nature genetics

# A method for calculating probabilities of fitness consequences for point mutations across the human genome

Brad Gulko[1], Melissa J Hubisz[2], Ilan Gronau[2,3] & Adam Siepel[1–3]

**We describe a new computational method for estimating the probability that a point mutation at each position in a genome will influence fitness. These 'fitness consequence' (fitCons) scores serve as evolution-based measures of potential genomic function. Our approach is to cluster genomic positions into groups exhibiting distinct 'fingerprints' on the basis of high-throughput functional genomic data, then to estimate a probability of fitness consequences for each group from associated patterns of genetic polymorphism and divergence. We have generated fitCons scores for three human cell types on the basis of public data from ENCODE. In comparison with conventional conservation scores, fitCons scores show considerably improved prediction power for *cis* regulatory elements. In addition, fitCons scores indicate that 4.2–7.5% of nucleotides in the human genome have influenced fitness since the human-chimpanzee divergence, and they suggest that recent evolutionary turnover has had limited impact on the functional content of the genome.**

During the past decade, two major developments—the emergence of massively parallel, ultra-cheap DNA sequencing technologies and the use of these technologies as digital readouts for functional genomic assays—have led to a profusion of data describing various features of genomes, epigenomes and transcriptomes[1,2]. However, investigators still have only rudimentary tools for integrating these diverse sources of information to obtain useful insights about genomic function and evolution. The limitations of current methods are particularly evident in the vast noncoding regions of eukaryotic genomes, which, despite recent progress[3–6], remain poorly annotated and understood. These limitations hamper progress in many areas, including molecular genetics, disease association and personalized medicine[7].

Many computational methods for the functional analysis of sequence data are based on the simple but profound observation that functionally important nucleotides tend to remain unchanged over evolutionary time because mutations at these sites generally reduce fitness and are therefore eliminated by natural selection[7–15]. A major strength of these conservation- or constraint-based approaches is that they sidestep thorny questions about the relationship between the outcomes of biochemical experiments and fitness-influencing functional roles[16–19] by getting at fitness directly through observations of evolutionary change. In essence, the 'experiment' considered by these methods is the one conducted directly on genomes by nature over millennia, and the outcomes of interest are the presence or absence of fixed mutations.

These conservation-based methods, however, depend critically on the assumption that genomic elements are present at orthologous locations and maintain similar functional roles over relatively long evolutionary time periods. Evolutionary turnover may cause inconsistencies between sequence orthology and functional homology that substantially limit this type of analysis. Consequently, investigators have developed two major alternative strategies for the identification and characterization of functional elements. The first strategy is to augment information about interspecies conservation with information about genetic polymorphism[20–28]. The shorter evolutionary time scales associated with intraspecies variation make this approach more robust to evolutionary turnover and less sensitive to errors in alignment and orthology detection. Polymorphic sites tend to be sparse along the genome, however, so this approach requires some type of pooling of information across genomic positions, which can be problematic in the absence of high-quality genomic annotations. The second strategy is to forgo the use of evolutionary information and to instead predict functional roles from genomic data alone, typically with machine learning methods for supervised classification[29,30] or clustering followed by labeling based on known examples[31–33]. This approach has the limitation that it depends strongly on previously characterized elements, which in noncoding regions are typically few and perhaps unrepresentative of the genome.

In this report, we introduce a method for genomic analysis that combines many of the strengths of these polymorphism-based and functional genomic approaches. Like functional genomic methods, our approach groups genomic regions according to functional genomic fingerprints across multiple assays. Instead of relying on known examples for classification, however, we characterize each group by a probability of mutational fitness consequences—or fitCons score—inferred from patterns of genetic variation. These fitCons scores are estimated using a recently developed statistical method, called Inference of Natural Selection from Interspersed Genomically Coherent Elements (INSIGHT), that contrasts patterns of polymorphism

[1]Graduate Field of Computer Science, Cornell University, Ithaca, New York, USA. [2]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, USA. [3]Present addresses: Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA (A.S.) and Efi Arazi School for Computer Science, Interdisciplinary Center, Herzliya, Israel (I.G.). Correspondence should be addressed to A.S. (asiepel@cshl.edu) or I.G. (ilan.gronau@idc.ac.il).

Received 13 August 2014; accepted 19 December 2014; published online 19 January 2015; doi:10.1038/ng.3196

and divergence for a collection of dispersed genomic sites with those for nearby neutrally evolving sites, accounting for negative and positive selection[34]. Thus, the method integrates both evolutionary and functional data in characterizing the potential functional importance of genomic regions. We demonstrate that these fitCons scores are useful for visualization, for prediction of *cis* regulatory elements and for measurement of the global influence of recent natural selection across the genome.

