Interpreting noncoding genetic variation in complex traits and human disease

Lucas D Ward[1,2] & Manolis Kellis[1,2]

Association studies provide genome-wide information about the genetic basis of complex disease, but medical research has focused primarily on protein-coding variants, owing to the difficulty of interpreting noncoding mutations. This picture has changed with advances in the systematic annotation of functional noncoding elements. Evolutionary conservation, functional genomics, chromatin state, sequence motifs and molecular quantitative trait loci all provide complementary information about the function of noncoding sequences. These functional maps can help with prioritizing variants on risk haplotypes, filtering mutations encountered in the clinic and performing systems-level analyses to reveal processes underlying disease associations. Advances in predictive modeling can enable data-set integration to reveal pathways shared across loci and alleles, and richer regulatory models can guide the search for epistatic interactions. Lastly, new massively parallel reporter experiments can systematically validate regulatory predictions. Ultimately, advances in regulatory and systems genomics can help unleash the value of whole-genome sequencing for personalized genomic risk assessment, diagnosis and treatment.

Understanding the genetic basis of disease can transform medicine by elucidating relevant biochemical pathways for drug targets and by enabling personalized risk assessments[1,2]. With the evolution of technologies over the past century, geneticists are no longer limited to studying Mendelian disorders and can tackle complex phenotypes. The resulting discovered associations have broadened from individual variants primarily in coding regions to much richer disease architectures, including noncoding variants, wider allelic spectra, numerous loci and weak effect sizes (**Table 1**). In the past few years, a new wave of technological advances has intensified the shift toward investigating more complex genetic architectures and uncovering the molecular mechanisms underlying them.

In the early twentieth century, several metabolic disorders were shown to be genetic and Mendelian, and later positional cloning allowed the identification of many such loci, such as those curated by the Online Mendelian Inheritance in Man database (OMIM)[3,4]. Starting in the 1980s, linkage analysis was used to correlate the inheritance of traits in families with the inheritance of mapped polymorphic markers that could be assayed through restriction-fragment-length polymorphism (RFLP) analysis[5,6]. However, the regions mapped by linkage analysis were necessarily large, and before the completion of the Human Genome Project, cloning candidate genes for follow-up association studies, resequencing and functional assays required the application of painstaking molecular techniques[7]. In addition, complex phenotypes were not amenable to linkage because of the large sample sizes needed to detect loci with modest effects above the genomic background[8]. The long haplotype structure of the human genome, and its systematic mapping by the HapMap Project[9], allowed single-nucleotide polymorphisms (SNPs) to be used as markers for common haplotypes, which could be genotyped using chip technology. The stage was set for a flood of unbiased, genome-wide association studies (GWAS) to search across unrelated individuals[10] for common variants associated with complex disease and diverse molecular phenotypes (**Fig. 1**; **Table 2**).

Relative to linkage analysis and sequencing, GWAS have less power in cases where different rare mutations act in different families or individuals at the same locus (allelic heterogeneity). However, they are far more sensitive than family studies to complex polygenic associations in which a phenotype is associated with the joint effect of many weakly contributing variants across different loci (locus heterogeneity). In this sense, GWAS have been a resounding success, identifying thousands of disease-associated loci for further study[11] and revealing previously unknown mechanisms for conditions such as Crohn's disease, macular degeneration and type 2 diabetes[2]. However, the pursuit of GWAS has also received criticism (**Box 1**) because of the structure of the knowledge it has been producing relative to the determinism of highly penetrant Mendelian genetic discoveries[2,12,13]. The current tension mirrors the intellectual rift in the early 1900s between Mendelians, who modeled inheritance of discrete traits as being carried by single genes, and the biometrician adherents of Galton, who studied the inheritance of continuous traits; the fields were reconciled by R.A. Fisher, who proposed that quantitative traits' heritability was owed to the contribution of many genes with small effect[14,15].

In this review, we discuss both the computational challenges and the opportunities presented by the large number of noncoding disease-associated variants being discovered through GWAS and medical resequencing. We first survey the types of regulatory annotations available, including those from functional and comparative genomics as well as quantitative trait loci (QTLs) and allele-specific events, and the ways in which these can be used to dissect disease-associated haplotypes to

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. Correspondence should be addressed to L.D.W. (lukeward@mit.edu) or M.K. (manoli@mit.edu).

Received 14 September; accepted 16 October; published online 8 November 2012; doi:10.1038/nbt.2422

Table 1 The diversity of genetic architectures underlying human phenotypes

Architecture	Notes and examples	Role of computational and regulatory genomics
Classic monogenic traits	The earliest human genes characterized were those leading to inborn errors in metabolism, which were shown by Garrod in the early 1900s to follow Mendelian inheritance[130,131]. The modern study of human disease genes began with the cloning of loci responsible for high-penetrance monogenic disorders with Mendelian inheritance patterns, such as phenylketonuria and cystic fibrosis[130,132,133], that were most amenable to classical mapping approaches. Variants associated with monogenic traits were also the first to be identified through positional cloning in the 1980s, a classic success being the *CFTR* mutations responsible for most cases of cystic fibrosis[3,132,133].	As the underlying mutations tend to alter protein structure, the computational challenge in predicting their effect lies in molecular modeling and structural studies.
Monogenic traits with multiple disease alleles	Even monogenetic diseases differ greatly in the extent to which a single risk allele predominates among affected individuals (allelic heterogeneity). On one end of the spectrum, the F508del allele of *CFTR* is found in about 70% of patients with cystic fibrosis[134], even though thousands of alleles are known. In contrast, phenylketonuria is extremely heterogeneous, with different *PAH* alleles predominating among affected individuals in different populations[135]. A majority of mutations in this class are missense or nonsense coding mutations[3].	As noted above, for protein-coding mutations, the relevant problem is predicting the biochemical effect of the amino acid substitution. In cases of allele heterogeneity, the observed substitutions may be too numerous to characterize experimentally, necessitating computational models (**Fig. 3c**).
Multiple loci with independent contributions ('oligogenic')	Many variants increase or decrease the risk of a disease, with the final phenotype relying on the genotype at many loci (locus heterogeneity). One example well studied through linkage analysis is Hirschprung disease, a complex disorder with low sex-dependent penetrance for which at least ten underlying genes are involved, including the tyrosine kinase receptor *RET* and the gene *GDNF* which encodes its ligand[136]. Interestingly, the most common variant in the main susceptibility gene *RET* is noncoding, a SNP in an enhancer. Both coding and noncoding variants are involved typically in one or a few well-defined pathways.	Oligogenic traits, in which a handful of well-characterized loci contribute to the phenotype, may present the best opportunity to observe and quantify epistatic interactions. In cases where noncoding regions are implicated, these haplotypes can be functionally mapped to isolate the most likely causal variants (**Fig. 2**).
Large numbers of variants jointly contributing weakly to a complex trait	GWAS of complex traits are also discovering many weakly contributing loci. For example, a recent meta-analysis of several height studies found 180 loci reaching genome-wide significance[15,103,137], enriched near genes already known to underlie skeletal growth defects. In the height study and in a study of psychiatric disorders, it has been shown that polygenic association extends to thousands of common variants, extending far beyond genome-wide-significant loci[137,138].	In contrast to the variants underlying monogenic traits, the variants involved in complex traits are overwhelmingly not associated with missense or nonsense coding mutations, suggesting that their mechanisms are primarily regulatory[11]. Large sets of regulatory variants can be combined with reference annotations to elucidate relevant pathways and tissues (**Fig. 3b**, **Table 5**).
Variants regulating a 'molecular trait' with unknown effect on organismal phenotype or fitness	Variants are rapidly being discovered that directly affect molecular quantitative traits, such as gene expression or chromatin state, many of which may have no effect on organismal phenotype or fitness[38].	QTLs and allele-specific analyses are needed to characterize these variants (**Fig. 1b**,**c**). As the studies performed to date sample only a small fraction of the cell types in which a variant may have an effect, and variant-expression associations are highly tissue specific[139], it is possible that many such regulatory variants remain to be discovered.
Variants causing no known molecular phenotype and no effect on organismal phenotype or fitness	The idea that the majority of mutations are neutral from an adaptive perspective was controversial when first proposed, and now is widely accepted[140–142].	Although it is straightforward to calculate from the genetic code what fraction of protein-coding mutations will cause an amino acid change, an analogous estimate for other molecular phenotypes is far more challenging and requires comprehensive regulatory models at the nucleotide level.
Private and somatic variants	Somatic mutations within an organism are frequent driver mutations selected in cancer formation[143].	The interpretation of private and somatic variations (**Fig. 3d**) will also benefit tremendously from a systematic regulatory annotation, as they are likely to exploit existing regulatory pathways, even though they are subject to cellular, rather than organismal, selective pressures.

identify the most promising causal variants at a locus. We then discuss the utility of these regulatory annotations for performing systems-level analysis of GWAS, revealing relevant cell types and regulatory mechanisms. Finally, we describe a variety of bioinformatics hurdles and computational challenges that lie ahead for the field, such as discovering epistatic interactions, connections between molecular and organismal phenotypes, and patterns that must be mined from potentially sensitive medical data.

