Functional interpretation of non-coding sequence variation: Concepts and challenges

Dirk S. Paul[1)], Nicole Soranzo[2)3)] and Stephan Beck[1)]*

Understanding the functional mechanisms underlying genetic signals associated with complex traits and common diseases, such as cancer, diabetes and Alzheimer's disease, is a formidable challenge. Many genetic signals discovered through genome-wide association studies map to non-protein coding sequences, where their molecular consequences are difficult to evaluate. This article summarizes concepts for the systematic interpretation of non-coding genetic signals using genome annotation data sets in different cellular systems. We outline strategies for the global analysis of multiple association intervals and the in-depth molecular investigation of individual intervals. We highlight experimental techniques to validate candidate (potential causal) regulatory variants, with a focus on novel genome-editing techniques including CRISPR/Cas9. These approaches are also applicable to low-frequency and rare variants, which have become increasingly important in genomic studies of complex traits and diseases. There is a pressing need to translate genetic signals into biological mechanisms, leading to prognostic, diagnostic and therapeutic advances.

Keywords:
complex traits; chromatin; genome editing; gene regulation; GWAS; regulatory variants

Introduction

The quest for identifying sequence variants associated with complex traits including common diseases has been greatly facilitated by technological progress in high-throughput DNA analysis, such as genotyping arrays and next-generation sequencing, complemented by advances in bioinformatics [1]. In recent years, researchers have systematically assayed millions of common genetic variants across hundreds of thousands of individuals in genome-wide association studies (GWAS). In GWAS, the allele frequencies of a set of sequence variants are statistically compared between individuals with a phenotype of interest (such as a clinical condition) and the general population. This results in the detection of sequence variants that show association with the phenotype. Despite the remarkable success of GWAS, there is a substantial gap between the plethora of associated sequence variants and our understanding of how most of these variants contribute to complex trait biology [2–4]. At least three key issues have impeded the functional translation of GWAS signals.

First, GWAS have focused on common SNPs (MAF $>$ 5%). SNPs either individually or in combination typically explain only a small fraction of the genetic variance of most complex traits [5, 6]. As a consequence, phenotypic effects due to the perturbation of trait-associated SNPs are likely to be subtle. This implies that large numbers of independent studies are required to estimate quantitatively the phenotypic impact of each genetic variant. Furthermore, highly sensitive assays in sufficiently large sample sizes may be required for their downstream assessment and validation.

DOI 10.1002/bies.201300126

[1)] UCL Cancer Institute, University College London, London, United Kingdom
[2)] Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom
[3)] Department of Haematology, University of Cambridge, Cambridge, United Kingdom

***Corresponding author:**
Dirk S. Paul
E-mail: d.paul@ucl.ac.uk

Abbreviations:
ChIP, chromatin immunoprecipitation; **CRISPR,** clustered regularly interspaced short palindromic repeats; **ENCODE,** Encyclopedia of DNA Elements; **eQTL,** expression quantitative trait locus; **GWAS,** genome-wide association study; **iPSC,** induced pluripotent stem cell; **LD,** linkage disequilibrium; **MAF,** minor allele frequency; **SNP,** single-nucleotide polymorphism; **TALEN,** transcription activator-like effector nuclease.

Second, the associated sequence variant identified in GWAS may in fact only be linked to, rather than itself be, the causal variant. This phenomenon is known as linkage disequilibrium (LD) [7]. The alleles of the index (lead) SNP, i.e. the sequence variant showing the strongest association with a trait of interest, are correlated with the alleles of multiple nearby proxy SNPs. The combination of these alleles form haplotypes along the chromosome and are transmitted together. Importantly, such haplotype structures are population-specific [8–10]. On genotyping platforms, only a few selected index SNPs per LD region are measured. In fact, the platforms exploit LD patterns in a way that the selected SNPs capture most of the genetic variation at any given locus [8]. In individuals of Northern European ancestry, LD structure extends to around 50 kilobases (with substantial variation [10]) and usually harbors several genes and transcripts. Because of LD, it may not be possible to discriminate statistically between multiple variants associated with a phenotype at a determined locus. As discussed in this review, the identification of suitable functional genome maps may be particularly helpful in prioritizing efforts for these loci.

Third, the vast majority (over 90%) of associated variants have been found to localize outside of known protein-coding sequences, thus impeding the direct interpretation of their functional effects [11]. To this end, the trait-associated variant may be involved in regulation of gene expression, which is chiefly dependent on cell type identity, developmental stage and environmental factors [12]. The variant may reside at gene regulatory elements, such as promoters, enhancers, silencers, and insulators, where it perturbs binding sites of transcription factors, local chromatin structure or co-factor recruitment, ultimately resulting in changes of transcriptional output of the nearby gene(s) [12]. Trait-associated variants that are located distal to transcription start sites at gene-dense regions are particularly difficult to interpret. Here, literature search or more precisely, experimental validation may determine selection of a candidate gene. In our article, we mainly discuss the influence of sequence variation on regulatory elements of protein-coding genes, but we recognize that sites affecting the transcription of non-protein coding RNAs may also play an important role in gene regulation.

The challenge ahead is to carve out suitable strategies to gain insights into cell type-specific molecular processes and pathways underlying the discovered GWAS signals. In this article, we (i) describe principles for interpreting non-coding genetic signals using public genome annotation resources; (ii) discuss considerations when using annotation data sets in primary cells, cell lines, and – looking ahead – differentiated induced pluripotent stem cells (iPSCs); and (iii) outline current and emerging strategies to prioritize and experimentally validate candidate regulatory variants and genes.

Genome annotation resources can guide the interpretation of genetic variation

The genomic positions of GWAS index SNPs and proxy SNPs can be compared to the positions of biochemical events referenced in publicly available annotation resources to help tease out the functional variant(s) from the vast number of trait-associated variants in LD. Such biochemical events include sites of transcription factor and micro RNA (miRNA) binding, chromatin accessibility and modifications, DNA methylation, and many other types. A trait-associated variant that overlaps with a regulatory element may be functionally relevant. The overlap also directly suggests a hypothesis with respect to the mechanism underlying the association, which can be tested in experimental assays. However, there are a number of caveats to this approach that need to be correctly evaluated.

