nature genetics

A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher[1,5], Daniela M Witten[2,5], Preti Jain[3,4], Brian J O'Roak[1,4], Gregory M Cooper[3] & Jay Shendure[1]

Current methods for annotating and interpreting human genetic variation tend to exploit a single information type (for example, conservation) and/or are restricted in scope (for example, to missense changes). Here we describe Combined Annotation–Dependent Depletion (CADD), a method for objectively integrating many diverse annotations into a single measure (C score) for each variant. We implement CADD as a support vector machine trained to differentiate 14.7 million high-frequency human-derived alleles from 14.7 million simulated variants. We precompute C scores for all 8.6 billion possible human single-nucleotide variants and enable scoring of short insertions-deletions. C scores correlate with allelic diversity, annotations of functionality, pathogenicity, disease severity, experimentally measured regulatory effects and complex trait associations, and they highly rank known pathogenic variants within individual genomes. The ability of CADD to prioritize functional, deleterious and pathogenic variants across many functional categories, effect sizes and genetic architectures is unmatched by any current single-annotation method.

A strength of genomic approaches in studying disease is the ability to replace informed but biased hypotheses with unbiased but generic ones, such as the equal treatment of all genetic variants in genome-wide association studies (GWAS). However, for both rare variants of large effect and common variants of weak effect, the use of prior knowledge can be critical for disease gene discovery[1–4]. For example, exome sequencing is an effective discovery strategy because it focuses on protein-altering variation, which is enriched for causal effects[5].

Although many existing annotation methods are useful for prioritizing causal variants to boost discovery power (for example, PolyPhen[6], SIFT[7] and GERP[8]), current approaches tend to suffer from one or more of four major limitations. First, annotation methods vary widely with respect to both inputs and outputs. For example, conservation metrics[8–10] are defined across the genome but do not use functional information and are not allele specific, whereas protein-based metrics[6,7] apply only to coding and often only to missense variants, thereby excluding >99% of human genetic variation. Second, each annotation method has its own metric, and these metrics are rarely comparable, making it difficult to evaluate the relative importance of distinct variant categories or annotations. Third, annotation methods trained on known pathogenic mutations are subject to major ascertainment biases and may not be generalizable. Fourth, it is a major practical challenge to obtain, let alone to objectively evaluate or combine, the existing panoply of partially correlated and partially overlapping annotations; this challenge will only increase in size as large-scale projects such as the Encyclopedia of DNA Elements (ENCODE)[11] continually increase the amount of relevant data available. The net result of these limitations is that many potentially relevant annotations are ignored, while the annotations that are used are applied and combined in *ad hoc* and subjective ways that undermine their usefulness.

Here we describe a general framework, Combined Annotation–Dependent Depletion (CADD), for integrating diverse genome annotations and scoring any possible human single-nucleotide variant (SNV) or small insertion-deletion (indel) event. The basis of CADD is to contrast the annotations of fixed or nearly fixed derived alleles in humans with those of simulated variants. Deleterious variants—that is, variants that reduce organismal fitness—are depleted by natural selection in fixed but not simulated variation. CADD therefore measures deleteriousness, a property that strongly correlates with both molecular functionality and pathogenicity[12]. Notably, metrics of deleteriousness, in contrast to pathogenicity or molecular functionality, have major advantages. Whereas the latter are limited in scope to a small set of genetically or experimentally well-characterized mutations and are subject to major ascertainment biases, deleteriousness can be measured systematically across the genome assembly (see refs. 8–10 and below). Further, selective constraint on genetic variants is related to the totality of their phenotype-relevant effects rather than to any individual molecular or phenotypic consequence. Measures of deleteriousness can therefore provide, in principle, a genome-wide, data-rich, functionally generic and organismally relevant estimate of variant effect.

RESULTS

Implementation of CADD

We identified differences between human genomes and the inferred human-chimpanzee ancestral genome[13] where humans carry a

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Department of Biostatistics, University of Washington, Seattle, Washington, USA. [3]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA. [4]Present address: Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, Oregon, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to J.S. (shendure@uw.edu) or G.M.C. (gcooper@hudsonalpha.org).

Received 13 July 2013; accepted 13 January 2014; published online 2 February 2014; doi:10.1038/ng.2892

derived allele with a frequency of at least 95% (14.9 million SNVs and 1.7 million indels). Nearly all of these events are fully fixed in the human lineage, with fewer than 5% appearing as nearly fixed polymorphisms in the 1000 Genomes Project[14] variant catalog (derived allele frequency (DAF) ≥ 95%). To simulate an equivalent number of *de novo* mutations, we used an empirical model of sequence evolution with CpG dinucleotide–specific rates and mutation rates locally estimated on a 1-Mb scale (**Supplementary Note**). Mutation rate parameters as well as the size distribution for indels were estimated from six-way primate genome alignments[15].

To generate annotations, we used the Ensembl Variant Effect Predictor[16] (VEP), data from the ENCODE Project[11] and information from UCSC Genome Browser tracks[17] (**Supplementary Table 1**). Annotations spanned a range of data types, including conservation metrics such as GERP[8], phastCons[9] and phyloP[10]; regulatory information[11] such as genomic regions of DNase I hypersensitivity[18] and transcription factor binding[19]; transcript information such as distance to exon-intron boundaries or expression levels in commonly studied cell lines[11]; and protein-level scores such as those generated with Grantham[20], SIFT[7] and PolyPhen[6]. The resulting variant-by-annotation matrix contained 29.4 million variants (half fixed or nearly fixed human-derived alleles ('observed') and half simulated *de novo* mutations ('simulated')) and 63 distinct annotations, some of which were composites that summarized many underlying annotations (**Supplementary Tables 1** and **2**, and **Supplementary Note**).