Systematic annotation of the noncoding genome

Interpretation of the molecular mechanisms of disease-associated loci can be a great challenge. Even though protein biochemistry has been used to characterize missense and nonsense coding mutations that most often underlie monogenic traits, the frequency with which loss-of-function mutations and rare coding variants are being discovered in healthy individuals[16,17] suggests our understanding is far from complete. The challenge of interpretation is even greater for noncoding variants, given the diversity of noncoding functions, the incomplete annotation of regulatory elements and the potential existence of still unknown mechanisms of regulatory control. Several pioneering studies have provided a model for the types of systematic regulatory annotations needed, by revealing the diverse mechanisms of action underlying human disease, including those at the transcriptional, mRNA splicing and translational levels (**Table 3**).

In each of these cases, extensive experimental follow-up was needed to uncover the molecular mechanisms responsible for the disease association signal. Many more disease-associated variants remain uncharacterized, emphasizing the need for systematic methods of annotating regulatory regions, their functional nucleotides and their interconnections.

Springing from recognition of the need for systematic interpretation of noncoding disease-associated variants, several large-scale projects are currently underway to enhance the annotation of the noncoding genome (**Fig. 2**). These rely on reference annotation maps using both functional genomics and comparative genomics, and they can dramatically increase

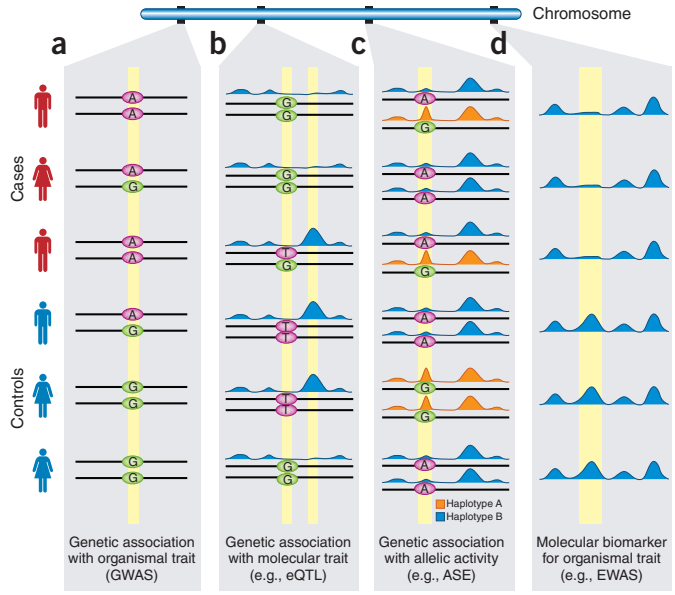

Figure 1 Four types of association tests. (**a**) Genetic association with organismal traits is performed in genome-wide association studies (GWAS); at the locus shown, the A allele is associated with disease. The effect of GWAS-discovered variants is mediated through many layers of molecular processes, some of which can also be interrogated at a genome-wide scale. (**b**) Rather than organismal traits, molecular traits can be used, leading to the discovery of local regulatory variants, such as expression quantitative trait loci (eQTLs). In this example, a local molecular signal, such as a region of open chromatin, varies across the individuals and is shown to covary with presence of the T allele; thus, this allele may influence a *cis*-regulatory motif. (**c**) Heterozygous sites in an individual can be used to interrogate allele-specific effects (ASE); unlike for molecular QTLs discovered across individuals, these studies control for variation in the *trans* genetic background. In this example, the G allele not only is associated with the presence of a transcription factor binding peak at that locus, but also, in heterozygous individuals, is overrepresented in ChIP-seq reads originating from that locus, suggesting that the transcription factor binds specifically to the G allele. (**d**) Functional genomics data can be directly compared between cases and controls to discover biomarkers for disease, such as in epigenome-wide association studies (EWAS), without necessarily attributing genetic causes to these molecular changes. Indeed, these biomarkers may be caused by *trans* genetic factors, by environmental factors or by the disease itself.

the annotation of regulatory elements, which can have a strong impact for interpreting both existing GWAS and individual personal genomes.

Reference functional genomics and chromatin state maps. Massively parallel short-read sequencing technologies have obviated the need for the extremely expensive tiling microarrays previously used to map biochemically active regions of the human genome. This has enabled chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) to map transcription factor binding, chromatin regulators or histone modification marks[18] as well as the mapping of DNA methylation using bisulfite sequencing (BS-seq)[19] and of accessible chromatin regions by DNase hypersensitivity analysis (DNase-seq)[20]. Computational integration of these data sets through supervised or unsupervised machine learning enables mapping of functional noncoding elements, such as distal enhancers, transcription factor binding sites and regulatory RNA genes, on a genome-wide scale. For example, the Encyclopedia of DNA Elements (ENCODE) project has released comprehensive maps of chromatin states, transcription factor binding and transcription for a selection of cell lines and DNase maps for many primary cells[21], and the US National Institutes of Health (NIH; Bethesda, Maryland) Epigenomics Roadmap Project[22] and BluePrint project[23] both aim to construct reference epigenome maps of hundreds of primary cells and cultured cells. Regulatory maps can then guide the way toward the most likely causal regulators on a haplotype (**Fig. 2a**).

Nucleotide-resolution regulatory annotations. Although maps of regulatory regions can be highly informative, increasing their resolution from hundreds of nucleotides to single nucleotides requires additional computational or experimental developments. This can leverage systematic efforts that seek to elucidate the binding specificities of transcription factors[24,25] and splicing regulators[26,27] and also to discover regulatory motifs genome wide based on their enrichment and conservation properties[28,29]. Similarly, new technologies have been applied to enhance existing techniques, such as digital genomic footprinting using DNase-seq[30], dynamic application of micrococcal nuclease (MNase)[31] or the use of lambda exonuclease (ChIP-exo)[32], dramatically increasing the mapping resolution of regulatory elements, even without knowledge of the specific motifs involved.

Predictive models of variant effects. Even when the functional elements and motifs are known, we need models to distinguish how mutations in

Table 2 Computational tools for association analyses[a]

Class of analysis	Tool	Notes
Genome-wide association between genotype and phenotype (GWAS)	SNPTEST[144]	Incorporates imputation
	Bim-Bam[145]	Bayesian regression approach combining imputation and association probabilities
	EIGENSTRAT[146]	Models ancestry differences between cases and controls using principal-components analysis (PCA)
	PLINK[147]	Large package including tools to impute and to control for population stratification, as well as hybrid methods such as family-based association and population-based linkage
Local association between genotype and molecular trait (e.g., eQTL)	eQTNMiner[148]	Tests a Bayesian hierarchical model incorporating priors based on transcription start site (TSS) distance
	Matrix eQTL[149]	Fast association testing of continuous or categorical genotype values with expression
Allele-specific expression and binding	ChIP-SNP[82]	For ChIP-chip data
	AlleleSeq[150]	For ChIP-seq and RNA-seq data
Genome-wide association between molecular trait and phenotype (e.g., differential expression, EWAS)	limma[151]	For expression microarray data
	edgeR[152]	For RNA-seq data

[a]Analyses using genotype information require tools to call variants, such as BirdSeed[153] on array data or GATK[154] on sequencing data, and tools to impute genotypes, such as MaCH[155].

Box 1 Potential and limitations of genome-wide association studies

Although several predominant criticisms of GWAS have been voiced, responses to each can guide future studies. Below, we discuss each critique in turn and discuss ways forward.

Cumulative predictive power. In general, the loci found to reach genome-wide significance have weak additive predictive power for specific phenotypes, which for some traits limits their clinical relevance at present[192–194]. However, risk prediction using the loci discovered for complex disease using GWAS often performs similarly to risk prediction using classic clinical tests and also has unique and useful properties, such as stability over the lifespan[195]. Predictors that jointly use hundreds or thousands of weakly contributing loci have also been shown to explain a larger proportion of variance than was appreciated when the approach was first used[138,196]. Integration of these discoveries into clinical protocols is in its infancy and should be expected to mature.

Noncoding variants with unknown effect. Most of the loci are noncoding, and many are far from discovered genes and, because of linkage disequilibrium, encompass many variants; therefore, they are not immediately informative or biochemically tractable for experimental work. Assigning a prior probability to the deleteriousness of a noncoding mutation is a challenging task[197]. To address this challenge, noncoding sequence is being annotated at a rapid pace through such systematic efforts as the ENCODE Project[21] and the Roadmap Epigenomics Mapping Consortium[22] as well as through studies of the impact of common variants on genomewide molecular phenotypes.

