguel P, Ingelsson E, Lind L, Groop L, Laakso M, Collins FS, Jukema JW, Palmer CNA, Grallert H, Metspalu A, Dehghan A, Köttgen A, Abecasis GR, Meigs JB, Rotter JI, Marchini J, Pedersen O, Hansen T, Langenberg C, Wareham NJ, Stefansson K, Gloyn AL, Morris AP, Boehnke M, McCarthy MI. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet.* 2018;50(11):1505-1513. PMID: 30297969.

10. Onengut-Gumuscu S, Chen WM, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, Farber E, Bonnie JK, Szpak M, Schofield E, Achuthan P, Guo H, Fortune MD, Stevens H, Walker NM, Ward LD, Kundaje A, Kellis M, Daly MJ, Barrett JC, Cooper JD, Deloukas P; Type 1 Diabetes Genetics Consortium, Todd JA, Wallace C, Concannon P, Rich SS. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat Genet.* 2015;47(4):381-6. PMID: 25751624.

11. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, Ma C, Fontanillas P, Moutsianas L, McCarthy DJ, Rivas MA, Perry JRB, Sim X, Blackwell TW, Robertson NR, Rayner NW, Cingolani P, Locke AE, Tajes JF, Highland HM, Dupuis J, Chines PS, Lindgren CM, Hartl C, Jackson AU, Chen H, Huyghe JR, van de Bunt M, Pearson RD, Kumar A, Müller-Nurasyid M, Grarup N, Stringham HM, Gamazon ER, Lee J, Chen Y, Scott RA, Below JE, Chen P, Huang J, Go MJ, Stitzel ML, Pasko D, Parker SCJ, Varga TV, Green T, Beer NL, Day-Williams AG, Ferreira T, Fingerlin T, Horikoshi M, Hu C, Huh I, Ikram MK, Kim BJ, Kim Y, Kim YJ, Kwon MS, Lee J, Lee S, Lin KH, Maxwell TJ, Nagai Y, Wang X, Welch RP, Yoon J, Zhang W, Barzilai N, Voight BF, Han BG, Jenkinson CP, Kuulasmaa T, Kuusisto J, Manning A, Ng MCY, Palmer ND, Balkau B, Stančáková A, Abboud HE, Boeing H, Giedraitis V, Prabhakaran D, Gottesman O, Scott J, Carey J, Kwan P, Grant G, Smith JD, Neale BM, Purcell S, Butterworth AS, Howson JMM, Lee HM, Lu Y, Kwak SH, Zhao W, Danesh J, Lam VKL, Park KS, Saleheen D, So WY, Tam CHT, Afzal U, Aguilar D, Arya R, Aung T, Chan E, Navarro C, Cheng CY, Palli D, Correa A, Curran JE, Rybin D, Farook VS, Fowler SP, Freedman BI, Griswold M, Hale DE, Hicks PJ, Khor CC, Kumar S, Lehne B, Thuillier D, Lim WY, Liu J, van der Schouw YT, Loh M, Musani SK, Puppala S, Scott WR, Yengo L, Tan ST, Taylor HA Jr, Thameem F, Wilson G Sr, Wong TY, Njølstad PR, Levy JC, Mangino M, Bonnycastle LL, Schwarzmayr T, Fadista J, Surdulescu GL, Herder C, Groves CJ, Wieland T, Bork-Jensen J, Brandslund I, Christensen C, Koistinen HA, Doney ASF, Kinnunen L, Esko T, Farmer AJ, Hakaste L, Hodgkiss D, Kravic J, Lyssenko V, Hollensted M, Jørgensen ME, Jørgensen T, Ladenvall C, Justesen JM, Käräjämäki A, Kriebel J, Rathmann W, Lannfelt L, Lauritzen T, Narisu N, Linneberg A, Melander O, Milani L, Neville M,

Orho-Melander M, Qi L, Qi Q, Roden M, Rolandsson O, Swift A, Rosengren AH, Stirrups K, Wood AR, Mihailov E, Blancher C, Carneiro MO, Maguire J, Poplin R, Shakir K, Fennell T, DePristo M, de Angelis MH, Deloukas P, Gjesing AP, Jun G, Nilsson P, Murphy J, Onofrio R, Thorand B, Hansen T, Meisinger C, Hu FB, Isomaa B, Karpe F, Liang L, Peters A, Huth C, O'Rahilly SP, Palmer CNA, Pedersen O, Rauramaa R, Tuomilehto J, Salomaa V, Watanabe RM, Syvänen AC, Bergman RN, Bharadwaj D, Bottinger EP, Cho YS, Chandak GR, Chan JCN, Chia KS, Daly MJ, Ebrahim SB, Langenberg C, Elliott P, Jablonski KA, Lehman DM, Jia W, Ma RCW, Pollin TI, Sandhu M, Tandon N, Froguel P, Barroso I, Teo YY, Zeggini E, Loos RJF, Small KS, Ried JS, DeFronzo RA, Grallert H, Glaser B, Metspalu A, Wareham NJ, Walker M, Banks E, Gieger C, Ingelsson E, Im HK, Illig T, Franks PW, Buck G, Trakalo J, Buck D, Prokopenko I, Mägi R, Lind L, Farjoun Y, Owen KR, Gloyn AL, Strauch K, Tuomi T, Kooner JS, Lee JY, Park T, Donnelly P, Morris AD, Hattersley AT, Bowden DW, Collins FS, Atzmon G, Chambers JC, Spector TD, Laakso M, Strom TM, Bell GI, Blangero J, Duggirala R, Tai ES, McVean G, Hanis CL, Wilson JG, Seielstad M, Frayling TM, Meigs JB, Cox NJ, Sladek R, Lander ES, Gabriel S, Burtt NP, Mohlke KL, Meitinger T, Groop L, Abecasis G, Florez JC, Scott LJ, Morris AP, Kang HM, Boehnke M, Altshuler D, McCarthy MI. The genetic architecture of type 2 diabetes. *Nature*. 2016;536(7614):41-47. PMID: 27398621.

12. Flannick J, Mercader JM, Fuchsberger C, Udler MS, Mahajan A, Wessel J, Teslovich TM, Caulkins L, Koesterer R, Barajas-Olmos F, Blackwell TW, Boerwinkle E, Brody JA, Centeno-Cruz F, Chen L, Chen S, Contreras-Cubas C, Córdova E, Correa A, Cortes M, DeFronzo RA, Dolan L, Drews KL, Elliott A, Floyd JS, Gabriel SB, Garay-Sevilla Eugenia M, García-Ortiz H, Gross M, Han S, Heard-Costa NL, Jackson AU, Jørgensen ME, Kang Min H, Kelsey M, Kim B, Koistinen HA, Kuusisto J, Leader JB, Linneberg A, Liu C, Liu J, Lyssenko V, Manning AK, Marcketta A, Malacara-Hernandez Manuel J, Martínez-Hernández A, Matsuo K, Mayer-Davis E, Mendoza-Caamal E, Mohlke KL, Morrison AC, Ndungu A, Ng C MY, O'Dushlaine C, Payne AJ, Pihoker C, Post WS, Preuss M, Psaty BM, Vasan RS, Rayner William N, Reiner AP, Revilla-Monsalve C, Robertson NR, Santoro N, Schurmann C, So Yee W, Soberón X, Stringham HM, Strom TM, Tam CTH, Thameem F, Tomlinson B, Torres JM, Tracy RP, Van Dam RM, Vujkovic M, Wang S, Welch RP, Witte DR, Wong T, Atzmon G, Barzilai N, Blangero J, Bonnycastle LL, Bowden DW, Chambers JC, Chan E, Cheng C, Cho Shin Y, Collins FS, De Vries PS, Duggirala R, Glaser B, Gonzalez C, Elena Gonzalez M, Groop L, Kooner JS, Kwak Heon S, Laakso M, Lehman DM, Nilsson P, Spector TD, Tai Shyong E, Tuomi T, Tuomilehto J, Wilson JG, Aguilar-Salinas CA, Bottinger EP, Burke B, Carey DJ, Chan J, Dupuis J, Frossard P, Heckbert SR, Hwang Yeong M, Kim Jin Y, Kirchner Lester H, Lee J-Y, Lee J, Loos RJF, Ma RCW, Morris AD, O'Donnell CJ, Palmer CNA, Pankow JS, Park Soo K, Rasheed A, Saleheen D, Sim X, Small KS, Teo Ying Y, Haiman CA, Hanis CL, Henderson BE, Orozco L, Tusié-Luna T, Dewey FE, Baras A, Gieger C, Meitinger T, Strauch K, Lange LA, Grarup N, Hansen T, Pedersen OB, Zeitler P, Dabelea D, Abecasis GR, Bell GI, Cox NJ, Seielstad M, Sladek R, Meigs JB, Rich SS, Rotter JI, DiscovEHR Collaboration, CHARGE, LuCamp, ProDiGY, GoT2D, ESP, SIGMA-T2D, T2D-GENES, AMP-T2D-GENES, Altshuler DM, Burtt NP, Scott LJ, Morris AP, Florez JC, McCarthy MI, Boehnke M. Genetic discovery and translational decision support from exome sequencing of 20,791 type 2 diabetes cases and 24,440 controls from five ancestries. *Nature*. 2019. Jun;570(7759):71-76. PMID: 31118516.

13. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. Genome-wide polygenic scores for

common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet.* 2018;50(9):1219-1224. PMID: 30104762.

14. Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance—a population-based twin study. Diabetologia. 1999;42(2):139-45. PubMed PMID: 10064092.

15. Willemsen G, Ward KJ, Bell CG, Christensen K, Bowden J, Dalgård C, Harris JR, Kaprio J, Lyle R, Magnusson PK, Mather KA, Ordoňana JR, Perez-Riquelme F, Pedersen NL, Pietiläinen KH, Sachdev PS, Boomsma DI, Spector T. The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium. Twin Res Hum Genet. 2015;18(6):762-71. PMID: 26678054.

16. Kyvik KO, Green A, Beck-Nielsen H. Concordance rates of insulin dependent diabetes mellitus: a population based study of young Danish twins. BMJ. 1995;311(7010):913-7. PMID: 7580548;

17. Aylward A, Chiou J, Okino ML, Kadakia N, Gaulton KJ. Shared genetic risk contributes to type 1 and type 2 diabetes etiology. *Hum Mol Genet.* 2018 Nov 7. doi: 10.1093/hmg/ddy314. [Epub ahead of print] PMID: 30407494.

18. Todd JA. Intolerable secretion and diabetes in tolerant transgenic mice, revisited. *Nat Genet.* 2016;48(5):476-7. PMID: 27120442.

19. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segrè AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Bengtsson Boström K, Bravenboer B, Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PR, Jørgensen T, Kao WH, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieverse A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JR, Petersen AK, Platou C, Proença C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shrader P, Sigurdsson G, Sparsø T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van Herpt T, van Vliet-Ostaptchouk JV, Walters GB, Weedon MN, Wijmenga C, Witteman J, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllensten U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Mohlke KL, Morris AD, Palmer CN, Pramstaller PP, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Wareham NJ, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Hu FB, Meigs JB, Pankow JS, Pedersen O, Wichmann HE, Barroso I, Florez JC, Frayling TM, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI; MAGIC investigators; GIANT Consortium. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet.* 2010;42(7):579-89. PMID: 2058182.

20. Mahajan A, Wessel J, Willems SM, Zhao W, Robertson NR, Chu AY, Gan W, Kitajima H, Taliun D, Rayner NW, Guo X, Lu Y, Li M, Jensen RA, Hu Y, Huo S, Lohman KK, Zhang W, Cook JP, Prins BP, Flannick J, Grarup N, Trubetskoy VV, Kravic J, Kim YJ, Rybin DV, Yaghootkar H, Müller-Nurasyid M, Meidtner K, Li-Gao R, Varga TV, Marten J, Li J, Smith AV, An P, Ligthart S, Gustafsson S, Malerba G,

Demirkan A, Tajes JF, Steinthorsdottir V, Wuttke M, Lecoeur C, Preuss M, Bielak LF, Graff M, Highland HM, Justice AE, Liu DJ, Marouli E, Peloso GM, Warren HR; ExomeBP Consortium; MAGIC Consortium; GIANT Consortium, Afaq S, Afzal S, Ahlqvist E, Almgren P, Amin N, Bang LB, Bertoni AG, Bombieri C, Bork-Jensen J, Brandslund I, Brody JA, Burtt NP, Canouil M, Chen YI, Cho YS, Christensen C, Eastwood SV, Eckardt KU, Fischer K, Gambaro G, Giedraitis V, Grove ML, de Haan HG, Hackinger S, Hai Y, Han S, Tybjærg-Hansen A, Hivert MF, Isomaa B, Jäger S, Jørgensen ME, Jørgensen T, Käräjämäki A, Kim BJ, Kim SS, Koistinen HA, Kovacs P, Kriebel J, Kronenberg F, Läll K, Lange LA, Lee JJ, Lehne B, Li H, Lin KH, Linneberg A, Liu CT, Liu J, Loh M, Mägi R, Mamakou V, McKean-Cowdin R, Nadkarni G, Neville M, Nielsen SF, Ntalla I, Peyser PA, Rathmann W, Rice K, Rich SS, Rode L, Rolandsson O, Schönherr S, Selvin E, Small KS, Stančáková A, Surendran P, Taylor KD, Teslovich TM, Thorand B, Thorleifsson G, Tin A, Tönjes A, Varbo A, Witte DR, Wood AR, Yajnik P, Yao J, Yengo L, Young R, Amouyel P, Boeing H, Boerwinkle E, Bottinger EP, Chowdhury R, Collins FS, Dedoussis G, Dehghan A, Deloukas P, Ferrario MM, Ferrières J, Florez JC, Frossard P, Gudnason V, Harris TB, Heckbert SR, Howson JMM, Ingelsson M, Kathiresan S, Kee F, Kuusisto J, Langenberg C, Launer LJ, Lindgren CM, Männistö S, Meitinger T, Melander O, Mohlke KL, Moitry M, Morris AD, Murray AD, de Mutsert R, Orho-Melander M, Owen KR, Perola M, Peters A, Province MA, Rasheed A, Ridker PM, Rivadineira F, Rosendaal FR, Rosengren AH, Salomaa V, Sheu WH, Sladek R, Smith BH, Strauch K, Uitterlinden AG, Varma R, Willer CJ, Blüher M, Butterworth AS, Chambers JC, Chasman DI, Danesh J, van Duijn C, Dupuis J, Franco OH, Franks PW, Froguel P, Grallert H, Groop L, Han BG, Hansen T, Hattersley AT, Hayward C, Ingelsson E, Kardia SLR, Karpe F, Kooner JS, Köttgen A, Kuulasmaa K, Laakso M, Lin X, Lind L, Liu Y, Loos RJF, Marchini J, Metspalu A, Mook-Kanamori D, Nordestgaard BG, Palmer CNA, Pankow JS, Pedersen O, Psaty BM, Rauramaa R, Sattar N, Schulze MB, Soranzo N, Spector TD, Stefansson K, Stumvoll M, Thorsteinsdottir U, Tuomi T, Tuomilehto J, Wareham NJ, Wilson JG, Zeggini E, Scott RA, Barroso I, Frayling TM, Goodarzi MO, Meigs JB, Boehnke M, Saleheen D, Morris AP, Rotter JI, McCarthy MI. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat Genet.* 2018;50(4):559-571. PMID: 29632382.

21. Small KS, Todorčević M, Civelek M, El-Sayed Moustafa JS, Wang X, Simon MM, Fernandez-Tajes J, Mahajan A, Horikoshi M, Hugill A, Glastonbury CA, Quaye L, Neville MJ, Sethi S, Yon M, Pan C, Che N, Viñuela A, Tsai PC, Nag A, Buil A, Thorleifsson G, Raghavan A, Ding Q, Morris AP, Bell JT, Thorsteinsdottir U, Stefansson K, Laakso M, Dahlman I, Arner P, Gloyn AL, Musunuru K, Lusis AJ, Cox RD, Karpe F, McCarthy MI. Regulatory variants at *KLF14* influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. *Nat Genet.* 2018;50(4):572-580. PMID: 29632379.

22. Thomsen SK, Raimondo A, Hastoy B, Sengupta S, Dai XQ, Bautista A, Censin J, Payne AJ, Umapathysivam MM, Spigelman AF, Barrett A, Groves CJ, Beer NL, Manning Fox JE, McCarthy MI, Clark A, Mahajan A, Rorsman P, MacDonald PE, Gloyn AL. Type 2 diabetes risk alleles in *PAM* impact insulin release from human pancreatic β-cells. *Nat Genet.* 2018;50(8):1122-1131. PMID: 30054598.