First, it must be noted that a large degree of non-functional overlap can be expected, because of the widespread distribution of the biochemical events. Thus, it is necessary to use unbiased approaches to evaluate which of the overlaps are functionally relevant and which occur by chance [13].

Second, regulatory elements that influence gene expression may operate in a spatial- and temporal-dependent manner [14]. Therefore, annotation data should ideally be retrieved for a cell type and developmental stage that is most relevant to the trait under investigation. For example, genetic signals associated with type 2 diabetes may be annotated using data sets obtained in pancreatic islets [13, 15]. However, for many complex traits including common diseases, the identity of relevant cell types is not obvious (e.g. for height or longevity), or appropriate cell types and tissues are difficult – or even impossible – to obtain for experimental assays (e.g. cells from the cerebral cortex in Alzheimer's disease). Several common disorders have been associated with distinct developmental phases such as fetal stages in metabolic syndrome and adult stages in age-related diseases [16]. Indeed, a recent study showed enrichment of GWAS signals associated with cardiovascular disease at regulatory regions in fetal tissue, and depletion of signals linked to breast cancer and Alzheimer's disease [11].

More and more genome annotation data sets are becoming available, including those created by the Encyclopedia of DNA Elements (ENCODE) Project Consortium [12], NIH Roadmap Epigenomics Mapping Consortium [17], and the recently launched BLUEPRINT Consortium [18]. While the ENCODE data comprises mainly transformed cell lines (e.g. due to practical reasons such as their wide availability and capacity to produce large numbers of cells), the Roadmap Epigenomics and BLUEPRINT data almost exclusively consist of a broad selection of primary, ex vivo tissues corresponding to normal tissues and organ systems that are involved in human disease processes. These consortia have also contributed towards setting standards for experimental protocols, reagents, and bioinformatic tools. In particular, genome browsers are valuable for accessing and visualizing the consortia's multi-dimensional data sets [19, 20].

The growing number of genome annotation data sets will enable the functional interpretation of GWAS signals in an increasingly context-specific manner. Alongside these, it will be necessary to refine computational methods to distil the vast amount of data into discrete segments of interpretable biological function. Segmentation approaches determine patterns and similarities between individual chromatin data sets to summarize them into a small set of "chromatin states" [21–23]. Nonetheless, the functionality of the predicted chromatin states, such as "strong" and "weak