## RESULTS

### General features of the prediction problem

Information about genetic variation can be used to estimate probabilities of fitness consequences for moderately large groups of genomic positions but not for individual loci, owing to the sparsity of informative sites along the genome. This property of 'group-wise' but not 'individual' predictivity is common to many statistical problems, but it is complicated in our case by two additional features. First, an appropriate scheme for grouping or stratification is not clear a priori here because genomic correlates of fitness consequences are incompletely understood. Second, the outcomes of interest in our problem—the fitness consequences of point mutations—are not directly evident from the data. To highlight these challenges, consider the simpler problem of estimating the expected risk of an automobile accident. This problem must also be addressed at the level of groups (either explicitly, through stratification of drivers, or implicitly, through regression), but in this case the relevant features—such as the age, sex and number of traffic violations of the driver—are generally plain to the analyst. In addition, the outcomes of interest—the occurrences and costs of accidents—are directly observed. In our problem, the genomic 'risk factors' for fitness-influencing mutations, particularly in unannotated noncoding regions of the genome, are much less clear. Furthermore, once a grouping is determined, it is still not possible to read off the associated fitness consequences of mutations; instead, they must be inferred from patterns of genetic variation using an evolutionary model.

### Calculation of fitCons scores

We have addressed these challenges using the following strategy. Beginning with genome-wide functional genomic data sets obtained from each cell type (**Fig. 1**, first step), we first cluster genomic positions by their joint functional genomic fingerprints (**Fig. 1**, second step). We focus on three highly informative and largely orthogonal functional genomic data types—DNase I digestion and sequencing (DNase-seq) data, RNA sequencing (RNA-seq) data and chromatin immunoprecipitation and sequencing (ChIP-seq) data describing histone modifications—which describe DNA accessibility, transcription and chromatin states, respectively. We divide genomic positions into 3 levels of DNase-seq signal, 4 levels of RNA-seq signal and 26 distinct chromatin states on the basis of the ChromHMM method[31,33]. In addition, we distinguish between sites that fall outside or within annotated protein-coding sequences (CDSs). We then consider all possible combinations of these 4 types of assignments, obtaining $3 \times 4 \times 26 \times 2 = 624$ distinct functional genomic classes. We apply this clustering step separately to three karyotypically normal cell types:

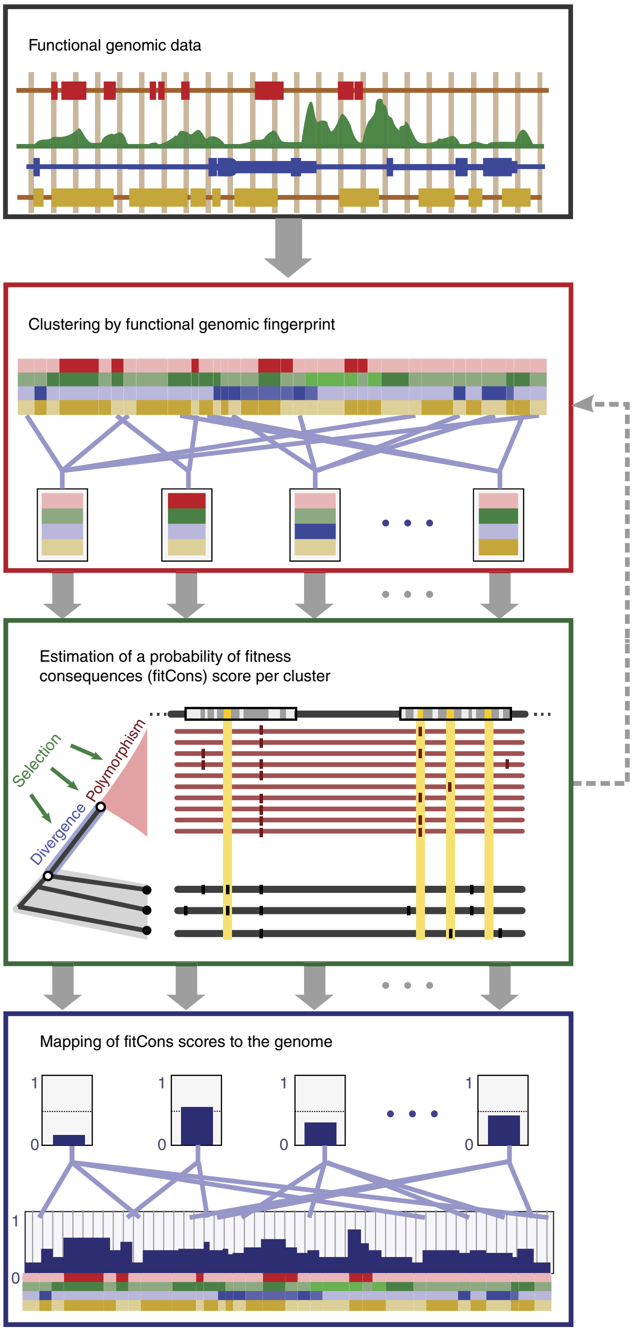
Functional genomic data
Clustering by functional genomic fingerprint
Estimation of a probability of fitness consequences (fitCons) score per cluster
Selection
Polymorphism
Divergence
Mapping of fitCons scores to the genome
1
0
1
0
1
0
1
0
1
0

