Genomic annotation of disease-associated variants reveals shared functional contexts

Yasuhiro Kyono^{1,2,3} · Jacob O. Kitzman^{1,2} · Stephen C. J. Parker^{1,2}

Received: 7 August 2018 / Accepted: 27 November 2018 / Published online: 12 February 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

Variation in non-coding DNA, encompassing gene regulatory regions such as enhancers and promoters, contributes to risk for complex disorders, including type 2 diabetes. While genome-wide association studies have successfully identified hundreds of type 2 diabetes loci throughout the genome, the vast majority of these reside in non-coding DNA, which complicates the process of determining their functional significance and level of priority for further study. Here we review the methods used to experimentally annotate these non-coding variants, to nominate causal variants and to link them to diabetes pathophysiology. In recent years, chromatin profiling, massively parallel sequencing, high-throughput reporter assays and CRISPR gene editing technologies have rapidly become indispensable tools. Rather than treating individual variants in isolation, we discuss the importance of accounting for context, both genetic (such as flanking DNA sequence) and environmental (such as cellular state or environmental exposure). Incorporating these features shows promise in terms of revealing biologically convergent molecular signatures across distant and seemingly unrelated loci. Studying regulatory elements in the proper context will be crucial for interpreting the functional significance of disease-associated variants and applying the resulting knowledge to improve patient care.

Keywords Chromatin · Diabetes · Epigenome · Gene expression · Genetics · Genome-wide association study · Human · Reporter assay · Review · Transcription

Abbreviations

ATAC-seq	Assay for transposase-accessible chromatin sequencing
ChIP-seq	Chromatin immunoprecipitation sequencing
dCas9	Dead CRISPR-associated protein 9
DHS	DNase I hypersensitive site

eQTL	Expression quantitative trait loci
GATA1	GATA-binding factor 1
GFP	Green fluorescent protein
GWAS	Genome-wide association studies
LPS	Lipopolysaccharides
MPRA	Massively parallel reporter assay
RFX	Regulatory factor X
SE	Stretch enhancer
sgRNA	Single guide RNA
SNP	Single-nucleotide polymorphism
STARR-seq	Self-transcribing active regulatory region sequencing

Jacob O. Kitzman and Stephen C. J. Parker contributed equally to this review.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00125-019-4823-3>) contains a slideset of the figures for download, which is available to authorised users.

✉ Stephen C. J. Parker
scjp@umich.edu

- ¹ Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, 2049 Palmer Commons Building, Ann Arbor, MI 48109, USA
- ² Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA
- ³ Present address: Institute for Genomics and Systems Biology, University of Chicago, Chicago, IL, USA

Introduction

Type 2 diabetes is a growing public health concern worldwide, as the number of people who suffer from the disease is expected to rise from 425 million in 2017 to 629 million by 2045 [1]. Two of the hallmarks of type 2 diabetes are the development of insulin resistance in peripheral tissues (liver, skeletal muscle and adipose) and dysfunction of insulin-secreting beta cells in

the pancreatic islets [2]. While environmental factors such as diet and exercise level contribute to type 2 diabetes, genetic predisposition also plays a key role. Genome-wide association studies (GWAS) have identified more than 200 loci associated with type 2 diabetes and related metabolic traits [3] (Fig. 1a). These genetic findings have already identified targets for known drugs such as thiazolidinediones [6], which serves as a positive control for the approach. At most such genomic loci, however, the variant with the strongest statistical association with type 2 diabetes ('lead single-nucleotide polymorphism' [SNP]) is found outside of protein-coding regions, suggesting a primary role for non-coding *cis*-regulatory elements in genetic risk for type 2 diabetes. Such variants may confer risk by altering transcription factor binding sites that propagate signals from upstream transcription factors to influence downstream target gene expression. It is challenging to determine the functional effects of non-coding variants for several reasons. First, lead SNPs are not necessarily the causal variants, which could

instead be among the many other variants in strong linkage equilibrium (LD) (Fig. 1b). Second, even after mapping a causal regulatory variant, it is often difficult to identify its downstream target gene(s), which may be distant. For example, an obesity-associated variant located within the first intron of the *FTO* gene affects expression of the genes *IRX3* and *IRX5*, located over a megabase away from the variant, without affecting the expression of the closest *FTO* gene [7, 8]. To bridge the disconnect between genetic factors and disease processes, there is a critical need to experimentally and bioinformatically annotate non-coding variants to better understand the biological mechanisms that contribute to genetic predisposition to type 2 diabetes. A refined view of the genetics of type 2 diabetes may allow for personalised risk scores [9] and stratification of patients by different underlying pathophysiology [10].