23. Rusu V, Hoch E, Mercader JM, Tenen DE, Gymrek M, Hartigan CR, DeRan M, von Grotthuss M, Fontanillas P, Spooner A, Guzman G, Deik AA, Pierce KA, Dennis C, Clish CB, Carr SA, Wagner BK, Schenone M, Ng MCY, Chen BH; MEDIA Consortium; SIGMA T2D Consortium, Centeno-Cruz F, Zerrweck C, Orozco L, Altshuler DM, Schreiber SL, Florez JC, Jacobs SBR, Lander ES. Type 2 diabetes

variants disrupt function of SLC16A11 through two distinct mechanisms. *Cell*. 2017;170(1):199-212.e20. PMID: 28666119.

24. Florez JC, Hirschhorn J, Altshuler D. The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits. *Annu Rev Genomics Hum Genet.* 2003;4:257-91. PMID: 14527304.

25. Maller J, George S, Purcell S, Fagerness J, Altshuler D, Daly MJ, Seddon JM. Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat Genet.* 2006;38(9):1055-9. PMID: 16936732.

26. Weedon MN, McCarthy MI, Hitman G, Walker M, Groves CJ, Zeggini E, Rayner NW, Shields B, Owen KR, Hattersley AT, Frayling TM. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Med.* 2006;3(10):e374. PMID: 17020404.

27. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet.* 2003;33(2):177-82. PMID: 12524541.

28. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol.* 2008;32(4):381-5. PMID: 18348202.

29. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19(9):581-590. PMID: 29789686.

30. Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, Berglund G, Altshuler D, Nilsson P, Groop L. Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N Engl J Med.* 2008;359(21):2220-32. PMID: 19020324.

31. Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB Sr, Cupples LA. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med.* 2008;359(21):2208-19. PMID: 19020323.

32. Lango H; UK Type 2 Diabetes Genetics Consortium, Palmer CN, Morris AD, Zeggini E, Hattersley AT, McCarthy MI, Frayling TM, Weedon MN. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes*. 2008;57(11):3129-35. PMID: 18591388.

33. Chatterjee N, Shi j, Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* 2016;17:392–406

34. Yaghootkar H, Scott RA, White CC, Zhang W, Speliotes E, Munroe PB, Ehret GB, Bis JC, Fox CS, Walker M, Borecki IB, Knowles JW, Yerges-Armstrong L, Ohlsson C, Perry JR, Chambers JC, Kooner JS, Franceschini N, Langenberg C, Hivert MF, Dastani Z, Richards JB, Semple RK, Frayling TM. Genetic evidence for a normal-weight "metabolically obese" phenotype linking insulin resistance, hypertension, coronary artery disease, and type 2 diabetes. *Diabetes*. 2014;63(12):4369-77. PMID: 25048195;

35. Scott RA, Fall T, Pasko D, Barker A, Sharp SJ, Arriola L, Balkau B, Barricarte A, Barroso I, Boeing H, Clavel-Chapelon F, Crowe FL, Dekker JM, Fagherazzi G, Ferrannini E, Forouhi NG, Franks PW, Gavrila D, Giedraitis V, Grioni S, Groop LC, Kaaks R, Key TJ, Kühn T, Lotta LA, Nilsson PM, Overvad K, Palli D, Panico S, Quirós JR, Rolandsson O, Roswall N, Sacerdote C, Sala N, Sánchez MJ, Schulze MB, Siddiq A, Slimani N, Sluijs I, Spijkerman AM, Tjonneland A, Tumino R, van der A DL, Yaghootkar H; RISC study group; EPIC-InterAct consortium, McCarthy MI, Semple RK, Riboli E, Walker M, Ingelsson E, Frayling TM, Savage DB, Langenberg C, Wareham NJ. Common genetic variants highlight the role of insulin resistance and

body fat distribution in type 2 diabetes, independent of obesity. *Diabetes*. 2014;63(12):4378-4387. PubMed PMID: 24947364.

36. Lotta LA, Gulati P, Day FR, Payne F, Ongen H, van de Bunt M, Gaulton KJ, Eicher JD, Sharp SJ, Luan J, De Lucia Rolfe E, Stewart ID, Wheeler E, Willems SM, Adams C, Yaghootkar H; EPIC-InterAct Consortium; Cambridge FPLD1 Consortium, Forouhi NG, Khaw KT, Johnson AD, Semple RK, Frayling T, Perry JR, Dermitzakis E, McCarthy MI, Barroso I, Wareham NJ, Savage DB, Langenberg C, O'Rahilly S, Scott RA. Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nat Genet.* 2017;49(1):17-26. PMID: 27841877.

37. Semple RK, Savage DB, Cochran EK, Gorden P, O'Rahilly S. Genetic syndromes of severe insulin resistance. *Endocr Rev.* 2011;32(4):498-514. PMID: 21536711.

38. Udler MS, 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, Glaser B, Mercader JM, Gaulton K, Flannick J, Getz G, Florez JC. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS Med.* 2018;15(9):e1002654. PMID: 30240442.

39. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, Nathan DM; Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med.* 2002;346(6):393-403. PMID: 11832527

40. Lindström J, Louheranta A, Mannelin M, Rastas M, Salminen V, Eriksson J, Uusitupa M, Tuomilehto J; Finnish Diabetes Prevention Study Group. The Finnish Diabetes Prevention Study (DPS): Lifestyle intervention and 3-year results on diet and physical activity. *Diabetes Care*. 2003;26(12):3230-6. PMID: 14633807.

41. Vassy JL, Hivert MF, Porneala B, Dauriz M, Florez JC, Dupuis J, Siscovick DS, Fornage M, Rasmussen-Torvik LJ, Bouchard C, Meigs JB. Polygenic type 2 diabetes prediction at the limit of common variant detection. *Diabetes*. 2014;63(6):2172-82. PubMed PMID: 24520119.

42. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet.* 2013;45(4):400-5. PMID: 23455638.

43. Multhaup ML, Kita R, Krock B, Eriksson N, Fontanillas P, Aslikbekyan S, Del Gobbo L, Shelton JF, Tennen RI, Lehman A, Furlotte NA, Koelsch BL. White Paper 23-19 The science behind 23andMe's Type 2 Diabetes report Estimating the likelihood of developing type 2 diabetes with polygenic models. 23andMe. Published March 2019. Available from https://research.23andme.com/wp-content/uploads/2019/03/23_19-Type2Diabetes_March2019.pdf

44. Scott RA, Scott LJ, Mägi R, Marullo L, Gaulton KJ, Kaakinen M, Pervjakova N, Pers TH, Johnson AD, Eicher JD, Jackson AU, Ferreira T, Lee Y, Ma C, Steinthorsdottir V, Thorleifsson G, Qi L, Van Zuydam NR, Mahajan A, Chen H, Almgren P, Voight BF, Grallert H, Müller-Nurasyid M, Ried JS, Rayner NW, Robertson N, Karssen LC, van Leeuwen EM, Willems SM, Fuchsberger C, Kwan P, Teslovich TM, Chanda P, Li M, Lu Y, Dina C, Thuillier D, Yengo L, Jiang L, Sparso T, Kestler HA, Chheda H, Eisele L, Gustafsson S, Frånberg M, Strawbridge RJ, Benediktsson R, Hreidarsson AB, Kong A, Sigurðsson G, Kerrison ND, Luan J, Liang L, Meitinger T, Roden M, Thorand B, Esko T, Mihailov E, Fox C, Liu CT, Rybin D, Isomaa B, Lyssenko V, Tuomi T, Couper DJ, Pankow JS, Grarup N, Have CT,

