APPLICATIONS OF NEXT-GENERATION SEQUENCING

Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data

Gregory M. Cooper and Jay Shendure‡*

Abstract | Genome and exome sequencing yield extensive catalogues of human genetic variation. However, pinpointing the few phenotypically causal variants among the many variants present in human genomes remains a major challenge, particularly for rare and complex traits wherein genetic information alone is often insufficient. Here, we review approaches to estimate the deleteriousness of single nucleotide variants (SNVs), which can be used to prioritize disease-causal variants. We describe recent advances in comparative and functional genomics that enable systematic annotation of both coding and non-coding variants. Application and optimization of these methods will be essential to find the genetic answers that sequencing promises to hide in plain sight.

Private
A genetic variant that is confined to a single individual, family or population.

Prior probability
Otherwise simply known as the 'prior', this is the probability of a hypothesis (or parameter value) without reference to the available data. Priors can be derived from first principles or be based on general knowledge or previous experiments.

**HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA.*
‡Department of Genome Sciences, University of Washington, Seattle, Washington 98115, USA.
e-mails: gcooper@hudsonalpha.org; shendure@u.washington.edu
doi:10.1038/nrg3046

The precise delineation of causal variants that alter human phenotypes, particularly diseases, is a fundamental goal of human genetics, providing crucial insights into the biology connecting genotype and phenotype and potentially facilitating the prediction of disease onset. Recently, 'next-generation' DNA sequencing[1] has provided a means to define nearly comprehensive maps of genetic variation, including the several million single nucleotide variants (SNVs), hundreds of thousands of small insertion or deletion events and thousands of structural variants[2] in typical human genomes. Most of these are common[3], but individual genomes also contain many thousands of rare and effectively private genetic variants.

Despite new methods to comprehensively catalogue human genetic variation, the identification of variants that are causal for disease or other traits remains a difficult challenge. Genetic approaches (such as linkage analysis and genome-wide association studies (GWASs)) can identify candidate variants but are often insufficiently powered to specifically identify causal variants. For example, GWASs have identified associations between ~1,300 loci and ~200 diseases or traits[4] (see the US National Human Genome Research Institute Catalog of Published Genome-Wide Association Studies). However, owing in part to weak effect sizes and correlations among neighbouring variants, precise identification of causal variants has been achieved for only a handful of these loci, and many legitimate associations remain undiscovered. Even for diseases suspected to result from highly penetrant or Mendelian mutations, locus heterogeneity and limitations in sample or pedigree sizes often limit the discovery power and resolution of purely genetic studies. Now, with the use of exome and genome sequencing in disease genetics, the challenge of whittling down a list of candidate variants to those that are causal becomes particularly important. In fact, the primary roadblock faced by the field is increasingly one of variant interpretation, rather than data acquisition.

The interpretive challenge presented by 'next-generation genetics' is, in fact, a long-standing one in quantitative genetics[5,6] and is often described as a 'multiple hypothesis testing problem' in which nominal 'significance' thresholds (for example, $P < 0.05$) yield many false discoveries when applied to many tests. However, true hypotheses are true, and false hypotheses are false, regardless of how many are tested. As such, the actual 'multiple testing burden' depends on the proportion of true and false hypotheses in any given set: that is, the 'prior probability' that any given hypothesis is true, rather than the number of tests per se. This challenge can thus be viewed as a 'naive hypothesis testing' problem — that is, when in reality only one or a few variants are causal for a given phenotype, but all (or many) variants are *a priori* equally likely candidates, the prior probability of any given variant being causal is miniscule. As a consequence, extremely convincing data

are required to support causality, which is potentially unachievable for some true positives.

Defining the challenge in terms of hypothesis quality rather than quantity, however, points to a solution. Specifically, experimental or computational approaches that provide assessments of variant function can be used to better estimate the prior probability that any given variant is phenotypically important, and these approaches thereby boost discovery power. As a simple illustration of this concept, we and others have developed methods to identify causative genes for Mendelian disorders using exome sequencing[7–10]. Specifically, rare non-synonymous variants are considered candidates, whereas common, synonymous or non-coding variants are ignored. This strategy is effective not because it reduces the number of tests per se, but because rare non-synonymous variants are intrinsically more plausible disease candidates. This method and related ones thus implicitly define non-uniform prior probabilities for candidate variants and use them to better interpret the significance of the observed genetic associations. Importantly, for prioritizing variants that have disease consequences, which typically reduce survival and reproductive success, these priors can be defined on the basis of expected deleteriousness.

In this Review, we survey strategies to estimate the deleteriousness of human genetic variants in order to better identify disease-causal variants when genetic information is insufficient: that is, when the haystack has been cleared away only to reveal a large pile of needles (FIG. 1). We focus on approaches to address two related but separable questions: whether a given variant has a functional effect at the molecular level and, if so, whether that functional alteration is deleterious to the organism. We discuss computational methods that use comparative genomics to predict deleteriousness in both coding and non-coding DNA and that in some cases incorporate knowledge of protein biochemistry and structure in the assessment of protein-coding variants. We also describe complementary experimental approaches for assessing deleteriousness and how such approaches are now being scaled to genome-wide levels. Although we are primarily concerned with applications in human disease research, these approaches can often be applied to other species, especially model organisms with high-quality genome assemblies and related resources. We further note that, although there is a distinction between *a priori* estimation of variant causality and *a posteriori* confirmation that a given candidate variant is truly causal, most of the approaches we describe can be used in both settings, albeit not in the same analysis. Finally, although this Review emphasizes analytical methods for SNVs, it will be crucial to develop methods to estimate the deleteriousness of all classes of variation that have an impact on genomes, ranging from SNVs to small insertion or deletion events to large structural changes.

The diversity of approaches that are relevant to variant interpretation is illustrated in FIG. 2 for the β-haemoglobin (*HBB*) locus, in which some mutations give rise to thalassaemia. We note at the outset that perhaps the greatest challenge for this field is the integration of these various strategies into a unified, quantitative, predictive framework that spans functional categories of variation and incorporates both experimental and computational information. At the end of this Review, we discuss challenges to be addressed before such a framework can be effective.

Computational approaches

Evolution as the best measure of deleteriousness. Most computational methods to estimate deleteriousness exploit the fact that sequences observed among living organisms are those that have not been removed by natural selection. Indeed, if we consider evolution to be the ultimate mutagenesis experiment, comparative sequence analysis is a powerful source of information regarding deleteriousness. In particular, by quantifying evolutionary changes in genes or genomes, conserved positions that have evolved too slowly to be neutral can be identified. These are sites in which past mutations were removed by purifying selection because they were deleterious and are therefore highly likely to be sites where recently occurring or new mutations are also deleterious.

Although the methodological details governing how this concept is exploited vary greatly, two considerations are essential. First, sequence conservation is not a predictor of deleteriousness per se, but rather it is conservation in excess of neutral expectations that is used to infer constraint[11]. This relationship is quantitative and spans a broad range of deleteriousness, manifesting as a continuous spectrum of evolutionary rates across wide ranges of evolutionary time over all domains of life. Second, the 'phylogenetic scope' of the compared sequences has substantial effects on the use and interpretability of these analyses[12–15]. Broad scopes — for example, human to yeast — afford extreme specificity, as neutral divergence is so large that any detectable sequence conservation indicates strong constraint. However, reductions in shared biology result in a loss of sensitivity for all of those functional sequences that have emerged or changed since the most recent common ancestor. Conversely, narrow scopes (for example, among primates) capture larger amounts of shared biology and functional sequence but at the expense of specificity. For example, at the per-nucleotide level, ~98.8% of sites are conserved between human and chimpanzee genomes[16], although only a small minority are under constraint[17–19]; comparisons of numerous primate sequences are required to achieve useful specificity within this scope[20,21].

Constraint-based approaches to annotate deleterious genomic positions assume that such positions will have a detectable history of purifying selection, but there are important drawbacks to this assumption. In particular, functional divergence will lessen the correlation between past constraint and present-day deleteriousness. For example, humans have acquired many genes[22,23] and regulatory elements[24] in recent evolutionary history; such sequences will have fewer, less divergent orthologues and are therefore given lower

Deleterious
A genetic variant that lowers the fitness of an organism: that is, it decreases survival or reproductive success.

Conserved
Shared identity of either protein or nucleotide sequences, which can be indicative of constraint.

Neutral
Sequences that are free to evolve in the absence of natural selection and are therefore subject only to random mutational and genetic drift processes.

Phylogenetic scope
The taxonomic range captured by a given comparative sequence analysis — for example, mammals or eukaryotes.