Here, we review genome-wide approaches to map regulatory elements, highlighting studies that have successfully applied these approaches to add gene regulatory context to type 2

Fig. 1 Functional mapping of diabetes-associated variants using tissue-specific regulatory maps. GWAS have identified loci associated with risk for type 2 diabetes, with strength of association ($-\log_{10} p$ value) shown throughout the genome in a Manhattan plot (a, data from [4, 5]). Each genome-wide significant region (above the horizontal red line) can then be explored using a locus-zoom plot (b), which shows one of the type 2 diabetes-associated loci (overlapping the gene *WFS1*) as an example [5]. In the locus zoom plot, each dot represents a variant associated with type 2 diabetes, and its colour represents the level of linkage disequilibrium (LD), with the lead variant (reference variant [Ref Var]) highlighted in purple. Most SNPs occur in non-coding regions, where chromatin state analyses (c) help identify locations of tissue-specific regulatory regions. While some enhancer regions may be shared across tissues, there are others that are unique. This figure is available as part of a [downloadable slideset](#)

diabetes GWAS variants. We discuss how integrating these maps may identify convergent mechanisms across distant loci, and shared biology underpinning type 2 diabetes pathogenesis. Finally, we propose future studies that could provide additional mechanistic information.

Tissue-specific maps of chromatin state identify relevant regulatory regions

In eukaryotes, DNA is wrapped around histone proteins to form nucleosomes; strings of nucleosomes then form a higher-order structure called chromatin in the cell nucleus. Beyond structural roles in packaging DNA, histones contribute to establishing and maintaining cell-type-specific gene expression programs by signalling through post-translational modifications and by regulating the accessibility of DNA to transcription factors.

Histone modifications Protruding ends of histone proteins can be post-translationally modified with various covalent marks (e.g. methylation, acetylation). Genome-wide maps for various histone marks have been generated for human cell lines and primary tissues using chromatin immunoprecipitation and sequencing (ChIP-seq) [11, 12], revealing that histones in different regions of the genome are decorated with distinct marks, reflecting the regulatory activity of those regions [13]. For example, transcription start sites are marked by trimethylation of histone H3 lysine 4 (H3K4me3), while enhancers are marked by both mono-methylation of H3K4 (H3K4me1) and acetylation of H3K27 (H3K27ac). Chromatin segmentation analyses integrate combinations of histone marks to annotate the genome into discrete ‘states’, each of which can be given labels that describe the underlying regulatory activity such as promoters and enhancers [14] (Fig. 1c). Parker, Stitzel and colleagues constructed chromatin state maps for pancreatic islets, and identified islet-specific stretch enhancers (SEs), which are long (≥ 3 kb) segments of the genome that are continuously decorated with enhancer-associated histone marks. SEs were found near to critical pancreatic islet genes (e.g. *INS1*), and were enriched for GWAS variants associated with type 2 diabetes and related traits [15]. Similar observations have been made for disease-relevant cell types in other disease models [16–18]. These studies collectively represent the first level of functional convergence in which disease-relevant variants across the genome are enriched in a set of large enhancers active in specific tissues. However, while chromatin state analysis is useful for narrowing down the regions of interest to a small subset of regulatory regions, the resolution of analysis is approximately 200 bp (a consequence of the fact that each nucleosome contains about 147 bp of DNA wrapped around the histones), which is still too coarse to pinpoint the underlying sequence motif(s) that could be mediating a genetic regulatory effect.

Accessible/open chromatin regions Transcription factor binding can create focal changes to chromatin architecture such that nucleosomes are displaced and the surrounding DNA becomes more physically accessible (‘open’). Information about chromatin accessibility can then be used to infer binding of individual transcription factors by looking for small regions that are protected (‘footprints’) relative to the more accessible flanking regions. Some of the first genome-wide maps of open chromatin regions in human pancreatic islets used formaldehyde-assisted isolation of regulatory elements (FAIRE-seq) [19] or DNase I digestion coupled to DNA sequencing (DNase-seq) [20]. By comparing these data to maps from other cell types, these studies identified islet-specific open chromatin regions that coincided with evolutionarily conserved binding sites for key islet transcription factors near to genes of critical importance in pancreatic islets (e.g. *PDX1* and *NKX6-1*). Subsequently, Pasquali and colleagues identified a subclass of open chromatin regions (which they referred to as C3 regions) enriched for enhancer-associated histone marks, bound by multiple key islet transcription factors, and in long-range physical interactions with nearby islet-specific gene promoters [21]. In recent years, a new open chromatin profiling method, the assay for transposase-accessible chromatin sequencing (ATAC-seq) [22], has enabled more routine analysis of scarce samples, such as human pancreatic islets,