Detection of rare variants. Significant loci tend to additively explain only a small proportion of the narrow-sense heritability of phenotypes[12], suggesting that rare rather than common variants may underlie their genetics, which will only be discovered through whole-exome and whole-genome sequencing or family-based studies[13]. Many explanations for 'hidden heritability' among the discovered common-variant associations have been proposed[12]. The relative importance of rare and common variants is a topic of intense debate[198], with arguments including the contentions that associations with common variants are in fact driven by synthetic associations with large-effect rare variants in long-range linkage disequilibrium[199], that common associations of weak effect contribute to heritability well beyond the threshold of statistical significance[137] and that narrow-sense heritability may be overestimated in many twin studies because of epistasis disguised as additivity[98].

Reproducibility. GWAS sometimes are not replicated across studies or populations[130], leading to the report of false positives and suspicion of the validity of novel associations, especially when they involve noncoding sequence. This could be partly due both to difficulties in imputing genotypes, which will benefit from an increased understanding of common human variation, and to the poor definition of organismal phenotypes[130], which could benefit from molecular disease biomarkers. Moreover, although the specific loci involved may differ across populations, they may reflect the same underlying molecular pathways, and thus regulatory annotations may be more reproducible across populations. Focusing on molecular phenotypes may improve reproducibility by isolating potential socioeconomic or other environmental factors that occur downstream of molecular phenotypes and can strongly affect organismal phenotypes.

different positions of a regulatory motif or element will affect its function. These models can be used to distinguish silent from deleterious mutations, as is possible within protein-coding regions. This requires integrative models of sequence motifs, chromatin state and expression patterns[24,33–36], which can be trained on experimentally tractable tissues or through *in vitro* experiments and applied to predict the effect of newly observed rare and 'private' mutations. The massive scale of regulatory predictions, encompassing hundreds of regulators and millions of regulatory motif instances, demands correspondingly massively parallel methods to validate them. Such methods, which exploit emerging large-scale DNA synthesis and sequencing technologies, are being developed both in model organisms and in cultured human cells[37–39], and they make it possible to test mechanistic hypotheses about causal variants at unprecedented scales (**Fig. 2b**).

Comparative genomics between related species. Even when a regulatory element is rarely used and its activity unobserved in the cell types and tissues sampled, its effect on fitness can still be recognized based on its preferential conservation across multiple related species. Genome-wide comparative analysis of many mammals has revealed a high-resolution map of constrained elements spanning 4.5% of the human genome[40,41], identifying millions of likely new elements, including individual transcription factor binding sites, whose nucleotides have been preserved across evolutionary time. Beyond the overall level of evolutionary constraint, the specific evolutionary signatures encoded in the patterns of substitutions, insertions and deletions across related species can provide information for the type of molecular function likely to be encoded by the constrained elements[41–44]. Together, constraint and evolutionary signatures can pinpoint functional transcription factor binding motifs and individual binding sites (**Fig. 2c**), noncoding RNA genes and structures, microRNAs and their targets, and yet-uncharacterized sequence elements that confer selective advantage.

Evolutionarily conserved biochemical activity. Even in the absence of conserved sequence, the conservation of biochemical activity can be indicative of conserved functional elements, even when the corresponding sequence features are not detectable by traditional alignment and constraint measures owing to turnover[45,46]. Because some fraction of protein binding and RNA transcription may be nonfunctional 'noise,' cross-species analysis of transcription factor binding[47] or gene expression[48] can help reveal the subset of elements that are most likely to be functional. However, lineage-specific elements may nevertheless be important and may not be captured through this method.

Interpreting variants using functional genomic annotations

For protein-coding mutations, knowledge of protein structure and function, and the unambiguous nature of the genetic code, have allowed the development of a class of predictive algorithms that can score the severity of missense and nonsense variants[49–52]. Reference annotations are needed to bring functional data sets to bear on understanding the molecular roles of disease-associated common variants in individual regions, especially for noncoding variants (**Fig. 2**). In addition, new methods are needed to define the relationship between global genetic architectures and genome-wide functional landscapes.

Tools for prioritizing variants. An immediate concern for practitioners of GWAS is the interpretation and prioritization of noncoding variants[53]. Several resources, including HaploReg[54] (L.D.W. and M.K.),

Table 3 Mechanisms through which noncoding variants influence human disease

Noncoding element disrupted	Molecular function and effect of mutations	Disease association
Splice junction and splicing enhancer	Splicing of mRNA is constitutive for some transcripts and highly tissue specific for others, relying on both canonical sequences at the exon-intron junction and weakly specified sequence motifs distributed throughout the transcript. Mutations affecting constitutive splice sites can have an effect similar to nonsense or missense mutations, resulting in aberrantly included introns or skipped exons, sometimes resulting in nonsense-mediated decay (NMD).	Splicing regulatory variants are implicated in several diseases[156,157].
		A recent analysis suggests that the majority of disease-causing point mutations in OMIM may exert their effects by altering splicing[158].
		Alternative splice site variants in the *WT1* gene are involved in Frasier syndrome (FS)[159].
		Skipping of exon 7 of the *SMN* gene is involved in spinal muscular atrophy (SMA)[160].
Sequences regulating translation, stability, and localization	Sequences in the 5′ untranslated regions (UTRs) of mRNAs can influence translation regulation, such as upstream open reading frames (ORFs), premature AUG or AUC codons, and palindromic sequences that form inhibitory stem loops[161]. Sequence motifs in the 3′ UTR are recognized by microRNAs and RNA-binding proteins (RBPs).	Loss-of-function mutations in the 5′ UTR of *CDKN2A* predispose individuals to melanoma[162].
		A rare mutation that creates a binding site for the miRNA hs-miR-189 in the transcript of the gene *SLITRK1* is associated with Tourette's syndrome[163].
Genes encoding *trans*-regulatory RNA	Noncoding RNAs participate in a panoply of regulatory functions; these RNAs range from the well-understood transfer and ribosomal RNA to the recently discovered long noncoding RNAs[164,165].	Both rare and common mutations in the gene *RMRP*, encoding an RNA component of the mitochondrial RNA processing RNase, have been associated with cartilage-hair hypoplasia[166].
		Noncoding RNA mutations can cause many other diseases[167].
Promoter	Promoter regions are an essential component of transcription initiation and the assembly of RNA polymerase and associated regulators. Mutations can affect binding of activators or repressors, chromatin state, nucleosome positioning, and also looping contacts of promoters with distal regulatory elements. Genes with coding disease mutations can also harbor independently associated regulatory variants that correlate with expression, are bound by proteins in an allele-specific manner, and disrupt or create regulatory motifs[168].	Mutations in the promoter of the HIV-1 progression–associated gene *CCR5* are correlated with expression of the receptor it encodes and bind differentially to at least three transcription factors[169,170].
		APOE promoter mutations are associated with Alzheimer's disease[171,172].
		Heme oxygenase-1 (*HO1*) promoter mutations lead to expression changes and are associated with many diseases[173].
Enhancer	Enhancers are distal regulatory elements that often lie 10,000–100,000 nucleotides from the start of their target gene. Mutations within them can disrupt sequence motifs for sequence-specific transcription factors, chromatin regulators and nucleosome positioning signals. Structural variants including inversions and translocations can disrupt their regulatory activity by moving them away from their targets, disrupting local chromatin conformation, or creating interactions with insulators or repressors that can hinder their action. Although it is thought that looping interactions with promoter regions play a role, the rules of enhancer-gene targeting are still poorly understood.	The role of distal enhancers in disease was suggested even before the development of GWAS by the many Mendelian disorders for which some patients had translocations or other structural variants far from the promoter[174–176].
		In one early study, point mutations were mapped in an unlinked locus in the intron of a neighboring gene, a million nucleotides away from the developmental gene *Shh*[177]; this distal locus acted as an enhancer of *Shh* and recapitulated the polydactyly phenotype in mouse.
		A number of GWAS hits have been validated as functional enhancers[178]: for example, common variants associated with cancer susceptibility map to a gene desert on chromosome 8, with one SNP demonstrated to disrupt a TCF7L2 binding site and to inhibit long-range activation of the oncogene *MYC*[179–181].
Synonymous mutations within protein-coding sequences	All of the aforementioned regulatory elements can also be encoded within the protein-coding exons themselves. Thus, synonymous mutations within protein-coding regions may be associated with noncoding functions, acting pre-transcriptionally at the DNA level or post-transcriptionally at the RNA level.	A synonymous variant in the dopamine receptor gene *DRD2* associated with schizophrenia and alcoholism has been shown to modulate receptor production through differences in mRNA folding and stability[182].