Constrained
Sequences that are under purifying selection to maintain function, which often, but not always, results in sequence conservation.

scores or are filtered away by constraint-based measures. This can be ameliorated by additional sequencing of closely related species to boost statistical power within narrower phylogenetic scopes (for example, REF. 20) and constitutes a strong argument to generate assemblies for many, if not all, primate species. Additionally, adaptive functionality often results from mutations in previously highly constrained sites[25–27], suggesting that constraint-based measures are effective even for regions with species-specific functional alterations. Nevertheless, the detection of causal variants that result from a gain-of-function mutation within previously non-functional (for example, REF. 28) or rapidly evolving sequences (for example, REF. 23) will be rendered disadvantageous by the evolutionary frameworks discussed here.

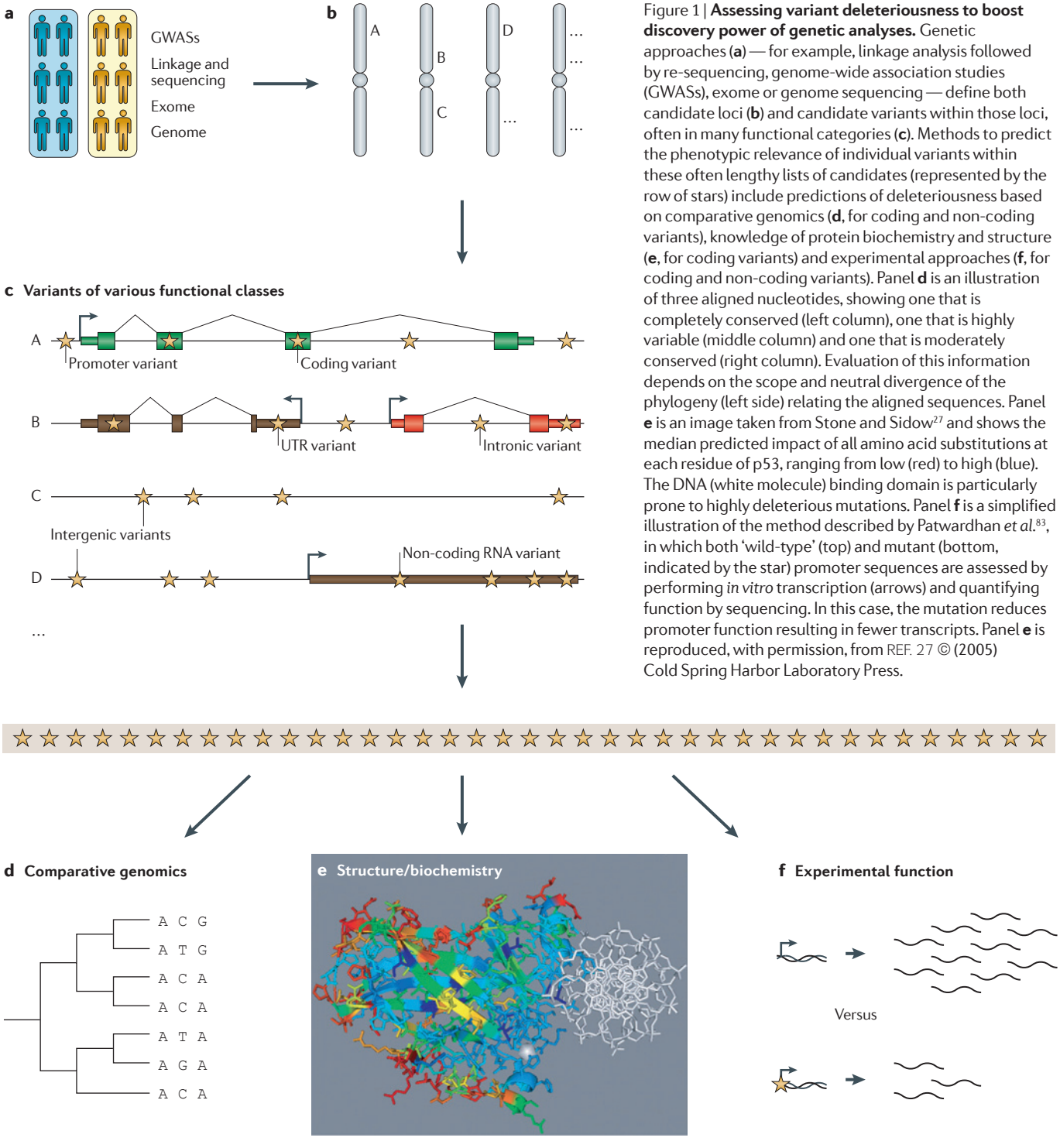

Figure 1 | **Assessing variant deleteriousness to boost discovery power of genetic analyses.** Genetic approaches (**a**) — for example, linkage analysis followed by re-sequencing, genome-wide association studies (GWASs), exome or genome sequencing — define both candidate loci (**b**) and candidate variants within those loci, often in many functional categories (**c**). Methods to predict the phenotypic relevance of individual variants within these often lengthy lists of candidates (represented by the row of stars) include predictions of deleteriousness based on comparative genomics (**d**, for coding and non-coding variants), knowledge of protein biochemistry and structure (**e**, for coding variants) and experimental approaches (**f**, for coding and non-coding variants). Panel **d** is an illustration of three aligned nucleotides, showing one that is completely conserved (left column), one that is highly variable (middle column) and one that is moderately conserved (right column). Evaluation of this information depends on the scope and neutral divergence of the phylogeny (left side) relating the aligned sequences. Panel **e** is an image taken from Stone and Sidow[27] and shows the median predicted impact of all amino acid substitutions at each residue of p53, ranging from low (red) to high (blue). The DNA (white molecule) binding domain is particularly prone to highly deleterious mutations. Panel **f** is a simplified illustration of the method described by Patwardhan *et al.*[83], in which both 'wild-type' (top) and mutant (bottom, indicated by the star) promoter sequences are assessed by performing *in vitro* transcription (arrows) and quantifying function by sequencing. In this case, the mutation reduces promoter function resulting in fewer transcripts. Panel **e** is reproduced, with permission, from REF. 27 © (2005) Cold Spring Harbor Laboratory Press.