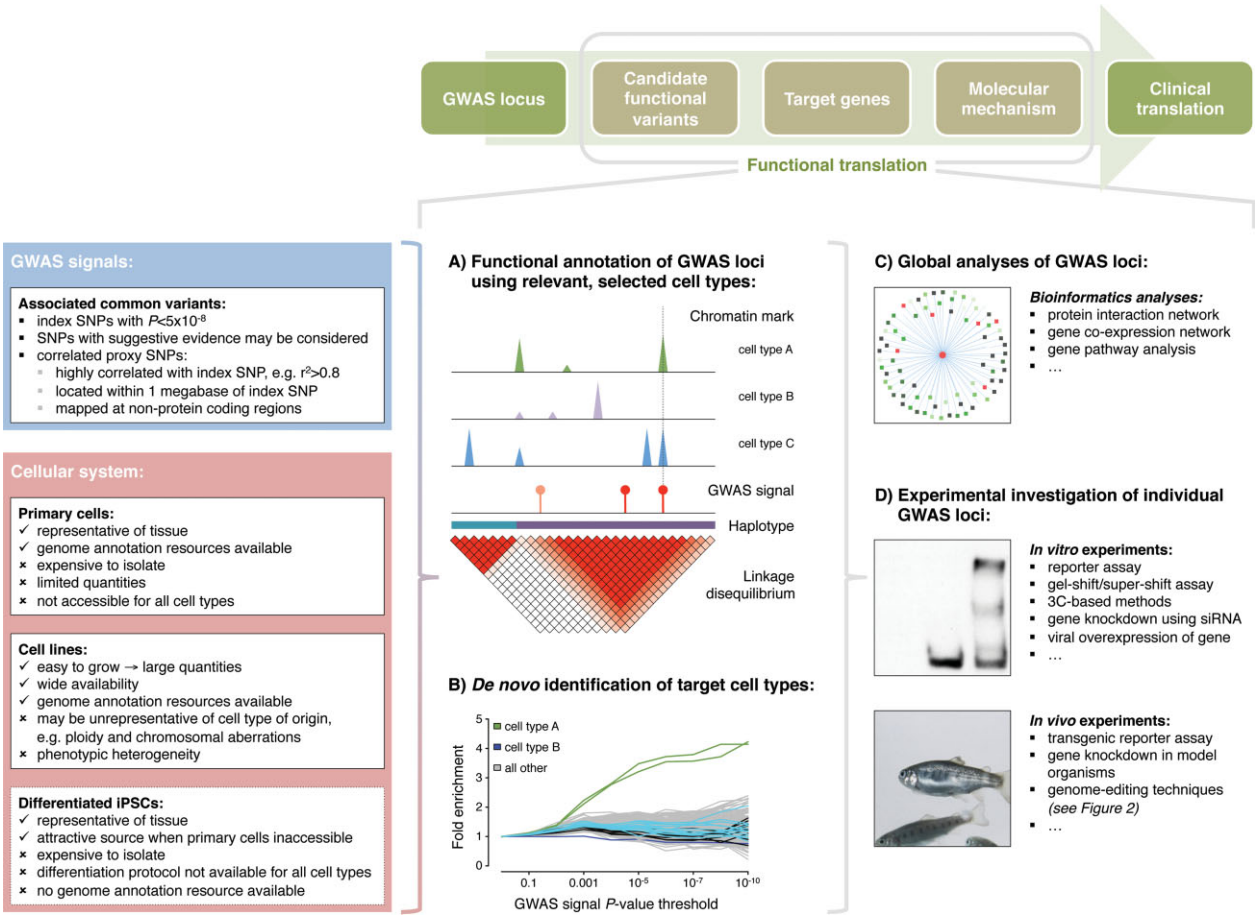

Figure 1. Genome annotation data in different cellular systems guide the functional interpretation of genetic variation. The pros and cons of annotation data sets obtained in different cellular systems are indicated on the left panel. Currently, publicly available resources mainly consist of data sets derived from primary cell cultures and transformed cell lines, while iPSCs may become more prominent in the future. Indeed, iPSC-derived cells may represent the key technological advance for assaying inaccessible cell types for which primary cultures cannot be obtained. Functional annotation of genetic signals at GWAS intervals can be restricted to selected cell types that are most relevant to the trait of interest (**A**), or unrestricted using all available cell types in an annotation resource (**B**). The latter approach may be valuable if target cell types of the trait are not yet established. To gain biological insights into the genetic architecture of complex traits, all GWAS signals may be collectively analyzed and associated with gene pathways and networks using bioinformatic tools (**C**). In parallel, individual GWAS intervals may be studied in depth using a range of in vitro and in vivo experimental assays (**D**). The use of emerging genome-editing techniques is illustrated in detail in Fig. 2. Abbreviations: LD, linkage disequilibrium; siRNA, small interfering RNA; TALEN, transcription activator-like effector nuclease; CRISPR, clustered regularly interspaced short palindromic repeats; 3C, chromosome conformation capture.

enhancer", as well as their predicted cell type specificity, need experimental validation.

Choosing a suitable cellular system for the annotation of GWAS loci

Genome-wide annotation data sets can be obtained in several cellular systems, mainly from primary cell cultures and transformed cell lines, while in the future, differentiated iPSCs are likely to become more prominent for this application [24]. Each cellular system has its pros and cons when applied to the prioritization of candidate regulatory variants at GWAS loci (Fig. 1).

Primary cells are directly representative of the tissues and organs from which they were isolated. However, the isolation of homogeneous cell populations is challenging, involving preparative procedures such as fluorescent-activated sorting of cells. Therefore, many primary cell populations are available in limited quantities for experimental assays. In addition, chromatin maps generated in primary cells and tissues represent only a snapshot of the development stage during which they were isolated.

Cultured cell lines and transformed (immortalized) cell lines mostly retain the characteristics of the primary tissue from which they were derived. However, chromosomal rearrangements, changes to chromatin structure, DNA methylation, and gene expression profiles may arise through, for example, serial subculturing of the cell lines [25]. These non-physiological transformations can lead to the appearance of artificial biochemical activities and misleading annotation

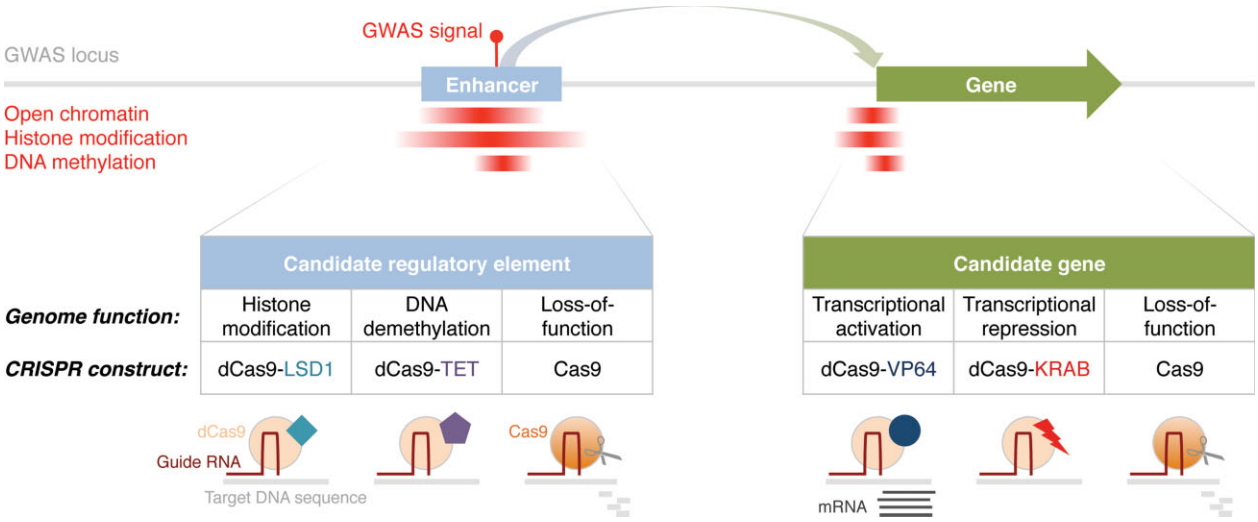

Figure 2. Investigating molecular consequences of candidate functional variants using CRISPR/Cas9. The advent of novel genome-editing techniques, such as CRISPR/Cas9, enables exciting new opportunities for validating GWAS candidate regulatory sites and genes. CRISPR/Cas9 in conjunction with customizable guide RNA can be used to precisely target genomic sites of interest to induce loss-of-function alterations. In addition, CRISPR-associated catalytically inactive Cas9 protein (dCas9) can be fused to different effector domains, including VP64 (activation), KRAB (repression), LSD1 (histone demethylation, specifically H3K4me2 and H3K27ac), and TET family proteins (DNA demethylation). Upon introduction of the CRISPR/(d)Cas9-complex into a cellular system, the molecular consequences of the genome editing can be further investigated.

of genetic variants. Indeed, we recently compared enrichment patterns of hematological trait-associated variants at sites of open chromatin in a selection of primary hematopoietic cells and their representative cell lines [26]. We observed a much larger number of both open chromatin sites and overlaps of these sites with trait-associated variants in cell lines compared to primary cells. The enrichment patterns in cell lines showed decreased specificity with respect to cell type identity.

The prioritization of candidate regulatory variants at GWAS loci by means of publicly available annotation resources suffers from a trade-off between sensitivity and specificity [27]. That is, GWAS signals can be annotated using functional genome maps derived from either selected cell types that are of most relevance to the phenotype of interest (Fig. 1A) or from across all available cell types (Fig. 1B). Limiting the annotation to selected cell types may only be considered if extensive prior knowledge about the target cell and tissue type exists. For example, this applies to some extent to the cellular components of the well-characterized hematopoietic system [26, 28]. Conversely, de novo identification of target cell types (i.e. out of all available cell types collected in an annotation resource) has emerged as key application of the functional annotation of GWAS signals and has already revealed novel target cell types for a number of common diseases. These include IL-17-producing T helper (TH17) cells in Crohn's disease and $CD19^{+}/CD20^{+}$ B cells in multiple sclerosis. Indeed, this approach may recognize pathogenic cell types without prior knowledge of disease-relevant molecular processes [11].