**Figure 1** Illustration of the procedure for calculating fitCons scores. Functional genomic data, such as DNase-seq, RNA-seq and histone modification data, are arranged along the genome sequence in tracks (first panel). Nucleotide positions in the genome are clustered by joint patterns across these functional genomic tracks (second panel). For example, one cluster might contain genomic positions with a high DNase-seq signal, a moderate RNA-seq signal and high signals for monomethylation of histone H3 at lysine 4 (H3K4me1) and acetylation of histone H3 at lysine 27 (H3K27ac), suggesting transcribed enhancers. Another might contain positions with a low DNase-seq signal, a high RNA-seq signal and a signal for trimethylation of histone H3 at lysine 36 (H3K36me3), suggesting actively transcribed gene bodies. Note that clusters will generally contain genomic positions dispersed along the genome sequence. Patterns of polymorphism and divergence are analyzed using INSIGHT[34] to obtain an estimate of the fraction of nucleotides under natural selection ($\rho$) in each cluster (third panel). This quantity is interpreted as the probability that each nucleotide position influences the fitness of the organism that carries it, or a fitness consequence (fitCons) score. The fitCons score for each cluster is assigned to all genomic positions that were included in the cluster (fourth panel). In this way, all nucleotide positions are assigned a score, but there can be no more distinct scores than there are clusters. Note that, in our initial work here, the clustering of genomic positions is accomplished by a simple exhaustive partitioning scheme that produces 624 distinct clusters. In future work, however, it may be desirable to iterate between clustering and calculating scores (dashed arrow).

