Minireview

Challenges and progress in interpretation of non-coding genetic variants associated with human disease

Yizhou Zhu[1], Cagdas Tazearslan[1] and Yousin Suh[1,2,3]

[1]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA; [2]Department of Ophthalmology & Visual Sciences, Albert Einstein College of Medicine, Bronx, NY 10461, USA; [3]Department of Medicine, Albert Einstein College of Medicine, Bronx, NY 10461, USA

Corresponding author: Yousin Suh. Email: yousin.suh@einstein.yu.edu

Impact statement

Most signals from genome-wide association studies (GWASs) map to the non-coding genome, and functional interpretation of these associations remained challenging. We reviewed recent progress in methodologies of studying the non-coding genome and argued that no single approach allows one to effectively identify the causal regulatory variants from GWAS results. By illustrating the advantages and limitations of each method, our review potentially provided a guideline for taking a combinatorial approach to accurately predict, prioritize, and eventually experimentally validate the causal variants.

Abstract

Genome-wide association studies have shown that the far majority of disease-associated variants reside in the non-coding regions of the genome, suggesting that gene regulatory changes contribute to disease risk. To identify truly causal non-coding variants and their affected target genes remains challenging but is a critical step to translate the genetic associations to molecular mechanisms and ultimately clinical applications. Here we review genomic/epigenomic resources and in silico tools that can be used to identify causal non-coding variants and experimental strategies to validate their functionalities.

Keywords: Causal variants, enhancers, functional genomics, genome-wide association studies, non-coding variants, variant annotation

***Experimental Biology and Medicine* 2017; 242: 1325–1334. DOI: 10.1177/1535370217713750**

Introduction

As of February 2017, there are about 3000 genome-wide association studies (GWASs) reporting more than 30,000 unique SNP-disease associations.[1,2] While many of these associated variants confer a rather small increase in risk individually, recent meta-analysis has shown that as a group, targets based on evidence from GWAS-associated loci are twice as likely to be therapeutically valid as those that are not.[3] Thus, it is important to delineate the mechanisms underlying disease-associated sequence variants at a molecular level. Biological insights can then be utilized to improve clinical outcomes, including developing effective strategies for disease prevention and/or therapeutics.

Interpretation of GWAS results, however, is challenging due to the fact that most variants found to be associated with disease lie outside of protein-coding regions. This observation remains true even after fine mapping around the associated loci.[4] These results suggest that disease-associated variants impose risk by altering functional DNA elements that regulate gene expression. Indeed, variation in gene expression has been shown to be highly heritable and a significant determinant of human disease susceptibility.[5] However, GWAS detect only statistical associations, not functional signals, resulting in ambiguity in determining the causal genes for associated non-coding variants. Thus, identifying target genes affected by non-coding variants remains challenging. A common empirical practice is to assign the non-coding GWAS variants to the nearest gene, which may not necessarily reflect the real situation.[6,7] In certain cases, this issue can be solved by incorporating complementary information, such as QTL and tissue-specific expression patterns of local genes.[8,9] When such information is not available, determination of the causal gene is more difficult. Furthermore, GWAS take advantage of linkage disequilibrium (LD) in the genome, a property of non-random chromosomal segregation, to cost-efficiently estimate the genotype with a relatively small number of tag SNPs. As the trade-off, all variants linked to the significantly associated disease tag SNPs can potentially be responsible for the detected association, while only a few of them play functional causal roles. Therefore, identification of truly causal GWAS variants and elucidating how they cause

dysregulation of target gene expression remain a significant challenge in the postgenomic era.

Non-coding variants may play regulatory roles for gene expression through multiple mechanisms. Variants in promoters can impose direct impact on transcription initiation and elongation.[10] Intronic and UTR variants can potentially affect the property of mRNAs, leading to altered stability or splicing patterns. In addition, variants may alter function or expression of multiple classes of non-coding RNAs, including long non-coding RNAs and small RNAs such as micro RNAs and small nucleolar RNAs.[11] Integrated genomic and epigenomic annotation studies suggested that GWAS variants were rather enriched in evolutionarily conserved putative enhancer regions, suggesting the significant role of enhancer variants in conferring disease risks.[12–14]

Variants in enhancers have predictable function through modulation of transcription factor (TF) binding motifs. However, the large size of TF pool and highly tissue- and context-dependent TF regulation hurdles the complete knowledge of function of enhancers and regulatory variants in enhancers. In this review, we will focus on enhancer GWAS variants. We will discuss current progress towards *in silico* and experimental identification, and validation of causal variants that interfere with enhancer function, thereby conferring disease risk through dysregulation of gene expression.

Enhancers

Enhancers are the principal regulatory components of the genome that enable cell-type and cell-state specificities of gene expression. Enhancers were initially defined as DNA elements that act over a distance to positively regulate expression of protein encoding target genes, independent of orientation and direction with respect to the target gene promoters.[15] The human genome is estimated to encode ~1 million enhancer elements and distinct sets of approximately 30,000–40,000 enhancers are active in a particular cell type,[16,17] vastly outnumbering protein-coding genes and promoters. Enhancer activation entails the presence of specific recognition sequences required for the cooperative recruitment of TFs that initially activate and subsequently permit signal-dependent regulation of gene expression in a spatial and temporal fashion.[18] By contrast, genetic variations in enhancer sequences that alter TF binding would predispose to 'improper' gene expression and ultimately susceptibility to diseases.[19,20] The enhancer-bound TFs facilitate chromatin accessibility by recruitment of nucleosome remodeling complexes with the core 80–120 basepairs representing the sites for binding of the activating/regulatory TFs.