"Scientists, choose your weapons"

Besides choosing a cellular system that is biologically relevant for the screening of candidate functional variants at GWAS loci, as outlined above, attention has to be paid to choosing a suitable genome annotation mark. In this respect, we advocate the use of a *general* hallmark of regulatory potential, such as open chromatin provided by deoxyribonuclease I (DNase I)-seq, formaldehyde-assisted isolation of regulatory elements (FAIRE)-seq, or the recently introduced assay for transposase-accessible chromatin (ATAC)-seq [29], as opposed to an *individual* mark, such as transcription factor binding provided by chromatin immunoprecipitation (ChIP)-seq.

Sites of open chromatin are associated with most classes of active gene regulatory elements, and as such, are specific to cell type, developmental stage, and other influencing factors [30, 31]. In contrast, in transcription factor ChIP-seq experiments, antibodies against a distinct DNA-binding protein are applied. The application of open chromatin as a screening tool for candidate functional variants is both informative and cost-effective, because one would need to obtain a substantial number of transcription factor ChIP-seq experiments to obtain a comparable information value. Besides, ChIP-seq data thus far generated by the ENCODE Consortium are scarce with respect to the number of cell types in which most transcription factors have been assayed. In addition, antibodies against particular transcription factors may not be available or validated for ChIP. The drawback of open chromatin assays is their lack of specificity, as the identified sites typically correlate with the binding sites of many different transcription factors [32]. Therefore, additional annotation data sets are required to verify the presence and type of a regulatory element (e.g. enhancer vs. promoter) and transcription factor-binding site. Here, in silico predictions may guide the identification of the specific transcription factor involved.

Alternatively, ChIP-seq of components of the gene expression machinery may also be useful for the annotation

Bioessays 36: 191–199, © 2013 The Authors. Bioessays published by WILEY Periodicals, Inc.

of candidate regulatory variants. For example, ChIP-seq of the transcriptional co-activator protein p300 has been shown to associate with active, cell type-specific enhancer sequences [33]. ChIP-seq of histone modifications that mark active promoters and enhancers are informative, but their typical broader peaks may impede precise screening and identification of candidate functional variants.

Picking strategies for the prioritization of candidate functional variants

To gain biological understanding of the genetic variants associated with complex traits including disease susceptibility and outcome, at least two strategies have been pursued. First, the combined analysis and interpretation of all genetic signals identified in a GWAS (Fig. 1C) and second, the in-depth analysis of selected, individual GWAS loci (Fig. 1D). The two strategies are described in detail in the following sections.

Global analyses of GWAS loci

Much biological insight from GWAS can be gained when multiple association signals are collectively co-analyzed. Rather than aiming to identify distinct molecular mechanisms, these bioinformatic approaches focus on connecting a selection of target genes and their products with knowledge databases, e.g. concerning tissue-specific gene expression signature, intracellular localization, citation in the literature (PubMed abstracts) or gene ontology terms [34–36]. Potential target genes in proximity to all index SNPs ($p < 5 \times 10^{-8}$) or a subset of SNPs reaching a certain significance threshold (e.g. $p < 0.05$) may be tested for association with shared gene pathways and networks. Of course, identified associations are not necessarily causal and require further validation. Findings may also be biased towards well-studied and frequently reported gene pathways. Nonetheless, these bioinformatic tools may provide a powerful means of generating novel hypotheses regarding molecular processes involved in disease etiology. For example, an integrated pathway analysis approach recently highlighted the striking role for host responses to mycobacteria in inflammatory bowel disease [37].

Annotating GWAS signals with information on gene expression can yield a better understanding of the regulatory networks underlying the association. In studies on expression quantitative trait loci (eQTLs), the alleles of the index or proxy SNP are correlated with variation in transcript levels that are quantitatively measured in unrelated individuals using gene expression arrays or – more sensitively and accurately – RNA-seq. SNPs that show a strong correlation with expression levels for a specific gene are likely to mark an eQTL for that gene. Both local- and distal-acting eQTLs can be identified, but the identification of distal-acting eQTLs has been largely unfruitful due to the inherent limited statistical power of the approach [38]. Systematic studies have demonstrated the practicality and efficiency of eQTLs as screening tool for candidate regulatory variants [39, 40]. As a relatively large number of individuals need to be sampled to gain statistical confidence in the observed association, easily accessible cell types (e.g. lymphoblastoid cells or monocytes) or cell types that have been extensively characterized, are usually chosen over the most relevant ones. However, it is important to note that a substantial proportion of eQTLs identified are cell type-restricted [41]. Furthermore, SNPs associated with transcript levels may not be causal, as eQTL studies suffer from LD structure. Although most eQTL studies have focused on protein-coding RNAs, non-coding RNAs (i.e. large intergenic non-coding RNAs; lincRNAs) are equally relevant candidates [42].

Instead of measuring gene expression levels across individuals, allele-specific expression (ASE) analysis measures transcript abundance within an individual [43]. In this powerful approach, transcript levels are assessed using RNA samples derived from individuals that are heterozygous at a particular eQTL SNP of interest. Transcripts that deviate from the expected 1:1 ratio at heterozygous alleles (i.e. show "allelic imbalance") are likely candidate transcripts. For example, ASE analysis has been applied to the asthma and autoimmune disease risk locus on chromosome 17q12-q21 pinpointing potential causal sequence variants [44].

Experimental investigation of individual GWAS loci

Ascertaining an exhaustive account of common as well as low-frequency variants is an important, initial step in the in-depth molecular analysis of selected, individual GWAS intervals. However, only a fraction of all genetic variants are examined in GWAS. Genotype imputation exploits known LD patterns and haplotype frequencies from reference data sets to estimate genotypes for additional SNPs not directly assayed in the initial genome-wide scan [45]. In addition to genotype imputation, established association intervals may be fine-mapped using dense, custom genotyping arrays. Such arrays are based on the deep sequencing catalogue of the 1,000 Genomes Project [10] and contain essentially all common and low-frequency variants at selected GWAS loci of a group of related clinical conditions such as autoimmune and inflammatory diseases [46, 47]. Selected GWAS intervals may be resequenced to enable sequencing-based genotyping. However, the costs for such an experiment may be considerable (e.g. due to the relatively deep coverage and large number of subjects required). We therefore argue for the application of sequencing data from the 1,000 Genomes and UK10K Projects (http://www.uk10k.org/), which should give sufficient account of low-frequency and rare sequence variation.