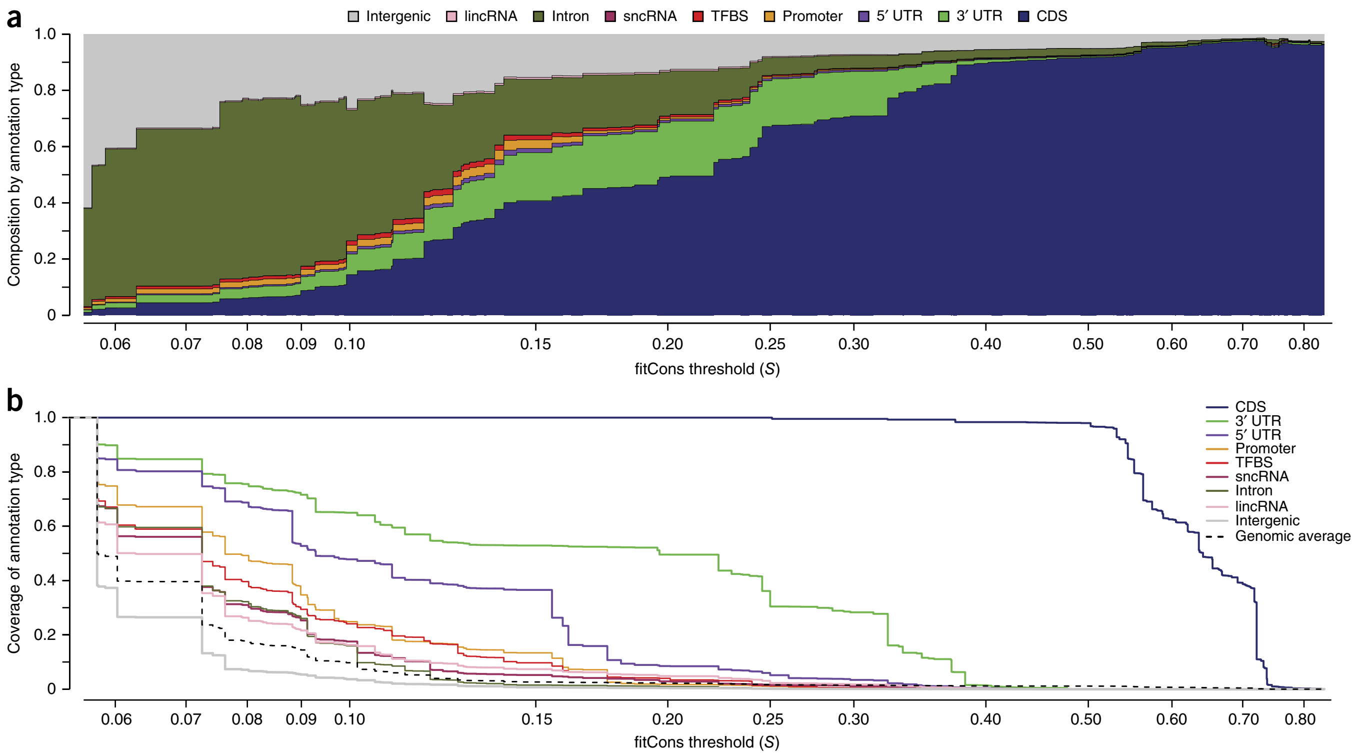

**Figure 2** Composition and coverage of high-scoring genomic regions according to fitCons. (**a**) Composition by annotation type in regions that exceed a fitCons score threshold of *S*, as *S* is varied across the range of possible scores. Each vertical cross-section of the plot can be thought of as a narrow 'stacked bar' representation of the composition by annotation type of all genomic positions at which the fitCons score is >*S*. At the left side of the plot, when *S* is small, the composition by annotation type is representative of the genome as a whole. As the threshold *S* increases, CDSs are increasingly enriched and intergenic sequences are increasingly depleted. Regions experiencing moderate levels of selection, such as UTRs, promoters, sncRNAs and introns, are most enriched at intermediate scores. Note the logarithmic scale for the *x* axis. TFBS, transcription factor binding site. (**b**) Coverage of the same annotation types by genomic regions having fitCons score > *S*, with an *x* axis matching that in **a**. The dashed line indicates the genome-wide average. At each value of *S*, the relative height of a given curve in comparison to the dashed line indicates the enrichment (or depletion) of the corresponding annotation type in genomic regions having score > *S*. The legend at the right lists the annotation types in order of decreasing enrichment. When multiple annotations applied to a single nucleotide position, one was selected in the following order: CDS, transcription factor binding site, promoter, sncRNA, lincRNA, 5′ UTR, 3′ UTR, intron and intergenic. These figures summarize data at 2.9 billion genomic sites.

human umbilical vein epithelial cells (HUVECs), H1 human embryonic stem cells (H1 hESCs) and lymphoblastoid cells (GM12878), resulting in 443–447 usable classes of sites with median numbers of 165,000 to 224,000 sites per class (see **Supplementary Table 1** and the Online Methods for details).

Next, we use INSIGHT to estimate the probabilities of mutational fitness consequences within each of these classes on the basis of patterns of polymorphism and divergence (**Fig. 1**, third step). This step yields an estimate of the fraction of sites under selection ($\rho$) for each of the analyzed classes, which serves as the fitCons score for that class. Finally, we assign to each nucleotide position in the genome the score estimated for the corresponding functional genomic class (**Fig. 1**, fourth step). Each genomic position is thus assigned a value between 0 and 1, representing the probability that the nucleotide at that position influences fitness, as estimated from patterns of variation at all genomic sites displaying the same functional genomic fingerprint. A vital property of these fitCons scores is that they integrate information from both evolutionary data and cell type–specific functional genomic data.

## Genomic distribution of fitCons scores

To obtain a general overview of the genomic distribution of fitCons scores, we first considered the composition and coverage of nucleotide sites of various annotation types as a variable threshold *S* was applied to the fitCons score, focusing on HUVECs (see the Discussion for a summary of other cell types). When *S* is zero, all sites are considered and the composition of annotations reflects the overall genomic distribution (**Fig. 2a**). As *S* increases, however, sites in known functional classes become strongly enriched relative to intergenic and intronic sites. Regions such as 5′ and 3′ UTRs, promoters and introns are most enriched at intermediate scores, reflecting moderate levels of natural selection in these regions, whereas CDSs dominate at the highest scores. Coverage properties (**Fig. 2b**) are best for CDSs, 3′ UTRs and 5′ UTRs (in that order), but they are also considerably elevated above the intergenic background for promoters, transcription factor binding sites, long intergenic noncoding RNAs (lincRNAs) and small noncoding RNAs (sncRNAs). Notably, the enrichment for functionally annotated genomic regions at high scores occurs despite no use of genomic annotations in the scoring scheme (except for CDS annotations). Instead, these elevated scores reflect differences in patterns of polymorphism and divergence that arise naturally from the fitness consequences of mutations in these regions and become evident after clustering on the basis of functional genomic data. The fitCons scores for each cell type are displayed across the genome as tracks in the Cold Spring Harbor Laboratory mirror of the UCSC Genome Browser (**Fig. 3** and **Supplementary Fig. 1**).

fitCons scores generally depend in expected ways on the marginal signals of functional genomic covariates, but they are also capable of