Genomic annotation of enhancers has been greatly facilitated by the development of high-throughput methods, providing surrogate markers for enhancer activity at an unprecedented resolution.[12,21–23] Enhancers are typically characterized by the presence of histone modifications (detected by ChIP-seq) such as H3K27Ac and H3K4me1/2.[24] Notably, the H3K27Ac positive enhancers showed high enhancer activity and co-occupancy with linage-specific TFs.[25] Thus, it has been proposed that H3K27Ac distinguished active enhancers from the primed or poised ones. Binding of TFs to enhancers results in depletion of nucleosome, making the region detectable by DNase-seq and ATAC-seq.[26,27] In addition, active enhancers are also indicated by expression of enhancer RNAs (eRNAs), which can be detected by deep RNA-seq, Global Run-On Sequencing, or Cap Analysis Gene Expression (CAGE).[22,28,29] Recent studies suggested that eRNAs could play a role in chromatin looping for interaction with the target gene promoter.[30] Finally, enhancers are hypomethylated at CpG dinucleotides, and hence can be detected by bisulfite sequencing.[31]

By collectively using these techniques, several epigenome consortia, such as ENCODE, Roadmap Epigenomics Project, and BLUEPRINT Hematopoietic Epigenome Project have had considerable achievements in identifying enhancers in a wide range of tissues and cell types.[12,32,33] These databases utilize standardized protocols to provide reproducible position information for enhancers, and hence have been applied in numerous meta-analyses studies. Since these databases provide 'surrogate' information on enhancer activity based on correlative evidence in steady states, it is critically important to conduct validation studies of the candidate enhancer elements and their GWAS variants within to test the functional relevance.

A key feature of enhancers is their ability to activate the transcription of a gene from a great distance. One classical example is a distal enhancer, when mutated is responsible for preaxial polydactyly.[34] The enhancer is located at intronic region of Limb Development Membrane Protein 1, yet has been strikingly found to be involved in regulation of sonic hedgehog located 1 Mb away, the true causal gene for the disease phenotype. A significant challenge, thus, is to define the targets of enhancers. Currently, non-coding GWAS variants are assigned to the nearest gene. However, the recent studies developing contact maps on a genome-wide scale indicates that many enhancer-like regions skip over the nearest gene and make contacts with more distant targets.[35] Therefore, accurate interpretation of the effects of non-coding genetic variation requires methods that allow correct assignment of regulatory elements to their target genes. Indeed, by employing a combined approach of expression quantitative trait loci (eQTL), circular chromatin conformation capture (4C), and genome editing, it was found that IRX3 and IRX5 were more plausible target genes of the obesity-associated variants in the FTO locus.[6,7] These two homeobox TFs are located 0.5 and 1 million bp away, respectively, from the GWAS signal. It has been demonstrated that a functional enhancer variant in the FTO GWAS locus (located in FTO intron 1) disrupts binding of a transcriptional repressor (ARID5B) in mesenchymal preadipocyte-specific enhancer, resulting in upregulation of both IRX3 and IRX5, which in turn shifts cell fate of adipocyte precursor toward white adipocyte and lipid storage.

Whether enhancers can be functionally classified remains an open hot topic. Answering this critical question belongs to the field of machine learning. Several pioneering studies reported that TF binding motifs were predictors of enhancer activity and tissue specificity. For example,

Yanez-Cuna *et al.*[36] reported that GATA and E-box motifs were functionally important for Drosophila S2 cell-specific enhancer function, whereas Ahmad *et al.*[37] found that Myb was crucial in activity of contractile cardial cells. On the other hand, Young's lab proposed the concept of super-enhancers, which were characterized by densely clustered enhancers and occupied with high levels of mediator complex.[38,39] These enhancers are believed to play central roles in cell fate determination, binding of lineage master regulators, and cell type-specific gene expression. Multiple diseases have GWAS associations in super-enhancer regions, such as Alzheimer's disease and multiple sclerosis.[40] More recently, direct evidence suggested that super-enhancers are involved in specific disease processes such as oncogenesis.[41–43] For example, in 8q24 locus, the non-coding regions near the *MYC* gene gained distinct super-enhancers in several cancer cell lines (Figure 1), indicating a possible model where distinct tissue-specific super-enhancers were responsible for misregulation of the oncogene in different cancers.

In summary, to identify regulatory variants in enhancers and to test their functionality and disease relevance require multifaceted and integrated approaches that capture the highly dynamic nature of enhancer function. These include *in silico* analysis to annotate and predict potentially causal enhancer variants and specific experimental systems to validate the role of selected enhancer variants in conferring disease risk.

In silico analysis: Prediction of functional enhancer variants

Multiple meta-analyses studies suggested enrichment of GWAS variants in close vicinity of enhancers.[44,45] Notably, the enrichment seemed to preferentially occur in disease-related cell types. For example, risk variants of type 1 diabetes and other autoimmune diseases show a significant enrichment in lymphocyte-specific enhancers,[46,47] whereas variants—associated with electrocardiographic-related traits and insulin levels were found to be enriched in super-enhancers specific to heart and adipose tissue, respectively.[40] Moreover, Alzheimer's disease-associated variants are found to be enriched in immune-cell-specific enhancers rather than neuron-specific ones, suggesting that immune processes may play a role in the pathogenesis of the disease.[48] Taken together, these studies not only provided evidence that at least a substantial portion of GWAS variants contributes to disease risks by interfering with enhancers, but also offered biological insights into the pathophysiology of complex diseases involving multiple cell types.

