Type 2 diabetes: genetic data sharing to advance complex disease research

Jason Flannick[1,2] *and Jose C. Florez*[1–4]

Abstract | As with other complex diseases, unbiased association studies followed by physiological and experimental characterization have for years formed a paradigm for identifying genes or processes of relevance to type 2 diabetes mellitus (T2D). Recent large-scale common and rare variant genome-wide association studies (GWAS) suggest that substantially larger association studies are needed to identify most T2D loci in the population. To hasten clinical translation of genetic discoveries, new paradigms are also required to aid specialized investigation of nascent hypotheses. We argue for an integrated T2D knowledgebase, designed for a worldwide community to access aggregated large-scale genetic data sets, as one paradigm to catalyse convergence of these efforts.

Heritability
The proportion of phenotypic variance in a population owing to genetic differences, as opposed to environmental differences.

[1]*Programs in Metabolism and Medical and Population Genetics, Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, Massachusetts 02142, USA.*
[2]*Center for Human Genetic Research, Department of Medicine, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02062, USA.*
[3]*Diabetes Research Center, Diabetes Unit, Department of Medicine, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02062, USA.*
[4]*Department of Medicine, Harvard Medical School, 25 Shattuck Street, Boston, Massachusetts 02115, USA.*
Correspondence to J.C.F.
JCFLOREZ@mgh.harvard.edu

doi:10.1038/nrg.2016.56
Published online 11 Jul 2016

Genetic studies of human disease seek insight into disease processes or novel therapies from naturally occurring heritable variation in a population. Characterizing the degree to which disease-informative variants exist, identifying them and their effects, and understanding how they might guide preventive or therapeutic interventions are thus related but distinct goals of genetic association studies.

Type 2 diabetes mellitus (T2D) is a complex, heritable disease with a sibling relative risk of approximately 2 (REF. 1) and a heritability estimated at 30–70%[2]. For decades, analyses of natural genetic variation have been used to understand T2D mechanisms or to improve prognoses for the approximately 415 million people who suffer from its effects; these include prolonged hyperglycaemia from insulin resistance and relative insulin deficiency[3], numerous micro- and macrovascular complications[4] and an increased risk of early death[5].

Before common variant genome-wide association studies (GWAS), it was not known to what extent such 'experiments of nature' could lend insight into T2D. Early genetic mapping studies for T2D involved comparatively small-scale candidate gene and linkage studies. Although these approaches localized several disease genes (for example, *PPARG*[6], *KCNJ11* (REF. 7) and *TCF7L2* (REF. 8)), they were on the whole unsuccessful[9]. By contrast, GWAS yielded — for the first time — a substantial number of genomic loci reproducibly linked to T2D risk, suggesting previously unsuspected biological pathways[10,11]. And yet, the modest fraction of heritability attributable to GWAS associations, together with insufficient resolution to identify specific causal variants or effector transcripts, led to suggestions that the endeavour produced minimal insights[12,13].

In the era that has followed these early GWAS, studies have become much larger and more diverse, technologies assay rarer variants across the entire exome or genome, and biological experiments have begun to translate reported associations to insights into causal variants and disease processes. Here, we review the history and recent progress towards mapping genes for T2D and consequent implications for disease genetic architecture. We then review studies to understand pathogenic mechanisms that may point to new therapeutic modalities. We argue that future research into each area will be most powerful and synergistic with a new model for open and interactive sharing of data and results between previously loosely coupled communities.

Mapping disease loci

Common variant association studies. Motivated by the increased power of association studies over prior family studies[14], as well as the common disease common variant hypothesis (CDCV hypothesis)[15], GWAS exploited inexpensive single-nucleotide polymorphism (SNP) microarrays and reference catalogues of population-level variation[16] to analyse through linkage disequilibrium nearly all common variants in the genome. The first T2D GWAS, which studied a few thousand samples each[17–21], collectively identified 10 genomic loci, and many loci were reported by multiple studies. These findings validated the GWAS design but suggested that common variants of large effects were mostly absent from the population: all associated variants affected risk by less than 40%, and most affected risk by closer to 15%.

In order to identify common variants of weaker effects, collaboration and data sharing were necessary to increase power (FIG. 1). The Diabetes Genetics Replication

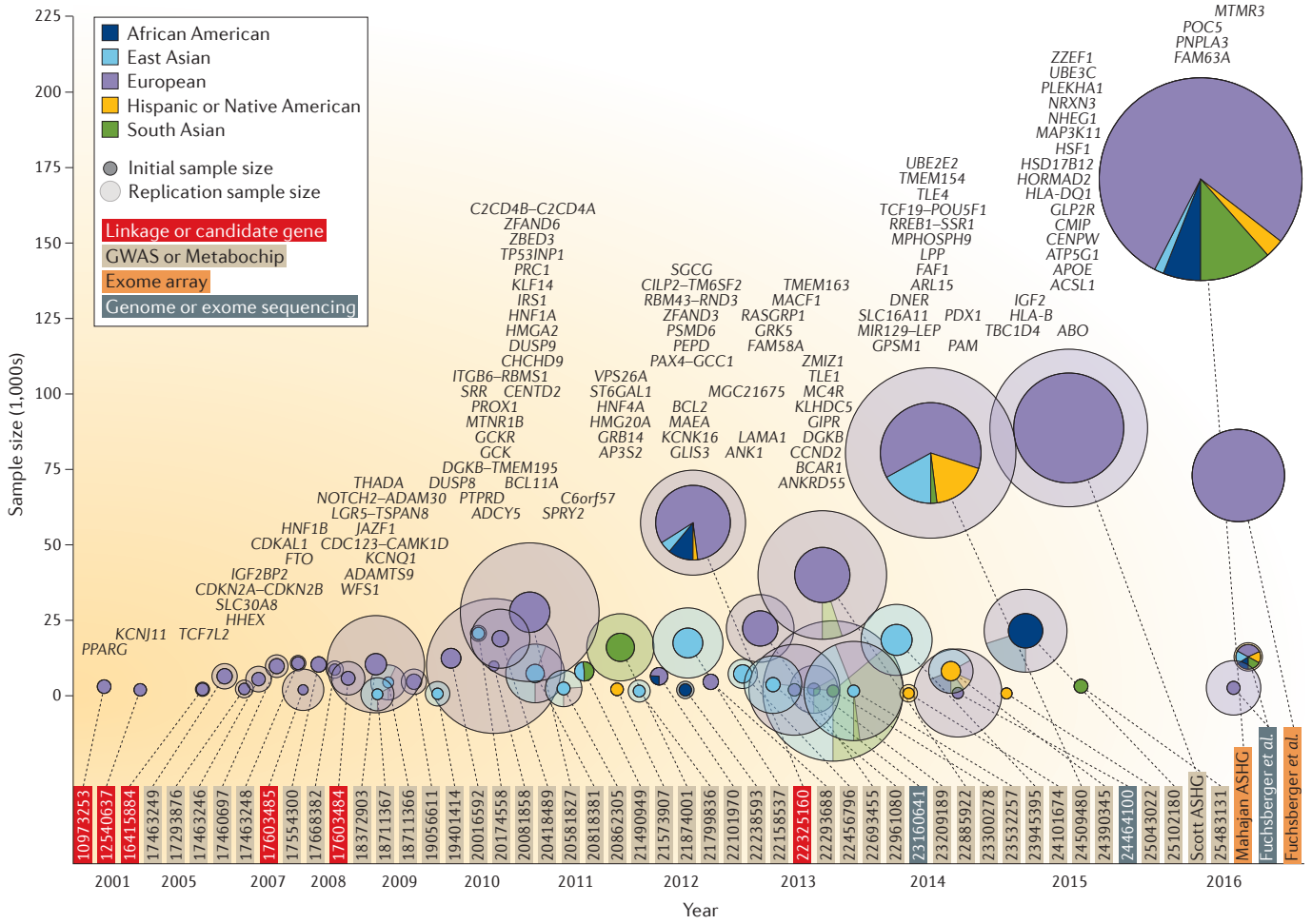

Figure 1 | **The history of T2D GWAS.** Over the years, type 2 diabetes mellitus (T2D) genome-wide association studies (GWAS), which typically consist of a two-stage discovery and replication study design, have increased in size and diversity. Plotted are circles representing T2D GWAS, as specified by the US National Human Genome Research Institute–European Bioinformatics Institute (NHGRI–EBI) GWAS catalogue[163], as well as additional candidate gene or sequencing studies of note. The *x*-axis shows the year of publication, whereas the *y*-axis shows discovery sample size. The inner (darker) circles are also scaled in proportion to discovery sample size, whereas the outer (lighter) circles are scaled in proportion to total (discovery + replication) sample size. Circles are coloured according to ethnic composition of the sample set: African American (dark blue), East Asian (light blue), European (purple), Hispanic or Native American (yellow) or South Asian (green). PubMed identifiers (or first author name) for each study are shown at the base of the figure and connected to the corresponding circle with a dotted line. Identifiers are coloured according to the technology used in the study: linkage or candidate gene studies (red), GWAS or Metabochip (beige), exome array (orange) or sequencing (dark grey). Additionally, T2D loci are listed directly above the first reporting study. With some exceptions based on retrospective knowledge, a *P* value of 5×10^{-8} is used as a threshold for significance, and loci are named as originally reported.

Genome-wide association studies
(GWAS). An approach for genetic mapping that compares frequencies of variants across the genome between disease cases and matched controls. This is a paradigm for identifying genes or biological processes that are relevant to a phenotype by identifying correlations between polymorphic genetic markers and the presence of the phenotype.

and Meta-analysis (DIAGRAM) Consortium increased sample size above 10,000 (with approximately 4,500 T2D cases), identifying a further six loci[22]. Enabled by a willingness of researchers to collaborate rather than to compete, an availability of resources and tools to harmonize data sets[23] and a means to share statistics while protecting study participants[24], GWAS meta-analysis consortia quickly became the standard for T2D and other complex traits[25,26]. The second DIAGRAM analysis included 45,000 samples (approximately 8,100 cases), identifying 12 additional T2D loci[27], whereas other consortia analysed quantitative glycaemic traits[28] and other T2D-relevant phenotypes[29–31] in either overlapping samples or more diverse ethnic groups[32–34]. The similar goals, organization and methodology of these consortia, as well as of those investigating other metabolic traits, led to the joint design of a custom genotyping array[35] that facilitated analysis of still larger sample sizes. A T2D association study used this array to analyse almost 150,000 individuals (approximately 34,800 cases) and identified ten further loci[36], whereas a glycaemic trait association study with over 133,000 individuals led to a similar number of discoveries[37]. GWAS thus catalysed a clear breakthrough in the identification of genomic loci reproducibly associated with disease.

Causal variants
Specific mutations underlying the molecular cascade that produces a phenotypic trait; by design, in most genetic mapping studies, the associated genetic marker is merely correlated with the underlying causal variant.

Effector transcripts
The specific RNA transcript (for example, mRNA transcribed from a gene) for which the function or expression is altered by the causal variant, leading to a phenotypic difference.