Fig. 2 Pinpointing individual *cis*-regulatory elements within broader regulatory regions. **(a)** Open chromatin regions can be identified by DNase-seq or ATAC-seq. Motif analysis within open chromatin regions may identify bound cognate transcription factors. Here we have shown how searching a transcription factor binding site motif database for the motifs identified in the open chromatin region can nominate a motif associated with a specific transcription factor. **(b)** Expression quantitative loci (eQTL) analyses, which use statistical associations between genetic variation and gene expression at a population level, can identify variants that influence expression of downstream target genes, for example, by activating or disrupting transcription factor binding sites. In this example, the blue ‘C’ allele disrupts the red motif and is associated with decreased expression of the hypothetical target ‘gene X’. This figure is available as part of a [downloadable slideset](#)

because of its lower minimum sample size (Fig. 2a). Recent islet ATAC-seq analyses identified significant enrichment for type 2 diabetes GWAS SNPs within the footprints of transcription factors belonging to the regulatory factor X (RFX) family [23]. Interestingly, for all nine SNPs found across five independent loci, the risk alleles disrupted highly preferred bases within the RFX motifs, suggesting a model in which cumulative disruption of RFX binding sites leads to increased risk of type 2 diabetes. Notably, mutations in the DNA-binding domain of RFX6 result in Mitchell–Riley syndrome, an autosomal recessive neonatal form of diabetes [24].

Compared with analysis of histone marks, open chromatin analyses (especially ATAC-seq) have a higher resolution, permitting the identification of specific transcription factor motifs that may be systematically altered by risk alleles. These findings represent a higher-resolution form of convergence: not only are islet enhancers enriched for overlap with type 2 diabetes GWAS variants, but the specific RFX motifs within these larger islet enhancers are systematically disrupted by risk alleles. The next step towards identifying the putative target gene of the regulatory motif can be accomplished with expression quantitative trait loci (eQTL) studies, which look at population-level statistical associations between gene expression and genetic variation to assign SNPs to target genes (Fig. 2b). Several such studies have been conducted across diverse and diabetes-relevant human tissues, such as adipose tissue, islets, liver and skeletal muscle [23, 25–28], and larger emerging studies promise to be a valuable sources of eQTLs. Additional layers of regulatory annotation could reveal additional signatures of convergence.

Activity-based functional genomics to nominate causal variants

While mapping histone marks and open chromatin regions can identify candidate regulatory regions, complementary approaches are needed to functionally validate the effect of individual genetic variants on enhancer activity. One such method is to narrow down putative causal variants to a subset through statistical genetic fine-mapping (reviewed in [29]). However, even with large cohorts and diverse ancestries, there frequently remain a number of plausible candidate regulatory variants. Because statistical genetic fine-mapping techniques and functional genomics may yield discordant results, it is important to compare the subsets of variants that emerge from these approaches.

Massively parallel reporter assay Reporter assays provide a complementary set of methods to test the regulatory activity of enhancers and promoters by cloning them into an episomal (extra-chromosomal) plasmid containing a reporter gene (e.g. green fluorescent protein [GFP], luciferase) [30]. Prior studies have successfully applied these assays to identify type 2 diabetes

risk variants that alter enhancer activity (for specific loci, see ref. [31]). However, these approaches conventionally require one-by-one testing of each enhancer fragment and/or allele, making them too low-throughput to accommodate the rapidly growing number of loci associated with type 2 diabetes [3]. Massively parallel reporter assays (MPRAs) have recently emerged as a powerful tool to study the activity of tens of thousands of cloned DNA fragments (for a review see [32]), which allow researchers to construct, transfect and test the activity of many putative enhancers all in a single, pooled experimental format (Fig. 3a, b). Here we highlight some of the studies that have successfully used MPRAs to identify regulatory variants. However, since none of the MPRA studies conducted to date have focused on type 2 diabetes, we will discuss studies in other disease models to illustrate the utility of these methods.