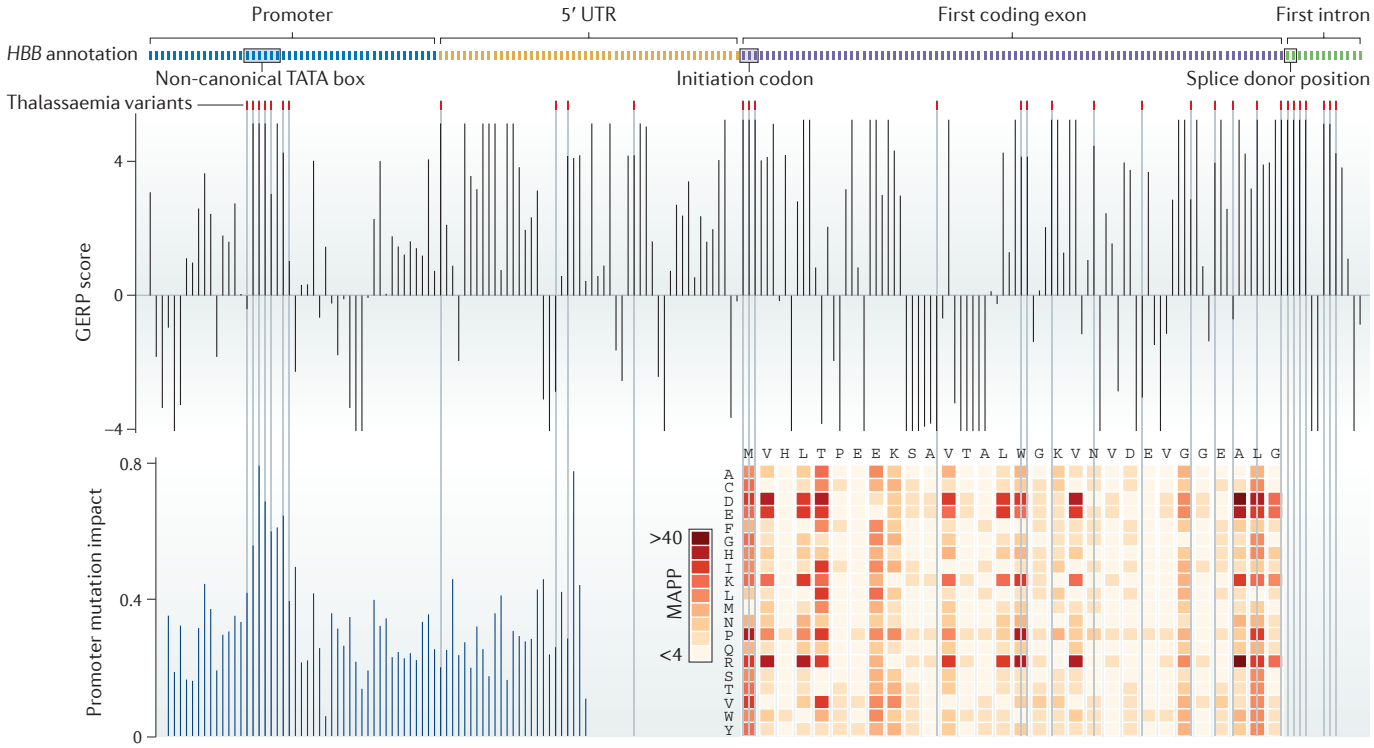

Figure 2 | **Functional and evolutionary annotations highlight disease variation at the *HBB* locus.** Shown is an integrated view of the nucleotides (each coloured bar at the top is one position), disease mutation locations, constraint scores, promoter mutagenesis data and protein-based mutation impact predictions for the proximal promoter, first exon and first intron of β-haemoglobin (*HBB*). Sites known to harbour disease mutations (red vertical bars) are from HbVar[98] and include all β-thalassaemia-associated point substitutions (insertion or deletion variants are not shown); shaded grey lines extend below each mutation site to the bottom of the figure. Beneath the HbVar annotations, Genomic Evolutionary Rate Profiling (GERP)[67] scores — measured as 'rejected substitutions' (higher scores indicate higher conservation) — are plotted as vertical bars ranging from −4 (values below this are capped) to 5.2. At the bottom left, the average absolute values of the mutational impact (over all non-reference mutations) as measured by Patwardhan *et al.*[83] are shown as vertical blue bars for each assayed position. Note that disease mutations cluster within the non-canonical TATA box, which shows the highest degree of promoter impairment and also high constraint scores. At the bottom right, predictions made by Multivariate Analysis of Protein Polymorphism (MAPP)[27] of the effects of all possible amino acid replacements at each codon are shown, binned into nine separate colours ranging from little to no predicted impact (light beige shading) to very high impact (dark red shading). The β-globin protein sequence encoded by *HBB* is plotted along the top of the MAPP prediction matrix. Note the cluster of disease mutations that disrupt the initiation codon and have large MAPP-predicted effects and high constraint scores. Also highlighted is the canonical splice donor position immediately downstream of the first exon, which also exhibits both disease mutations and high GERP scores.

Predicting the effects of protein-coding sequence changes. Methods for predicting the deleteriousness of protein-altering variants are the richest and most detailed of the available approaches, capable of leveraging both evolutionary and biochemical information. Nonsense and frameshift mutations are the most obvious candidates, as they are predicted to result in a loss of protein function and are heavily enriched among disease-causal variation (for example, see REFS 29,30). However, this class of variation is not unambiguously deleterious, in some cases allowing functional protein production or resulting in a loss of protein that is apparently not harmful[31]. Considering non-synonymous variants, the simplest and earliest approaches to estimate deleteriousness use discrete biochemical categorizations such as 'radical' versus 'conservative' amino acid changes[32]. However, there are now numerous more sophisticated approaches to classify non-synonymous variants on both quantitative and discrete scales (TABLE 1). Although detailed summaries of each available method are outside the scope of this Review (but see REF. 33), three general considerations are important.

First, it is important to differentiate first-principles approaches from trained classifiers. First-principles approaches explicitly define a biological property (typically evolutionary) of deleterious variants and make predictions on the basis of similarity or dissimilarity to that definition. By contrast, trained classifiers generate prediction rules by identifying heuristic combinations of many potentially relevant properties that optimally differentiate a set of true positives and negatives. First-principles approaches have the advantage of greater interpretability; for example, radical and conservative annotations of amino acid substitutions

Table 1 | **Tools for protein-sequence-based prediction of deleteriousness**

Name	Type	Information	URL	Refs
MAPP	Constraint-based predictor	Evolutionary and biochemical	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	27
SIFT	Constraint-based predictor	Evolutionary and biochemical (indirect)	http://sift.bii.a-star.edu.sg/	39
PANTHER	Constraint-based predictor	Evolutionary and biochemical (indirect)	http://www.pantherdb.org/	41
MutationTaster*	Trained classifier	Evolutionary, biochemical and structural	http://www.mutationtaster.org/	40
nsSNP Analyzer	Trained classifier	Evolutionary, biochemical and structural	http://snpanalyzer.uthsc.edu/	44
PMUT	Trained classifier	Evolutionary, biochemical and structural	http://mmb2.pcb.ub.es:8080/PMut/	38
polyPhen	Trained classifier	Evolutionary, biochemical and structural	http://genetics.bwh.harvard.edu/pph2/	35
SAPRED	Trained classifier	Evolutionary, biochemical and structural	http://sapred.cbi.pku.edu.cn/	42
SNAP	Trained classifier	Evolutionary, biochemical and structural	http://www.rostlab.org/services/SNAP/	36
SNPs3D	Trained classifier	Evolutionary, biochemical and structural	http://www.snps3d.org/	51
PhD-SNP	Trained classifier	Evolutionary and biochemical (indirect)	http://gpcr2.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP_Help.html	37

*Also makes predictions for synonymous and non-coding variant effects: for example, splicing. MAPP, Multivariate Analysis of Protein Polymorphism; polyPhen, polymorphism phenotyping.

have a straightforward biochemical interpretation. Additionally, these approaches will not be misled by 'gold-standard' data sets that are contaminated with erroneous annotations and/or are not representative of the general population of true positives and negatives[34]. However, first-principles methods are only as good as the assumption (or assumptions) that they make and do not model all of the relevant factors. Conversely, a trained classifier approach effectively yields a 'black-box' prediction and will be prone to the biases and errors of the gold-standard data. However, these approaches have the advantage of being specifically tunable to the desired task (that is, predicting disease causality) and are capable of incorporating many sources of information without requiring a detailed understanding of how that information is relevant.

Second, nearly all of these methods use alignments of homologous proteins to estimate mutational deleteriousness[27,35–46]. These strategies generally exploit both the level of sequence conservation (that is, sites with fewer observed substitutions are inferred to be under tighter constraints and will have more deleterious effects when mutated) and the patterns of observed substitutions (that is, sites that are observed to tolerate a subset of amino acids are likely to be more deleteriously affected by amino acids outside the subset). In either case, the phylogenetic scope and quality of the alignment are essential but often overlooked factors. Alignments with lower diversity offer less power, as both compatible and incompatible substitutions may be absent owing to insufficient divergence. Conversely, inclusion of more distant sequences may suggest a wider range of diversity than is actually tolerated. For example, assessments of mutations on lactose operon inhibitor (LacI) function are less accurate when based on alignments that include paralogues relative to alignments restricted to orthologues, despite the fact that the former capture greater sequence diversity[27]. This observation is likely to reflect the relaxed constraint and increased potential for acquiring novel functions of duplicated genes[47]. Interestingly, a recent analysis of mutational effects on human methylenetetrahydrofolate reductase activity showed that inclusion of even distant orthologues in a sequence alignment can reduce power[48]. This points to the impact of orthologous protein functional divergence over long evolutionary periods and/or the accumulation of deleterious mutations that are offset by compensatory changes[43].

Third, most protein-sequence-based methods also exploit biochemical data, including amino acid properties (such as charge), sequence information (such as the presence of a binding site) and structural information (such as the presence of a β-sheet). The integration of these data with comparative sequence analysis can significantly improve predictions of deleteriousness[49–53]. As one example, consider the Multivariate Analysis of Protein Polymorphism (MAPP) method[27], which quantifies constraint in terms of biochemical properties (such as hydropathy and polarity) rather than substitutions per se. Mutational impact is estimated by measuring the distance between the properties of the new amino acid and the phylogenetically weighted average properties of the aligned residues, and these distances are normalized to the observed evolutionary variability for each property. For example, if a given position exhibits many

Table 2 | **Tools for nucleotide-sequence-based prediction of deleteriousness**

Name	Type	Information	URL	Refs
phastCons	Phylogenetic HMM	Evolutionary	http://compgen.bscb.cornell.edu/phast/	60
GERP	Single-site scoring	Evolutionary	http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html	67
Gumby	Single-site scoring	Evolutionary	http://pga.jgi-psf.org/gumby/	21
phyloP	Single-site scoring	Evolutionary	http://compgen.bscb.cornell.edu/phast/	66
SCONE	Single-site scoring	Evolutionary	http://genetics.bwh.harvard.edu/scone/	68
binCons	Sliding-window scoring	Evolutionary	http://zoo.nhgri.nih.gov/binCons/index.cgi	69
Chai Cons	Sliding-window scoring	Evolutionary and structural	http://research.nhgri.nih.gov/software/chai	71
VISTA	Visualization tool (various scores)	Evolutionary	http://genome.lbl.gov/vista/index.shtml	70

GERP, Genomic Evolutionary Rate Profiling; HMM, hidden Markov model; SCONE, Sequence Conservation Evaluation.

substitutions but all of the observed amino acids are small, then MAPP would predict that a mutation would be tolerated if the new amino acid were small but not tolerated if it were large. Such predictions are at least semi-quantitative: the magnitude of the biochemical deviation between the evolutionarily observed and mutant amino acids correlates with the extent of functional impairment (for LacI mutants) and disease severity (for anaemia-causing *HBB* mutants)[27]. In this way, both the rates and biochemical properties of changes that are observed (or not observed) throughout evolution are informative.

