Published in final edited form as:
Nat Biotechnol. 2012 November ; 30(11): 1095–1106. doi:10.1038/nbt.2422.

Interpreting non-coding variation in complex disease genetics

Lucas D. Ward[1,2] and **Manolis Kellis**[1,2]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

[2]The Broad Institute of MIT and Harvard

Abstract

Association studies provide genome-wide information about the genetic basis of complex disease, but medical research has primarily focused on protein-coding variants, due to the difficulty of interpreting non-coding mutations. This picture has changed with advances in the systematic annotation of functional non-coding elements. Evolutionary conservation, functional genomics, chromatin state, sequence motifs, and molecular quantitative trait loci all provide complementary information about non-coding function. These functional maps can help prioritize variants on risk haplotypes, filter mutations encountered in the clinic, and perform systems-level analyses to reveal processes underlying disease associations. Advances in predictive modeling can enable dataset 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 revolutionize medicine by elucidating relevant biochemical pathways for drug targets and by enabling personalized risk assessments[1,2]. As technologies evolved 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 non-coding variants, wider allelic spectra, numerous loci, and weak effect sizes (Table 1). In the last few years, a new wave of technological advances has intensified the shift towards tackling 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 which could be assayed through restriction fragment length polymorphism (RFLP) analysis[5,6]. However, the regions mapped by linkage

Correspondence should be addressed to L.D.W. (lukeward@mit.edu) and M.K. (manoli@mit.edu).

analysis were necessarily large, and cloning candidate genes for follow-up association studies, resequencing, and functional assays required the application of painstaking molecular techniques before the completion of the Human Genome Project[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], has 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 where 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 diseases 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].

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.

Cumulative predictive power. Generally, the discovered loci reaching genome-wide significance have weak additive predictive power for specific phenotypes, which limits their clinical relevance for some traits at present[130–132]. However, risk prediction using the loci discovered for complex disease using GWAS often performs similarly to using classical clinical tests, and has unique properties, such as stability over the lifespan[133]. Predictors that jointly use hundreds or thousands of weakly-contributing loci have also been shown to explain a larger proportion of variance than was initially appreciated[134,135]. Integrating these discoveries into clinical protocols is in its infancy, and should be expected to mature.

Non-coding variants with unknown effect. Most of the loci are non-coding and many are far from discovered genes, and, because of linkage disequilibrium (LD), encompass many variants; therefore, they are not immediately informative or biochemically tractable for experimental work. Assigning a prior probability to the deleteriousness of a non-coding mutation is challenging[136]. To address this challenge, non-coding sequence is being annotated at a rapid pace through systematic efforts such as the ENCODE Project[21] and the Roadmap Epigenomics Mapping Consortium[22], and through studies of the impact of common variants on genomewide molecular phenotypes, discussed below.

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[137], ranging from arguments that associations with common variants are in fact driven by synthetic associations with large-effect rare variants in long-range LD[138], that common associations of weak effect contribute to heritability well beyond the threshold of statistical significance[139], and that narrow-sense heritability may be overestimated in many twin studies due to epistasis disguised as additivity[98].

Reproducibility. GWAS sometimes do not replicate across studies or populations[140], leading to the report of false positives and suspicion of the validity of novel associations, especially when they are non-coding. This could be partly due to the difficulties both in imputing genotypes, which will benefit from an increased understanding of common human variation, and to the poor definition of organismal phenotypes[140], which can benefit from molecular disease biomarkers discussed below. Moreover, while 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 socio-economic or other environmental factors that occur downstream of molecular phenotypes and can strongly affect organismal phenotypes.

In this review, we discuss both the computational challenges and the opportunities presented by the large number of non-coding 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 identify the most promising causal variants at a locus. We then discuss the utility of these regulatory annotations to perform systems-level analysis of GWAS and allelic spectra, revealing relevant cell types and regulatory mechanisms. Finally, we present a variety of bioinformatics hurdles and computational challenges that lie ahead for the field, such as discovering epistatic interactions, connections between molecular and organismal phenotype, and patterns that must be mined from potentially sensitive medical data.

Systematic annotation of the non-coding 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 non-coding variants, given the diversity of non-coding functions, the incomplete annotation of regulatory elements, and potentially 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 at the transcriptional, splicing, and translational level (Table 3).

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

Recognizing the need for systematic interpretation of non-coding disease-associated variants, several large-scale projects are currently underway to enhance the annotation of the non-coding genome (Fig. 2). These rely on reference annotation maps using both functional genomics and comparative genomics, and can dramatically increase 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) applied to map transcription factor binding, chromatin regulators, or histone modification marks[18], mapping of DNA methylation using bisulfite sequencing (BS-Seq)[19] and mapping of accessible chromatin regions by DNase hypersensitivity analysis (DNase-Seq)[20]. Computational integration of these datasets through supervised or unsupervised machine learning enables mapping of functional non-coding 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 is releasing comprehensive maps of chromatin states, TF binding, and transcription for a selection of cell lines and DNase maps for many primary cells[21], and the NIH 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 towards the most likely causal regulators on a haplotype (Fig. 2a).

Nucleotide-resolution regulatory annotations

While 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 to also 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 different positions of a regulatory motif or element will affects 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 exploit emerging large-scale synthesis and sequencing technologies are being developed both in model organisms and cultured human cells[37–39], and enable testing 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], revealing 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 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), non-coding RNA genes and structures, microRNAs and their targets, and yet uncharacterized sequence elements that confer a selective advantage.

Evolutionarily conserved biochemical activity

Even in 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 due 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 not 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, has 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 datasets to bear on understanding the molecular roles of disease-associated common variants in individual regions, especially for non-coding 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 non-coding variants[53]. A number of resources, including HaploReg[54] (L.D.W. and M.K.), RegulomeDB[55], and ENSEMBL's Variant Effect Predictor[56] aim to annotate non-coding 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/exome sequencing data, and 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 utility of high-resolution functional genomics maps to serve as a type of pathway annotation. Similar results have been seen using DNase hypersensitivity maps across a large number of cell types[63], and by examining concordance between expression quantitative trait loci (eQTLs) and 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 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 tests; 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 and filters[71]; the additional use of functional genomic maps has recently been proposed[72]. Improved annotation of non-coding regions will obviously 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

While until this point we have discussed regulatory annotations from reference cell lines, biochemical activity is itself genotype-dependent, and thus a single reference annotation fails to capture the complexity of the regulatory genome. Moreover, we treated LD as a property of the human genome, while it is in fact population specific, and patterns of LD and selection have varied across both geography and time. This increased complexity can in fact be leveraged to 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 non-coding loci that affect molecular phenotypes: association studies and allele-specificity studies. Association studies (Fig. 1b) have been used to discover non-coding 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 (such as steady-state mRNA level corresponding to a gene) in the same cell type isolated across unrelated individuals, and search for genetic regulators 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 which should be broadly applicable to testing hypotheses about cis regulatory interactions from genomics models.

