**Progress in defining the genetic contribution to type 2 diabetes susceptibility**

Andrew P Morris

Department of Biostatistics, University of Liverpool, Liverpool, UK

*Contact information*

Prof. Andrew Morris

Department of Biostatistics, University of Liverpool

Block F, Waterhouse Building

1-5 Brownlow Street

Liverpool L69 3GL

United Kingdom

E-mail: apmorris@liverpool.ac.uk

**Abstract.** Candidate gene, genome-wide association, exome array and sequencing studies have identified more than 140 loci associated with type 2 diabetes (T2D) susceptibility. In this review, progress in understanding the genetic architecture of T2D susceptibility across diverse populations and in localising potential causal variants for the disease through fine-mapping studies is discussed. The additional insights gained from these genetic studies into novel molecular mechanisms and pathophysiology underlying T2D susceptibility are described, and the prospects for future genomic investigations of the disease are considered.

Short title: The genetic contribution to type 2 diabetes susceptibility.

Number of words: 3,016.

**Introduction.** Type 2 diabetes (T2D) mellitus is a common complex disease that currently affects more than 400 million people across the globe. The disease is characterised by insulin resistance and beta cell dysfunction [1], and can lead to a range of microvascular and macrovascular complications [2], increased risk of cardiovascular disease [3] and reduced life-expectancy [4]. Whilst T2D most often occurs as a downstream consequence of obesity, the disease is observed to run in families, with a sibling relative risk of ~2.8 [5] and estimated heritability of 30-70% [6]. Improved awareness of the genetic contribution to T2D susceptibility has thus been considered an important avenue to understanding the underlying pathophysiology of the disease that will help to address the limitations of currently available preventive and therapeutic options.

The earliest genetic investigations of T2D susceptibility focussed on small-scale studies of candidate genes, often identified through family-based linkage analyses or relevant biological insight from monogenic forms of the disease. The results of these studies were often difficult to validate, and success was limited to the identification of variants in a handful of loci including *PPARG* [7], *KCNJ11* [8] and *TCF7L2* [9]. In contrast, the first well-designed genome-wide association studies (GWAS) of T2D susceptibility made use of a “hypothesis-free” approach that interrogates (many) hundreds of thousands of common single nucleotide polymorphisms (SNPs) across the genome, typically defined to have minor allele frequency of at least 5%. These GWAS substantially increased the number of confirmed disease loci, including *CDKAL1*, *CDKN2A-B*, *HHEX-IDE*, *IGF2BP2*, *SLC30A8*, *WFS1*, and *FTO* (primary effect on risk via obesity) [10-16]. However, it is important to note that GWAS loci have historically been named by the gene that maps closest to the lead SNP (i.e. that with strongest association signal), but that this does not imply causality.

Early endeavours to understand the genetic contribution to T2D susceptibility were enhanced through international collaborative efforts that increased power to detect disease loci through meta-analysis of GWAS for the disease and related glycemic traits by the DIAGRAM Consortium and MAGIC Investigators [17-19] in European ancestry populations. Such meta-analyses typically involve the provision of association summary statistics for each SNP (such as allelic odds-ratios and confidence intervals, *p*-values, and allele frequencies) to a “central analysis team” across GWAS contributing to the consortium, without the need to exchange individual-level genotype and phenotype data. These efforts expanded the genetic landscape of T2D susceptibility to more than 40 loci at genome-wide significance (*p*<5x10-8), but also revealed important novel biology, with effects of associated SNPs on both beta-cell function and insulin action, and enrichment for association signals in/near genes involved in cell cycle regulation.

Despite the success of early GWAS meta-analyses in identifying novel loci contributing to T2D susceptibility, challenges remain in the utility of the findings of these studies for clinical translation through informing development of novel therapeutics for the disease and enabling implementation of personalised medicine strategies [20]. First, association signals at GWAS loci explained relatively little (~10%) of the heritability of T2D susceptibility, inhibiting personalised risk prediction. Second, association signals typically extended over large genomic intervals because of linkage disequilibrium between common SNPs in European ancestry populations, limiting the resolution of fine-mapping efforts to localise the causal variant(s) at each locus, and identify the effector transcript(s) through which their effects are mediated. In this article, progress in addressing these challenges will be reviewed, focussing on larger GWAS meta-analyses of common SNPs across diverse populations, and whole-genome and whole-exome sequencing efforts to access lower frequency genetic variants.