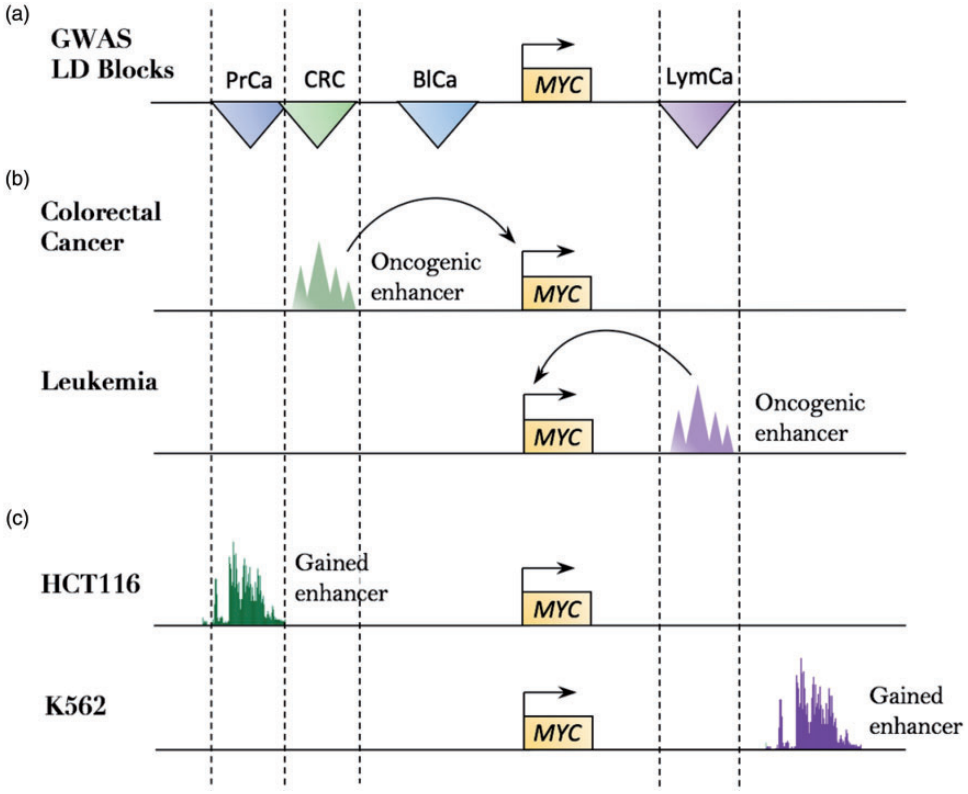

Figure 1 The activity of enhancers and super-enhancers is cell- and tissue-specific. (a) Landscapes of GWAS associations in the neighboring genomic regions of *MYC* locus in Chr. 8q24. PrCa: prostate cancer, CRC: colorectal cancer, BlCa: bladder cancer, LymCa: and lymphoma. (b) A hypothetical model that may explain the genetic association patterns: different types of cancer cells may gain tissue-specific oncogenic enhancers/super-enhancers, resulting in misregulation of the *MYC* oncogene in a tissue-specific manner. (c) However, in actual case, the gained super-enhancers in tumors were found outside the corresponding LD regions in colorectal cancer (HCT116) and leukemia K562) cell lines. This may suggest a complex mechanism underlying the GWAS association, such as the presence of functional variants that alter the enhancer–target gene interaction network rather than directly affecting enhancer's capability to facilitate promoter activity.
GWAS: genome-wide association studies. (A color version of this figure is available in the online journal.)

A central question in enhancer annotation is how to precisely identify the TF binding regions. While most enhancers have predicted lengths of kilobase (kb), the actual region bound by TFs might be much smaller in size. In fact, the CAGE study has demonstrated that enhancers produce bidirectional eRNAs, and the region in between these transcripts, typically ~200 bp in length, possesses the highest enhancer activities.[22] The underlying message could be that even if a variant falls into an enhancer region, there is a good chance that it is not functional. To reduce the false positive predictions, a common strategy is to consider whether the variant falls into specific TF binding motifs. However, motif prediction by probability matrix is also prone to high false positive rates, as the motifs are short (typically <10 bp), and many TFs allow sequence variations in certain positions of the motifs. The motif prediction can be largely improved by overlaying with ChIP-seq data.[49] However, the assay can only probe for one out of thousands of TFs at a time and is subjected to technical limitations such as very large number of required cells and availability of antibodies. Altogether, the key barriers to the high accuracy mapping of causal variants are in biology rather than bioinformatics.

Despite the noted challenges, considerable efforts have been made for functional annotations of non-coding variants (Table 1). Databases such as RegulomeDB and HaploReg have recently been developed by incorporating epigenomic annotation from multiple sources and attempted to provide comprehensive information of underlying enhancers for the query non-coding variants.[54,55] FORGE is a convenient tool that evaluates tissue-specific enhancer enrichment of a query GWAS SNP list.[56] Other tools with similar principles but alternative scoring algorithms, such as GWAVA and CADD, allow prioritization from a large list of variants.[57,58] More recent studies including Finucane *et al.*[13] and Farh *et al.*[46] considered LD in scoring the likelihood of causality of variants. Although intuitively variants in high LD are more likely to be causal, the inference can easily be confounded by other factors: r^2 is biased to variants with similar allele frequencies and multiple linked variants may have combinatorial effects. Moreover, multiple candidate variants will still be inevitably sharing probabilities of being the causal variant in high LD regions. To reduce the LD background, one interesting approach is to identify conserved GWAS associations from distinct ethnicity background, a method known as trans-ethnic analysis.[59] While the method has been proven to improve the overall prediction of causality, such approach may lose certain ethnicity-specific GWAS signals originating from ethnic-specific heterozygosity of the region.

In addition to canonical mapping methods based on genetic and epigenetic information, an alternative approach to examine the variant functionality is through quantitative trait loci (QTLs) analysis. This includes QTL with eQTL, splicing (sQTL), methylation (meQTL), protein/proteome, and all epigenomic signals from DNase and ChIP-seq assays.[60–62] Due to lower requirement for input material, currently more information is available for eQTL and meQTL. An example of rapidly expanding databases of eQTL and sQTL is Genotype-Tissue Expression portal. The database currently includes 53 tissue types from 554 donors (449 genotyped), and the project ultimately aims to profile transcriptome data from >900 genotyped individuals.[51] The National Heart, Lung, and Blood Institute also presented Genome-Wide Repository of Associations, a collection database for all published genotype–phenotype association results including GWAS, eQTL, and meQTL data.[52] The database was updated in 2015 to V2.0, collecting about 8.87 million SNP associations from 2082 studies.[1] Since enhancer function is often tissue dependent, such comprehensive databases are invaluable in identifying variants correlated with differences in transcription levels. Comprehensive eQTL data also allows proper pairing between variants and their target genes. In the FTO locus, for example, variants associated with obesity do not show association with the expression of FTO but with IRX3 and IRX5 in multiple cell types including the primary human preadipocytes.[7] Indeed, meta-analyses showed that eQTLs were gene centric and enriched in both putative regulatory elements and GWAS SNPs, suggesting a possible general model where GWAS variants modulate enhancer function and affect nearby transcribed genes.[63,64] One general concern for QTL studies is the extraordinarily high dimensionality of the data. For example, the total number of parameters is equal to the product of the number of

Table 1 Methods for studying the functionality of non-coding GWAS variants. *In silico* approaches for functional enhancer variants. After imputation, list of candidate GWAS variants can be prioritized based on predicted function from the publicly available data resources. Candidate target genes, cell/tissue types, and mechanisms of TF binding interruption can be inferred to assist design of specific validation assays. See text for detailed explanation for each method.