One such study sought to functionally validate variants associated with erythrocyte traits. They tested a library of 75 GWAS lead SNPs and 2,681 others in high linkage disequilibrium ($r^2 \geq 0.8$) [34]. Sequences with significant activity in their MPRA were more likely to originate from highly accessible regions of chromatin termed DNase I hypersensitive sites (DHSs) in erythroid cell types *in vivo*, and were also more likely to overlap with the binding sites for a critical erythroid transcription factor GATA-binding factor 1 (GATA1) and its cofactor T cell acute lymphocytic leukaemia 1 (TAL1), as compared with sequences that did not exhibit activity in their reporter assay. There were 32 variants with a concordant effect on regulatory activity and chromatin accessibility, i.e. alleles that were more active in MPRA exhibited higher chromatin accessibility at their native loci. In another recent study, an MPRA library was tested in two different human immortalised cell lines (HepG2 and K562), identifying motifs and variants predictive of cell-type-specific activity [35]. Thus, reporter assays at least partially reflect the native transcriptional regulatory environment, and employing these assays in distinct cell types provides cellular and developmental contexts needed to understand the biological effects of genetic variants in different tissues and organs.

Self-transcribing active regulatory region sequencing (STARR-seq) is another promising MPRA approach [36], in which cloned candidate DNA fragments stimulate their own transcription, and the resulting enhancer activities are measured by RNA sequencing (RNA-seq). One recent application of STARR-seq examined putative enhancers overlapping GWAS loci associated with cancer risk [37], and observed that ~18% of fragments tested had activity above background. Active fragments were enriched for those bearing active histone marks at their endogenous loci, which suggests that the regulatory signatures needed to establish these chromatin states *in vivo* is at least partially encoded on these fragments, and may be disrupted by variants within them.

There are several limitations associated with current MPRA approaches. First, fragments may have different activities on episomal plasmids as compared with their endogenous

Fig. 3 Functionally dissecting *cis*-regulatory elements. **(a)** Tiling MPRA may identify sequence fragments that can act as promoters or enhancers, either basally (grey) or after an environmental exposure (orange). **(b)** Allele-specific MPRA is used to systematically analyse the impact of nucleotide substitutions on the regulatory activity of selected elements. Analysis of the MPRA data with quantitative allele-specific analysis of reads (QuASAR) can help take into account base-calling errors and over dispersion [33]. One of the limitations of MPRA is that the elements may show different activities in DNA fragments in isolation compared with their endogenous loci, and **(c)** CRISPR mutagenesis can be used to test the necessity of regulatory elements in the native chromatin context. This figure is available as part of a [downloadable slideset](#)

loci, where they are packaged into chromatin and flanked by native sequence. Second, the activity of some of the enhancers may require the presence of multiple transcription factor binding motifs (reviewed in [38]), such that MPRA may fail to detect activity for individual fragments in isolation. Third, some reporter plasmids use synthetic, heterologous elements such as a minimal promoter to test enhancer fragments. Prior studies have suggested there may be a requirement for promoters and enhancers to be of compatible ‘types’ [39, 40]; therefore, if candidate enhancers are incompatible with the type of promoters used in the MPRA, this may lead to false-negative or false-positive results. Finally, MPRA is only intended to test whether individual *cis*-regulatory elements (or allelic variants) are sufficient to activate gene expression; a key complementary question, which we discuss next, is whether individual sites are necessary for enhancer activity.

In situ mutagenesis of regulatory elements using CRISPR/Cas9

Genome and epigenome engineering methods provide powerful new tools for studying gene dysregulation in type 2 diabetes. Gene or regulatory element knockout cells or animals can now be routinely derived using CRISPR/Cas9-directed

mutagenesis (Fig. 3c). In addition, modulation of target gene expression has been demonstrated by fusing catalytically dead CRISPR-associated protein 9 (dCas9) protein to various transcriptional effectors (reviewed in [41]). These tools allow researchers to test enhancer function in the native chromosomal context (i.e. in situ) to circumvent some of the limitations associated with MPRA, and to create animal or cell line models for selected alleles.