**Additional T2D susceptibility loci identified through GWAS.** Since 2010, the most widespread approach to the discovery of additional loci contributing to T2D susceptibility has been through GWAS. These efforts have been bolstered by improved efficiency of GWAS genotyping technology, enabling interrogation of larger numbers of SNPs that better cover common genetic variation across populations in increased sample sizes. Furthermore, methodological innovations, such as imputation [21], which enables prediction of genotypes at SNPs not typed on GWAS arrays, but which are present in high-density reference panels of whole-genome sequence data, such as those from the 1000 Genomes Project [22] and Haplotype Reference Consortium [23], allow association testing at millions of variants across the genome (**Table 1**).

As described above, the largest GWAS meta-analyses for T2D susceptibility, to date, have been undertaken by the DIAGRAM Consortium in populations of European ancestry [24,25\*], predominantly because of existing infrastructure, sample availability, and relatively poor coverage of common genetic variation in other major ethnic groups by early genotyping arrays [26]. However, European ancestry populations do not fully characterise T2D risk variants in other ethnic groups, and GWAS meta-analyses have also been undertaken in East Asians [27,28], South Asians [29], Hispanics/Latinos [30,31], and African Americans [32]. These studies have highlighted overlap in T2D susceptibility loci between ethnicities and provided evidence that genetic risk scores derived from European ancestry GWAS are transferrable to other population groups [33]. Investigations across ethnic groups have revealed minimal evidence of heterogeneity in allelic odds-ratios on T2D susceptibility at lead SNPs, despite substantial differences in allele frequencies [34,35\*]. T2D risk alleles across the genome also demonstrate consistent directions of effect across diverse populations at loci with only nominal association with disease susceptibility [34]. These observations imply that common causal variants at many T2D susceptibility loci are shared across ancestry groups, and arose prior to human population migration out of Africa, prompting trans-ethnic GWAS meta-analyses to maximise sample size and power for additional discoveries [34,36\*]. On the other hand, lower-frequency variation is more likely to be ethnic- or even population-specific, highlighting the important of undertaking genetic investigations of T2D susceptibility across diverse populations. For example, GWAS undertaken in the isolated Greenlandic population highlighted a missense variant in *TBC1D4* that confers muscle insulin resistance and T2D risk, which is rare or monomorphic in other ancestry groups [37]. Furthermore, differences in LD structure between common variants across ethnicities is beneficial for localising causal variants underlying association signals, as described below.

**The genetic architecture of type 2 diabetes susceptibility.** GWAS have continued to be extremely successful in identifying loci contributing to T2D susceptibility. Association signals at the majority of these loci are defined by common lead SNPs with modest effects on the disease (allelic odds-ratios less than 1.5). As a consequence, these loci jointly explain no more than ~20% of the heritability of the disease, and thus have limited predictive power for future disease in unaffected individuals when compared with traditional clinical and lifestyle risk factors [38]. However, genetic risk scores that are derived from SNPs, including those with nominal associations with T2D susceptibility that map outside of loci attaining genome-wide significance, do show promise for disease prediction, after accounting for body-mass index, age and sex [39].

GWAS are not designed to capture rare genetic variation (typically defined to have minor allele frequency less than 0.5%), even after supplementation with imputation. There has therefore been considerable support for the notion that much of the “missing heritability” of complex human traits, including T2D susceptibility, could be explained by lower frequency variation, which can best be assayed through sequencing. To investigate this hypothesis, the GoT2D/T2D-GENES Consortium undertook the largest sequencing study (whole-genome and whole-exome) of T2D susceptibility, to date, in more than 15,000 individuals of diverse ancestry [40\*\*]. A coding variant (*PAX4* p.Arg192His) attained genome-wide significance at the *GCC1* locus that was common in East Asian ancestry populations (minor allele frequency ~10%), but virtually absent from other ethnic groups. These data provided no evidence to support the “synthetic association” hypothesis [41], under which common GWAS SNP signals can be explained by rare variants that are not interrogated through genotyping and imputation. Modelling of disease architecture demonstrated that: (i) the whole-genome association data were consistent with a “common polygenic” model in which large numbers of common variants of modest effect explain about 75% of T2D heritability; and (ii) across the exome, the overall contribution of rare and low-frequency coding variants (minor allele frequency in the range of 0.1% to 5%) to T2D risk was just 2.9%, compared to 6.3% for common coding variants.