We first assessed the validity of our general approach by constructing a series of univariate models that contrast observed and simulated variants using each of the 63 annotations as individual predictors (**Supplementary Note**). Nearly all models were highly predictive for distinguishing observed and simulated variants (**Supplementary Tables 3–5**) and were consistent with expectation. For example, we found a nearly 20-fold depletion of nonsense variants, a 2-fold depletion of missense variants and no depletion of intergenic or upstream or downstream variants (**Supplementary Table 6**). Nonsense and missense mutations that occurred near the start sites of coding DNA were more depleted than those occurring near the ends (**Supplementary Table 7**), and variants within 20, and especially within 2, nucleotides of splice junctions were also depleted (**Supplementary Fig. 1**).

The best-performing individual annotations were protein-level metrics such as PolyPhen[6] and SIFT[7], but these evaluated only missense variants (0.63% of all variants in the training data are missense; of these, 88% had defined PolyPhen values and 90% had defined SIFT values). Conservation metrics were the strongest individual genome-wide annotations (**Supplementary Table 3**).

We also examined correlations between annotations (**Supplementary Fig. 2**) and the value of adding interaction terms between annotations (**Supplementary Fig. 3**). Many annotations were correlated, and many interactions had area under the curve (AUC) values above 0.5, but only a handful of interacting pairs meaningfully improved a simple additive model. Overall, these analyses demonstrate that substantial biological differences are present between the observed and simulated variants with respect to the 63 annotations and that linear models capture much of this information.

We next trained a support vector machine[21] (SVM) with a linear kernel on features derived from the 63 annotations, supplemented by a limited number of interaction terms (**Supplementary Fig. 4**, **Supplementary Tables 1** and **2**, and **Supplementary Note**). Ten models, independently trained on observed variants and different samples of simulated variants, were highly correlated (all pairwise Spearman rank correlations > 0.99; **Supplementary Fig. 5**). An average of these models was applied to score all 8.6 billion possible SNVs of the human reference genome (GRCh37). To simplify interpretation in some contexts, we also defined phred-like[22] scores (scaled C scores) on the basis of the rank of the C score of each variant relative to all 8.6 billion possible SNVs, ranging from 1 to 99 (**Supplementary Note**). For example, substitutions with the highest 10% (10^{-1}) of all scores—that is, those least likely to be observed human alleles under our model—were assigned values of 10 or greater ('≥C10'), whereas variants in the highest 1% (10^{-2}), 0.1% (10^{-3}), etc. were assigned scores '≥C20', '≥C30', etc.

Genome-wide properties of C scores

We first calculated the proportion of all possible substitutions with a given scaled C score having specific functional consequences (**Fig. 1**

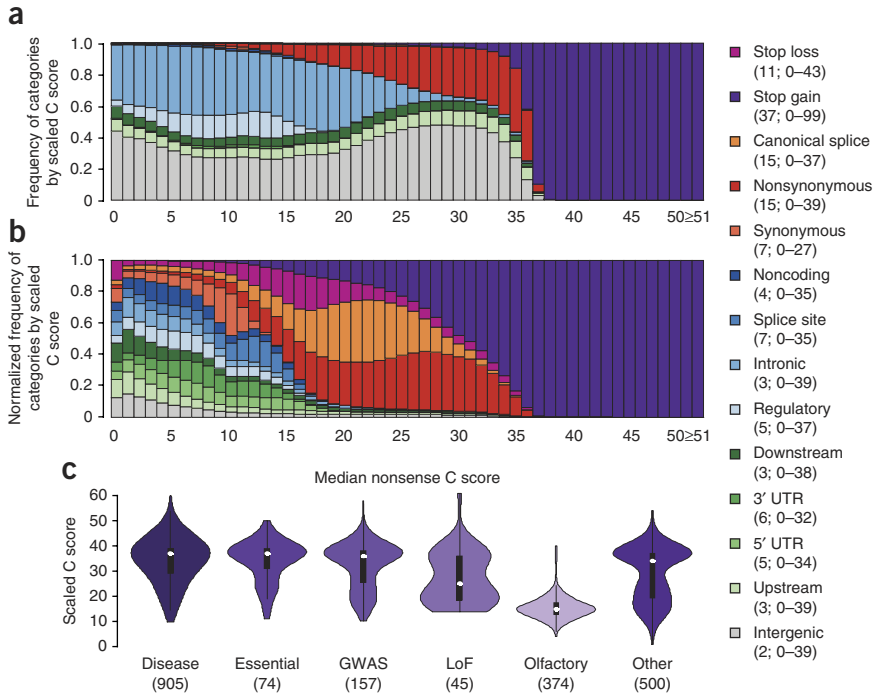

Figure 1 Relationship of scaled C scores and categorical variant consequences. (**a**) Proportion of substitutions with a specific consequence for each scaled C score bin. (**b**) Proportion of substitutions with a specific consequence after first normalizing by the total number of variants observed in that category. The legend includes in parentheses the median and range of scaled C score values for each category. Consequences were obtained from Ensembl VEP[16] (**Supplementary Note**); for example, noncoding refers to changes in annotated noncoding transcripts. Detailed counts of functional assignments in each C score bin are provided in **Supplementary Table 8**. (**c**) Violin plots of the median C scores of potential nonsense (stop-gain) variants for genes that harbor at least 5 known pathogenic mutations[48] (disease); are predicted to be essential[23]; harbor variants associated with complex traits[41] (GWAS); harbor at least 2 loss-of-function mutations in 1000 Genomes Project data[49] (LoF); encode olfactory receptor proteins; or are in a random selection of 500 genes (other; **Supplementary Note**).

and **Supplementary Table 8**). Although trained solely on differences between observed and simulated variants rather than on sets of known disease-causing variants that might introduce ascertainment bias, C scores were highest for potential nonsense variants (median of 37) and were next highest for missense and canonical splice-site variants (median of 15), whereas intergenic variants comprised the variants with the lowest C scores (median of 2). However, 76% of potential SNVs with C score of ≥20 were noncoding (falling into categories other than missense, nonsense, canonical splice site or stop loss), whereas 74% of potential missense and 18% of potential nonsense SNVs had C scores <20. Further, within each functional class, there were distinctions that are biologically relevant and are likely predictively useful. For example, potential nonsense variants—often treated as a homogeneous group in disease studies—in olfactory receptor genes had lower scores than variants in other genes, whereas potential nonsense variants in genes found previously to be essential[23] had higher scores (**Fig. 1**, bottom, and **Supplementary Fig. 6**). C scores thus capture a considerable amount of information, both in comparisons of functional categories and analysis within specific functional categories. Of note, these distinctions were absent or muted with other measures, either owing to missingness (for example, for missense-only measures) or lack of functional awareness (for example, conservation measures cannot distinguish between a nonsense and a missense allele at a given position).