After the comprehensive assessment of genetic variation at GWAS intervals, the variants are then overlapped with genome annotation marks in order to identify candidate functional variants. Importantly, mere annotation of genetic variants using epigenomic data sets does not prove molecular function and causality (that is, impact on organismal phenotype). Candidate regulatory variants that overlap with one or more annotation marks require substantial experimental validation. This should involve an integrated approach of multiple experimental methods to gain confidence in the observed effect. Most frequently applied in vitro cellular

assays include luciferase reporter assays [15, 48–51], gel-shift and super-shift assays [49, 50, 52], as well as allele-specific chromatin assays [44, 53], which should be performed in relevant cell types to avoid misleading biological interpretations (see above). Candidate regulatory sites may also be tested using in vivo assays. For example, the activity of tissue-specific enhancer sequences can be assessed in transgenic mouse assays [54].

Regulatory variants potentially lie great distances from the gene(s) they control, functioning through long-range regulatory interactions [11, 55, 56]. Chromosome conformation capture (3C) and descended methods (e.g. circular 3C, 4C; enhanced 4C, e4C; and Hi-C), as well as chromatin interaction paired-end tagging (ChIA-PET) techniques, examine long-range physical interactions between distal gene regulatory elements and promoter regions of target genes and have already been successfully applied to non-coding GWAS signals [11, 55, 56]. These experimental tools, while technically challenging to perform, provide an unprecedented view of the interplay between regulatory variants and genes at GWAS intervals.

Once established, target genes may be further characterized with respect to the trait of interest. Here, traditional assays to characterize gene function can be used including gene knockdown using small interfering RNA (siRNA) or gene overexpression using adeno-associated viral vectors [50]. For a number of complex traits, e.g. hematological traits such as platelet counts and volume, gene knockdown in zebrafish (*Danio rerio*) embryos using morpholinos (antisense oligonucleotides) has proved particularly insightful. In systematically switching-off candidate genes at GWAS intervals, several novel genes implicated in platelet formation were identified and successfully validated [57].

Investigating the molecular mechanism of individual GWAS loci is arduous, and arguably, scalable in vitro approaches are needed to experimentally validate candidate functional variants in a high-throughput manner. Indeed, progress has been made in massively parallel reporter assays, which use large-scale DNA synthesis and next-generation sequencing to simultaneously measure the reporter activity of many thousands of enhancer variants [58, 59].

Revolutionizing the functional translation of GWAS signals by genome engineering

Despite remarkable technological advances for the discovery of genetic variation, the experimental tools to study the molecular mechanisms of candidate functional variants have seen only little progress thus far. To this end, we envisage the application of site-specific genome-editing techniques to be game-changing. Transcription activator-like effector nucleases (TALENs) [60–63] and clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems [64, 65] are novel classes of genome-editing techniques. These methods enable the modification of any genomic sequence of interest in mammalian cells and model organisms.

TALENs comprise a FokI nuclease domain, which cleaves DNA in a non-sequence-specific manner, fused to a modular DNA-binding domain. The DNA-binding domain is composed of highly conserved amino acid repeats, transcription activator-like effectors (TALEs), which can be engineered to recognize specific DNA sequences. The engineered nucleases bind as a dimer to a target site, where they induce a DNA double-strand break. In turn, DNA damage response pathways are triggered, such as non-homologous end-joining (NHEJ) or homology-directed repair (HDR), which enable the precise introduction, exclusion or alteration of gene alleles at the target site. Bauer et al. applied TALENs to modulate the activity of an enhancer sequence critical for erythroid expression of *BCL11A*, a gene implicated in hemoglobin disorders [66, 67]. The lineage-specific enhancer contains common sequence variants identified through GWAS, which impact erythroid transcription factor binding. The authors suggest this GWAS-identified enhancer as potential therapeutic target in hemoglobinopathies.

CRISPR/Cas systems have recently emerged as an alternative to TALENs, vastly improving its cleavage efficiency and ease of implementation at reduced cost [65, 68]. Type II CRISPR/Cas systems use Cas9 nucleases that are guided to a genomic sequence of interest via synthetic RNA molecules. Thus, CRISPR's application of guide RNAs supersedes the need for engineering custom proteins. In addition to the disruption of genomic sequence through nucleases, CRISPR/Cas9 may be provided with effector domains that exert distinct regulatory functions. For example, the CRISPR-associated catalytically inactive Cas9 protein, termed dCas9, can be fused to activator domains [69, 70], repressor domains [69] or potentially domains that alter different epigenetic states [71]. Such modified CRISPR/dCas9-fusion proteins, together with guide RNA, can then be introduced to control the activity of candidate regulatory elements that harbor GWAS signals or candidate genes. In Fig. 2, we have summarized different types of CRISPR/Cas9 constructs and their potential application to examine regulatory elements at GWAS intervals.

For genome-editing techniques, the issue of specificity and delivery is paramount. The desired genome modification should ideally occur with high frequency in the cell population and with no detectable off-target effects. Although rapid progress has been made in this respect [72], further improvements in molecular design and experimental procedures are needed to please sceptical reviewers. Here, we suggest light-inducible transcriptional effectors (LITEs) to offer exciting possibilities for studying the function and regulation of mammalian genomes in the near future. LITE modules consist of the light-sensitive photoreceptor cryptochrome 2 (CRY2) that is fused to a customizable DNA-binding domain based on the TALE or CRISPR/Cas9 systems. The construct allows spatially and temporally precise, graded, reversible and non-invasive modulation of gene transcription and epigenetic states [73, 74].

CRISPR/Cas9 constructs can be introduced into human somatic cells and iPSCs (via transfection) as well as various model organisms (e.g. via injection of mouse zygotes [75]). For application in human somatic cells, we suggest the use of cells of defined genotype. Resources of healthy individuals, who can be recalled on the basis of their genotype for donating cellular components (e.g. immune effector cells) and subsequent functional studies, have proven valuable [51, 76].