**Fine-mapping type 2 diabetes susceptibility loci.** GWAS loci are typically defined by a single lead SNP, a representation that does not take account of the possibility that the T2D association derives from multiple causal variants acting independently of each other or through haplotype effects. Moreover, the association often extends over hundreds of kilobases, including multiple genes through which the effect on T2D susceptibility may be mediated. It has become common, therefore, to name loci according to the gene mapping closest to the lead SNP, unless there is a more compelling biological candidate nearby. However, with the exception of loci such as *SLC30A8* and *GCKR*, where the lead SNP is a coding variant, and functional studies have supported T2D association signals, these labels are effectively arbitrary and offer no insight into relevant disease biology [42]. Comprehensive fine-mapping studies require an exhaustive catalogue of genetic variation across a locus, which can best be achieved through sequencing, although this remains prohibitively expensive in the large sample sizes needed for complex human traits. On the other hand, imputation into existing GWAS can also provide near complete coverage of common and low-frequency variation (with minor allele frequency as low as 0.1%) across a locus, but is limited to variants that are present in the reference panel.

The first step in fine-mapping GWAS loci is to undertake conditional analyses to “delineate” distinct association signals (i.e. statistically independent) arising from multiple causal variants in the same region. Across T2D susceptibility loci, such analyses have revealed that distinct association signals are widespread, with the most dramatic delineation observed for the region flanking *KCNQ1* (**Figure 1**), where up to six index variants have been reported at locus-wide significance (*p*<10-5) in European ancestry and trans-ethnic meta-analyses, three of which map to a less than 50kb intronic region of the gene, and all showing homogenous allelic odds-ratios across diverse populations [25\*,35\*,43\*\*]. These meta-analyses also reported multiple distinct signals of T2D association at loci mapping to/near *CCND2*, *DGKB*, *GIPR*, *HNF1A*, *HNF4A* and *MC4R*, and also at *CDKN2A-B*, where the index variants are located in a 12kb inter-genic recombination interval and represent an established haplotype effect on disease risk [11,44,45].

The second step in fine-mapping is then to localise likely causal variants underlying each distinct T2D association signal. The most widely used approach to localisation is to construct 99% (or 95%) credible sets of variants for each signal that are the smallest set that together account for 99% (or 95%) of the posterior probability of driving the association [45]. To date, the most comprehensive fine-mapping study focussed on localising causal variants driving 49 distinct association signals across 37 T2D susceptibility loci in GWAS of European ancestry genotyped with the Metabochip, and supplemented with imputation up to reference panels from the 1000 Genomes Project [43\*\*]. From these analyses, the 99% credible set included no more than ten variants for ten distinct association signals, mapping to nine loci. The most precise localisation was observed at the *MTNR1B* locus, where the credible set included only the index variant, rs10830963, which accounted for more than 99.8% of the posterior probability of driving the T2D association signal. Small 99% credible sets, consisting of just three variants, were also observed for the association at the *TCF7L2* locus (mapping to 4.3kb), and one of the signals at the *KCNQ1* locus (mapping to just 200bp).

Fine-mapping efforts have been greatly enhanced through trans-ethnic meta-analysis of GWAS by taking advantage of differences in LD structure between diverse populations [26,46,47]. This was demonstrated in an investigation of four T2D susceptibility loci (*CDKAL1*, *CDKN2A-B*, *IGF2BP2* and *KCNQ1*), interrogated via GWAS in populations of European, Asian, African American and Hispanic/Latino ancestry, fine-mapping resolution at distinct association signals was improved after trans-ethnic meta-analysis when compared with ethnic-specific analyses (**Figure 2**), in terms of the number of SNPs reported in the credible set and/or the genomic interval covered [35\*].