Pooled in situ approaches combine Cas9 nuclease with libraries of single guide RNAs (sgRNAs) that densely tile a target locus to map functional regulatory elements that are required for target gene expression. One of the first uses of this approach targeted DHSs surrounding the *BCL11A* gene, and discovered several critical regions that, upon deletion, resulted in reduced expression of the *BCL11A* gene and a concomitant increase in expression of fetal haemoglobin (which is normally repressed by *BCL11A*). Within these critical regions, the authors also identified a binding site for the key erythroid transcription factor GATA1 [42]. In this example, maps of erythroid-specific chromatin accessibility narrowed the region of interest of potentially important sites, but a systematic knockout screen was needed to define the regulatory grammar.

Small indels (insertion or deletion of bases) induced by non-homologous end-joining may not be sufficient to disrupt enhancer function; therefore, more recent approaches use pairs of sgRNAs in each cell to delete larger DNA fragments [43, 44]. An alternative strategy to investigate the impact of regulatory variants using in situ mutagenesis involves the introduction of precise single-nucleotide mutations at the target locus, which can be achieved by providing an exogenous DNA template with desired mutations for homology-directed repair.

Current pooled in situ mutagenesis approaches require a functional phenotype that can be perturbed and selected for. One such example is the expression level of an endogenous gene (e.g. fetal haemoglobin in ref. [42]), or a tagged gene product (e.g. GFP). However, it is difficult to apply these schemes to broad collections of enhancers, since the vast majority of their target genes are not known, or even if they were, would require laborious construction of many bespoke reporter cell lines. A recently developed approach, MOsaic Single-cell Analysis by Indexed CRISPR Sequencing (MOSAIC-seq) provides a promising general alternative for in situ enhancer mutagenesis, which combines the targeting of a dCas9-Krüppel associated box (KRAB) transcriptional repressor to candidate regulatory loci (i.e. CRISPR interference, or CRISPRi), with a read-out provided by single-cell RNA sequencing (scRNA-seq) to measure the resulting change in gene expression [45]. Besides conducting a proof-of-principle experiment by targeting the β -globin locus in K562 cells, the authors demonstrated its utility by targeting constituent enhancers within 15 different super-enhancers to dissect the relative contribution of each constituent on target gene expression. Establishing appropriate cellular models and read-outs remains a challenge for applying these techniques to type 2 diabetes, but, nevertheless, they hold promise for finely mapping the individual *cis*-regulatory sites and establishing their grammar.

Beyond the CRISPR modification and expression profiling experiments described above, proximity in the 3D chromatin environment of the nucleus provides another signal to pair *cis*-regulatory SNPs with their target genes. One such example in islets is a C3 region and *ISL* gene promoter [21]. A potential gold-standard for linking a candidate *cis*-regulatory SNP to its target gene would be the observation of consistent results across these approaches, from statistical association to experimental profiling and perturbation.

***Cis*-regulatory elements operate in a context-specific manner**

Factors that modulate the nuclear *trans* environment (e.g. transcription factor abundance and localisation) greatly influence how cells execute *cis*-regulatory programs. Such factors could

be intrinsic properties of different cell types established during development, or could be modulated by extrinsic stimuli, such as stress or hormone signalling. However, most functional genomic maps and reporter screens carried out to date have been obtained under steady-state (or basal) conditions. Therefore, integrating these maps and screens with developmental or treatment-induced dynamics represents an important direction for future studies. Here we present examples that illustrate how studying the impact of genetic variants under the proper context may be crucial for revealing functional convergence of disease-associated variants (Fig. 3a).

Environmental perturbation may be required to reveal the activity of some regulatory elements. For example, one study described ‘latent enhancers’ in mouse bone marrow-derived macrophages, which under basal conditions do not exhibit either histone marks typically associated with enhancer or chromatin accessibility but rapidly acquire these marks in response to an inflammatory agent (lipopolysaccharides [LPS]) or inflammatory cytokines such as IL-4 and IFN γ [46]. Similar observations were reported in human primary monocytes upon stimulation, and genetic associations were uncovered, with gene expression which varied when different immune stimuli were applied, such that both treatment and genotype interacted to affect gene expression [47]. In a particularly striking example, at one SNP associated with *HIP1* expression, the direction of association reversed under treatment, with the selected allele negatively correlated with *HIP1* expression in unstimulated cells but positively correlated after stimulation with LPS. Another example from the same study highlighted how dynamic gene expression kinetics required selecting the proper experimental time points: when cells were stimulated with LPS for 2 h, SNP rs2275888 was only associated with expression of one gene; however, the same SNP became associated with expression of five others after 24 h of stimulation.