Allele-specificity activity

In contrast, allele specificity tests look at heterozygous sites in individuals and look for a skew in the molecular signal towards 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 genomewide. While 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 LD patterns, 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 non-coding loci show multiple lines of evidence for local adaptation[90].

Utilizing population structure and relatedness

Ultimately, linkage analysis and GWAS are sensitive to complementary genetic architectures, but a wide spectrum of diseases likely exhibit both locus and allele 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 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 non-coding 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 considering genome-wide enrichments of functional annotations in disease-associated regions, the aforementioned methods have so far considered each locus as acting independently and considered 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]. While 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 by 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 search for epistasis

Some methods propose to limit the search space for interactions by only searching 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 is with prior knowledge from gene annotations and protein-protein interactions[104–106]. Again, epigenomic maps and improved regulatory annotation holds promise for zeroing 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]. While 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 inter-individual variation can also being used to help infer regulatory networks[111], which unlike evidence learned solely from expression patterns provide 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 genetic architecture of cancer.

Artificial selection and drug response experiments in model organisms

While 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

While genotyping and sequencing is 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], and 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 differentially-methylated sites 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 in order 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 can be unclear between the molecular trait and the organismal trait. A recent study used this method to challenge the idea that raising 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 (Figure 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 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, analogous to the problem 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 is both great power but 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 datasets, each of these topics has relied on extensive data sharing between genomics and medical genetics researchers. However, sharing is still limited due to privacy concerns and informatics challenges of database interoperability. These challenges are even greater for non-genomic datasets 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

In order to facilitate integrative analysis, GWAS investigators should report the association of all variants, not just those that are most significant. The editorial board of Nature Genetics recently articulated a policy to this effect[124], but concerns remain about sufficiently de-identifying association results in order to protect subject privacy[125]. Procedures in place at central archives such as the NCBI’s 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 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 in order to mine patient data for genetic associations with biomarkers and disease, and discover novel patterns of disease heterogeneity[129].

Integration of medical and pharmacogenomics datasets

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, towards enabling genomics as commonplace for clinical practice.

CONCLUSIONS

Data from GWAS and whole-genome sequencing continue to expand the catalog of non-coding 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. 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 as well as 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 phenotype. In addition, an explosion of functional genomics data has been facilitated by high-throughput sequencing technology, allowing “intermediate” molecular phenotypes to be correlated with both 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 NSF CAREER grant 0644282.

References

1. Collins F. Has the revolution arrived? Nature. 2010; 464:674–675. [PubMed: 20360716]
2. Lander ES. Initial impact of the sequencing of the human genome. Nature. 2011; 470:187–197. [PubMed: 21307931]
3. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genetics. 2003; 33:228–237. [PubMed: 12610532]
4. Hamosh A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research. 2004; 33:D514–D517. [PubMed: 15608251]

5. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980; 32:314–331. [PubMed: 6247908]

6. Lander ES, Botstein D. Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. Genetics. 1989; 121:185–199. [PubMed: 2563713]

7. Watson JD. The Human Genome Project: Past, Present, and Future. Science. 1990; 248:44–49. [PubMed: 2181665]

8. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet. 1995; 11:241–247. [PubMed: 7581446]

9. Gibbs RA, et al. The International HapMap Project. Nature. 2003; 426:789–796. [PubMed: 14685227]

10. McCarthy MI, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics. 2008; 9:356–369.

11. Hindorff LA, et al. Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits. PNAS. 2009; 106:9362–9367. [PubMed: 19474294]

12. Manolio TA, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]

13. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics. 2010; 11:415–425.

14. Fisher R. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society Edinburgh. 1918; 52:399–433.

15. Visscher PM, McEVOY B, Yang J. From Galton to GWAS: Quantitative Genetics of Human Height. Genetics Research. 2010; 92:371–379. [PubMed: 21429269]

16. MacArthur DG, et al. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science. 2012; 335:823–828. [PubMed: 22344438]

17. Nelson MR, et al. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science. 201210.1126/science.1217876

18. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics. 2009; 10:669–680.

19. Meissner A, et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucl Acids Res. 2005; 33:5868–5877. [PubMed: 16224102]

20. Boyle AP, et al. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell. 2008; 132:311–322. [PubMed: 18243105]

21. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

22. Bernstein BE, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2010; 28:1045–1048. [PubMed: 20944595]

23. Adams D, et al. BLUEPRINT to decode the epigenetic signature written in blood. Nature Biotechnology. 2012; 30:224–226.

24. Bussemaker HJ, Foat BC, Ward LD. Predictive Modeling of Genome-Wide mRNA Expression: From Modules to Molecules. Annual Review of Biophysics and Biomolecular Structure. 2007; 36:329–347.

25. Tompa M, et al. Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology. 2005; 23:137–144.

26. Barash Y, et al. Deciphering the splicing code. Nature. 2010; 465:53–59. [PubMed: 20445623]

27. Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA. 2008; 14:802–813. [PubMed: 18369186]

28. Xie X, et al. Systematic discovery of regulatory motifs in human promoters and 3|[prime]| UTRs by comparison of several mammals. Nature. 2005; 434:338–345. [PubMed: 15735639]

29. Moses A, 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 Biology. 2004; 5:R98. [PubMed: 15575972]

30. Hesselberth JR, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nature Methods. 2009; 6:283–289. [PubMed: 19305407]

31. Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S. Epigenome characterization at single base-pair resolution. PNAS. 2011; 108:18318–18323. [PubMed: 22025700]

32. Rhee HS, Pugh BF. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell. 2011; 147:1408–1419. [PubMed: 22153082]

33. Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell. 2004; 117:185–198. [PubMed: 15084257]

34. Roy S, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010; 330:1787–1797. [PubMed: 21177974]

35. Gerstein MB, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. in press.

36. Davidson EH, et al. A Genomic Regulatory Network for Development. Science. 2002; 295:1669–1678. [PubMed: 11872831]

37. Patwardhan RP, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012; 30:265–270. [PubMed: 22371081]

38. Sharon E, et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol. 2012; 30:521–530. [PubMed: 22609971]

39. Melnikov A, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012; 30:271–277. [PubMed: 22371084]

40. Davydov EV, et al. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++ PLoS Comput Biol. 2010; 6:e1001025. [PubMed: 21152010]

41. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011; 478:476–482. [PubMed: 21993624]

42. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 2003; 423:241–254. [PubMed: 12748633]

43. Stark A, et al. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature. 2007; 450:219–232. [PubMed: 17994088]

44. Papatsenko D, Kislyuk A, Levine M, Dubchak I. Conservation patterns in different functional sequence categories of divergent Drosophila species. Genomics. 2006; 88:431–442. [PubMed: 16697139]

45. Dermitzakis ET, Clark AG. Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover. Mol Biol Evol. 2002; 19:1114–1121. [PubMed: 12082130]

46. Meader S, Ponting CP, Lunter G. Massive turnover of functional sequence in human and other mammalian genomes. Genome Res. 2010; 20:1335–1343. [PubMed: 20693480]

47. Schmidt D, et al. Five-Vertebrate ChIP-Seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. Science. 2010; 328:1036–1040. [PubMed: 20378774]

48. Brawand D, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011; 478:343–348. [PubMed: 22012392]

49. Ng PC, Henikoff S. SIFT: Predicting Amino Acid Changes That Affect Protein Function. Nucl Acids Res. 2003; 31:3812–3814. [PubMed: 12824425]

50. Yue P, Melamud E, Moult J. SNPs3D: Candidate gene and SNP selection for association studies. BMC Bioinformatics. 2006; 7:166. [PubMed: 16551372]

51. Ramensky V, Bork P, Sunyaev S. Human Non-synonymous SNPs: Server and Survey. Nucl Acids Res. 2002; 30:3894–3900. [PubMed: 12202775]

52. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nature Methods. 2010; 7:248–249. [PubMed: 20354512]

53. Baker M. Functional genomics: The changes that count. Nature. 2012; 482:257–262. [PubMed: 22318607]

54. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012; 40:D930–934. [PubMed: 22064851]

55. Boyle AP, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012; 22:1790–1797. [PubMed: 22955989]

56. McLaren W, et al. Deriving the Consequences of Genomic Variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010; 26:2069–2070. [PubMed: 20562413]

57. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucl Acids Res. 2010; 38:e164–e164. [PubMed: 20601685]

58. Yandell M, et al. A Probabilistic Disease-Gene Finder for Personal Genomes. Genome Res. 201110.1101/gr.123158.111

59. Subramanian A, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. 2005; 102:15545–15550. [PubMed: 16199517]

60. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet. 2010; 11:843–854. [PubMed: 21085203]

61. McKinney BA, Pajewski NM. Six Degrees of Epistasis: Statistical Network Models for GWAS. Front Genet. 2012; 2

62. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011; 473:43–49. [PubMed: 21441907]

63. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337:1190–1195. [PubMed: 22955828]

64. Nica AC, et al. Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. PLoS Genet. 2010; 6:e1000895. [PubMed: 20369022]

65. Nicolae DL, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010; 6:e1000888. [PubMed: 20369019]

66. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. Am J Hum Genet. 2010; 86:6–22. [PubMed: 20074509]

67. Knight J, Barnes MR, Breen G, Weale ME. Using Functional Annotation for the Empirical Determination of Bayes Factors for Genome-Wide Association Study Analysis. PLoS ONE. 2011; 6:e14808. [PubMed: 21556132]

68. Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. Genet Epidemiol. 2007; 31:871–882. [PubMed: 17654612]

69. Chen GK, Witte JS. Enriching the analysis of genomewide association studies with hierarchical modeling. Am J Hum Genet. 2007; 81:397–404. [PubMed: 17668389]

70. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011; 21:1109–1121. [PubMed: 21536720]

71. Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. Genetic Epidemiology. 2011; 35:S12–S17. [PubMed: 22128052]

72. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics. 2010; 11:773–785.

73. Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. PLoS Genet. 2011; 7:e1001316. [PubMed: 21383968]

74. Degner JF, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012; 482:390–394. [PubMed: 22307276]

75. Kasowski M, et al. Variation in Transcription Factor Binding Among Humans. Science. 2010; 328:232–235. [PubMed: 20299548]

76. Majewski J, Pastinen T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends in Genetics. 2011; 27:72–79. [PubMed: 21122937]

77. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]

78. Lappalainen T, Montgomery SB, Nica AC, Dermitzakis ET. Epistatic Selection between Coding and Regulatory Variation in Human Evolution and Disease. Am J Hum Genet. 2011; 89:459–463. [PubMed: 21907014]

79. Kerkel K, et al. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nat Genet. 2008; 40:904–908. [PubMed: 18568024]

80. Prendergast JG, Tong P, Hay DC, Farrington SM, Semple CA. A genome-wide screen in human embryonic stem cells reveals novel sites of allele-specific histone modification associated with known disease loci. Epigenetics & Chromatin. 2012; 5:6. [PubMed: 22607690]

81. McDaniell R, et al. Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans. Science. 2010; 328:235–239. [PubMed: 20299549]

82. Maynard ND, Chen J, Stuart RK, Fan JB, Ren B. Genome-wide mapping of allele-specific protein-DNA interactions in human cells. Nature Methods. 2008; 5:307–309. [PubMed: 18345007]

83. Ge B, et al. Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. Nat Genet. 2009; 41:1216–1222. [PubMed: 19838192]

84. Ng PC, Murray SS, Levy S, Venter JC. An agenda for personalized medicine. Nature. 2009; 461:724–726. [PubMed: 19812653]

85. Patterson N, et al. Methods for high-density admixture mapping of disease genes. Am J Hum Genet. 2004; 74:979–1000. [PubMed: 15088269]

86. Consortium T. 1000 G. P. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]

87. Coop G, et al. The Role of Geography in Human Adaptation. PLoS Genet. 2009; 5:e1000500. [PubMed: 19503611]

88. Hernandez RD, et al. Classic Selective Sweeps Were Rare in Recent Human Evolution. Science. 2011; 331:920–924. [PubMed: 21330547]

89. Sabeti PC, et al. Positive natural selection in the human lineage. Science. 2006; 312:1614–1620. [PubMed: 16778047]

90. Grossman SR, et al. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. Science. 2010; 327:883–886. [PubMed: 20056855]

91. Ott J, Kamatani Y, Lathrop M. Family-based designs for genome-wide association studies. Nature Reviews Genetics. 2011; 12:465–474.