**Insight from genetic studies into novel molecular mechanisms and pathophysiology underlying T2D susceptibility.** Improved fine-mapping of T2D susceptibility loci offers exciting opportunities to improve understanding of disease biology, both in terms of causal molecular mechanisms and effector genes through which the effects of association signals are mediated, which might point to novel targets (genes and pathways) for drug development. The most direct route to biological inference occurs when a T2D association signal is localised to a coding variant. For example, *GCKR* p.Pro446Leu is the lead SNP at the *GCKR* locus, and accounts for a substantial proportion of the posterior probability of driving the T2D association signal. The T2D risk allele at this variant has also been reproducibly associated with increased fasting glucose and decreased triglycerides [18,48]. Functional studies have demonstrated that the variant influences hepatic glucose metabolism, offering a plausible mechanism to explain the observed associations with T2D and related metabolic traits [49,50]. Coding index variants for distinct signals at *HNF4A* (p.Thr139Ile) and *HNF1A* (p.Ile27Leu and p.Ala98Val) are also highly likely to be driving T2D associations, pointing to likely effector transcripts at these loci, which is also supported by the known impact of rare, loss of function mutations in these genes on maturity onset diabetes of the young [51,52]. Early GWAS [13] also identified a common missense variant driving the T2D association signal at *SLC30A8* (p.Trp325Arg), which is also associated with fasting glucose and fasting proinsulin [18,53]. The evidence supporting *SLC30A8* as the effector transcript at this locus has been further bolstered by the observation of significantly reduced T2D risk amongst carriers of low-frequency loss-of-function mutations in the gene [54].

It is important to remember, however, that not all coding variant associations will be causal, but might instead reside on haplotypes that incorporate many non-coding alleles because of LD between common SNPs, and thus might lead to incorrect inference of the relevant effector transcript at the locus. It is therefore essential to consider the association of coding variants identified through exome-sequencing or exome-array genotyping in their regional context through fine-mapping of the locus with available GWAS data [55\*\*]. Indeed, comprehensive fine-mapping studies of T2D susceptibility loci have highlighted that association signals map predominantly to non-coding sequence, and are most likely to be mediated by effects on gene regulation [25\*,35\*,43\*\*].

One approach to understand the regulatory mechanisms through which T2D association signals impact on disease is through integration of credible set variants with data from molecular profiling studies of diverse cell types, including international collaborative efforts such as Epigenome Roadmap Project and the ENCODE Project [56,57], and investigations focussed on more relevant disease-related tissues, including pancreatic islets and liver [58,59]. After fine-mapping, T2D association signals have been demonstrated to be enriched in pancreatic islet active enhancers and/or promoters, and binding sites for FOXA2, EZH2, NKX2.2 and PDX1 in pancreatic islets and across tissues [25\*,43\*\*].

At the *MTNR1B* locus, the lead non-coding SNP, rs10830963, accounts for more than 99% of the posterior probability of driving the T2D association. The SNP overlaps a FOXA2 binding site and maps to an enhancer element in islets and liver-derived cells, indicating that it is located in a transcriptionally active region, and the risk allele is predicted to create a recognition motif that matches the consensus sequence of NEUROD1. Functional experimentation demonstrated that the risk allele preferentially binds NEUROD1 in islet-derived cells, *in vitro*, and increases FOXA2-bound enhancer activity in human islets [42\*\*]. These results suggest altered NEUROD1 binding in islets contributes to T2D susceptibility at this locus and are consistent with previous reports correlating the risk allele with higher *MTNR1B* expression [60,61].

Investigation of the impact of lead SNPs at T2D susceptibility loci on diabetes-related quantitative phenotypes, such as fasting glycemic traits, anthropometric measures and lipid profiles, can provide insight into relevant biology underlying the disease. For example, hierarchical clustering of metabolic trait association *Z*-scores, aligned to the T2D risk alleles at lead SNPs, highlighted three major clades of loci with primary impact on: (i) insulin secretion and processing (including *TCF7L2*, *CDKAL1* and *IGF2BP2*); (ii) insulin action (including *PPARG* and *IRS1*); and (iii) body-mass index and lipid levels (including *MC4R* but not *FTO*) [25\*]. The clusters also demonstrated tissue-specific enrichments in regulatory elements that were consistent with the observed metabolic trait associations.