Genome-wide regulatory maps made under different treatments illustrate the widespread impact of environmental exposures on gene regulation. For example, analyses in the mouse liver showed that 24 h of fasting induced changes in chromatin accessibility and H3K27ac signals around thousands of DHSs located nearby fasting-induced genes. Combining RNA-seq and ChIP-seq analyses for key transcription factors whose motifs were enriched within fasting-induced DHSs, the authors identified the glucocorticoid receptor as a critical factor that makes fasting-induced enhancers accessible so that other factors such as cAMP responsive element binding protein 1 (CREB1) can bind and activate gluconeogenesis programs in the liver [48].

While the above studies clearly highlight the importance of studying gene regulation under diverse environmental conditions, the application of this emerging concept is still limited in the diabetes genomics literature; a few recent examples are described here. To identify glucose-responsive regulatory elements in pancreatic islet beta cells, the INS-1E rat pancreatic

islet beta cell line was treated with glucose for 2 and 12 h and genome-wide changes in occupancy of MED1 protein (a subunit of the mediator complex that is involved in long-range interaction between enhancers and promoters), DHS and enhancer RNA transcription were measured [49]. Clustering analysis based on temporal dynamics identified six different patterns, which correlated with temporal dynamics of nearby glucose-responsive genes in their RNA-seq data. Motif enrichment analyses within these glucose-responsive regulatory elements identified the motif for carbohydrate response element binding protein (ChREBP), a transcription factor that was previously implicated in glucose-induced gene regulation in pancreatic islet beta cells. Two recent studies focused on a type 2 diabetes variant (rs508419) that overlaps with a skeletal muscle-specific promoter region at the *ANK1* locus [28, 50]. Human skeletal muscle eQTL data indicate that risk allele dosages result in higher *ANK1* expression [28]. Testing of the SNP region by luciferase reporter assays in the C2C12 mouse skeletal muscle myoblast cell line showed that the risk allele exhibited higher promoter activity than the non-risk counterpart [50]. Interestingly, however, the researchers were able to detect impairment only when they treated cells with insulin; under basal conditions, increased *ANK1* protein did not affect glucose uptake. Of note, prior islet eQTL studies showed that the risk allele of the same variant (rs508419) is associated with reduced expression of the transcription factor *NKX6-3* [23, 25], representing a tissue-dependent effect of regulatory variants, and potentially more complicated genetic architecture at this locus that is yet to be revealed. These examples highlight the importance and the challenges of modelling environmental stimuli in functional genomic studies of diabetes.

Conclusions

GWAS continue to identify genomic loci contributing to type 2 diabetes risk; however, interpretation of these signals remains challenging because most GWAS variants occur outside protein-coding genes. In recent years, massively parallel sequencing, high-throughput reporter assays and CRISPR gene editing technologies have quickly become indispensable tools for researchers to further understand the molecular basis of complex human diseases such as type 2 diabetes. In this review, we have considered how these approaches may be employed to further resolve GWAS-detected loci to identify individual variants and their functional effects. While the data generated so far have provided deeper insight into the gene regulation of type 2 diabetes risk variants, our understanding of the tissue specificity of these variants, and their interplay with environmental stimuli remains limited. Since enhancers integrate and transduce environmental signals to execute gene expression programs, studying the impact of genetic variants under diverse conditions will be crucial for furthering our understanding of disease-associated

variants. Moving forward, we believe that generating functional annotations in different environmental contexts and genetic perturbations will help partition swathes of GWAS signals into coherent, tissue-specific subsets to shed light on underlying pathophysiology. In summary, by employing the approaches discussed, additional convergent functional contexts are likely to emerge, and this information would enable higher-resolution patient stratification and determination of individualised risk.

Acknowledgements We thank members of the Kitzman and Parker laboratories, and associated collaborators, for invaluable discussions. We apologise in advance to authors whose work we were unable to cite or discuss because of space limitations.

Funding Work in the laboratories of SCJP is supported by the American Diabetes Association Pathway to Stop Diabetes Initiator Award 1-14-INI-07 (SCJP) and NIH/NIDDK grants R00 DK099240 and R01 DK117960 (SCJP).

Duality of interest The authors declare that there is no duality of interest associated with this manuscript.