We next compared scaled C scores with levels of genetic diversity, finding that C scores were negatively correlated with the DAFs of variants listed by the 1000 Genomes Project[14] or the Exome Sequencing Project[24] (ESP) (**Fig. 2a** and **Supplementary Figs. 7–9**), depletion of human genetic variation from the 1000 Genomes Project catalog (**Fig. 2b**) and depletion of chimpanzee-derived variants (**Fig. 2c**). Notably, these validation data sets had minimal overlap with the observed subset for the training data, which consisted only of fixed or nearly fixed (DAF > 95%) human-derived alleles. Furthermore, although we could not fully eliminate confounding by these factors, the negative correlation between C scores and DAFs for standing variation was robust to controlling for variation in background selection, local GC content, local CpG density and site-based conservation (**Supplementary Fig. 9**).

C scores of functional or pathogenic variants

We next sought to assess the usefulness of CADD in prioritizing functional and disease-relevant variation within five distinct contexts.

First, for *KMT2D* (*MLL2*), the gene mutated in Kabuki syndrome, C scores enabled the discrimination of a diverse set of disease-associated alleles[25] from rare, likely benign variants listed in ESP[24] (Wilcoxon rank-sum test $P = 9.9 \times 10^{-94}$; $n = 210$ disease associated/679 likely benign). Other metrics were markedly inferior in terms of accuracy or comprehensiveness (**Supplementary Fig. 10**).

Second, for *HBB*, the gene mutated in β-thalassemia, C scores of disease-associated alleles[26]—a set of indels ($n = 93$) and SNVs ($n = 119$) with regulatory/upstream ($n = 54$), splicing ($n = 37$), missense ($n = 22$), nonsense ($n = 18$) and other effects—were significantly and more strongly correlated with 3 levels of phenotypic severity than other measures (Kruskal-Wallis rank-sum test $P = 2.4 \times 10^{-7}$; $n = 48$ mild/65 intermediate/99 severe; **Supplementary Fig. 11**).

Third, pathogenic variants curated by the US National Institutes of Health (NIH) ClinVar database[27] were well separated from likely benign alleles (ESP[24] DAF ≥ 5%) matched to the same categorical consequences (Wilcoxon rank-sum test $P < 1 \times 10^{-300}$, $n = 8{,}174$ pathogenic/8,174 likely benign; **Fig. 3** and **Supplementary Figs. 12–16**). We note that there was substantial overlap between ClinVar and the training data underlying PolyPhen. When the corresponding sites were excluded from the test data set or when PolyPhen was excluded as a training feature from CADD, C scores continued to outperform all or nearly all missense-only metrics and conservation measures (**Supplementary Fig. 12**).

Fourth, C scores strongly correlated with the number of observed somatic cancer mutations in *TP53* (encoding p53) reported to the International Agency for Research on Cancer (IARC) (Spearman rank correlation = 0.38, $P = 6 \times 10^{-73}$, $n = 2{,}068$; **Supplementary Note**).

Fifth, we examined two enhancers[28] and one promoter[29] in which we previously performed saturation mutagenesis. C scores were significantly correlated with experimentally measured fold change in absolute expression from individual variants and were overall more significantly correlated than measures of sequence conservation

