BRIEFINGS IN BIOINFORMATICS. VOL 16. NO 3. 393–412
Advance Access published on 10 June 2014
doi:10.1093/bib/bbu018

Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression

Mulin Jun Li, Bin Yan, Pak Chung Sham and Junwen Wang

Submitted: 22nd January 2014; Received (in revised form): 23rd April 2014

Abstract

Understanding the genetic basis of human traits/diseases and the underlying mechanisms of how these traits/diseases are affected by genetic variations is critical for public health. Current genome-wide functional genomics data uncovered a large number of functional elements in the noncoding regions of human genome, providing new opportunities to study regulatory variants (RVs). RVs play important roles in transcription factor bindings, chromatin states and epigenetic modifications. Here, we systematically review an array of methods currently used to map RVs as well as the computational approaches in annotating and interpreting their regulatory effects, with emphasis on regulatory single-nucleotide polymorphism. We also briefly introduce experimental methods to validate these functional RVs.

Keywords: *regulatory variant; genetic mapping; transcriptional gene regulation; chromatin state; functional prediction; function validation*

INTRODUCTION

The advance in next-generation sequencing projects, such as the 1000 Genomes Project and the Personal Genome Project, have identified tens of millions of human DNA polymorphisms in populations and millions of variants per individual [1–3]. Nevertheless, the biological function of these variants, including both germline and somatic mutations, is largely unknown. In the next step, it is important to interpret the underlying molecular function, evolution and pathways that link these variants to diseases/traits. It has been well established that variants altering the amino acids of protein-coding genes play an important role in molecular pathogenesis [4]. However, by looking at the genomic location of the associated variants detected in recent genome-wide association study (GWAS), ~88% of them fall outside of protein-coding regions [5, 6], which indicates the significance of studying the function of these variants.

The functions of genetic variations, which do not directly change the protein sequence, are quite diversified in its genomic loci, and are involved in almost all processes of gene regulation, from

Corresponding author. Junwen Wang, Room 1-05E, The Hong Kong Jockey Club Building for Interdisciplinary Research, 5 Sassoon Road, Pokfulam, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China. Tel.: +852-28315075; Fax: +852-28551254; E-mail: junwen@alumni.upenn.edu

Mulin Jun Li is a PhD student at The University of Hong Kong. His research interests include bioinformatics and functional genetics, and he has developed many tools in annotating and prioritizing genetic variants.

Bin Yan is a research assistant professor at Hong Kong Baptist University, Hong Kong SAR, China. His research interests are transcriptional and epigenetic regulatory networks underlying cancer, stem cells and plant stress resistance.

Pak Chung Sham is a chair professor at The University of Hong Kong. His research interests include statistical genetics and epidemiology of psychiatric disorders.

Junwen Wang is an associate professor at The University of Hong Kong. His research interests include bioinformatics and computational genomics.

transcription to posttranslation [7, 8]. One type of variants that are intensively investigated recently locates near the splicing sites in either exon or intron, which directly affect the splicing events and result in aberrant transcript isoform abundance [9–11]. Recent reviews have summarized the association and mechanism between these splicing variants and disease [12, 13]. The noncoding variations located at RNA-producing regions can also influence the biological activities in posttranscriptional processing and translational initiation. The variations altering RNA secondary structures can cause different activities of alternative splicing, RNA folding, functional site accessibility, and the natural selection for specific RNA shapes many of variable sites including 3′-, 5′-untranslated region (3′-, 5′-UTR), microRNA (miRNA) binding site and RNA-binding proteins recognition motif genome-widely [14]. An increasing number of studies have demonstrated that genetic variants in the 3′-UTR or coding region of messager RNA (mRNA) [15], or miRNA body [16] can disrupt existing or create new mRNA–miRNA binding sites, and promote disease development and cancer pathogenesis. Genetic variants in the 5′-UTR also frequently change the global structure of the untranslated region and affect the recognition of translation initiation complex [17, 18]. In addition, genetic variants can affect the function of long intergenic non-coding RNAs (lincRNA) in a tissue-specific manner [19, 20]. Furthermore, the synonymous single nucleotide polymorphisms (sSNPs) that are present in the mature mRNA regions may affect different translational activities. sSNP may change the translational rate according to differential transfer RNA abundance determined by codon usage bias [21–23]; it can also alter the secondary structure of mRNA and the efficiency of protein expression [24], as well as the co-translational folding [25]. Importantly, with the explosive growth of next-generation sequencing (NGS) studies, a distinct group of genetic variants that affect gene expression have been identified. Because of their significant roles in regulating gene expression levels, these regulatory variants (RVs) attracted great interest of functional geneticists over the past decade [26, 27].

RVs play many roles in transcription, including transcription factor binding, chromatin states and epigenetic modifications [28]. Trait/disease-associated SNPs (TASs) detected by GWAS are significantly enriched in the regions that harbor functional elements, such as transcriptional factor binding sites (TFBSs), DNase I hypersensitive sites (DHSs), conservative regions and expression quantitative trait loci (eQTLs) [29–32]. These results imply that RVs in the linked regions of TASs may directly or indirectly connect to different transcriptional regulation events. Also, late studies showed that important chromatin marks, such as H3K4me3 and H3K27ac, and putative enhancer regions are phenotypically cell type specific and are likely associated with TAS loci in the relevant cell type of same disease/trait [33, 34]. Furthermore, population genetics studies demonstrated that RVs are under strong natural selection and affect gene expressions via local adaptation [35, 36]. Taken together, all these findings shed light on importance of function investigation of RVs in affecting gene regulation.

Recently, technological innovation, such as NGS and high-throughput genotyping platforms [37, 38], enable us to perform genome-wide genetic mapping and large-scale functional profilings in multiple individuals/cell lines. In addition, many computational methods are developed to help researchers in detecting, annotating and prioritizing RVs with satisfactory sensitivity and specificity [39–43]. In this review, we will focus on current approaches to identify functional RVs affecting transcriptional gene regulation. As shown in Figure 1, various methods, from upstream variant genetic mapping to downstream annotation, as well as the final functional validation, will be discussed for RV identification. We summarize prevalent techniques of RV genetic mapping, prioritization strategies using bioinformatics analysis of genomic data, and we also briefly introduce current experimental methods on functional RV validation *in vitro* and *in vivo*.

REGULATORY VARIANTS GENETIC MAPPING

Genetic mapping is the primary step to associate variant/gene markers with specific disease/trait, based on genotype information from disease and control groups. The methodology of RV genetic mapping follows the same procedure of conventional mapping. We will first summarize these evolving genetic mapping methods and highlight the differences and difficulties of RV fine-mapping.

Whole-genome association test

Most RVs are conventionally thought to cooperatively influence human complex diseases and traits

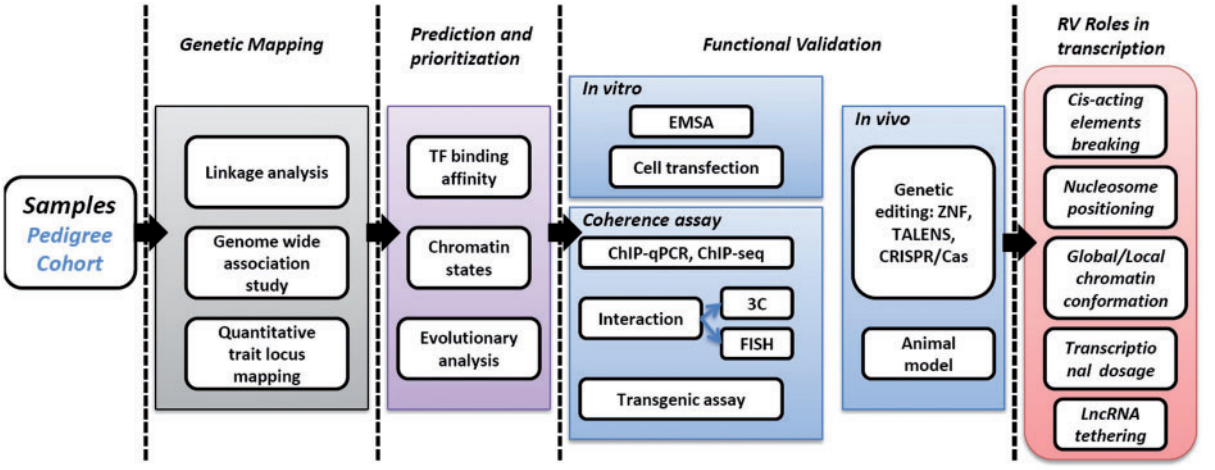

Figure 1: The general framework of RVs in genetic mapping, prediction, prioritization and functional validation.

together with environmental factors [44]. Under the common disease–common variant hypothesis, genetics association studies are usually performed to map the relationship between a phenotype and specific genomic loci under case-control design. The merit of case-control association test, compared with family-based design, comes from the convenience and feasibility to discover genetic susceptible loci on a group of unrelated individuals with the same trait. Technology advances have enabled us to carry out high-throughput GWAS for many common genetic variants [37]. Modern GWAS genotyping chips typically contain 300 000–5 000 000 tag SNPs that are selected from segments of human genome with high linkage disequilibrium (LD), which make it possible to identify susceptive genetic loci without genotyping every SNP in the whole chromosome [45, 46]. However, in most of the time, tag SNPs are not the functional ones for the investigated phenotype, and the GWAS effect size only represents the significance of that locus in which tag SNPs locate. Therefore, it can reasonably postulate that the immediate gene contains or is close to the tag SNPs (with *P*-value <5E-8) could be the causal gene associated with targeted disease/trait, but it is hard to find out which SNP is the causal mutation.