Contribution statement All authors were responsible for drafting the article and revising it critically for important intellectual content. All authors approved the version to be published.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. International Diabetes Federation (2017) IDF diabetes atlas, 8th edn. IDF, Brussels Available from www.diabetesatlas.org
2. DeFronzo RA, Ferrannini E, Groop L et al (2015) Type 2 diabetes mellitus. *Nat Rev Dis Primers* 1:15019
3. Mahajan A, Taliun D, Thurner M et al (2018) Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *bioRxiv*. <https://doi.org/10.1101/245506>
4. Morris AP, Voight BF, Teslovich TM et al (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 44(9):981–990. <https://doi.org/10.1038/ng.2383>
5. Type 2 Diabetes Knowledge Portal. Available from www.type2diabetesgenetics.org/gene/geneInfo/WFS1. Accessed 5 November 2018
6. Thomsen SK, Gloyn AL (2017) Human genetics as a model for target validation: finding new therapies for diabetes. *Diabetologia* 60(6):960–970. <https://doi.org/10.1007/s00125-017-4270-y>
7. Smemo S, Tena JJ, Kim K-H et al (2014) Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507(7492):371–375. <https://doi.org/10.1038/nature13138>
8. Claussnitzer M, Dankel SN, Kim K-H et al (2015) FTO obesity variant circuitry and adipocyte browning in humans. *N Engl J Med* 373(10):895–907. <https://doi.org/10.1056/NEJMoa1502214>
9. Khera AV, Chaffin M, Aragam KG et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 50(9):1219–1224. <https://doi.org/10.1038/s41588-018-0183-z>
10. Udler MS, Kim J, von Grotthuss M et al (2018) Type 2 diabetes genetic loci informed by multi-trait associations point to disease