Genetic architecture
The number, frequencies and effects on disease of genetic variants in a population.

Common disease common variant hypothesis
(CDCV hypothesis). The hypothesis that, owing to historical human population growth, some disease loci for common diseases may harbour alleles common in the population.

Linkage disequilibrium
Correlations among nearby variants, owing to historical patterns of demography and recombination, exploited by genome-wide association studies to map common variant associations.

Glycaemic
Traits pertaining to the physiology of blood glucose regulation, usually involving measures of glucose, insulin or other related hormones.

Rare variant models
A model of genetic architecture in which rare variants (for example, those with a frequency < 1 %) explain most of the heritability.

Synthetic associations
A hypothesis based on simulations that multiple causal rare variants of strong effects might cause a common variant statistical association.

Common variant models
Models of genetic architecture in which common variants (for example, those with a frequency > 1 %) explain most of the heritability.

Models for the genetic architecture of T2D. Despite the successes described above, the use of GWAS findings was debated, as the identified associations explained less than 10% of T2D heritability[38]. Whereas GWAS proponents cited previously unsuspected disease processes implicated by newly identified associations[39], counter-arguments stated that modest-effect common variants offered little value towards understanding or predicting disease[13] and might even be misleading if caused, in fact, by distant rare variants[40] or population artefacts[12]. Using qualitative arguments[12,41] or simulations[40,42], many argued for future studies of rare variants, which were expected from evolutionary arguments to have much larger effects on disease risk[12,41] and potentially to explain much of disease heritability[38].

However, closer analysis of GWAS results revealed that evidence for rare variant models was far from conclusive. Simulations suggesting that rare variants (or synthetic associations) might explain many GWAS signals[40,43] were countered by empirical examples inconsistent with that model. These included an absence of linkage signals[44], a common-shifted distribution of GWAS variant frequencies[45] and replicated associations across populations[46]. Furthermore, new methods for estimating heritability from all variants analysed in a GWAS, rather than only those reaching significance, suggested that common variants might explain more than half of T2D heritability, leaving hundreds or thousands of smaller-effect associations to be detected[36,47].

This collection of analyses, examining similar data, thus offered contrasting models for the genetic architecture of T2D. In fact, a comprehensive simulation-based analysis suggested that, before 2012, T2D genetic studies and epidemiological parameters were consistent with either rare variant or common variant models[48]. Adjudicating between them required higher resolution studies to interrogate variants across the entire frequency spectrum.

Rare variant association studies. The first study to investigate the contribution of lower-frequency (<5%) variants to T2D or related traits appeared in 2012, when coding variant analysis in 8,229 individuals (using the Illumina exome array) identified five low-frequency variants associated with glycaemic traits (three at previously unidentified loci)[49]. Shortly thereafter, an exome sequencing study in 2,000 Danes (with 1,000 cases) identified two T2D-associated common coding variants at previously unidentified loci, although no low-frequency variants were identified[50]. In 2014, a genome sequencing study of 2,630 Icelanders, followed by statistical imputation of variants into 278,554 additional Icelanders, identified three more T2D-associated low-frequency variants[51], all within loci that had previously been reported for T2D or related traits. Together with several studies that followed for T2D[52], glucose[53,54] or insulin levels[55], these reports demonstrated the ability of exome- or genome-wide analyses to identify lower-frequency variants relevant to T2D.

In parallel, candidate gene experiments were also successful, in limited contexts, at identifying broader allelic series within genes initially implicated through common variants in T2D. An analysis of rare variants in *MTNR1B* demonstrated a strong effect of functional variants (according to a series of molecular or cellular assays) on T2D risk; non-functional variants had comparatively weaker effects[56]. A similar pattern was observed for rare variants in *PPARG*[57]. In *SLC30A8*, rare variants predicted to cause protein truncation were associated with protection from T2D[58], which contrasted with expectations from previous animal[59] or cellular work[60]. The contribution of these variants to T2D heritability, however, was small.

Deciphering genetic architecture through larger-scale sequencing studies. Although these newly reported associations were valuable, the extent to which they supported rare or common variant genetic architectural models for T2D was left unanswered. Analysis of the 2,000 Danish exomes rejected extreme models, with T2D explained by a small number (<20) of large-effect, low-frequency, non-synonymous variants[61], but the power was insufficient for further conclusions. To characterize T2D genetic architecture to a much higher resolution, a more comprehensive set of sequencing studies was designed, spanning 12,940 multi-ethnic exome sequences (6,504 cases) with low-frequency variants genotyped in 79,854 additional individuals (28,305 cases), as well as 2,657 whole-genome sequences (1,326 cases) with all variants imputed into 44,414 additional individuals (12,971 cases)[62]. Strikingly, only a single low-frequency variant, which had been previously reported in the Icelandic sequencing study, was associated with T2D at genome-wide significance, despite high power to detect low-frequency variants of even moderate effect. These results suggested that low-frequency variation — and coding variation of any frequency — has, at most, limited roles in the genetic architecture of T2D. Consequently, larger common-variant GWAS would probably continue to identify additional T2D loci, whereas rare variant association studies would require comparatively much larger sample sizes[48,63,64].

Indeed, larger GWAS meta-analyses have continued to expand the number of T2D-associated loci. A 2014 *trans*-ethnic meta-analysis of more than 110,000 individuals (approximately 26,500 cases), which was motivated by a consistency of common variant associations observed across different populations[46], identified seven new loci[65], whereas analysis of almost 160,000 European individuals (approximately 26,700 cases) subsequently identified 18 more[66]. GWAS of previously unstudied populations, such as Mexicans[67], African Americans[68] or the Inuit[69], have also yielded new associations, in many cases because different population histories have led to different allelic spectra and consequent gains in power. Conversely, even in larger samples (that is, >232,000 individuals, approximately 56,600 cases), coding variant analyses have identified few associations with low-frequency variants[70]. These findings, together with new methods to assess polygenicity from GWAS[71] and excess concordance observed for even nonsignificant common variant associations across ancestries[65], continue to support a common variant model for the larger fraction of T2D genetic architecture.

Box 1 | **Population risk versus biologically relevant variants and genes**

Genetic studies use 'experiments of nature' — that is, naturally occurring variation — to characterize both the genetic basis of disease in the population (on which stratification for intervention might be predicated) as well as genes or biological pathways that might guide insight into disease processes. These two goals are in many senses complementary: identifying most variants responsible for disease provides a catalogue of disease-relevant genes, and variants of disproportionate impact in the population may point to pathways of disproportionate relevance to disease. However, insight into disease biology is not predicated on a complete description of the genetic basis of disease, nor does explanation of heritability in the population necessarily lead to improved biological understanding.

These distinct goals thus lie at the heart of one notable debate in recent years: the degree to which genome-wide association studies (GWAS) have been 'successful'. One interpretation of the first GWAS findings for type 2 diabetes mellitus (T2D) and other complex diseases was a focus on 'missing heritability': that is, the inability of GWAS findings to explain most of the genetic risk of disease in the population or to provide sufficient prognostic or stratifying information[13,38]. The modest effects of common variants, responsible for the limited fraction of heritability explained, were claimed to provide limited insight into disease biology, and rarer variants were claimed to offer a more profitable explanation for the remaining genetic basis of disease[13].

GWAS, however, were not designed to explain the entirety of disease heritability but rather to identify the subset of variants that by chance may have reached common frequencies in the population[15]. The identification of some — let alone many — such variants offered key insights into T2D biology, such as the confirmation that most risk variants act through reduced β-cell function or mass, rather than many other previously hypothesized pathways[27], and the demonstration that most molecular mechanisms of disease risk are regulatory[110,113,162]. Furthermore, the effects of genetic variants are constrained by population history and natural selection, suggesting that stronger perturbations of the same or different genes in a common pathway might lead to larger phenotypic effects. In this sense, rare variants can be of great value, even if they in fact explain less heritability than common variants: for example, the finding of protective rare loss-of-function variants in *SLC30A8* (REF. 58), which explained less heritability than the common variant by several orders of magnitude[36], nonetheless provided an important suggestion about the direction of disease risk from protein inactivation. Similarly, *KCNJ11–ABCC8* and *PPARG* exemplify viable drug targets that could, in principle, have initially been identified by modest-effect common-variant associations[6,7].

Conversely, recent studies that have succeeded in explaining missing heritability have not necessarily led to increased biological insight. Polygenic score[47] and mixed linear modelling[36] analyses have shown that GWAS variants in fact 'tag' most T2D heritability, but these analyses by design do not identify specific variants from which biological insight might be derived. Furthermore, an explanation of disease heritability does not imply an accurate model for disease risk prediction[97,98], as the environmental component of T2D risk introduces substantial and inherent noise. Similarly, characterizing the genetic architecture of T2D — or the number, frequencies and effect sizes of variants that contribute to heritability — is, in some senses, easier than identifying specific causal variants and can greatly affect genetic study designs or genetic risk prediction.

Thus, the two goals of explaining disease heritability and identifying biologically relevant genes require, in many cases, distinct approaches by distinct communities. Diverse analyses of shared genetic data sets are one future route to empower both types of study.

Imputation
A technique to infer the unknown genotype of a variant in an individual based on correlations with nearby genotyped variants.

Allelic series
A number of alleles of a gene or locus with a range of phenotypic and/or molecular effects that are of use to infer a genetic–phenotypic dose–response curve.

Polygenicity
An idealized model in which a phenotype is caused by a large number of variants, each with small and normally distributed phenotypic effects.

Transcriptomic
The study of the expression levels of all transcripts in a cell.

Epigenomic
The study of all epigenetic modifications of a cell, including DNA methylation and histone modifications, which are largely responsible for the genes expressed in a specific tissue at a given developmental stage or metabolic state.

Homeostasis model assessments
A method based on fasting measures of glucose and insulin levels that is used to estimate β-cell function or insulin resistance.

Thus, additional T2D loci will probably be discovered mostly by continued increases in the size and diversity of common-variant GWAS, incorporating new populations, additional phenotypes[72,73] and new resources for genotype imputation[74,75]. The collaborative networks and centralized analysis plans of GWAS consortia, refined over the past half-decade, will probably play a key part in the organization of these efforts.

Understanding disease biology

Characterizing the alleles that contribute to population disease risk is only one goal of genetic studies (BOX 1). Arguably more consequential is using the mapped associations to develop and test hypotheses about disease biology. Progress towards this latter goal has been variable, in part because experiments to understand the molecular, cellular and physiological mechanisms behind an association are typically highly domain-specific and thus are less systematized than the analysis protocols now deployed in GWAS. Moreover, the opportunity to use genetic findings to validate or disprove hypotheses from animal or cellular models is often impeded by the cumbersome access to results and opaque language of genetic studies.

Recent biological and clinical studies increasingly suggest one possible means to increase the translational use of human genetic findings: through convergence on common resources and workflows (FIG. 2; TABLE 1). Paradigms have begun to emerge for investigations of physiological mechanisms through deep patient phenotyping and stratification; for the analysis of potential causal variants and effector transcripts through the use of genetic, transcriptomic and epigenomic maps; and for the elucidation of molecular or cellular mechanisms through functional experiments, new techniques and genomic resources.