RegulomeDB[55] and ENSEMBL's Variant Effect Predictor[56] aim to annotate noncoding common variants from association studies using conservation, functional genomics and regulatory motif data. Databases, such as ANNOVAR[57] and VAAST[58], are specialized for annotating whole-genome and whole-exome sequencing data, and they leverage population-level negative selection to identify extremely rare coding alleles that are most likely to be functional. None of these tools presently brings together all of the available annotation resources listed in the previous section, however, and they will need to be continuously updated to reflect the exponential growth of regulatory knowledge (**Table 4**).

Gene set enrichment analysis. Prior knowledge of gene interrelationships has been leveraged in studies of gene expression to discover differentially regulated pathways, even where single genes in those pathways change expression too little to rise to statistical significance[59]. These methods for gene-set enrichment analysis (GSEA) are being applied to GWAS, where, similarly, genetic risk is expected to be concentrated along biological pathways and multiple testing diminishes the statistical significance of associations considered individually. Dozens of methods have been developed to use prior knowledge from gene functional annotation databases to perform pathway analysis on GWAS[60,61] (**Fig. 3a**).

Regulatory element enrichment analysis. A recent study used chromatin state maps to discover an enrichment of cell type–specific enhancers among the top associations in several GWAS[62] (L.D.W., M.K. and colleagues), demonstrating the ability of high-resolution functional genomics maps to serve as a type of pathway annotation. Similar results have been seen using DNase hypersensitivity maps and protein-binding maps across a large number of cell types[63,89,191] and by examining concordance between expression quantitative trait loci (eQTLs) and

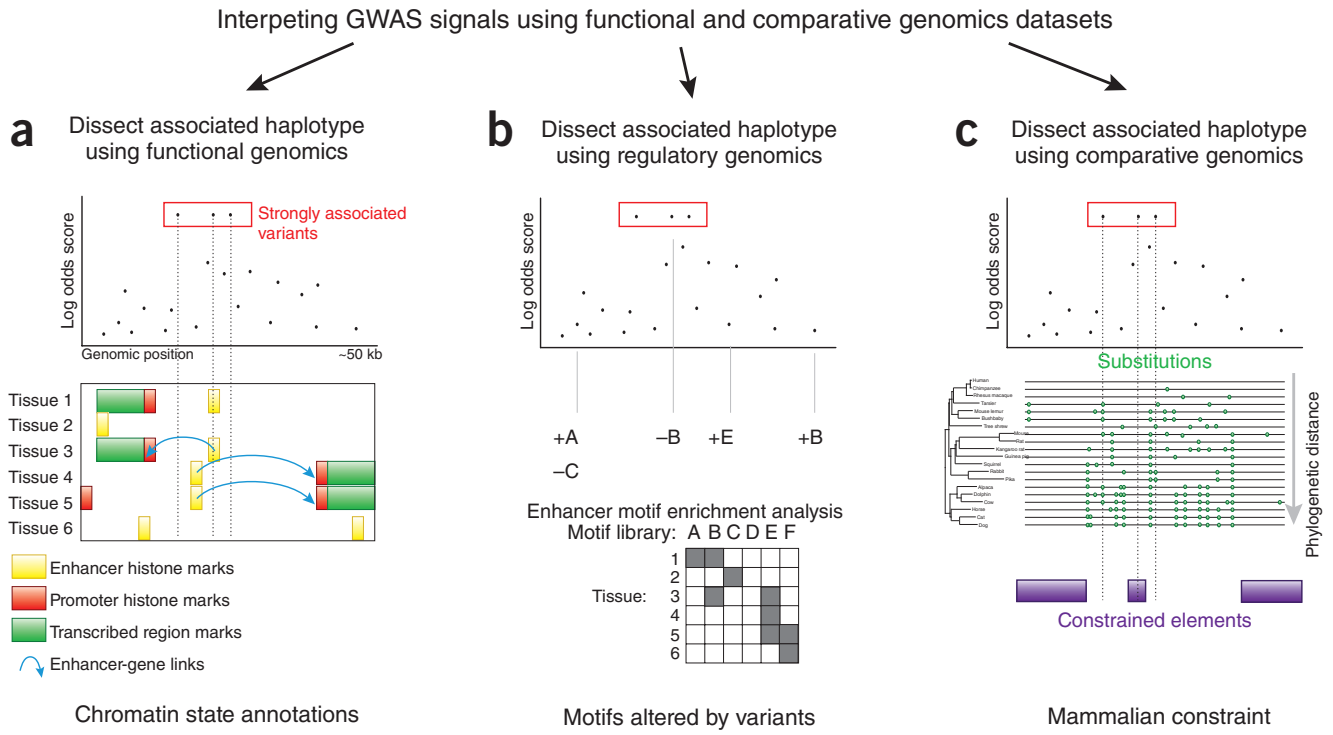

Figure 2 Dissecting haplotypes discovered through association tests. These three examples are ways to annotate loci containing several linked SNPs (in this case, three) to discover those most likely to be causal. (**a**) Functional genomics techniques are being developed to discover putative regulatory elements and link these elements to their target genes. Here, the middle SNP lies in an enhancer in tissue 1 and tissue 3, and regulates a gene to its left. (**b**) Regulatory genomics information leads to prediction of sequence motifs active in classes of enhancers, and this can be combined with the motif creation and/or disruption caused by variants. In this case, the middle SNP deletes a match to motif B, which is predicted to be active in enhancers found in both tissues 1 and 3. (**c**) Comparative genomics identifies regions of evolutionary constraint in noncoding sequence. Here, sequence surrounding only the middle SNP is constrained across mammals.

GWAS[64,65]. These approaches have demonstrated the power of reference epigenomes to identify relevant tissues for further study (**Fig. 3b**). Another way to use prior knowledge about variant function is to incorporate the information into the association study itself through Bayesian methods[61,66–69] or by using boosting to prioritize disease networks[70]. However, it is difficult to evaluate the utility of these weighting schemes, which essentially discard loci about which there is the least functional data.

Burden and aggregation tests for dealing with heterogeneity. For potentially causal rare variants discovered through whole-genome sequencing, a class of techniques has been developed that deal successfully with allelic heterogeneity and low allele frequencies by pooling mutations across individuals by genes, pathways or other functional annotations

Table 4 Comparison of recent tools to systematically annotate variants[a]

Tool	Type	Input method	Protein annotation	Regulatory annotation	Other
SeattleSeq (http://snp.gs.washington.edu/SeattleSeqAnnotation/)	Server	Variants	Deleteriousness scores	Conservation scores	dbSNP clinical association data
ANNOVAR[57]	Software	Variants, regions	User defined: user downloads desired variation, conservation, coding and noncoding functional annotations		
ENSEMBL VEP[56]	Server	Variants, regions	Deleteriousness scores	Regulatory motif alteration scores	OMIM, GWAS data
VAAST[58]	Software	Variants	Deleteriousness scores	Conservation scores	Aggregation to discover rare variants in case-control studies
HaploReg[54]	Server	Variants, studies	dbSNP consequence data	Chromatin state, protein binding, DNase, conservation, regulatory motif alteration scores	GWAS data, eQTL, LD calculation, enrichment analysis per study
RegulomeDB[55]	Server	Variants, regions	Not applicable	Histone modification, protein binding, DNase, conservation, regulatory motif alteration scores	eQTL, reporter assays, combined score analysis per variant

[a]Many such tools have been released as databases or software in the past decade; a sampling of the most recent are listed here.

