HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2017 September 13.

Published in final edited form as:

Nat Genet. 2017 April ; 49(4): 618–624. doi:10.1038/ng.3810.

Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data

Yi-Fei Huang[1], **Brad Gulko**[1,2], and **Adam Siepel**[1,*]

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

[2]Graduate Field of Computer Science, Cornell University, Ithaca, NY 14853, USA

Abstract

Many genetic variants that influence phenotypes of interest are located outside of protein-coding genes, yet existing methods for identifying such variants have poor predictive power. Here, we introduce a new computational method, called LINSIGHT, that substantially improves the prediction of noncoding nucleotide sites at which mutations are likely to have deleterious fitness consequences, and which therefore are likely to be phenotypically important. LINSIGHT combines a generalized linear model for functional genomic data with a probabilistic model of molecular evolution. The method is fast and highly scalable, enabling it to exploit the "Big Data" available in modern genomics. We show that LINSIGHT outperforms the best available methods in identifying human noncoding variants associated with inherited diseases. In addition, we apply LINSIGHT to an atlas of human enhancers and show that the fitness consequences at enhancers depend on cell type, tissue specificity, and constraints at associated promoters.

Introduction

In the human genome, most nucleotides that are associated with diseases or other phenotypes, or that show signatures of natural selection, fall outside of protein-coding genes[1-3]. Many of these nucleotides appear to fall in *cis*-regulatory elements, including promoters, enhancers, and insulators. Similar observations hold across most animals and plants[4-7]. Recent efforts to characterize noncoding sequences using high-throughput biochemical assays have produced a wealth of data, identified many regulatory elements, and clarified general aspects of gene regulation[8-12]. Nevertheless, a substantial gap remains between the outcomes of these experiments and a detailed understanding of noncoding

*Correspondence should be addressed to A.S. (asiepel@cshl.edu).

URLs. LINSIGHT program, https://github.com/CshlSiepelLab/LINSIGHT; UCSC Genome Browser, http://genome.ucsc.edu/; Cold Spring Harbor Laboratory Mirror of UCSC Genome Browser, http://genome-mirror.cshl.edu/; Complete Genomics human variation data, http://www.completegenomics.com/public-data/69-Genomes/; SPIDEX database, http://www.deepgenomics.com/spidex/.

Author Contributions.

YFH and AS conceived and designed the study. YFH designed and implemented the LINSIGHT method. YFH and BG analyzed the data. AS supervised the research. YFH and AS wrote the manuscript with review and feedback from BG.

Competing Financial Interests. The authors declare no competing financial interests.

function, for several reasons. First, these assays generally measure genomic and epigenomic features roughly correlated with, but not directly indicative of, regulatory function. Second, they generally have relatively low resolution along the genome, identifying regions hundreds of nucleotides long, rather than pinpointing single nucleotides. Third, these measures are highly condition-specific, and data have only been generated for a small subset of cell types and conditions.

As a consequence, there is a pressing need for computational methods that more precisely predict regulatory function by jointly considering the results of numerous such assays together with complementary data, such as annotations of protein-coding genes and measures of evolutionary conservation across species. The development of statistical and machine-learning methods that attempt to address this integrative prediction challenge has emerged as an active, fast-moving area of research. Recently published methods in this area can be roughly divided intothree categories: (1) machine-learning classifiers that attempt to separate known disease variants from putatively benign variants using a variety of genomic features (e.g., GWAVA[13] and FATHMM-MKL[14]); (2) sequence- and motif-based predictors for the impact of noncoding variants on cell-type-specific molecular phenotypes, such as chromatin accessibility or histone modifications (e.g., DeepBind[15], DeepSEA[16] and Basset[17]); and (3) evolutionary methods that consider data on genetic variation together with functional genomic data and aim to predict the effects of noncoding variants on fitness (e.g., CADD[18], DANN[19], FunSeq2[20], and fitCons[3]). A limitation of methods of the first class is that they depend strongly on the available training data, which may be limited and may not be representative of the broader class of regulatory sequences of interest. Methods of the second class have the limitation that the significance of molecular phenotypes at the organismal level is often unclear. Evolutionary methods, by contrast, obtain their signal not primarily from previously assigned class labels, but instead from signatures of natural selection over many generations. They are therefore both less data limited, and more focused on phenotypes that truly influence fitness, than the other methods. This approach is likely to be particularly powerful for detecting regulatory variants that tend to be under strong purifying selection, such as rare variants associated with severe diseases. Evolution-based methods also naturally integrate over cell types, an important strength when the relevant tissue- or cell-types for a condition of interest are unknown.

Among the available evolution-based methods, fitCons[3] is unique in explicitly characterizing the influence of natural selection at each genomic site of interest using a full probabilistic evolutionary model and patterns of genetic variation within and between species. FitCons makes a distinction between functional genomic and comparative genomic data, first defining several hundred clusters of genomic positions with distinct functional genomic “signatures,” and then estimating the fraction of nucleotides under natural selection within each cluster from polymorphism and divergence data. These estimates are obtained using the INSIGHT evolutionary model[21,22], and are interpreted as the probabilities that mutations in each cluster of genomic sites will have fitness consequences (fitCons scores). In this manner, fitCons aggregates information about natural selection from large numbers of sites with similar functional profiles based on evolutionary first principles. A major limitation of the method, however, is that it scales poorly with the available functional genomic data. In particular, the number of clusters considered by the method increases exponentially with the

number of functional genomic annotations, which keeps it from taking advantage of the growing body of functional genomic data. A related problem is that the restriction to small numbers of genomic features leads to a relatively coarse-grained, blocky pattern of scores along the genome, which does not allow for fine distinctions among nearby nucleotide sites.

In this paper, we describe a new method, *Linear* INSIGHT (LINSIGHT; pronounced *lin-site*), that is based on the existing INSIGHT/fitCons framework but has vastly improved speed, scalability, genomic resolution, and prediction power. The main idea behind LINSIGHT is to bypass the clustering step of fitCons and instead couple the probabilistic INSIGHT model directly to a generalized linear model for genomic features. This strategy results in a more streamlined model that scales linearly, rather than exponentially, with the available data, and can make direct use of the input data, with no need for discretization. By integrating a large number of genomic features, LINSIGHT provides a precise, high-resolution description of the fitness consequences of noncoding mutations in the human genome. We demonstrate that LINSIGHT outperforms state-of-the-art prediction methods in the task of prioritizing noncoding disease variants from the Human Gene Mutation database (HGMD)[23] and the NCBI ClinVar database[24]. Furthermore, we use LINSIGHT to show that the evolutionary constraints on human enhancers depend on their associated tissue types, degree of tissue specificity, and associated promoters, which has important implications for understanding the evolution of *cis*-regulatory elements and for improving variant prioritization methods. Our LINSIGHT scores are available as a track on the Cold Spring Harbor Laboratory mirror of the UCSC Genome Browser (hg19 assembly). The LINSIGHT software is available from our public laboratory GitHub repository.

Results

LINSIGHT combines INSIGHT with a scalable linear model

The original INSIGHT and fitCons methods[3,21,22] infer the selective pressure on noncoding sites, and hence the likely fitness consequences of noncoding mutations, by contrasting patterns of genetic variation at each focal site with the patterns at nearby genomic regions that are likely to be free from the influence of selection (“neutrally evolving sites”). To address the problem that genetic variation within species and between closely related species (such as the human and chimpanzee) are sparse across the genome, fitCons pools information across the thousands of genomic sites assigned to each discrete cluster.