Jørgensen ME, Jørgensen T, Linneberg A, Cornelis MC, van Dam RM, Hunter DJ, Kraft P, Sun Q, Edkins S, Owen KR, Perry JRB, Wood AR, Zeggini E, Tajes-Fernandes J, Abecasis GR, Bonnycastle LL, Chines PS, Stringham HM, Koistinen HA, Kinnunen L, Sennblad B, Mühleisen TW, Nöthen MM, Pechlivanis S, Baldassarre D, Gertow K, Humphries SE, Tremoli E, Klopp N, Meyer J, Steinbach G, Wennauer R, Eriksson JG, Männistö S, Peltonen L, Tikkanen E, Charpentier G, Eury E, Lobbens S, Gigante B, Leander K, McLeod O, Bottinger EP, Gottesman O, Ruderfer D, Blüher M, Kovacs P, Tonjes A, Maruthur NM, Scapoli C, Erbel R, Jöckel KH, Moebus S, de Faire U, Hamsten A, Stumvoll M, Deloukas P, Donnelly PJ, Frayling TM, Hattersley AT, Ripatti S, Salomaa V, Pedersen NL, Boehm BO, Bergman RN, Collins FS, Mohlke KL, Tuomilehto J, Hansen T, Pedersen O, Barroso I, Lannfelt L, Ingelsson E, Lind L, Lindgren CM, Cauchi S, Froguel P, Loos RJF, Balkau B, Boeing H, Franks PW, Barricarte Gurrea A, Palli D, van der Schouw YT, Altshuler D, Groop LC, Langenberg C, Wareham NJ, Sijbrands E, van Duijn CM, Florez JC, Meigs JB, Boerwinkle E, Gieger C, Strauch K, Metspalu A, Morris AD, Palmer CNA, Hu FB, Thorsteinsdottir U, Stefansson K, Dupuis J, Morris AP, Boehnke M, McCarthy MI, Prokopenko I; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes.* 2017;66(11):2888-2902. PMID: 28566273;

45. https://www.genomicsplc.com/wp-content/uploads/2019/03/Genomics-plc-PRS-details.pdf

46. Redondo MJ, Jeffrey J, Fain PR, Eisenbarth GS, Orban T. Concordance for islet autoimmunity among monozygotic twins. *N Engl J Med.* 2008;359(26):2849-50. PMID: 19109586

47. Hyttinen V, Kaprio J, Kinnun L, Koskenvuo M, Tuomilehto J. Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. Diabetes. 2003; 52(4):1052-5. PMID: 12663480.

48. Kuo CF, Chou IJ, Grainge MJ, Luo SF, See LC, Yu KH, Zhang W, Doherty M, Valdes M. Familial aggregation and heritability of type 1 diabetes mellitus and coaggregation of chronic diseases in affected families. Clin Epidemiol. 2018; 10: 1447-1455. PMID: 30349392.

49. Noble JA, Valdes AM, Cook M, Klitz W, Thomson G, Ehrlich HA.The Role of HLA Class 11 Genes in Insulin-Dependent Diabetes Mellitus: Molecular Analysis of 180 Caucasian, Multiplex Families. *Am J Hum Genet.* 1996;59:1134-1148.

50. Redondo MJ, Steck AK, Pugliese A. Genetics of type 1 diabetes. *Pediatr Diabetes.* 2018;19(3):346-353. PMID: 29094512.

51. Vafiadis P, Bennett ST, Todd JA, Nadeau J, Grabs R, Goodyer CG, Wickramasinghe S, Colle E, Polychronakos C. Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. *Nat Genet.* 1997;15(3):289-92. PMID: 9054944.

52. Pugliese A, Zeller M, Fernandez A Jr, Zalcberg LJ, Bartlett RJ, Ricordi C, Pietropaolo M, Eisenbarth GS, Bennett ST, Patel DD. The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. *Nat Genet.* 1997;15(3):293-7. PMID: 9054945.

53. Onengut-Gumuscu S, Ewens KG, Spielman RS, Concannon P. A functional polymorphism (1858C/T) in the PTPN22 gene is linked and associated with type I diabetes in multiplex families. *Genes Immun.* 2004;5(8):678-80. PMID: 15526003.

54. Aly TA, Ide A, Jahromi MM, Barker JM, Fernando MS, Babu SR, Yu L, Miao D, Erlich HA, Fain PR, Barriga KJ, Norris JM, Rewers MJ, Eisenbarth GS. Extreme

genetic risk for type 1A diabetes. *Proc Natl Acad Sci U S A.* 2006;103(38):14074-9. PMID: 16966600;

55. Steck AK, Dong F, Wong R, Fouts A, Liu E, Romanos J, Wijmenga C, Norris JM, Rewers MJ. Improving prediction of type 1 diabetes by testing non-HLA genetic variants in addition to HLA markers. *Pediatr Diabetes*. 2014;15(5):355-62. PubMed PMID: 25075402.

56. Winkler C, Krumsiek J, Buettner F, Angermüller C, Giannopoulou EZ, Theis FJ, Ziegler AG, Bonifacio E. Feature ranking of type 1 diabetes susceptibility genes improves prediction of type 1 diabetes. *Diabetologia*. 2014;57(12):2521-9. PMID: 25186292.

57. Oram RA, Patel K, Hill A, Shields B, McDonald TJ, Jones A, Hattersley AT, Weedon MN. A Type 1 Diabetes Genetic Risk Score Can Aid Discrimination Between Type 1 and Type 2 Diabetes in Young Adults. *Diabetes Care.* 2016;39(3):337-44. PMID: 26577414.

58. Redondo MJ, Oram RA, Steck AK. Genetic Risk Scores for Type 1 Diabetes Prediction and Diagnosis. *Curr Diab Rep*. 2017;17(12):129. PMID: 29080981.

59. Bonifacio E, Beyerlein A, Hippich M, Winkler C, Vehik K, Weedon MN, Laimighofer M, Hattersley AT, Krumsiek J, Frohnert BI, Steck AK, Hagopian WA, Krischer JP, Lernmark Å, Rewers MJ, She JX, Toppari J, Akolkar B, Oram RA, Rich SS, Ziegler AG; TEDDY Study Group. Genetic scores to stratify risk of developing multiple islet autoantibodies and type 1 diabetes: A prospective study in children. *PLoS Med.* 2018;15(4):e1002548. PubMed PMID: 29614081.

60. Sharp SA, Rich SS, Wood AR, Jones SE, Beaumont RN, Harrison JW, Schneider DA, Locke JM, Tyrrell J, Weedon MN, Hagopian WA, Oram RA. Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis. *Diabetes Care*. 2019;42(2):200-207. PMID: 30655379.

61. Bingley PJ, Bonifacio E, Williams AJ, Genovese S, Bottazzo GF, Gale EA. Prediction of IDDM in the general population: strategies based on combinations of autoantibody markers. *Diabetes*. 1997;46(11):1701-10. PMID: 9356015.

62. Greenbaum CJ, Anderson AM, Dolan LM, Mayer-Davis EJ, Dabelea D, Imperatore G, Marcovina S, Pihoker C; SEARCH Study Group. Preservation of beta-cell function in autoantibody-positive youth with diabetes. *Diabetes Care.* 2009;32(10):1839-44. PMID: 19587365.

63. Shields BM, Hicks S, Shepherd MH, Colclough K, Hattersley AT, Ellard S. Maturity-onset diabetes of the young (MODY): how many cases are we missing? *Diabetologia*. 2010;53(12):2504-8. PMID: 20499044.

64. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661-78. PMID: 17554300.

65. Grubb AL, McDonald TJ, Rutters F, Donnelly LA, Hattersley AT, Oram RA, Palmer CNA, van der Heijden AA, Carr F, Elders PJM, Weedon MN, Slieker RC, 't Hart LM, Pearson ER, Shields BM, Jones AG. A Type 1 Diabetes Genetic Risk Score Can Identify Patients With GAD65 Autoantibody-Positive Type 2 Diabetes Who Rapidly Progress to Insulin Therapy. *Diabetes Care.* 2019;42(2):208-214. PMID: 30352895.

66. Thomas NJ, Jones SE, Weedon MN, Shields BM, Oram RA, Hattersley AT. Frequency and phenotype of type 1 diabetes in the first six decades of life: a cross-sectional, genetically stratified survival analysis from UK Biobank. *Lancet Diabetes Endocrinol*. 2018;6(2):122-129. PubMed PMID: 29199115;

67. Patel KA, Oram RA, Flanagan SE, De Franco E, Colclough K, Shepherd M, Ellard S, Weedon MN, Hattersley AT. Type 1 Diabetes Genetic Risk Score: A Novel Tool to Discriminate Monogenic and Type 1 Diabetes. *Diabetes*. 2016;65(7):2094-2099. PMID: 27207547.

68. Johnson MB, Patel KA, De Franco E, Houghton JAL, McDonald TJ, Ellard S, Flanagan SE, Hattersley AT. A type 1 diabetes genetic risk score can discriminate monogenic autoimmunity with diabetes from early-onset clustering of polygenic autoimmunity with diabetes. *Diabetologia*. 2018;61(4):862-869. PMID: 29417186.