**Concluding remarks: prospects for genetic studies of T2D susceptibility.** Looking ahead, GWAS of T2D susceptibility will undoubtedly continue to expand the catalogue of loci contributing to the disease. The increasing availability of population biobanks, such as the UK Biobank and Million Veteran Program, with high-density genotype data (from GWAS and sequencing), demographic and lifestyle information, and linkage to electronic medical records, will enhance discovery efforts, particularly through application of multi-phenotype methods that model the correlation between T2D and diabetes-related metabolic traits. Assessment of the contribution of lower frequency variation to T2D susceptibility will be augmented through larger-scale whole-genome and whole-exome sequencing, and the availability of improved statistical methods to aggregate rare variants to evaluate their joint effects on the disease. The opportunities for fine-mapping will continue to be enhanced with the increasing availability of GWAS in diverse populations, and larger and more dense reference panels for imputation. Improved genomic annotation, particularly in non-coding regions, and expression data from densely genotyped samples in diabetes-relevant tissues will offer increased understanding of the biological mechanisms underlying T2D association signals, and will be ameliorated through methodological development to allow integration of these diverse data resources. Finally, it will be essential to develop high-throughput, tractable animal models and *in vitro* assays to enable exhaustive assessment of the functional impact of potential causal genes identified through genetic studies of T2D susceptibility. Together, these efforts offer promising avenues for the “holy grail” of clinical translation of genetic discoveries from GWAS and sequencing, through identification of “allelic series” of common and rare variants, such as the missense and loss-of-function variants mutations reported in *SLC30A8* [54], to the assessment of their effects on gene function or expression in diabetes-relevant tissues, which will support the targeted development or repurposing of therapeutics. Through co-ordinated collaboration between researchers across a wide range of disciplines, including human genetics, functional genomics and computational biology, the community should be optimistic that genetic studies of T2D susceptibility will offer exciting and realistic prospects for the prevention and treatment of the disease.