To identify the true functional SNP from tag SNP (or GWAS SNP), we have to consider the haplotype structure of the investigated population. The haplotype blocks of different human populations are quite distinct. For example, European-descent populations have more highly correlated SNPs and longer haplotype blocks than other populations such as African or Asian. The recent International HapMap Project and the 1000 Genomes Project have produced high-quality genotyping data in large sample size of different human populations [1, 47]. The influence of these projects, along with the evolving genotyping technologies, led to wide-spread GWAS that have identified over thousands of common variants associated with many traits and diseases [5]. It also enables us to systematically hunt for functional regulatory single-nucleotide polymorphisms (rSNPs) given a GWAS SNP, by using population-specific haplotype structures [48, 49]. Several statistical inference methods have been developed to fine-map strongly associated SNPs in the corresponding LD proxy. Genotype imputation can be used to infer the untyped or missing SNPs [50–52], and to exclude less significant SNPs in the same LD according to likelihood ratio test [53]. With accumulation of GWAS on the same or similar traits, meta-analysis methods are frequently used to improve the power of association study [54, 55].

GWAS is a routine method of association mapping to discover associated variants for common diseases/traits; it has some distinct features in fine mapping of rSNPs. First, after statistical fine-mapping, only a small fraction of associated SNPs are nonsynonymous single nucleotide variants (nsSNV) that can easily be linked to protein function, which indicates that most of common disorders are caused by gene regulatory mechanism and those rSNPs may exert their function in transcriptional gene regulation [6]. On the other hand, by mapping all GWAS significant SNPs (*P*-value <5E-8) to the dataset of HapMap3 and comparing the occurrences of these SNPs in each function type, Li et al. found disease/trait-associated rSNPs are less frequent than

associated nsSNV [56]. It reflects that GWAS has less power to detect the putatively causal SNPs in the regulatory regions, especially for intergenic SNPs. Also, the haplotype frequency of a small effect size rSNP locus could be low in the studied population. Hence, we usually need a large sample size to achieve genome-wide significance of rSNP even for common phenotypes. Again, the function of an rSNP on a trait can be modulated by one or more other variants in long-range LD or unrelated locus by epistasis, which makes independent test even difficult.

As large-scale genome sequencing projects have revealed large amount of rare variants in human populations, the genetic effects of these low frequency variants that were not included in current GWAS chips have been proposed as main causes of the 'missing heritability' [57–59]. With the continuously reduced sequencing cost, whole-exome sequencing (WES) and whole-genome sequencing (WGS) have currently emerged to overcome the restrictions of GWAS chips, and are used to discover rare and *de novo* disease-causal mutations for both Mendelian disorders and complex diseases [60, 61]. Many new disease-associated variants/genes have been identified by these technologies [62, 63]. Evidence has shown that rare variants involved in complex disease etiology are more likely to be functional than common variants [58, 59]. Many factors can affect the statistical power to identify the disease causal variant with low frequency including variant effect size, sample size, genetic inheritance mode and minor allele frequency (MAF). To efficiently test the disease association of rare variants that are detected by WES, WGS or low frequency variant genotyping arrays, there are many statistical methods developed by using burden tests, C-alpha test and their derivatives [64–68]. It is anticipated that a large number of common and rare regulatory variants with disease association will be identified in the near future. However, interpreting the function of these disease-associated rSNPs is still a big challenge, particularly in downstream analysis and functional validation. It requires different integrative resources to represent, annotate and prioritize those TASs in the post-GWAS era [56]. Some recent reviews have emphasized on those parts [49, 69–71].

Linkage analysis

Linkage analysis has been originally applied to assess excess co-transmissions of marker variants with monogenic disease in families with multiple affected members. Recent applications of WES significantly speed up this process of Mendelian disease-casual genes identification [61, 72]. Exome sequencing-based linkage studies are usually carried out on affected and unaffected individuals from a family or unrelated group. Recent target enrichment solutions of WES, such as 'All Exon Kits' provided by Agilent, capture almost all protein-coding regions as well as the nearby regulatory loci [73]. After the sequencing step, the read fragments will first be mapped to reference genome, and the sequence variants will be called at specific positions according to a series of quality controls and recalibrations [74–77]. Bioinformatics tools have been developed to reduce the biases and errors in the genome mapping and variant calling [78]. Given large amount of called sequence variants, filtration and prioritization are usually conducted to select putatively disease-causal mutations for further replication and validation [79]. Because Mendelian genetic disorders have varied penetrance and complicated inheritance, it entails different analysis strategies to isolate the causal mutations efficiently from neutral sequence variants [80]. For example, given a trio in which there are two unaffected parents and an affected son, at least four strategies can be used: candidate linkage regions strategy [performing identical by descent (IBD) scan], runs of homozygosity strategy (when consanguineous mating happens, searching long range of continuous homozygous genotypes in patients), double-hit gene strategy [hitting deleterious allele for homozygous genotypes (one from maternal and other from paternal)], and *de novo*-mutation strategy (searching for *de novo*-mutation only existing in child but not in parents). If no other sample and disease information are available, researchers have to try each aforementioned strategy one by one and then rely on their expertize to select several most likely causal variants into the final prioritization list for follow-up validation.

Although WES can efficiently map disease-causal variants/genes at genome-wide level, it only targets a small fraction (~1%) of the entire genome. Meanwhile, the efforts are mainly focus on two types of variants for further validation: nsSNVs that disrupt protein-coding sequences and variants that affect the transcript splicing. However, studies on variations in noncoding regions (here refer to RVs) that could be highly deleterious in monogenic disease development are significantly hampered by the

limited genomic coverage of WES. It is notable that synonymous or nonsynonymous changes in the protein-coding regions have been reported to hold another hidden role in gene regulation. It was reported that ~14% of the codons within 86.9% of human genes are occupied by the binding events of transcription factors (TFs) in 81 diverse cell types. The concept of 'duons' has been proposed for genetic variants located in the exon regions, which play both protein coding and regulatory roles [81, 82]. In this capacity, there are great possibilities to identify RVs that affect TF binding and other regulatory incidents in exonic regions. On the other hand, WGS, with its constant decreasing cost, provides unprecedented chance to genome-wide screen causal RV of Mendelian disease in typical pedigrees using linkage-based method, which promotes the full understanding of biological mechanisms of inherited disorders [83, 84]. The method of whole-genome mapping of RVs is consistent with that of WES, but there are no universal criteria for RV prioritization, which will be discussed later. A recent study discovered six different recessive mutations in a previously uncharacterized enhancer region located 25 kb downstream of *PTF1A* by integrating combined WGS, linkage analysis and epigenomic profiling data [85]. The mutations affect the binding of transcription factors FOXA2 and PDX1, and increase the susceptibility of isolated pancreatic agenesis. In this study, researchers used runs of homozygous strategy to search causal variants over long runs of homozygosity regions on six affected subjects and one unaffected subject from three unrelated consanguineous families. Another study used WGS to exploit the casual factors among amyotrophic lateral sclerosis pedigree. Researchers adopted linkage-based strategy to fine-map shared haplotype in a two-generation pedigree and discovered a non-coding pathogenic hexanucleotide repeat expansion that contributes the disease susceptibility [86].

Mapping quantitative trait locus

Quantitative traits, such gene expression, DNA methylation and histone modification, are thought to be largely heritable during the evolution of species [87]. Most of quantitative trait locus (QTLs) only account for a small fraction of the total genetic variations in the population, exert relatively small effect size and jointly contribute to a complex trait [88]. Understanding the correlation between genetic variations of DNA sequence and the phenotypic changes of the quantitative traits is not only important for the identification of disease molecular mechanism, but also for the functional interpretation of the genetic variants, which is hard to detect by conventional genetic mapping, especially for rSNP that locates outside of protein-coding regions. Many QTLs mapping studies have unveiled new susceptible loci and provided significant insights into the human genetics and medicine, and the method becomes an important complement of linkage analysis and association study [89, 90]. The power to map QTLs is determined by their genetic effects, the allele frequencies and the pattern of LD. Large numbers of individual and genetic markers per individual are required to locate the true effect site [88]. Original QTL mapping first defines a series of genomic markers co-segregate with a QTL and then generates recombinant inbred lines from two parents who differ in a trait. Significant relationship between each marker and investigated trait, includes single-marker mapping [91], interval mapping [92], composite interval mapping [93, 94] and multiple trait mapping [95], can be determined by different statistics tests such as *t*-test, ANOVA or regression analysis. However, the iterative mappings need be further performed to identify the high-resolution region containing the QTL [96].

Recent advances on large-scale genotyping and sequencing of human genome have enabled us efficiently to map high-resolution QTLs using SNP markers. The International HapMap Consortium made great efforts to catalog all common genetic variation across different ethnic groups (at least 5% MAF) [47, 97]. Also, the 1000 Genomes Project aimed to sequence 2500 individuals and identify rare variants with a MAF of <1% [1]. These population genetics data provide a foundation for mapping the exact locus underlies quantitative traits in human. On the other hand, high-throughput microarray and NGS-coupled profiling have produced a large number of genomic, epigenomic and transcriptomic data, which drive the evolution of QTL mapping methodology and provide great opportunities to dissect the genetic variations of complex phenotypes. To date, several human quantitative traits, including gene expression, protein expression, non-coding RNA expression, alternative splicing, chromatin accessibility, DNA methylation, histone modification and translational efficiency, have been used to measure their genetic associations under different cell types/tissues, and most of these genetic associations are controlled by the rSNPs that exert