- mechanisms and subtypes: A soft clustering analysis. *PLoS Med* 15(9):e1002654. <https://doi.org/10.1371/journal.pmed.1002654>
11. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74. <https://doi.org/10.1038/nature11247>
 12. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W et al (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518:317–330
 13. Kouzarides T (2007) Chromatin modifications and their function. *Cell* 128(4):693–705. <https://doi.org/10.1016/j.cell.2007.02.005>
 14. Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 28(8):817–825. <https://doi.org/10.1038/nbt.1662>
 15. Parker SCJ, Stitzel ML, Taylor DL et al (2013) Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A* 110(44):17921–17926. <https://doi.org/10.1073/pnas.1317023110>
 16. Maurano MT, Humbert R, Rynes E et al (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337(6099):1190–1195. <https://doi.org/10.1126/science.1222794>
 17. Farh KK-H, Marson A, Zhu J et al (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518(7539):337–343. <https://doi.org/10.1038/nature13835>
 18. Trynka G, Sandor C, Han B et al (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* 45(2):124–130. <https://doi.org/10.1038/ng.2504>
 19. Gaulton KJ, Nammo T, Pasquali L et al (2010) A map of open chromatin in human pancreatic islets. *Nat Genet* 42(3):255–259. <https://doi.org/10.1038/ng.530>
 20. Stitzel ML, Sethupathy P, Pearson DS et al (2010) Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab* 12(5):443–455. <https://doi.org/10.1016/j.cmet.2010.09.012>
 21. Pasquali L, Gaulton KJ, Rodríguez-Seguí SA et al (2014) Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* 46(2):136–143. <https://doi.org/10.1038/ng.2870>
 22. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10(12):1213–1218. <https://doi.org/10.1038/nmeth.2688>
 23. Varshney A, Scott LJ, Welch RP et al (2017) Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc Natl Acad Sci U S A* 114(9):2301–2306. <https://doi.org/10.1073/pnas.1621192114>
 24. Smith SB, Qu H-Q, Taleb N et al (2010) Rfx6 directs islet formation and insulin production in mice and humans. *Nature* 463(7282):775–780. <https://doi.org/10.1038/nature08748>
 25. van de Bunt M, Manning Fox JE, Dai X et al (2015) Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. *PLoS Genet* 11(12):e1005694. <https://doi.org/10.1371/journal.pgen.1005694>
 26. Civelek M, Wu Y, Pan C et al (2017) Genetic regulation of adipose gene expression and cardio-metabolic traits. *Am J Hum Genet* 100(3):428–443. <https://doi.org/10.1016/j.ajhg.2017.01.027>
 27. GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group et al (2017) Genetic effects on gene expression across human tissues. *Nature* 550:204–213
 28. Scott LJ, Erdos MR, Huyghe JR et al (2016) The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nat Commun* 7(1):11764. <https://doi.org/10.1038/ncomms11764>
 29. Schaid DJ, Chen W, Larson NB (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* 19(8):491–504. <https://doi.org/10.1038/s41576-018-0016-z>
 30. Banerji J, Rusconi S, Schaffner W (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27(2):299–308. [https://doi.org/10.1016/0092-8674\(81\)90413-X](https://doi.org/10.1016/0092-8674(81)90413-X)
 31. Stitzel ML, Kycia I, Kursawe R, Ucar D (2015) Transcriptional regulation of the pancreatic islet: implications for islet Function. *Curr Diab Rep* 15(9):66. <https://doi.org/10.1007/s11892-015-0635-0>
 32. Inoue F, Ahituv N (2015) Decoding enhancers using massively parallel reporter assays. *Genomics* 106(3):159–164. <https://doi.org/10.1016/j.ygeno.2015.06.005>
 33. Kalita CA, Moyerbrailean GA, Brown C, Wen X, Luca F, Pique-Regi R (2018) QuASAR-MPRA: accurate allele-specific analysis for massively parallel reporter assays. *Bioinformatics* 34(5):787–794. <https://doi.org/10.1093/bioinformatics/btx598>
 34. Ulirsch JC, Nandakumar SK, Wang L et al (2016) Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* 165(6):1530–1545. <https://doi.org/10.1016/j.cell.2016.04.048>
 35. Ernst J, Melnikov A, Zhang X et al (2016) Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat Biotechnol* 34(11):1180–1190. <https://doi.org/10.1038/nbt.3678>
 36. Arnold CD, Gerlach D, Stelzer C et al (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339(6123):1074–1077. <https://doi.org/10.1126/science.1232542>
 37. Liu S, Liu Y, Zhang Q et al (2017) Systematic identification of regulatory variants associated with cancer risk. *Genome Biol* 18(1):194. <https://doi.org/10.1186/s13059-017-1322-z>
 38. Long HK, Prescott SL, Wysocka J (2016) Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* 167(5):1170–1187. <https://doi.org/10.1016/j.cell.2016.09.018>
 39. Zabidi MA, Arnold CD, Schernhuber K et al (2015) Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* 518(7540):556–559. <https://doi.org/10.1038/nature13994>
 40. Arnold CD, Zabidi MA, Pagani M et al (2017) Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nat Biotechnol* 35(2):136–144. <https://doi.org/10.1038/nbt.3739>
 41. Montalbano A, Canver MC, Sanjana NE (2017) High-throughput approaches to pinpoint function within the noncoding genome. *Mol Cell* 68(1):44–59. <https://doi.org/10.1016/j.molcel.2017.09.017>
 42. Canver MC, Smith EC, Sher F et al (2015) BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* 527(7577):192–197. <https://doi.org/10.1038/nature15521>
 43. Gasperini M, Findlay GM, McKenna A et al (2017) CRISPR/Cas9-mediated scanning for regulatory elements required for hprt1 expression via thousands of large, programmed genomic deletions. *Am J Hum Genet* 101(2):192–205. <https://doi.org/10.1016/j.ajhg.2017.06.010>
 44. Diao Y, Fang R, Li B et al (2017) A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Methods* 14(6):629–635. <https://doi.org/10.1038/nmeth.4264>
 45. Xie S, Duan J, Li B et al (2017) Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Mol Cell* 66:285–299.e5
 46. Ostuni R, Piccolo V, Barozzi I et al (2013) Latent enhancers activated by stimulation in differentiated cells. *Cell* 152(1–2):157–171. <https://doi.org/10.1016/j.cell.2012.12.018>
 47. Fairfax BP, Humburg P, Makino S et al (2014) Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* 343(6175):1246949. <https://doi.org/10.1126/science.1246949>

48. Goldstein I, Baek S, Presman DM et al (2017) Transcription factor assisted loading and enhancer dynamics dictate the hepatic fasting response. *Genome Res* 27(3):427–439. <https://doi.org/10.1101/gr.212175.116>
49. Schmidt SF, Madsen JGS, Frafjord KØ et al (2016) Integrative genomics outlines a biphasic glucose response and a ChREBP-ROR γ axis regulating proliferation in β cells. *Cell Rep* 16(9): 2359–2372. <https://doi.org/10.1016/j.celrep.2016.07.063>
50. Yan R, Lai S, Yang Y et al (2016) A novel type 2 diabetes risk allele increases the promoter activity of the muscle-specific small ankyrin 1 gene. *Sci Rep* 6(1):25105. <https://doi.org/10.1038/srep25105>