Physiological investigations. To gain physiological insight into T2D associations, many studies have investigated the many T2D-related phenotypes that are available in large numbers of genotyped samples. Analysing 31 T2D-associated variants for associations with homeostasis model assessments of β-cell function (HOMA-B) or insulin resistance (HOMA-IR) produced early evidence that most T2D variants increase risk through pancreatic β-cell dysfunction[27]. Ensuing similar approaches have defined finer-grained clusters of T2D-associated variants[76],

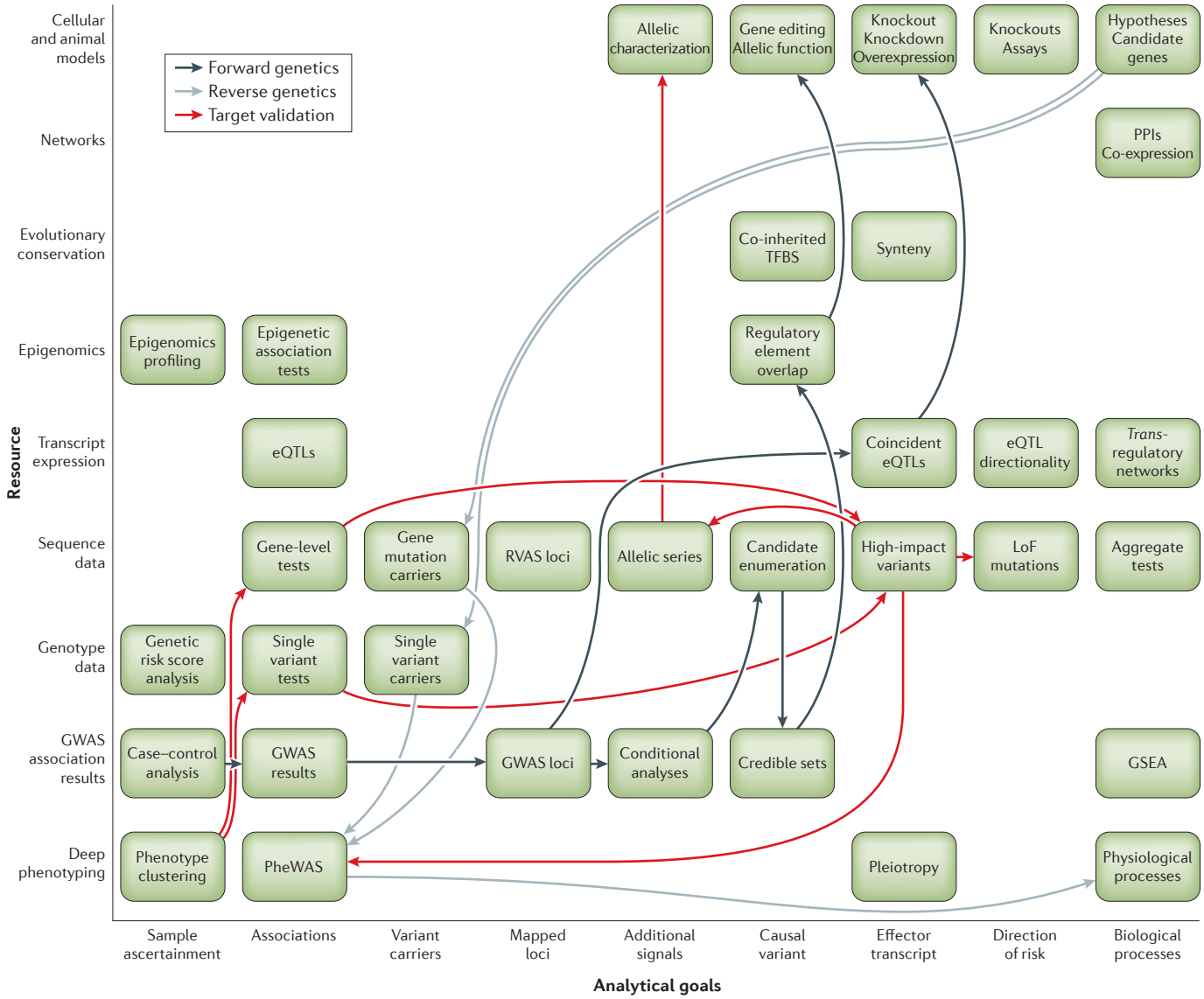

Figure 2 | **Common resources, analyses and workflows for understanding T2D biology.** Although investigations of biological mechanisms for type 2 diabetes mellitus (T2D) are typically highly context-dependent, in recent years common approaches have emerged. Example goals or lines of investigation (*x*-axis) include ascertaining samples for genetic analysis (sample ascertainment), computing genotype–phenotype associations (associations), identifying genetic variant carriers (variant carriers), mapping T2D loci (mapped loci), identifying additional signals at mapped loci (additional signals), identifying the causal variant for an association signal (causal variant), identifying the transcript that mediates an association (effector transcript), elucidating the directional relationship between molecular activity and disease risk (direction of risk) and understanding or identifying T2D-relevant biological processes (biological processes). Example resources (*y*-axis) include deep human phenotypic measurements (deep phenotyping), catalogues of associations reported from genome-wide association studies (GWAS association results), genotypes at select variants (genotype data), full sequence data (sequence data), catalogues of transcript expression (transcript expression), catalogues of epigenomic marks (epigenomics), measurements of cross-species conservation (evolutionary conservation), networks of gene–gene or protein–protein interactions (PPIs) or associations (networks) and experimental results from model systems (cellular and animal models). Analyses that use a class of resource to address a specific question are shown as boxes; examples of these analyses are described in TABLE 1. Three common workflows, or series of analyses, are indicated by sets of arrows between boxes. An example 'forward genetics' workflow (dark grey) carries out a GWAS of T2D cases and matched controls, identifies new loci and then predicts causal variants (by fine mapping and mechanistic prediction from epigenomic annotations) or effector transcripts (by analysis of expression quantitative trait loci (eQTLs)). An example 'reverse genetics' workflow (light grey) investigates the function of a candidate gene by identifying carriers of high-impact variants and analysing their phenotypes. An example 'target validation' workflow (red) identifies a potential target through association analysis for a specific disease subtype (by stratifying individuals based on a range of phenotypic criteria) and then uses additional genetic and experimental analyses to inform the directional relationship between molecular and phenotypic effect, a therapeutic 'dose–response' curve and effects of target perturbation on a range of deeper phenotypes. GSEA, genome set enrichment analysis; LoF, loss of function; PheWAS, phenome-wide association study; RVAS, rare variant association study; TFBS, transcription factor binding site.

Table 1 | **Example analyses for understanding T2D biology**

Goal	Analysis	Description/rationale	Examples
Sample ascertainment	Phenotype clustering	Stratification of samples to identify variants for disease subtypes	Ascertainment based on T2D-related phenotypes[89] or high-dimensional EMRs[91]
	Case–control analysis	Classic association analysis	Most GWAS[65] or sequencing studies[62]
	Genetic risk score analysis	Measure aggregate effects of all variants carried by a patient	Weighted scores from GWAS variants for T2D[97] or related traits[78,98]
	Epigenomics profiling	Use epigenomics rather than genetics to stratify patients	Prediction based on blood methylation levels[96]
Associations	PheWAS	Test for association with all available phenotypes for a given variant	Associations with T2D-related traits[27,104]; glucose tolerance tests[164]; fasting versus 2-hour post-OGTT[69]; glucose monitoring[165]; insulin sensitivity and signalling[87]; body mass distribution[78]
	GWAS results	Standard output of a GWAS	Variants reaching genome-wide significance in T2D GWAS[65]
	Single-variant tests	More nuanced association tests not typical of a standard GWAS	Replication of preliminary single-variant associations[51,58]; custom analyses[89]; custom conditional analyses[62,90]; recessive models[69]
	Gene-level tests	Aggregate association analysis, commonly used in RVAS	Aggregate LoF tests[58]; tests of functional variants[56,57]
	eQTLs	Associations with transcript levels	eQTL analysis of islet RNA-seq data[119,120]
	Epigenetic association tests	Associations with epigenomic differences	Methylation differences between cases and controls[96,122,124,125] or monozygotic twins[124,125]
Variant carriers	Single-variant carriers	Carriers of a specific variant	Identification of low-frequency variant[51,104] or homozygous variant carriers[69]
	Gene mutation carriers	Carriers of any variant meeting specified criteria within a gene	Diagnostic mutation screening[87,165]
Mapped loci	GWAS loci	The loci identified by GWAS	Reported from common variant GWAS[27,36,65]
	RVAS loci	The genes identified by RVAS	Reported from RVAS[51]
Additional signals	Conditional analysis	Used by GWAS or fine mapping to identify multiple signals at a locus	Conditional analysis on GWAS signals[62,70,102]
	Allelic series	Identify variants of a range of effects	Aggregate rare variant analysis for known genes[56–58] or gene sets[62,99]
	Allelic characterization	Functionally validate and calibrate the effects of an allelic series	Missense variant screening with functional assays[56,57]
Causal variant	Credible sets	Localize common variant association to individual variant	Credible set analysis of European[105] or multi-ethnic samples[65], using dense genotyping[102] or sequencing[62]
	Candidate enumeration	Use sequence data to ensure all variants analysed for credible set	Candidates from 1000 Genomes Project[74] or large-scale sequencing[62]
	Regulatory element overlap	Prioritize variants based on predicted regulatory effect	Enrichment of epigenomic annotations[102,110,111,113–116] or causal variant prioritization[102,115,127,129,135,136]
	Co-inherited TFBS	Novel method to use evolutionary patterns to predict causal variant	Phylogenetic conservation of co-occurring TFBS[138]
	Gene editing, allelic function	Verify variant effect on regulatory landscape or expression	Allele-specific reporter assays[102,113,115,116,129,135,136] or EMSA[102,113,129,135,136]; CRISPR[127,138]
Effector transcript	Pleiotropy	Effects on multiple traits can hypothesize effector transcript	Multiple variants near *RREB1* associated with related traits[62]
	High-impact variants	Elevated prior likelihood of causality	Missense variant associations in *TM6SF2* (REF. 62), *PPARG*[6] or *SLC30A8* (REF. 19)
	Coincident eQTLs	Technique used to suggest regulatory effect of association	Association coincident with eQTL from public data sets[19], custom data sets[51,126,127,131] or islet RNA-seq[120]; allelic expression profiling[128,129]
	Synteny	Co-inheritance of variant and gene	Conserved genomic regulatory blocks[130]
	Knockout, knockdown, overexpression	Experimentally measure effects of gene perturbation	Gene overexpression or knockdown in human cell[126,127] or mouse models[127]
Direction of risk	LoF mutations	Observe 'human knockouts'	Effect of gene inactivation[58]
	eQTL directionality	Altered regulation in humans	Direction of expression change[51,120]
	Knockouts, assays	Effect of knockout in model systems	Inference from cellular[126,127] or animal phenotypes[127]