92. Minichiello MJ, Durbin R. Mapping trait loci by use of inferred ancestral recombination graphs. Am J Hum Genet. 2006; 79:910–922. [PubMed: 17033967]

93. Wu Y. Association mapping of complex diseases with ancestral recombination graphs: models and efficient algorithms. J Comput Biol. 2008; 15:667–684. [PubMed: 18651799]

94. Asthana S, et al. Widely Distributed Noncoding Purifying Selection in the Human Genome. PNAS. 2007; 104:12410–12415. [PubMed: 17640883]

95. Ward LD, Kellis M. Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. Science. 201210.1126/science.1225057

96. Hill WG, Goddard ME, Visscher PM. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. PLoS Genet. 2008; 4:e1000008. [PubMed: 18454194]

97. Shao H, et al. Genetic Architecture of Complex Traits: Large Phenotypic Effects and Pervasive Epistasis. PNAS. 2008; 105:19910–19914. [PubMed: 19066216]

98. Zuk O, Hechter E, Sunyaev SR, Lander ES. The Mystery of Missing Heritability: Genetic Interactions Create Phantom Heritability. PNAS. 201210.1073/pnas.1119675109

99. Costanzo M, et al. The Genetic Landscape of a Cell. Science. 2010; 327:425–431. [PubMed: 20093466]

100. Cordell HJ. Detecting gene|[ndash]|gene interactions that underlie human diseases. Nature Reviews Genetics. 2009; 10:392–404.

101. Musani SK, et al. Detection of gene x gene interactions in genome-wide association studies of human population data. Hum Hered. 2007; 63:67–84. [PubMed: 17283436]

102. Lou XY, et al. A Combinatorial Approach to Detecting Gene-Gene and Gene-Environment Interactions in Family Studies. Am J Hum Genet. 2008; 83:457–467. [PubMed: 18834969]

103. Allen HL, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–838. [PubMed: 20881960]

104. Emily M, Mailund T, Hein J, Schauser L, Schierup MH. Using biological networks to search for interacting loci in genome-wide association studies. Eur J Hum Genet. 2009; 17:1231–1240. [PubMed: 19277065]

105. Mechanic LE, Luke BT, Goodman JE, Chanock SJ, Harris CC. Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions. BMC Bioinformatics. 2008; 9:146. [PubMed: 18325117]

106. Pattin KA, Moore JH. Exploiting the Proteome to Improve the Genome-Wide Genetic Analysis of Epistasis in Common Human Diseases. Hum Genet. 2008; 124:19–29. [PubMed: 18551320]

107. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002; 295:1306–1311. [PubMed: 11847345]

108. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289–293. [PubMed: 19815776]

109. Fullwood MJ, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009; 462:58–64. [PubMed: 19890323]

110. Cheng C, et al. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. PLoS Comput Biol. 2011; 7:e1002190. [PubMed: 22125477]

111. Zhu J, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature Genetics. 2008; 40:854–861. [PubMed: 18552845]

112. Barretina J, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483:603–607. [PubMed: 22460905]

113. Burke MK, et al. Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature. 2010; 467:587–590. [PubMed: 20844486]

114. Gresham D, et al. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. PLoS Genet. 2008; 4:e1000303. [PubMed: 19079573]

115. Perlstein EO, Ruderfer DM, Roberts DC, Schreiber SL, Kruglyak L. Genetic basis of individual differences in the response to small-molecule drugs in yeast. Nature Genetics. 2007; 39:496–502. [PubMed: 17334364]

116. Quackenbush J. Microarray analysis and tumor classification. N Engl J Med. 2006; 354:2463–2472. [PubMed: 16760446]

117. Liberzon A, et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011; 27:1739–1740. [PubMed: 21546393]

118. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nature Reviews Genetics. 2011; 12:529–541.

119. Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. Nature. 2010; 465:721–727. [PubMed: 20535201]

120. Chen LS, Emmert-Streib F, Storey JD. Harnessing naturally randomized transcription to infer regulatory relationships among genes. Genome Biol. 2007; 8:R219. [PubMed: 17931418]

121. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med. 2008; 27:1133–1163. [PubMed: 17886233]

122. Voight BF, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. The Lancet. 10.1016/S0140-6736(12)60312-2

123. Chen R, et al. Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes. Cell. 2012; 148:1293–1307. [PubMed: 22424236]

124. Asking for more. Nature Genetics. 2012; 44:733–733. [PubMed: 22735581]

125. Homer N, et al. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. PLoS Genet. 2008; 4:e1000167. [PubMed: 18769715]

126. Salathé M, et al. Digital epidemiology. PLoS Comput Biol. 2012; 8:e1002616. [PubMed: 22844241]

127. Brownstein JS, Sordo M, Kohane IS, Mandl KD. The Tell-Tale Heart: Population-Based Surveillance Reveals an Association of Rofecoxib and Celecoxib with Myocardial Infarction. PLoS ONE. 2007; 2

128. Roque FS, et al. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. PLoS Comput Biol. 2011; 7:e1002141. [PubMed: 21901084]

129. Wilke RA, et al. The emerging role of electronic medical records in pharmacogenomics. Clin Pharmacol Ther. 2011; 89:379–386. [PubMed: 21248726]

130. Kraft P, Hunter DJ. Genetic Risk Prediction — Are We There Yet? New England Journal of Medicine. 2009; 360:1701–1703. [PubMed: 19369656]

131. Yngvadottir B, MacArthur DG, Jin H, Tyler-Smith C. The promise and reality of personal genomics. Genome Biol. 2009; 10:237. [PubMed: 19723346]

132. Roberts NJ, et al. The Predictive Capacity of Personal Genome Sequencing. Sci Transl Med. 2012; 4:133ra58–133ra58.

133. Jostins L, Barrett JC. Genetic risk prediction in complex disease. Hum Mol Genet. 2011; 20:R182–188. [PubMed: 21873261]

134. Stahl EA, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nature Genetics. 2012; 44:483–489. [PubMed: 22446960]

135. Purcell SM, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460:748–752. [PubMed: 19571811]

136. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature Reviews Genetics. 2011; 12:628–640.