69. Frayling TM, Evans JC, Bulman MP, Pearson E, Allen L, Owen K, Bingham C, Hannemann M, Shepherd M, Ellard S, Hattersley AT. beta-cell genes and diabetes: molecular and clinical characterization of mutations in transcription factors. *Diabetes*. 2001;50(Suppl 1):S94-100. PMID: 11272211.

70. Flannick J, Beer NL, Bick AG, Agarwala V, Molnes J, Gupta N, Burtt NP, Florez JC, Meigs JB, Taylor H, Lyssenko V, Irgens H, Fox E, Burslem F, Johansson S, Brosnan MJ, Trimmer JK, Newton-Cheh C, Tuomi T, Molven A, Wilson JG, O'Donnell CJ, Kathiresan S, Hirschhorn JN, Njølstad PR, Rolph T, Seidman JG, Gabriel S, Cox DR, Seidman CE, Groop L, Altshuler D. Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat Genet.* 2013;45(11):1380-5. PMID: 24097065.

71. Lango Allen H, Johansson S, Ellard S, Shields B, Hertel JK, Raeder H, Colclough K, Molven A, Frayling TM, Njølstad PR, Hattersley AT, Weedon MN. Polygenic risk variants for type 2 diabetes susceptibility modify age at diagnosis in monogenic HNF1A diabetes. *Diabetes*. 2010;59(1):266-71. PMID: 19794065.

72. Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, Dadaev T, Leongamornlert D, Anokian E, Cieza-Borrella C, Goh C, Brook MN, Sheng X, Fachal L, Dennis J, Tyrer J, Muir K, Lophatananon A, Stevens VL, Gapstur SM, Carter BD, Tangen CM, Goodman PJ, Thompson IM Jr, Batra J, Chambers S, Moya L, Clements J, Horvath L, Tilley W, Risbridger GP, Gronberg H, Aly M, Nordström T, Pharoah P, Pashayan N, Schleutker J, Tammela TLJ, Sipeky C, Auvinen A, Albanes D, Weinstein S, Wolk A, Håkansson N, West CML, Dunning AM, Burnet N, Mucci LA, Giovannucci E, Andriole GL, Cussenot O, Cancel-Tassin G, Koutros S, Beane Freeman LE, Sorensen KD, Orntoft TF, Borre M, Maehle L, Grindedal EM, Neal DE, Donovan JL, Hamdy FC, Martin RM, Travis RC, Key TJ, Hamilton RJ, Fleshner NE, Finelli A, Ingles SA, Stern MC, Rosenstein BS, Kerns SL, Ostrer H, Lu YJ, Zhang HW, Feng N, Mao X, Guo X, Wang G, Sun Z, Giles GG, Southey MC, MacInnis RJ, FitzGerald LM, Kibel AS, Drake BF, Vega A, Gómez-Caamaño A, Szulkin R, Eklund M, Kogevinas M, Llorca J, Castaño-Vinyals G, Penney KL, Stampfer M, Park JY, Sellers TA, Lin HY, Stanford JL, Cybulski C, Wokolorczyk D, Lubinski J, Ostrander EA, Geybels MS, Nordestgaard BG, Nielsen SF, Weischer M, Bisbjerg R, Røder MA, Iversen P, Brenner H, Cuk K, Holleczek B, Maier C, Luedeke M, Schnoeller T, Kim J, Logothetis CJ, John EM, Teixeira MR, Paulo P, Cardoso M, Neuhausen SL, Steele L, Ding YC, De Ruyck K, De Meerleer G, Ost P, Razack A, Lim J, Teo SH, Lin DW, Newcomb LF, Lessel D, Gamulin M, Kulis T, Kaneva R, Usmani N, Singhal S, Slavov C, Mitev V, Parliament M, Claessens F, Joniau S, Van den Broeck T, Larkin S, Townsend PA, Aukim-Hastie C, Gago-Dominguez M, Castelao JE, Martinez ME, Roobol MJ, Jenster G, van Schaik RHN, Menegaux F, Truong T, Koudou YA, Xu J, Khaw KT, Cannon-Albright L, Pandha H, Michael A, Thibodeau SN, McDonnell SK, Schaid DJ, Lindstrom S, Turman C, Ma J, Hunter DJ, Riboli E, Siddiq A, Canzian F, Kolonel LN, Le Marchand L, Hoover RN, Machiela MJ, Cui Z, Kraft P, Amos CI, Conti DV, Easton DF, Wiklund F, Chanock SJ, Henderson BE, Kote-Jarai Z, Haiman CA, Eeles RA; Profile Study;

Australian Prostate Cancer BioResource (APCB); IMPACT Study; Canary PASS Investigators; Breast and Prostate Cancer Cohort Consortium (BPC3); PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium; Cancer of the Prostate in Sweden (CAPS); Prostate Cancer Genome-wide Association Study of Uncommon Susceptibility Loci (PEGASUS); Genetic Associations and Mechanisms in Oncology (GAME-ON)/Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE) Consortium. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. ***Nat Genet.*** **2018;50(7):928-936. PMID: 29892016.**

73. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, Tyrer JP, Chen TH, Wang Q, Bolla MK, Yang X, Adank MA, Ahearn T, Aittomäki K, Allen J, Andrulis IL, Anton-Culver H, Antonenkova NN, Arndt V, Aronson KJ, Auer PL, Auvinen P, Barrdahl M, Beane Freeman LE, Beckmann MW, Behrens S, Benitez J, Bermisheva M, Bernstein L, Blomqvist C, Bogdanova NV, Bojesen SE, Bonanni B, Børresen-Dale AL, Brauch H, Bremer M, Brenner H, Brentnall A, Brock IW, Brooks-Wilson A, Brucker SY, Brüning T, Burwinkel B, Campa D, Carter BD, Castelao JE, Chanock SJ, Chlebowski R, Christiansen H, Clarke CL, Collée JM, Cordina-Duverger E, Cornelissen S, Couch FJ, Cox A, Cross SS, Czene K, Daly MB, Devilee P, Dörk T, Dos-Santos-Silva I, Dumont M, Durcan L, Dwek M, Eccles DM, Ekici AB, Eliassen AH, Ellberg C, Engel C, Eriksson M, Evans DG, Fasching PA, Figueroa J, Fletcher O, Flyger H, Försti A, Fritschi L, Gabrielson M, Gago-Dominguez M, Gapstur SM, García-Sáenz JA, Gaudet MM, Georgoulias V, Giles GG, Gilyazova IR, Glendon G, Goldberg MS, Goldgar DE, González-Neira A, Grenaker Alnæs GI, Grip M, Gronwald J, Grundy A, Guénel P, Haeberle L, Hahnen E, Haiman CA, Håkansson N, Hamann U, Hankinson SE, Harkness EF, Hart SN, He W, Hein A, Heyworth J, Hillemanns P, Hollestelle A, Hooning MJ, Hoover RN, Hopper JL, Howell A, Huang G, Humphreys K, Hunter DJ, Jakimovska M, Jakubowska A, Janni W, John EM, Johnson N, Jones ME, Jukkola-Vuorinen A, Jung A, Kaaks R, Kaczmarek K, Kataja V, Keeman R, Kerin MJ, Khusnutdinova E, Kiiski JI, Knight JA, Ko YD, Kosma VM, Koutros S, Kristensen VN, Krüger U, Kühl T, Lambrechts D, Le Marchand L, Lee E, Lejbkowicz F, Lilyquist J, Lindblom A, Lindström S, Lissowska J, Lo WY, Loibl S, Long J, Lubiński J, Lux MP, MacInnis RJ, Maishman T, Makalic E, Maleva Kostovska I, Mannermaa A, Manoukian S, Margolin S, Martens JWM, Martinez ME, Mavroudis D, McLean C, Meindl A, Menon U, Middha P, Miller N, Moreno F, Mulligan AM, Mulot C, Muñoz-Garzon VM, Neuhausen SL, Nevanlinna H, Neven P, Newman WG, Nielsen SF, Nordestgaard BG, Norman A, Offit K, Olson JE, Olsson H, Orr N, Pankratz VS, Park-Simon TW, Perez JIA, Pérez-Barrios C, Peterlongo P, Peto J, Pinchev M, Plaseska-Karanfilska D, Polley EC, Prentice R, Presneau N, Prokofyeva D, Purrington K, Pylkäs K, Rack B, Radice P, Rau-Murthy R, Rennert G, Rennert HS, Rhenius V, Robson M, Romero A, Ruddy KJ, Ruebner M, Saloustros E, Sandler DP, Sawyer EJ, Schmidt DF, Schmutzler RK, Schneeweiss A, Schoemaker MJ, Schumacher F, Schürmann P, Schwentner L, Scott C, Scott RJ, Seynaeve C, Shah M, Sherman ME, Shrubsole MJ, Shu XO, Slager S, Smeets A, Sohn C, Soucy P, Southey MC, Spinelli JJ, Stegmaier C, Stone J, Swerdlow AJ, Tamimi RM, Tapper WJ, Taylor JA, Terry MB, Thöne K, Tollenaar RAEM, Tomlinson I, Truong T, Tzardi M, Ulmer HU, Untch M, Vachon CM, van Veen EM, Vijai J, Weinberg CR, Wendt C, Whittemore AS, Wildiers H, Willett W, Winqvist R, Wolk A, Yang XR, Yannoukakos D, Zhang Y, Zheng W, Ziogas A; ABCTB Investigators; kConFab/AOCS Investigators; NBCS Collaborators, Dunning AM, Thompson DJ, Chenevix-Trench G, Chang-Claude J, Schmidt MK, Hall P, Milne RL, Pharoah PDP, Antoniou AC, Chatterjee N, Kraft P, García-Closas M,