Table 1 cont. | **Example analyses for understanding T2D biology**

Goal	Analysis	Description/rationale	Examples
Biological processes	Physiological processes	Understanding of physiology from patient investigation	Insight from carrier phenotypes[69,78,87,164,165]
	GSEA	Pathways enriched for GWAS variants	Commonly reported in GWAS[27,36]
	Aggregate tests	Tests of variants across gene set	Aggregate variant effects in monogenic genes[62,99]
	Regulatory networks	Infer networks from variant perturbations of expression patterns	*Trans*-eQTL analyses[133] or enrichment[132]
	Co-expression, PPIs	Use interactome to organize variants	Networks to rank genes[146] or test hypotheses[145]
	Hypotheses, candidate genes	Classic functional studies and starting point for 'reverse genetics'	Experiments based on biological hypotheses[142–144]

Listed are examples of commonplace analyses, depicted in FIG. 1, to gain insight into T2D biology. For each line of investigation (first column) and analysis (second column) depicted in FIG. 1, listed is a more specific description of the analysis (third column) and examples, not intended as comprehensive, from the literature (fourth column). EMR, electronic medical record; EMSA, electrophoretic mobility shift assay; eQTL, expression quantitative trait loci; GSEA, gene set enrichment analysis; GWAS, genome-wide association study; LoF, loss of function; OGTT, oral glucose tolerance test; PheWAS, phenome-wide association study; PPI, protein–protein interaction; RNA-seq, RNA sequencing; RVAS, rare variant association study; TFBS, transcription factor binding site; T2D, type 2 diabetes mellitus.

classified glycaemic trait associations[77,78] and assessed cellular phenotypes[79]. Recently, Mendelian randomization[80] has gained favour as an approach to evaluate causal relationships between T2D-relevant endophenotypes or biomarkers, supporting causality between body mass index (BMI)[81] or bilirubin levels[82] and T2D, but rejecting causal relationships between triglyceride[83], high-density lipoprotein (HDL) cholesterol[84], adiponectin[85] or uric acid[86] levels and T2D.

For some genes, deeper phenotyping of variant carriers has led to quite specific mechanistic insights. For example, insulin resistance measures and muscle and fat biopsy samples of *PTEN* mutation carriers demonstrated a link with T2D via enhanced phosphatidylinositol 3-kinase (PI3K)–protein kinase B (PKB; also known as AKT) pathway signalling and hence increased insulin sensitivity[87]. Similarly, in the study that identified a common but large-effect Inuit-specific T2D variant in *TBC1D4*, access to measures of fasting and 2-hour post-oral glucose tolerance test (OGTT) glucose and insulin showed an association with glucose-stimulated insulin secretion but not with fasting levels of glucose[69]. This finding is consistent with mouse models and a proposed mechanism of action related to glucose uptake by skeletal muscle in the post-prandial setting[88].

These distinct phenotypic spectra support a perspective widely held by clinicians, who contend with the heterogeneity of T2D on a daily basis, of the existence of distinct T2D subtypes. Indeed, stratification of patients into more homogeneous subgroups has produced novel associations, such as near *LAMA1*, through analysis of lean T2D cases[89] or additional associations with insulin after controlling for BMI[90]. The striking difference in association strength between homozygous *TBC1D4* variants and T2D as defined by glycated haemoglobin levels ($P = 0.008$) versus World Health Organization standards ($P = 1.6 \times 10^{-24}$) further emphasizes how phenotype definition can affect association results[69]. In addition, attempts to formalize specific disease clusters have begun to yield fruit[91], while intriguing recent studies have suggested novel markers, such as somatic clonal mosaic events (CMEs)[92], the gut microbiome[93], metabolite levels[94,95] or epigenomic marks[96], to predict or classify T2D subtypes. The clinical benefit of genetic scores to predict diabetes risk, however, remains modest at best, not only for T2D[97,98], but also for monogenic forms of diabetes mellitus[99]. By contrast, genetic risk scores for type 1 diabetes, in which the *HLA* region explains a substantial portion of heritability, have proved useful to distinguish type 1 from monogenic diabetes[100] or even T2D[101] in specific clinical scenarios.

Prediction of causal variants and effector transcripts. To link predicted physiological mechanisms to causal variants or effector transcripts that mediate a T2D association, the first step is typically to fine-map the locus in large sample sizes. Fine mapping starts with conditional analysis to determine the number of independent signals at the locus. Some loci, in fact, harbour a substantial number of signals (for example, five at the *KCNQ1* locus)[102], which can collectively discern independent risk and protective signals (at the 9p21 locus)[103] or resolve directional relationships between different phenotypes (such as T2D and glucose at the *G6PC2* locus)[53]. Additional signals can also implicate effector transcripts[51] or highlight new phenotypic associations[51,104], both of which are exemplified by a second association at the *CCND2* locus[51].

For each signal, fine-mapping approaches then construct credible sets[105] that quantify the likelihood of causality for each variant. Large-scale analyses have, in some cases, identified a single causal candidate with near certainty, such as at the *CDKN2A–CDKN2B*[65] or *MTNR1B*[102] loci. In addition, the transferability of associations between ethnic groups[46,65,68] can enable multiethnic fine mapping studies to increase credible set resolution through differential linkage disequilibrium patterns: for example, at the *JAZF1* locus[65]. Large sample sizes, as well as comprehensive catalogues of variation from efforts such as the 1000 Genomes Project[74], are crucial to the success of fine mapping[62].

Among the variants that are identified by fine mapping, perhaps the most valuable are coding variants that have a clear molecular impact on protein function.

Mendelian randomization
A technique that uses genetic variation to infer causal relationships between correlated phenotypes.

Fine mapping
An approach to localize common variant association signals to potentially causal variants, using exhaustive candidate enumeration and genotyping in large case–control samples.

Common coding variants in linkage disequilibrium with T2D associations have an elevated prior likelihood of causality and facilitate hypotheses about effector transcripts, such as for *RREB1* at the *RREB1–SSR1* locus and *TM6SF2* at the *CILP2–TM6SF2* locus[62]. An allelic series of coding variants can provide further insights into gene function. For SLC30A8, the identification of protective protein-truncating variants hypothesized an unsuspected relationship between decreased, rather than increased, activity and T2D protection[58]. In genes for monogenic forms of diabetes[62,99], rare risk-increasing coding variants suggested mechanistic links between T2D and comparatively well-understood Mendelian disease processes[62,101]. For *MTNR1B*[56] or *PPARG*[57], the ability to functionally discriminate T2D-associated from neutral variants provided a potential foundation to develop disease-relevant assays for therapeutic development[106].

Resources for prioritization of causal non-coding variants. Credible sets for most T2D associations do not include coding variants. Instead, they span non-coding regions of the genome and suggest causal variant regulatory effects. Although early fine-mapping studies[103,105] had limited access to noncoding functional annotations, efforts such as the ENCODE[107], Epigenomics Roadmap[108] and GTeX[109] projects have since provided reference regulatory landscape maps of numerous human tissues. Indeed, GWAS variants have since been shown to cluster within regulatory elements[110,111], and T2D-associated variants show the strongest localization to islet enhancers[112,113]. More recent studies have also shown enrichment of T2D-associated variants within transcription-factor binding sites, most notably for forkhead box A2 (FOXA2) in islets[102] and PPARγ in fat[114].

To further capture the range of developmental stages and metabolic states of specific relevance to T2D pathophysiology, numerous smaller-scale efforts have increasingly characterized the pancreatic islet regulatory landscape along multiple axes. Early formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE–seq), DNase sequencing (DNase–seq) and chromatin immunoprecipitation followed by sequencing (ChIP–seq) studies identified thousands of islet-selective open chromatin regions[115] and regulatory elements[116]. RNA sequencing (RNA–seq) transcriptome maps have since catalogued long non-coding RNAs (lncRNAs)[117], microRNAs[118], and expression quantitative trait loci (eQTLs)[119,120] that are specific to pancreatic islets or β-cells[121]. Islet methylation profiling has been carried out in patients with T2D[122,123] and monozygotic twin pairs[124,125]. Progress has also been made towards integrated catalogues of these resources, identifying regulatory element clusters that are of high relevance to islet function[112,113].

Experimental investigation of molecular and cellular risk mechanisms. These new epigenomic and transcriptomic resources have helped to guide functional experiments for many T2D GWAS loci. To support hypothesized effector transcripts, genotype-dependent expression can provide key evidence, as exemplified by studies of the *ADCY5* (REF. 126) and *FTO*[127] loci. T2D GWAS thus routinely investigate whether new signals coincide with *cis*-eQTLs[27,36,65]: these analyses are limited by the tissues available in public data sets but can be successful, for example at the *KLF14* (REF. 27) locus. Analyses of eQTLs in custom data sets, such as pancreatic islets, have shown further success, predicting and experimentally validating *ZMIZ1* as a gene involved in glucose homeostasis[120].

Alternatives have been used to address limitations of *cis*-eQTL mapping, such as potential environmental confounders or restrictions to transcripts near the original signal. Allelic expression profiling is less frequently subject to confounders and has been used in some studies[128], for example, to implicate *ARAP1* (REF. 129) as an effector transcript (although more recent studies of the locus provide evidence in favour of *STARD10* (REF. 120)). Mapping of chromatin interactions has been used to identify longer-range eQTL relationships, exemplified by implication of the megabase-distant *IRX3* transcript at the *FTO* locus[127,130,131]. Finally, *trans*-eQTLs can be a valuable resource to investigate global regulatory effects of T2D-associated variants[132], showing, for example, that *KLF14* affects multiple metabolic traits as a master *trans*-regulator of adipose gene expression[133].

In analogy to transcriptomic maps hypothesizing effector transcripts, epigenomic annotations have proven valuable in hypothesizing causal variants. An early example was a variant near *TCF7L2*, which was shown to lie within islet-selective open chromatin[115] and later an islet-specific enhancer[113]. This variant suggested that allele-specific reporter assays could be used in β-cell lines[115,116] to confirm a previously suggested[134] β-cell-specific risk mechanism. Variants within predicted islet enhancers at the *ZFAND3* (REF. 113) and *MTNR1B*[102] loci have also been verified to have an impact on enhancer activity; in both cases, islet transcription factor binding site (TFBS) maps motivated further experiments to verify a molecular mechanism of decreased neurogenic differentiation factor 1 (NEUROD1) binding[102,113]. Similarly, at the *JAZF1* (REF. 135), *CDC123–CAMK1D*[136], *ARAP1* (REF. 129) and *FTO*[127] loci, maps of open chromatin prioritized variants subsequently tested for enhancer activity, allele-specific expression and transcription factor binding.