The key idea behind LINSIGHT is instead to accomplish this pooling of information across sites indirectly, using a generalized linear model (Figure 1 and Table 1; see Supplementary Note and Supplementary Tables 1 and 2 for complete details). In particular, the parameters of the INSIGHT model that describe natural selection (ρ and γ) are determined as linear-sigmoid functions of the genomic features local to each site. (The third selection parameter from INSIGHT, η, is omitted because positive selection has a negligible effect in this setting; see Supplementary Note.) Thus, the probability of fitness consequences of mutations at each site i, denoted ρ_i, is assumed to depend on genomic features at that site, such as its RNA expression level (RNA-seq read depth), chromatin accessibility (DNase-I hypersensitive sites), and histone modifications or bound transcription factors (ChIP-seq peaks), as well as on features based on annotations (e.g., distance to nearest transcription start site, match to

known TFBS motif) and comparative genomics (e.g., phyloP[25] or phastCons[4] scores). We refer to ρ_i as the **LINSIGHT score** at site *i*. This scoring strategy has several major advantages: it requires no clustering and no discretization, and it scales linearly with the available genomic features, allowing hundreds of features to be considered. In contrast to fitCons, the scalability of the method enables data to be pooled across cell types, and it allows the scores to reach single-nucleotide resolution along the genome. Nevertheless, LINSIGHT continues to benefit from the advantages of the probabilistic INSIGHT model of molecular evolution.

All parameters of the LINSIGHT model are estimated simultaneously from genome-wide data by maximum likelihood using an online stochastic gradient descent algorithm (Methods). The gradients for the feature weights are efficiently computed by the back-propagation method widely used in neural network training[26]. Indeed, the model can be considered a type of neural network, albeit one without hidden layers. Its main disadvantage relative to fitCons—the assumption of an additive, linear relationship between features and selection parameters—could be addressed by adding hidden layers to the neural network, although we have found its performance to be excellent without this extension. Notably, the amount of data available for training is large in comparison to the number of free parameters and we have not yet found regularization to be necessary, but it could easily be added if necessary.

LINSIGHT scores across the human genome are generally consistent with, but often improve on, previous measures of evolutionary conservation

We applied LINSIGHT to a large public data set consisting of complete genome sequences for multiple human individuals and nonhuman primates, comparative genomic data for mammals and vertebrates, and a wide variety of functional genomic data, and we generated LINSIGHT scores for all positions across human reference genome (Methods). We considered a total of 48 genomic features, falling in three general classes: conservation scores, predicted binding sites, and regional annotations (Table 2 and Supplementary Table 3).

The distributions of INSIGHT scores in annotated regions of the noncoding genome are generally consistent with previous observations based on conservation scores[1,4,25]. For example, splice sites are very highly constrained (median LINSIGHT score of 0.956, indicating a 95.6% probability of fitness consequences due to mutations at these nucleotide sites), whereas annotated TFBSs show reduced, but still substantial, constraint (median score of 0.240 for TFBSs shared across species, median score of 0.106 for all TFBSs from the Ensembl Regulatory Build[27]; Figure 2a). Other promoter regions (median score of 0.073) and untranslated regions (UTRs; median scores of 0.128 and 0.076 for 5' and 3' UTRs, respectively) are somewhat less constrained, and unannotated intronic and intergenic regions exhibit the least constraint (median scores of 0.044–0.048). As observed previously, 5' UTRs show somewhat more constraint than 3' UTRs, although both types of UTRs contain subsets of sites subject to strong selection (LINSIGHT score > 0.8)[4,25]. The estimate for the more conserved TFBSs (0.240) is roughly similar to, if slightly lower than, previous estimates directly obtained from experimentally defined TFBSs (~30-40% of sites under selection[22,28]), despite that it was obtained indirectly in this case via the generalized linear model. The genome-wide average of the LINSIGHT scores is about 0.07, suggesting that about

7% of noncoding sites are under evolutionary constraint, consistent with numerous previous studies[3,4,29-31].

Across all noncoding positions in the genome, the LINSIGHT scores are fairly well correlated with those from other recently published methods particularly within conserved elements, which are enriched for regulatory function (see Supplementary Note and Supplementary Figure 1). On the task of identifying likely regulatory elements, the methods that make use of functional genomic data generally perform better than pure conservation methods, and LINSIGHT is among the best at this task (see Supplementary Note). For example, LINSIGHT has good power to identify transcription factor binding sites from the ORegAnno database[32] (AUC = 0.926), outperformed only by the DeepSEA functional significance score (AUC = 0.965) and FunSeq2 (AUC = 0.950) (Supplementary Figure 2). Thus, despite that it relies on an evolutionary objective function, LINSIGHT maintains good performance in the prediction of regulatory elements.

Consistent with these general trends, LINSIGHT highlights many of the regions identified by conservation methods such as phastCons[4], phyloP[25], and GERP++[33], but also identifies some regions that have relatively low conservation scores yet are likely to have important biological functions. An example is HGMD variant CR065653 in a putative enhancer, associated with upregulation of the telomerase reverse transcriptase (*TERT*) gene, which obtains an elevated LINSIGHT score, but is not identified by phastCons, phyloP, or GERP++ as being under constraint (Figure 2b). This example also demonstrates that the genomic resolution of the LINSIGHT scores is dramatically better than that of fitCons, and approaches the nucleotide resolution of phyloP and GERP++. LINSIGHT can identify functional variants not only in enhancers but also in promoter regions (Supplementary Figure 3a) and associated with splicing (Supplementary Figure 3b). Thus, it is useful as a general predictor of functional noncoding sites under evolutionary constraint.

LINSIGHT accurately identifies disease-associated variants in noncoding regions

We tested the ability of LINSIGHT to identify noncoding nucleotide positions that are associated with inherited human diseases, using the HGMD[23] and ClinVar[24] databases to define positive examples, and common polymorphisms (MAF > 1%), which are unlikely to be functionally important, to define negative examples. For comparison, we evaluated the CADD[18], Eigen[34], DeepSEA[16], FunSeq2[20], GWAVA[13], and phyloP[25] methods on the same task. For each scoring method, we computed false positive vs. true positive rates for the complete range of score thresholds, displaying the results as receiver operating characteristic (ROC) curves and measuring prediction power by the area-under-the-curve (AUC) statistic. Because the results of these tests can be highly sensitive to the criteria for selecting negative examples, we considered three schemes of increasing stringency (following ref. [13]): a random sample of negative examples (unmatched), negative examples matched by distance to the nearest transcription start site (matched TSS), and negative examples matched by specific genomic region (matched region; see Methods for details). In all cases, equal numbers of positive and negative examples were considered.

Overall, LINSIGHT outperformed all other methods in all comparisons (Figure 3). Its absolute prediction power varied across matching schemes in a predictable manner, being highest in

the unmatched comparison (e.g., AUC = 0.897 for HGMD) and decreasing in the matched TSS (AUC = 0.759) and matched region (AUC = 0.661) comparisons. The same effect also occurred for most other methods, but the methods that make heavier use of regional information, such as FunSeq2, suffered more as the matching stringency increased. These observations highlight the difficulty of distinguishing functional sites from nearby nonfunctional sites, which is considerably harder than separating regions enriched in functional sites from the genomic background. Nevertheless, LINSIGHT has some power for this challenging task. In almost all cases, the AUCs were considerably higher for ClinVar than for HGMD, apparently because ClinVar is heavily enriched for variants in splice sites, which are relatively easy to identify (Supplementary Figure 4). An exception to this rule was GWAVA, which performs exceptionally well on HGMD (cross-validation AUCs of 0.71–0.97)[13] and much more poorly on ClinVar (AUCs of 0.734–0.884), but GWAVA was trained using HGMD[13] and its performance on that data set appears to reflect overfitting (it is not shown in the HGMD ROC plots for this reason). This dependency on the training set for GWAVA demonstrates one of the pitfalls of pure classification strategies, and highlights a strength of the evolution-based strategy, which does not require a training set. Nevertheless, phyloP performs quite poorly on the HGMD data set, showing that scores based exclusively on evolution are of limited usefulness in this task.