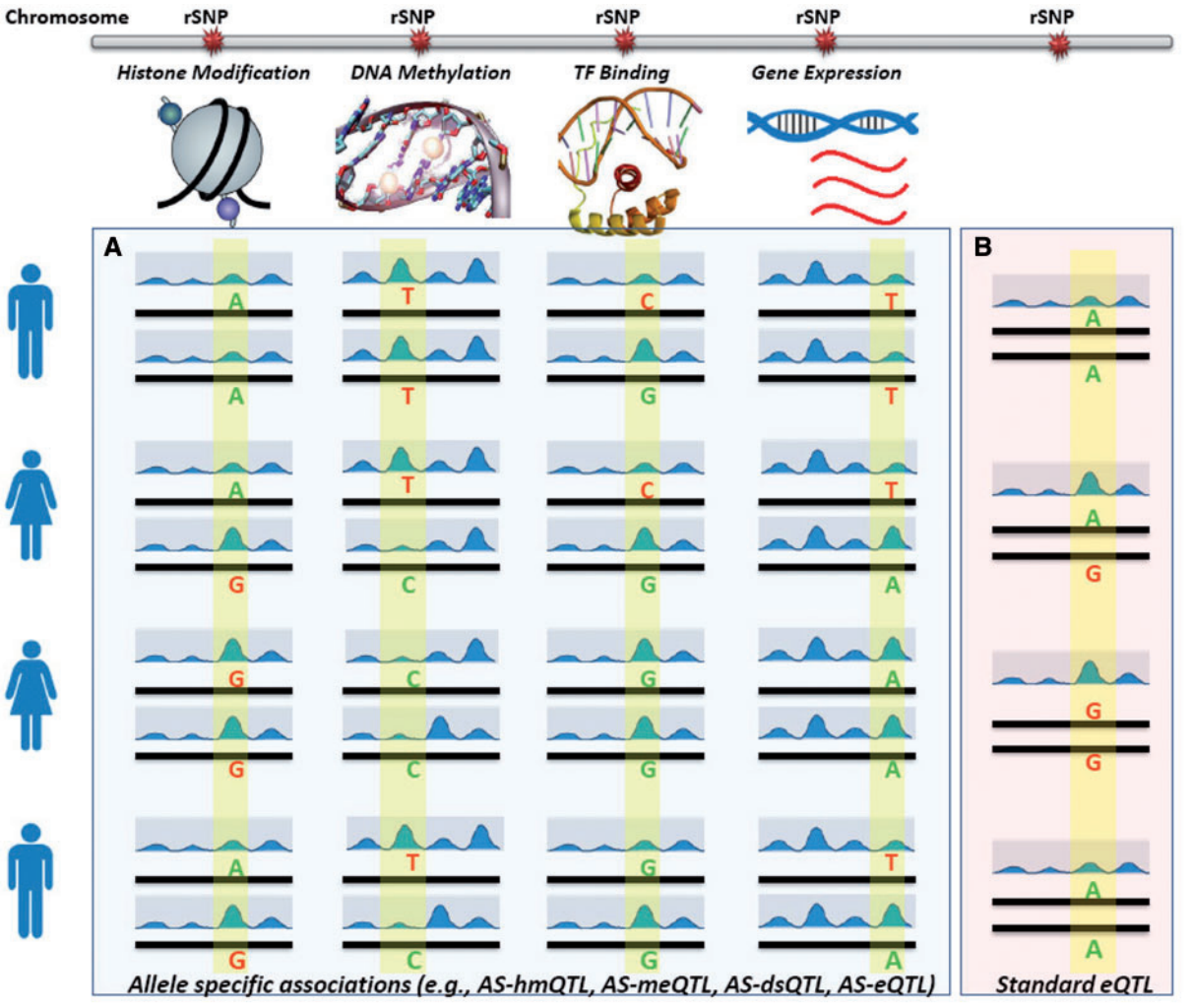

Figure 2: Mapping the QTL according to different genomic and epigenomic signals. (**A**) QTLs mapping that correlates allelic effect to different molecular phenotypes of transcriptional gene regulation including histone modification, DNA methylation, TF binding and gene expression. Molecular phenotypes express allele-specific expression pattern and require haplotype-based QTL mapping other than (**B**) standard QTL using only genotype and overall expression.

allele-specific function in highly dynamic chromatin states and transcriptional activities (Figure 2). Compared with standard QTL mapping (Figure 2b) in which the genotype of target SNPs are treated as the independent variable, the haplotype-based QTL mapping (Figure 2a) are recently used to pinpoint the allele-specific transcriptional activities of target SNP [98].

Gene transcript eQTLs

eQTL mapping is one of the most prominent directions in the studies of quantitative traits and has been extensively applied to unravel the genetic variants that explain the variation in gene expression levels. Typical eQTL mapping requires both genetic (genotyping or sequencing per individual) and gene expression data (microarray or RNA-Seq per individual). These methods measure direct association between genetic variants and gene expression levels in a cohort of individuals (tens or hundreds) from the same population. Recent works have successfully revealed large number of genome-wide eQTLs for different ancestry [99] or different tissues/cell types [100], and those results highlight the highly dynamic gene regulation in condition-specific manner and provide a comprehensive view of linking the rSNPs to their direct gene targets. Furthermore, similar experiments have been carried out on the expression of lincRNAs across different tissues to study the association between genetic variants and

lincRNA abundances [19]. However, questions are raised on the exact molecular mechanisms of those associated loci, in which the regulatory variants have been characterized as either *cis* or *trans* acting, demonstrating the functional complexities in terms of physical distance with their target genes [101]. Although one can immediately estimate that eQTL may affect the activity of *cis*-regulatory element (CRE), such as promoter and distant enhancer, which directly control the expression of neighboring gene [102], or function to influence upstream regulators of target gene in an indirect way, such as TF and their target genes [103], miRNAs and their target genes [90, 104], it is still difficult to connect them to underlying phenotypes, especially for human disease. Researchers have used GWAS results to show that TASs are more likely to be eQTL [105] and to predict gene/SNP-disease associations by matching patterns of expression data [106, 107], but interpreting the functional relationship between those rSNPs and disease development yet requires in-depth investigation from regulatory perspective.

DNase I sensitivity quantitative trait loci

To identify the causal regulatory variants and further exploit the regulatory mechanisms of how eQTLs affect gene expression, DNase I sequencing has been used to measure chromatin accessibility in matched samples (such as Yoruba lymphoblastoid cell lines). Many DNase I sensitivity quantitative trait loci (dsQTLs) have been successfully inferred by correlating the DNase I sensitivity level with individual genotype, indicating that allele constituent of rSNPs can cause different levels of transcription factor binding or nucleosome occupancy at regulatory loci [108]. Joint dsQTL–eQTLs analysis also demonstrated that dsQTLs are dominant factors in affecting gene expression levels and most of eQTLs are also dsQTLs [108]. Therefore, rSNP affecting chromatin accessibility may be a major mechanism linking to associated changes in gene regulation and, ultimately, individual phenotype.

Histone modification quantitative trait loci

Recent genomic studies elucidated some specific post-translational modifications of histone (like H3K4me1, H3K4me2, H3K4me3, H3K27ac and H3K27me3) and TFs (like EP300, CTCF and cohesin) are associated with active or repressive chromatin states and could be regarded as the chromatin marks of CREs [109]. Followed by gene expression and DNase I hypersensitivity traits, researchers recently used multiple histone marks and specific TF-binding profiles to investigate whether the chromatin variability are genetically inheritable in a relatively small cohort from the same human population or in trios [110, 111]. Those works uncovered that large number of abundant allelic specificities is correlated with concordant trend of TF binding, histone modifications and transcription operation [112]. Several histone modification quantitative trait loci (hmQTLs) and TF-binding QTLs were successfully identified at both population and family levels, which indicates that the variances of critical molecular traits shape the phenotypic differences between individuals and ethnic groups through genetic operation [113, 114]. On another layer, the mapping of hmQTLs will also greatly facilitate the identification of regulatory variants affecting functional chromatin states.

Methylation quantitative trait loci

DNA methylation is a fundamental epigenetic mark that controls the switch of gene expression [115]. Nevertheless, the dependency of genomic sequence for DNA methylation level as well as the lineage specificity is largely unknown [116]. Same as with other QTL mapping methods, researchers correlated genome-wide DNA methylation profiles with individual genotypes on human cohorts to identify loci that affect DNA methylation. A lot of genetic loci were discovered that can explain differentially methylated CpG sites in population specific or cell type specific manner, although not all variations of DNA methylation can be interpreted according to genetic factors [117–120]. Therefore, the dynamic DNA methylation profile as well as their causal relationship with gene expression will be an essential part on studying the functional role of regulatory variants.

Splicing quantitative trait loci

Alternative splicing can produce different mature mRNA isoforms from same gene and more than 90% of human genes are alternatively spliced [121, 122]. Recent RNA-seq technology has provided effective solution to quantitatively measure the exon expression levels and canonical/novel splicing events [123]. It also enables us to look at the correlation between genetic variants and exon expression level on single nucleotide level. To this end, several studies applied linear regression models to detect splicing quantitative trait loci (sQTLs) in population cell lines [10, 11, 124]. Functional interpretation of

sQTLs will be a daunting task because of the complexity of splicing. Genetic variants may affect different splicing events, such as exon skipping, 3′ or 5′ alternative splice junction and intron retention. It has been uncovered that alternative splicing can also be modulated by the variation of RNA secondary structure [14].

Other QTLs, such as ribosomal associated gene transcript expression quantitative trait loci [23] and protein expression quantitative trait loci [125], were investigated recently and have been shown to explain many quantitative traits even under the circumstances of inadequate sample size. Those important QTLs function in posttranscriptional and translational level can also directly contribute to diversified phenotypes. Although QTL mapping can be an efficient solution to exploit the genetic loci that are associated with quantitative traits, it is still difficult to find out the true functional variants because of the limitations of LD structure, sample size and genotyping volume in investigated population. Therefore, fine mapping of regulatory variants from QTL studies need in-depth analysis and replication. In addition, context dependency is also a determinant when mapping regulatory QTL and rSNP in different cell lineages.

Higher-order mapping

Aforementioned genetic mapping methods assume that functional genetic variants contribute to a phenotype independently and follow an additive effect model. These methods inevitably miss the chance to detect the collaborative effect in which two or more variants work together. Many studies have showed that the genetic landscape of a cell is highly interactive and coherent between genes in eukaryote [126, 127]. The epistasis and its implications in human diseases are also well discussed and have been proposed to solve the problem of missing heritability in lots of association studies [128–131]. In this regard, researchers have developed many statistical and experimental methods for higher-order genetic mapping [132].