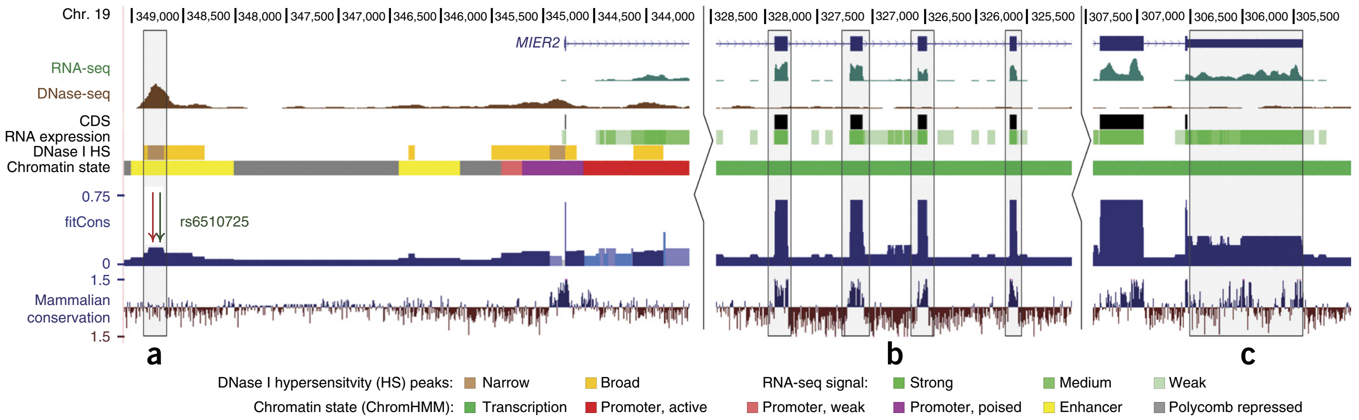

**Figure 3** Genome browser display showing functional genomic fingerprints and fitCons scores. Shown, from top to bottom, are the exons of the *MIER2* gene; the raw RNA-seq and DNase-seq signals; the 4 discretized tracks used to define the 624 functional genomic fingerprints, including annotation-based CDSs, RNA-seq signal, DNase-seq signal and chromatin modifications; the fitCons scores based on those fingerprints (dark blue, with lighter blues less statistically significant); and, for comparison, phyloP-based conservation scores for mammals. (**a**) An apparent enhancer, marked by a combination of enhancer-associated chromatin modifications and a strong DNase-seq signal, displays elevated fitCons scores but no elevation in conservation scores. Many regulatory elements display such a pattern, either because they have arisen recently in evolutionary time or because errors in orthology detection or alignment result in spuriously low conservation scores. Here a ChIP-seq–supported transcription factor binding site for AP-1 (red arrow) and a lung cancer–associated SNP (green arrow) are highlighted. (**b**) CDS exons show elevated scores according to both fitCons and phyloP. (**c**) The 3′ UTR, marked by transcription-associated chromatin modifications, a high RNA-seq signal and an absence of DNase I hypersensitivity or CDS annotations, displays moderately elevated fitCons scores and patches of evolutionary conservation. fitCons scores are fairly well correlated with phyloP conservation scores[15] across the genome, with some notable exceptions in noncoding regions (**Supplementary Fig. 1**). Browser tracks are publicly available on the Cold Spring Harbor Laboratory mirror of the UCSC Genome Browser (hg19 assembly).

capturing complex, non-additive relationships among covariates. For example, the scores outside of CDSs increase with marginal DNase-seq (**Fig. 4a**) and RNA-seq (**Fig. 4b**) signals, as expected; yet, a closer examination shows that the scores actually decrease with DNase-seq intensity in the presence of high RNA-seq intensity, owing to an implicit partitioning of 5′ and 3′ UTRs by DNase-seq data (**Fig. 4c**). This example demonstrates that our exhaustive partitioning scheme allows the method to capture unanticipated relationships between functional genomic covariates and natural selection.

## Predictive power for *cis* regulatory loci

We evaluated the predictive power of fitCons scores for known cell type–specific regulatory elements in comparison with three widely used phylogenetic conservation scoring methods, the phastCons[12], phyloP[15] and Genomic Evolutionary Rate Profiling (GERP)[13] programs. In addition, we considered a new program, called Combined Annotation-Dependent Depletion (CADD)[35], that estimates the relative levels of pathogenicity of potential human variants using a support vector machine (SVM), many different genomic annotations and simulations of nucleotide divergence rates. Where appropriate, we also considered RegulomeDB, a scoring system for the regulatory potential of variant sites based on combined experimental and computational data[36], and EnhancerFinder, a kernel-based predictor for developmental enhancers based on multiple data types[37]. We evaluated the performance of these methods in predicting three types of functional elements that have putative roles in transcriptional regulation on the basis of different data sets: (i) binding sites for various transcription factors supported by ChIP-seq data from the Encyclopedia of DNA Elements (ENCODE) Project[3,28]; (ii) high-resolution expression quantitative trait loci (eQTLs)