The performance advantage of LINSIGHT was maintained when performance was measured using precision-recall curves in place of standard ROC curves (Supplementary Figure 5) and when rare variants were used in place of common variants as negative examples (Supplementary Figures 6 & 7). These performance advantages are statistically significant in most cases, with a few exceptions mostly stemming from the small size of the ClinVar data set (Supplementary Tables 4 and 5). In addition, a more detailed comparison with CADD showed that training CADD's logistic regression model using LINSIGHT's features resulted in improved performance, but not enough to make it competitive with LINSIGHT (Supplementary Table 6). Thus, the excellent performance of INSIGHT in these tests appears to derive both from its use of a broad collection of informative features along the genome and its probabilistic model of evolution.

To gain insight into which genomic features were most informative, we systematically omitted groups of related features and reassessed the prediction performance of LINSIGHT (Supplementary Note). Briefly, we found that regional features, such as ChIP-seq peaks and DNase-I hypersensitive sites (Table 2), were broadly useful in distinguishing genomic regions enriched for functional variants from the genomic background, but conservation scores were most important in separating functional sites from nearby nonfunctional sites (Supplementary Figure 8). Predicted binding sites were most informative in promoter regions.

The evolutionary constraints on enhancers are context-dependent

LINSIGHT is also potentially useful for studying the influence of natural selection on noncoding sequences. Compared with other measures of selection, LINSIGHT has the advantages of considering both functional genomic and population genomic data, of detecting the influence of selection on relatively recent time scales (e.g., since the human/

chimpanzee divergence), and of providing a model-based, easily interpretable measure of fitness consequences. With these advantages in mind, we used LINSIGHT to gauge the degree of evolutionary constraint on enhancers in the human genome, considering in particular the relationships of constraint with the number and type of active cell types, and with constraint at the target promoter of each enhancer. We analyzed nearly 30,000 enhancers (median length 293 bp) from a recently published atlas of active enhancers in dozens of human cell types and tissues, which were identified based on their transcriptional signatures[35]. This approach of annotating enhancers based on enhancer-associated RNAs (eRNAs) has been shown to identify elements having active roles in gene regulation in a cell-type-specific fashion with high genomic resolution[35-37].

First, we examined the relationship between the LINSIGHT scores and the number of cell types in which each enhancer is active. We found that the LINSIGHT scores were significantly positively correlated with the number of active cell types (Spearman's $\rho = 0.284$, $p < 10^{-15}$; Figure 4a), indicating that a broader spectrum of activity across cell types is associated with stronger purifying selection. To ensure that this observation reflected real differences in selective pressure and not simply correlations with the epigenomic features considered by LINSIGHT, we retrained LINSIGHT using only conservation scores and predicted binding sites and obtained essentially identical results (Supplementary Figure 9a). Furthermore, a partial correlation test indicated that the LINSIGHT scores were still strongly correlated with the number of cell types when controlling for eRNA expression level averaged across all FANTOM5 libraries (partial Spearman's $\rho = 0.24$; $p < 10^{-15}$). These findings parallel similar findings for protein-coding genes[38-40] and TFBSs[22] and likely reflect a general correlation between pleiotropy and constraint (see Discussion).

Second, we examined the relationship between the LINSIGHT score and the tissue type in which each enhancer is active, focusing on enhancers active in a single tissue type. We found that tissue-specific enhancers associated with sensory perception (olfactory region and parotid gland), the immune system (lymph node), digestion (stomach), and male reproduction (penis and testis) had the lowest LINSIGHT scores, whereas tissue-specific enhancers associated with tissues such as smooth muscle, the skin, and the urinary tract and bladder had the highest LINSIGHT scores (Supplementary Figure 10). These findings are also broadly consistent with findings for protein-coding genes, which have indicated that sensory, immune, dietary, and male reproductive genes are associated with relaxation of constraint and/or positive selection[40,41]. Interestingly, enhancers active in tissues associated with female reproduction (e.g., uterus, female gonad, and vagina) appeared to be under substantially more constraint than those active in tissues associated with male reproduction. Finally, we compared the LINSIGHT scores at enhancer/promoter pairs predicted from co-expression across tissues[35]. The LINSIGHT scores for these paired enhancers and promoters are weakly but significantly correlated (Figure 4b and Supplementary Figure 9b), indicating that the same types of evolutionary pressures tend to act at both members of each pair. Together, these results indicate that the evolutionary constraints on enhancers are dependent on several factors, including their degree of tissue specificity, the particular tissues in which they are active, and the evolutionary constraints associated with their target promoters.

Discussion

As sequencing costs fall and appreciation for regulatory variation grows, whole genome sequencing is rapidly supplanting exome sequencing as the primary technique for identifying and characterizing genetic variants that have phenotypic consequences. Hence, there is an increasing need for computational methods that can effectively prioritize noncoding variants based on their likelihood of phenotypic importance. In this paper, we address this problem with a new computational method, called LINSIGHT, that combines the evolutionary model of our previously developed INSIGHT method with a generalized linear model for functional genomic data and genome annotations, resulting in substantially improved scalability, resolution, and power. We have generated LINSIGHT scores across the human genome, making use of a large collection of publicly available population, comparative, and functional genomic data, and we find the scores to be consistent with previously available scores in many respects, but to improve on them in others. In particular, on the task of identifying human disease-associated variants from the HGMD and ClinVar databases, LINSIGHT offered the best performance of several methods we tested, across a range of types of variants and test designs. Importantly, LINSIGHT requires no training set of known regulatory or disease variants and therefore is expected to have better generalization properties than “supervised” machine-learning classifiers (see Introduction).

In conceptual terms, the new LINSIGHT method is closely related to our previous fitCons method[3], with the primary difference being that LINSIGHT pools data across sites implicitly through the use of its generalized linear model, whereas fitCons pools data by explicitly clustering sites according to discretized functional genomic signatures. In effect, LINSIGHT trades the restrictions of a linearity assumption for the benefits of computational speed, a reduced parameterization, and scalability to very large numbers of genomic features. Notably, the new model design also has a number of important side benefits. First, it avoids the need for discretization of the genomic features. In addition, as the number of features grows larger, the genomic resolution of the scores naturally becomes much finer, approaching the nucleotide-level resolution of conservation scores. Finally, the generalized linear model can readily be extended to a “deep” neural network through the addition of hidden layers. While it remains to be seen how much this extension will help in practice, in principle it can capture the types of nonlinearity and interactions between features that have been observed in this setting (for examples, see references [3] and [42]).

Our approach to characterizing noncoding variants is based on the premise that natural selection in the past, at individual nucleotide sites, provides useful information about phenotypic importance in the present. This assumption clearly will not hold in all cases. For example, variants that increase the risk for post-reproductive diseases or that influence phenotypes dependent on the modern human environment will not necessarily show signs of historical purifying selection. In addition, traits dependent on highly epistatic loci or on the aggregate contributions of large numbers of loci may have difficult-to-detect marginal contributions to fitness at individual nucleotides. Nevertheless, our results indicate that the evolution-based approach is useful for many phenotypes of interest. Furthermore, in comparison to the available high-throughput experimental methods, evolution-based

methods have the crucial advantage of measuring the importance of genetic variants in real organisms in their natural environments over many generations.