However, we advocate the use of CRISPR/Cas9 in conjunction with iPSC technology. iPSCs are generated by reprogramming mature adult cells, such as fibroblasts, into cells that hold properties of embryonic stem cells. From this state of immaturity, iPSCs have the potential to differentiate in vitro into a wide range of specialized cell types. For example, detailed protocols have been published describing the differentiation of iPSCs into human cerebral cortex neurons [77, 78] – cells that cannot be obtained from primary sources. By differentiating iPSCs into distinct lineages, the effect of CRISPR/Cas9 can be assessed under different genetic programmes in the same cellular system. For example, an annotated regulatory element containing a GWAS signal associated with Alzheimer's disease may be disrupted using CRISPR/Cas9 in iPSCs. Following differentiation of the iPSCs into cerebral cortex neurons, the functional consequences of that genome alteration can be compared with the control iPSC population, e.g. through transcriptome analysis. Furthermore, multiple implicated GWAS signals, regulatory elements and genes can potentially be disrupted in combination to examine their interaction or synergistic effects. Lastly, this experimental design may be expanded to iPSCs derived from patients, enabling the possibility to "cure" disruptive effects of GWAS signals at the cellular level. The latter may be boosted by, for example, the newly established UK Human Induced Pluripotent Stem Cells Initiative (HipSci; http://www.hipsci.org/). This open-access resource will create iPSC lines derived from over 1,000 healthy individuals and individuals with genetic diseases by 2016. Moreover, the proposed approach would make currently available tools for the correction of disease-causing mutations more efficient, such as zinc finger nucleases (ZFNs) combined with piggyBac transposons in iPSCs [79].

Striving towards the functional interpretation of rare non-coding sequence variation

Low-frequency (MAF, 1–5%) and rare (MAF $< 1\%$) genetic variants (not captured by GWAS) may explain a substantial fraction of the genetic component of complex traits including common diseases [80, 81]. Costs of next-generation sequencing applications have plummeted over the last years and with innovative sequencing methods on the horizon, notably nanopore DNA sequencing, genotyping-based methods will likely be replaced by sequencing-based methods for detection of trait-associated variants. Arguably, this will entail not only challenges for the design of association analyses of low-frequency and rare alleles linked to complex traits and diseases (e.g. extreme trait resequencing or family-based studies [80]), but also for the functional interpretation of the implicated alleles. Compounded by both allelic and locus heterogeneity [82], the sheer number of low-frequency, rare and private variants will make identification of the causal variants challenging. To increase statistical power, studies thus far have therefore centred exclusively on variants of apparent functional consequence, i.e. missense, nonsense, frame-shift or splicing variants. However, it can be expected that some of the causal sites will reflect gene regulatory variants [83]. Akin to the functional classification of coding variants, non-coding variants may be recognized as, for example, disrupting transcription factor binding sites. However, such sites are usually around 200 bp in size when identified using ChIP-seq and sometimes only a fraction of these sites harbor the known binding motif of the transcription factor. In contrast, nucleotide-resolution techniques, such as DNase I and ATAC-seq footprinting, allow for identification of the exact location of the transcription factor binding site [29, 84]. Therefore, we argue that these assays in particular are ideally suited to systematically prioritize low-frequency and rare alleles, due to the reduction of sequence space with likely biological importance. However, we note that in contrast to coding regions where rare variants that are likely to be deleterious can be combined for statistical testing, it is currently unclear how to categorize and test a large number of rare variants across regulatory regions with uncertain functional consequences.

Conclusions and outlook

We believe there is still substantial scope for performing GWAS in the coming years. Well-powered meta-analyses of GWAS detect novel small-effect association regions, and fine-mapping approaches as well as studies in ethnic subgroups refine existing ones. In parallel, we call for a boost in the number of investigations into the molecular mechanisms of confirmed associations. There is a pressing need to translate genetic signals of complex traits and diseases into molecular mechanisms, through both global meta-analysis of multiple GWAS intervals and in-depth mechanistic studies of transcription, chromatin structure and DNA methylation at individual GWAS intervals. This functional translation is crucial for the identification of novel "druggable" or reversible components and pathogenic pathways. This in turn has the potential to empower clinical care through, for example, improved risk prediction, biomarker identification, disease subclassification, drug development and dosing [85].

Despite the identification of over 2,000 robust associations with more than 300 complex traits and diseases [85], only a negligible fraction of discovered GWAS intervals have been followed up in experimental studies. In most cases, the causal variant(s) as well as gene(s) are unknown. With the advent of clinical (http://www.genomicsengland.co.uk/) and personal (http://www.personalgenomes.org/) whole-genome sequencing, an overwhelming number of sequence variants will be identified at non-coding regions with potential regulatory effects. Thus, efficient strategies for functional translation are urgently needed. To this end, we expect that CRISPR/Cas9 will play a pivotal role in unravelling molecular mechanisms for a significant number of trait-associated genetic variants. Although not discussed in detail in this article, we suggest that proteomic tools could also be integrated in the functional translation of GWAS findings. For example, the interaction dynamics between a trait-associated variant and a transcription factor complex may be characterized by mass spectrometry [86]. It is important to note that the subtle phenotypic effects of trait-associated variants discovered through GWAS

should not be inferred as indicative of subtle effects of the regulatory element at which they reside. In contrast, they may potentially have a strong phenotypic effect as part of a gene regulatory complex [66].

Non-genetic factors, e.g. epigenetic variation, have been suggested to have a substantial impact on complex trait etiology. Similar to GWAS, such epigenetic variation (specifically, DNA methylation) can be assayed across many individuals and tested for association with a complex trait of interest in epigenome-wide association studies (EWAS) [87]. EWAS may be used to explore genetic risk alleles that mediate their effects through epigenetic mechanisms [88, 89]. Thus, integration of GWAS and EWAS presents a promising strategy to unravel the underlying biological mechanisms of complex traits and diseases [90–92]. Here, the newly formed Genetics of DNA Methylation Consortium (GoDMC; http://www.godmc.org.uk/) will bring together scientists studying the genetic basis of DNA methylation and provide a centralized hub for coordinating data analyses.