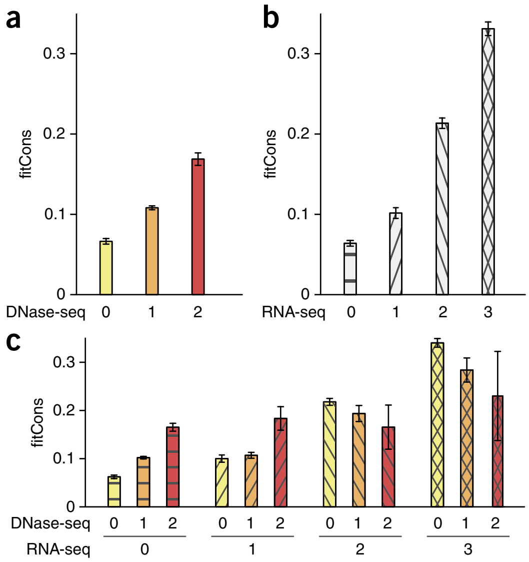

**Figure 4** Average fitCons scores as a function of DNase-seq and RNA-seq intensity. Results represent averages across all non-CDS clusters having the marginal or joint property of interest. Error bars represent standard errors of the aggregated scores (Online Methods). (**a**) fitCons scores increase with DNase-seq intensity, probably owing to an increasing density of *cis* regulatory elements: 0, no DNase-seq signal; 1, broad peaks; 2, narrow peaks. (**b**) fitCons scores increase with RNA-seq intensity: 0, no RNA-seq reads; 1–3, weak to strong RNA-seq signal (Online Methods). (**c**) fitCons scores behave in a non-additive manner as joint combinations of DNase-seq and RNA-seq intensity are considered. In particular, at medium to high RNA-seq read depth (classes 2 and 3), fitCons scores decrease (rather than increase) with increasing DNase-seq signal. This unexpected pattern is explained by enrichment for DNase I hypersensitivity near the 5′ ends of genes. Conditional on a high RNA-seq signal, a high DNase-seq signal tends to be associated with the 5′ UTRs and upstream regions of genes, which are under fairly weak selection, whereas a low DNase-seq signal is associated with 3′ UTRs, which are under stronger selection. Each bar in **a** summarizes 104 clusters, each bar in **b** summarizes 78 clusters and each bar in **c** summarizes 26 clusters.

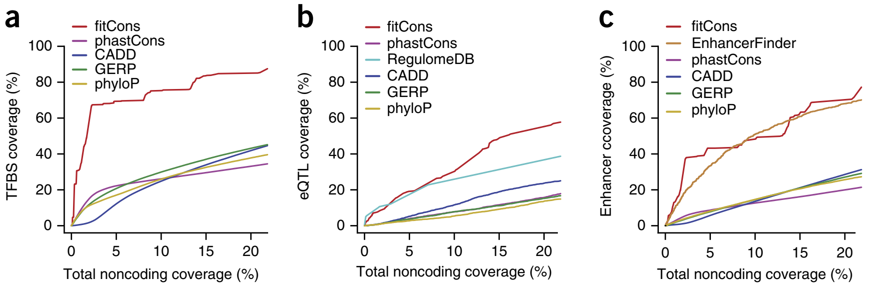

**Figure 5** Coverage of active *cis* regulatory elements as a function of total coverage of the noncoding genome. Coverage of each type of element is shown as the score threshold is adjusted to alter the total coverage of noncoding sequences in the genome, excluding sites annotated as CDSs or UTRs. fitCons is compared with scores from the CADD[35], GERP[13], phastCons[12] and phyloP[15] programs (Online Methods). (**a**) Coverage of 55,844 transcription factor binding sites detected by ChIP-seq in HUVECs[28]. (**b**) Coverage of high-resolution eQTLs identified in a recent large-scale study[6], restricted to 3,662 eQTLs associated with genes transcribed in HUVECs. Coverage of eQTLs is also shown for classification of single-nucleotide variants by RegulomeDB[36]. The divergence-based scores (phastCons, phyloP, GERP and CADD) all perform poorly on the eQTL data set, probably because the ascertainment for segregating sites creates a bias against evolutionary conservation. Note also that the apparent performance of RegulomeDB, particularly at low total noncoding coverage, is somewhat influenced by consideration of eQTL data in its scoring scheme. (**c**) Coverage of 462 enhancers identified by characteristic chromatin marks[38] assayed in HUVECs. Coverage of these enhancers by EnhancerFinder[37] predictions is also shown. In all three plots, the *x* axis represents coverage at 2.8 billion noncoding positions.

identified in a recent large-scale study[6]; and (iii) enhancers identified on the basis of characteristic chromatin marks[38] (see the Online Methods for details).

To place the different predictors on equal footing, we plotted the base-wise coverage of each type of regulatory element as a function of the total coverage of the noncoding genome, varying score thresholds to include 0–20% of noncoding sites (**Fig. 5**). This strategy allowed us to measure the extent to which the elements of interest displayed signals that rose above the background of the noncoding genome, in a uniform manner across scoring methods. By this test, the fitCons scores showed dramatically better sensitivity for noncoding elements than almost all of the other methods considered. For example, at a total noncoding coverage of 2.5%, fitCons scores achieved nearly 70% coverage of transcription factor binding sites, whereas the other methods all had less than 20% coverage. Similarly, the coverage of enhancers was about 40% at 2.5% noncoding coverage, whereas most other scoring methods showed almost no signal above background. Only EnhancerFinder, which is specifically designed for this task, showed comparable prediction performance on enhancers. We also performed a more traditional evaluation of the tradeoff between sensitivity and specificity using receiver operating characteristic (ROC) curves and found that fitCons scores were considerably better predictors of regulatory function than all other methods considered (Online Methods and **Supplementary Fig. 2**).