**Conflict of interest.** The author declares no conflict of interest.

**References**

1. Kahn SE *et al.* (2014). Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. *Lancet* 383: 1068-83.
2. Fowler MJ, *et al.* (2008). Microvascular and macrovascular complications of diabetes. *Clin Diabetes* 26: 77-82.
3. Rao Kondapally Seshasai S, *et al.* (2011). Diabetes mellitus, fasting glucose, and risk of cause-specific death. *N Engl J Med* 364: 829-41.
4. Tancredi M, *et al.* (2015). Excess mortality among persons with type 2 diabetes. *N Engl J Med* 373: 1720-32.
5. Hemminki K, *et al.* (2010). Familial risks for type 2 diabetes in Sweden. *Diabetes Care* 33: 293-7.
6. Almgren P, *et al.* (2011). Heritability and familiarity of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* 54: 2811-9.
7. Altshuler D, *et al.* (2000). The common *PPARγ* Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* 26: 76-80.
8. Gloyn AL, *et al.* (2003). Large-scale association studies of variants in genes encoding the pancreatic β-cell KATP channel subunits Kir6.2 (*KCNJ11*) and SUR1 (*ABCC8*) confirms that the *KCNJ11* E23K variant is associated with type 2 diabetes. *Diabetes* 52: 568-72.
9. Grant SF, *et al.* (2006). Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nat Genet* 38: 320-3.
10. Diabetes Genetics Initiative (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316: 1331-6.
11. Zeggini E, *et al.* (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316: 1336-41.
12. Scott LJ, *et al.* (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316: 1341-5.
13. Sladek R, *et al.* (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445: 881-5.
14. Steinthorsdottir V, *et al.* (2007). A variant in *CDKAL1* influences insulin response and risk of type 2 diabetes. *Nat Genet* 39: 770-5.
15. Frayling TM, *et al.* (2007). A common variant in the *FTO* gene is associated with body-mass index and predisposes to childhood and adult obesity. *Science* 316: 889-94.
16. Sandhu MS, *et al*. (2007). Common variants in WFS1 confer risk of type 2 diabetes. *Nat Genet* 39: 951-3.
17. Zeggini E, *et al.* (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 40: 638-45.
18. Dupuis J, *et al.* (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 42: 105-16.
19. Voight BF, *et al.* (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 42: 579-89.
20. McCarthy MI, *et al.* (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9: 356-69.
21. Marchini J, Howie B (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11: 499-511.
22. The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526: 68-74.
23. McCarthy S, *et al.* (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 48: 1279-83.
24. Morris AP, *et al.* (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 44: 981-90.
25. **Scott RA, *et al.* (2017). An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* 66: 2888-902. \*Largest and most comprehensive GWAS of T2D susceptibility in European ancestry populations that identified 13 novel loci and revealed tissue-specific enrichment of regulatory annotations.**
26. Rosenberg NA, *et al.* (2010). Genome-wide association studies in diverse populations. *Nat Rev Genet* 11: 356-66.
27. Cho YS, *et al.* (2012). Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in East Asians. *Nat Genet* 44: 67-72.
28. Imamura M, *et al.* (2016). Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. *Nat Commun* 7: 10531.
29. Kooner JS, *et al.* (2011). Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* 43: 984-9.
30. Parra EJ, *et al.* (2011). Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American samples from Starr County, Texas. *Diabetologia* 54: 2038-46.
31. SIGMA Type 2 Diabetes Consortium (2014). Sequence variants in *SLC16A11* are a common risk factor for type 2 diabetes in Mexico. *Nature* 506: 97-101.
32. Ng MC, *et al.* (2014). Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet* 20: e1004517.
33. Waters KM, *et al.* (2010). Consistent association of type 2 diabetes risk variants found in Europeans in diverse racial and ethnic groups. *PLoS Genet* 6: e1001078.
34. Mahajan A, *et al.* (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* 46: 234-44.
35. **Horikoshi M, *et al.* (2016). Transancestral fine-mapping of four type 2 diabetes susceptibility loci highlights potential causal regulatory mechanisms. *Hum Mol Genet* 25: 2070-81. \*Fine-mapping study of four T2D susceptibility loci that demonstrated the importance of GWAS meta-analysis in diverse populations for localising causal variants underlying association signals.**
36. **Zhao W, *et al.* (2017). Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat Genet* 49: 1450-7. \*Largest trans-ethnic GWAS meta-analysis for T2D susceptibility that identified 16 novel loci and co-localisation with association signals for coronary heart disease.**
37. Moltke I, *et al.* (2014). A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature* 512: 190-3.
38. Wang X, *et al.* (2016). Genetic markers of type 2 diabetes: progress in genome-wide association studies and clinical application for risk prediction. *J Diabet* 8: 24-35.
39. Läll K, *et al.* (2017). Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genet Med* 19: 322-9.
40. **Fuchsberger C, *et al.* (2016). The genetic architecture of type 2 diabetes. *Nature* 536: 41-7. \*\*Largest sequencing study of T2D susceptibility in diverse populations which enabled first large-scale investigation of the contribution of lower frequency variation to the disease.**
41. Dickson SP, *et al.* (2010). Rare variants create synthetic genome-wide associations. *PLoS Biol* 8: e1000294.
42. Ng HJ, Gloyn AL (2013). Bridging the gap between genetic associations and molecular mechanisms for type 2 diabetes. *Curr Diab Rep* 13: 778.
43. **Gaulton KJ, *et al.* (2015). Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat Genet* 47: 1415-25. \*\*Comprehensive fine-mapping study across 39 T2D susceptibility loci, which revealed potential causal regulatory mechanism at the *MTNR1B* locus.**
44. Shea J, *et al.* (2011). Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat Genet* 43: 801-5.
45. Maller JB, *et al.* (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* 44: 1294-1301.
46. Morris AP (2011). Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* 35: 809-22.
47. Mägi R, *et al.* (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum Mol Genet* 26: 3639-50.
48. Orho-Melander M, *et al.* (2008). Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations. *Diabetes* 57: 3112-21.
49. Beer NL, *et al.* (2009). The P446L variant in *GCKR* associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Hum Mol Genet* 18: 4081-8.
50. Rees MG, *et al.* (2012). Cellular characterisation of the *GCKR* P446L variant associated with type 2 diabetes risk. *Diabetologia* 55: 114-22.
51. Yamagata K, *et al.* (1996). Mutations in the hepatocyte nuclear factor-1 alpha gene in maturity-onset diabetes of the young (MODY3). *Nature* 384: 455-8.
52. Yamagata K, *et al.* (1996). Mutations in the hepatocyte nuclear factor-4 alpha gene in maturity-onset diabetes of the young (MODY1). *Nature* 384: 458-60.
53. Strawbridge RJ, *et al.* (2011). Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* 60: 2624-34.
54. Flannick J, *et al.* (2014). Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat Genet* 46: 357-63.
55. **Mahajan A, *et al.* (2017). Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat Genet* (under review), https://www.biorxiv.org/content/early/2017/05/31/144410.1. \*\*Largest study of the contribution of coding variation to T2D susceptibility in diverse populations, highlighting the importance of fine-mapping in determining causal variants and effector genes.**
56. Bernstein BE, *et al.* (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotech 28: 1045-8.
57. The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57-74.
58. Soccio RE, *et al.* (2011). Species-specific strategies underlying conserved functions of metabolic transcription factors. Mol Endocrinol 25: 694-706.
59. Pasquali L, *et al.* (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* 46: 136-43.
60. Fadista J, *et al.* (2014). Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc Natl Acad Sci USA* 111: 13924-9 (2014).
61. van de Bunt M, *et al.* (2015). 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* 11: e1005694.
62. Albrechtsen A, *et al.* (2013). Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia* 56: 298-310.
63. Tabassum R, *et al.* (2013). Genome-wide association study for type 2 diabetes in Indians identifies a new susceptibility locus at 2q21. *Diabetes* 62: 977-86.
64. Palmer ND, *et al.* (2012). A genome-wide association search for type 2 diabetes genes in African Americans. *PLoS One* 7: e29202.
65. Qi L, *et al.* (2010). Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Hum Mol Genet* 19: 2706-15.
66. Yamauchi T, *et al.* (2010). A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at *UBE2E2* and *C2CD4A-C2CD4B*. Nat Genet 42: 864-8.
67. Steinthorsdottir V, *et al.* (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* 46: 294-8.
68. Hara K, *et al.* (2014). Genome-wide association study identifies three novel loci for type 2 diabetes. *Hum Mol Genet* 23: 239-46.
69. Tsai F-J, *et al.* (2010). A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLoS Genet* 6: e1000847.
70. Li H, *et al.* (2013). A genome-wide association study identifies *GRK5* and *RASGRP1* as type 2 diabetes loci in Chinese Hans. *Diabetes* 62: 91-8.
71. Kong A, *et al.* (2009). Parental origin of sequence variants associated with complex diseases. *Nature* 462: 868-74.
72. Saxena R, *et al.* (2013). Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in Sikhs of Punjabi origin from India. *Diabetes* 62: 1746-55.
73. Shu XO, *et al.* (2010). Identification of new genetic risk variants for type 2 diabetes. *PLoS Genet* 6: e1001127.
74. Perry JRB, *et al.* (2012). Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in *LAMA1* and enrichment for risk variants in lean compared to obese cases. *PLoS Genet* 8: e1002741.
75. Cook JP, Morris AP (2016). Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *Eur J Hum Genet* 24: 1175-80.