The case for non-coding variation analysis. Published analyses that predict deleteriousness of genetic variation within individual genomes have largely focused on protein-altering variants, as these are the most amenable to functional interpretation. However, non-coding variants constitute the overwhelming majority of human genetic variation[3]. Furthermore, collective GWAS evidence shows that ~88% of trait-associated variants of weak effect are non-coding[54]. Although only a few are defined at the molecular level (for example, REF. 55), this supports the hypothesis that most weak-effect causal variants are non-coding. This hypothesis is further supported by an abundance of heritable factors altering gene expression[56] that are enriched within trait-associated loci[57] and the observation that regulatory variation is a major driver of morphological differences between closely related species[58,59]. Additionally, evolutionary analyses demonstrate that approximately fivefold more non-coding positions exist than coding positions in human genomes that have been subject to purifying selection[17,18,60]. Finally, although likely to explain a proportionally larger fraction of more severe phenotypes[29], estimates of the fraction of Mendelian diseases caused by protein alterations are skewed upwards by ascertainment bias and an absence of failed protein-centric disease studies in the publication record. Conversely, the existence of numerous large-effect regulatory variants[61–63] — despite the increased difficulties associated with their discovery — confirms at least that a substantial fraction of mutations underlying Mendelian diseases is non-coding. Analytical restrictions to coding mutations are therefore untenable in the long term.

Nucleotide-sequence-based predictions in non-coding and coding DNA. As for protein-altering variants, comparative genomics is a central component in deleteriousness prediction for non-coding variants. However, the phylogenetic scope for non-coding sequence comparisons is typically narrower than that for proteins. Whereas many human proteins have homologues in bacteria and most have homologues in vertebrates, only a small fraction of non-coding bases in human genomes align to fish genomes[64], and there is no detectable conservation outside vertebrates[65]. Consequently, the primary scope used for human non-coding sequence analysis is mammalian, and the first global insights regarding non-coding deleteriousness emerged from comparisons with the mouse genome. These efforts estimated that ~5% of positions in human genomes, a minority of which are coding, exhibit evidence of purifying selection throughout mammalian evolution[18]. Subsequently, sequencing of dozens of mammalian genomes has refined this estimate upwards to 7–8%[17,60] and has begun to identify constraints on specific nucleotides with increasing resolution[66–68].

There are now numerous methods to infer nucleotide-level constraints in genomic sequence alignments[20,21,60,66–71] (TABLE 2). These methods differ in detail but are united by the principle of estimating observed rates of evolutionary change and contrasting these estimates with rates expected for neutral positions; sites with fewer substitution events receive higher scores, which are indicative of an increased likelihood and/or intensity of constraint. One important distinction to consider in applying any given approach is context dependency. Some methods — such as binCons[69] and phastCons[60] — use sliding windows or hidden Markov models and consequently the score at each position depends partially on the score of its neighbours. This is in contrast with other methods — such as Genomic Evolutionary Rate Profiling (GERP)[67], Gumby[21], phyloP[66] and Sequence Conservation Evaluation (SCONE)[68] — that consider each position independently. The methods that use sliding windows have the advantage of generating 'smoother' score distributions and are necessary when the signal-to-noise ratio for any given position is low. However, they will tend to

Box 1 | **GERP and polyPhen score distribution for non-synonymous variants**

polyPhen	**Benign**	**Possibly**	**Probably**	**No prediction**
	69.9%	16.1%	10.2%	3.8%
GERP	**<3**	**3–5**	**>5**	**No prediction**
	61.0%	27.8%	9.1%	2.0%

In the figure, we summarize distributions of protein impact estimates, defined by 'polymorphism phenotyping' (polyPhen, version 1)[35,45], and constraint scores, defined by Genomic Evolutionary Rate Profiling (GERP)[17,67,72], for non-synonymous variants in human exome data that were reported Ng *et al.*[8] These data are meant to provide a general sense for the effects of variant filtering with commonly used definitions of deleteriousness, and other approaches (TABLES 1,2) could be used similarly. Scores from polyPhen are binned into three groups, 'benign', 'possibly damaging' and 'probably damaging', and are based on a trained classifier using both biochemical and evolutionary information. GERP scores are based on genomic sequence alignments and measured as 'rejected substitutions' (note that the bins were defined arbitrarily), indicating the difference between observed and expected (assuming neutrality) rates of evolution. Neutral sites tend to score near zero, whereas constrained sites generally score positively. The maximum genome-wide score (~5.8) only applies to sites that are perfectly conserved across all sequenced mammals, whereas 'no prediction' applies to sites that are aligned to none or few species (mostly repetitive sequences).

confound predictions for neighbouring sites: consider, for example, a truly neutral synonymous position flanking a highly constrained non-synonymous position or a crucial nucleotide in a transcription factor binding site adjacent to a degenerate nucleotide. Furthermore, comparative genomic data sets now afford substantial amounts of per-position information. Alignments of more than 30 mammalian genome draft sequences capturing ~6 substitutions per neutral site are readily available (see the University of California Santa Cruz (UCSC) Genome Bioinformatics website) with steady expansions anticipated (for example, the Genome 10K Project). Given such data sets, single-position scoring methods based on mammalian sequence alignments are generally preferable for human SNV analysis.

The relative utility of nucleotide- and protein-based approaches can be compared using exome data (for example, BOX 1). Recently, we showed that nucleotide constraint scores defined by GERP[67] provide enrichments for deleterious causal mutations that are similar to those provided by protein-based annotations for two Mendelian diseases[72]. This was seen in spite of the fact that synonymous variants were also evaluated. In addition, this approach facilitated quantitative ranking of candidates and placed the known causal genes at or near the top of candidate lists. This constraint-based ranking is unavailable with methods that generate discrete functional predictions (for example, 'benign' versus 'damaging'). Surprisingly, these observations suggest that nucleotide-level metrics are as powerful as protein-based metrics, even without knowledge of protein biochemistry. There are at least three potential explanations for this finding. First, although the absence of functional information may reduce predictive power, this may be offset by the elimination of paralogues and functionally divergent orthologues that are often part of the more diverse alignments used for protein analyses[33] (indeed, a recent study[43] showed that choice of phylogenetic scope is one explanation for discrepant predictions made by protein-based algorithms). Second, absence of functional information is actually beneficial when that information is misleading; this may occur, for example, when a non-synonymous variant appears to be benign at the amino acid level but has a deleterious effect on splicing[73]. Finally, quantitative assessments in general have advantages over discrete predictions. Importantly, each of these explanations may be exploited to improve protein-based metrics through more careful sequence collection and alignment curation (for example, using syntenic relationships to differentiate orthologues from paralogues), more comprehensive descriptions of molecular function (see below) and by generating quantitative predictions.

It is clear that single-nucleotide constraint scores have a practical utility for exome studies and that they can identify small, functionally enriched subsets of both exomes and genomes (BOX 2). However, their use in identifying causal variants in non-coding regions of whole genomes remains unproven. Encouragingly, numerous studies have identified correlations between increasing constraint score and decreasing genetic diversity at both the population[74,75] and individual[76] levels. Such observations demonstrate that non-coding sites under constraint throughout mammalian evolution have been subject to recent purifying selection, which is indicative of mutational deleteriousness. Strong correlations between constraint scores and experimental measures of regulatory functionality also exist. These include many examples of constraint-predicted enhancers with crucial roles in embryonic development[64,77] and more general correlations between non-coding molecular functionality and evolutionary constraint[19,78]. As such, comparative genomic annotations convey a significant amount of site-specific information that correlates with both population genetic and *in vitro* assessments of non-coding deleteriousness and thus offer clear promise for causal-variant discovery in whole genomes.