Three major strategies can be used to search the genome-wide epistasis effect, including main and interaction effects, using genotype data within cases and controls [133, 134]. The exhaustive strategy iteratively enumerates all possible interactions among SNPs, and then evaluates the statistical significance under an assumed distribution. It is extremely computational demanding when applying this method to a whole GWAS data set even for many optimized algorithms [135, 136]. The second strategy selects the valid SNPs combination randomly in the candidate space and tests them in a well-trained model. But this strategy may introduce biases such as sampling error and model over-fitting [137, 138]. Third, heuristic strategy searches for the valid combination under the given conditions according to prior knowledge and defined rule [139, 140]. Apart from the SNP-SNP interaction, the interactions between genetic variants and environment factors can be used to understand the genetic basis of disease development beyond heritability [141, 142]. However, the power and true discoveries of association testing for multi-SNP and SNP-environment interaction are largely restricted by the allele frequency, the significance level, sample size, the number of typed SNPs and disease penetrance, which raise challenges in statistical corrections of multiple testing. Importantly, rSNPs that take up large proportion of interactions could interpret many variable transcriptional activities by their *cis*-acting loci and eQTLs [143, 144].

Perspectives on genetic mapping of regulatory variants

Fine mapping of RVs are more difficult than mapping the protein-coding variants. First, the search space is larger because we have to search significant signals across whole genome, which is ∼30–50 times larger than the protein-coding regions. More computing resources are needed in phasing, imputation and association testing. For mapping RVs using family-based WGS on rare disease, it still requires large efforts in searching linked regions and filtering unqualified variants. Secondly, the possible biological functions of RVs can be complex and involve many processes of transcriptional gene regulation (see next section). Therefore, unlike nsSNVs that directly change the protein function, it is difficult to pinpoint the molecular mechanism of RVs only from genotype information alone. Fortunately, the recent advent of high-throughput NGS-coupled technologies enables us to measure genomic, epigenomic and transcriptomic traits, such as transcript expression, TF binding, chromatin accessibility, histone modification and DNA methylation. This genome wide data significantly facilitate the interpretation of RVs' effects on those phenotypes by methods such as eQTL, dsQTL and hmQTL. In summary, fine-mapping of RVs and estimation of their functional roles in gene regulation are complicated but feasible.

GWAS, linkage analysis and QTL are complementary solutions that collaboratively provide means to decipher the effect of genetic variants and the etiology of human diseases/traits.

Function prediction and prioritization of regulatory variants

Genetic mapping from either population level or family-based study cannot always capture the true causal variants. For most of GWAS and QTL analysis, identifying the causal variant from a GWAS SNP is hard unless we perform expensive and time-consuming experiments. The linkage analysis also has limited resolution in detecting the pathogenic mutations from large linkage peaks for most of genetic inheritance patterns. Even if the result of fine mapping can be acquired by statistical step, the critical issue is how to accurately elucidate the biological mechanism these variants act. Sophisticated algorithms for prioritizing nsSNVs have been developed, as the variants can directly alter protein function and many relevant information are available to facilitate rightly deleterious estimation, including phylogenesis, amino acid physicochemical properties and conformation information. Recent efforts further improved the performance of nsSNV pathogenic prioritization by combined prediction model with more functional scores of existed tools [145–147]. However, for the RVs that have elusive functions in transcriptional gene regulation, the functional interpretation will be largely complicated.

Gene transcription is governed by many spatial and temporal factors such as global or local chromatin states, nucleosome positioning, TF binding, and enhancer/promoter activities. RVs altering any one of these processes may change the gene regulation and result in the phenotypic abnormality. The influencing effect size of RVs can be very diverse in terms of the variant properties. An rSNP may only change the motif sequence of the *cis*-acting regulatory element and consequently affects the transcription regulation performance. In contrast, a deletion or insertion of DNA sequence may completely deplete the motif that a specific regulator binds. Also, the copy number changes of DNA fragments could result in big chromosome conformation change and abnormal transcriptional level. Many continuous genomic loci that can recruit the binding of core TFs and function as super enhancers have been identified [148, 149]. Genetic variants, especially Indels and CNVs, in functional chromatin stretches will have a great chance to impact the processing of condition-dependent gene transcription [34]. We outlined the biological mechanisms of RVs influencing transcriptional gene regulation in Table 1. Although the molecular mechanisms of RVs are elusive, current genomic studies have found many genuine RVs in human genome according to specific genetic and epigenetic features. Here, we highlight some prevalent and new emerging computational methods for detecting and prioritizing RVs.

Transcriptional regulator activity profiling

Early studies have revealed a batch of TF-binding motifs using systematic evolution of ligands by exponential enrichment or other high-throughput motif enrichment methods, which are generally represented by position weight matrices (PWMs). Computationally, by applying motif scanning for these PWMs stored in public database including TRANSFAC [171], JASPAR [172] and UniPROBE [173], one could easily judge the putative DNA-binding affinity for each TF given a DNA sequence [174]. As TFBSs are short (usually 6–20 bp) and degenerate, mutations in there are more likely to impact binding-affinity changes [175]. Many diseases and traits can be attributed to the allele-specific TF binding, and most of these binding alterations are caused by sequence variations in the DNA functional elements including promoters, enhancers, silencers and insulators [176, 177]. It is straightforward to quantitatively measure the difference of binding affinities (gain or loss) between alleles by calculating the log-odds of binding probabilities for each motif-given paired DNA sequences contain SNP. Researchers have developed various bioinformatics tools with the modifications to estimate the variants effect and prioritize the rSNPs based on binding affinity changes (Table 2).

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is an effective method to analyze protein's interaction with DNA, and has been frequently used to investigate the genome-wide binding pattern of a specific TF. The Encyclopedia of DNA Elements (ENCODE) project has performed more than 500 ChIP-seq experiments of hundreds TFs across a number of human cell lines [26, 189]. Those data provide unprecedented resources to study the TFs dynamic activities considering the specificity of TF binding in different cell types. Therefore, to facilitate the identification of

Table 1: Mechanisms of regulatory variants influencing transcriptional gene regulation

Known processes	Molecular mechanisms
Affecting promoter activity	Mutations disrupt or promote the transcription initiation and the assembly of transcriptional factory by directly affecting the binding of critical activators, repressors and other transcriptional units [150–153].
Altering enhancer/silencer function	Genetic variants located in enhancer/silencer region can affect the binding motif of transcription factors, chromatin regulators and other distal transcriptional factors, which disturb the interaction between enhancer/silencer and its target gene [154–156].
Altering insulator, other distal *cis*-regulatory elements	The function of insulator could be abolished by genetic variants that disrupt the CTCF binding [157, 158].
Influencing nucleosome positioning	Mutations in specific DNA sequence can affect the packing efficiency of nucleosomes and the transitions between euchromatin and heterochromatin [159–161].
Disrupting distal/proximal interaction between functional elements	The genetic variants, include SNP and indel, can affect the normal chromatin structure and result in improper chromosome interactions [162–164, 165].
Breaking global chromosome structure	Large Indel and CNV can destroy the normal chromatin structure and result in improper chromosome conformation [166].
Changing transcriptional dosage	gene expression can be influenced by higher and lower gene dosages through insertions or deletions of duplicate gene or transcriptional unit into cell [167, 168].
Affecting noncoding RNA tethering	The interaction between long noncoding RNA (such as Xist) and chromatin may be lost due to a DNA mutation disrupt the RNA–DNA recognition [169, 170].

Table 2: Current function prediction and prioritization tools for detecting regulatory variants in transcriptional gene regulation

Theoretical basis	Name	Type	Available information	Functions
TF-binding affinity	is-rSNP [178]	Web server	TF motifs	Scoring two alleles with a TF PWM
	TRAP [179]	Web server	TF motifs	Calculate binding affinity change of two alleles by PWM scanning
Chromatin states	HaploReg [42]	Database	Chromatin state, DNase, TF binding, TF motif, eQTL, Conservation	Data integration, combined GWAS and LD
	RegulomeDB [41]	Database	Histone modifications, DHSs, TF binding, TF motif, Conservation	Data integration, category scoring
	GWAS3D [43]	Web server	Histone modifications, DHSs, TF binding, TF motif, chromosome conformation, conservation, combined *P*-value	Data integration, real-time calculation, combined GWAS and LD, statistical test, large-scale annotations
	ChroMoS [180]	Web server	Predicted chromatin state, TF motif	Data integration, scoring two alleles with TF PWM
	rSNPBase [181]	Database	Histone modifications, TF binding, eQTL, distal interaction	Data integration, LD information
Evolution	VAAST [182]	Software	Phylogenetic conservation, Amino acid substitution	Unified likelihood model
	GERP++ [183]	Database	Rejected substitutions, Neutral rate, Base-wise score	Probability model based on evolutionary tree
	PhyloP [184]	Software	Base-wise conservation scores and *P*-value	Probability based on an alignment and a model of neutral evolution
	dbPSHP [185]	Database	Positive selection scores	Comprehensive data integration from population genetic, positive selection models
Combined	GWASrap [56]	Web server	Combining functional prediction scores from transcription to translation	Data integration, large-scale annotations, additive model
	FunSeq [186]	Software	Histone modifications, DHSs, TF binding, TF motif, distal regulatory module, conservation, functional score	Data integration, large-scale annotations, cancer and personal genomes
	CADD [187]	Software	88 annotations for genomic and epigenomic data, C-score	Data integration, large-scale annotations, machine learning method
	GWAVA [188]	Software	Histone modifications, TF binding, DHSs, RNA polymerase binding, conservation	Data integration, large-scale annotations, machine learning method

RVs that actually function in specific condition, we have to use cell type specific TF-binding site data and filter the spurious binging events. Fortunately, recent resources, including HaploReg [42] and RegulomeDB [41], have collected massive TF-binding events from ENCODE and public data sets to annotate the putative RVs and mapped the TASs to those signals, which significantly narrow down the RVs identification (Table 2).