137. Gibson G. Rare and common variants: twenty arguments. Nature Reviews Genetics. 2012; 13:135–145.

138. Goldstein DB. The Importance of Synthetic Associations Will Only Be Resolved Empirically. PLoS Biol. 2011; 9

139. Yang J, et al. Common SNPs explain a large proportion of heritability for human height. Nat Genet. 2010; 42:565–569. [PubMed: 20562875]

140. Nebert DW, Zhang G, Vesell ES. From Human Genetics and Genomics to Pharmacogenetics and Pharmacogenomics: Past Lessons, Future Directions. Drug Metab Rev. 2008; 40:187–224. [PubMed: 18464043]

141. Garrod AE, Harris H. Inborn errors of metabolism. 1909

142. Woo SL, Lidsky AS, Güttler F, Chandra T, Robson KJ. Cloned human phenylalanine hydroxylase gene allows prenatal diagnosis and carrier detection of classical phenylketonuria. Nature. 1983; 306:151–155. [PubMed: 6316140]

143. Riordan JR, et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science. 1989; 245:1066–1073. [PubMed: 2475911]

144. Audrézet M, et al. Genomic rearrangements in the CFTR gene: Extensive allelic heterogeneity and diverse mutational mechanisms. Human Mutation. 2004; 23:343–357. [PubMed: 15024729]

145. Zschocke J. Phenylketonuria mutations in Europe. Human Mutation. 2003; 21:345–356. [PubMed: 12655544]

146. Amiel J, et al. Hirschsprung Disease, Associated Syndromes and Genetics: A Review. J Med Genet. 2008; 45:1–14. [PubMed: 17965226]

147. Nica AC, et al. The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. PLoS Genet. 2011; 7:e1002003. [PubMed: 21304890]

148. King JL, Jukes TH. Non-Darwinian Evolution. Science. 1969; 164:788–798. [PubMed: 5767777]

149. Kimura M. Evolutionary rate at the molecular level. Nature. 1968; 217:624. [PubMed: 5637732]

150. Ohno S. So much ‘junk’ DNA in our genome. Brookhaven symposia in biology. 1972; 23:366–370. [PubMed: 5065367]

151. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458:719–724. [PubMed: 19360079]

152. Korn JM, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature Genetics. 2008; 40:1253–1260. [PubMed: 18776909]

153. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–1303. [PubMed: 20644199]

154. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010; 34:816–834. [PubMed: 21058334]

155. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics. 2007; 39:906–913. [PubMed: 17572673]

156. Servin B, Stephens M. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. PLoS Genet. 2007; 3:e114. [PubMed: 17676998]

157. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]

158. Purcell S, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]

159. Veyrieras J-B, et al. High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. PLoS Genet. 2008; 4

160. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012; 28:1353–1358. [PubMed: 22492648]

161. Rozowsky J, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011; 7:522. [PubMed: 21811232]

162. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004; 3:Article 3.

163. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–140. [PubMed: 19910308]

164. Faustino NA, Cooper TA. Pre-mRNA Splicing and Human Disease. Genes Dev. 2003; 17:419–437. [PubMed: 12600935]

165. Cáceres JF, Kornblihtt AR. Alternative splicing: multiple control mechanisms and involvement in human disease. Trends in Genetics. 2002; 18:186–193. [PubMed: 11932019]

166. López-Bigas N, Audit B, Ouzounis C, Parra G, Guigó R. Are splicing mutations the most frequent cause of hereditary disease? FEBS Letters. 2005; 579:1900–1903. [PubMed: 15792793]

167. Barbaux S, et al. Donor splice-site mutations in WT1 are responsible for Frasier syndrome. Nat Genet. 1997; 17:467–470. [PubMed: 9398852]

168. Lorson CL, Hahnen E, Androphy EJ, Wirth B. A Single Nucleotide in the SMN Gene Regulates Splicing and Is Responsible for Spinal Muscular Atrophy. PNAS. 1999; 96:6307–6311. [PubMed: 10339583]

169. Cazzola M, Skoda RC. Translational Pathophysiology: A Novel Molecular Mechanism of Human Disease. Blood. 2000; 95:3280–3288. [PubMed: 10828006]

170. Bisio A, et al. Functional analysis of CDKN2A/p16INK4a 5′-UTR variants predisposing to melanoma. Hum Mol Genet. 2010; 19:1479–1491. [PubMed: 20093296]

171. Abelson JF, et al. Sequence Variants in SLITRK1 Are Associated with Tourette's Syndrome. Science. 2005; 310:317–320. [PubMed: 16224024]

172. Guttman M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009; 458:223–227. [PubMed: 19182780]

173. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell. 2009; 136:629–641. [PubMed: 19239885]

174. Bonafé L, et al. Evolutionary Comparison Provides Evidence for Pathogenicity of RMRP Mutations. PLoS Genet. 2005; 1:e47. [PubMed: 16244706]

175. Cooper TA, Wan L, Dreyfuss G. RNA and Disease. Cell. 2009; 136:777–793. [PubMed: 19239895]

176. Knight JC. Regulatory polymorphisms underlying complex disease traits. Journal of Molecular Medicine. 2004; 83:97–109. [PubMed: 15592805]

177. Martin MP, et al. Genetic Acceleration of AIDS Progression by a Promoter Variant of CCR5. Science. 1998; 282:1907–1911. [PubMed: 9836644]

178. Bream JH, et al. CCR5 Promoter Alleles and Specific DNA Binding Factors. Science. 1999; 284:223–223. [PubMed: 15224670]

179. Bray NJ, et al. Allelic expression of APOE in human brain: effects of epsilon status and promoter haplotypes. Hum Mol Genet. 2004; 13:2885–2892. [PubMed: 15385439]

180. St George-Hyslop PH, Petit A. Molecular biology and genetics of Alzheimer's disease. C R Biol. 2005; 328:119–130. [PubMed: 15770998]

181. Exner M, Minar E, Wagner O, Schillinger M. The role of heme oxygenase-1 promoter polymorphisms in human disease. Free Radical Biology and Medicine. 2004; 37:1097–1104. [PubMed: 15451051]

182. Kleinjan DA, van Heyningen V. Long-Range Control of Gene Expression: Emerging Mechanisms and Disruption in Disease. The American Journal of Human Genetics. 2005; 76:8–32. [PubMed: 15549674]

183. Noonan JP, McCallion AS. Genomics of Long-Range Regulatory Elements. Annual Review of Genomics and Human Genetics. 2010; 11:1–23.

184. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009; 461:199–205. [PubMed: 19741700]

185. Lettice LA, et al. A Long-Range Shh Enhancer Regulates Expression in the Developing Limb and Fin and Is Associated with Preaxial Polydactyly. Hum Mol Genet. 2003; 12:1725–1735. [PubMed: 12837695]

186. Sakabe NJ, Savic D, Nobrega MA. Transcriptional enhancers in development and disease. Genome Biology. 2012; 13:238. [PubMed: 22269347]

187. Pomerantz MM, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nature Genetics. 2009; 41:882–884. [PubMed: 19561607]

188. Tuupanen S, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nature Genetics. 2009; 41:885–890. [PubMed: 19561604]

189. Wasserman NF, Aneas I, Nobrega MA. An 8q24 Gene Desert Variant Associated with Prostate Cancer Risk Confers Differential in Vivo Activity to a MYC Enhancer. Genome Res. 2010; 20:1191–1197. [PubMed: 20627891]

190. Duan J, et al. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum Mol Genet. 2003; 12:205–216. [PubMed: 12554675]

191. SeattleSeq Annotation. at <http://snp.gs.washington.edu/SeattleSeqAnnotation/>

192. Burgner D, et al. A Genome-Wide Association Study Identifies Novel and Functionally Related Susceptibility Loci for Kawasaki Disease. PLoS Genet. 2009; 5:e1000319. [PubMed: 19132087]

193. Emilsson V, et al. Genetics of gene expression and its effect on disease. Nature. 2008; 452:423–428. [PubMed: 18344981]

194. Segrè AV, Groop L, Mootha VK, Daly MJ, Altshuler D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. 2010; 6

195. Raychaudhuri S, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 2009; 5:e1000534. [PubMed: 19557189]

196. 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. 2010; 19:3482–3488. [PubMed: 20601676]

197. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature Biotechnology. 2010; 28:817–825.

198. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012; 22:1748–1759. [PubMed: 22955986]

199. John S, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nature Genetics. 2011; 43:264–268. [PubMed: 21258342]

200. Cowper-Sal-lari R, et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nat Genet. 201210.1038/ng.2416

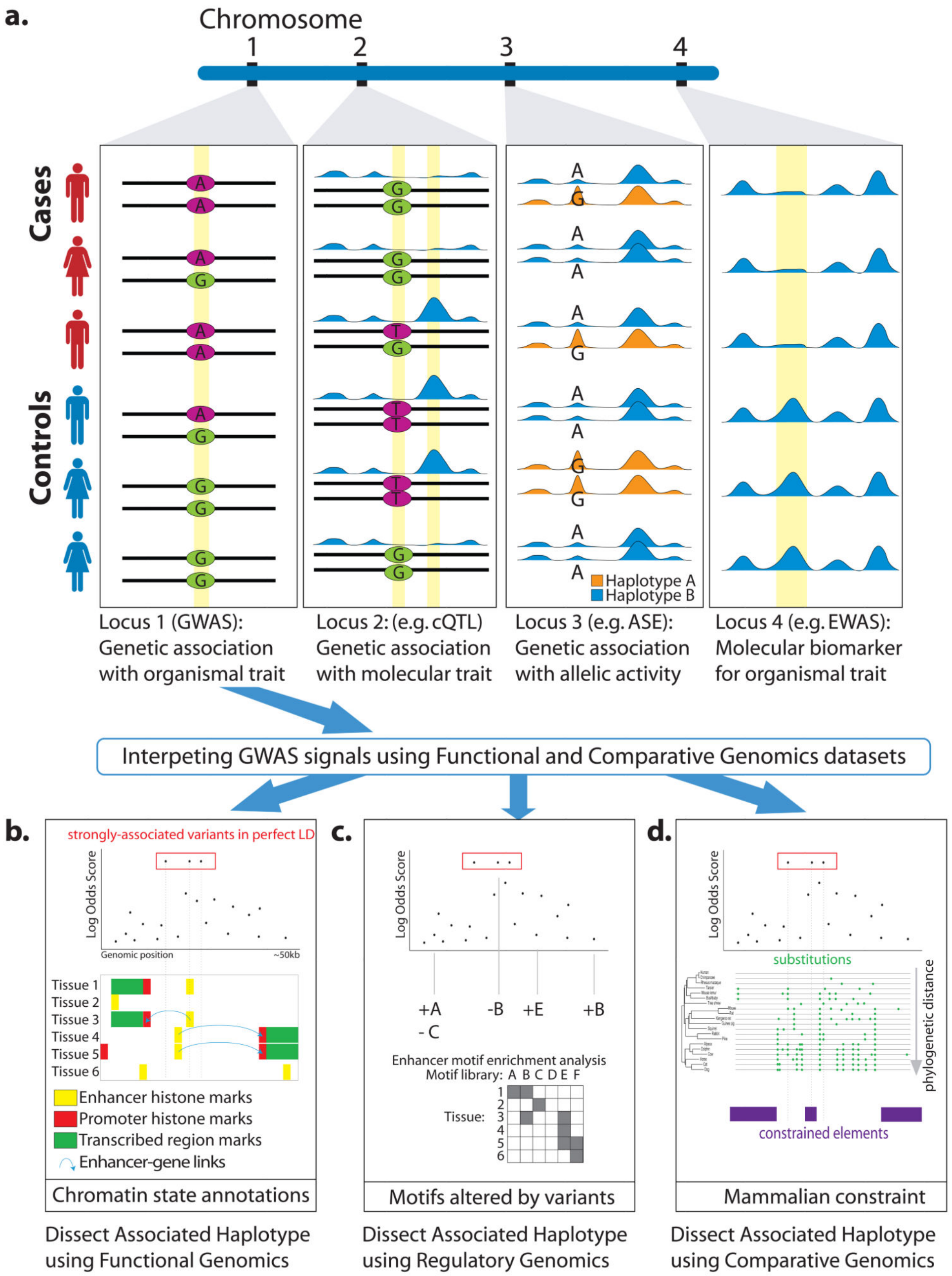
a.
Chromosome
1
2
3
4
Cases
Controls
A
G
T
Haplotype A
Haplotype B
Locus 1 (GWAS): Genetic association with organismal trait
Locus 2: (e.g. cQTL) Genetic association with molecular trait
Locus 3 (e.g. ASE): Genetic association with allelic activity
Locus 4 (e.g. EWAS): Molecular biomarker for organismal trait
Interpeting GWAS signals using Functional and Comparative Genomics datasets
b.
strongly-associated variants in perfect LD
Log Odds Score
Genomic position
~50kb
Tissue 1
Tissue 2
Tissue 3
Tissue 4
Tissue 5
Tissue 6
Enhancer histone marks
Promoter histone marks
Transcribed region marks
Enhancer-gene links
Chromatin state annotations
Dissect Associated Haplotype using Functional Genomics
c.
Log Odds Score
+A
- C
-B
+E
+B
Enhancer motif enrichment analysis
Motif library: A B C D E F
Tissue: 1 2 3 4 5 6
Motifs altered by variants
Dissect Associated Haplotype using Regulatory Genomics
d.
Log Odds Score
substitutions
phylogenetic distance
constrained elements
Mammalian constraint
Dissect Associated Haplotype using Comparative Genomics

Figure 1. Four types of next-generation association tests

(a) Genetic association with organismal traits is performed in genome-wide association studies (GWAS); at the locus shown, the G 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 genomewide 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 co-vary with presence of the T allele; this allele may influence a *cis*-regulatory motif of chromatin. (c) Heterozygous sites in individual cells can be used to interrogate allele-specific effects;

unlike molecular QTLs discovered across individuals, these studies control for variation in *trans* genetic background. In this example, the G allele is not only associated with the presence of a TF binding peak at that locus, but in heterozygous individuals is over-represented in ChIP-seq reads originating from that locus, suggesting that the TF binds specifically to the G allele. (d) Functional genomics data can be directly compared between cases and controls to discover biomarkers for disease, without necessarily attributing genetic causes to these molecular changes. Indeed, these biomarkers may be caused by *trans* genetic factors, environmental factors, or by the disease itself.