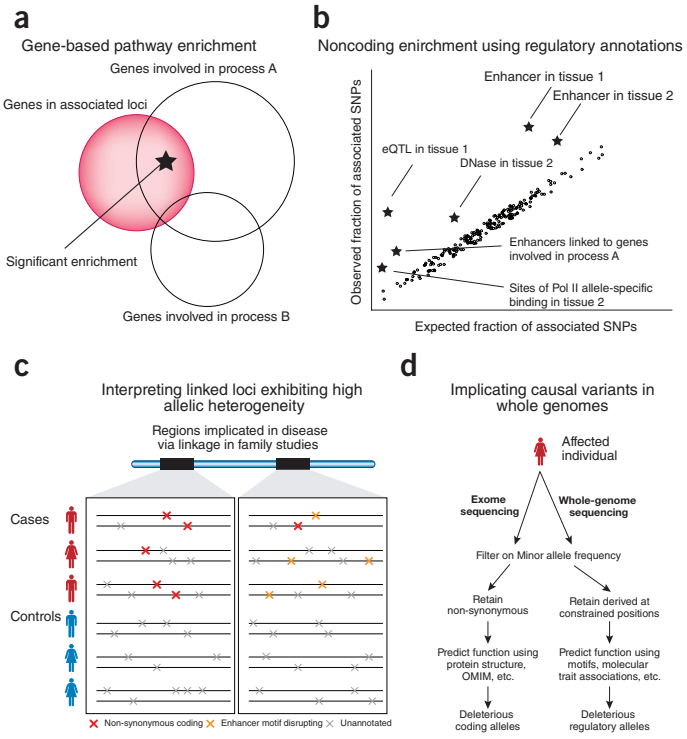

Figure 3 Systems-level analyses beyond isolated common haplotypes. (**a**) Gene-based enrichment analysis of genetic architecture. A typical analysis of GWAS results will compare the set of genes near associated loci with prior knowledge about those genes, leading to hypotheses about the pathways involved (in this example, process A but not process B). (**b**) Noncoding enrichment analysis of genetic architecture using regulatory annotations. High-resolution maps of diverse regulatory annotations can also be intersected with GWAS results. Examples are shown in which tissue-associated enhancers, eQTLs, DNase peaks or allele-specific polymerase binding are enriched among the results of a GWAS. In addition, regulatory annotations can be combined with gene-based annotations and linking information, in this case discovering an enrichment for enhancers linked to the genes involved in process A. (**c**) Interpreting linked loci exhibiting high allelic heterogeneity. In some cases only rare mutations at a locus contribute to its genetic mechanism, and these regions may only be discovered through classical linkage analysis. These regions can now be interrogated through exome or whole-genome sequencing, and an imbalanced burden of putatively deleterious alleles may be observed in cases (as in the left example). With regulatory annotations, these burden and aggregation tests can now be extended to noncoding regions (as in the right example). (**d**) Interpreting causal variants in whole genomes. Personal genomes pose the challenge of exposing potentially causal variants that were too rare or low in penetrance to have been associated with a phenotype through association or linkage studies. For coding alleles, prior knowledge is currently used in several ways when analyzing personal genomes: knowledge of the genetic code (to filter on nonsynonymous variants), inference of negative selection from population panels (to filter out common variants) and models developed from biophysical principles (to focus on those amino acid substitutions most likely to alter protein structure and function). Similar pipelines will need to be developed for regulatory regions. We propose using both population-level and cross-species signals of selection (to filter out not only common variants, but those that are not constrained across mammals) and all of the regulatory models previously mentioned (predicted regulatory elements and the motifs active within them, molecular trait associations, including eQTLs, etc.). Such a pipeline will be crucial to interpreting the flood of sequencing data that will be collected in both clinical and research settings.

and filters[71]; the additional use of functional genomic maps has recently been proposed[72]. Improved annotation of noncoding regions obviously will empower this type of analysis (**Fig. 3c**). **Table 5** lists examples of new insights from computational methods integrating regulatory elements with GWAS.

Interpreting variants using population variation in molecular phenotypes

Although until this point we have discussed regulatory annotations from reference cell lines, biochemical activity is itself dependent on genotype, and thus a single reference annotation fails to capture the complexity of the regulatory genome. Moreover, we treated linkage disequilibrium as a property of the human genome, although it is in fact population specific, and patterns of linkage disequilibrium and selection have varied across both geography and time. This increased complexity can in fact be leveraged to both gain additional insights into genome regulation and provide additional power for the aforementioned analyses.

Genotype-associated molecular activity. Two powerful tools have emerged to identify noncoding loci that affect molecular phenotypes: association studies and allele-specificity studies. Association studies (**Fig. 1b**) have been used to discover noncoding *cis* regulators of methylation (meQTLs)[73], DNase I sensitivity (dsQTLs)[74], transcription factor binding[75], gene expression (eQTLs)[76] and alternative splicing[77]. In the same manner as GWAS on organism-level quantitative traits, these studies consider a phenotype associated with a particular genomic locus (for example, steady-state mRNA level corresponding to a gene) in the same cell type isolated across unrelated individuals and then search for genetic modulators of those molecular processes. A recent related study used eQTL data to reveal selective signatures of epistasis between deleterious coding variants and the regulatory variants that modulate their penetrance[78], a method that should be broadly applicable to testing hypotheses about *cis* regulatory interactions from genomics models.

Allele-specific activity. Instead of variation between individuals, allele specificity tests evaluate molecular activity at heterozygous sites within each individual and look for a skew in the molecular signal toward one of the alleles (**Fig. 1c**). Allele-specific methylation[79], histone modification[80], DNase I sensitivity[81], protein binding[82] and expression[83] have been surveyed across the genome. Although association studies have the advantage of identifying regulatory variants that may be acting at some genetic distance from the regulated locus and can include homozygous individuals in the sample, allele-specific studies can be performed on single individuals and inherently control for possible *trans*-regulatory differences caused by individuals' genetic background.

Importance of population-specific effects. Causal variants within associated haplotypes should be identified not only for further research, but also for genetic counseling; because of variations in patterns of linkage disequilibrium, a SNP that marks a risk haplotype efficiently in one population may not in another[84]. Computational methods that explicitly model ethnic background in admixed populations can increase their power by exploiting their shared ancestry[85].

Population differentiation and positive selection. Haplotype structure and allele frequencies from the HapMap project[9] and 1000 Genomes Project[86] provide evidence of both positive and negative selection currently acting on the human lineage. Although the relative importance of population structure and selective sweeps in recent human history is debated[87–89], many noncoding loci show multiple lines of evidence for local adaptation[90].

Using population structure and relatedness. Ultimately, linkage analysis and GWAS are sensitive to complementary genetic architectures, but it is likely that a wide spectrum of diseases exhibit both locus and allele

Table 5 Examples of regulatory enrichment analyses of genetic associations

Class of test	Finding	Computational tools used
Gene-set enrichment near associated loci	Regulatory network of five proteins is implicated in Kawasaki disease[183]	Ingenuity Pathway Analysis (closed source)
	Genes differentially expressed in adipose overlap with genetic associations with obesity[184]	Microarray analysis of differential expression
	TGF-β pathway and Hedgehog signaling pathway are enriched among height GWAS loci[103]	GSEA using MAGENTA[185], network from text-mining using GRAIL[186], known disease genes from OMIM[4] and eQTL enrichment
Concordance with eQTL results	eQTL prioritization during replication facilitates validation of two Crohn's disease susceptibility loci[187]	eQTL enrichment
	GWAS involving immune system show enrichment for lymphoblastoid eQTL[64]	eQTL enrichment (RTC[64])
Chromatin state enrichment	Many GWAS show enrichment for enhancers in biologically relevant cell types[62]	ChromHMM to define discrete chromatin states (M.K. and colleagues[188]); and enrichment analysis
TF binding site and DNase hypersensitivity enrichment	Many GWAS show enrichment for ENCODE-annotated DNase and ChIP sites[189]	Enrichment analysis
	Many GWAS show enrichment for DNase in biologically relevant cell types[63]	Hot-spot algorithm to define discrete hypersensitive sites[190]; enrichment analysis
	FOXA1 and estrogen receptor binding sites are enriched among breast cancer GWAS loci[191]	Variant Set Enrichment (VSE[191])

heterogeneity. Because the genomically distributed signals of association with complex disease are weak, the potential confounding effects of population stratification and cryptic relatedness become especially important to control. Family-based methods, such as linkage analysis and the transmission disequilibrium test (TDT), are free of these complications, and these have been combined with association tests in a new class of methods[91]. In addition, new methods in phylogenomics and ancestral recombination graph reconstruction provide an opportunity to enhance association studies by explicitly taking population structure and region-specific relatedness into account[92,93].

Aggregate measures of purifying selection. Modeling of allele frequency data[94,95] and sequence divergence data[46] suggests that a large amount of negative selection is occurring outside of mammalian conserved elements, evidence for widespread noncoding function. These same forces can maintain disease-associated alleles at lower frequency in the population dependent on their penetrance and expressivity.

Identifying higher-order relationships between variants

Even when analyzing genome-wide enrichments of functional annotations in disease-associated regions, the aforementioned methods have so far considered each locus as acting independently and treated their effects as additive. Functional genomics should enable us to consider higher-order interactions between these individual loci, by leveraging functional and variation information to build interaction and regulatory networks. These networks can then guide the search for epistatic effects.

Detecting epistasis *de novo*. Substantial disagreement exists over the relative importance of epistasis in the genetic basis of complex disease[96–98]. Although genetic interactions have been systematically mapped in yeast[99], and cases have been identified in human[66], testing for all possible interactions remains impossible; understandably, detecting epistasis in association studies is an area of intense theoretical interest[66,100,101]. One method[102] successfully discovered epistasis between two taste receptor genes affecting nicotine dependence using a multifactor dimensionality reduction (MDR) method integrated with linkage information from a pedigree disequilibrium test, similar to the hybrid linkage-association studies described previously[91].