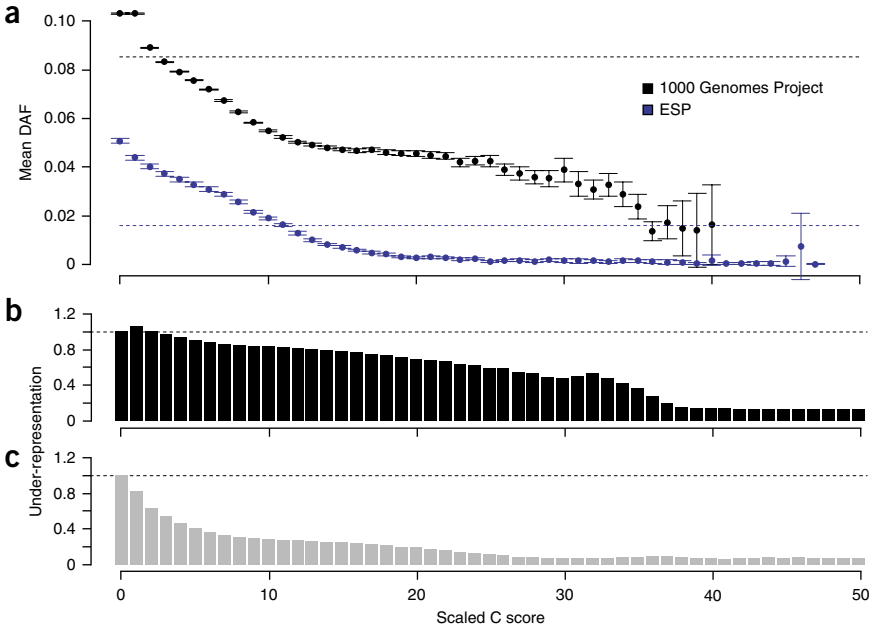

Figure 2 Relationship between scaled C scores and genetic variation. (**a**) Mean DAF by scaled C score for variants listed by the 1000 Genomes Project[14] or ESP[24]. Dashed lines indicate mean DAF values, and confidence intervals indicate 1.96 × s.e.m. for DAFs in each bin. (**b**) Under-representation of polymorphic sites in 1000 Genomes Project data. (**c**) Under-representation of chimpanzee lineage–derived variants. Under-representation is defined as the proportion of 1000 Genomes Project (**b**) or chimpanzee-derived (**c**) variants in a specific scaled C score bin divided by the frequency with which that scaled C score is observed for all possible mutations of the human reference assembly ($10^{\text{C score}/-10}$). The stronger under-representation of chimpanzee-derived variants relative to 1000 Genomes Project variants is expected given that the former are mostly fixed or high-frequency variants (and have survived many generations of purifying selection), whereas the latter are mostly low-frequency variants. Depletion values in **b**,**c** for C score bins other than 0 are significantly different from expectation (binomial proportion test, all $P < 1 \times 10^{-11}$).

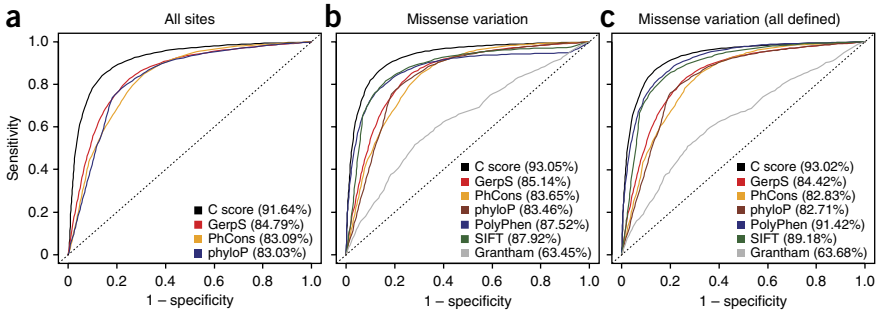

Figure 3 Sensitivity of methods in distinguishing pathogenic and benign variants. Receiver operating characteristics (ROCs) are shown discriminating curated, pathogenic mutations defined by the ClinVar database[27] from matched, likely benign ESP alleles (DAF ≥ 5%)[24] with the same categorical consequence. (**a**) Genome-wide variants for which GerpS, PhCons and phyloP scores are defined (n = 16,334). (**b**) Analysis limited to missense changes (n = 15,154), with missing values imputed to an upper limit of each score. (**c**) Analysis limited to missense changes for which PolyPhen, SIFT and Grantham scores are all defined (n = 13,358). Versions of the plot in **c** that exclude overlap between PolyPhen training data and the ClinVar database or use a CADD model trained without PolyPhen as a feature are shown in **Supplementary Figure 12**. Area under the curve (AUC) values are provided for each of the scores used.

(Spearman rank correlation of combined data = 0.31, $P = 1.9 \times 10^{-65}$, n = 2,847; **Supplementary Fig. 17**).

Collectively, these analyses demonstrate that CADD is quantitatively predictive of deleteriousness, pathogenicity and molecular functionality, both protein altering and regulatory, in a variety of experimental and disease contexts. In each of these contexts, the predictive usefulness of CADD was much better than measures of sequence conservation, the only comprehensive type of variant score, and also tended to be better, in most cases substantially so, than function-specific metrics when restricted to the appropriate variant subsets.

Application of CADD to human genetics

We next considered how CADD might be useful in evaluating candidate variation within exome or genome-wide studies.

First, we analyzed *de novo* exome variants (SNVs and indels) identified in children with autism spectrum disorders[30–34] (ASD) and intellectual disability[35,36] along with unaffected siblings or controls, considering 88 nonsense, 1,015 missense, 359 synonymous, 32 canonical splice-site and 150 other variants, including indels. Variants in affected children were significantly more deleterious than variants in unaffected siblings or controls when each disorder was considered separately (**Supplementary Table 9**) or in combination (ASD + intellectual disability Wilcoxon rank-sum test $P = 2.0 \times 10^{-4}$, n = 1,130 probands/514 controls). Additionally, *de novo* variants in probands with intellectual disability were significantly more deleterious than variants in probands with ASD ($P = 4.7 \times 10^{-5}$, n = 170 intellectual disability/960 ASD), suggesting a more deleterious global mutation burden in intellectual disability, which is consistent with the observation of increased sizes and numbers of copy number variants in intellectual disability relative to ASD[37].

Second, it is well established that annotations such as PolyPhen and conservation scores are valuable in the sequencing-based identification of disease-causal genes by virtue of their ability to highly rank pathogenic variants[1,2,38]. We therefore examined the distribution of C scores for variants in the genomes of 11 individuals representing diverse populations[39,40], finding that CADD highly ranked known disease-causal (ClinVar pathogenic) variants within the complete spectrum of variation in personal genomes (**Fig. 4**, **Supplementary Fig. 16** and **Supplementary Tables 10** and **11**). Furthermore, CADD was both more quantitative and more comprehensive in this task (for example, ~27% of pathogenic ClinVar SNVs were not scored by PolyPhen because of missing values or the restriction of PolyPhen to missense variation). Given its considerable superiority over the best available protein-based and conservation metrics in terms of ranking known pathogenic variants in the complete spectrum of variation within personal genomes, CADD will likely improve the power of sequence-based disease studies beyond that achieved with current standard approaches.