Methods	Target gene	Functional cell type	Mechanism	Causal variant	Database
Enhancer annotation		✓		✓	ENCODE, Roadmap, BLUEPRINT[12,32,33]
TF ChIP		✓	✓	✓	ENCODE
Motif prediction			✓	✓	JASPAR, ENCODE[50]
eRNA (CAGE)	✓	✓		✓	FANTOM5[22]
eQTL, sQTL, meQTL	✓	✓			GTEx GRASP[51,52]
Hi-C	✓	✓			Hi-C browser[53]

GWAS: genome-wide association studies; TF: transcription factor.

transcripts and variants. This essentially prohibits the genome-wide multiple corrections and forces correction strategies based on local chromatin sections, which in turn results in discrepancies in definition of significant associations among different studies.[63]

Experimental strategies: Validation of causal enhancer variants

One of the most critical, but often lagging, steps in identifying and testing the functional relevance of the non-coding variants detected in GWAS is the functional validation of the candidate enhancer variants. With the advent of gene-editing tools and high-throughput sequencing, it is now more feasible to accomplish this goal on a genomic scale. Here we highlight several novel technologies that can be used in functionalization of enhancer variants (summarized in Table 2), thereby establishing the causality of GWAS variants in conferring disease risks.

The massive parallel reporter assay (MPRA) allows examination of a large number of enhancers and enhancer variants within a single experiment.[65–67] Typically, thousands of candidate enhancer regions are synthesized and cloned into a mammalian expression vector, where the co-synthesized barcodes or the enhancers themselves are transcribed as identifiers for each construct. The mixed reporter library is then transfected to a cell line, and the vector DNA and the reporter RNA transcripts are individually collected and sequenced. The enhancer activity of the constructs can then be represented by the read count ratio between RNA and DNA. Several groups have reported success in using MPRA to identify causal GWAS variants.[68,69] To reduce a considerable level of background variation and improve the consistency of the results, it is recommended to increase the number of barcodes per construct and replication experiments.[70]

Since enhancer function depends on the local chromatin context, genome-editing tools are indispensable for studies of enhancer mechanisms in the endogenous genome. CRISPR/Cas9 is a recently developed technology that allows efficient and scalable targeted genome editing. CRISPR/Cas9 recognizes target sequence by binding to a roughly 20 basepair-long complementary guide RNA, allowing highly cost-efficient and simplified assay designs compared to its predecessors, zinc finger nuclease and transcription activator-like effector nuclease, which require full length synthesis of DNA-binding domain for each target. While the function of Cas9 can be best characterized as a sequence-specific endonuclease, the CRISPR/Cas9 technology is highly versatile in applications for enhancer studies. Depending on the design, wild-type Cas9 can facilitate targeted sequence modification through non-homologous end joining, such as complete deletion of chromatin segments (knock-out), or site-specific DNA integration (knock-in) to remove enhancers or modify their function, respectively.[30,71,72] The nuclease-disabled Cas9 (dCas9) has been applied to manipulate target enhancers by coupling with specific TFs. In a pioneer study from Gilbert *et al.*,[73] dCas9 was fused with a transcriptional activator VP64 or a repressive KRAB domain to activate or repress the activity of particular enhancers, respectively, to determine their roles in tumor cell proliferation and myeloid differentiation. Additionally, His-tagged dCas9 can also be used as a sequence indicator to pull down specific enhancers and study its protein composition by mass spectrometry, a valuable approach for identifying the responsible TFs when a candidate causal variant is given.[74,75]

While the CRISPR/Cas9-mediated gene editing has revolutionized the functional validation of enhancers, the throughput of the method is typically low, requiring prioritization of candidate variants by other approaches. One exception, though, is that if the phenotypic outcome is either related to cellular survival or detectable by cell sorting, it will be possible to design CRISPR-based screening assays by creating complex viral libraries and infecting cells with low density. The principles of such screening assays have been well demonstrated by several studies performing Genome-Scale CRISPR Knock-Out.[76,77] Recently, Horlbeck *et al.*[78] described an enhancer-version of the assay by utilizing dCas9 activator and inhibitor. In addition to the throughput, another concern, when performing genome editing, is the choice of a suitable model. Since most GWAS variants are associated with complex traits, *in vivo* studies should be intuitively preferred. However, performing studies using animal models is usually limited due to poor conservation of non-coding sequences between species. Studies of enhancers across species suggested their conservation at a functional level rather than nucleotide sequence.[79,80] Therefore, although modeling the effect of a particular variant could be difficult, the underlying functional enhancer would be more likely conserved and available for *in vivo* studies. Epigenetic annotations of regulatory elements in model organisms, such as modENCODE and mouse ENCODE, were available to search for enhancer

Table 2 Methods for studying the functionality of non-coding GWAS variants. Experimental methods for validating the functionality of non-coding variants predicted from *in silico* analysis. See text for detailed explanation for each method

Methods	Target gene	Functional cell type	Mechanism	Causal variant	High throughput
Reporter assay (MPRA)		✓	✓	✓	✓
EMSA/ChIP		✓	✓	✓	✓
CRISPR/Cas9	✓	✓	✓	✓	
3C/4C/capture Hi-C	✓	✓	✓		✓

GWAS: genome-wide association studies.

candidates.[81,82] An alternative strategy is to perform functional studies in relevant human cell lines, although in such case the cell line and phenotypic output must be carefully selected to be relevant to the pathophysiology of disease with which non-coding variants are associated.