Experimental approaches

Experimental analyses are generally performed to add support to a limited number of candidate causal variants identified using other information and can provide powerful support of causality for a given phenotype. For example, this may consist of the *in vitro* demonstration of molecular consequences (such as disruption of expression or protein folding) or the *in vivo* recapitulation of the human phenotype in a model organism. However, experimental approaches are not yet widely used for prioritizing large numbers of candidate variants, primarily because experimental analyses of variants are challenging to implement and costly to scale. Assessment of coding mutations in particular requires the identification of a measurable property that associates with function, which remains difficult as little is known about the function of most human proteins. There are, however, at least three emerging paradigms for large-scale experimental assessment of genetic variants.

Box 2 | **Distribution of GERP scores within subsets of human genomes**

	GERP < 0	>0	>1	>2	>3	> 4	>5	No prediction
Exome	19.5%	79.3%	73.5%	65.7%	55.9%	42.3%	20.8%	1.1%
Genome	27.0%	29.2%	18.9%	8.9%	3.8%	1.6%	0.55%	43.8%
Non-coding variation*	43.4%	39.9%	24.5%	10.9%	4.0%	1.4%	0.33%	16.7%
Synonymous variation‡	61.4%	37.6%	29.1%	20.4%	12.4%	5.6%	1.5%	0.95%
Non-synonymous variation‡	31.3%	66.6%	58.3%	48.1%	36.8%	23.6%	9.1%	2.0%

*Non-coding variants observed by Keinan *et al.*[117]. ‡Coding variants observed by Ng *et al.*[8].

The table provides an overview of Genomic Evolutionary Rate Profiling (GERP) score distributions for the genome, exome and different categories of single nucleotide variants (SNVs); these data provide a general sense for the effects of constraint and the utility of constraint scores to prioritize subsets of genomes and variants. For example, a threshold of rejected substitutions >5 includes 0.55% of the genome but 20.8% of the exome, reflecting the strong enrichment for sites under constraint within protein-coding exons. However, as there are ~16.9 million total genomic sites with rejected substitutions >5 (0.55% of ~3.08 billion) but only ~6.9 million in the exome (20.8% of ~33 million), most sites at this threshold are non-coding. Similar arguments apply to variant collections. Note that the depletion for 'no prediction' in all three variation categories relative to the genome is in part due to regions that are both difficult to align to other species and difficult to re-sequence (for example, repetitive sequences).

First, projects such as the Encyclopedia of DNA Elements (ENCODE) are applying diverse assays in many cell types and conditions to generate functional annotations at a genome-wide scale, including protein-coding genes, non-coding RNAs and *cis*-regulatory elements[19,79]. These data facilitate hypothesis generation to guide and prioritize variants based on their overlap with molecular features or interactions. For example, a candidate causal variant may be observed to disrupt a consensus sequence motif within a known binding site for a particular transcription factor, rapidly defining a specific and testable hypothesis. Additionally, such data sets may be sufficient in themselves to identify testable associations between common polymorphisms and gene expression or protein binding through expression quantitative trait loci (eQTLs) and related analyses[80–82].

Second, several groups are developing strategies whereby variants in regulatory sequences[83], RNAs[84] and proteins[85] can be studied in a highly multiplexed fashion. In these methods, synthetic approaches are used to construct mutagenized libraries of a sequence of interest that are subjected to massively parallel functional characterization (FIG. 3). For example, Patwardhan *et al.*[83] generated synthetic promoter libraries that included all possible single-nucleotide mutations relative to a reference sequence by cleaving off oligonucleotides from a custom microarray (FIG. 3a). They then performed *in vitro* transcription and used high-throughput RNA sequencing (RNA-seq) of barcodes linked to variant promoters as a measure of their transcriptional activity. Fowler *et al.*[85] demonstrated high-resolution mapping of protein sequence–function relationships by subjecting a degenerate library of protein variants of a human WW domain to successive rounds of phage display and peptide ligand binding and used massively parallel DNA sequencing of the library at each stage as the readout (FIG. 3b). Such methods and their derivatives may prove useful for the experimental 'pre-evaluation' of large numbers of potential mutations in loci of clinical relevance.

Third, detailed but generically assayed molecular phenotypes may be useful to capture and measure protein function. For example, cells may be perturbed by overexpression or knockdown of specific genes and subsequently subjected to high-throughput assessments, such as RNA-seq of transcriptional activity[86] or chromatin immunoprecipitation followed by sequencing (ChIP–seq) of transcription factor binding[87]. The extent to which the overexpression of mutant forms of the gene recapitulate the phenotype associated with overexpression of the wild-type gene or rescue the phenotype associated with knockdown of the wild-type gene may be informative with respect to the impact of the mutation on function. A recent analysis of deletion mutations of the transcription factor E2F1, for example, demonstrated that only loss of the DNA binding domain (DBD) changed its genomic binding profile in MCF7 breast cancer cells[88]. Consistent with the observation that the DBD is the most highly evolutionarily constrained region of the protein, these data suggest that mutations of the DBD are likely to be more deleterious to E2F1 function than mutations elsewhere. As the costs of variant gene synthesis and sequencing-based assays continue to decline, such molecular phenotyping will increasingly provide functional assessments for many mutations in many genes, including those for which little or no prior knowledge is available.

Interpretive difficulties for experimental characterization. It is important to recognize that experimental predictions, as with computational predictions, are informative but often not definitive. For negative results, an inevitable concern is whether the experiment was performed in the appropriate context[89], including: genomic context (that is, dependencies on flanking sequence or chromatin state); developmental context (that is, dependency on cell type or developmental stage); or organismal context (that is, functional consequences of a mutation that are species-specific).