The tests above were based on regulatory elements that are putatively active in the cell type for which the scores were produced, to highlight the benefits of using cell type–specific functional data. To evaluate how well these advantages extended across cell types, we created an integrated fitCons score by combining information from three cell types (Online Methods) and evaluated the performance of this score in predicting regulatory elements pooled from multiple cell types. We found that, in this less favorable setting, the fitCons scores still had better predictive performance for *cis* regulatory elements than any of the other scoring methods (**Supplementary Fig. 3**).

To address possible deficiencies of these tests, we carried out two additional sets of validation experiments. First, we performed a second round of experiments on ChIP-seq–supported transcription factor binding sites that considered only the subset of nucleotide positions at which base preferences were especially strong, which should be enriched for bases having fitness consequences. The ROC curves based on this more stringent test were very similar to the original curves (**Supplementary Fig. 4**), demonstrating that the apparent performance of the fitCons scores was not artificially inflated by the coarse-grained nature of our scores and transcription factor binding sites. Second, we examined an alternative set of predicted enhancers for GM12878 cells identified on the basis of characteristic patterns of divergent transcription initiation[39]. Unlike the chromatin-based enhancer predictions described above, these predictions were based on data completely independent from those underlying the fitCons scores. Nevertheless, the fitCons scores still displayed excellent predictive power for this set, better than all other methods besides EnhancerFinder (**Supplementary Fig. 5**).

## Proportion of the human genome under selection

The proportion of nucleotides in the human genome that directly influence fitness—sometimes called the 'share under selection' (SUS)—has primarily been estimated using methods that consider divergence patterns among mammals, for which turnover of functional elements might be an important confounding factor[40–44]. In addition to being useful as predictors of function, the fitCons scores could be useful in obtaining estimates of the SUS that are less sensitive to turnover because they measure natural selection over much shorter time scales.

An initial estimate of the SUS can be obtained by simply averaging the fitCons scores across all nucleotide positions in the genome. Because each score represents a probability that an individual nucleotide influences fitness, the average of these scores represents an expected fraction of nucleotides in the genome having fitness-influencing functions, or an expected SUS. This approach yielded an estimate of 7.5% (±0.1%) for HUVECs or 7.5–7.8% across the cell types. These estimates are largely consistent with but on the high end of those based on cross-species divergence, which generally have fallen between 3 and 8% (refs. 12,40,44–46). Among the sites under selection, we estimate that 9.0% are in CDSs, 2.2% are in 3′ UTRs, 35.2% are in introns, 51.7% are in intergenic regions and <1% are in each of several other noncoding annotation classes (**Supplementary Table 2**). Our estimates of the SUS are somewhat lower than previous estimates that have explicitly allowed for evolutionary turnover, most of which have been two to three times higher than the pan-mammalian estimates of ~5% (refs. 26,44,46–48). However, they are similar to a recent estimate of 7.1–9.2% based on improved alignments and a new model for turnover[49].

Violations of modeling assumptions will tend to bias fitCons scores upward, particularly for functional classes for which the true fraction is close to zero (**Supplementary Note**). To address this problem, we performed a parallel calculation for 'neutral' sites that intersected the large class of genomic positions having a 'null' functional genomic