An accurate interpretation of the effects of non-coding genetic variation requires methods that allow correct assignment of regulatory elements with their target genes. A crucial method to correctly assign non-coding variants to target genes is chromatin conformation capture (3C) that delineate long- and short-range chromatin interactions. The original 3C was designed for detecting 'one-to-one' interaction (chromatin looping) between two sites on the chromosome. With advanced high-throughput sequencing, its derivatives, 4C, 5C, and Hi-C, were developed to, respectively, characterize 'one-to-all,' 'many-to-many,' and 'all-to-all' interactions that present more comprehensive information of high order chromatin structure.[83] In 2014, Rao *et al*.[84] greatly improved the resolution of Hi-C to the kilobase level with a protocol utilizing *in situ* restriction enzyme digestion. The study showed that the chromatin was organized in unit of blocks, i.e. topology-associated domains (TADs), maintained by boundary proteins such as CTCF. Recent studies further indicated that these boundaries were responsible for restraining enhancer–promoter interactions within the TADs, and disruption of boundaries would cause abnormally gained interactions, which could be responsible for certain disease phenotypes such as limb malfunctions.[85–87] Taken together, the chromatin structure has emerged as a critical component in transcription regulation, and disease variants altering the chromatin interaction networks are equally likely to yield functional impacts as those interfering with enhancer machinery.

Although all chromatin looping assays are similar in principle, they can be classified into two classes, one of them being more qualitative and the other more quantitative. For 3C, original 4C, and 5C, the interaction libraries are amplified with pairs of specific primers, rendering read count quantification susceptible to PCR bias. Library amplification involving at least one random sonication end, including Hi-C, capture Hi-C, and recently available NG Capture-C and UMI-4C, enables sequencing deduplication during alignment and thus are more quantitative.[84,88–90] Clearly, the quantitative methods should be preferred for testing variant effects on chromatin interaction, since GWAS variants with moderate effects are less likely to cause all or none changes. Among these assays, Hi-C probes for the interactions at the genome-wide level and represents the most comprehensive information. However, since the total number of possible genome-wide interactions is gigantic (proportional to the square of available restriction enzyme sites), the available read count from each interaction is often too small for robust quantifications. Alternative methods such as promoter capture Hi-C were designed to overcome the issue and yielded high resolution chromatin interactomes.[88,91] Comparatively, UMI-4C probes for the interaction status of limited number of genomic regions of interest (viewpoints), but consequentially produces data with high sequencing depth and read counts (reported 10,000 for each viewpoint).[89] Thus, for examination of a particular variant or locus, better outcomes should be obtained with targeted approaches.

Summary and perspectives

In recent years, a remarkable progress has been made in methodologies of studying the non-coding genome. The expeditious advancements in techniques have been accompanied with rapid expansion of data resources and development of sophisticated prediction tools for functional characterization of non-coding variants. Still, no single approach allows one to effectively identify the causal enhancer variants from GWAS results. While increasingly comprehensive knowledge of the non-coding genome may eventually allow much simplified workflows for more effective interpretation of non-coding variants; currently, though, the best strategy seems to integrate the results from multiple methods to accurately predict, prioritize, and eventually experimentally validate the causal variants.

While our review focused on the current optimal strategy to identify a most likely causal non-coding variant, among many associated candidate variants, underlying GWAS signal it has been demonstrated that multiple variants in combination contribute to a GWAS signal.[92] In GWAS loci with multiple disease associations falling into distinct LD patterns, such as 8q24, presence of multiple causal variants is expected. An integrated *in silico* analysis followed by systematic experimental validation studies by step-wise, co-modulation of multiple variants will shed light on the role of multiple variants.

The study of the non-coding genome is also benefited from increasingly more complete clinical networks. A novel branch of association studies, phenotype-wide association studies (PheWAS), is rapidly developing along with the electronic medical records and genomics network.[93] In contrast to GWAS, PheWAS reports spectra of phenotypes associated to probed variants, providing insights into the phenotypic outcomes of genetic variations. Combining comprehensive medical records with genome-wide genetic, genomic, and epigenomic data available in human tissues banks will provide an invaluable platform for identifying disease-related epigenomic changes in the non-coding genome, especially for regulatory elements and sequences that are not conserved across species.[94]

With increasingly cost-effective high throughput sequencing, more association studies using whole genome sequencing (WGS) data will be available in the near future. The major motivation for large-scale WGS is to identify disease-associated rare variants, as demonstrated by several studies.[95–97] Rare variants that were causal for GWAS associations are expected to have much larger effect compared to common variants. However, the large number of rare variants and their requirement of large sample size to reach statistical power raised additional challenge for their functional characterization.[98] Still, it is predictable that a considerable fraction of these novel associations will fall in to the non-coding genome, demanding functional prediction tools with higher precision and functional assays with higher throughputs.

Authors' contributions: The article was drafted by YZ and critically revised by CT. YS revised and edited the manuscript.

ACKNOWLEDGEMENTS

This work was funded by NIH grants: AG017242, GM104459, and CA180126 (Suh).