Guiding the search for epistasis. Some methods propose to limit the search space for interactions by searching only among the most significant independently associated loci; this method failed to discover any interactions among the 180 loci reported to be associated with height[103]. Another proposed limit on the search space exploits prior knowledge from gene annotations and protein-protein interactions[104–106]. Again, epigenomic maps and improved regulatory annotation hold promise for focusing in on relevant combinations of SNPs that might be expected to interact.

Linking enhancers to their target genes using physical interaction data. Unlike promoters, enhancers pose the dual challenge of both pinpointing their location in vast nonfunctional sequences and linking them to their target genes. These distal regulatory elements often interact physically with promoters, and technologies to detect these interactions, such as chromatin conformation capture (3C, Hi-C)[107,108] and chromatin interaction paired-end tagging (ChIA-PET)[109], are advancing rapidly.

Linking enhancers to their target genes using cell-to-cell variability. Another way of detecting enhancer-gene relationships is to measure the correlation of these elements' activity with expression across multiple cell types and conditions. This technique is being used to infer gene regulatory networks in human[35] and model organisms[99,110]. Although protein-protein interaction and metabolic networks are the most common types of prior knowledge integrated into existing algorithms, these regulatory networks may provide a more useful starting point in the search for epistasis.

Inferring networks from individual-to-individual variability. Molecular QTL data discovered from interindividual variation can also be used to help infer regulatory networks[111], and unlike evidence obtained solely from expression patterns, this provides unambiguous directionality for causality.

Inferring networks from systematic perturbations. Chemical perturbations of cultured cells have been used for network inference. These experiments are useful not only for their relevance to understanding pharmacological mechanisms, but also for revealing the difference in network topology between normal and cancerous cells[112], including gene-gene and gene-drug interactions relevant to interpreting the genetic architecture of cancer.

Artificial selection and drug response experiments in model organisms. Although human genetic history and selective pressures are closely intertwined, model organisms offer an opportunity to measure the global effects of selection and the resulting genetic interactions in a controlled setting[113,114]. Model organisms have also proven useful for testing gene-gene[99] and gene-drug[115] interactions on a scale that is impossible in humans.

Functional genomics in a medical setting

Although genotyping and sequencing are already becoming commonplace for discovery of disease loci and increasingly for diagnostics in a clinical setting, in the future the democratization of genome-wide molecular profiling technologies will further enable cohort-level molecular association studies and personal functional genomics in a medical setting. These can complement existing genetic and chemical biomarkers with molecular-level diagnostics of disease state.

Functional genomics of disease cohorts. One of the major clinical applications of DNA microarrays was to identify disease-involved genes and to classify disease subtypes by genome-wide expression signatures[116]. Disease-associated gene sets from microarrays and now RNA-seq can be used to define biological pathways, such as those in the Molecular Signatures Database (MSigDB)[117]. Similarly, chromatin maps can be compared across lineages or between disease and normal tissue to define sets of regulating loci (**Fig. 1d**). These sets can be used for enrichment and pathway analysis of GWAS, as described previously.

Epigenome–phenotype association. Microarray-based assays for methylation are now allowing for the first time 'epigenome-wide association studies' (EWAS)[118], which identify sites whose epigenomic modifications are associated with disease without taking into account genotype (**Fig. 1d**). Such studies may bypass some of the environmental variability that lowers the penetrance of genetic factors[119]. Integrating family members into EWAS studies may be especially useful to test for imprinting and other parent-of-origin effects.

Genetic association with molecular phenotypes for determining causality. One important future use of molecular QTLs may be to empower Mendelian randomization studies[120,121]. Molecular traits—expression, epigenetic state or biomarkers—can be important stepping stones between genetic variation and complex phenotypes, but the direction of causality between the molecular trait and the organismal trait can be unclear. A recent study has used this method to challenge the idea that raising high-density lipoprotein (HDL) cholesterol levels reduces risk of myocardial infarction, showing that alleles for higher HDL did not convey the genetic protection from heart disease that would be expected if cholesterol were causal[122].

Predicting molecular consequences of rare and private mutations. Once these regulatory mechanisms are predicted from functional genomics and molecular variation, the next challenge is applying this knowledge to rare variants discovered by whole-genome sequencing (**Fig. 2d**). A goal for regulatory genomics should be to develop models that predict the effect of novel regulatory variants with the same accuracy as existing methods for predicting effects of novel protein-coding variants.

Functional genomics of individuals. Some expression signatures of disease subtypes or progression are already being used clinically, and their use promises to grow. However, in a problem analogous to that of rare variants discovered through sequencing, clinical functional genomics samples will also exhibit patterns too rare in the population to have been correlated with disease. As a recent pilot study on an individual demonstrates[123], there are both great power and also many challenges associated with interpreting such personal-omics profiling, and new computational models are needed that can generalize from the effects of common genetic and functional variation to personal genetics and functional genomics.

Hurdles in biomedical informatics and interoperability

In addition to these conceptual challenges of statistical and computational integration of disparate data sets, each of these topics has relied on extensive data sharing between genomics and medical genetics researchers. Practically, however, such sharing is still limited owing to privacy concerns and informatics challenges of database interoperability. These challenges are even greater for nongenomic data sets, such as medical records and drug response, resulting in treasure troves of information remaining unused. To complete the integration of genomics into the drug discovery and target validation pipelines, several additional hurdles need to be overcome.

GWAS *P*-value sharing. To facilitate integrative analysis, GWAS investigators should report the association of all variants, not just those that are most significant. *Nature Genetics* recently articulated a policy to this effect[124], but concerns remain about sufficiently de-identifying association results to protect subject privacy[125]. Procedures in place at central archives such as the National Center for Biotechnology Information's (Bethesda, MD) database of Genotypes and Phenotypes (dbGaP) and the European Genome-Phenome Archive (EGA) are crucial to balancing the rights of human subjects with the principles of scientific openness.

Database integration. The interoperability of databases remains paramount to integrative analysis. Continuing efforts by the University of California, Santa Cruz (UCSC) Genome Browser and the ENSEMBL Genome Browser have facilitated integration of epigenomic and variation data, but better connections to domain-specific knowledge bases such as the GTex eQTL Browser, dbGaP analyses and the NHGRI GWAS Catalog[11] would broaden the scope of connections available to geneticists.

Medical record standardization. Medical records have been successfully mined to discover epidemiological patterns[126], adverse drug reactions[127] and disease risk factors and heterogeneity[128]. As electronic medical records become populated with genetic data, cooperation with clinicians will be needed to mine patient data for genetic associations with biomarkers and disease as well as discover novel patterns of disease heterogeneity[129].

Integration of medical and pharmacogenomics data sets. Ultimately, informatics challenges will need to be resolved in order to connect the resulting molecular predictions to patient records, environmental variables, drug screening and response databases so as to move toward enabling genomics as commonplace for clinical practice.

Conclusions

Data from GWAS and whole-genome sequencing continue to expand the catalog of noncoding variants implicated in human disease, and data from epigenome mapping consortia complemented with regulatory modeling are needed to prioritize candidate causal variants and candidate affected tissues. To date, from the large numbers of noncoding variants associated with human disease, the molecular mechanisms of action for only very few have been successfully characterized. However, thoughtful integration of systematic and manual annotations of gene sets along with higher-resolution functional maps may hold the key to implicating pathways and cell types, both through joint consideration

of the many weak additive associations discovered in GWAS and in the search for epistatic interactions between variants. Clinically relevant regulatory interactions may then be tested experimentally in the tissues or *in vitro* experimental conditions that are predicted to recapitulate the relevant phenotypes. In addition, an explosion of functional genomics data has been facilitated by high-throughput sequencing technology, allowing 'intermediate' molecular phenotypes to be correlated both with organismal phenotype and with genotype. This new type of data can be combined with genetic associations to decipher the mechanisms underlying complex disease.