**Figure legends**

**Figure 1. Five distinct T2D association signals attaining locus-wide significance at the *KCNQ1* locus in 27,206 cases and 57,574 controls of European ancestry.** Association summary statistics are presented from conditional analyses after adjusting for all other index SNPs at the locus. Each point represents a SNP, plotted with their conditional *p*-value (on a -log10 scale) as a function of genomic position (NCBI build 37). In each plot, the index SNP is represented by the purple symbol. The colour coding of all other variants indicates LD with the index variant in European ancestry haplotypes from the 1000 Genomes Project reference panel: red *r*2≥0.8; gold 0.6≤*r*2<0.8; green 0.4≤*r*2<0.6; cyan 0.2≤*r*2<0.4; blue *r*2<0.2; grey *r*2 unknown. The shape of the symbol corresponds to the annotation of the variant: upward triangle for frameshift, stop or splice; downward triangle for non-synonymous; square for synonymous or UTR; and circle for intronic or non-coding. Recombination rates are estimated from Phase II HapMap and gene annotations are taken from the UCSC genome browser.

**Figure 2. Trans-ethnic fine-mapping of the T2D association signal at the *CDKAL1* locus in 22,086 cases and 42,539 controls from diverse populations.** Association summary statistics are presented on the basis of: GWAS in East Asian ancestry populations and European, Hispanic and South Asian ancestry populations (top); and trans-ethnic meta-analysis of GWAS from all ancestry groups (bottom). Each point represents a variant, plotted with their log10 Bayes’ factor (BF) as a function of genomic position (NCBI Build 37). In each plot, the index SNP from the trans-ethnic meta-analysis is represented by the purple symbol. The colour coding of all other variants indicates LD with the index SNP (estimated from 1000 Genomes Project reference haplotypes by EUR *r*2 for the trans-ethnic meta-analysis and European, Hispanic and South Asian ancestry GWAS, and by ASN *r*2 for the East Asian ancestry GWAS): red *r*2≥0.8; gold 0.6≤*r*2<0.8; green 0.4≤*r*2<0.6; cyan 0.2≤*r*2<0.4; blue *r*2<0.2; grey *r*2 unknown. The shape of the plotting symbol corresponds to the annotation of the variant: upward triangle for frameshift, stop or splice; downward triangle for non-synonymous; square for synonymous or UTR; and circle for intronic or non-coding. Recombination rates are estimated from Phase II HapMap and gene annotations are taken from the UCSC genome browser. The genomic interval covered by the 99% credible set of variants for the association signal from the trans-ethnic and ethnic-specific meta-analyses are highlighted by the red region.