Other functional experiments show the increasingly diverse genomic resources used in mechanistic characterization. Imprinting patterns led to predictions of effector transcripts at the *KCNQ1* locus[137] and potential explanations for seemingly conflicting phenotype associations at the *GRB10* locus[73]. Patterns of cross-species conservation predicted *IRX3* as the effector transcript at the *FTO* locus[130] and a regulatory variant as causal at the *PPARG* locus[138]. Elucidation of a molecular, cellular and physiological mechanism that underlies obesity risk at the *FTO* locus[127] demonstrates to date perhaps the most comprehensive use of genomic information in a functional setting: epigenomic annotations guided enhancer assays to suggest adipocyte precursors as the affected cell type; long-range chromatin interactions guided eQTL experiments to predict *IRX3* as the effector transcript;

Protein-truncating variants
Variants, such as nonsense, frameshift, readthrough or splice site mutations, that lead to incomplete protein sequences and possibly non-functional proteins.

Expression quantitative trait loci
(eQTLs). Associations between a genetic marker and expression levels of a transcript.

cis-eQTLs
Expression quantitative trait loci (eQTLs) on the same chromosome and typically near the location of the gene that encodes the associated transcript.

trans-eQTLs
Expression quantitative trait loci (eQTLs) in a different chromosome from the gene encoding the associated transcript.

Table 2 | **Potential users of an integrated T2D knowledge base**

	Statistical geneticists	Biologists (informed by genetics)	Biologists (uninformed by genetics)	Pharmaceutical researchers	Clinicians
Major goals	• Identify novel associations • Explain heritability	Elucidate molecular, cellular and physiological mechanisms of association	Examine human genetic support for hypothesis from model system or pathway	• Identify potential novel targets • Obtain support for targets from human genetics	Estimate phenotypic effect for variant of interest
Relevant data sets	• Large collections of genotype and sequence data • Basic phenotypes	• Association catalogue • Credible set results • Epigenomic and transcriptomic data sets	• Coding variation • Annotation of functional impact • Rich phenotypes	• GWAS associations • Coding variation • Annotations • Rich phenotypes	• Large sequence datasets • Rich phenotypes
Typical workflow	Carry out novel association analysis on entire body of data	Examine overlap of functional annotations with variants in credible set	Identify high-impact variants in gene of interest and test for association with range of phenotypes	• Predict causal gene, directionality from GWAS signal • Assess phenotypic associations for LoF variants in gene of interest	Obtain variant frequencies in individuals with various phenotypes
Preferred interface	• Programmatic APIs • Exploratory 'sandbox'	Visualizations or genome browser within a portal	Query builders and analysis modules within a portal	• Portal with support for workflows • Programmatic APIs	• Simple portal • Programmatic APIs
Knowledge of genetics	Expert	Intermediate	Basic or none	Basic or intermediate	Basic
Needed curated content	Description of methods applied to data and results	• Qualitative interpretation of statistics • Potential caveats • Knowledge relevant to gene or pathway	• Guide for workflow to run • Relevant statistics and their interpretation • Potential caveats	• Summary of validated genetic results • Qualitative interpretation of statistics	Documentation of ascertainment protocol and phenotypic measurements
Type of finding made possible	LoF variants in *SLC30A8* (REF. 58)	Causal variants, effector transcripts or mechanisms at *MTNR1B*[102] or *ZMIZ1* (REF. 120) loci	Associations with T2D for mutations in *LIN28–let-7* (REF. 143) pathway or PPARγ binding sites[114]	Protective LoF variants in *SLC30A8* (REF. 58), assay calibration with *PPARG* variants[57], deep phenotyping of *PTEN* variant carriers[87]	Investigation of phenotypic effects of variants in monogenic diabetes genes[99]

Many communities might benefit from analyses that can be carried out on integrated genomic data sets. The table shows five potential user groups that might find use from a type 2 diabetes mellitus (T2D) knowledge base, together with their major analytical goals, data sets of high relevance, an example workflow that might be supported, the preferred interface to the knowledge base, the expected level of genetics expertise, the curated content that might be necessary to augment the analysis and the type of finding that a knowledge base might facilitate in the future. The groups listed are statistical geneticists who typically carry out genome-wide association studies (GWAS) or complex genetic association analyses, biologists who aim to translate genetic associations into mechanistic insight, biologists who investigate hypotheses from animal or cellular models, pharmaceutical researchers who seek to identify or prioritize targets based on human genetics, and clinicians who seek to interpret genetic variants identified in patient genome sequences. Both the examples and classification of users are simplistic and not intended to be comprehensive. Furthermore, a typical researcher may in reality move between different user groups depending on the line of research. API, application programming interface; LoF, loss of function; PPARγ, peroxisome proliferator activated receptor-γ.

co-expression and *trans*-eQTL analysis guided cellular phenotyping, gene perturbation and mouse models to predict adipocyte browning as a cellular mechanism; and cross-species conservation of TFBS patterns guided CRISPR–Cas9 editing to predict the causal variant.

These recent experiments thus augment more traditional functional investigations of genetic associations, such as using gene knockdown to validate *ATM* in metformin response[29], mouse models to investigate *CDKAL1* (REF. 139), mRNA expression analysis to study the tissue of action for *SLC2A2* (REF. 140), or biochemical assays to study the function of *GCKR*[141]. Conversely, genetic associations and genomic resources can also be used to test predictions that were originally made by functional studies; for example, a prediction from mouse adipose tissue that variants in PPARγ-binding DNA elements account for strain-specific expression patterns and drug responses was tested through enrichment of metabolic trait associations within human PPARγ-binding elements[114]. Other predictions from mice, such as the relevance of *Rfx6* (REF. 142), the *Lin28–let-7* pathway[143] and *Imp2* (REF. 144) to glycaemic traits or diabetes, could also be investigated through phenotypic analysis of human mutation carriers. The same can be said for predictions from functional genomic approaches, such as interactome analysis[145,146] or knockout screens from new assays[147].

A need for a public T2D genetics knowledge base

The state of genetic mapping and functional studies for T2D highlight two different but complementary trends. The need for larger consortium-led GWAS and sequencing studies will produce centralized analyses of large genomic data sets. The need for increasingly diverse and specialized approaches to investigate these findings will produce a need for common genomic resources. And yet, today, the value of genetic data sets is only partially realized. For example, protective loss-of-function mutations in *SLC30A8* were undetected in a consortium-led genome-wide analysis of 13,000 exomes[62] and only identified after years of focused genetic and functional follow up[58].

For the biological community to make fuller and more intelligent use of the data produced by genomics consortia, barriers must be removed. Current consortia

CRISPR–Cas9 editing
A technique for precise and efficient editing of genetic information within a cell.

Interactome
The study of all protein–protein interactions in a system.

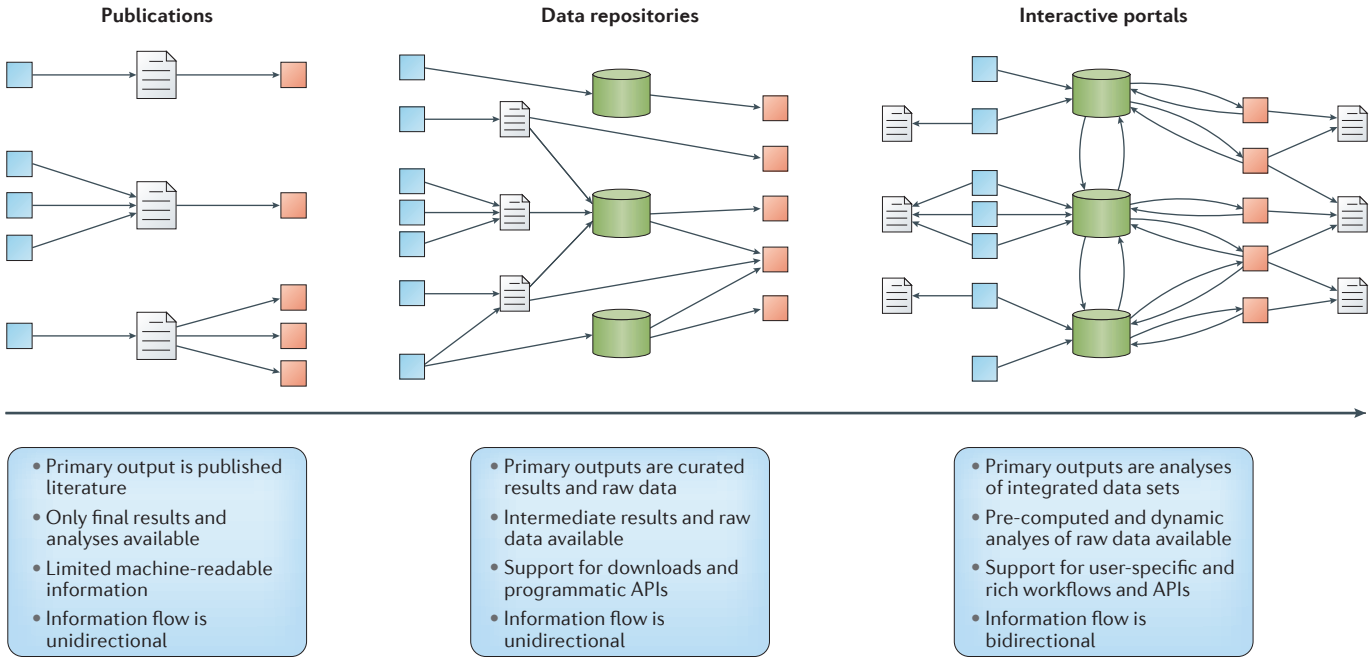

Figure 3 | **The evolution of human genetic knowledge bases.** For experimentalists to make wider use of the analyses and data produced by genetics consortia (FIG. 1), new access mechanisms are necessary. Traditionally, the results produced by genome-wide association study (GWAS) consortia have been primarily accessible through publications, which list loci that meet stringent genome-wide significance. Catalogues of associations require manual curation from the literature. The left panel shows the flow of information from data producers (blue boxes on the left) to publications (white papers), which are then read by data consumers (pink boxes on the right). Over time, many studies have begun to make available a wider range of intermediate results alongside publications, such as files that contain associations for every analysed variant. Genomic resources commonly used in experiments to understand GWAS associations (FIG. 2) have also been made accessible through databases and portals. The middle panel shows a revised flow of information, in which not only published results but also intermediate analyses or raw data are made available through multiple knowledge bases (green cylinders). In the future, we argue that much more information could be extracted from genetic analyses if portals made available results as well as facilitated novel analyses on data sets integrated across multiple studies. The right panel shows a future flow of information, in which consumers of data specify complex analytical workflows, which are carried out on data within interconnected knowledge bases. By enabling a broad community of users to carry out custom analyses, progress towards understanding type 2 diabetes mellitus (T2D) biology may accelerate (shown by publications on the right). API, application programming interface.

settings are seldom the ideal context for producing the analyses required for any specific experiment; although custom analyses can be requested, disparate data set locations and multiple analysts can make it challenging to compute results in the most rigorous and expedited fashion. Furthermore, experimentalists or other 'consumers' of genomic data have varied levels of expertise and may be unable to interpret current analytical practices (TABLE 2).