Finally, we analyzed CADD scores for SNPs identified by GWAS of complex traits, contrasting them with scores for nearby control

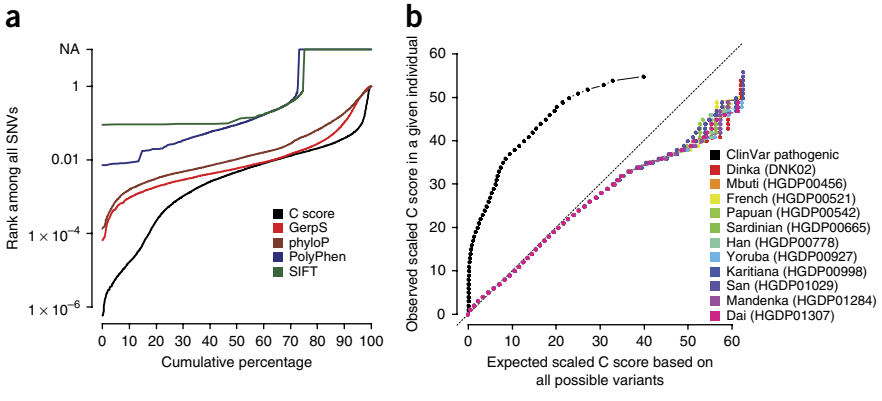

Figure 4 Ranking of pathogenic ClinVar variants among the variants identified by whole-genome sequencing in 11 human individuals from diverse populations. (**a**) Cumulative distribution of the rankings of 9,831 pathogenic ClinVar variants when 'spiked' into each of 11 personal genomes. For example, C scores of ~30% for ClinVar variants rank in the top 0.1% of all variants within a personal genome, and most rank in the top 1%. About 25% of pathogenic ClinVar SNVs are not scored by PolyPhen or SIFT because of missing values or the restriction of these methods to missense variation; note also that rankings for PolyPhen and SIFT are computed among missense variants only and are therefore derived from far fewer total variants (see a plot restricted to missense variation in **Supplementary Fig. 16**). (**b**) Quantile-quantile plot of C scores for the SNVs identified in the 11 individual genomes and pathogenic ClinVar SNVs. For a given scaled C score observed in an individual, the fraction of that individual's variants with a C score at least that high was computed (y axis). The C score corresponding to this quantile of the distribution of all possible variants is displayed on the x axis. High C scores are under-represented compared to the set of all possible variants. In contrast, known disease-causal variants from ClinVar have large C scores relative to the set of all possible variants. This fact can be exploited to prioritize causal variants identified from whole-genome sequencing of individual genomes as in **a** (see also **Supplementary Tables 10** and **11**).

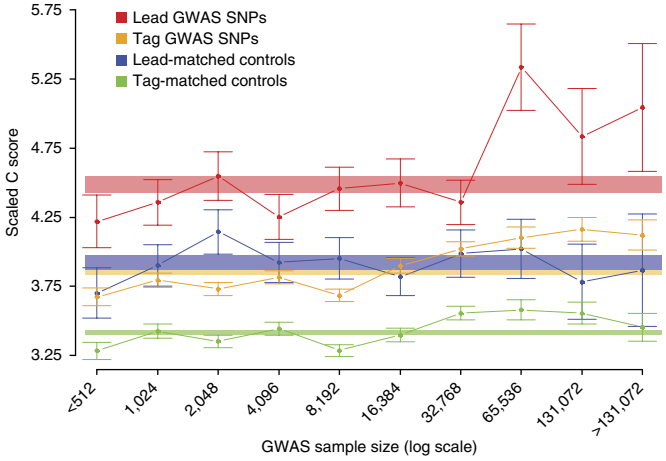

Figure 5 C scores for GWAS SNPs are higher than for nearby control SNPs and are dependent on study sample size. The average scaled C score (*y* axis) is plotted for each category of SNPs, as indicated by color, relative to the sample size of the association study in which the SNP was identified (*x* axis). Sample size bins are $\log_2$ scaled and mutually exclusive; for example, the bin labeled 1,024 represents all SNPs from studies with between 512 and 1,024 samples. Error bars, ±1 s.e.m. Each shaded rectangle represents overall (across all sample sizes) scaled C score mean ± 1 s.e.m. for each category as indicated by color.

SNPs matched for allele frequency and genotyping array availability (**Fig. 5** and **Supplementary Note**). We found that lead GWAS SNPs had significantly higher C scores than control SNPs (one-sided Wilcoxon rank-sum test $P = 1.3 \times 10^{-12}$, $n = 5{,}498$ GWAS/5,498 control); nearby SNPs in linkage disequilibrium with lead SNPs (tag SNPs) scored lower on average than lead SNPs but also had significantly higher scores than their matched controls ($P = 5.1 \times 10^{-107}$). Differences in C score remained significant after controlling for properties such as gene-body effect, gene expression level, conservation and regulatory element overlap; each of these properties was significantly different (all $P < 0.01$) for associated and control SNPs, but none could fully explain discrepancies in C score (**Supplementary Table 12** and **Supplementary Note**). C scores for trait-associated SNPs furthermore correlated with the size of the underlying association study and with the statistical significance of the association itself (**Fig. 5**, **Supplementary Fig. 18** and **Supplementary Note**), probably owing to the increased ability of larger studies and stronger association statistics to enrich for causal variants. Although for the most part not causal, GWAS-identified SNPs, especially strongly associated lead SNPs from large studies, were found by our analysis to be enriched for causal variants, consistent with previously observed GWAS enrichments for individual annotations[11,41–44].

DISCUSSION

With CADD, we describe a generic, expandable framework for integrating information contained in diverse annotations of genetic variation into a single score. We demonstrate that in a variety of contexts this approach is better, in some cases modestly but in many cases dramatically so, than other widely used annotations prioritizing functional and pathogenic variants. Further, beyond usefulness in any one setting, there are practical and conceptual advantages to CADD that should prove of major value to genetic studies of human disease. First, the information content of many individual annotations is objectively merged into a single value, which is far preferable to *ad hoc* approaches for combining annotations and is likely to improve performance, consistent with the benefits seen for consensus methods in missense mutation–specific annotation[45]. Second, CADD can readily incorporate expansions to existing annotations and entirely new annotations. This ability to indefinitely and readily integrate new information is crucial in light of projects such as ENCODE, which are continuously and rapidly expanding available annotations[11]. Third, CADD combines the generality of conservation-based metrics with the specificity of subset-relevant functional metrics (for example, PolyPhen), exploiting the advantages of both approaches while attenuating their respective disadvantages.