Using LINSIGHT, we examined the influence of negative selection on enhancers, considering the relationships between constraint on enhancers and numbers of active cell types, tissue of activity, and constraint at associated promoters. LINSIGHT is potentially useful for addressing these questions because it should be much more robust to evolutionary turnover than conventional conservation-based methods, and some classes of enhancers are known to turn over more quickly than others[43]. We found that, in general, the trends in constraint at enhancers parallel those previously reported for protein-coding genes. For example, constraint increases with breadth of activity across cell types and decreases in tissues associated with rapid evolution, such as olfactory regions, the immune system, and male reproduction. Constraint also appears to be correlated at enhancer/promoter pairs. These observations about the specific ways in which evolutionary constraints on enhancers depend on genomic context may be useful in improving the prediction power for the fitness consequences of noncoding mutations.

As has been suggested for protein-coding genes[38], it seems plausible that the positive correlation between the strength of constraint and the number of active cell types can be explained by pleiotropy: enhancers active in more cell types are more likely to participate in multiple regulatory networks, perhaps with distinct roles involving the binding of different factors and/or the use of different binding sites within each enhancer. As a result, they may be subject to greater constraint. Nevertheless, many open questions remain about the influences of constraint on enhancers, and it will be important to examine these questions further in light of rapidly improving enhancer annotations, data describing enhancer-promoter interactions[44-46], and observations of complex evolutionary behavior at enhancers[47].

Online Methods

Genomic features

The genomic features used by LINSIGHT can be divided into three categories: conservation scores, predicted binding sites, and regional annotations (Table 2 and Supplementary Table 3). Conservation scores included phyloP scores[25], phastCons elements[4], SiPhy omega elements[48,49], and CEGA elements[50]. Except for SiPhy, each score type was represented by multiple data tracks—for example, phastCons tracks for vertebrate, mammalian, and primate alignments (Supplementary Table 3). Predicted binding sites included transcription factor binding sites (TFBS) and RNA binding sites. Predicted TFBSs were obtained from the conserved TFBS track in the UCSC Genome Browser[51], the rVISTA database[52], SwissRegulon[53], FunSeq2[20], and the Ensembl Regulatory Build[27]. RNA binding sites include splice sites predicted by SPIDEX[54] and miRNA target sites predicted by TarBase[55]. The regional annotations were based a variety of sources, including ChIP-seq and RNA-seq data from the ENCODE[11] and Roadmap Epigenomics[12] projects, enhancers from FANTOM5[35], predicted distal regulatory modules from FunSeq2[20], and the distances to nearest TSSs based on GENCODE gene models[56]. All features and the resulting LINSIGHT scores were expressed in genomic coordinates for the hg19 assembly of the human genome.

Polymorphism and divergence data

The polymorphism and divergence data used by the INSIGHT component of the LINSIGHT model were borrowed from previous analyses[3,21,22]. Briefly, we obtained human single nucleotide polymorphisms from high-coverage genome sequences for 54 unrelated individuals from the "69 Genomes" data set from Complete Genomics, eliminating nucleotide sites with more than two alleles. Outgroup alleles were defined by the aligned chimpanzee, orangutan, and rhesus macaque reference genomes from UCSC. Several filters were applied to these data to reduce technical errors from alignment, sequencing, and genotype inference; for example, we removed simple repeats, recent transposable elements, recent segmental duplications, and regions not in syntenic alignment across primates[22]. Putatively neutral regions were identified by starting with all aligned regions, then removing coding sequences, conserved noncoding sequences, and their proximal flanking regions. These regions were used to estimate neutral divergence and polymorphism rates in the human lineage in a block-wise manner across the genome, to account for local variation in mutation rates[21]. To allow for uncertainty in the human-chimpanzee most recent common ancestor (MRCA), we integrated over a distribution of ancestral alleles inferred after fitting a standard phylogenetic model to the outgroup sequences[21].

Generalized linear model

The selection parameters in the INSIGHT model, ρ and γ, were defined as linear-sigmoid functions of the local genomic features at each nucleotide site i. Specifically, if $\mathbf{D}_i$ is a column vector of genomic features at site i, then

$$\rho_i = g(\mathbf{W}_\rho \mathbf{D}_i) \quad and \quad \gamma_i = h(\mathbf{W}_\gamma \mathbf{D}_i), \quad (1)$$

where the row vectors $\mathbf{W}_\rho$ and $\mathbf{W}_\gamma$ consist of feature weights (free parameters in the model) and $g()$ and $h()$ are sigmoid functions that map all input values to the range **(0,1)**. For $h()$, we used the standard logistic function, $h(x) = 1/(1 + e^{-x})$. For $g()$, however, we used the asymmetric Gompertz sigmoid function[57], $g(x) = \exp[-3\exp(-x)]$, which ensured that gradients were not too small when ρ_i is close to zero and accelerated convergence during model fitting.

Fitting the LINSIGHT model to the data

The weights for all genomic features were estimated by approximately maximizing the log likelihood of the INSIGHT model with respect to our genome-wide data set. We began by considering all genomic positions not excluded by our data-quality filters. Because our focus was on noncoding regions, we additionally excluded coding regions annotated by GENCODE (release 19). Instead of a traditional "batch" learning algorithm, which would require either storing all data in memory or reading it from disk many times, we used an "online" stochastic gradient descent algorithm[58]. The algorithm processed the genome sequentially, in "minibatches" of 100 successive nucleotides, each time updating the parameter vector in the direction of the gradient of the log likelihood function, with learning rates of 0.001 and 0.01 for ρ and γ, respectively. Gradients were computed analytically, by propagating partial derivatives through the linear-sigmoid component of the model using the

chain rule (back-propagation). The learning procedure was truncated after 20 passes through the entire data set. The entire process took less than one day on a desktop computer. The genome-wide LINSIGHT scores are available from the Cold Spring Harbor Laboratory mirror of the UCSC Genome Browser (hg19 assembly).

Comparison with other methods

Our benchmarking scheme for prioritization of disease-associated variants closely followed the one introduced in ref. [13]. The HGMD and ClinVar noncoding disease variants and three sets of negative controls were obtained directly from this study. The negative controls consisted of: (1) a randomly selected subset of human common variants which is 100-fold larger than the set of HGMD variants (unmatched); (2) a subset of human common variants matched to the disease variants by exact distance-to-nearest-TSS (matched TSS) (although each negative example is not necessarily near the same TSS as the matched disease variant); and (3) a subset of human common variants required to be within 1-kb of the matched disease variants (matched region). The two matched sets account for the enrichment of known disease variants near coding genes. We later defined three additional sets of negative controls by the same strategy but using singleton variants from the 1000 Genomes Project phase 3 data[59] instead of common variants, to ensure that our results were not driven by differences in allele frequency between the disease variants and negative controls. In all cases, we subsampled the negative sets to balance the numbers of positive and negative sets. To reduce stochasticity, subsampling was performed 100 times and average performance statistics were reported.

For comparison, we downloaded precomputed CADD[18] (v1.3), GWAVA[13] (v1.0), FunSeq2[20] (v2.1.0), and Eigen[34] (Oct. 11, 2015) scores from the source websites. In all cases, we used GWAVA scores based on training with variants matched by distance-to-nearest-TSS were used[13]. In addition, we obtained mammalian phyloP[25] scores based on the 46-way whole-genome alignment for hg19 from the UCSC Genome Browser[51], and we computed DeepSEA functional significance scores for both disease variants and negative controls using the online DeepSEA web service[16] (computed on Nov 3, 2016). The DeepSEA functional significance scores integrate individual tissue-specific DeepSEA scores based on polymorphism data; these were used in all comparisons because the tissue types associated with disease variants and ORegAnno TFBSs are typically unknown. Note that two of the methods considered, CADD and DeepSEA, provide allele-specific predictions, whereas the other methods assign identical scores to all alternative variants. When evaluating CADD and DeepSEA on the ClinVar data set, we used the score corresponding to the annotated disease-associated allele. When evaluating these methods on the HGMD data set, however, no disease-associated allele was provided, so we used the maximum score for the three alternative alleles.