The increased convergence of experimental approaches and resources, however, suggests that many investigators could benefit from access to genomic data sets through a tractable number of user-friendly workflows. A logical mechanism, recently embraced by the wider T2D community[148], is a web-based portal to an integrated T2D genetics knowledge base, designed explicitly to aid translation of association results to biological insights. Other genotype–phenotype databases have become commonplace, most notably for Mendelian diseases[149,150], and the inherent challenges in building and supporting them are well reviewed elsewhere[151]. Additional reference epigenomic[107,108], transcriptomic[109] and genetic[152] data sets commonly used in functional experiments are also increasingly available through public web portals: in some cases through integrated analyses[153,154].

To facilitate functional studies of T2D loci, a portal must be designed to support experimentalists, rather than to serve as a repository for consortium-produced results (FIG. 3). This will require increased interactivity, data drawn from multiple interconnected sources and access to on-demand analyses of sensitive genotype and phenotype data. Such a portal would forge new connections among GWAS consortia and experimentalists in the academic, government or private sectors, diversifying the downstream uses of genetic data sets.

Building a knowledge base

To realize the vision of a shared T2D knowledge base and portal (FIG. 4), several efforts offer insight. Modern portals, such as those for the GTEx or WashU Epigenome Browser projects, are increasingly interactive

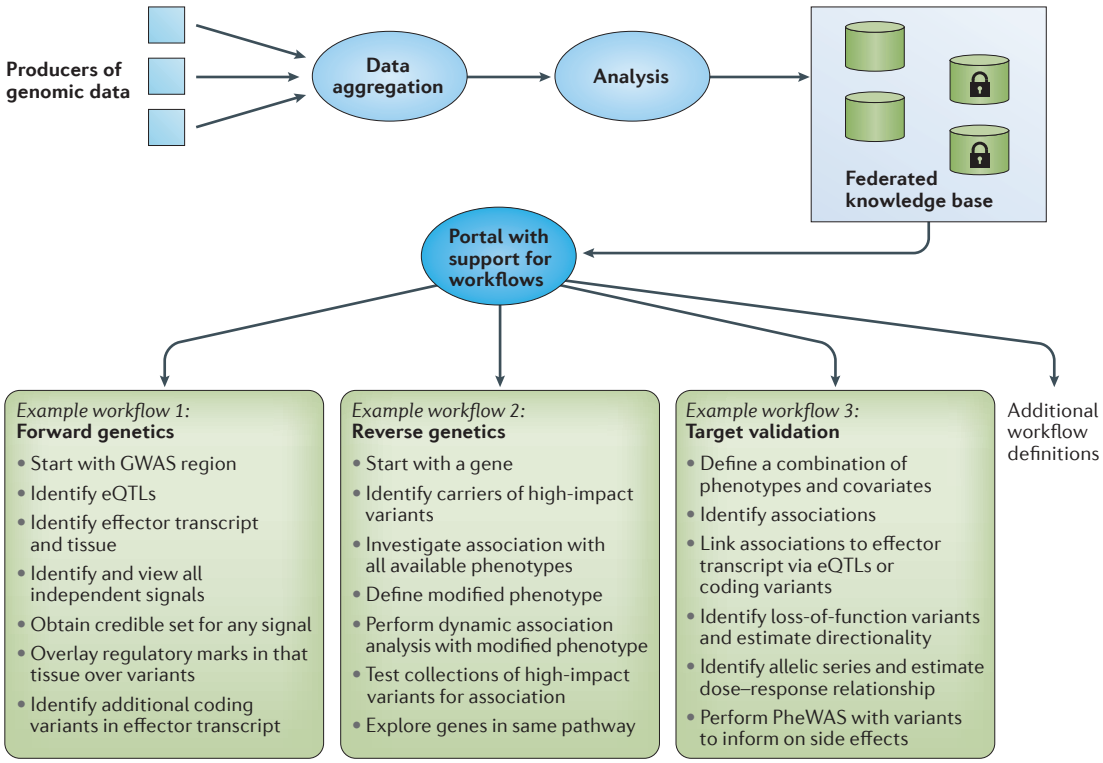

Figure 4 | **An integrated knowledge base of T2D genetics.** An integrated knowledge base of type 2 diabetes mellitus (T2D) genetics requires a new approach to aggregate, analyse and democratize genomic data sets. The most foundational need is to encourage cohort investigators to share genotypic and phenotypic information. Where regulations permit, these data could be aggregated at a central location, whereas federated knowledge bases will be necessary to integrate data sets that must be kept at their original locations. To enable a broad community of users to access information within these data, the methods and approaches developed by genome-wide association study (GWAS) consortia must be systematically applied to compute the needed analyses, either 'offline' in advance or 'online' on demand. Interactive web portals are then necessary for users to access these results, probably through workflows that span multiple questions or analyses (FIG. 2). eQTL, expression quantitative trait locus; PheWAS, phenome-wide association study.

and responsive. Collaborations such as the exome aggregation consortium[155] have pioneered approaches to aggregate raw sequence data from independent projects, whereas more aspirational efforts, such as the Global Alliance for Genomics and Health (GA4GH), seek to create standards and incentives for responsible genomic and clinical data sharing. Web platforms such as GenomeSpace and Galaxy[156] allow a wide community of (even non-expert) researchers to reproducibly carry out analyses, whereas tools such as the Michigan or Sanger imputation services and xBrowse allow secure upload and automated expert analysis of user data. Increasingly, new initiatives recognize the value in collaboration across prior boundaries, such as between industry and academia in the case of the Accelerating Medicines Partnership (AMP)[157], Center for Therapeutic Target Validation (CTTV)[158] and Innovative Medicines Initiative (IMI)[159].

What is needed in order to exploit these trends to advance genetic or biological understanding of T2D? On the basis of recent T2D genetic studies, extremely large sample sizes will be necessary to map or to characterize substantially more loci in the population[62]. Rich phenotype information will be needed for physiological characterizations, and additional epigenomic and transcriptomic data sets will be necessary to support molecular or cellular experiments. The foundational task for a T2D knowledge base is thus to support large-scale genetic data aggregation, harmonization and integration with other resources. Although 'business intelligence' approaches commonly address such challenges through 'big data' analytics or 'data warehouses' (REF. 160), applying analogous computational techniques to human genetic data sets will require substantial research and investment.

As a T2D knowledge base will contain sensitive patient data, secure mechanisms will be necessary to enforce access to only consented data sets. Where regulations prevent transfer to a central location, federated databases — in which data reside locally but support central analyses — will need development. Incentives such as increased access to cutting-edge analytics or fair publication and data embargo policies must be developed to inspire investigators, who typically spend years or decades collecting patient data, to contribute to the knowledge base for the sake of global discovery. Potentially, a shared knowledge base might enable entirely new categories of data sets, such as those from clinical trials

Business intelligence
A term, commonly used in business, that denotes a set of techniques for transforming raw data into meaningful insights.

Big data
A term for data sets that are so large or complex that new paradigms are needed to extract meaningful insights from them.

Data warehouses
A system for carrying out integrated analyses across multiple initially disparate data sources.

with rich information on patient drug responses[161], to be integrated with genetic association analyses for the first time.

Data sets in a knowledge base must be further analysed to produce broadly applicable results. Consortium-developed methods and workflows for genetic association analysis must be made sufficiently robust and efficient to analyse large and diverse data set collections. 'Offline' analyses will need regular (and automatic, where possible) application to update pre-computed statistics, while 'online' analyses will be needed for queries that cannot be pre-computed, such as novel patient stratifications to define homogeneous disease subtypes or aggregate association tests[63] of custom variant sets from functional assays[56,57]. Some online analyses may be complex, offering an opportunity for users to create and share custom workflows with the community. Where necessary, standards or new approaches will be needed to extend the analyses offered by a knowledge base to a federated setting.

Association results will also require additional processing to produce interpretable answers to biological queries. Statistical significance must be carefully conveyed to non-technical investigators to limit both type 1 errors, which arise from the multiple testing burden of 'online' analyses, and type 2 errors, which arise from false equation of absence of evidence with evidence of absence. Analyses from the knowledge base should also be packaged to suggest actionable hypotheses or worthwhile experiments, such as affected cell types or ranked lists of causal variants. Ideally, researchers who investigate these hypotheses should be encouraged to contribute their experimental results back to the knowledge base to empower future analyses by others.

Finally, these data and analyses must be made available through a public portal. Analytical results will need accompanying summaries and visualizations that answer simple but specific biological questions. Users will need the capabilities to carry out interactive workflows and have the ability to save work or rerun analyses reproducibly even if the underlying data are updated. Clear documentation and educational content must accompany the data and methods and must be tailored to a broad community with varied levels of expertise.

In summary, an ideal T2D knowledge base must be comprehensive, secure, compliant, automated and rigorous, yet also interpretable and inviting. If successful, it would greatly enhance the value of genetic association studies by generating a synergistic link between a network of data contributors and a community of experimentalists. It can therefore not only serve as a central repository of a large trove of genomic information but also aspire to 'democratize genetics' to the global community, thus becoming a transformative engine for discovery and paradigm for other disease fields.

Conclusions

Two trends have emerged from the study of T2D genetics in recent years. First, evidence has mounted that T2D genetic architecture is likely to be polygenic and characterized by many loci that are detectable only in hundreds of thousands of samples, arguing that larger and larger collections of genetic data will be necessary to discover disease-relevant variants in the population. Second, the increasing number of genes or processes that are linked to T2D will only increase the diversity of approaches necessary to translate these associations to mechanistic insight, although common resources and workflows have begun to emerge.

We have argued that a T2D knowledge base and portal could exploit both of these trends and create additional value from the consortium-led analyses that will probably continue to be the main paradigm for genetic mapping. Through an interactive portal that connects researchers around the world, the current — mostly unidirectional — flow of information from genetic discovery to functional characterization may one day be augmented by a cycle in which functional experiments inspire additional genetic experiments as well. One such portal, the Type 2 Diabetes Knowledge Portal, represents the efforts of over 100 investigators to increase access to genetic analyses of tens to hundreds of thousands of samples. Whatever the means to facilitate collaboration, a diverse collection of communities using increasingly common resources will be necessary to continue to advance our understanding of T2D biology and help one day to improve patient treatment and outcomes.