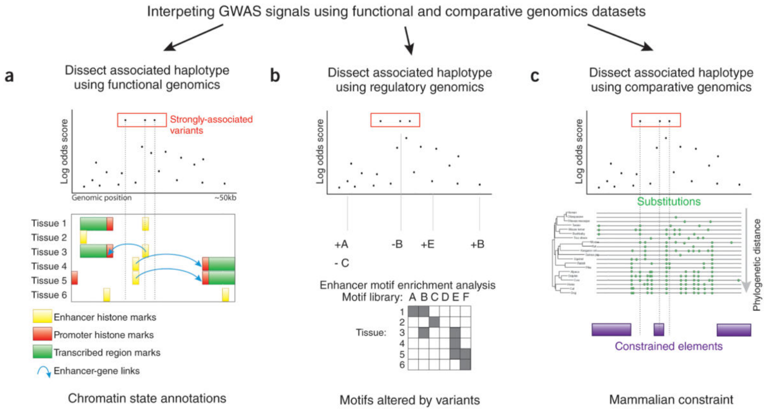
Interpeting GWAS signals using functional and comparative genomics datasets
a
Dissect associated haplotype using functional genomics
Log odds score
Strongly-associated variants
Genomic position
~50kb
Tissue 1
Tissue 2
Tissue 3
Tissue 4
Tissue 5
Tissue 6
Enhancer histone marks
Promoter histone marks
Transcribed region marks
Enhancer-gene links
Chromatin state annotations
b
Dissect associated haplotype using regulatory genomics
Log odds score
+A
- C
-B
+E
+B
Enhancer motif enrichment analysis
Motif library: A B C D E F
Tissue:
1
2
3
4
5
6
Motifs altered by variants
c
Dissect associated haplotype using comparative genomics
Log odds score
Substitutions
Phylogenetic distance
Constrained elements
Mammalian constraint

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/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 Tissue 1 and Tissue 3. (c) Comparative genomics identifies regions of evolutionary constraint in non-coding sequence. Here, sequence surrounding only the middle SNP is constrained across mammals.

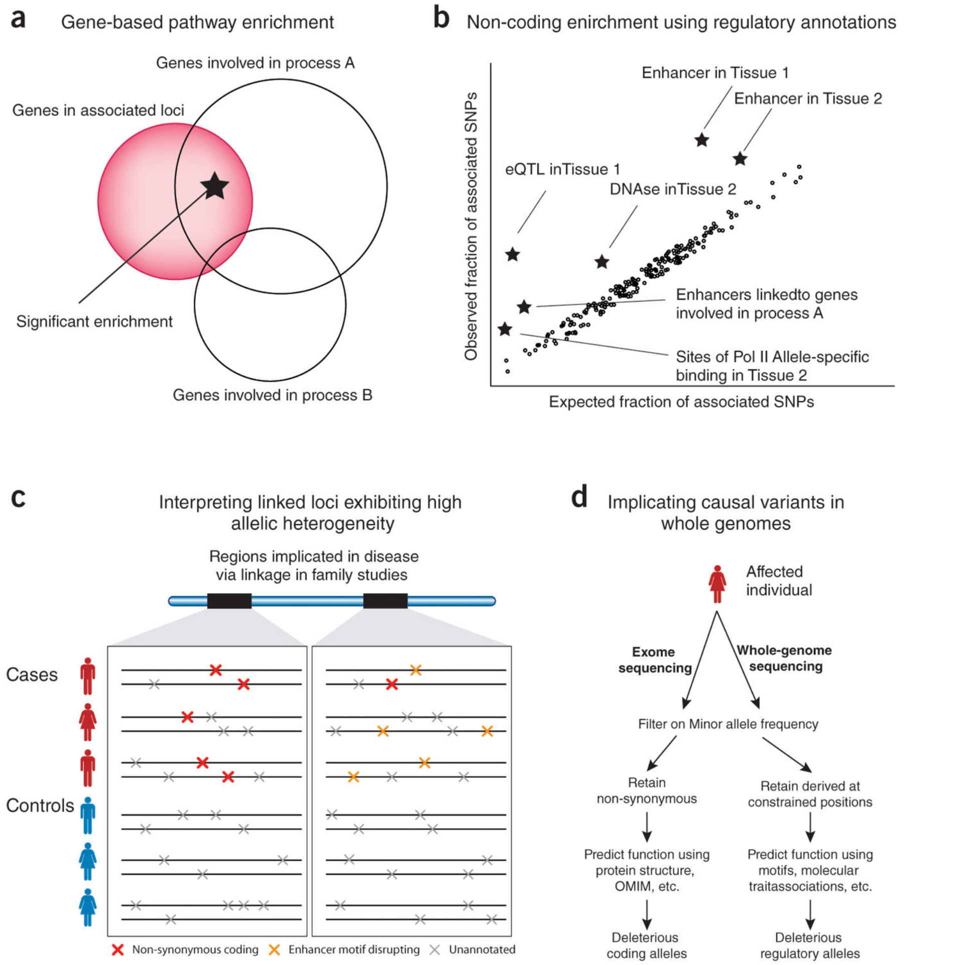
a
Gene-based pathway enrichment
Genes involved in process A
Genes in associated loci
Significant enrichment
Genes involved in process B
b
Non-coding enirchment using regulatory annotations
Observed fraction of associated SNPs
Enhancer in Tissue 1
Enhancer in Tissue 2
eQTL inTissue 1
DNAse inTissue 2
Enhancers linkedto genes involved in process A
Sites of Pol II Allele-specific binding in Tissue 2
Expected fraction of associated SNPs
c
Interpreting linked loci exhibiting high allelic heterogeneity
Regions implicated in disease via linkage in family studies
Cases
Controls
Non-synonymous coding
Enhancer motif disrupting
Unannotated
d
Implicating causal variants in whole genomes
Affected individual
Exome sequencing
Whole-genome sequencing
Filter on Minor allele frequency
Retain non-synonymous
Retain derived at constrained positions
Predict function using protein structure, OMIM, etc.
Predict function using motifs, molecular traitassociations, etc.
Deleterious coding alleles
Deleterious regulatory alleles

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) Non-coding 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 where 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 will only be discovered through classical linkage analysis. These regions can now be interrogated through WES/WGS, and an imbalanced burden of putatively deleterious alleles can be observed in cases (as in the left example). With regulatory annotations, these burden tests can now be extended to non-coding 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-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 such as 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.