Acknowledgements

We would like to thank Augusto Rendon (Department of Haematology, University of Cambridge, and NHS Blood and Transplant, Cambridge, UK) for fruitful discussions during the manuscript preparation and Laura Phipps for proofreading the manuscript. This work is supported by the EU-FP7 Project BLUEPRINT (282510) and the Wellcome Trust (098051 and 99148).

References

1. **Mardis ER**. 2011. A decade's perspective on DNA sequencing technology. *Nature* **470**: 198–203.
2. **McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB**, et al. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**: 356–69.
3. **Donnelly P**. 2008. Progress and challenges in genome-wide association studies in humans. *Nature* **456**: 728–31.
4. **Visscher PM, Brown MA, McCarthy MI, Yang J**. 2012. Five years of GWAS discovery. *Am J Hum Genet* **90**: 7–24.
5. **Manolio TA, Collins FS, Cox NJ, Goldstein DB**, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747–53.
6. **Eichler EE, Flint J, Gibson G, Kong A**, et al. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**: 446–50.
7. **Weir BS**. 2008. Linkage disequilibrium and association mapping. *Annu Rev Genomics Hum Genet* **9**: 129–42.
8. **The International HapMap Consortium**. 2005. A haplotype map of the human genome. *Nature* **437**: 1299–320.
9. **The International HapMap Consortium**. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–61.
10. **The 1000 Genomes Project Consortium**. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.
11. **Maurano MT, Humbert R, Rynes E, Thurman RE**, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**: 1190–5.
12. **The ENCODE Project Consortium**. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.
13. **Trynka G, Sandor C, Han B, Xu H**, et al. 2013. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* **45**: 124–30.
14. **Maston GA, Evans SK, Green MR**. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**: 29–59.
15. **Gaulton KJ, Nammo T, Pasquali L, Simon JM**, et al. 2010. A map of open chromatin in human pancreatic islets. *Nat Genet* **42**: 255–9.
16. **Symonds ME, Sebert SP, Hyatt MA, Budge H**. 2009. Nutritional programming of the metabolic syndrome. *Nat Rev Endocrinol* **5**: 604–10.
17. **Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B**, et al. 2010. The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol* **28**: 1045–8.
18. **Adams D, Altucci L, Antonarakis SE, Ballesteros J**, et al. 2012. BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol* **30**: 224–6.
19. **Zhou X, Maricque B, Xie M, Li D**, et al. 2011. The human epigenome browser at Washington University. *Nat Methods* **8**: 989–90.
20. **Karnik R, Meissner A**. 2013. Browsing (epi)genomes: a guide to data resources and epigenome browsers for stem cell researchers. *Cell Stem Cell* **13**: 14–21.
21. **Ernst J, Kellis M**. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**: 817–25.
22. **Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N**, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43–9.
23. **Hoffman MM, Ernst J, Wilder SP, Kundaje A**, et al. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucl Acids Res* **41**: 827–41.
24. **McKernan R, Watt FM**. 2013. What is the point of large-scale collections of human induced pluripotent stem cells? *Nat Biotechnol* **31**: 875–7.
25. **Masters JRW**. 2000. Human cancer cell lines: fact and fantasy. *Nat Rev Mol Cell Biol* **1**: 233–6.
26. **Paul DS, Albers CA, Rendon A, Voss K**, et al. 2013. Maps of open chromatin highlight cell type-restricted patterns of regulatory sequence variation at hematological trait loci. *Genome Res* **23**: 1130–41.
27. **Boyle AP, Hong EL, Hariharan M, Cheng Y**, et al. 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**: 1790–7.
28. **Nürnberg ST, Rendon A, Smethurst PA, Paul DS**, et al. 2012. A GWAS sequence variant for platelet volume marks an alternative DNM3 promoter in megakaryocytes near a MEIS1 binding site. *Blood* **120**: 4859–68.
29. **Buenrostro JD, Giresi PG, Zaba LC, Chang HY**, et al. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*, in press, doi: 10.1038/nmeth.2688.
30. **Song L, Zhang Z, Grasfeder LL, Boyle AP**, et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* **21**: 1757–67.
31. **Thurman RE, Rynes E, Humbert R, Vierstra J**, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82.
32. **Ernst J, Kellis M**. 2013. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Res* **23**: 1142–54.
33. **May D, Blow MJ, Kaplan T, McCulley DJ**, et al. 2012. Large-scale discovery of enhancers from human heart tissue. *Nat Genet* **44**: 89–93.
34. **Zhong H, Yang X, Kaplan LM, Molony C**, et al. 2010. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet* **86**: 581–91.
35. **Wang K, Li M, Hakonarson H**. 2010. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* **11**: 843–54.
36. **Croft D, O'Kelly G, Wu G, Haw R**, et al. 2011. Reactome: a database of reactions, pathways and biological processes. *Nucl Acids Res* **39**: D691–7.
37. **Jostins L, Ripke S, Weersma RK, Duerr RH**, et al. 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**: 119–24.
38. **Grundberg E, Small KS, Hedman AK, Nica AC**, et al. 2012. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**: 1084–9.
39. **Nica AC, Montgomery SB, Dimas AS, Stranger BE**, et al. 2010. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* **6**: e1000895.
40. **Nicolae DL, Gamazon E, Zhang W, Duan S**, et al. 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**: e1000888.
41. **Dimas AS, Deutsch S, Stranger BE, Montgomery SB**, et al. 2009. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**: 1246–50.
42. **Kumar V, Westra H-J, Karjalainen J, Zhernakova DV**, et al. 2013. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet* **9**: e1003201.
43. **Pastinen T**. 2010. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* **11**: 533–8.