Classification of disease-associated variants by genomic location

For analyses that considered the genomic locations of disease-associated variants, we divided the variants in the HGMD and ClinVar databases into four categories based on their locations relative to gene models from GENCODE (release 19). These categories were: (1) “promoter” variants, located within 1 kb upstream of the 5’-most annotated transcription

start site of any protein-coding gene; (2) “splicing” variants, located within 20 bp of any annotated splice junction; (3) “UTR” variants, located within the annotated 5’ or 3’ UTR of any protein-coding gene; and (4) all “other” variants. Each variant was assigned to the first category whose criteria it fulfilled in the order *splicing* > *UTR* > *promoter* > *other*.

Quantification of the contributions of genomic feature classes

We measured the relative contributions of the conservation scores, predicted binding sites, and regional annotations by removing all features of each class (see Table 2), retraining the LINSIGHT model without those features, and evaluating the AUC of the reduced model. The *contribution* of each class of features was defined as the AUC for the full model minus the AUC for the reduced model, averaged across 100 independent subsamples of negative controls described above. Notice that, while this difference in AUCs is generally positive, it may be negative due to stochastic effects. This analysis was performed on a merged set of HGMD and ClinVar variants, separately for promoter, splicing, UTR, and other regions.

Analysis of evolutionary constraints on enhancers

To study evolutionary constraints on enhancers, we used the comprehensive atlas of human enhancers based on enhancer RNAs (eRNAs) that was recently provided by the FANTOM5 project[35]. The evolutionary constraint for each enhancer was quantified by taking the average LINSIGHT score across all nucleotide sites in the enhancer. We examined the relationship between this measure of constraint and the number of cell types in which each enhancer was active (according to a detectable eRNA signature). We also defined a subset of enhancers as tissue-specific, based on apparent activity in only a single tissue type, and examined the relationship between tissue of activity and degree of constraint. Finally, we obtained putative enhancer-TSS pairs (based on correlated patterns of expression across tissues) from the FANTOM5 website, and examined the correlation in constraint at the enhancer and promoter in each pair, defining the promoter as the 1 kb region upstream of the TSS. In cases where an enhancer was associated with multiple TSSs, the TSS with highest correlation coefficient was selected.

Statistical analysis

To examine the relationship between evolutionary constraints on enhancers and tissue specificity, Spearman’s rank correlation coefficient was calculated between the average LINSIGHT score for each enhancer and its number of active cell types. To quantify the statistical significance of the correlation, a two-tailed p-value was computed using the standard asymptotic t approximation implemented in the “cor.test” function in R ($p < 10^{-15}$; $n = 29{,}303$). The same method was used to quantify the statistical significance of the correlation between the average LINSIGHT scores at enhancer/promoter pairs ($p < 10^{-15}$; $n = 25{,}067$). Furthermore, to investigate the relationship between the average LINSIGHT score in an enhancer and the number of active cell types when controlling for average eRNA expression level, the partial Spearman’s ρ and a two-tailed p-value were computed using the *ppcor* package[60] ($p < 10^{-15}$; $n = 29{,}303$). To investigate whether the difference between two AUCs is statistically significant, the DeLong test was used to compute two-tailed p-values[61].

Code availability

The LINSIGHT code is available at https://github.com/CshlSiepelLab/LINSIGHT.

Data availability

The training data and pre-computed LINSIGHT scores are available at http://compgen.cshl.edu/~yihuang/LINSIGHT/.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Ilan Gronau for comments on the manuscript and members of the Siepel Laboratory for helpful discussions. This research was supported by US National Institutes of Health (NIH) grants GM102192 and HG008901. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

References

1. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002; 420:520–562. [PubMed: 12466850]
2. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–9367. [PubMed: 19474294]
3. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nature Genetics. 2015; 47:276–283. [PubMed: 25599402]
4. Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research. 2005; 15:1034–1050. [PubMed: 16024819]
5. Wang Y, et al. Sequencing and comparative analysis of a conserved syntenic segment in the Solanaceae. Genetics. 2008; 180:391–408. [PubMed: 18723883]
6. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011; 478:476–482. [PubMed: 21993624]
7. Haudry A, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet. 2013; 45:891–898. [PubMed: 23817568]
8. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
9. Gerstein MB, et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010; 330:1775–1787. [PubMed: 21177976]
10. Roy S, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010; 330:1787–1797. [PubMed: 21177974]
11. Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
12. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]
13. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nature Methods. 2014; 11:294–296. [PubMed: 24487584]
14. Shihab HA, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015; 31:1536–1543. [PubMed: 25583119]
15. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015; 33:831–838. [PubMed: 26213851]

16. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learningbased sequence model. Nature Methods. 2015; 12:931–934. [PubMed: 26301843]
17. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016
18. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics. 2014; 46:310–315. [PubMed: 24487276]
19. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015; 31:761–763. [PubMed: 25338716]
20. Fu Y, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biology. 2014; 15:480. [PubMed: 25273974]
21. Gronau I, Arbiza L, Mohammed J, Siepel A. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. Molecular Biology and Evolution. 2013; 30:1159–1171. [PubMed: 23386628]
22. Arbiza L, et al. Genome-wide inference of natural selection on human transcription factor binding sites. Nature Genetics. 2013; 45:723–729. [PubMed: 23749186]
23. Stenson PD, et al. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Human Genetics. 2013; 133:1–9.
24. Landrum MJ, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research. 2014; 42:D980–D985. [PubMed: 24234437]
25. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research. 2010; 20:110–121. [PubMed: 19858363]
26. Rumelhart D, Hinton G, Williams R. Learning representations by back-propagating errors. Nature. 1986; 323:533–536.
27. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The Ensembl Regulatory Build. Genome Biology. 2015; 16:1–8. [PubMed: 25583448]
28. Gaffney DJ, Blekhman R, Majewski J. Selective constraints in experimentally defined primate regulatory regions. PLoS Genet. 2008; 4:e1000157. [PubMed: 18704158]
29. Chiaromonte F, et al. The share of human genomic DNA under selection estimated from human-mouse genomic alignments. Cold Spring Harbor Symp Quant Biol. 2003; 68:245–254. [PubMed: 15338624]
30. Meader SJ, Ponting CP, Lunter G. Massive turnover of functional sequence in human and other mammalian genomes. Genome Research. 2010; 20:1335–1343. [PubMed: 20693480]
31. Rands CM, Meader S, Ponting CP, Lunter G. 8.2% of the human genome is constrained: Variation in rates of turnover across functional element classes in the human lineage. PLoS Genetics. 2014; 10:e1004525. [PubMed: 25057982]
32. Lesurf R, et al. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. Nucleic Acids Res. 2016; 44:D126–132. [PubMed: 26578589]
33. Davydov EV, et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++ PLoS Computational Biology. 2010; 6:e1001025. [PubMed: 21152010]
34. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature Genetics. 2016; 48:214–220. [PubMed: 26727659]
35. Andersson R, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014; 507:455–461. [PubMed: 24670763]
36. Core LJ, et al. Analysis of nascent RNA identifies a unified architecture of transcription initiation regions at mammalian promoters and enhancers. Nat Genet. 2014; 46:1311–1320. [PubMed: 25383968]
37. Andersson R, Sandelin A, Danko CG. A unified architecture of transcriptional regulatory elements. Trends Genet. 2015; 31:426–433. [PubMed: 26073855]
38. Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 2000; 17:68–74. [PubMed: 10666707]

39. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 2005; 102:14338–14343. [PubMed: 16176987]
40. Kosiol C, et al. Patterns of positive selection in six mammalian genomes. PLoS Genetics. 2008; 4:e1000144. [PubMed: 18670650]
41. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biology. 2006; 4:e72. [PubMed: 16494531]
42. Spivakov M, et al. Analysis of variation at transcription factor binding sites in Drosophila and humans. Genome Biology. 2012; 13:1–15.
43. Villar D, et al. Enhancer evolution across 20 mammalian species. Cell. 2015; 160:554–566. [PubMed: 25635462]
44. Rao SS, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159:1665–1680. [PubMed: 25497547]
45. Guo Y, et al. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell. 2015; 162:900–910. [PubMed: 26276636]
46. Tang Z, et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell. 2015; 163:1611–1627. [PubMed: 26686651]
47. Wunderlich Z, et al. Kruppel expression levels are maintained through compensatory evolution of shadow enhancers. Cell Rep. 2015; 12:1740–1747. [PubMed: 26344774]
48. Garber M, et al. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009; 25:i54–i62. [PubMed: 19478016]
49. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011; 478:476–482. [PubMed: 21993624]
50. Dousse A, Junier T, Zdobnov EM. CEGA–a catalog of conserved elements from genomic alignments. Nucleic Acids Research. 2016; 44:D96–D100. [PubMed: 26527719]
51. Kent WJ, et al. The human genome browser at UCSC. Genome Research. 2002; 12:996–1006. [PubMed: 12045153]
52. Dubchak I, et al. Whole-genome rVISTA: a tool to determine enrichment of transcription factor binding sites in gene promoters from transcriptomic data. Bioinformatics. 2013; 29:2059–2061. [PubMed: 23736530]
53. Pachkov M, Balwierz PJ, Arnold P, Ozonov E, van Nimwegen E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic Acids Research. 2013; 41:D214–D220. [PubMed: 23180783]
54. Xiong HY, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2014; 347:1254806. [PubMed: 25525159]
55. Vlachos IS, et al. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Research. 2015; 43:D153–D159. [PubMed: 25416803]
56. Harrow J, et al. GENCODE: The reference human genome annotation for the ENCODE Project. Genome Research. 2012; 22:1760–1774. [PubMed: 22955987]
57. Gompertz B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philosophical Transactions of the Royal Society of London. 1825; 115:513–583.
58. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research. 2011; 12:2121–2159.
59. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]
60. Kim S. ppcor: An R package for a fast calculation to semi-partial correlation coefficients. Communications for Statistical Applications and Methods. 2015; 22:665–674. [PubMed: 26688802]
61. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics. 1988; 44:837–845. [PubMed: 3203132]

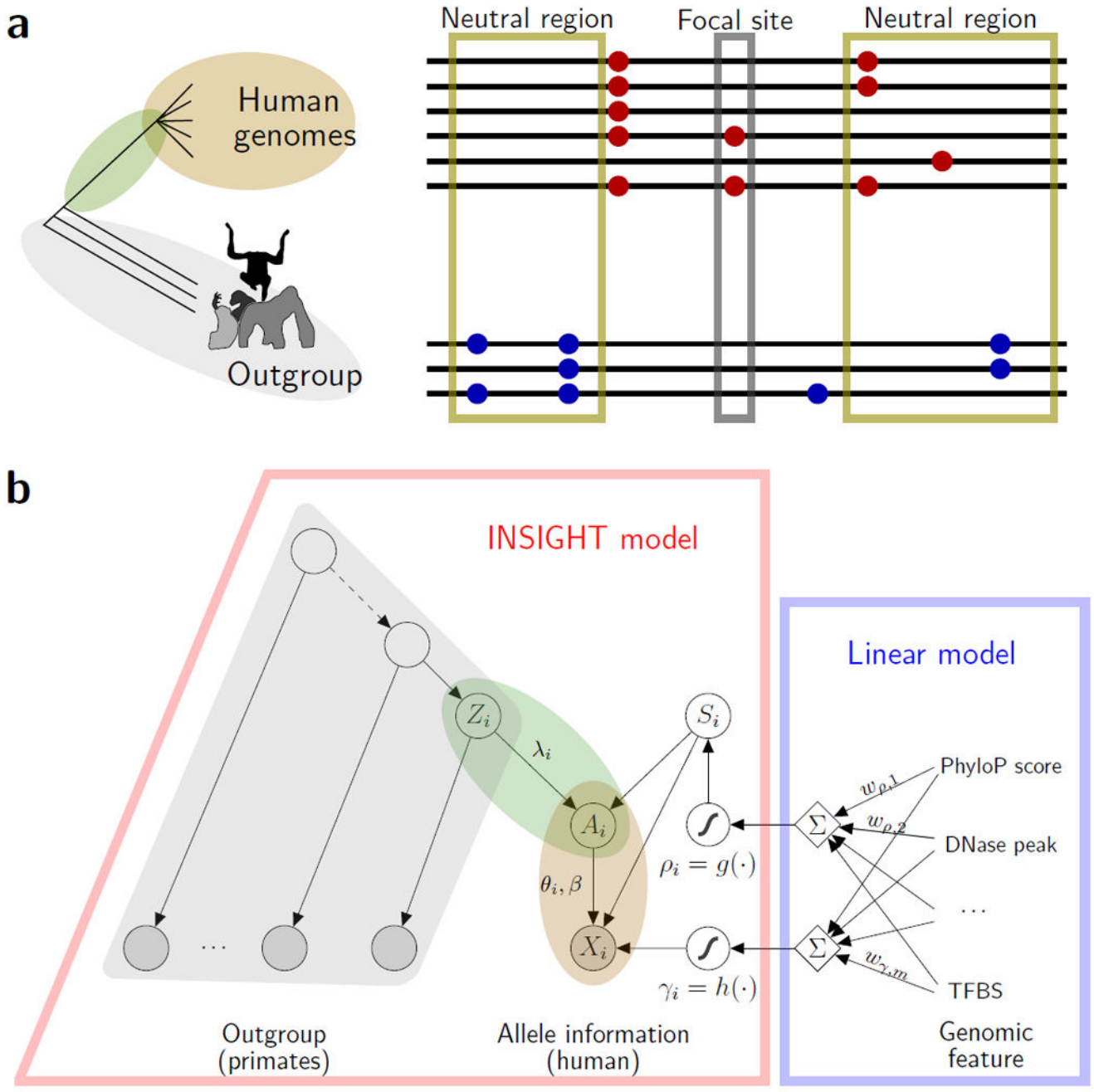

Fig. 1.

Conceptual overview of LINSIGHT. **(a)** Like the fitCons method[3], LINSIGHT estimates probabilities that mutations at each genomic site will have fitness consequences, based on patterns of genetic polymorphism within a species (here, humans) and patterns of divergence from closely related outgroup species (chimpanzee, orangutan, and rhesus macaque). Patterns of genetic variation at the focal site and other sites like it are contrasted with those in neutrally evolving regions nearby. Red circles indicate human single nucleotide polymorphisms and blue circles indicate nucleotide substitutions between species. **(b)**

LINSIGHT combines the probabilistic graphical model from INSIGHT[21,22] with a generalized linear model. The selection parameters from INSIGHT, ρ and γ, are defined in a sitewise manner by linear combinations of local genomic features, followed by sigmoid transformations. The figure summarizes the behavior at a particular focal site *i*. The matching shaded regions in the left of **(a)** and the left of **(b)** indicate corresponding portions of the INSIGHT model and the phylogeny and sequence data. See Table 1 for definitions of all parameters and variables.