CADD also has a number of limitations that may restrict its usefulness in certain analyses or may represent areas for improvement. First, C scores measure reductions in variation, which correlate with deleteriousness but are also affected by the local mutation rate, background selection, biased gene conversion and other phenomena, potentially limiting accuracy. Second, C scores reflect the proportion of variants with a given annotation pattern that are visible to selection but may not capture differences in selective intensity; other approaches, such as polymorphism-to-divergence comparisons, may be more accurate for estimating selective coefficients[46]. Third, there is a great need for more gold-standard data, particularly for noncoding regions of the genome, the current paucity of which limits the development of better annotations as well as our ability to validate predictions. Fourth, it is at present not possible to precisely calibrate the relationship between CADD-estimated deleteriousness and the likelihood that a variant is pathogenic. As such, C scores are best interpreted in terms of likelihood of deleteriousness rather than likelihood of pathogenicity: for example, the quantifiable extent of depletion of a given C score from chimpanzee-derived alleles (**Fig. 2c** and **Supplementary Table 11**). Especially in discovering causal variants, CADD scores should be treated as one piece of information contributing to the totality of evidence for pathogenicity and should be evaluated as a supplement, not a replacement, for genetic information.

The one-stop nature of CADD is likely to be of great practical and conceptual value to future sequencing studies. It will minimize the scope and diversity of annotations that have to be generated, tracked and evaluated by a laboratory or project and will reduce the need for *ad hoc* combinations of filters, scores and parameters as is now routinely carried out. For example, a standard approach in exome studies is to merge missense (with or without an annotation of damaging or a given level of conservation), nonsense and splice-disrupting variants into a single, internally unranked list of protein-altering variants before genetic analysis[5]. With CADD, one might avoid arbitrary filters or thresholds altogether, including both coding and noncoding variants on a single, meaningfully ranked list. For example, a recent study of recessive, non-syndromic pancreatic agenesis identified five causal noncoding variants that disrupted the function of a distal enhancer of *PTF1A*[47]. C scores for these noncoding, disease-causal variants (scaled scores between 23.2 and 24.5) rank them higher than 99.5% of all possible human SNVs, higher than 97% of missense SNVs in a typical exome and higher than 56% of pathogenic SNVs in ClinVar[27].

Both in research and in the clinic, the capacity to define catalogs of genetic variants exceeds our ability to systematically evaluate their potential effects. This disparity will deepen as sequencing accelerates, genomes displace exomes and the array of functional categories and annotations expands. A unified, quantitative and scalable framework capable of exploiting many genomic annotations will be essential to meet the challenge posed. We anticipate that the model described here and the accompanying freely available precomputed scores for all possible GRCh37/hg19 SNVs (see URLs) will immediately be

broadly useful and will improve over time, enabling better interpretation of variants of uncertain significance in a clinical setting and improving discovery power for genetic studies of both mendelian and complex diseases.

URLs. CADD, http://cadd.gs.washington.edu/; Genome Variation Server (GVS), http://gvs.gs.washington.edu/GVS137/; CADD simulator, http://cadd.gs.washington.edu/simulator; GWAS catalog, http://www.genome.gov/gwastudies.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank P. Green and members of the Shendure laboratory for helpful discussions and suggestions. Our work was supported by US NIH grants U54HG006493 (to J.S. and G.M.C.), DP5OD009145 (to D.M.W.) and DP1HG007811 (to J.S.).

AUTHOR CONTRIBUTIONS

G.M.C. and J.S. designed the study. M.K. processed the annotation data and scores and developed and implemented the simulator and scripts required for scoring. P.J. and B.J.O. prepared and provided data sets and annotations. D.M.W. and M.K. developed the model and performed model training. D.M.W. performed the analysis of individual features and interactions. M.K., D.M.W., G.M.C. and J.S. analyzed the model's performance on different data sets. G.M.C. analyzed the GWAS data. J.S., G.M.C., M.K. and D.M.W. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Cooper, G.M. *et al.* Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods* **7**, 250–251 (2010).
2. Cooper, G.M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
3. Musunuru, K. *et al.* From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
4. Ward, L.D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–1106 (2012).
5. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
6. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
7. Ng, P.C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
8. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
9. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
10. 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).
11. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
12. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge and New York, 1983).
13. Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843 (2008).
14. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
15. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18**, 1814–1828 (2008).
16. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
17. Meyer, L.R. *et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.* **41**, D64–D69 (2013).
18. Boyle, A.P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
19. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
20. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
21. Franc, V. & Sonnenburg, S. Optimized cutting plane algorithm for large-scale risk minimization. *J. Mach. Learn. Res.* **10**, 2157–2192 (2009).
22. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
23. Liao, B.Y. & Zhang, J. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc. Natl. Acad. Sci. USA* **105**, 6987–6992 (2008).
24. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
25. Makrythanasis, P. *et al. MLL2* mutation detection in 86 patients with Kabuki syndrome: a genotype-phenotype study. *Clin. Genet.* doi:10.1111/cge.12081 (16 January 2013).
26. Giardine, B. *et al.* HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. *Hum. Mutat.* **28**, 206 (2007).
27. Baker, M. One-stop shop for disease genes. *Nature* **491**, 171 (2012).
28. Patwardhan, R.P. *et al.* Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nat. Biotechnol.* **30**, 265–270 (2012).
29. Patwardhan, R.P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
30. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nat. Genet.* **43**, 585–589 (2011).
31. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
32. Sanders, S.J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
33. Neale, B.M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
34. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
35. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
36. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
37. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
38. Ng, S.B. *et al.* Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–793 (2010).
39. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
40. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
41. 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).
42. 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).
43. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
44. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).
45. González-Pérez, A. & Lopez-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.* **88**, 440–449 (2011).
46. Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nat. Genet.* **45**, 723–729 (2013).
47. Weedon, M.N. *et al.* Recessive mutations in a distal *PTF1A* enhancer cause isolated pancreatic agenesis. *Nat. Genet.* **46**, 61–64 (2014).
48. Stenson, P.D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).
49. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).