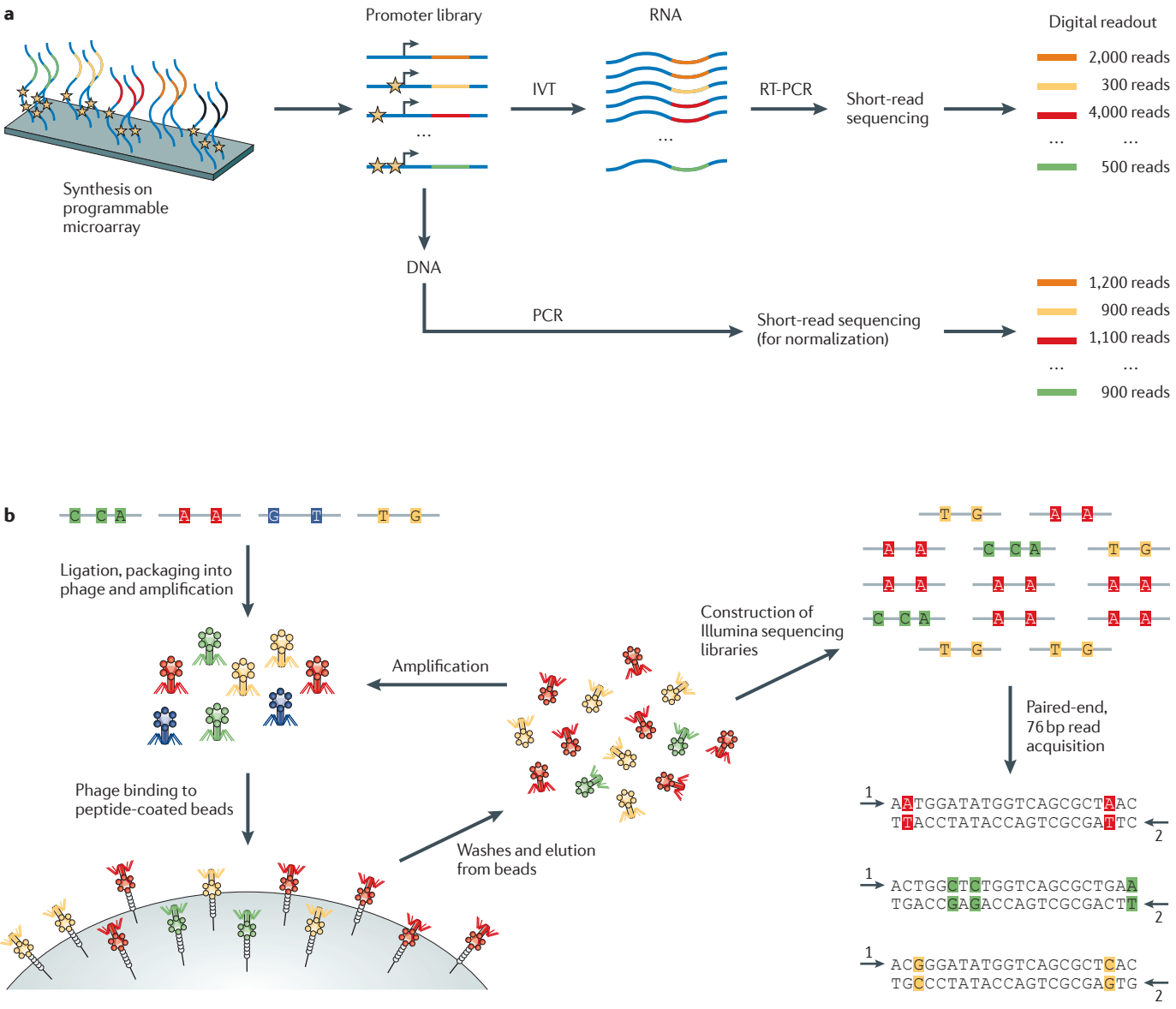

Figure 3 | **High-throughput experimental assessment of variant function.** **a** | High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Mutated versions of a regulatory element are synthesized in parallel on a programmable microarray, linked in *cis* to a transcribed barcode and subjected to *in vitro* transcription (IVT). Deep sequencing of barcode-derived RNAs enables measurement of the functional impact of non-coding mutations at single-nucleotide resolution. **b** | High-resolution mapping of protein sequence–function relationships by phage display of a complex library of sequences encoding variants of a single WW domain and multiple rounds of selection consisting of phage binding to a peptide ligand via the expressed WW domain. Deep sequencing of the resulting library of variant WW domains is applied after each round of selection. This method quantitatively assesses the impact of non-synonymous mutations on the binding efficiency of the WW domain to its peptide ligand. **c** | The binding of a WW domain (blue ribbon) to its peptide ligand, with the primers used for sequencing numbered and labelled with arrows. RT-PCR, reverse transcriptase PCR. Panel **a** is adapted, with permission, from REF. 83 © (2009) Macmillan Publishers Ltd. Panel **b** is adapted, with permission, from REF. 85 © (2010) Macmillan Publishers Ltd. All rights reserved.

Conversely, positive experimental results do not necessarily establish causality or deleteriousness. For example, a mutation that happens to fall in a promoter or enhancer has a reasonable chance of influencing transcriptional regulation, but it is difficult to assess the significance of a functional non-coding mutation based solely on the magnitude of its impact on transcriptional regulation. The same concern applies to coding sequences, in which the experimental observation of a proximal molecular impact does not necessarily imply causality with respect to a phenotype at the organismal level (for example, nonsense variants that exist at high frequencies in many genes[31]). In both cases, experimentally demonstrable molecular functionality does not necessarily equate to organismal deleteriousness (a similar concern also applies to computational assessments).

Challenges and unanswered questions

A major goal will be to develop a unified, quantitative, predictive framework to estimate the prior probabilities for any given mutation to be both functionally relevant and disease relevant, accounting for both computational and experimental sources of information. A number of challenges must be met for such a framework to succeed, but the following are particularly important.

Evolutionary versus experimental annotations. A better understanding of the relative value of evolutionary annotations versus functional annotations, and discrepancies between them, will be important for accurate assessments of deleteriousness. ENCODE, for example, found only modest correlation between constraint estimates and molecular functions in human cells[19,78]. The existence of constrained nucleotides without experimental annotation is not surprising, as these are likely to be important positions that have not been appropriately assayed yet; given the number of possible combinations of developmental time, cell type and molecular functionality, it is inevitable that many functional sites remain experimentally uncharacterized. However, the converse — namely, experimentally functional sites without evidence for constraint — is more problematic. This in part reflects truly important functions that appear unconstrained. There are structural features of DNA, for example, that are under constraint that is undetectable by primary sequence conservation[71], and there are also clear examples of important regulatory elements that fail constraint-based measures[90]. However, it is likely that there are also many molecular events in cells that are biochemically functional but biologically inert, lacking even in phenotypic relevance to the cell, let alone the organism[12,19,78]. For example, although transcription factors have been shown to bind to thousands or tens of thousands of sites in any given cell type, the expression levels of relatively few nearby genes change when the binding patterns are modified[91], suggesting that many, or perhaps most, individual binding events have minimal downstream effects.

Coding versus non-coding variants. The relative importance of coding versus non-coding variants is unknown, but it is essential to define appropriate weighting schemes for each in disease studies. As discussed above, enrichments for coding variants in Mendelian disease are biased upwards. However, the estimated 5/1 ratio of constrained non-coding to constrained coding bases[17,18,60] decreases as constraint thresholds increase[67], suggesting that the ratio of non-coding to coding causal variation will decline relative to increasing mutational penetrance and disease severity. Better quantification of both the relative abundances and penetrances of coding and non-coding causal variants is therefore essential. Note that such estimates are not necessarily a factor in deciding between exome and genome sequencing for any given disease. Current cost-to-benefit ratios clearly support exome sequencing as an approach to efficiently study the many traits that primarily result from coding alterations (for example, REFS 8,29), in much the same way that microarray-based detection of copy number variants (CNVs) can be used to identify obviously pathogenic variants for some phenotypes without any sequencing[92]. Rather, this question bears on interpretation and assessment of whole-genome sequences when costs decrease to the point that such data are routinely obtained.

However, our emphasis here on differential considerations for coding and non-coding SNVs understates the complexity of both the genome and of genetic variation. For example, 'non-coding' spans a diverse collection of functional consequences, including both transcribed variants (such as non-coding RNA) and non-transcribed variants (such as promoters) (FIG. 1). Furthermore, many SNVs have more than one potential functional consequence, such as a non-synonymous change that also disrupts a splicing enhancer[73]. This functional diversity, as well as the differential extent to which evolutionary constraint operates on each class of sequence, renders our simple dichotomy between coding and non-coding SNVs problematic. Furthermore, causative mutations include not only SNVs but also structural changes, such as small insertion or deletion events, large CNVs, retrotransposition events and inversions[3,92,93,94]. A unified predictive framework must be able to estimate deleteriousness for these types of mutations as well.

Tests for accuracy. A major challenge to the development and interpretation of deleteriousness predictions is accuracy assessment ('benchmarking'), which requires large collections of true positive (deleterious) and true negative (neutral) mutations, ideally collected in a manner that is unbiased relative to the methodology being evaluated. Several data sources are currently used as benchmarks. One such source is mutagenesis experiments of individual proteins that generate many possible mutations and assess their function, as has been performed for several bacterial and viral proteins[95–97]. A second benchmark is gene-specific collections of mutations that associate with disease: for

example, *HBB* variants that result in thalassaemia[98] (FIG. 2) and *TP53* mutations that are observed in tumours[99]. Finally, pathogenic variants that have been identified for various genes and diseases are collected in large-scale databases such as Online Mendelian Inheritance in Man (OMIM), Swiss-Prot[100] and the Human Gene Mutation Database (HGMD)[62]. Although valuable, each of these resources suffers from drawbacks[34]. For example, mutagenesis studies of individual proteins provide robust mutational assessments for those proteins, but the extent to which those estimates generalize to other proteins is unclear, especially when using a single viral or bacterial enzyme to study the functionally diverse human proteome. Conversely, databases of known pathogenic variants relate directly to human genes and diseases but only include mutations that manifest in particular clinical phenotypes; mutations that give rise to lethality, subclinical disease or a distinct phenotype are often not sampled. Benchmarking resources are especially limited for non-coding variants, as there are relatively few known pathogenic non-coding mutations compared with the situation for coding variants, and those that exist often reside in regulatory sequences immediately adjacent to a small number of genes. Recent technological developments are likely to improve the depth and breadth of experimental assessments of both protein and regulatory function, as well as catalogues of defined pathogenic variants. Such resources will be needed to improve the extent to which deleteriousness predictions can be benchmarked.

Regulatory element vocabularies. In the same way that knowledge of the genetic code greatly increases the power to predict the impact that mutations have within proteins, richer and more comprehensive functional vocabularies for regulatory elements will be needed to assess the impact of non-coding variants. Some examples of this are beginning to emerge. Sequence features that are related to splicing regulation[73] are generally part of exome studies[7,8] and are also incorporated into some predictive methods[40]. Additionally, ENCODE[19] is generating motif definitions for a wide range of regulatory element features. These data show, for example, that binding of a transcription factor to a genomic site harbouring a heterozygous SNV in a given individual favours the allele that more closely matches the consensus sequence (Timothy E. Reddy and Richard M. Myers, HudsonAlpha Institute for Biotechnology, personal communication). These findings are consistent with observations from model organisms that rates of evolution of individual nucleotides within protein binding sites correlate strongly with the level of degeneracy of those nucleotides in consensus binding motifs[101–103]. In addition, progress is being made in identifying long-range regulatory elements (for example, REFS 90,104) and matching them to their target genes through techniques such as chromosome conformation capture[105]. Accurately pairing regulatory elements and genes is particularly important for gene-level prediction methods that identify disease genes on the basis of aggregate enrichments for rare deleterious variants in affected individuals[106,107].