Simard J, Easton DF. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet.* 2019104(1):21-34. PMID: 30554720

74. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet.* 2017;100(4):635-649. PMID: 28366442.

75. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet.* 2019;51(4):584-591. PMID: 30926966.

76. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, Boyle EA, Zhang X, Racimo F, Pritchard JK, Coop G. Reduced signal for polygenic adaptation of height in UK Biobank. *Elife*. 2019;8:e39725. PMID: 30895923.

77. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, Chiang CW, Hirschhorn J, Daly MJ, Patterson N, Neale B, Mathieson I, Reich D, Sunyaev SR. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife*. 2019;8:e39702. PMID: 30895926

78. Hivert MF, Jablonski KA, Perreault L, Saxena R, McAteer JB, Franks PW, Hamman RF, Kahn SE, Haffner S; DIAGRAM Consortium, Meigs JB, Altshuler D, Knowler WC, Florez JC; Diabetes Prevention Program Research Group. Updated genetic score based on 34 confirmed type 2 diabetes Loci is associated with diabetes incidence and regression to normoglycemia in the diabetes prevention program. *Diabetes*. 2011;60(4):1340-8. PMID: 21378175.

79. Grant RW, O'Brien KE, Waxler JL, Vassy JL, Delahanty LM, Bissett LG, Green RC, Stember KG, Guiducci C, Park ER, Florez JC, Meigs JB. Personalized genetic risk counseling to motivate diabetes prevention: a randomized trial. Diabetes Care. 2013;36(1):13-9. PMID 22933432.

80. Martens FK, Tonk ECM, Janssens ACJW. Evaluation of polygenic risk models using multiple performance measures: a critical assessment of discordant results. *Genet Med.* 2019;21(2):391-397. PMID: 29895851;

81. Hollands GJ, French DP, Griffin SJ, Prevost AT, Sutton S, King S, Marteau TM. The impact of communicating genetic risks of disease on risk-reducing health behaviour: systematic review with meta-analysis. BMJ. 2016;352:i1102. PMID: 26979548;

82. Franks PW, McCarthy MI. Exposing the exposures responsible for type 2 diabetes and obesity. *Science*. 2016;354(6308):69-73. PMID: 27846494.

83. Dimas AS, Lagou V, Barker A, Knowles JW, Mägi R, Hivert MF, Benazzo A, Rybin D, Jackson AU, Stringham HM, Song C, Fischer-Rosinsky A, Boesgaard TW, Grarup N, Abbasi FA, Assimes TL, Hao K, Yang X, Lecoeur C, Barroso I, Bonnycastle LL, Böttcher Y, Bumpstead S, Chines PS, Erdos MR, Graessler J, Kovacs P, Morken MA, Narisu N, Payne F, Stancakova A, Swift AJ, Tönjes A, Bornstein SR, Cauchi S, Froguel P, Meyre D, Schwarz PE, Häring HU, Smith U, Boehnke M, Bergman RN, Collins FS, Mohlke KL, Tuomilehto J, Quertemous T, Lind L, Hansen T, Pedersen O, Walker M, Pfeiffer AF, Spranger J, Stumvoll M, Meigs JB, Wareham NJ, Kuusisto J, Laakso M, Langenberg C, Dupuis J, Watanabe RM, Florez JC, Ingelsson E, McCarthy MI, Prokopenko I; MAGIC Investigators. Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes*. 2014;63(6):2158-71. PubMed PMID: 24296717.

84. Tan VY, Fevotte C. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. *IEEE Trans Pattern Anal Mach Intell.* 2013;35(7):1592-605.

85. Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences.* 1984;10(2-3):191-203.
86. Speliotes EK, Yerges-Armstrong LM, Wu J, Hernaez R, Kim LJ, Palmer CD, Gudnason V, Eiriksdottir G, Garcia ME, Launer LJ, Nalls MA, Clark JM, Mitchell BD, Shuldiner AR, Butler JL, Tomas M, Hoffmann U, Hwang SJ, Massaro JM, O'Donnell CJ, Sahani DV, Salomaa V, Schadt EE, Schwartz SM, Siscovick DS; NASH CRN; GIANT Consortium; MAGIC Investigators, Voight BF, Carr JJ, Feitosa MF, Harris TB, Fox CS, Smith AV, Kao WH, Hirschhorn JN, Borecki IB; GOLD Consortium. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. PLoS Genet. 2011;7(3):e1001324. PMID: 21423719;
87. Kozlitina J, Smagris E, Stender S, Nordestgaard BG, Zhou HH, Tybjærg-Hansen A, Vogt TF, Hobbs HH, Cohen JC. Exome-wide association study identifies a *TM6SF2* variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet.* 2014;46(4):352-6. PMID: 24531328.
88. Strawbridge RJ, Dupuis J, Prokopenko I, Barker A, Ahlqvist E, Rybin D, Petrie JR, Travers ME, Bouatia-Naji N, Dimas AS, Nica A, Wheeler E, Chen H, Voight BF, Taneera J, Kanoni S, Peden JF, Turrini F, Gustafsson S, Zabena C, Almgren P, Barker DJ, Barnes D, Dennison EM, Eriksson JG, Eriksson P, Eury E, Folkersen L, Fox CS, Frayling TM, Goel A, Gu HF, Horikoshi M, Isomaa B, Jackson AU, Jameson KA, Kajantie E, Kerr-Conte J, Kuulasmaa T, Kuusisto J, Loos RJ, Luan J, Makrilakis K, Manning AK, Martínez-Larrad MT, Narisu N, Nastase Mannila M, Ohrvik J, Osmond C, Pascoe L, Payne F, Sayer AA, Sennblad B, Silveira A, Stancáková A, Stirrups K, Swift AJ, Syvänen AC, Tuomi T, van 't Hooft FM, Walker M, Weedon MN, Xie W, Zethelius B; DIAGRAM Consortium; GIANT Consortium; MuTHER Consortium; CARDIoGRAM Consortium; C4D Consortium, Ongen H, Mälarstig A, Hopewell JC, Saleheen D, Chambers J, Parish S, Danesh J, Kooner J, Ostenson CG, Lind L, Cooper CC, Serrano-Ríos M, Ferrannini E, Forsen TJ, Clarke R, Franzosi MG, Seedorf U, Watkins H, Froguel P, Johnson P, Deloukas P, Collins FS, Laakso M, Dermitzakis ET, Boehnke M, McCarthy MI, Wareham NJ, Groop L, Pattou F, Gloyn AL, Dedoussis GV, Lyssenko V, Meigs JB, Barroso I, Watanabe RM, Ingelsson E, Langenberg C, Hamsten A, Florez JC. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes*. 2011;60(10):2624-34. PMID: 21873549.
89. Gaulton KJ, Ferreira T, Lee Y, Raimondo A, Mägi R, Reschen ME, Mahajan A, Locke A, Rayner NW, Robertson N, Scott RA, Prokopenko I, Scott LJ, Green T, Sparso T, Thuillier D, Yengo L, Grallert H, Wahl S, Frånberg M, Strawbridge RJ, Kestler H, Chheda H, Eisele L, Gustafsson S, Steinthorsdottir V, Thorleifsson G, Qi L, Karssen LC, van Leeuwen EM, Willems SM, Li M, Chen H, Fuchsberger C, Kwan P, Ma C, Linderman M, Lu Y, Thomsen SK, Rundle JK, Beer NL, van de Bunt M, Chalisey A, Kang HM, Voight BF, Abecasis GR, Almgren P, Baldassarre D, Balkau B, Benediktsson R, Blüher M, Boeing H, Bonnycastle LL, Bottinger EP, Burtt NP, Carey J, Charpentier G, Chines PS, Cornelis MC, Couper DJ, Crenshaw AT, van Dam RM, Doney AS, Dorkhan M, Edkins S, Eriksson JG, Esko T, Eury E, Fadista J, Flannick J, Fontanillas P, Fox C, Franks PW, Gertow K, Gieger C, Gigante B, Gottesman O, Grant GB, Grarup N, Groves CJ, Hassinen M, Have CT, Herder C, Holmen OL, Hreidarsson AB, Humphries SE, Hunter DJ, Jackson AU, Jonsson A, Jørgensen ME, Jørgensen T, Kao WH, Kerrison ND, Kinnunen L, Klopp N, Kong A, Kovacs P, Kraft P, Kravic J, Langford C, Leander K, Liang L, Lichtner P, Lindgren CM, Lindholm E, Linneberg A, Liu CT, Lobbens S, Luan J, Lyssenko V, Männistö S, McLeod O, Meyer J, Mihailov E,