1. Hemminki, K., Li, X., Sundquist, K. & Sundquist, J. Familial risks for type 2 diabetes in Sweden. *Diabetes Care* **33**, 293–297 (2010).
2. Almgren, P. *et al.* Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* **54**, 2811–2819 (2011).
3. Kahn, S. E., Cooper, M. E. & Del Prato, S. Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. *Lancet* **383**, 1068–1083 (2014).
4. Fowler, M. J. Microvascular and macrovascular complications of diabetes. *Clin. Diabetes* **26**, 77–82 (2008).
5. Tancredi, M. *et al.* Excess mortality among persons with type 2 diabetes. *N. Engl. J. Med.* **373**, 1720–1732 (2015).
6. Altshuler, D. *et al.* The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* **26**, 76–80 (2000).
7. Gloyn, A. L. *et al.* Large-scale association studies of variants in genes encoding the pancreatic β-cell K_{ATP} channel subunits Kir6.2 (*KCNJ11*) and SUR1 (*ABCC8*) confirm that the *KCNJ11* E23K variant is associated with type 2 diabetes. *Diabetes* **52**, 568–572 (2003).
8. Grant, S. F. *et al.* Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323 (2006).
9. Guan, W., Pluzhnikov, A., Cox, N. J. & Boehnke, M. Meta-analysis of 23 type 2 diabetes linkage studies from the international type 2 diabetes linkage analysis consortium. *Hum. Hered.* **66**, 35–49 (2008).
10. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
11. McCarthy, M. I. Genomics, type 2 diabetes, and obesity. *N. Engl. J. Med.* **363**, 2339–2350 (2010).
12. McClellan, J. & King, M. C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).
13. Goldstein, D. B. Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 1696–1698 (2009).
14. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
15. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
16. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
17. Diabetes Genetics Initiative. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
18. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
19. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
20. Steinthorsdottir, V. *et al.* A variant in *CDKAL1* influences insulin response and risk of type 2 diabetes. *Nat. Genet.* **39**, 770–775 (2007).
21. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
22. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40**, 638–645 (2008).
This is the first example of collaboration and data sharing to identify new association signals for T2D.

23. de Bakker, P. I. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–R128 (2008).
24. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
25. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
26. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
27. Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
This is an early illustration of the paradigm now established for GWAS analyses.
28. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105–116 (2010).
29. Zhou, K. *et al.* Common variants near *ATM* are associated with glycaemic response to metformin in type 2 diabetes. *Nat. Genet.* **43**, 117–120 (2011).
30. Saxena, R. *et al.* Genetic variation in *GIPR* influences the glucose and insulin responses to an oral glucose challenge. *Nat. Genet.* **42**, 142–148 (2010).
31. Strawbridge, R. J. *et al.* 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**, 2624–2634 (2011).
32. Cho, Y. S. *et al.* Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in East Asians. *Nat. Genet.* **44**, 67–72 (2012).
33. Kooner, J. S. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.* **43**, 984–989 (2011).
34. Yamauchi, T. *et al.* A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at *UBE2E2* and *C2CD4A–C2CD4B*. *Nat. Genet.* **42**, 864–868 (2010).
35. Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8**, e1002793 (2012).
36. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
37. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycaemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
38. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
A thorough summary of the debate around different models of genetic architecture is presented here, following the results of early GWAS.
39. Hirschhorn, J. N. Genomewide association studies — illuminating biologic pathways. *N. Engl. J. Med.* **360**, 1699–1701 (2009).
40. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
41. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40**, 695–701 (2008).
42. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
43. Goldstein, D. B. The importance of synthetic associations will only be resolved empirically. *PLoS Biol.* **9**, e1001008 (2011).
44. Anderson, C. A., Soranzo, N., Zeggini, E. & Barrett, J. C. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biol.* **9**, e1000580 (2011).
45. Wray, N. R., Purcell, S. M. & Visscher, P. M. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biol.* **9**, e1000579 (2011).
46. Waters, K. M. *et al.* Consistent association of type 2 diabetes risk variants found in Europeans in diverse racial and ethnic groups. *PLoS Genet.* **6**, e1001078 (2010).
47. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44**, 483–489 (2012).
48. Agarwala, V., Flannick, J., Sunyaev, S., Go, T. D. C. & Altshuler, D. Evaluating empirical bounds on complex disease genetic architecture. *Nat. Genet.* **45**, 1418–1427 (2013).
This is a comprehensive study of the support from empirical data sets for different models of T2D genetic architecture.
49. Huyghe, J. R. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.* **45**, 197–201 (2013).
50. Albrechtsen, A. *et al.* Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia* **56**, 298–310 (2013).
51. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298 (2014).
52. Estrada, K. *et al.* Association of a low-frequency variant in *HNF1A* with type 2 diabetes in a Latino population. *JAMA* **311**, 2305–2314 (2014).
53. Mahajan, A. *et al.* Identification and functional characterization of *G6PC2* coding variants influencing glycaemic traits define an effector transcript at the *G6PC2-ABCB11* locus. *PLoS Genet.* **11**, e1004876 (2015).
54. Wessel, J. *et al.* Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat. Commun.* **6**, 5897 (2015).
55. Manning, A. K. *et al.* A low frequency *AKT2* coding variant enriched in the Finnish population is associated with fasting insulin levels. (Abstract #56) *The 64th Annual Meeting of The American Society of Human Genetics, San Diego, California* http://www.ashg.org/2014meeting/pdf/2014_ASHG_Meeting_Platform_Abstracts.pdf (18–22 Oct 2014).
56. Bonnefond, A. *et al.* Rare *MTNR1B* variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat. Genet.* **44**, 297–301 (2012).
This is an early example of the ability of functional assays to filter benign from deleterious alleles and improve the power of aggregate association tests.
57. Majithia, A. R. *et al.* Rare variants in *PPARG* with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc. Natl Acad. Sci. USA* **111**, 13127–13132 (2014).
58. Flannick, J. *et al.* Loss-of-function mutations in *SLC30A8* protect against type 2 diabetes. *Nat. Genet.* **46**, 357–363 (2014).
This is one of the first studies to identify protective rare variants associated with T2D, demonstrating the need for large-scale data aggregation.
59. Rutter, G. A. Think zinc: new roles for zinc in the control of insulin secretion. *Islets* **2**, 49–50 (2010).
60. Nicolson, T. J. *et al.* Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. *Diabetes* **58**, 2070–2083 (2009).
61. Lohmueller, K. E. *et al.* Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. *Am. J. Hum. Genet.* **93**, 1072–1086 (2013).
62. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* http://dx.doi.org/10.1038/nature18642 (2016).
A comprehensive characterization is presented here of the genetic architecture of T2D using various sequencing approaches.
63. Moutsianas, L. *et al.* The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet.* **11**, e1005165 (2015).
This simulation study shows the need for potentially numerous aggregate association analyses to identify disease genes with different allelic spectra.
64. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl Acad. Sci. USA* **111**, E455–E464 (2014).
Approaches for rare variant association studies are thoroughly explained and examined.
65. Mahajan, A. *et al.* Genome-wide *trans*-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
66. Scott, R. A. *et al.* Genome-wide association study imputed to 1000 Genomes reveals 18 novel associations with type 2 diabetes. (Abstract 53). *The 64th Annual Meeting of The American Society of Human Genetics, San Diego, California* http://www.ashg.org/2014meeting/pdf/2014_ASHG_Meeting_Platform_Abstracts.pdf (18–22 Oct 2014).
67. Sigma Type 2 Diabetes Consortium. Sequence variants in *SLC16A11* are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97–101 (2014).
68. Ng, M. C. Y. *et al.* Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet.* **10**, e1004517 (2014).
69. Moltke, I. *et al.* A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193 (2014).
The largest effect association observed for T2D to date is described here, demonstrating the power of studying population isolates as well as insights that can be learned from variant carrier phenotyping.
70. Mahajan, A. *et al.* Large-scale exome chip association analysis identifies novel type 2 diabetes susceptibility loci and highlights candidate effector genes. (Abstract #299). *The 65th Annual Meeting of The American Society of Human Genetics, Baltimore, Maryland* http://www.ashg.org/2015meeting/pdf/57715_Platform.pdf (6–10 Oct 2015).
71. Bulik-Sullivan, B. K. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
72. Iyengar, S. K. *et al.* Genome-wide association and trans-ethnic meta-analysis for advanced diabetic kidney disease: Family Investigation of Nephropathy and Diabetes (FIND). *PLoS Genet.* **11**, e1005352 (2015).
73. Prokopenko, I. *et al.* A central role for *GRB10* in regulation of islet function in man. *PLoS Genet.* **10**, e1004235 (2014).
74. Horikoshi, M. *et al.* Discovery and fine-mapping of glycaemic and obesity-related trait loci using hhigh-density imputation. *PLoS Genet.* **11**, e1005230 (2015).
75. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *bioRxiv* http://dx.doi.org/10.1101/035170 (2015).
76. Dimas, A. S. *et al.* Impact of type 2 diabetes susceptibility variants on quantitative glycaemic traits reveals mechanistic heterogeneity. *Diabetes* **63**, 2158–2171 (2014).
77. Ingelsson, E. *et al.* Detailed physiologic characterization reveals diverse mechanisms for novel genetic loci regulating glucose and insulin metabolism in humans. *Diabetes* **59**, 1266–1275 (2010).
78. Scott, R. A. *et al.* Common genetic variants highlight the role of insulin resistance and body fat distribution in type 2 diabetes, independent of obesity. *Diabetes* **63**, 4378–4387 (2014).
79. Rosengren, A. H. *et al.* Reduced insulin exocytosis in human pancreatic β-cells with gene variants linked to type 2. *Diabetes* **61**, 1726–1733 (2012).
80. Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
81. Fall, T. *et al.* The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. *PLoS Med.* **10**, e1001474 (2013).
82. Abbasi, A. *et al.* Bilirubin as a potential causal factor in type 2 diabetes risk: a Mendelian randomization study. *Diabetes* **64**, 1459–1469 (2015).
83. De Silva, N. M. *et al.* Mendelian randomization studies do not support a role for raised circulating triglyceride levels influencing type 2 diabetes, glucose levels, or insulin resistance. *Diabetes* **60**, 1008–1018 (2011).
84. Haase, C. L., Tybjærg-Hansen, A., Nordestgaard, B. G. & Frikke-Schmidt, R. H.D. L. Cholesterol and risk of type 2 diabetes: a Mendelian randomization study. *Diabetes* **64**, 3328–3333 (2015).
85. Yaghootkar, H. *et al.* Mendelian randomization studies do not support a causal role for reduced circulating adiponectin levels in insulin resistance and type 2 diabetes. *Diabetes* **62**, 3589–3598 (2013).
86. Sluijs, I. *et al.* A mendelian randomization study of circulating uric acid and type 2 diabetes. *Diabetes* **64**, 3028–3036 (2015).