ACKNOWLEDGMENTS

L.D.W. and M.K. were funded by NIH grants R01HG004037 and RC1HG005334 and US National Science Foundation CAREER grant 0644282.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nbt.2422.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Collins, F. Has the revolution arrived? *Nature* **464**, 674–675 (2010).
2. Lander, E.S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011).
3. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33**, 228–237 (2003).
4. Hamosh, A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
5. Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
6. Lander, E.S. & Botstein, D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199 (1989).
7. Watson, J.D. The Human Genome Project: past, present, and future. *Science* **248**, 44–49 (1990).
8. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**, 241–247 (1995).
9. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
10. McCarthy, M.I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
11. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
The NHGRI GWAS Catalog reported here laid the groundwork for systematic intersection of functional annotations with disease-associated regions, and highlighted the preponderance of noncoding disease associations.
12. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
This paper reports the deliberations of the NHGRI's expert working group on the sources of unexplained heritability, and their suggestions for future research strategies.
13. Cirulli, E.T. & Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).
14. Fisher, R. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
15. Visscher, P.M. McEvoy, B. & Yang, J. From Galton to GWAS: quantitative genetics of human height. *Genet. Res.* **92**, 371–379 (2010).
16. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
17. Nelson, M.R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
18. Park, P.J. ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).
19. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33**, 5868–5877 (2005).
20. Boyle, A.P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
21. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
The ENCODE consortium scale-up datasets represent the most comprehensive annotation of the noncoding genome at the time of this review.
22. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
23. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.* **30**, 224–226 (2012).
24. Bussemaker, H.J., Foat, B.C. & Ward, L.D. Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 329–347 (2007).
25. Tompa, M. *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23**, 137–144 (2005).
26. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465**, 53–59 (2010).
27. Wang, Z. & Burge, C.B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802–813 (2008).
28. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
29. Moses, A.M., Chiang, D., Pollard, D., Iyer, V. & Eisen, M. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.* **5**, R98 (2004).
30. Hesselberth, J.R. *et al.* Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).
31. Henikoff, J.G., Belsky, J.A., Krassovsky, K., MacAlpine, D.M. & Henikoff, S. Epigenome characterization at single base-pair resolution. *Proc. Natl. Acad. Sci. USA* **108**, 18318–18323 (2011).
32. Rhee, H.S. & Pugh, B.F. Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419 (2011).
33. Beer, M.A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).
34. Roy, S. *et al.* Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
35. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
36. Davidson, E.H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
37. Patwardhan, R.P. *et al.* Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nat. Biotechnol.* **30**, 265–270 (2012).
38. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
39. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
40. Davydov, E.V. *et al.* Identifying a high fraction of the human to be under selective constraint using GERP++. *PLOS Comput. Biol.* **6**, e1001025 (2010).
41. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
Conserved elements were shown to be enriched among disease-associated variants, motivating the use of conservation to guide candidate causal SNP selection.
42. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
43. Stark, A. *et al.* Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**, 219–232 (2007).
44. Papatsenko, D., Kislyuk, A., Levine, M. & Dubchak, I. Conservation patterns in different functional sequence categories of divergent *Drosophila* species. *Genomics* **88**, 431–442 (2006).
45. Dermitzakis, E.T. & Clark, A.G. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19**, 1114–1121 (2002).
46. Meader, S., Ponting, C.P. & Lunter, G. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res.* **20**, 1335–1343 (2010).
47. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).
48. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
49. Ng, P.C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
50. Yue, P., Melamud, E. & Moult, J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* **7**, 166 (2006).
51. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894–3900 (2002).
52. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
53. Baker, M. Functional genomics: the changes that count. *Nature* **482**, 257–262 (2012).
54. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
55. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
56. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. *Bioinformatics* **26**, 2069–2070 (2010).
57. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
58. Yandell, M. *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Res.* 10.1101/gr.123158.111 (2011).
59. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).

60. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* **11**, 843–854 (2010).
61. McKinney, B.A. & Pajewski, N.M. Six degrees of epistasis: statistical network models for GWAS. *Front. Genet* **2**, 109 (2012).
62. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
This was the first demonstration that cross-tissue enhancer maps can link noncoding variants from GWAS to relevant cell types and candidate regulatory mechanisms.
63. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
64. Nica, A.C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
This study uses eQTLs to investigate the tissue specificity of gene regulatory mechanisms, and suggests that assaying many tissues will be critical to developing a cis-regulatory map of the human genome.
65. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
66. Cantor, R.M., Lange, K. & Sinsheimer, J.S. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* **86**, 6–22 (2010).
The authors present an extensive review of how biological annotations are being used in association studies and to interpret their results. They show how knowledge of molecular pathways can be used to enhance discovery, test for epistasis and aggregate results.
67. Knight, J., Barnes, M.R., Breen, G. & Weale, M.E. Using functional annotation for the empirical determination of Bayes factors for genome-wide association study analysis. *PLoS ONE* **6**, e14808 (2011).
68. Lewinger, J.P., Conti, D.V., Baurley, J.W., Triche, T.J. & Thomas, D.C. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet. Epidemiol.* **31**, 871–882 (2007).
69. Chen, G.K. & Witte, J.S. Enriching the analysis of genomewide association studies with hierarchical modeling. *Am. J. Hum. Genet.* **81**, 397–404 (2007).
70. Lee, I., Blom, U.M., Wang, P.I., Shim, J.E. & Marcotte, E.M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**, 1109–1121 (2011).
71. Dering, C., Hemmelmann, C., Pugh, E. & Ziegler, A. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet. Epidemiol.* **35**, S12–S17 (2011).
72. Bansal, V., Libiger, O., Torkamani, A. & Schork, N.J. Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* **11**, 773–785 (2010).
73. Pai, A.A., Bell, J.T., Marioni, J.C., Pritchard, J.K. & Gilad, Y. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet.* **7**, e1001316 (2011).
74. Degner, J.F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
75. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).
76. Majewski, J. & Pastinen, T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.* **27**, 72–79 (2011).
77. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
78. Lappalainen, T., Montgomery, S.B., Nica, A.C. & Dermitzakis, E.T. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am. J. Hum. Genet.* **89**, 459–463 (2011).
79. Kerkel, K. *et al.* Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat. Genet.* **40**, 904–908 (2008).
80. Prendergast, J.G., Tong, P., Hay, D.C., Farrington, S.M. & Semple, C.A. A genome-wide screen in human embryonic stem cells reveals novel sites of allele-specific histone modification associated with known disease loci. *Epigenetics Chromatin* **5**, 6 (2012).
81. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235–239 (2010).
In this study, the authors demonstrated that both genomic protein binding and DNase I hypersensitivity were heritable, and therefore under genetic control.
82. Maynard, N.D., Chen, J., Stuart, R.K., Fan, J.-B. & Ren, B. Genome-wide mapping of allele-specific protein-DNA interactions in human cells. *Nat. Methods* **5**, 307–309 (2008).
83. Ge, B. *et al.* Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.* **41**, 1216–1222 (2009).
84. Ng, P.C., Murray, S.S., Levy, S. & Venter, J.C. An agenda for personalized medicine. *Nature* **461**, 724–726 (2009).
85. Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74**, 979–1000 (2004).
86. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
87. Coop, G. *et al.* The role of geography in human adaptation. *PLoS Genet.* **5**, e1000500 (2009).
88. Hernandez, R.D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920–924 (2011).
89. Sabeti, P.C. *et al.* Positive natural selection in the human lineage. *Science* **312**, 1614–1620 (2006).
90. Grossman, S.R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883–886 (2010).
91. Ott, J., Kamatani, Y. & Lathrop, M. Family-based designs for genome-wide association studies. *Nat. Rev. Genet.* **12**, 465–474 (2011).
92. Minichiello, M.J. & Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.* **79**, 910–922 (2006).
93. Wu, Y. Association mapping of complex diseases with ancestral recombination graphs: models and efficient algorithms. *J. Comput. Biol.* **15**, 667–684 (2008).
94. Asthana, S. *et al.* Widely distributed noncoding purifying selection in the human genome. *Proc. Natl. Acad. Sci. USA* **104**, 12410–12415 (2007).
95. Ward, L.D. & Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 10.1126/science.1225057 (2012).
96. Hill, W.G., Goddard, M.E. & Visscher, P.M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4**, e1000008 (2008).
97. Shao, H. *et al.* Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proc. Natl. Acad. Sci. USA* **105**, 19910–19914 (2008).
98. Zuk, O., Hechter, E., Sunyaev, S.R. & Lander, E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. USA* 10.1073/pnas.1119675109 (2012).
99. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
100. Cordell, H.J. Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.* **10**, 392–404 (2009).
101. Musani, S.K. *et al.* Detection of gene x gene interactions in genome-wide association studies of human population data. *Hum. Hered.* **63**, 67–84 (2007).
102. Lou, X.-Y. *et al.* A combinatorial approach to detecting gene–gene and gene–environment interactions in family studies. *Am. J. Hum. Genet.* **83**, 457–467 (2008).
103. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
104. Emily, M., Mailund, T., Hein, J., Schauser, L. & Schierup, M.H. Using biological networks to search for interacting loci in genome-wide association studies. *Eur. J. Hum. Genet.* **17**, 1231–1240 (2009).
105. Mechanic, L.E., Luke, B.T., Goodman, J.E., Chanock, S.J. & Harris, C.C. Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions. *BMC Bioinformatics* **9**, 146 (2008).
106. Pattin, K.A. & Moore, J.H. Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. *Hum. Genet.* **124**, 19–29 (2008).
107. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
108. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
109. Fullwood, M.J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
110. Cheng, C. *et al.* Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLOS Comput. Biol.* **7**, e1002190 (2011).
111. Zhu, J. *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* **40**, 854–861 (2008).
112. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
113. Burke, M.K. *et al.* Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* **467**, 587–590 (2010).
114. Gresham, D. *et al.* The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet.* **4**, e1000303 (2008).
115. Perlstein, E.O., Ruderfer, D.M., Roberts, D.C., Schreiber, S.L. & Kruglyak, L. Genetic basis of individual differences in the response to small-molecule drugs in yeast. *Nat. Genet.* **39**, 496–502 (2007).
116. Quackenbush, J. Microarray analysis and tumor classification. *N. Engl. J. Med.* **354**, 2463–2472 (2006).
117. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
118. Rakyan, V.K., Down, T.A., Balding, D.J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* **12**, 529–541 (2011).
The authors review the challenges and promise of EWAS, and how their results can be used in conjunction with GWAS.
119. Petronis, A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* **465**, 721–727 (2010).
120. Chen, L.S., Emmert-Streib, F. & Storey, J.D. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol.* **8**, R219 (2007).
121. Lawlor, D.A., Harbord, R.M., Sterne, J.A.C., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
122. Voight, B.F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* **380**, 572–580.
123. Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293–1307 (2012).
124. Anonymous. Asking for more. *Nat. Genet.* **44**, 733 (2012).
125. Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4**, e1000167 (2008).
126. Salathé, M. *et al.* Digital epidemiology. *PLOS Comput. Biol.* **8**, e1002616 (2012).
127. Brownstein, J.S., Sordo, M., Kohane, I.S. & Mandl, K.D. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial Infarction. *PLoS ONE* **2**, e840 (2007).