Mirza G, Mühleisen TW, Müller-Nurasyid M, Navarro C, Nöthen MM, Oskolkov NN, Owen KR, Palli D, Pechlivanis S, Peltonen L, Perry JR, Platou CG, Roden M, Ruderfer D, Rybin D, van der Schouw YT, Sennblad B, Sigurðsson G, Stančáková A, Steinbach G, Storm P, Strauch K, Stringham HM, Sun Q, Thorand B, Tikkanen E, Tonjes A, Trakalo J, Tremoli E, Tuomi T, Wennauer R, Wiltshire S, Wood AR, Zeggini E, Dunham I, Birney E, Pasquali L, Ferrer J, Loos RJ, Dupuis J, Florez JC, Boerwinkle E, Pankow JS, van Duijn C, Sijbrands E, Meigs JB, Hu FB, Thorsteinsdottir U, Stefansson K, Lakka TA, Rauramaa R, Stumvoll M, Pedersen NL, Lind L, Keinanen-Kiukaanniemi SM, Korpi-Hyövälti E, Saaristo TE, Saltevo J, Kuusisto J, Laakso M, Metspalu A, Erbel R, Jöcke KH, Moebus S, Ripatti S, Salomaa V, Ingelsson E, Boehm BO, Bergman RN, Collins FS, Mohlke KL, Koistinen H, Tuomilehto J, Hveem K, Njølstad I, Deloukas P, Donnelly PJ, Frayling TM, Hattersley AT, de Faire U, Hamsten A, Illig T, Peters A, Cauchi S, Sladek R, Froguel P, Hansen T, Pedersen O, Morris AD, Palmer CN, Kathiresan S, Melander O, Nilsson PM, Groop LC, Barroso I, Langenberg C, Wareham NJ, O'Callaghan CA, Gloyn AL, Altshuler D, Boehnke M, Teslovich TM, McCarthy MI, Morris AP; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat Genet.* 2015;47:1415-25. PMID: 26551672.

90. Varshney A, Scott LJ, Welch RP, Erdos MR, Chines PS, Narisu N, Albanus RD, Orchard P, Wolford BN, Kursawe R, Vadlamudi S, Cannon ME, Didion JP, Hensley J, Kirilusha A; NISC Comparative Sequencing Program, Bonnycastle LL, Taylor DL, Watanabe R, Mohlke KL, Boehnke M, Collins FS, Parker SC, Stitzel ML. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc Natl Acad Sci U S A.* 2017;114(9):2301-2306. PMID: 28193859.

91. Thurner M, van de Bunt M, Torres JM, Mahajan A, Nylander V, Bennett AJ, Gaulton KJ, Barrett A, Burrows C, Bell CG, Lowe R, Beck S, Rakyan VK, Gloyn AL, McCarthy MI. Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes susceptibility loci. *Elife.* 2018;7:e31977. PMID: 29412141.

92. van de Bunt M, Manning Fox JE, Dai X, Barrett A, Grey C, Li L, Bennett AJ, Johnson PR, Rajotte RV, Gaulton KJ, Dermitzakis ET, MacDonald PE, McCarthy MI, Gloyn AL. Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. *PLoS Genet.* 2015;11(12):e1005694. PMID: 26624892.

93. Fernández-Tajes J, Gaulton KJ, van de Bunt M, Torres J, Thurner M, Mahajan A, Gloyn AL, Lage K, McCarthy MI. Developing a network view of type 2 diabetes risk pathways through integration of genetic, genomic and functional data. *Genome Med.* 2019;11(1):19 PMID: 30914061.

94. Sandholm N, Van Zuydam N, Ahlqvist E, Juliusdottir T, Deshmukh HA, Rayner NW, Di Camillo B, Forsblom C, Fadista J, Ziemek D, Salem RM, Hiraki LT, Pezzolesi M, Trégouët D, Dahlström E, Valo E, Oskolkov N, Ladenvall C, Marcovecchio ML, Cooper J, Sambo F, Malovini A, Manfrini M, McKnight AJ, Lajer M, Harjutsalo V, Gordin D, Parkkonen M; The FinnDiane Study Group, Tuomilehto J, Lyssenko V, McKeigue PM, Rich SS, Brosnan MJ, Fauman E, Bellazzi R, Rossing P, Hadjadj S, Krolewski A, Paterson AD; The DCCT/EDIC Study Group, Florez JC, Hirschhorn JN, Maxwell AP; GENIE Consortium, Dunger D, Cobelli C, Colhoun HM, Groop L, McCarthy MI, Groop PH; SUMMIT Consortium. The Genetic Landscape of Renal Complications in Type 1 Diabetes. *J Am Soc Nephrol.* 2017;28(2):557-574. PMID: 27647854.

95. van Zuydam NR, Ahlqvist E, Sandholm N, Deshmukh H, Rayner NW, Abdalla M, Ladenvall C, Ziemek D, Fauman E, Robertson NR, McKeigue PM, Valo E, Forsblom C, Harjutsalo V; Finnish Diabetic Nephropathy Study (FinnDiane), Perna A, Rurali E, Marcovecchio ML, Igo RP Jr, Salem RM, Perico N, Lajer M, Käräjämäki A, Imamura M, Kubo M, Takahashi A, Sim X, Liu J, van Dam RM, Jiang G, Tam CHT, Luk AOY, Lee HM, Lim CKP, Szeto CC, So WY, Chan JCN; Hong Kong Diabetes Registry Theme-based Research Scheme Project Group, Ang SF, Dorajoo R, Wang L, Clara TSH, McKnight AJ, Duffy S; Warren 3 and Genetics of Kidneys in Diabetes (GoKinD) Study Group, Pezzolesi MG; GENIE (GEnetics of Nephropathy an International Effort) Consortium, Marre M, Gyorgy B, Hadjadj S, Hiraki LT; Diabetes Control and Complications Trial (DCCT)/Epidemiology of Diabetes Interventions and Complications (EDIC) Research Group, Ahluwalia TS, Almgren P, Schulz CA, Orho-Melander M, Linneberg A, Christensen C, Witte DR, Grarup N, Brandslund I, Melander O, Paterson AD, Tregouet D, Maxwell AP, Lim SC, Ma RCW, Tai ES, Maeda S, Lyssenko V, Tuomi T, Krolewski AS, Rich SS, Hirschhorn JN, Florez JC, Dunger D, Pedersen O, Hansen T, Rossing P, Remuzzi G; SUrrogate markers for Micro- and Macrovascular hard endpoints for Innovative diabetes Tools (SUMMIT) Consortium, Brosnan MJ, Palmer CNA, Groop PH, Colhoun HM, Groop LC, McCarthy MI. A Genome-Wide Association Study of Diabetic Kidney Disease in Subjects With Type 2 Diabetes. *Diabetes*. 2018;67(7):1414-1427. PMID: 29703844.

96. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang T, Paige E, Surendran P, Oliver-Williams C, Kamat MA, Prins BP, Wilcox SK, Zimmerman ES, Chi A, Bansal N, Spain SL, Wood AM, Morrell NW, Bradley JR, Janjic N, Roberts DJ, Ouwehand WH, Todd JA, Soranzo N, Suhre K, Paul DS, Fox CS, Plenge RM, Danesh J, Runz H, Butterworth AS. Genomic atlas of the human plasma proteome. Nature. 2018 ;558(7708):73-79. PMID: 29875488.