ONLINE METHODS

Simulated and observed variants. The basis of the CADD framework is to capture correlates of selective constraint as manifested in differences between simulated variants and observed human-derived changes. For simulated variants, we developed a genome-wide simulator of *de novo* germline variation. The simulator was motivated by the parameters of the General Time Reversible (GTR) model[50], but, because the standard GTR model does not naturally accommodate asymmetric CpG-specific mutation rates, we used a fully empirical model of sequence evolution with a separate rate for CpG dinucleotides and local adjustment of mutation rates (**Supplementary Note**). Simulation parameters were obtained from Ensembl Enredo-Pecan-Ortheus (EPO)[13,15] whole-genome alignments of six primate species (Ensembl Compara release 66). A custom script and associated rate matrices underlying the genome-wide simulator have been made available (CADD simulator; see URLs). We applied these parameters to simulate SNV and indels on the basis of the human reference sequence (GRCh37).

For observed human-derived changes, we extracted sites where the human reference genome differed from the inferred human-chimp ancestral genome from the Ensembl EPO six-primate alignments defined above, excluding variants in the most recent data from the 1000 Genomes Project[14] (variant release 3 of 20110521) with a frequency of greater than 5% and including variants where the human reference carried an ancestral allele (i.e., matching the inferred human-chimpanzee ancestral sequence) but where the derived allele was observed with frequency of greater than 95% in 1000 Genomes Project data. We identified a total of 14,893,290 SNVs, 627,071 insertions and 1,107,414 deletions (less than 50 bp in length).

Variant annotation matrix. We used Ensembl VEP (Ensembl Gene annotation v68)[16] to obtain gene model annotation for single-nucleotide and indel variants. For SNVs within coding sequence, we also obtained SIFT[7] and PolyPhen-2 (ref. 6) scores from VEP. We combined output lines describing MotifFeatures with the other annotation lines, reformatted to a pure tabular format, reduced the different Consequence output values to 17 levels, and implemented a 4-level hierarchy in case of overlapping annotations (**Supplementary Note**). To the 6 VEP input-derived columns (chromosome, start, reference allele, alternative allele, variant type (SNV/INS/DEL) and length) and 26 actual VEP output-derived columns, we added 56 columns providing diverse annotations (for example, mappability scores and segmental duplication annotation as distributed by UCSC[51,52]; PhastCons and phyloP conservation scores[53] for 3 multi-species alignments[9] excluding the human reference sequence in score calculation; GERP++ single-nucleotides scores, element scores and *P* values[54], also defined from alignments with the human reference excluded; background selection scores[40,55]; expression values, acetylation at histone H3 lysine 27 (H3K27), methylation at histone H3 lysine 4 (H3K4), trimethylation at H3K4, nucleosome occupancy and open chromatin tracks provided for ENCODE cell lines in the UCSC super tracks[52]; genomic segment type assignment from Segway[56]; predicted transcription factor binding sites and motifs[11]; and overlapping ENCODE chromatin immunoprecipitation and sequencing (ChIP-seq) transcription factors[11], 1000 Genome Project[14] and ESP[57] variant status and frequencies and Grantham scores[20] associated with a reported amino acid substitution). A full description is provided in the **Supplementary Note**, and all columns of the obtained annotation matrix are listed in **Supplementary Table 1**.

Imputation and final training data set. In the annotations described above, some columns were not useful for model training or needed to be excluded from training as they differed between the simulated variants and the human-chimpanzee ancestor differences for technical reasons (see the **Supplementary Note** for a complete list; note that no allele frequency information was used in model training). We imputed missing values in genome-wide measures by the genome average obtained from the simulated data or set missing values to 0 where appropriate (**Supplementary Table 2**). Further, we created an "undefined" category for categorical annotations to accommodate missing values. To deal with missing values in annotations that were not defined for a subset of variants (for example, information only available for protein-coding genes), we set the missing values to 0 and also created indicator variables that contained a 1 if the corresponding variant was undefined and a 0 otherwise. Because insertions and deletions may produce arbitrary length Ref/Alt and reference (oAA)/variant (nAA) amino acid sequence columns (and thus not a fixed number of categorical levels), these values were set to "*N*" for Ref/Alt and to "undefined" for nAA/oAA.

Sites from the simulation were labeled +1, and human-derived variants were labeled −1. Only insertions and deletions shorter than 50 bp were considered for model training, and the length column was capped at 49 bp for the prediction of longer events. The ratio of indel events to SNV events obtained for the simulation was 1:8.46.

Model training. We generated 10 training data sets by sampling an equal number of 13,141,299 SNVs, 627,071 insertions and 926,968 deletions from both the simulated variant and observed variant data sets. To train each SVM model, processed data were converted to a sparse matrix representation after converting all *n*-level categorical values to *n* individual Boolean flags. We randomly selected 1% of sites (~132,000 SNVs, 6,000 insertions and 9,000 deletions) for a test data set. All other sites were used to train linear SVMs using the LIBOCAS v0.96 library[21]. The SVM model fits a hyperplane as defined below. $X_1, \ldots, X_n$ represent the 63 annotations described above (which expand to 166 features owing to the treatment of categorical annotations), $W_1,\ldots,W_{11}$ represent the Boolean features that indicate whether a given feature (out of cDNApos, relcDNApos, CDSpos, relCDSpos, protPos, relProtPos, Grantham, PolyPhenVal and SIFTval, as well as Dst2Splice ACCEPTOR and DONOR) is undefined, $1_{\{A\}}$ is an indicator variable for whether event *A* holds, and *D* is the set of bStatistic, cDNApos, CDSpos, Dst2Splice, GerpN, GerpS, mamPhCons, mamPhyloP, minDistTSE, minDistTSS, priPhCons, priPhyloP, protPos, relcDNApos, relCDSpos, relProtPos, verPhCons and verPhyloP. Because of the coding of categorical values using Boolean variables, the total number of features in this model is 949.