44. **Verlaan DJ, Berlivet S, Hunninghake GM, Madore A-M**, et al. 2009. Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. *Am J Hum Genet* **85**: 377–93.
45. **Li Y, Willer C, Sanna S, Abecasis G**. 2009. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**: 387–406.
46. **Trynka G, Hunt KA, Bockett NA, Romanos J**, et al. 2011. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* **43**: 1193–201.
47. **Liu JZ, Almarri MA, Gaffney DJ, Mells GF**, et al. 2012. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat Genet* **44**: 1137–41.
48. **Pomerantz MM, Ahmadiyeh N, Jia L, Herman P**, et al. 2009. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41**: 882–4.
49. **Tuupanen S, Turunen M, Lehtonen R, Hallikas O**, et al. 2009. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* **41**: 885–90.
50. **Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE**, et al. 2010. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**: 714–9.
51. **Lee JC, Espeli M, Anderson CA, Linterman MA**, et al. 2013. Human SNP links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway. *Cell* **155**: 57–69.
52. **Paul DS, Nisbet JP, Yang T-P, Meacham S**, et al. 2011. Maps of open chromatin guide the functional follow-up of genome-wide association signals: application to hematological traits. *PLoS Genet* **7**: e1002139.
53. **Smith AJP, Howard P, Shah S, Eriksson P**, et al. 2012. Use of allele-specific FAIRE to determine functional regulatory polymorphism using large-scale genotyping arrays. *PLoS Genet* **8**: e1002908.
54. **Visel A, Blow MJ, Li Z, Zhang T**, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**: 854–8.
55. **Davison LJ, Wallace C, Cooper JD, Cope NF**, et al. 2012. Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. *Hum Mol Genet* **21**: 322–33.
56. **Harismendy O, Notani D, Song X, Rahim NG**, et al. 2011. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* **470**: 264–8.
57. **Gieger C, Radhakrishnan A, Cvejic A, Tang W**, et al. 2011. New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**: 201–8.
58. **Melnikov A, Murugan A, Zhang X, Tesileanu T**, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**: 271–7.
59. **Kheradpour P, Ernst J, Melnikov A, Rogov P**, et al. 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23**: 800–11.
60. **Moscou MJ, Bogdanove AJ**. 2009. A simple cipher governs DNA recognition by TAL effectors. *Science* **326**: 1501.
61. **Boch J, Scholze H, Schornack S, Landgraf A**, et al. 2009. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**: 1509–12.
62. **Miller JC, Tan S, Qiao G, Barlow KA**, et al. 2011. A TALE nuclease architecture for efficient genome editing. *Nat Biotechnol* **29**: 143–8.
63. **Hockemeyer D, Wang H, Kiani S, Lai CS**, et al. 2011. Genetic engineering of human pluripotent cells using TALE nucleases. *Nat Biotechnol* **29**: 731–4.
64. **Cong L, Ran FA, Cox D, Lin S**, et al. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**: 819–23.
65. **Mali P, Yang L, Esvelt KM, Aach J**, et al. 2013. RNA-guided human genome engineering via Cas9. *Science* **339**: 823–6.
66. **Hardison RC, Blobel GA**. 2013. GWAS to therapy by genome edits? *Science* **342**: 206–7.
67. **Bauer DE, Kamran SC, Lessard S, Xu J**, et al. 2013. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**: 253–7.
68. **Ding Q, Regan SN, Xia Y, Oostrom LA**, et al. 2013. Enhanced efficiency of human pluripotent stem cell genome editing through replacing TALENs with CRISPRs. *Cell Stem Cell* **12**: 393–4.
69. **Gilbert LA, Larson MH, Morsut L, Liu Z**, et al. 2013. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**: 442–51.
70. **Maeder ML, Linder SJ, Cascio VM, Fu Y**, et al. 2013. CRISPR RNA-guided activation of endogenous human genes. *Nat Methods* **10**: 977–9.
71. **Mali P, Esvelt KM, Church GM**. 2013. Cas9 as a versatile tool for engineering biology. *Nat Methods* **10**: 957–63.
72. **Carroll D**. 2013. Staying on target with CRISPR-Cas. *Nat Biotechnol* **31**: 807–9.
73. **Möglich A, Hegemann P**. 2013. Biotechnology: programming genomes with light. *Nature* **500**: 406–8.
74. **Konermann S, Brigham MD, Trevino A, Hsu PD**, et al. 2013. Optical control of mammalian endogenous transcription and epigenetic states. *Nature* **500**: 472–6.
75. **Wang H, Yang H, Shivalila CS, Dawlaty MM**, et al. 2013. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**: 910–8.
76. **Dendrou CA, Plagnol V, Fung E, Yang JHM**, et al. 2009. Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nat Genet* **41**: 1011–5.
77. **Shi Y, Kirwan P, Smith J, Robinson HPC**, et al. 2012. Human cerebral cortex development from pluripotent stem cells to functional excitatory synapses. *Nat Neurosci* **15**: 477–86.
78. **Shi Y, Kirwan P, Livesey FJ**. 2012. Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nat Protoc* **7**: 1836–46.
79. **Yusa K, Rashid ST, Strick-Marchand H, Varela I**, et al. 2011. Targeted gene correction of α1-antitrypsin deficiency in induced pluripotent stem cells. *Nature* **478**: 391–4.
80. **Cirulli ET, Goldstein DB**. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**: 415–25.
81. **Singleton AB, Hardy J, Traynor BJ, Houlden H**. 2010. Towards a complete resolution of the genetic architecture of disease. *Trends Genet* **26**: 438–42.
82. **Ward LD, Kellis M**. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**: 1095–106.
83. **Albers CA, Paul DS, Schulze H, Freson K**, et al. 2012. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nat Genet* **44**: 435–9.
84. **Neph S, Vierstra J, Stergachis AB, Reynolds AP**, et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**: 83–90.
85. **Manolio TA**. 2013. Bringing genome-wide association findings into clinical use. *Nat Rev Genet* **14**: 549–58.
86. **Butter F, Davison L, Viturawong T, Scheibe M**, et al. 2012. Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS Genet* **8**: e1002982.
87. **Rakyan VK, Down TA, Balding DJ, Beck S**. 2011. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* **12**: 529–41.
88. **Rakyan VK, Beyan H, Down TA, Hawa MI**, et al. 2011. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genet* **7**: e1002300.
89. **Liu Y, Aryee MJ, Padyukov L, Fallin MD**, et al. 2013. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol* **31**: 142–7.
90. **Bell CG, Finer S, Lindgren CM, Wilson GA**, et al. 2010. Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. *PLoS One* **5**: e14040.
91. **Birney E**. 2011. Chromatin and heritability: how epigenetic studies can complement genetic approaches. *Trends Genet* **27**: 172–6.
92. **Ke X, Cortina-Borja M, Silva BC, Lowe R**, et al. 2013. Integrated analysis of genome-wide genetic and epigenetic association data for identification of disease mechanisms. *Epigenetics* **8**: 1236–44.