87. Pal, A. *et al. PTEN* mutations as a cause of constitutive insulin sensitivity and obesity. *N. Engl. J. Med.* **367**, 1002–1011 (2012).
This paper provides a nice illustration of the power of deep phenotyping studies to gain molecular, cellular and physiological insights into disease pathways.
88. Wang, L. *et al.* PTEN deletion in pancreatic α-cells protects against high-fat diet–induced hyperglucagonemia and insulin resistance. *Diabetes* **64**, 147–157 (2015).
89. Perry, J. R. B. *et al.* 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 Genetics* **8**, e1002741 (2012).
This is an example in which phenotypic sample stratification identified a novel GWAS locus.
90. Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycaemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
91. Li, L. *et al.* Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **7**, 311ra174 (2015).
92. Bonnefond, A. *et al.* Association between large detectable clonal mosaicism and type 2 diabetes with vascular complications. *Nat. Genet.* **45**, 1040–1043 (2013).
93. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
94. Walford, G. A. *et al.* Metabolite traits and genetic risk provide complementary information for the prediction of future type 2 diabetes. *Diabetes Care* **37**, 2508–2514 (2014).
95. Wurtz, P. *et al.* Circulating metabolite predictors of glycemia in middle-aged men and women. *Diabetes Care* **35**, 1749–1756 (2012).
96. Chambers, J. C. *et al.* Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *Lancet Diabetes Endocrinol.* **3**, 526–534 (2015).
97. Talmud, P. J. *et al.* Sixty-five common genetic variants and prediction of type 2 diabetes. *Diabetes* **64**, 1830–1840 (2015).
98. Vassy, J. L. *et al.* Polygenic type 2 diabetes prediction at the limit of common variant detection. *Diabetes* **63**, 2172–2182 (2014).
99. Flannick, J. *et al.* Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat. Genet.* **45**, 1380–1385 (2013).
100. Patel, K., Weedon, M. N., Ellard, S., Oram, R. A. & Hattersley, A. T. Type 1 diabetes genetic risk score — a novel tool to differentiate monogenic diabetes from T1D. (Abstract 1746-P) *Diabetes* **64** (Suppl. 1), A453 (2015).
101. Oram, R. A. *et al.* A type 1 diabetes genetic risk score can aid discrimination between type 1 and type 2 diabetes in young adults. *Diabetes Care* **39**, 337–344 (2016).
102. Gaulton, K. J. *et al.* Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
This is a nice example of the common workflow to localize common variant associations to causal variants and molecular disease mechanisms.
103. Shea, J. *et al.* Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat. Genet.* **43**, 801–805 (2011).
104. Yaghootkar, H. *et al.* Association analysis of 29,956 individuals confirms that a low-frequency variant at *CCND2* halves the risk of type 2 diabetes by enhancing insulin secretion. *Diabetes* **64**, 2279–2285 (2015).
105. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
An early fine mapping study is presented here that established the now commonly used 'credible set' methodology.
106. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
107. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
108. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
109. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
110. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
This is the first prominent and clear demonstration that GWAS variants cluster within regulatory regions of the human genome.
111. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2012).
112. Parker, S. C. J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl Acad. Sci. USA* **110**, 17921–17926 (2013).
113. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
114. Soccio, Raymond, E. *et al.* Genetic variation determines PPARγ function and anti-diabetic drug response *in vivo*. *Cell* **162**, 33–44 (2015).
115. Gaulton, K. J. *et al.* A map of open chromatin in human pancreatic islets. *Nat. Genet.* **42**, 255–259 (2010).
This is one of the earliest studies that used epigenomic annotations to prioritize potentially causal variants at T2D GWAS loci.
116. Stitzel, M. L. *et al.* Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab.* **12**, 443–455 (2010).
117. Morán, I. *et al.* Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab.* **16**, 435–448 (2012).
118. van de Bunt, M. *et al.* The miRNA profile of human pancreatic islets and Beta-Cells and relationship to type 2 diabetes pathogenesis. *PLoS ONE* **8**, e55272 (2013).
119. Fadista, J. *et al.* Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl Acad. Sci. USA* **111**, 13924–13929 (2014).
120. van de Bunt, M. *et al.* Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycaemic traits to their downstream effectors. *PLoS Genet.* **11**, e1005694 (2015).
This paper shows how eQTL analysis can be used to hypothesize effector transcripts at GWAS loci.
121. Nica, A. C. *et al.* Cell-type, allelic, and genetic signatures in the human pancreatic beta cell transcriptome. *Genome Res.* **23**, 1554–1562 (2013).
122. Dayeh, T. *et al.* Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. *PLoS Genet.* **10**, e1004160 (2014).
123. Volkmar, M. *et al.* DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. *EMBO J.* **31**, 1405–1426 (2012).
124. Nilsson, E. *et al.* Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. *Diabetes* **63**, 2962–2976 (2014).
125. Yuan, W. *et al.* An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins. *Nat. Commun.* **5**, 5719 (2014).
126. Hodson, D. J. *et al.* ADCY5 couples glucose to insulin secretion in human islets. *Diabetes* **63**, 3009–3021 (2014).
127. Claussnitzer, M. *et al.* FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
This is a comprehensive investigation of the disease mechanisms responsible for a GWAS association, using nearly all relevant modern experimental and computation techniques.
128. Locke, J. M., Hysenaj, G., Wood, A. R., Weedon, M. N. & Harries, L. W. Targeted allelic expression profiling in human islets identifies *cis*-regulatory effects for multiple variants identified by type 2 diabetes genome-wide association studies. *Diabetes* **64**, 1484–1491 (2015).
129. Kulzer, Jennifer, R. *et al.* A common functional regulatory variant at a type 2 diabetes locus upregulates *ARAP1* expression in the pancreatic beta cell. *Am. J. Hum. Genet.* **94**, 186–197 (2014).
130. Ragvin, A. *et al.* Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to *HHEX*, *SOX4*, and *IRX3*. *Proc. Natl Acad. Sci. USA* **107**, 775–780 (2010).
131. Smemo, S. *et al.* Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. *Nature* **507**, 371–375 (2014).
132. Elbein, Steven, C. *et al.* Genetic risk factors for type 2 diabetes: a *trans*-regulatory genetic architecture? *Am. J. Hum. Genet.* **91**, 466–477 (2012).
133. Small, K. S. *et al.* Identification of an imprinted master *trans* regulator at the *KLF14* locus related to multiple metabolic phenotypes. *Nat. Genet.* **43**, 561–564 (2011).
This is a nice example of the power of *trans*-eQTL analysis to identify disease mechanisms responsible for a GWAS association.
134. Florez, J. C. *et al. TCF7L2* polymorphisms and progression to diabetes in the Diabetes Prevention Program. *N. Engl. J. Med.* **355**, 241–250 (2006).
135. Fogarty, M. P., Panhuis, T. M., Vadlamudi, S., Buchkovich, M. L. & Mohlke, K. L. Allele-specific transcriptional activity at type 2 diabetes — associated single nucleotide polymorphisms in regions of pancreatic islet open chromatin at the *JAZF1* locus. *Diabetes* **62**, 1756–1762 (2013).
136. Fogarty, M. P., Cannon, M. E., Vadlamudi, S., Gaulton, K. J. & Mohlke, K. L. Identification of a regulatory variant that binds, FOXA1 and FOXA2 at the *CDC123/CAMK1D* type 2 diabetes, GWAS locus. *PLoS Genet.* **10**, e1004633 (2014).
137. Travers, M. E. *et al.* Insights into the molecular mechanism for type 2 diabetes susceptibility at the *KCNQ1* locus from temporal changes in imprinting status in human islets. *Diabetes* **62**, 987–992 (2013).
138. Claussnitzer, M. *et al.* Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. *Cell* **156**, 343–358 (2014).
139. Wei, F.-Y. *et al.* Deficit of $tRNA^{Lys}$ modification by Cdkal1 causes the development of type 2 diabetes in mice. *J. Clin. Invest.* **121**, 3598–3608 (2011).
140. McCulloch, L. J. *et al.* GLUT2 (*SLC2A2*) is not the principal glucose transporter in human pancreatic beta cells: implications for understanding genetic association signals at this locus. *Mol. Genet. Metab.* **104**, 648–653 (2011).
141. Beer, N. L. *et al.* 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**, 4081–4088 (2009).
142. Smith, S. B. *et al.* Rfx6 directs islet formation and insulin production in mice and humans. *Nature* **463**, 775–780 (2010).
143. Zhu, H. *et al.* The *Lin28/let-7* axis regulates glucose metabolism. *Cell* **147**, 81–94 (2011).
144. Dai, N. *et al.* IGF2BP2/IMP2-Deficient mice resist obesity through enhanced translation of *Ucp1* mRNA and other mRNAs encoding mitochondrial proteins. *Cell Metab.* **21**, 609–621 (2015).
145. Mercader, J. M. *et al.* Identification of novel type 2 diabetes candidate genes involved in the crosstalk between the mitochondrial and the insulin signaling systems. *PLoS Genet.* **8**, e1003046 (2012).
146. Taneera, J. *et al.* Expression profiling of cell cycle genes in human pancreatic islets with and without type 2 diabetes. *Mol. Cell. Endocrinol.* **375**, 35–42 (2013).
147. Burns, S. M. *et al.* High-throughput luminescent reporter of insulin secretion for discovering regulators of pancreatic beta-cell function. *Cell Metab.* **21**, 126–137 (2015).
148. The American Diabetes Association. Genetics portal for type 2 diabetes debuts. *Diabetes Dispatch* http://www.diabetesdispatchextra.org/genetics-portal-for-type-2-diabetes-debuts (2015).
149. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
150. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).
151. Brookes, A. J. & Robinson, P. N. Human genotype-phenotype databases: aims, challenges and opportunities. *Nat. Rev. Genet.* **16**, 702–715 (2015).

152. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
153. Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681 (2015).
154. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
155. Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv* http://dx.doi.org/10.1101/030338 (2015).
156. Blankenberg, D. *et al.* in *Current Protocols in Molecular Biology* (eds Ausubel, F. M. *et al.*) (Wiley, 2010).
157. Reardon, S. Pharma firms join NIH on drug development. *Nature* http://dx.doi.org/10.1038/nature.2014.14672 (2014).
158. Barrett, J. C., Dunham, I. & Birney, E. Using human genetics to make new medicines. *Nat. Rev. Genet.* **16**, 561–562 (2015).
159. Goldman, M. The innovative medicines initiative: a European response to the innovation challenge. *Clin. Pharmacol. Ther.* **91**, 418–425 (2012).
160. Kimball, R., Ross, M., Thornthwaite, W., Mundy, J. & Becker, K. *The Data Warehouse Lifecycle Toolkit* 2nd edn (Wiley, 2008).
161. Espeland, M. A. *et al.* Consent for genetics studies among clinical trial participants: findings from Action for Health in Diabetes (Look AHEAD). *Clin. Trials* **3**, 443–456 (2006).
162. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
163. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
164. Pal, A. *et al.* Loss-of-function mutations in the cell-cycle control gene *CDKN2A* impact on glucose homeostasis in humans. *Diabetes* **65**, 527–533 (2015).
165. Torekov, S. S. *et al.* KCNQ1 long QT syndrome patients have hyperinsulinemia and symptomatic hypoglycemia. *Diabetes* **63**, 1315–1325 (2014).

Acknowledgements
The authors thank B. Alexander for help with figure design and creation, as well as helpful discussions.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
GenomeSpace: http://www.genomespace.org
Global Alliance for Genomics and Health (GA4GH): http://genomicsandhealth.org
GTEx Portal: http://gtexportal.org/home
Michigan Imputation Server: https://imputationserver.sph.umich.edu
Sanger Imputation Service: https://imputation.sanger.ac.uk
Type 2 Diabetes Knowledge Portal: http://www.type2diabetesgenetics.org
WashU Epigenome Browser: http://epigenomegateway.wustl.edu
xBrowse: https://atgu.mgh.harvard.edu/xbrowse

ALL LINKS ARE ACTIVE IN THE ONLINE PDF