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

1. Eicher JD, Landowski C, Stackhouse B, Sloan A, Chen W, Jensen N, Lien JP, Leslie R, Johnson AD. GRASP v2.0: an update on the genome-wide repository of associations between SNPs and phenotypes. *Nucleic Acids Res* 2015;**43**:D799–804
2. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014;**42**:D1001–6
3. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, Whittaker JC, Sanseau P. The support of human genetic evidence for approved drug indications. *Nat Genet* 2015;**47**:856–60
4. Wellcome Trust Case Control C, Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, Howson JM, Auton A, Myers S, Morris A, Pirinen M, Brown MA, Burton PR, Caulfield MJ, Compston A, Farrall M, Hall AS, Hattersley AT, Hill AV, Mathew CG, Pembrey M, Satsangi J, Stratton MR, Worthington J, Craddock N, Hurles M, Ouwehand W, Parkes M, Rahman N, Duncanson A, Todd JA, Kwiatkowski DP, Samani NJ, Gough SC, McCarthy MI, Deloukas P, Donnelly P. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* 2012;**44**:1294–1301
5. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nat Rev Genet* 2009;**10**:184–94
6. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gœmez-MarÚn C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF, Lee JH, Puviindran V, Tam D, Shen M, Son JE, Vakili NA, Sung HK, Naranjo S, Acemel RD, Manzanares M, Nagy A, Cox NJ, Hui CC, Gomez-Skarmeta JL, Nœbrega MA. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 2014;**507**:371–5
7. Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry JL, Puviindran V, Abdennur NA, Liu J, Svensson PA, Hsu YH, Drucker DJ, Mellgren G, Hui CC, Hauner H, Kellis M. FTO obesity variant circuitry and adipocyte browning in humans. *N Engl J Med* 2015;**373**:895–907
8. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JW, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E. Colocalization of GWAS and eQTL signals detects target genes. *Am J Hum Genet* 2016;**99**:1245–60
9. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, GTEx Consortium, Nicolae DL, Cox NJ, Im HK. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 2015;**47**:1091–98
10. Nurnberg ST, Rendon A, Smethurst PA, Paul DS, Voss K, Thon JN, Lloyd-Jones H, Sambrook JG, Tijssen MR, HaemGen Consortium, Italiano JE Jr, Deloukas P, Gottgens B, Soranzo N, Ouwehand WH. A GWAS sequence variant for platelet volume marks an alternative DNM3 promoter in megakaryocytes near a MEIS1 binding site. *Blood* 2012;**120**:4859–68
11. Hrdlickova B, de Almeida RC, Borek Z, Withoff S. Genetic variation in the non-coding genome: involvement of micro-RNAs and long non-coding RNAs in disease. *Biochim Biophys Acta* 2014;**1842**:1910–22
12. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74
13. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, Ripke S, Day FR, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, RACI Consortium, Purcell S, Stahl E, Lindstrom S, Perry JR, Okada Y, Raychaudhuri S, Daly MJ, Patterson N, Neale BM, Price AL. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 2015;**47**:1228–35
14. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B, Stahl E, Schizophrenia Working Group of the Psychiatric Genomics Consortium, SWE-SCZ Consortium, Kähler AK, Hultman CM, Purcell SM, McCarroll SA, Daly M, Pasaniuc B, Sullivan PF, Sullivan PF, Neale BM, Wray NR, Raychaudhuri S, Price AL, Schizophrenia Working Group of the Psychiatric Genomics Consortium, SWE-SCZ Consortium. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet* 2014;**95**:535–52
15. Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 1981;**27**:299–308
16. Bernstein BE, *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74
17. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo P, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30
18. ENCODE Project Consortium. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;**38**:576–89
19. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM, Pirruccello JP, Muchmore B, Prokunina-Olsson L, Hall JL, Schadt EE, Morales CR, Lund-Katz S, Phillips MC, Wong J, Cantley W, Racie T, Ejebe KG, Orho-Melander M, Melander O, Koteliansky V, Fitzgerald K, Krauss RM, Cowan CA, Kathiresan S, Rader DJ. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 2010;**466**:714–19
20. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, Frazer KA. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* 2011;**470**:264–68
21. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME. Widespread transcription at neuronal activity-regulated enhancers. *Nature* 2010;**465**:182–87
22. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jørgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, Negishi Y,

Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Müller F, FANTOM Consortium, Forrest AR, Carninci P, Rehli M, Sandelin A. An atlas of active enhancers across human cell types and tissues. *Nature* 2014;**507**:455–61
23. Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA, Glass CK, Rosenfeld MG, Fu XD. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 2011;**474**:390–4
24. Heintzman ND, *et al*. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 2007;**39**:311–8
25. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 2010;**107**:21931–6
26. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* 2015;**109**:21–29
27. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* 2010;**2010**:pdb prot5384
28. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* 2014;**46**:1311–20
29. Wu H, Nord AS, Akiyama JA, Shoukry M, Afzal V, Rubin EM, Pennacchio LA, Visel A. Tissue-specific RNA expression marks distant-acting developmental enhancers. *PLoS Genet* 2014;**10**:e1004610
30. Yang H, Wang H, Shivalila CS, Cheng AW, Shi L, Jaenisch R. One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* 2013;**154**:1370–9
31. Wiench M, John S, Baek S, Johnson TA, Sung MH, Escobar T, Simmons CA, Pearce KH, Biddie SC, Sabo PJ, Thurman RE, Stamatoyannopoulos JA, Hager GL. DNA methylation status predicts cell type-specific enhancer activity. *EMBO J* 2011;**30**:3028–39
32. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA. The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol* 2010;**28**:1045–8
33. Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, Bock C, Boehm B, Campo E, Caricasole A, Dahl F, Dermitzakis ET, Enver T, Esteller M, Estivill X, Ferguson-Smith A, Fitzgibbon J, Flicek P, Giehl C, Graf T, Grosveld F, Guigo R, Gut I, Helin K, Jarvius J, Küppers R, Lehrach H, Lengauer T, Lernmark Å, Leslie D, Loeffler M, Macintyre E, Mai A, Martens JH, Minucci S, Ouwehand WH, Pelicci PG, Pendeville H, Porse B, Rakyan V, Reik W, Schrappe M, Schübeler D, Seifert M, Siebert R, Simmons D, Soranzo N, Spicuglia S, Stratton M, Stunnenberg HG, Tanay A, Torrents D, Valencia A, Vellenga E, Vingron M, Walter J, Willcocks S. BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol* 2012;**30**:224–6
34. Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 2003;**12**:1725–35
35. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature* 2012;**489**:109–13
36. Yanez-Cuna JO, Arnold CD, Stampfel G, Boryń LM, Gerlach D, Rath M, Stark A. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res* 2014;**24**:1147–56
37. Ahmad SM, Busser BW, Huang D, Cozart EJ, Michaud S, Zhu X, Jeffries N, Aboukhalil A, Bulyk ML, Ovcharenko I, Michelson AM. Machine learning classification of cell-specific cardiac enhancers uncovers developmental subnetworks regulating progenitor cell division and cell fate specification. *Development* 2014;**141**:878–88
38. Pott S, Lieb JD. What are super-enhancers? *Nat Genet* 2015;**47**:8–12
39. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 2013;**153**:307–19
40. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. Super-enhancers in the control of cell identity and disease. *Cell* 2013;**155**:934–47
41. Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, Bradner JE, Young RA. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol Cell* 2015;**58**:362–70
42. Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD, Etchin J, Lawton L, Sallan SE, Silverman LB, Loh ML, Hunger SP, Sanda T, Young RA, Look AT. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a non-coding intergenic element. *Science* 2014;**346**:1373–7
43. Loven J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* 2013;**153**:320–34
44. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res* 2012;**22**:1748–59
45. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011;**473**:43–49
46. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJ, Shishkin AA, Hatan M, Carrasco-Alfonso MJ, Mayer D, Luckey CJ, Patsopoulos NA, De Jager PL, Kuchroo VK, Epstein CB, Daly MJ, Hafler DA, Bernstein BE. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 2015;**518**:337–43
47. Onengut-Gumuscu S, Chen WM, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, Farber E, Bonnie JK, Szpak M, Schofield E, Achuthan P, Guo H, Fortune MD, Stevens H, Walker NM, Ward LD, Kundaje A, Kellis M, Daly MJ, Barrett JC, Cooper JD, Deloukas P, Type 1 Diabetes Genetics Consortium, Todd JA, Wallace C, Concannon P, Rich SS. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat Genet* 2015;**47**:381–86
48. Gjoneska E, Pfenning AR, Mathys H, Quon G, Kundaje A, Tsai LH, Kellis M. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* 2015;**518**:365–9
49. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* 2014;**15**:272–86
50. Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R, Zhang AW, Parcy F, Lenhard B, Sandelin A, Wasserman WW. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2016;**44**:D110–15
51. Consortium GT. The Genotype-Tissue Expression (GTEx) Project. *Nat Genet* 2013;**45**:580–5
52. Leslie R, O'Donnell CJ, Johnson AD. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* 2014;**30**:i185–94
53. *3D Genome Browser*, http://www.3dgenome.org/ (accessed 7 March 2017)
54. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res* 2016;**44**:D877–81
55. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 2012;**22**:1790–7
56. Dunham IKE, Iotchkova V, Morganella S, Birney E. FORGE: a tool to discover cell specific enrichments of GWAS associated SNPs in regulatory regions. *F1000Res* 2015;**4**:18
57. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. *Nat Methods* 2014;**11**:294–6

58. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**:310–5
59. Li YR, Keating BJ. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med* 2014;**6**:9
60. Pai AA, Pritchard JK, Gilad Y. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet* 2015;**11**:e1004857
61. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* 2015;**16**:197–212
62. Melzer D, Perry JR, Hernandez D, Corsi AM, Stevens K, Rafferty I, Lauretani F, Murray A, Gibbs JR, Paolisso G, Rafiq S, Simon-Sanchez J, Lango H, Scholz S, Weedon MN, Arepalli S, Rice N, Washecka N, Hurst A, Britton A, Henley W, van de Leemput J, Li R, Newman AB, Tranah G, Harris T, Panicker V, Dayan C, Bennett A, McCarthy MI, Ruokonen A, Jarvelin MR, Guralnik J, Bandinelli S, Frayling TM, Singleton A, Ferrucci L. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet* 2008;**4**:e1000072
63. Zhang X, Gierman HJ, Levy D, Plump A, Dobrin R, Goring HH, Curran JE, Johnson MP, Blangero J, Kim SK, O'Donnell CJ, Emilsson V, Johnson AD. Synthesis of 53 tissue and cell line expression QTL datasets reveals master eQTLs. *BMC Genomics* 2014;**15**:532
64. Liu X, Finucane HK, Gusev A, Bhatia G, Gazal S, O'Connor L, Bulik-Sullivan B, Wright FA, Sullivan PF, Neale BM, Price AL. Functional architectures of local and distal regulation of gene expression in multiple human tissues. *Am J Hum Genet* 2017;**100**:605–16
65. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, Kellis M, Lander ES, Mikkelsen TS. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 2012;**30**:271–7
66. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 2013;**339**:1074–7
67. Inoue F, Ahituv N. Decoding enhancers using massively parallel reporter assays. *Genomics* 2015;**106**:159–64
68. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF, Sabeti PC. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* 2016;**165**:1519–29
69. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, Melnikov A, McDonel P, Do R, Mikkelsen TS, Sankaran VG. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* 2016;**165**:1530–45
70. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* 2013;**23**:800–11
71. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. Multiplex genome engineering using CRISPR/Cas systems. *Science* 2013;**339**:819–23
72. Wang L, Shao Y, Guan Y, Li L, Wu L, Chen F, Liu M, Chen H, Ma Y, Ma X, Liu M, Li D. Large genomic fragment deletion and functional gene cassette knock-in via Cas9 protein mediated genome editing in one-cell rodent embryos. *Sci Rep* 2015;**5**:17517
73. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, Qi LS, Kampmann M, Weissman JS. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* 2014;**159**:647–61
74. Waldrip ZJ, Byrum SD, Storey AJ, Gao J, Byrd AK, Mackintosh SG, Wahls WP, Taverna SD, Raney KD, Tackett AJ. A CRISPR-based approach for proteomic analysis of a single genomic locus. *Epigenetics* 2014;**9**:1207–11
75. Fujita T, Fujii H. Efficient isolation of specific genomic regions and identification of associated proteins by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. *Biochem Biophys Res Commun* 2013;**439**:132–6
76. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, Mero P, Dirks P, Sidhu S, Roth FP, Rissland OS, Durocher D, Angers S, Moffat J. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* 2015;**163**:1515–26
77. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* 2014;**343**:84–87
78. Horlbeck MA, Gilbert LA, Villalta JE, Adamson B, Pak RA, Chen Y, Fields AP, Park CY, Corn JE, Kampmann M, Weissman JS. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* 2016;**5**:e19760.
79. Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T, Stelhing-Sun S, Lee K, Thurman RE, Vong S, Bates D, Neri F, Diegel M, Giste E, Dunn D, Vierstra J, Hansen RS, Johnson AK, Sabo PJ, Wilken MS, Reh TA, Treuting PM, Kaul R, Groudine M, Bender MA, Borenstein E, Stamatoyannopoulos JA. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* 2014;**515**:365–70
80. Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* 2006;**312**:276–9
81. Mouse EC, ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T, Giste E, Johnson A, Zhang M, Balasundaram G, Byron R, Roach V, Sabo PJ, Sandstrom R, Stehling AS, Thurman RE, Weissman SM, Cayting P, Hariharan M, Lian J, Cheng Y, Landt SG, Ma Z, Wold BJ, Dekker J, Crawford GE, Keller CA, Wu W, Morrissey C, Kumar SA, Mishra T, Jain D, Byrska-Bishop M, Blankenberg D, Lajoie BR, Jain G, Sanyal A, Chen KB, Denas O, Taylor J, Blobel GA, Weiss MJ, Pimkin M, Deng W, Marinov GK, Williams BA, Fisher-Aylor KI, Desalvo G, Kiralusha A, Trout D, Amrhein H, Mortazavi A, Edsall L, McCleary D, Kuan S, Shen Y, Yue F, Ye Z, Davis CA, Zaleski C, Jha S, Xue C, Dobin A, Lin W, Fastuca M, Wang H, Guigo R, Djebali S, Lagarde J, Ryba T, Sasaki T, Malladi VS, Cline MS, Kirkup VM, Learned K, Rosenbloom KR, Kent WJ, Feingold EA, Good PJ, Pazin M, Lowdon RF, Adams LB. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* 2012;**13**:418
82. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston RH, modENCODE Consortium. Unlocking the secrets of the genome. *Nature* 2009;**459**:927–30
83. Pombo A, Dillon N. Three-dimensional genome architecture: players and mechanisms. *Nat Rev Mol Cell Biol* 2015;**16**:245–57
84. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**:1665–80
85. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y, Lu Y, Wu Y, Jia Z, Li W, Zhang MQ, Ren B, Krainer AR, Maniatis T, Wu Q. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* 2015;**162**:900–910
86. Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI, Zhao K, Young RA. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* 2014;**159**:374–87
87. Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, Santos-Simarro F, Gilbert-Dussardier B, Wittler L, Borschiwer M, Haas SA, Osterwalder M, Franke M, Timmermann B, Hecht J, Spielmann M, Visel A, Mundlos S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 2015;**161**:1012–25
88. Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, Nagano T, Andrews S, Wingett S, Kozarewa I, Assiotis I, Fenwick K, Maguire SL, Campbell J, Natrajan R, Lambros M, Perrakis E, Ashworth A, Fraser P, Fletcher O. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res* 2014;**24**:1854–68

89. Schwartzman O, Mukamel Z, Oded-Elkayam N, Olivares-Chauvet P, Lubling Y, Landan G, Izraeli S, Tanay A. UMI-4C for quantitative and targeted chromosomal contact profiling. *Nat Methods* 2016;**13**:685–91
90. Davies JO, Telenius JM, McGowan SJ, Roberts NA, Taylor S, Higgs DR, Hughes JR. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat Methods* 2016;**13**:74–80
91. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, Herman B, Happe S, Higgs A, LeProust E, Follows GA, Fraser P, Luscombe NM, Osborne CS. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* 2015;**47**:598–606
92. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sallari R, Lupien M, Markowitz S, Scacheri PC. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res* 2014;**24**:1–13
93. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E, Basford MA, Carrell DS, Peissig PL, Kho AN, Pacheco JA, Rasmussen LV, Crosslin DR, Crane PK, Pathak J, Bielinski SJ, Pendergrass SA, Xu H, Hindorff LA, Li R, Manolio TA, Chute CG, Chisholm RL, Larson EB, Jarvik GP, Brilliant MH, McCarty CA, Kullo IJ, Haines JL, Crawford DC, Masys DR, Roden DM. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013;**31**:1102–10
94. Dubord PJ, Mannis MJ. International eye banking and the Eye Bank Association of America (EBAA). *Refract Corneal Surg* 1991;**7**:478
95. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* 2010;**362**:1181–91
96. Carss KJ, Arno G, Erwood M, Stephens J, Sanchis-Juan A, Hull S, Megy K, Grozeva D, Dewhurst E, Malka S, Plagnol V, Penkett C, Stirrups K, Rizzo R, Wright G, Josifova D, Bitner-Glindzicz M, Scott RH, Clement E, Allen L, Armstrong R, Brady AF, Carmichael J, Chitre M, Henderson RH, Hurst J, MacLaren RE, Murphy E, Paterson J, Rosser E, Thompson DA, Wakeling E, Ouwehand WH, Michaelides M, Moore AT, NIHR-BioResource Rare Diseases Consortium, Webster AR, Raymond FL. Comprehensive rare variant analysis via whole-genome sequencing to determine the molecular pathology of inherited retinal disease. *Am J Hum Genet* 2017;**100**:75–90
97. Yu B, de Vries PS, Metcalf GA, Wang Z, Feofanova EV, Liu X, Muzny DM, Wagenknecht LE, Gibbs RA, Morrison AC, Boerwinkle E. Whole genome sequence analysis of serum amino acid levels. *Genome Biol* 2016;**17**:237
98. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med* 2015;**7**:16