Table 1

The diversity of genetic architectures underlying human phenotypes.

Architecture	**Notes**	**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[140,141]. 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[140,142,143], 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[3,142,143].	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[144], even though thousands of alleles are known. In contrast, phenylketonuria is extremely heterogeneous, with different *PAH* alleles predominating among affected individuals in different populations[145]. 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 (“oligogenetic”)	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 genes are involved, including the tyrosine kinase receptor *RET* and the gene *GDNF* which encodes its ligand[146]. Interestingly, the most common variant in the main susceptibility gene *RET* is non-coding, a single-nucleotide polymorphism (SNP) in an enhancer. Both coding and non-coding variants are involved typically in one or a small number of well-defined pathways.	Oligogenetic 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 non-coding 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 on 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[15,103,139], 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[135,139]	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].	QTL 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[147], 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[148–150].	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[151].	The interpretation of private and somatic variations (Fig 3d) will also benefit

Architecture	Notes	Role of computational and regulatory genomics
		tremendously from a systematic regulatory annotation, as they likely exploit existing regulatory pathways, even though they are subject to cellular, rather than organismal selective pressures.

Table 2

Computational tools for association analyses.

Class of analysis	Tool	Notes
Genome-wide association between genotype and phenotype (GWAS)	SNPTEST[155]	Incorporates imputation
	Bim-Bam[156]	Bayesian regression approach combining imputation and association probabilities
	EIGENSTRAT[157]	Models ancestry differences between cases and controls using principal components analysis
	PLINK[158]	Large package including tools to impute, control for population stratification, and hybrid methods such as family-based association and population-based linkage
Local association between genotype and molecular trait (e.g., eQTL)	eQTNMiner[159]	Tests a Bayesian hierarchical model incorporating priors based on TSS distance
	Matrix eQTL[160]	Fast association testing of continuous or categorical genotype values with expression
Allele-specific expression and binding	ChIP-SNP[82]	For ChIP-chip data
	AlleleSeq[161]	For ChIP-seq and RNA-seq data
Genome-wide association between molecular trait and phenotype (e.g., differential expression, EWAS)	limma[162]	For expression microarray data
	edgeR[163]	For RNA-seq data

Note: analyses using genotype information require tools to call variants, such as BirdSeed[152] on array data or GATK[153] on sequencing data, and tools to impute genotypes, such as MaCH[154].

Table 3

Mechanisms through which non-coding variants influence human disease.

Non-coding element disrupted	Molecular function and effect of mutations.	Disease association
Splice-junction and splicing-enhancer	Splicing is constitutive for some transcripts and highly tissue-specific for others, relying on both canonical sequences at the exon-intron junction as well as 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[164,165]. A recent analysis suggests that the majority of disease-causing point mutations in OMIM may exert their effects through splicing[166]. Alternative splice site variants in the *WT1* gene are involved in Frasier Syndrome (FS)[167] Skipping of exon 7 of the *SMN* gene is involved in spinal muscular atrophy (SMA)[168]
Sequences regulating translation, stability, and localization	Sequences in the 5′-untranslated regions (UTRs) of mRNAs can influence translation regulation, such as upstream ORFs, premature AUG or AUC codons, and palindromic sequences that form inhibitory stem loops[169]. 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[170]. 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[171].
Genes encoding trans-regulatory RNA	Non-coding RNAs participate in a panoply of regulatory functions, ranging from the well-understood transfer and ribosomal RNA to the recently-discovered long non-coding RNAs[172,173].	Both rare and common mutations in the gene *RMRP* encoding an RNA component of the mitochondrial RNA processing ribonuclease have been associated with cartilage-hair hypoplasia[174] Non-coding RNA mutations can cause many other diseases[175].
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[176].	Mutations in the promoter of the HIV1-progression associated gene *CCR5*, are correlated with expression of the receptor it encodes and bind differentially to at least three transcription factors[177,178] *APOE* promoter mutations are associated with Alzheimer's disease[179,180] Heme oxygenase-1 (*HO-1*) promoter mutations lead to expression changes and are associated with many diseases[181]
Enhancer	Enhancers are distal regulatory elements that often lie 10,000 to 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. While 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 GWAS by many Mendelian disorders for which some patients had translocations or other structural variants far from the promoter[182–184]. 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* [185]; this distal locus acted as an enhancer of

Non-coding element disrupted	Molecular function and effect of mutations.	Disease association
		Shh and recapitulated the polydactyly phenotype in mouse.
		A number of GWAS hits have been validated as functional enhancers[186]; 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*[187–189].
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 non-coding 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[190].

Table 4

Comparison of recent tools to systematically annotate variants

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

Tool	Type	Input method	Protein annotation	Regulatory annotation	Other
SeattleSeq[191]	server	variants	deleteriousness scores	conservation scores	dbSNP clinical association data
ANNOVAR[57]	software	variants, regions	User-defined: user downloads desired variation, conservation, coding and non-coding 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
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		Histone modification, protein binding, DNase, conservation, reguatory motif alteration scores	eQTL, reporter assays, combined score analysis per variant

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 implicated in Kawasaki disease[192]	Ingenuity Pathway Analysis (closed-source)
	Genes differentially expressed in adipose overlap with genetic associations with obesity[193]	Microarray analysis of differential expression
	TGF-β pathway, Hedgehog signaling pathway are enriched among height GWAS loci[103]	GSEA using MAGENTA[194], network from text-mining using GRAIL[195], known disease genes from OMIM[4], eQTL enrichment
Concordance with eQTL results	eQTL prioritization during replication facilitated validation of two Crohn's disease susceptibility loci[196]	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[197] (M.K. and colleagues); enrichment analysis
TF binding site and DNase hypersensitivity enrichment	Many GWAS show enrichment for ENCODE-annotated DNAse and ChIP sites[198]	Enrichment analysis
	Many GWAS show enrichment for DNAse in biologically-relevant cell types[63]	Hotspot algorithm to define discrete hypersensitive sites[199]; enrichment analysis
	FOXA1 and estrogen receptor binding sites are enriched among breast cancer GWAS loci[200]	Variant Set Enrichment (VSE[200])