Although direct TF-binding affinity scanning coupled with experimental data is a valid method to search functional rSNP candidates, the searching space can be extremely large when applying all known motifs even if incorporating ChIP-seq data for regional filtering. This situation can be relieved if both genetic mapping and TF-binding experiment are conducted beforehand, or limit the search only on several relevant TFs [190, 191]. In addition, multiple testing control and permutation are usually needed to fetch the statistically significant hits [192], which further aggravate the computational intensity. Importantly, current power of rSNP predication by TF-binding affinity scanning is still low because of the limited number of discovered motifs and TF-binding assays.

Chromatin state measurement

Many studies showed that specific chromatin states are inheritable following DNA replication in eukaryotes ranging from yeast to mammals [193–195]. However, how the genetic factors, such as SNP, Indel and CNV, connect to distinct chromatin structure and result in the spatiotemporal patterns of gene regulation is largely unknown. It was well established that chromatin states including histone modifications and DNA methylation underlie specific functional elements and represent regulatory processes other than TF regulation [115, 196]. Some chromatin marks are frequently used to pinpoint the distinct functional elements (like enhancer/insulator/promoter) in different cell types, which indicate active or repressive transcription events of euchromatin. In addition, active chromatin captured by DHSs sequencing usually exposes the DNA and produces accessible chromatin zones that are functionally related to transcriptional activity. Therefore, researchers began to incorporate chromatin marks for rSNP prediction. Same as with the TF-binding profiling, one can easily locate the putatively regulatory variants by mapping profiles of different chromatin marks (such as H3K4me1, H3K27ac, EP300, DHS for enhancers) and then filter the variants according to the marks occupancy. HaploReg [42], RegulomeDB [41], GWASrap [56] and rSNPBase [181] have collected and curated large number of cell type specific chromatin data for each SNP site in latest dbSNP and 1000 Genomes Project. Those data significantly scale down the searching space and facilitate rSNP identification (Table 2).

On the other hand, the spatial organization of chromosomes is not random and is pivotal to the spatiotemporal regulation of gene expression, DNA replication and repair and recombination. Chromosome conformation capture (3C) as a new emerging technology is used to analyze the organization of chromosomes of cell's natural state. The derivatives of 3C, such as Circularized Chromosome Conformation Capture (4C), Carbon-Copy Chromosome Conformation Capture (5C), Hi-C, ChIP-loop and Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET), greatly improve the power of 3C and aid the genetic and epigenetic studies of chromosomes [197]. It was reported that structure variations can disrupt the chromosome conformation by altering interaction of chromosome fragments [166, 198, 199]. Apart from overall chromosome structure, promoters and distal functional elements frequently act in looping interactions that have been implicated in transcriptional gene regulation, and many studies have shown that long-range interactions enhance or inhibit gene expression directly [200–202]. With the improvement of sensitivity of high-throughput conformation capture including high resolution Hi-C (up to 10 kb) and ChIA-PET (narrow peaks) in cell population or single cell level [203–205], we can investigate the genome-wide DNA interaction profile and their association to SNP function [206, 207]. Recent study used 3C- and 4C-Seq methods to show that obesity-associated variants within the *FTO* gene can affect another gene *IRX3's* expression, at megabase distances, by long-range interaction with the *IRX3* promoter [208], which is a perfect example to demonstrate the practicability and validity of chromosome conformation capture assay in identifying casual regulatory variants and target gene. Also, it highlights that the long-range interaction genomic data can be used to prioritize the functional regulatory variants in relevant cell type. In GWAS3D [43], researchers combined multiple domains data, especially distal interaction data, to annotate and predict the rSNP in its risk haplotype and in 3D chromatin

structure, which greatly increase the sensitivity and specificity of rSNP identification.

To quantitatively evaluate and prioritize functional consequence of rSNPs correlated with different chromatin signals, one direct measurement is to calculate the underlying TF-binding affinity changes between different alleles using aforementioned motif-scanning method. Both GWAS3D [43] and ChroMoS [180] adopted this manner in the final prioritization step (Table 2). Recent works have collected a large set of ENCODE TF motifs by incorporating different enrichment methods [209] and by integrating multilevel regulators [210], and those data further extend the TFs-binding library. However, the causal relationship between the TF binding and dynamic epigenomic modification is still a debatable question. hmQTLs and methylation quantitative trait loci studies have uncovered a batch of genetic loci that correlated with the changes of histone modifications or DNA methylation by specific TF binding [113, 114, 120]. Genetic variants that alter the TF binding within CRE may lead to heterogeneity and asymmetry of chromatin states between individuals or cells [211]. But whether and how the regulatory genetic variants change and form those differences of epigenomic modification are still open questions. Recently, several methods and tools were devised to systematically combine large-scale genomic and epigenomic data for noncoding variants prioritization [186–188] (Table 2).

Evolutionary methods

Comparative genomics approaches for RVs prediction assume that the DNA sequence harbor the RV locus remain conserved across different species at an extensive phylogenetic distance. Differing from protein-coding variants, it is usually required that RV locate in the range of 20–200 bp DNA sequence under purifying selection, such as conservative enhancer and promoter [212]. These conserved sections are interpreted as regulatory function units in which substitutions were rejected during natural selection and species evolution [213, 214]. Many genetic studies use evolutionarily conserved score, like phastCons and 28-way vertebrate alignment, as the putative benchmark for genomic regions that may have biological importance, even if the functional annotation of these regions is unknown [215–217]. To distinguish exact evolutionary signature for SNP site, base-wise scores for rejected substitutions are adopted including GERP++ and PhyloP [183, 184]. Those information help researchers to efficiently predict and prioritize the putative casual variants, especially for the RVs in the intergenic regions with inadequate functional annotations (Table 2).

However, comparative genomics approaches can only discover limited number of RVs in the whole genome, and it will miss many non-conserved regions that RVs locate [218]. Only a small subset of CREs is likely to be discovered by rigorous evolutionary constraints like high conservation across all species in mammals [219]. In this regard, the statistical power will be low for genome-wide RVs discovery according to the different level conservation information. Recent study has found that GATA1 binds site for an enhancer of GHP68 only in a species-specific manner, indicating that orthologous hitting will be invalid [220]. Also, lineage-specific elements that evolve in the recent time, as well as loci under adaptive selection, could also be important targets when mapping and prioritizing RVs from evolutionary perspective [221, 185] (Table 2).

Directions and strategies on prioritization of regulatory variants

Functional annotation, prediction and prioritization are crucial in the downstream analysis to identify causal variants. Accurately estimating the function effect and ranking the most pathogenic variants are still challenging. Many factors can influence the sensitivity and specificity of true causal RVs prioritization. First of all, the temporal and spatial biological process will affect RVs to exert their function among different tissues/cell types [222]. The transcriptional signals, such as gene expression, histone modification and chromatin state, have been shown to express distinct pattern around the eQTL and GWAS loci in different tissues/cell types [33, 223]. Those dynamic regulation patterns stress the importance of tissues/cell type specificity when predicting the function of RVs. Second, the lack of sufficient genomic data in multiple dimensions (TF-binding profiles, epigenomic signals and chromatin states) limit the method's usage in some tissues/cell types. For example, there are more genomic data generated for ENCODE tier 1 cell lines (GM12878, K562, H1 human embryonic stem cells) than other cell lines, which limit the power to detect causal RVs in less studied cell types. In addition, the population difference could be a reason that result in the difficulty of identification genetic causalities for population specific diseases/traits. The risk allele of a disease-causal

variant in one population may not produce functional effect in another population because of haplotype structure, epistasis and other environment factors. Therefore, the population specificity should be incorporated in the practice of personalized variants prioritization in the future. Lastly, although there are many methods and bioinformatics tools (Table 2) available to predict and prioritize the deleteriousness and pathogenicity of RVs, they rely on varied annotations and adopted distinct statistical methods, which will significantly affect the prediction performance in terms of consistence, sensitivity and specificity. Considering the functional complexity of RVs in gene regulation, combinatorial integration of results generated from a variety of computational methods could be a better strategy.

Functional validation of regulatory variants

We can use existing genomic features as well as bioinformatics methods to prioritize the most probably damaging variants that affect gene transcriptional regulation for the investigated traits and to predict the functional mechanism of these RVs. However, selected candidates still could be a false positive hit, as the predictive power and condition-dependent gene regulation can be distinct in terms of different traits, individuals and cell types/tissues. The experimental function validation is needed to interrogate the true effect of RVs according to several rigorous study designs. Evolving techniques have enabled researchers to perform functional experiments to decide the effects of RVs in various ways.

To initially check the TF-binding affinity between different alleles, experiments like electrophoretic mobility shift assay and construct transfection followed by luciferase expression are frequently used to validate function of promoter, enhancer or other CREs *in vitro* [150–153]. Recently, in continued succession of GWAS, researcher have successfully revealed a common non-coding SNP (rs12740374) at the 1p13 locus that functions as an enhancer to create a C/EBP (CCAAT/enhancer binding protein) binding site and alter the hepatic expression of the SORT1 gene, which finally affects plasma LDL-C and very low-density lipoprotein particle levels [154]. Also, similar enhancer reporter assay showed that a transcriptional enhancer element in which the G allele of a casual variant rs554219 reduces the cyclin D1 protein levels and increases the risk of breast cancer by abolishing the binding of ELK4 TF [155]. Chromatin immunoprecipitation coupled with real-time polymerase chain reaction (ChIP-qPCR), ChIP-chip, as well as ChIP-seq assays are effective strategies to quantitatively measure the DNA-protein binding interactions from one to multiple genomic loci for a specific TF. These techniques have been used to map the binding difference between cells with different genetic background [156, 162, 224]. On the other hand, to directly verify the SNP effect on three dimensions, such as enhancer-promoter interaction, 3C and fluorescence in situ hybridization-related methods are broadly used as the supplementary experiment after conventional TF-binding assay [162, 163].