Variant interactions. Finally, interactions between individual variants and between variants and the environment are clearly relevant to accurate genotype–phenotype predictions. Molecular characterizations of such interactions are emerging in yeast. For example, it has recently been shown that multiple transcription factor mutations and growth conditions interact to modulate sporulation efficiency[108,109]. However, although promising approaches are being developed (see, for example, REFS 106,110), such analyses are orders of magnitude more difficult in humans, as control and documentation of environmental exposures are limited and the combinatorial possibilities of even two-way genetic interactions are myriad. Comprehensive descriptions of molecular interactions, including protein–protein interaction[111] and gene co-expression networks[112], coupled with both literature and automated annotation of pathways and gene functions[113], are crucial to tackle this challenge. Such characterizations may take the form of identifying 'excess' deleteriousness within pathways or network modules, similar in some respects to the identification of significant excesses of large CNVs observed among individuals with neurological disease[93,94]. These assessments identify groupings of risk variants and provide a starting point to begin addressing the effects of individual variants and combinations thereof.

Concluding remarks

Next-generation sequencing has enabled 'next-generation genetics', wherein variant identification is no longer the rate-limiting step. However, the interpretive challenges preventing the optimal exploitation of these data are formidable. Although hardly unprecedented in quantitative genetics applications, these challenges are becoming increasingly important given the rapid expansion of genetic approaches for studying human disease, including the study of many phenotypes that were considered to be unapproachable until recently[6,114].

These challenges also highlight a great irony. The core strength of genomic approaches for understanding disease is the freedom to make discoveries in previously unexplored places, replacing informed but biased hypotheses with unbiased but generic ones. However, as a consequence, every result must be treated with the scepticism that is appropriate for an ignorant hypothesis, and bringing down all tests to the lowest common denominator comes at a cost of missing important discoveries. Importantly, the dichotomy between near universal assumptions of ignorance[115] or knowledge[116] is a false one. Full realization of the potential for genomics to characterize the genetic basis for human disease demands a better way to balance these opposites, removing biases driven by historical accidents, false premises or simple myopia but exploiting those that reflect genuine biology.

1. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature Biotech.* **26**, 1135–1145 (2008).
2. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
3. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
4. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011).
5. Manly, K. F., Nettleton, D. & Hwang, J. T. Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Res.* **14**, 997–1001 (2004).
This is a valuable review of the relationships between prior probability, statistical significance and false-discovery rates as they pertain to genome-wide analyses.
6. Morton, N. E. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7**, 277–318 (1955).
7. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature Genet.* **42**, 30–35 (2010).
8. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
This is the first demonstration of exome sequencing being used to identify the causal variants for a Mendelian disease. Protein-based annotations of functional deleteriousness were essential to this effort.
9. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl Acad. Sci. USA* **106**, 19096–19101 (2009).
10. Erlich, Y. *et al.* Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. *Genome Res.* **21**, 658–664 (2011).
11. Kimura, M. *The Neutral Theory Of Molecular Evolution* (Cambridge Univ. Press, New York, 1983).
12. Cooper, G. M. & Brown, C. D. Qualifying the relationship between sequence conservation and molecular function. *Genome Res.* **18**, 201–205 (2008).
13. McAuliffe, J. D., Jordan, M. I. & Pachter, L. Subtree power analysis and species selection for comparative genomics. *Proc. Natl Acad. Sci. USA* **102**, 7900–7905 (2005).
14. Stone, E. A., Cooper, G. M. & Sidow, A. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **6**, 143–164 (2005).
15. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3**, e10 (2005).
16. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
17. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++ . *PLoS Comput. Biol.* **6**, e1001025 (2010).
18. The Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
19. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
20. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
21. Prabhakar, S. *et al.* Close sequence comparisons are sufficient to identify human *cis*-regulatory elements. *Genome Res.* **16**, 855–863 (2006).
22. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326 (2010).
23. Johnson, M. E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
24. Wang, T. *et al.* Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl Acad. Sci. USA* **104**, 18613–18618 (2007).
25. Enard, W. *et al.* Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* **418**, 869–872 (2002).
26. Prabhakar, S. *et al.* Human-specific gain of function in a developmental enhancer. *Science* **321**, 1346–1350 (2008).
This study demonstrates that constraint-based measures may also identify sequences with human-specific functionality.
27. Stone, E. A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* **15**, 978–986 (2005).
The authors describe a combined phylogenetic and biochemical approach to predict the effects of amino acid substitutions. They demonstrate a quantitative relationship between past evolutionary rates of biochemical change and present day deleteriousness.
28. De Gobbi, M. *et al.* A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **312**, 1215–1217 (2006).
29. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet.* **33**, 228–237 (2003).
30. Ng, S. B. *et al.* Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nature Genet.* **42**, 790–793 (2010).
31. MacArthur, D. G. & Tyler-Smith, C. Loss-of-function variants in the genomes of healthy humans. *Hum. Mol. Genet.* **19**, R125–R130 (2010).
32. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
33. Ng, P. C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7**, 61–80 (2006).
34. Care, M. A., Needham, C. J., Bulpitt, A. J. & Westhead, D. R. Deleterious SNP prediction: be mindful of your training data! *Bioinformatics* **23**, 664–672 (2007).
35. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248–249 (2010).
36. Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35**, 3823–3835 (2007).
37. Capriotti, E., Calabrese, R. & Casadio, R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**, 2729–2734 (2006).
38. Ferrer-Costa, C., Orozco, M. & de la Cruz, X. Sequence-based prediction of pathological mutations. *Proteins* **57**, 811–819 (2004).
39. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
This describes SIFT (also see reference 46), a commonly used tool to predict the effects of amino acid substitutions and an early demonstration of the importance of sequence conservation to functional predictions.
40. Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods* **7**, 575–576 (2010).
41. Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).
42. Ye, Z. Q. *et al.* Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). *Bioinformatics* **23**, 1444–1450 (2007).
43. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).
44. Bao, L., Zhou, M. & Cui, Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.* **33**, W480–W482 (2005).
45. Sunyaev, S. *et al.* Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591–597 (2001).
This paper describes polymorphism phenotyping (polyPhen) (also see reference 35), a commonly used tool to predict the effects of amino acid substitutions, and illustrates the value of classifiers trained on numerous biochemical and evolutionary features.
46. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
47. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
48. Marini, N. J., Thomas, P. D. & Rine, J. The use of orthologous sequences to predict the impact of amino acid substitutions on protein function. *PLoS Genet.* **6**, e1000968 (2010).
49. Dobson, R. J., Munroe, P. B., Caulfield, M. J. & Saqi, M. A. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. *BMC Bioinformatics* **7**, 217 (2006).
50. Saunders, C. T. & Baker, D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.* **322**, 891–901 (2002).
51. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353**, 459–473 (2005).
52. Bao, L. & Cui, Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics* **21**, 2185–2190 (2005).
53. Li, Y. *et al.* Predicting disease-associated substitution of a single amino acid by analyzing residue interactions. *BMC Bioinformatics* **12**, 14 (2011).
54. 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).
55. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
This paper describes the precise identification of a common transcriptional regulatory variant that influences cholesterol levels and cardiovascular disease risk.
56. Storey, J. D. *et al.* Gene-expression variation within and among human populations. *Am. J. Hum. Genet.* **80**, 502–509 (2007).
57. 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).
This analysis demonstrated that expression-associated variants are enriched among trait-associated variants, suggesting that non-coding regulatory variants are causally relevant for many traits.
58. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
59. Carroll, S. B. Evolution at two levels: on genes and form. *PLoS Biol.* **3**, e245 (2005).
60. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
61. 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).
This study describes non-coding mutations that cause Mendelian limb defects by affecting enhancers important to developmental sonic hedgehog (*Shh*) gene regulation. A combination of evolutionary sequence conservation and mouse-based experimental assessments of variant function were used.
62. Stenson, P. D. *et al.* The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum. Genomics* **4**, 69–72 (2009).
63. Treisman, R., Orkin, S. H. & Maniatis, T. Specific transcription and RNA splicing defects in five cloned β-thalassaemia genes. *Nature* **302**, 591–596 (1983).
64. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2005).
65. Dehal, P. *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157–2167 (2002).
66. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
67. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
68. Asthana, S., Roytberg, M., Stamatoyannopoulos, J. & Sunyaev, S. Analysis of sequence conservation at nucleotide resolution. *PLoS Comput. Biol.* **3**, e254 (2007).
69. Margulies, E. H., Blanchette, M., Haussler, D. & Green, E. D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507–2518 (2003).