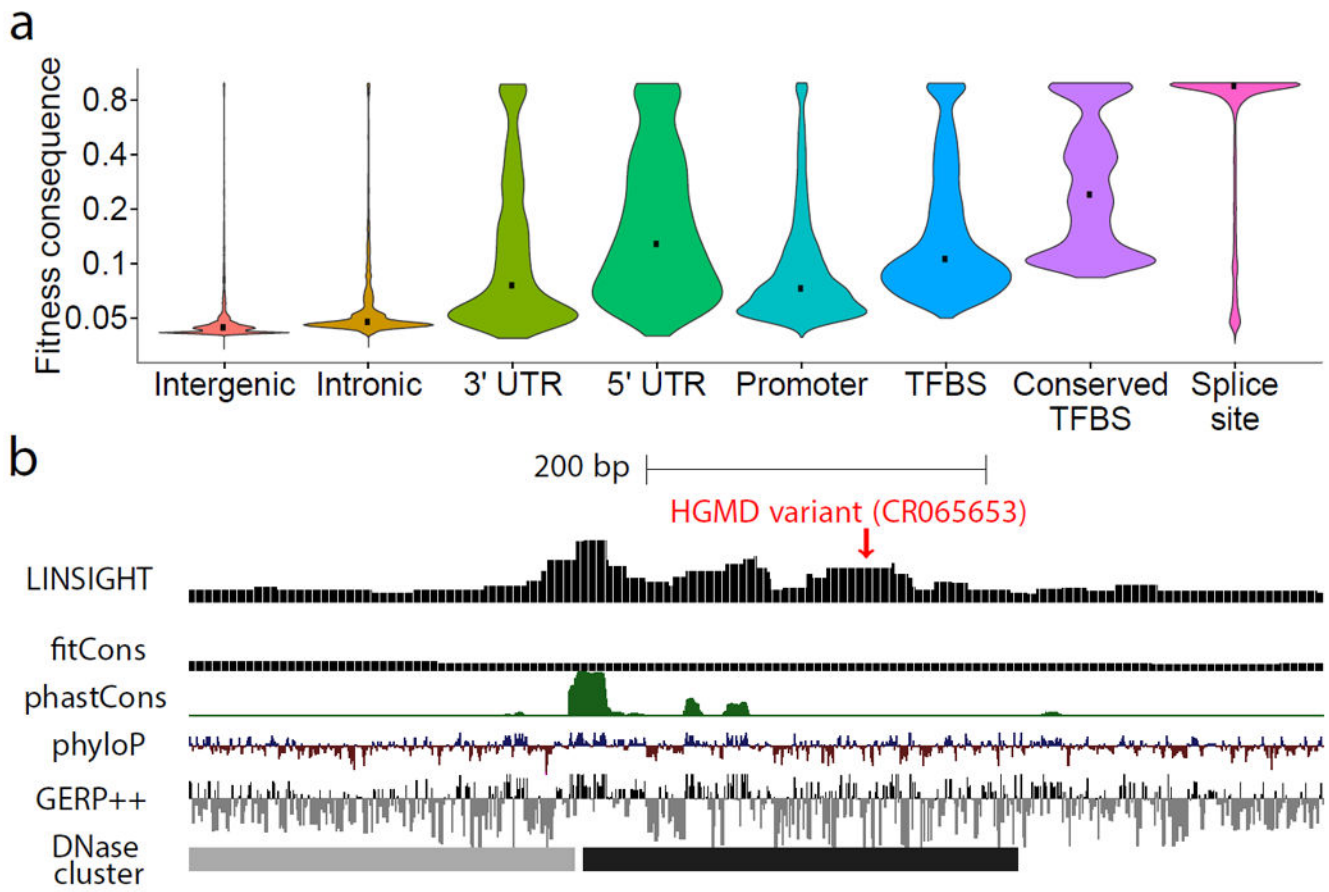

Fig. 2.

Summary of LINSIGHT scores across the noncoding human genome (3.001 billion nucleotide sites). **(a)** Distributions of LINSIGHT scores for various genomic regions. Intergenic, intronic, UTRs, and 1-kb promoters were defined based on GENCODE annotations (version 19); TFBSs were predicted from ChIP-seq peaks (Ensembl Regulatory Build); and conserved TFBSs were obtained from the UCSC Genome Browser. Within each violin plot, width represents density and black dot represents median LINSIGHT score. Note the logarithmic vertical scale. **(b)** UCSC Genome Browser display showing LINSIGHT scores alongside those from fitCons, phastCons, phyloP, and GERP++. LINSIGHT integrates functional genomic data together with conservation scores and other features to provide a high-powered, high-resolution measure of potential function. In this example, it is the only method to highlight a variant from HGMD (CR065653) that is associated with up-regulation of the telomerase reverse transcriptase (*TERT*) gene. See Supplementary Figure 3 for additional examples.

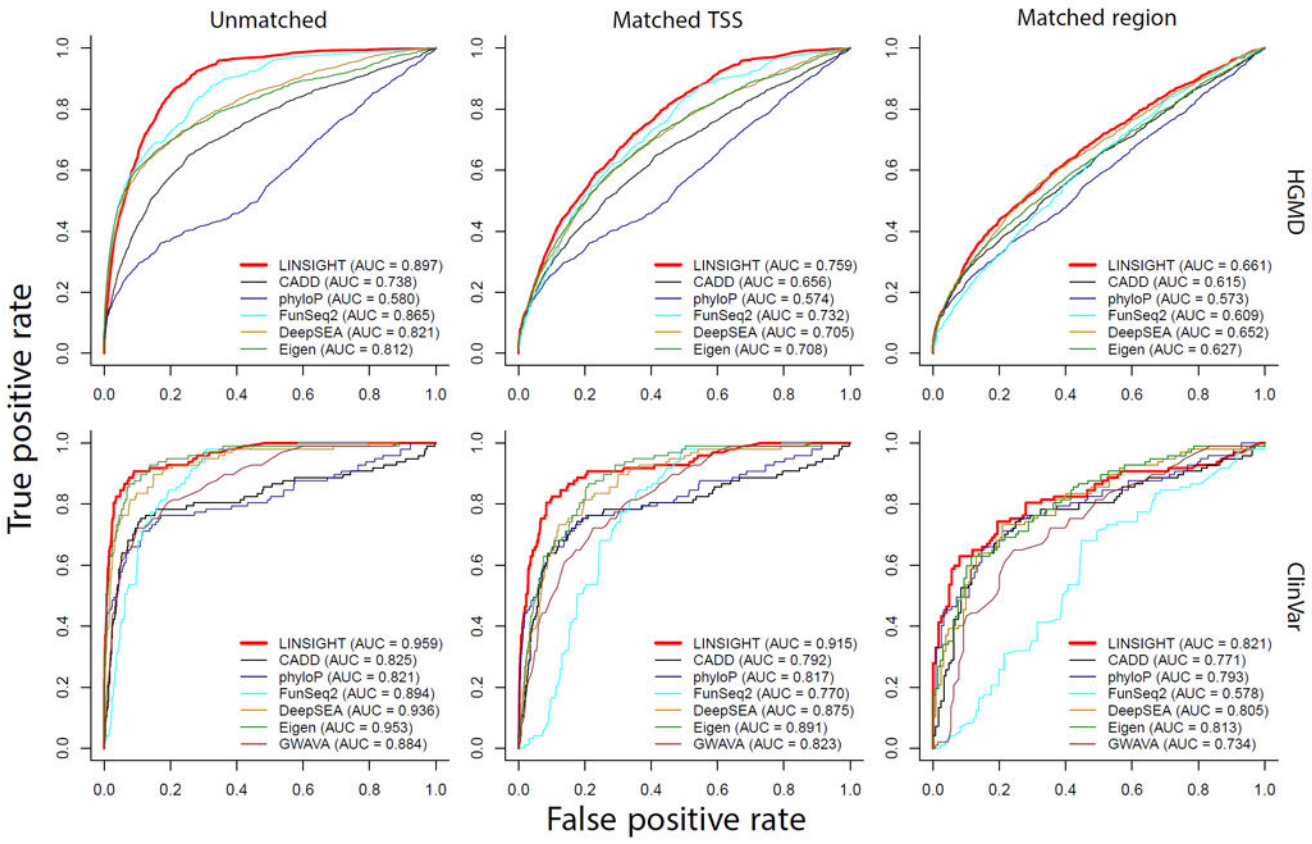

Fig. 3.