$$
\begin{aligned}
0 \;=\; & \beta_0 + \sum_{i=1}^{166} \beta_i X_i + \sum_{i=1}^{5}\sum_{j=1}^{5} \gamma_{ij} 1_{\{i\text{th Ref category and } j\text{th Alt category}\}} \\
& + \sum_{i=1}^{21}\sum_{j=1}^{21} \delta_{ij} 1_{\{i\text{th oAA category and } j\text{th nAA category}\}} \\
& + \sum_{i=1}^{11} \tau_i W_i + \sum_{i=1}^{17} \sum_{j \in \boldsymbol{D}} \alpha_{ij} 1_{\{i\text{th Consequence category}\}} X_j
\end{aligned}
$$

SVM models were trained, using various values for the generalization parameter (*C*), which assigns the cost of misclassifications. Model training convergence in 2,000 iterations (~70 h) for different settings of *C* is shown in **Supplementary Figure 4**. These results indicate that model training only converges within a reasonable amount of time for *C* values around 0.0025 and below. We therefore trained models for all ten training data sets with $C = 0.0025$. We determined the average of the model parameters and used the average model.

Model testing and validation. We annotated all 8.6 billion possible substitutions in the human reference genome (GRCh37) and applied the model to score all possible substitutions. When scoring sites with multiple VEP annotation lines, we scored all possible annotations first and report the one with the highest deleteriousness after applying the four hierarchy levels. We mapped C scores to a Phred-like scale (scaled C scores) ranging from 1 to 99 on the basis of their ranking relative to all possible substitutions in the human reference genome, i.e., $-10\log_{10}$ (rank/total number of substitutions).

We used several data sets extracted from the literature and public databases to examine the performance of model scores (see the **Supplementary Note** for details). (i) We determined C scores for specific gene classes motivated by the analysis performed by Khurana *et al.*[58] (i.e., HGMD[48], non-immune essential genes described by Liao *et al.*[23], GWAS genes as available from the catalog of published Genome-Wide Association Studies (GWAS catalog; see URLs), loss-of-function genes from MacArthur *et al.*[49] and olfactory genes from the Ensembl 68 gene build). (ii) We downloaded 210 mutations in *KMT2D* (*MLL2*) associated with Kabuki syndrome from Makrythanasis *et al.*[25]. We complemented these with 679 putatively benign variants observed in ESP[57]. (iii) We downloaded a total of 119 SNVs, 30 insertions and 63 deletions

doi:10.1038/ng.2892

(all required to be at most 50 bp in length) within or near *HBB* that gave rise to α-thalassemia from HbVar[26]. Disease categories were used as defined by HbVar, except that all types that were not "beta0" or "beta+" were pooled into one category, "other." (iv) We obtained the NCBI ClinVar[27] data set (released 16 June 2012) and extracted variants that were marked "pathogenic" or "non-pathogenic (benign)." We also selected a set of apparently benign variants (allele frequency ≥ 5%) from ESP that were matched to the pathogenic ClinVar sites in terms of their consequence annotations. In addition, we generated a data set where we matched ESP and ClinVar frequencies to three decimal places of the alternative allele frequency. Because of the overlap of ClinVar and ESP variants with the PolyPhen training data set, we trained a separate classifier without the PolyPhen features, and we also examined performance on the subset of ClinVar and ESP variants not used for PolyPhen training. To compare the performance of CADD with that of other publically available missense annotations not used in model training, we downloaded scores from dbNSFP 2.0 (ref. 59). (v) We combined high-confidence *de novo* mutations from 5 family-based autism exome sequencing studies[30–34], including a total of 948 probands with ASD and 590 unaffected siblings. Further, we obtained coding variants as described above for 2 family-based intellectual disability studies[35,36], including 151 families with intellectual disability and 20 unrelated control families. (vi) We obtained data on fold change in expression for each base substitution in *ALDOB* and *ECR11* enhancers from Patwardhan *et al.*[28]. This data set contained a total of 777 variants for *ALDOB* and 1,860 variants for *ECR11*. Further, we obtained the *HBB* promoter data of Patwardhan *et al.*[29]. The promoter data set contained a total of 210 variants associated with a fold change in expression. (vii) We obtained a list of 23,788 single-nucleotide somatic cancer mutations in *TP53* that were reported to IARC. These mutations correspond to 2,068 distinct variants; we recorded the number of times that each variant was reported. (viii) We obtained GATK VCF variant call files for all autosomes and the X chromosome from shotgun sequencing of 11 men originating from diverse human populations[40]. (ix) We accessed the National Human Genome Research Institute (NHGRI) GWAS catalog on 18 December 2012 and obtained 9,977 distinct SNP-trait associations spanning 7,531 unique SNPs in 1000 Genomes Project data; these variants are referred to as "lead SNPs." We used the Genome Variation Server (GVS; see URLs) to find all SNPs within 100 kb of a lead SNP that had a pairwise correlation of $r^2 \geq 0.8$ within Utah residents with ancestry from northern and western Europe (CEU). This resulted in the identification of an additional 56,538 unique SNPs, referred to as "tag SNPs." We also developed "control" SNP sets, selected to match trait-associated SNPs for a variety of features that may bias SNPs found by GWAS in the absence of any causal effect.

50. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17**, 57–86 (1986).
51. Fujita, P.A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* **39**, D876–D882 (2011).
52. Rosenbloom, K.R. *et al.* ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.* **40**, D912–D917 (2012).
53. Hubisz, M.J., Pollard, K.S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51 (2011).
54. 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).
55. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
56. Hoffman, M.M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).
57. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
58. Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput. Biol.* **9**, e1002886 (2013).
59. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* **32**, 894–899 (2011).