However, the true effect size of RV may vary between cultured cell lines and *in vivo* system, and this requires in-depth investigation on isogenic systems and animal models. Transgenic assay is able to construct a paired reporter genes with wild and mutated functional CREs linked in front of a low-activity promoter. Then an enzyme assays (stain for β-galactosidase) after cell transfection (usually inject into fertilized mouse egg or fly embryo) gives quantitative estimation of the CRE activities, and the difference of transcriptional activities explains the regulatory effects of RV [163, 225]. With the extensive application of genomic editing systems, including zinc finger nucleases, transcription-activator-like effector nucleases and clustered regularly interspaced short palindromic repeats (CRISPR/Cas), *in vivo* functional validation of RV according to powerful gene editing technologies will greatly speed up the understanding of trait-associated genetic variants in real gene transcription [226–228]. Finally, we can indirectly test the causal relationship between RV and its functional products *in vivo* in human samples, by correlating the genotype with actual target gene/protein expression on effective samples [156].

CONCLUSIONS

Complete identification of function relevant RV and its trait association require extensive investigations from upstream genetic mapping to downstream functional analysis and validation. Outcomes from GWAS, QTL and WES are continuously expanding the catalog of candidate RVs and greatly narrow down the searching space of truly functional loci in the risk haplotypes. The versatile consortia of genomic data by ENCODE, Roadmap Epigenomics

and GTEx projects significantly facilitate the annotating, interpreting and prioritizing RVs and its effect across different cell types and tissues. Up to now, only a small fraction of RVs and its exact function were successfully characterized, and most of functional explanations of RVs linking to human diseases almost focused on the change of TF-binding affinity and enhancer/promoter activity. However, the functional consequence of RVs acting on transcriptional gene regulation could be complex and relate to high-order chromatin state and epigenetic regulatory programs. Although several models and assumptions tried to understand molecular mechanisms of RVs, there is lack of valid, high-throughput and cost-efficient biotechnologies. Future works should extensively focus on RV functional validation from multiple perspectives.

Key Points

- Interpreting the functional role of genetic variants located in human genome regulatory regions, such as enhancers and promoters, is an indispensable step to understand molecular mechanism of human diseases/traits and evolution.
- Whole-genome association test, linkage analysis and quantitative trait locus mapping are three important methods to detect causal regulatory variants.
- Many human quantitative traits, including gene expression, protein expression, non-coding RNA expression, alternative splicing, chromatin accessibility, DNA methylation, histone modification and translational efficiency, can used to measure genotype–phenotype associations at cell type/tissue specific level.
- Data from large-scale genomic projects, such as ENCODE, Roadmap Epigenomics and GETx projects, will significantly promote functional annotation, prediction and prioritization of regulatory variants.

FUNDING

This work was supported by the Research Grants Council (781511M) of Hong Kong, genomic SRT of the University of Hong Kong, and NSFC (91229105) of China, Grant of Science Faculty of Hong Kong Baptist University (FRG1/13-14/008 and FRG2/12-13/066).

References

1. Abecasis GR, Auton A, Brooks LD, *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**:56–65.
2. Yngvadottir B, Macarthur DG, Jin H, *et al*. The promise and reality of personal genomics. *Genome Biol* 2009;**10**:237.
3. Peters BA, Kermani BG, Sparks AB, *et al*. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 2012;**487**:190–5.
4. Hamosh A, Scott AF, Amberger JS, *et al*. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**: D514–17.
5. Hindorff LA, Sethupathy P, Junkins HA, *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009;**106**:9362–7.
6. Li MJ, Wang P, Liu X, *et al*. GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res* 2012;**40**:D1047–54.
7. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* 2011;**12**:683–91.
8. Makrythanasis P, Antonarakis SE. Pathogenic variants in non-protein-coding sequences. *Clin Genet* 2013;**84**:422–8.
9. Benovoy D, Kwan T, Majewski J. Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments. *Nucleic Acids Res* 2008;**36**:4417–23.
10. Pickrell JK, Marioni JC, Pai AA, *et al*. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010;**464**:768–72.
11. Zhao K, Lu ZX, Park JW, *et al*. GLiMMPS: Robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biol* 2013;**14**:R74.
12. Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* 2007;**8**:749–61.
13. Fraser HB, Xie X. Common polymorphic transcript variation in human disease. *Genome Res* 2009;**19**:567–75.
14. Wan Y, Qu K, Zhang QC, *et al*. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 2014;**505**:706–9.
15. Sethupathy P, Collins FS. MicroRNA target site polymorphisms and human disease. *Trends Genet* 2008;**24**:489–97.
16. Carbonell J, Alloza E, Arce P, *et al*. A map of human microRNA variation uncovers unexpectedly high levels of variability. *Genome Med* 2012;**4**:62.
17. Halvorsen M, Martin JS, Broadaway S, *et al*. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet* 2010;**6**:e1001074.
18. Martin JS, Halvorsen M, Davis-Neulander L, *et al*. Structural effects of linkage disequilibrium on the transcriptome. *RNA* 2012;**18**:77–87.
19. Kumar V, Westra HJ, Karjalainen J, *et al*. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet* 2013;**9**:e1003201.
20. Bhartiya D, Jalali S, Ghosh S, *et al*. Distinct patterns of genetic variations in potential functional elements in long non-coding RNAs. *Hum Mut* 2013;**35**:192–201.
21. Cannarozzi G, Schraudolph NN, Faty M, *et al*. A role for codon order in translation dynamics. *Cell* 2010;**141**:355–67.
22. Fredrick K, Ibba M. How the sequence of a gene can tune its translation. *Cell* 2010;**141**:227–9.
23. Li Q, Makri A, Lu Y, *et al*. Genome-wide search for exonic variants affecting translational efficiency. *Nat Commun* 2013; **4**:2260.

24. Nackley AG, Shabalina SA, Tchivileva IE, *et al*. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* 2006;**314**:1930–3.
25. Tsai CJ, Sauna ZE, Kimchi-Sarfaty C, *et al*. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *J Mol Biol* 2008;**383**: 281–91.
26. Bernstein BE, Birney E, Dunham I, *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
27. Gerstein MB, Kundaje A, Hariharan M, *et al*. Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012;**489**:91–100.
28. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* 2012;**30**:1095–6.
29. Schaub MA, Boyle AP, Kundaje A, *et al*. Linking disease associations with regulatory information in the human genome. *Genome Res* 2012;**22**:1748–59.
30. Hardison RC. Genome-wide epigenetic data facilitate understanding of disease susceptibility association studies. *J Biol Chem* 2012;**287**:30932–40.
31. Maurano MT, Humbert R, Rynes E, *et al*. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 2012;**337**:1190–5.
32. Vernot B, Stergachis AB, Maurano MT, *et al*. Personal and population genomics of human regulatory variation. *Genome Res* 2012;**22**:1689–97.
33. Trynka G, Sandor C, Han B, *et al*. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* 2013;**45**:124–30.
34. Parker SC, Stitzel ML, Taylor DL, *et al*. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci USA* 2013;**110**:17921–6.
35. Arbiza L, Gronau I, Aksoy BA, *et al*. Genome-wide inference of natural selection on human transcription factor binding sites. *Nat Genet* 2013;**45**:723–29.
36. Fraser HB. Gene expression drives local adaptation in humans. *Genome Res* 2013;**23**:1089–96.
37. Gunderson KL, Steemers FJ, Lee G, *et al*. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* 2005;**37**:549–54.
38. Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet* 2010;**11**:31–46.
39. Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75.
40. Marchini J, Howie B, Myers S, *et al*. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;**39**:906–13.
41. Boyle AP, Hong EL, Hariharan M, *et al*. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 2012;**22**:1790–97.
42. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 2012;**40**:D930–4.
43. Li MJ, Wang LY, Xia Z, *et al*. GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Res* 2013;**41**:W150–8.
44. Hemminki K, Lorenzo Bermejo J, Forsti A. The balance between heritable and environmental aetiology of human disease. *Nat Rev Genet* 2006;**7**:958–65.
45. Wang WY, Barratt BJ, Clayton DG, *et al*. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005;**6**:109–18.
46. McCarthy MI, Abecasis GR, Cardon LR, *et al*. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008; **9**:356–69.
47. Altshuler DM, Gibbs RA, Peltonen L, *et al*. Integrating common and rare genetic variation in diverse human populations. *Nature* 2010;**467**:52–8.
48. On beyond GWAS. *Nat Genet* 2010;**42**:551.
49. Freedman ML, Monteiro AN, Gayther SA, *et al*. Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* 2011;**43**:513–18.
50. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009;**5**: e1000529.
51. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 2009;**84**:210–23.
52. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;**11**: 499–511.
53. Udler MS, Tyrer J, Easton DF. Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genet Epidemiol* 2010;**34**:463–8.
54. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;**26**:2190–1.
55. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. *Am J Hum Genet* 2013;**14**:379–89.
56. Li MJ, Sham PC, Wang J. Genetic variant representation, annotation and prioritization in the post-GWAS era. *Cell Res* 2012;**22**:1505–8.
57. Manolio TA, Collins FS, Cox NJ, *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.
58. Gorlov IP, Gorlova OY, Sunyaev SR, *et al*. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet* 2008;**82**:100–12.
59. Hunt KA, Mistry V, Bockett NA, *et al*. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* 2013;**498**:232–5.
60. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010;**11**:415–25.
61. Bamshad MJ, Ng SB, Bigham AW, *et al*. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011;**12**:745–55.
62. Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet* 2012;**21**:R1–9.
63. Tang H, Jin X, Li Y, *et al*. A large-scale screen for coding variants predisposing to psoriasis. *Nat Genet* 2013.

64. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008;**83**:311–21.
65. Wu MC, Lee S, Cai T, *et al*. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011;**89**:82–93.
66. Neale BM, Rivas MA, Voight BF, *et al*. Testing for an unusual distribution of rare variants. *PLoS Genet* 2011;**7**: e1001322.
67. Lee S, Emond MJ, Bamshad MJ, *et al*. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012;**91**:224–37.
68. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012;**13**: 762–75.
69. Edwards SL, Beesley J, French JD, *et al*. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet* 2013;**93**:779–97.
70. Hou L, Zhao H. A review of post-GWAS prioritization approaches. *Front Genet* 2013;**4**:280.
71. Zhang X, Bailey SD, Lupien M. Laying a solid foundation for Manhattan—'setting the functional basis for the post-GWAS era'. *Trends Genet* 2014;**30**:140–9.
72. Gilissen C, Hoischen A, Brunner HG, *et al*. Unlocking Mendelian disease using exome sequencing. *Genome Biol* 2011;**12**:228.
73. Asan , Xu Y, Jiang H, *et al*. Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol* 2011;**12**:R95.
74. Wang W, Wei Z, Lam TW, *et al*. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Sci Rep* 2011;**1**:55.
75. Bromberg Y. Building a genome analysis pipeline to predict disease risk and prevent disease. *J Mol Biol* 2013;**425**:3993–4005.
76. Pavlopoulos GA, Oulas A, Iacucci E, *et al*. Unraveling genomic variation from next generation sequencing data. *BioData Mining* 2013;**6**:13.
77. Xu F, Wang W, Wang P, *et al*. A fast and accurate SNP detection algorithm for next-generation sequencing data. *Nat Commun* 2012;**3**:1258.
78. DePristo MA, Banks E, Poplin R, *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;**43**:491–8.
79. Li MX, Gui HS, Kwan JS, *et al*. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res* 2012;**40**:e53.
80. Gilissen C, Hoischen A, Brunner HG, *et al*. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* 2012;**20**:490–7.
81. Weatheritt RJ, Babu MM. Evolution. The hidden codes that shape protein evolution. *Science* 2013;**342**:1325–6.
82. Stergachis AB, Haugen E, Shafer A, *et al*. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* 2013;**342**:1367–72.
83. Swami M. Whole-genome sequencing identifies Mendelian mutations. *Nat Rev Genet* 2010;**11**:313.
84. Cooper DN, Chen JM, Ball EV, *et al*. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Hum Mut* 2010;**31**:631–55.
85. Weedon MN, Cebola I, Patch AM, *et al*. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* 2013.
86. Herdewyn S, Zhao H, Moisse M, *et al*. Whole-genome sequencing reveals a coding non-pathogenic variant tagging a non-coding pathogenic hexanucleotide repeat expansion in C9orf72 as cause of amyotrophic lateral sclerosis. *Hum Mol Genet* 2012;**21**:2412–19.
87. McDaniell R, Lee BK, Song L, *et al*. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* 2010;**328**:235–9.
88. Mackay TF, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet* 2009; **10**:565–77.
89. Dubois PC, Trynka G, Franke L, *et al*. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* 2010;**42**:295–302.
90. Westra HJ, Peters MJ, Esko T, *et al*. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* 2013;**45**:1238–43.
91. Edwards MD, Stuber CW, Wendel JF. Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics* 1987;**116**:113–25.
92. Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 1989;**121**:185–99.
93. Zeng ZB. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc Natl Acad Sci USA* 1993;**90**:10972—6.
94. Jansen RC. Interval mapping of multiple quantitative trait loci. *Genetics* 1993;**135**:205–11.
95. Jiang C, Zeng ZB. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 1995;**140**:1111–27.
96. Doerge RW. Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* 2002;**3**:43–52.
97. Frazer KA, Ballinger DG, Cox DR, *et al*. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;**449**:851–61.
98. Sun W, Hu Y. eQTL mapping using RNA-seq data. *Stat Biosci* 2013;**5**:198–219.
99. Lappalainen T, Sammeth M, Friedlander MR, *et al*. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 2013;**501**:506–11.
100. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013;**45**:580–5.
101. Michaelson JJ, Loguercio S, Beyer A. Detection and interpretation of expression quantitative trait loci (eQTL). *Methods* 2009;**48**:265–76.
102. Brown CD, Mangravite LM, Engelhardt BE. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet* 2013;**9**:e1003649.
103. Qin J, Li MLJ, Wang PW, *et al*. ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. *Nucleic Acids Res* 2011;**39**:W430–6.
104. Qin J, Li MJ, Wang P, *et al*. ProteoMirExpress: inferring microRNA and protein-centered regulatory networks from high-throughput proteomic and mRNA expression data. *Mol Cell Proteomics* 2013;**12**:3379–87.

105. Nicolae DL, Gamazon E, Zhang W, *et al*. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 2010;**6**:e1000888.
106. He X, Fuller CK, Song Y, *et al*. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am J Hum Genet* 2013;**92**:667–80.
107. Conde L, Bracci PM, Richardson R, *et al*. Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma. *Am J Hum Genet* 2013;**92**:126–30.
108. Degner JF, Pai AA, Pique-Regi R, *et al*. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 2012;**482**:390–4.
109. Margueron R, Reinberg D. Chromatin structure and the inheritance of epigenetic information. *Nat Rev Genet* 2010; **11**:285–96.
110. Stower H. Gene regulation: from genetic variation to phenotype via chromatin. *Nat Rev Genet* 2013;**14**:824.
111. Furey TS, Sethupathy P. Genetics driving epigenetics. *Science* 2013;**342**:705–6.
112. Kilpinen H, Waszak SM, Gschwind AR, *et al*. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* 2013;**342**:744–7.
113. McVicker G, van de Geijn B, Degner JF, *et al*. Identification of genetic variants that affect histone modifications in human cells. *Science* 2013;**342**:747–9.
114. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, *et al*. Extensive variation in chromatin states across humans. *Science* 2013;**342**:750–2.
115. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012;**13**: 484–92.
116. Schubeler D. Molecular biology. Epigenetic islands in a genetic ocean. *Science* 2012;**338**:756–57.
117. Gibbs JR, van der Brug MP, Hernandez DG, *et al*. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* 2010;**6**: e1000952.
118. Heyn H, Moran S, Hernando-Herraez I, *et al*. DNA methylation contributes to natural human variation. *Genome Res* 2013;**23**:1363–72.
119. Muers M. Gene expression: Disentangling DNA methylation. *Nat Rev Genet* 2013;**14**:519.
120. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, *et al*. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* 2013;**2**: e00523.
121. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature* 2010;**463**:457–63.
122. Wang ET, Sandberg R, Luo S, *et al*. Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008; **456**:470–6.
123. Katz Y, Wang ET, Airoldi EM, *et al*. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* 2010;**7**:1009–15.
124. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, *et al*. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 2010;**464**:773–7.
125. Wu L, Candille SI, Choi Y, *et al*. Variation and genetic control of protein abundance in humans. *Nature* 2013;**499**: 79–82.
126. Costanzo M, Baryshnikova A, Bellay J, *et al*. The genetic landscape of a cell. *Science* 2010;**327**:425–31.
127. Park S, Lehner B. Epigenetic epistatic interactions constrain the evolution of gene expression. *Mol Syst Biol* 2013;**9**:645.
128. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Heredity* 2003;**56**:73–82.
129. Moore JH, Williams SM. Epistasis and its implications for personal genetics. *Am J Hum Genet* 2009;**85**:309–20.
130. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;**10**:392–404.
131. Zuk O, Hechter E, Sunyaev SR, *et al*. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* 2012;**109**:1193–8.
132. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am J Hum Genet* 2010;**86**:6–22.
133. Wang Y, Liu G, Feng M, *et al*. An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics* 2011;**27**:2936–43.
134. Shang J, Zhang J, Sun Y, *et al*. Performance analysis of novel methods for detecting epistasis. *BMC Bioinformatics* 2011;**12**:475.
135. Zhang X, Huang S, Zou F, *et al*. TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics* 2010;**26**:i217–27.
136. Wan X, Yang C, Yang Q, *et al*. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet* 2010;**87**:325–40.
137. Yang C, He Z, Wan X, *et al*. SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics* 2009;**25**: 504–11.
138. Wu J, Devlin B, Ringquist S, *et al*. Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genet Epidemiol* 2010;**34**:275–85.
139. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nat Genet* 2007;**39**: 1167–3.
140. Wan X, Yang C, Yang Q, *et al*. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics* 2010;**26**:30–7.
141. Lin X, Lee S, Christiani DC. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* 2013;**14**:667–81.
142. Thomas D. Gene—environment-wide association studies: emerging approaches. *Nat Rev Genet* 2010;**11**:259–72.
143. Huang Y, Wuchty S, Przytycka TM. eQTL epistasis—challenges and computational approaches. *Front Genet* 2013;**4**:51.
144. Zhang W, Zhu J, Schadt EE, *et al*. A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Computat Biol* 2010;**6**:e1000642.
145. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 2011;**88**:440–9.
146. Li MX, Kwan JS, Bao SY, *et al*. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet* 2013;**9**: e1003143.

147. Sifrim A, Popovic D, Tranchevent LC, *et al*. eXtasy: variant prioritization by genomic data fusion. *Nat Methods* 2013;**10**: 1083–4.

148. Whyte WA, Orlando DA, Hnisz D, *et al*. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 2013;**153**:307–19.

149. Hnisz D, Abraham BJ, Lee TI, *et al*. Super-enhancers in the control of cell identity and disease. *Cell* 2013;**155**:934–7.

150. Phornphutkul C, Anikster Y, Huizing M, *et al*. The promoter of a lysosomal membrane transporter gene, CTNS, binds Sp-1, shares sequences with the promoter of an adjacent gene, CARKL, and causes cystinosis if mutated in a critical region. *Am J Hum Genet* 2001;**69**:712–21.

151. Niimi T, Munakata M, Keck-Waggoner CL, *et al*. A polymorphism in the human UGRP1 gene promoter that regulates transcription is associated with an increased risk of asthma. *Am J Hum Genet* 2002;**70**:718–25.

152. Hu XZ, Lipsky RH, Zhu G, *et al*. Serotonin transporter promoter gain-of-function genotypes are linked to obsessive-compulsive disorder. *Am J Hum Genet* 2006;**78**:815–26.

153. Theuns J, Brouwers N, Engelborghs S, *et al*. Promoter mutations that increase amyloid precursor-protein expression are associated with Alzheimer disease. *Am J Hum Genet* 2006;**78**:936–46.

154. Musunuru K, Strong A, Frank-Kamenetsky M, *et al*. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 2010;**466**:714–19.