Prediction power of various computational methods for distinguishing disease-associated noncoding variants from variants not likely to have phenotypic effects. True positive and false positive rates are proportions of disease and neutral variants, respectively, having scores that exceed each threshold, as the threshold is varied. Power is quantified using the Area Under the Curve (AUC) statistic. Results are shown for positive examples from the HGMD[23] (1495 variants) and ClinVar[24] (101 variants not in HGMD) databases. Only autosomal variants were included and duplicated variants were removed. Common SNPs (MAF > 1%) were used as negative examples and were either randomly selected (unmatched), matched to positive examples by distance to nearest transcription start site (matched TSS), or matched to positive examples within 1 kb along the genome (matched region). The numbers of positive and negative examples were balanced by subsampling, which was performed 100 times to obtain average true positive and false positive rates. LINSIGHT is compared with CADD[18], phyloP[25], FunSeq2[20], DeepSEA[16], Eigen[34], and GWAVA[13]. FitCons is not included because it performs poorly on this task due to its low genomic resolution and cell-type specificity. GWAVA results are not shown for the HGMD data set because GWAVA was trained on this data set.

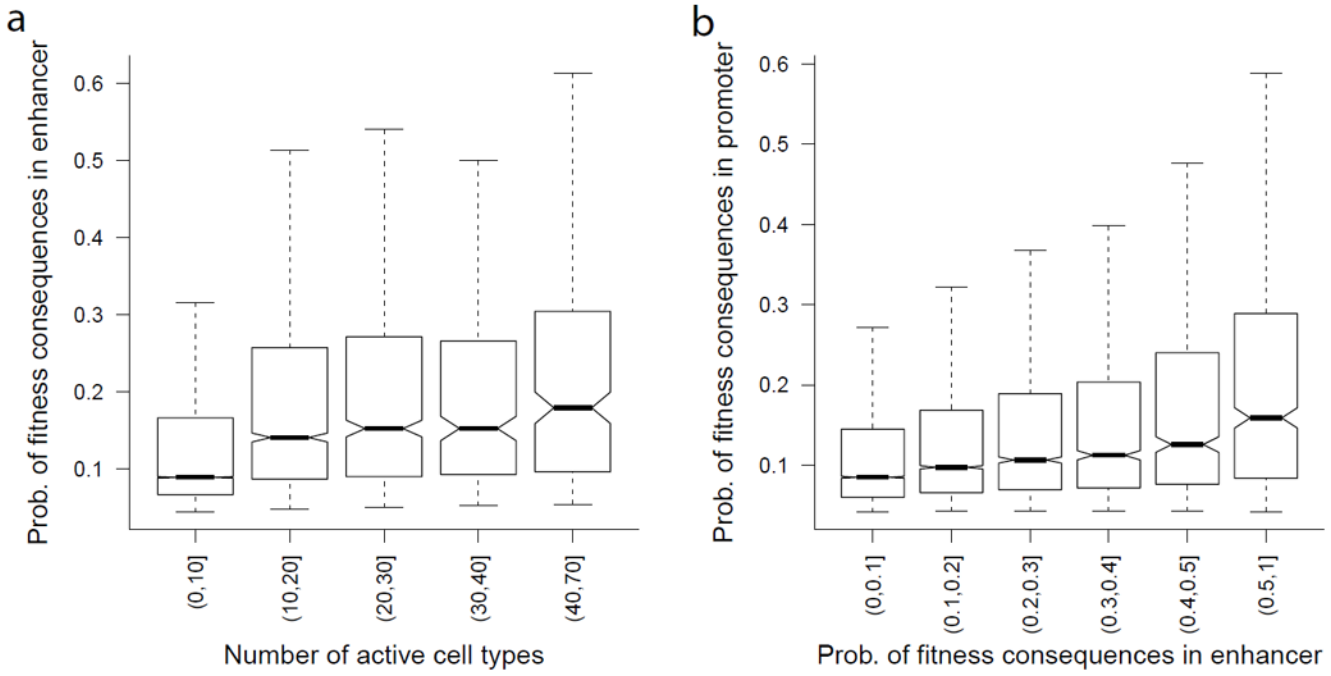

Fig. 4.

Evolutionary constraints on enhancers. **(a)** Probability of fitness consequences for mutations in enhancers (measured by average LINSIGHT score) is positively correlated with the number of cell types in which each enhancer is active (Spearman's rank correlation coefficient $\rho = 0.284$; two-tailed p-value $< 10^{-15}$). Results are shown for 29,303 enhancers in 69 cell types. **(b)** Probability of fitness consequences for mutations in enhancers is positively correlated with probability of fitness consequences for mutations in associated promoters (Spearman's rank correlation coefficient $\rho = 0.150$; two-tailed p-value $< 10^{-15}$). Results are shown for 25,067 enhancer-promoter pairs.

Table 1

Summary of key model parameters and variables

	Parameters inherited from INSIGHT[a]
ρ_i	Probability that site i is under selection. Interpreted as the **LINSIGHT** score for site i
γ_i	Expected relative rate of low-frequency derived alleles at site i given that it is under selection
λ_i	Neutral substitution rate at site i
θ_i	Neutral polymorphism rate at site i
$\beta = (\beta_1,\beta_2,\beta_3)$	Fractions of neutral polymorphisms with low-, intermediate-, and high-frequency derived alleles
	Variables inherited from INSIGHT[a]
$X_i = (X_i^{maj}, X_i^{min}, Y_i)$	Observed polymorphism data at site i, including major allele, minor allele, and minor-allele frequency class
Z_i	Human-chimpanzee ancestral allele at site i
A_i	Human ancestral allele at site i
S_i	Indicator for whether or not site i is under selection
	Components of LINSIGHT's generalized linear model[a]
$\mathbf{D}_i = (d_{i,1},\ldots,d_{i,m})$	Genomic feature vector at site i
$\mathbf{W}_\rho = (w_{\rho,1},\ldots,w_{\rho,m})$	Weight vector for ρ (free parameters)
$\mathbf{W}\gamma = (w_{\gamma,1},\ldots,w_{\gamma,m})$	Weight vector for γ (free parameters)
$g()$	Sigmoid function for ρ (Gompertz)
$h()$	Sigmoid function for γ (logistic)

[a]See Supplementary Note and Supplementary Table 1 for full details.

Table 2

Summary of genomic features used for LINSIGHT scores

Class	Genomic feature[a]	Spatial resolution
Conservation	phyloP score	High
	phastCons element	High
	SiPhy element	High
	CEGA element	High
Binding site	Conserved TFBS	High
	rVISTA TFBS	High
	SwissRegulon TFBS	High
	Predicted TFBS within ChIP-seq peak	High
	Conserved miRNA binding site	High
	Splicing site predicted by SPIDEX	High
Regional annotation	ChIP-seq peak of transcription factor	Low
	DNase-I hypersensitive site	Low
	UCSC FAIRE peak	Low
	RNA-seq signal	Low
	Histone modification peak	Low
	FANTOM5 enhancer	Low
	Predicted distal regulatory module	Low
	Distance to nearest TSS	Low

[a]Each “genomic feature” listed here may actually correspond to multiple features in the model. For example, four features are derived from phyloP scores: two from the mammalian phyloP scores and two from the vertebrate phyloP scores. See Supplementary Table 3 for complete details.