128. Roque, F.S. *et al.* Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLOS Comput. Biol.* **7**, e1002141 (2011).
129. Wilke, R.A. *et al.* The emerging role of electronic medical records in pharmacogenomics. *Clin. Pharmacol. Ther.* **89**, 379–386 (2011).
130. Nebert, D.W., Zhang, G. & Vesell, E.S. From human genetics and genomics to pharmacogenetics and pharmacogenomics: past lessons, future directions. *Drug Metab. Rev.* **40**, 187–224 (2008).
A critical review of current challenges in human genetics and the application of pharmacogenetic discoveries to clinical practice.
131. Garrod, A. E. & Harris, H. *Inborn Errors of Metabolism* (Henry Frowde and Hodder & Stoughton, London, 1909).
132. Woo, S.L., Lidsky, A.S., Güttler, F., Chandra, T. & Robson, K.J. Cloned human phenylalanine hydroxylase gene allows prenatal diagnosis and carrier detection of classical phenylketonuria. *Nature* **306**, 151–155 (1983).
133. Riordan, J.R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066–1073 (1989).
134. Audrézet, M.P. *et al.* Genomic rearrangements in the CFTR gene: extensive allelic heterogeneity and diverse mutational mechanisms. *Hum. Mutat.* **23**, 343–357 (2004).
135. Zschocke, J. Phenylketonuria mutations in Europe. *Hum. Mutat.* **21**, 345–356 (2003).
136. Amiel, J. *et al.* Hirschsprung disease, associated syndromes and genetics: a review. *J. Med. Genet.* **45**, 1–14 (2008).
137. Yang, J. *et al.* Common SNPs explain a large proportion of heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
138. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
139. Nica, A.C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7**, e1002003 (2011).
140. King, J.L. & Jukes, T.H. Non-Darwinian evolution. *Science* **164**, 788–798 (1969).
141. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
142. Ohno, S. So much 'junk' DNA in our genome. *Brookhaven Symp. Biol.* **23**, 366–370 (1972).
143. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719–724 (2009).
144. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
145. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* **3**, e114 (2007).
146. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
147. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
148. Veyrieras, J.-B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4**, e1000214 (2008).
149. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
150. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7**, 522 (2011).
151. Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 3 (2004).
152. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
153. Korn, J.M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).
154. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
155. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
156. Faustino, N.A. & Cooper, T.A. Pre-mRNA splicing and human disease. *Genes Dev.* **17**, 419–437 (2003).
157. Cáceres, J.F. & Kornblihtt, A.R. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.* **18**, 186–193 (2002).
158. López-Bigas, N., Audit, B., Ouzounis, C., Parra, G. & Guigó, R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.* **579**, 1900–1903 (2005).
159. Barbaux, S. *et al.* Donor splice-site mutations in WT1 are responsible for Frasier syndrome. *Nat. Genet.* **17**, 467–470 (1997).
160. Lorson, C.L., Hahnen, E., Androphy, E.J. & Wirth, B. A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc. Natl. Acad. Sci. USA* **96**, 6307–6311 (1999).
161. Cazzola, M. & Skoda, R.C. Translational pathophysiology: a novel molecular mechanism of human disease. *Blood* **95**, 3280–3288 (2000).
162. Bisio, A. *et al.* Functional analysis of CDKN2A/p16INK4a 5′-UTR variants predisposing to melanoma. *Hum. Mol. Genet.* **19**, 1479–1491 (2010).
163. Abelson, J.F. *et al.* Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* **310**, 317–320 (2005).
164. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
165. Ponting, C.P., Oliver, P.L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136**, 629–641 (2009).
166. Bonafé, L. *et al.* Evolutionary comparison provides evidence for pathogenicity of RMRP mutations. *PLoS Genet.* **1**, e47 (2005).
167. Cooper, T.A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* **136**, 777–793 (2009).
168. Knight, J.C. Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med.* **83**, 97–109 (2005).
169. Martin, M.P. *et al.* Genetic acceleration of AIDS progression by a promoter variant of CCR5. *Science* **282**, 1907–1911 (1998).
170. Bream, J.H. *et al.* CCR5 promoter alleles and specific DNA binding factors. *Science* **284**, 223 (1999).
171. Bray, N.J. *et al.* Allelic expression of APOE in human brain: effects of epsilon status and promoter haplotypes. *Hum. Mol. Genet.* **13**, 2885–2892 (2004).
172. St George-Hyslop, P.H. & Petit, A. Molecular biology and genetics of Alzheimer's disease. *C. R. Biol.* **328**, 119–130 (2005).
173. Exner, M., Minar, E., Wagner, O. & Schillinger, M. The role of heme oxygenase-1 promoter polymorphisms in human disease. *Free Radic. Biol. Med.* **37**, 1097–1104 (2004).
174. Kleinjan, D.A. & van Heyningen, V. Long-range control of gene expression: Emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32 (2005).
175. Noonan, J.P. & McCallion, A.S. Genomics of long-range regulatory elements. *Annu. Rev. Genomics Hum. Genet.* **11**, 1–23 (2010).
176. Visel, A., Rubin, E.M. & Pennacchio, L.A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
177. Lettice, L.A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and Fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
178. Sakabe, N.J., Savic, D. & Nobrega, M.A. Transcriptional enhancers in development and disease. *Genome Biol.* **13**, 238 (2012).
179. Pomerantz, M.M. *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.* **41**, 882–884 (2009).
180. Tuupanen, S. *et al.* The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat. Genet.* **41**, 885–890 (2009).
181. Wasserman, N.F., Aneas, I. & Nobrega, M.A. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res.* **20**, 1191–1197 (2010).
182. Duan, J. *et al.* Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.* **12**, 205–216 (2003).
183. Burgner, D. *et al.* A genome-wide association study identifies novel and functionally related susceptibility loci for Kawasaki disease. *PLoS Genet.* **5**, e1000319 (2009).
184. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
185. Segrè, A.V., Groop, L., Mootha, V.K., Daly, M.J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6**, e1001058 (2010).
186. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
187. Fransen, K. *et al.* Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. *Hum. Mol. Genet.* **19**, 3482–3488 (2010).
188. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).
189. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).
190. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).
191. Cowper-Sal·lari, R. *et al.* Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* **44**, 1191–1191 (2012).
192. Kraft, P. & Hunter, D.J. Genetic risk prediction—Are we there yet? *N. Engl. J. Med.* **360**, 1701–1703 (2009).
193. Yngvadottir, B., MacArthur, D.G., Jin, H. & Tyler-Smith, C. The promise and reality of personal genomics. *Genome Biol.* **10**, 237 (2009).
194. Roberts, N.J. *et al.* The predictive capacity of personal genome sequencing. *Sci. Transl. Med.* **4**, 133ra58 (2012).
195. Jostins, L. & Barrett, J.C. Genetic risk prediction in complex disease. *Hum. Mol. Genet.* **20**, R182–R188 (2011).
196. Stahl, E.A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44**, 483–489 (2012).
197. Cooper, G.M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
198. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2011).
199. Goldstein, D.B. The importance of synthetic associations will only be resolved empirically. *PLoS Biol.* **9**, e1001008 (2011).