155. French JD, Ghoussaini M, Edwards SL, *et al*. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *Am J Hum Genet* 2013;**92**:489–503.

156. Tuupanen S, Turunen M, Lehtonen R, *et al*. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* 2009;**41**:885–90.

157. Xie X, Mikkelsen TS, Gnirke A, *et al*. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci USA* 2007;**104**:7145–50.

158. Engel N, Raval AK, Thorvaldsen JL, *et al*. Three-dimensional conformation at the H19/Igf2 locus supports a model of enhancer tracking. *Hum Mol Genet* 2008;**17**: 3021–9.

159. Tolstorukov MY, Volfovsky N, Stephens RM, *et al*. Impact of chromatin structure on sequence variability in the human genome. *Nat Struct Mol Biol* 2011;**18**:510–15.

160. Prendergast JG, Semple CA. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome Res* 2011;**21**:1777–87.

161. Gaffney DJ, McVicker G, Pai AA, *et al*. Controls of nucleosome positioning in the human genome. *PLoS Genet* 2012;**8**:e1003036.

162. Cowper-Sal lari R, Zhang X, Wright JB, *et al*. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat Genet* 2012;**44**:1191–8.

163. Wasserman NF, Aneas I, Nobrega MA. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res* 2010;**20**:1191–7.

164. Davison LJ, Wallace C, Cooper JD, *et al*. Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. *Hum Mol Genet* 2012; **21**:322–33.

165. Wright JB, Brown SJ, Cole MD. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol Cell Biol* 2010;**30**:1411–20.

166. Camps J, Grade M, Nguyen QT, *et al*. Chromosomal breakpoints in primary colon cancer cluster at sites of structural variants in the genome. *Cancer Res* 2008;**68**: 1284–95.

167. Gonzalez E, Kulkarni H, Bolivar H, *et al*. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 2005;**307**:1434–40.

168. Zhou J, Lemos B, Dopman EB, *et al*. Copy-number variation: the balance between gene dosage and expression in Drosophila melanogaster. *Genome Biol Evol* 2011;**3**:1014–24.

169. Pasmant E, Sabbagh A, Vidaud M, *et al*. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J* 2011;**25**:444–8.

170. Zhang H, Zeitz MJ, Wang H, *et al*. Long noncoding RNA-mediated intrachromosomal interactions promote imprinting at the Kcnq1 locus. *J. Cell Biol* 2014;**204**:61–75.

171. Matys V, Fricke E, Geffers R, *et al*. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 2003;**31**:374–8.

172. Sandelin A, Alkema W, Engstrom P, *et al*. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 2004;**32**:D91–4.

173. Newburger DE, Bulyk ML. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* 2009;**37**:W202–8.

174. Bailey TL, Boden M, Buske FA, *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009;**37**:W202–8.

175. Zhang G, Chen X, Chan L, *et al*. An SNP selection strategy identified IL-22 associating with susceptibility to tuberculosis in Chinese. *Sci Reports* 2011;**1**:20.

176. Kasowski M, Grubert F, Heffelfinger C, *et al*. Variation in transcription factor binding among humans. *Science* 2010; **328**:232–5.

177. Williamson I, Hill RE, Bickmore WA. Enhancers: from developmental genetics to the genetics of common human disease. *Dev Cell* 2011;**21**:17–19.

178. Macintyre G, Bailey J, Haviv I, *et al*. is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics* 2010;**26**:i524–30.

179. Thomas-Chollier M, Hufton A, Heinig M, *et al*. Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nat Protoc* 2011;**6**:1860–9.

180. Barenboim M, Manke T. ChroMoS: an integrated web tool for SNP classification, prioritization and functional interpretation. *Bioinformatics* 2013;**29**:2197–8.

181. Guo L, Du Y, Chang S, *et al*. rSNPBase: a database for curated regulatory SNPs. *Nucleic Acids Res* 2014;**42**: D1033–9.

182. Hu H, Huff CD, Moore B, *et al*. VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Gene Epidemiol* 2013;**37**:622–34.

183. Davydov EV, Goode DL, Sirota M, *et al*. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010;**6**: e1001025.

184. Pollard KS, Hubisz MJ, Rosenbloom KR, *et al*. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 2010;**20**:110–21.

185. Li MJ, Wang LY, Xia Z, *et al*. dbPSHP: a database of recent positive selection across human populations. *Nucleic Acids Res* 2014;**42**:D910–16.

186. Khurana E, Fu Y, Colonna V, *et al*. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 2013;**342**:1235587.

187. Kircher M, Witten DM, Jain P, *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**:310–15.

188. Ritchie GR, Dunham I, Zeggini E, *et al*. Functional annotation of noncoding sequence variants. *Nat Methods* 2014; **11**:294–6.

189. Wang J, Zhuang J, Iyer S, *et al*. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res* 2013; **41**:D171–6.

190. Ameur A, Rada-Iglesias A, Komorowski J, *et al*. Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP. *Nucleic Acids Res* 2009;**37**:e85.

191. Bryzgalov LO, Antontseva EV, Matveeva MY, *et al*. Detection of Regulatory SNPs in Human Genome Using ChIP-seq ENCODE Data. *PloS One* 2013;**8**: e78833.

192. Li MJ, Sham PC, Wang J. FastPval: a fast and memory efficient program to calculate very low P-values from empirical distribution. *Bioinformatics* 2010;**26**:2897–9.

193. Ptashne M. On the use of the word 'epigenetic'. *Curr Biol* 2007;**17**:R233–6.

194. Beisel C, Paro R. Silencing chromatin: comparing modes and mechanisms. *Nat Rev Genet* 2011;**12**:123–35.

195. Moazed D. Mechanisms for the inheritance of chromatin states. *Cell* 2011;**146**:510–18.

196. Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 2011;**12**:7–18.

197. de Wit E, de Laat W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* 2012;**26**:11–24.

198. Huang L, Yu D, Wu C, *et al*. Copy number variation at 6q13 functions as a long-range regulator and is associated with pancreatic cancer risk. *Carcinogenesis* 2012; **33**:94–100.

199. De S, Michor F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat Biotechnol* 2011;**29**:1103–8.

200. Noordermeer D, de Wit E, Klous P, *et al*. Variegated gene expression caused by cell-specific long-range DNA interactions. *Nat Cell Biol* 2011;**13**:944–51.

201. Sanyal A, Lajoie BR, Jain G, *et al*. The long-range interaction landscape of gene promoters. *Nature* 2012;**489**:109–13.

202. Deng W, Lee J, Wang H, *et al*. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* 2012;**149**:1233–44.

203. Jin F, Li Y, Dixon JR, *et al*. A high-resolution map of the three–dimensional chromatin interactome in human cells. *Nature* 2013;**503**:290–4.

204. Nagano T, Lubling Y, Stevens TJ, *et al*. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013;**502**:59–64.

205. Zhang Y, Wong CH, Birnbaum RY, *et al*. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* 2013;**504**:306–10.

206. Duggal G, Wang H, Kingsford C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic Acids Res* 2014;**42**:87–96.

207. Pomerantz MM, Ahmadiyeh N, Jia L, *et al*. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* 2009;**41**:882–4.

208. Smemo S, Tena JJ, Kim KH, *et al*. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 2014;**507**:371–5.

209. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* 2013. doi: 10.1093/nar/gkt1249 (Advance online publication 13 December 2013).

210. Guan D, Shao J, Deng Y, *et al*. CMGRN: a web server for constructing multilevel gene regulatory networks using ChIP-seq and gene expression data. *Bioinformatics* 2014; **30**:1190–2.

211. Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, *et al*. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res* 2012;**22**:1735–47.

212. Pennacchio LA, Ahituv N, Moses AM, *et al*. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 2006;**444**:499–502.

213. Hardison RC. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* 2000;**16**: 369–72.

214. Dermitzakis ET, Reymond A, Antonarakis SE. Conserved non-genic sequences—an unexpected feature of mammalian genomes. *Nat Rev Genet* 2005;**6**:151–7.

215. Emison ES, McCallion AS, Kashuk CS, *et al*. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature* 2005;**434**: 857–63.

216. Siepel A, Bejerano G, Pedersen JS, *et al*. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;**15**:1034–50.

217. Miller W, Rosenbloom K, Hardison RC, *et al*. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* 2007;**17**:1797–808.

218. Chatterjee S, Bourque G, Lufkin T. Conserved and non-conserved enhancers direct tissue specific transcription in ancient germ layer specific developmental control genes. *BMC Dev Biol* 2011;**11**:63.

219. King DC, Taylor J, Zhang Y, *et al*. Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. *Genome Res* 2007;**17**:775–86.

220. Schmidt D, Wilson MD, Ballester B, *et al*. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 2010;**328**:1036–40.

221. Torkamani A, Kannan N, Taylor SS, *et al*. Congenital disease SNPs target lineage specific structural elements in protein kinases. *Proc Natl Acad Sci USA* 2008;**105**:9011–16.

222. Dimas AS, Deutsch S, Stranger BE, *et al*. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 2009;**325**:1246–50.

223. Fu J, Wolfs MG, Deelen P, *et al*. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet* 2012;**8**:e1002431.

224. Vernes SC, Spiteri E, Nicod J, *et al*. High-throughput analysis of promoter occupancy reveals direct neural targets of FOXP2, a gene mutated in speech and language disorders. *Am J Hum Genet* 2007;**81**:1232–50.

225. Bhatia S, Bengani H, Fish M, *et al*. Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. *Am J Hum Genet* 2013;**93**:1126–34.

226. Urnov FD, Rebar EJ, Holmes MC, *et al*. Genome editing with engineered zinc finger nucleases. *Nat Rev Genet* 2010; **11**:636–46.

227. Miller JC, Tan S, Qiao G, *et al*. A TALE nuclease architecture for efficient genome editing. *Nat Biotechnol* 2011;**29**: 143–8.

228. Cong L, Ran FA, Cox D, *et al*. Multiplex genome engineering using CRISPR/Cas systems. *Science* 2013; **339**:819–23.