97. Kettunen J, Demirkan A, Würtz P, Draisma HH, Haller T, Rawal R, Vaarhorst A, Kangas AJ, Lyytikäinen LP, Pirinen M, Pool R, Sarin AP, Soininen P, Tukiainen T, Wang Q, Tiainen M, Tynkkynen T, Amin N, Zeller T, Beekman M, Deelen J, van Dijk KW, Esko T, Hottenga JJ, van Leeuwen EM, Lehtimäki T, Mihailov E, Rose RJ, de Craen AJ, Gieger C, Kähönen M, Perola M, Blankenberg S, Savolainen MJ, Verhoeven A, Viikari J, Willemsen G, Boomsma DI, van Duijn CM, Eriksson J, Jula A, Järvelin MR, Kaprio J, Metspalu A, Raitakari O, Salomaa V, Slagboom PE, Waldenberger M, Ripatti S, Ala-Korpela M. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nat Commun. 2016;7:11122. PMID: 27005778;

98. Feng Y, Mao G, Ren X, Xing H, Tang G, Li Q, Li X, Sun L, Yang J, Ma W, Wang X, Xu X. Ser1369Ala variant in sulfonylurea receptor gene ABCC8 is associated with antidiabetic efficacy of gliclazide in Chinese type 2 diabetic patients. *Diabetes Care*. 2008;31(10):1939-44. PMID: 18599530.

99. Florez JC, Jablonski KA, Sun MW, Bayley N, Kahn SE, Shamoon H, Hamman RF, Knowler WC, Nathan DM, Altshuler D; Diabetes Prevention Program Research Group. Effects of the type 2 diabetes-associated PPARG P12A polymorphism on progression to diabetes and response to troglitazone. *J Clin Endocrinol Metab.* 2007;92(4):1502-9. PMID: 17213274.

100. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, Bottinger EP, Dudley JT. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med.* 2015;7(311):311ra174. PMID: 26511511.

101. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, Vikman P, Prasad RB, Aly DM, Almgren P, Wessman Y, Shaat N, Spégel P, Mulder H,

Lindholm E, Melander O, Hansson O, Malmqvist U, Lernmark Å, Lahti K, Forsén T, Tuomi T, Rosengren AH, Groop L. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol.* 2018;6(5):361-369. PMID: 29503172.

102. Dennis JM, Shields BM, Henley WE, Jones AG, Hattersely AT. Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. Lancet Diabetes Endocrinol. 2019 Jun;7(6):442-451. PMID: 31047901.

103. Janssens ACJW, Joyner MJ. Polygenic Risk Scores That Predict Common Diseases Using Millions of Single Nucleotide Polymorphisms: Is More, Better? *Clin Chem.* 2019 May;65(5):609-611. PMID: 30808642.

104. Curtis D. Clinical relevance of genome-wide polygenic score may be less than claimed. *Ann Hum Genet* 2019 Jul;83(4):274-277. PMID: 30906985.

105. Payne K, Gavan SP, Wright SJ, Thompson AJ. Cost-effectiveness analyses of genetic and genomic diagnostic tests. *Nat Rev Genet.* 2018;19(4):235-246. PMID: 29353875.

106. Health Education England. Preparing the healthcare workforce to deliver the digital future. Published 11 Feb 2019. https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf

107. Joly Y, Dyke SOM, Knoppers BM, Pastinen T. Are Data Sharing and Privacy Protection Mutually Exclusive? *Cell.* 2016;167(5):1150-1154. PMID: 27863233.

108. Rose G. Sick individuals and sick populations. *Int J Epidemiol.* 1985 Mar;14:32-8. PMID: 3872850

Figure 1: How polygenic scores are derived. For full explanation see text.

Figure 2. Comparison of rsPS and gePS for T2D using data from Mahajan et al, 2018b [9] rsPS and gePS were generated using a T2D GWAS meta-analysis of 455,313 European individuals and used to predict incident T2D in 13,480 cases and 311,390 controls from the UK Biobank. **a**) AUROC curves for models predicting incident T2D: each model was adjusted for genotyping array and the first six principal components of ancestry. **b**) Prevalence of T2D according to 40 groups binned according to the polygenic scores, with each grouping representing 2.5% of the population. **c**) Distribution of rsPS and gePS in the cases and controls. The x-axis represents polygenic score, with values scaled to a mean of 0 and standard deviation of 1. Both rsPS and gePS in UK Biobank individuals is normally distributed with a shift towards right, observed for T2D cases.

Table 1. Comparison of three published global extended polygenic scores for T2D. For the LDPred algorithm, the tuning parameter ρ reflects the proportion of polymorphisms assumed to be causal for the disease. For the pruning and thresholding strategy, r^2 reflects the degree of independence from other variants in the linkage disequilibrium, and *P* value reflects the *P* value threshold used for a selecting variants from the discovery GWAS. * Discovery GWAS from Mahajan et al. 2018b after removing UK Biobank samples [9]. Note the difference in testing dataset sample size from the published results in Mahajan et al. 2018b [9]. Results presented here are based on re-analysis of data after splitting UK Biobank samples into optimization and testing set. ** Logistic model adjusted for other technical covariates such as principal components. $^{\$}$ Subset of GWAS samples. $^{\#}$ Obtained through private communication with authors. LD: Linkage Disequilibrium.

	Study		
	Khera et al. 2018 [13]	Mahajan et al. 2018b [9]	23andMe [43]

Discovery GWAS	**Number of cases**	26,676	55,005	80,792
	Number of controls	132,532	400,308	1,479,116
	Reference	Scott et al. 2017 [44]	Mahajan et al. 2018b* [9]	Multhaup et al. 2019 [43]
Optimisation dataset	**Methods**	LDPred	Pruning and thresholding	Predetermined cut offs
	Number of cases	2,785	5,639	48,028
	Number of controls	120,280	112,307	893,692
	***P* value threshold**	-	0.1	$1x10^{-5}$
	LD pruning threshold	-	r^2>0.6	50 kb window
	Tuning parameter	ρ = 0.01	-	-
	Polymorphisms in risk score	6,917,436	171,249	1,244
	Reference	UK Biobank	UK Biobank	23andMe[$]
Testing dataset	**Number of cases**	5,853	13,480	9,008
	Number of controls	288,978	311,390	167,622
	Reference	UK Biobank	UK Biobank	23andMe
AUROC in testing dataset (Europeans)	**Not adjusted for age and sex****	0.64[#]	0.66	0.65
	Adjusted for age and sex	0.73	0.73	-
	Odds ratio of top 5% bin vs remainder population	2.75	2.75 without age and sex adjustment 4.52 with age and sex adjustment	2.76[#]

Table 2. Partitioned polygenic score clusters capturing etiological heterogeneity in T2D. Comparison of pPS clusters identified by Mahajan *et al.* [20] and Udler *et al.* [38]. TG: Triglycerides; BMI: Body mass index; WHR: waist hip ratio

Physiological impact		**Phenotypic features**	**Cluster name**		**Examples of T2D loci**
			Udler *et al.* 2018 [38]	**Mahajan *et al.* 2018a [20]**	
Adverse impact on β-cell function	High proinsulin	Low fasting insulin (+ High proinsulin)	Beta-Cell	Insulin Secretion 1	*ABO, ADCY5, HNF1A, HNF1B, MTNR1B, SLC30A8, TCF7L2*
	Low proinsulin	Low fasting insulin (+ Low proinsulin)	Proinsulin	Insulin Secretion 2	*IGF2BP2, CENTD2/ARAP1, CCND2*
Reduced insulin sensitivity	Mediation via obesity	High TG + High WHR + Low BMI	Lipodystrophy	Insulin Action	*MACF1, GRB14, IRS1, PPARG, ANKRD55, KLF14, LPL, CMIP*
	Mediation via fat distribution	High BMI + High WHR	Obesity	Adiposity	*NRXN3, FTO, MC4R*
	Mediation via lipid metabolism	Low TG	Liver/Lipid	Dyslipidaemia	*GCKR, TM6SF2/CILP2*
Undetermined		No striking phenotype association	No assignment	Mixed features	*BCL11A, TLE1, PLEKHA1, HMGA2, MTMR3*