70. Dubchak, I. *et al.* Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10**, 1304–1306 (2000).
71. Parker, S. C., Hansen, L., Abaan, H. O., Tullius, T. D. & Margulies, E. H. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324**, 389–392 (2009).
72. Cooper, G. M. *et al.* Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nature Methods* **7**, 250–251 (2010).
This paper demonstrated that functionally agnostic nucleotide-level constraint scores, defined by GERP (also see references 17 and 67), offer considerable utility for causal variant discovery in exome analyses.
73. Wang, G. S. & Cooper, T. A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Rev. Genet.* **8**, 749–761 (2007).
74. Drake, J. A. *et al.* Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genet.* **38**, 223–227 (2006).
75. Katzman, S. *et al.* Human genome ultraconserved elements are ultraselected. *Science* **317**, 915 (2007).
76. Goode, D. L. *et al.* Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. *Genome Res.* **20**, 301–310 (2010).
77. Pennacchio, L. A. *et al. In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
78. Margulies, E. H. *et al.* Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* **17**, 760–774 (2007).
79. The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
80. Ge, B. *et al.* Global patterns of *cis* variation in human cells revealed by high-density allelic expression analysis. *Nature Genet.* **41**, 1216–1222 (2009).
81. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
82. Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M. & Snyder, M. Genetic analysis of variation in transcription factor binding in yeast. *Nature* **464**, 1187–1191 (2010).
83. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotech.* **27**, 1173–1175 (2009).
This paper defined a method to exploit next-generation sequencing to comprehensively yet efficiently assay point mutations in transcriptional promoters.
84. Pitt, J. N. & Ferre-D'Amare, A. R. Rapid construction of empirical RNA fitness landscapes. *Science* **330**, 376–379 (2010).
85. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nature Methods* **7**, 741–746 (2010).
86. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods* **5**, 621–628 (2008).
87. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
88. Cao, A. R. *et al.* Genome-wide analysis of transcription factor E2F1 mutant proteins reveals that N- and C-terminal protein interaction domains do not participate in targeting E2F1 to the human genome. *J. Biol. Chem.* **286**, 11985–11996 (2011).
89. Botstein, D. & Shortle, D. Strategies and applications of *in vitro* mutagenesis. *Science* **229**, 1193–1201 (1985).
90. Blow, M. J. *et al.* ChIP-seq identification of weakly conserved heart enhancers. *Nature Genet.* **42**, 806–810 (2010).
91. Cheng, Y. *et al.* Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res.* **19**, 2172–2184 (2009).
92. Miller, D. T. *et al.* Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.* **86**, 749–764 (2010).
93. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
94. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).
95. Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S. & Miller, J. H. Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J. Mol. Biol.* **240**, 421–433 (1994).
96. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222**, 67–88 (1991).
97. Loeb, D. D. *et al.* Complete mutagenesis of the HIV-1 protease. *Nature* **340**, 397–400 (1989).
98. Hardison, R. C. *et al.* HbVar: a relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. *Hum. Mutat.* **19**, 225–233 (2002).
99. Olivier, M. *et al.* The IARC TP53 database: new online mutation analysis and recommendations to users. *Hum. Mutat.* **19**, 607–614 (2002).
100. Yip, Y. L. *et al.* The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum. Mutat.* **23**, 464–470 (2004).
101. Brown, C. D., Johnson, D. S. & Sidow, A. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* **317**, 1557–1560 (2007).
102. Kim, J., He, X. & Sinha, S. Evolution of regulatory sequences in 12 *Drosophila* species. *PLoS Genet.* **5**, e1000330 (2009).
103. Moses, A. M., Chiang, D. Y., Kellis, M., Lander, E. S. & Eisen, M. B. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol. Biol.* **3**, 19 (2003).
104. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
105. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
106. Liu, D. J. & Leal, S. M. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.* **6**, e1001156 (2010).
This paper describes an approach to assess the significance of correlations between gene or locus aggregates of rare variants and phenotypes and may also be useful in identifying significant variant interactions.
107. Yandell, M. *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Res.* 23 Jun 2011(doi:10.1101/gr.123158.111).
This paper defines a method, VAAST, to predict disease genes or loci on the basis of the total predicted deleteriousness of rare variants observed in affected individuals.
108. Gerke, J., Lorenz, K. & Cohen, B. Genetic interactions between transcription factors cause natural variation in yeast. *Science* **323**, 498–501 (2009).
109. Gerke, J., Lorenz, K., Ramnarine, S. & Cohen, B. Gene–environment interactions at nucleotide resolution. *PLoS Genet.* **6**, e1001144 (2010).
110. Bush, W. S. *et al.* A knowledge-driven interaction analysis reveals potential neurodegenerative mechanism of multiple sclerosis susceptibility. *Genes Immun.* (2011).
111. Rual, J. F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
112. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
113. The Gene Ontology Consortium. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
114. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
115. Ioannidis, J. P. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
116. Rothman, K. J. No adjustments are needed for multiple comparisons. *Epidemiology* **1**, 43–46 (1990).
117. Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genet.* **39**, 1251–1255 (2007).

Acknowledgements
We thank C. Brown and E. Stone for comments on an earlier draft and R. Patwardhan for sharing data.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Gregory M. Cooper's homepage: http://www.hudsonalpha.org/g-cooper-lab
Jay Shendure's homepage: http://krishna.gs.washington.edu
Encyclopedia of DNA Elements (ENCODE): http://www.genome.gov/10005107
The Genome 10K Project: http://genome10k.soe.ucsc.edu
University of California, Santa Cruz (UCSC) Genome Bioinformatics: http://genome.ucsc.edu
Human Gene Mutation Database (HGMD): http://www.hgmd.org
National Human Genome Research Institute Catalog of Published Genome-Wide Association Studies: http://www.genome.gov/gwastudies
Online Mendelian Inheritance in Man (OMIM): http://www.ncbi.nlm.nih.gov/omim

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Author Biographies

Gregory M. Cooper has been a Faculty Investigator at the HudsonAlpha Institute for Biotechnology since September 2010. He carried out his doctoral work with Arend Sidow at Stanford University, California, USA, and his postdoctoral work with Evan Eichler and Debbie Nickerson at the University of Washington, Seattle, USA. His laboratory's work centres on the use of functional genomics, expression genetics and comparative genomics to improve the interpretation of human genetic variation.

Jay Shendure has been on the Faculty in the Department of Genome Sciences at the University of Washington since September 2007. He carried out his doctoral work with George Church at Harvard Medical School, Boston, Massachusetts, USA. His laboratory is broadly focused on translational genomics and technology development for high-throughput genomics and molecular biology.

Online Summary

- Genome and exome sequencing yield extensive catalogues of genetic variation in many individuals, but purely genetic approaches are often insufficiently powered to specifically identify the few variants that are causally related to any given phenotype. Indeed, variant interpretation is an increasingly important challenge at the interface of genetics, statistics and biology.
- Non-uniform estimates of the prior probability for variants to be biologically functional will be required to address this challenge. For disease studies, this can be translated into the need to estimate variant deleteriousness.
- Nearly all computational methods to predict deleteriousness use comparative sequence analysis, exploiting the fact that natural selection removes deleterious variants and tends to conserve the identities of important positions within genes and genomes.
- Assessment of protein-altering variants leverages both biochemical and evolutionary information, whereas non-coding variation is more challenging to study, given a lack of understanding of the molecular functionality of non-coding sequences relative to coding sequences.
- Experimental assessments of the functional impact of variants have historically relied on low-throughput assays. However, projects such as the Encyclopedia of DNA Elements (ENCODE) and the clever use of next-generation sequencing technologies are increasingly facilitating large-scale, systematic experimental assessment of genomic variation of many types.
- Ultimately, unified predictive methods that are applicable to both coding and non-coding variants that leverage both functional and evolutionary information will be crucial for the meaningful interpretation of personal genomes. However, important unknowns and unsolved phenomena, including the relative abundance and penetrance of coding versus non-coding variants, disagreements between evolutionary and experimental definitions of molecular functionality, and the vocabularies that define transcriptional regulatory elements, must first be addressed.

Table of contents summary

Subject categories

Genomics, Bioinformatics, Disease Genetics, Nucleic Acid Sequencing