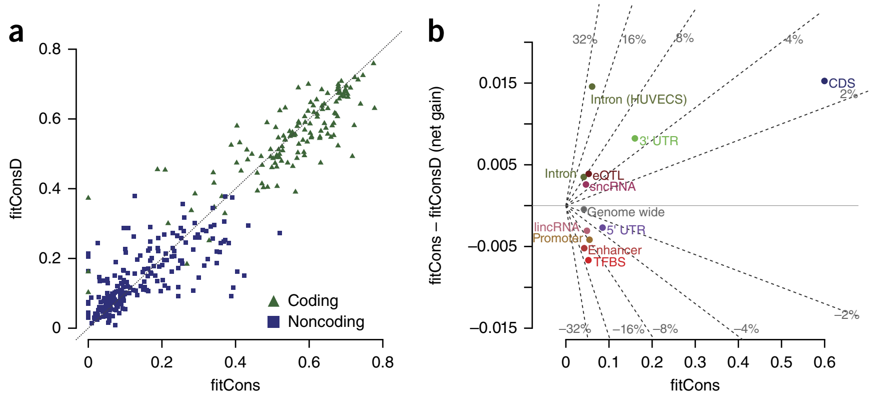

**Figure 6** Comparison between fitCons and fitConsD scores. fitConsD is an alternative estimate of fitness consequences, analogous to fitCons but based on an estimator of the fraction of sites under natural selection that considers divergence patterns across four primate genomes (Online Methods). (**a**) fitCons and fitConsD scores are shown for the clusters defined using functional genomic data from HUVECs. Scores are shown for the 348 clusters of size 10 kb or larger, distinguishing between coding clusters (green triangles) and noncoding clusters (blue squares). Both sets of scores are corrected by subtracting the possible contribution from model misspecification (Online Methods). Correlation between the two sets of scores is high overall ($R^2$ = 0.88) and is somewhat higher for coding ($R^2$ = 0.69) than for noncoding ($R^2$ = 0.51) clusters. (**b**) The net gain in the fraction of sites under selection on population genetic time scales relative to primate divergence time scales, computed by subtracting average fitConsD scores from average fitCons score for different classes of functional elements (negative values imply net loss). Net gain is plotted against average fitCons score, and lines of constant slope radiating from the origin represent constant values of a 'net gain rate' per functional site, computed as NGR = (fitCons – fitConsD)/fitCons. The NGR is small (≤10%) for almost all annotation classes considered, with the main exception being the introns of active genes (NGR > 20%; see "Intron (HUVECs)"), which are enriched in clusters that exhibit an absence of DNase-seq or RNA-seq signal and chromatin modifications, suggesting transcriptional elongation.

fingerprint (no DNase-seq, RNA-seq or histone modification signal). This calculation resulted in an estimate of 3.3%, which can be considered an upper bound on the contribution of error because these putatively neutral sites undoubtedly include some sites under selection. By subtracting this 3.3% from our naive estimate of 7.5%, we obtained an estimated lower bound for the SUS of 4.2%, with somewhat higher fractions of selected sites in CDSs and 3′ UTRs (**Supplementary Table 2**). (These estimates are for HUVECs, but the results for the other cell types were very similar.) Overall, our analysis of the SUS suggests that between 4.2 and 7.5% of nucleotides in the genome have direct fitness-influencing functions and that the ratio of noncoding to coding functional sites is between 5.4 and 10.1.

### Implications for evolutionary turnover of functional elements

To better understand the differences between fitCons scores and conventional divergence-based scores, we devised an alternative scoring system (denoted fitConsD) based on the same site clusters but an estimator of the fraction of nucleotides under selection that instead considers nucleotide divergence patterns across primates (Online Methods). Thus, the fitCons and fitConsD scores both represent probabilities of fitness consequences per nucleotide but over two different evolutionary time scales. Overall, these two measures were remarkably well correlated, with $R^2$ = 0.88 (**Fig. 6a**). Furthermore, a measure based on the difference between fitConsD and fitCons scores suggested relatively low amounts of turnover across annotation classes, accounting for no more than about 10% of all functional sites (**Fig. 6b**). These observations suggest that the main signal for selection has been maintained over long evolutionary time periods and that turnover has been modest during primate evolution but that there are some classes of sites that show stronger recent than ancient natural selection.

## DISCUSSION

The essential idea of our approach is to use functional genomic data to group sites into classes that are relatively homogeneous in terms of their functional roles, then to characterize the bulk influence of natural selection on these classes on the basis of their patterns of polymorphism and divergence. For our estimation of natural selection, we make use of a recently developed probabilistic model of evolution and efficient algorithms for genome-wide inference (INSIGHT). We interpret INSIGHT-based estimates of fractions of nucleotides under selection as probabilities that each nucleotide influences fitness, or fitness consequence (fitCons) scores. Even with a simple clustering scheme, these fitCons scores appear to be highly informative about genomic function.

According to our experiments, fitCons scores have excellent predictive performance for putative *cis* regulatory elements, outperforming several divergence-based methods (phastCons, phyloP, GERP and CADD) and one annotation-based method (RegulomeDB) by clear margins. They also performed slightly better in enhancer prediction than EnhancerFinder, a program specifically designed for this purpose, although it should be noted that EnhancerFinder was trained on other cell types. Notably, prediction performance does not appear to be sensitive to the choice of neutral sites used by INSIGHT (**Supplementary Figs. 6** and **7**). In part, the observed improvement in performance reflects the use of cell type–specific data (**Fig. 5** and **Supplementary Fig. 2**), but fitCons scores also show a clear performance advantage when considering all annotated elements rather than just active ones (**Supplementary Fig. 3**). Thus, the approach of grouping genomic sites by functional genomic signatures and then measuring group-wise fitness consequences on the basis of patterns of genetic variation appears to offer real benefits for the prediction of regulatory function, as compared with methods that consider either genetic divergence or functional genomic data alone.

Interestingly, the recently published CADD method performed no better on our tests than conventional conservation scores, despite reports by the authors of substantial advantages over phyloP, phastCons, GERP and other methods[35]. This inconsistency appears to reflect several important differences between our validation experiments and those they reported. First, our tests focused specifically on putative *cis* regulatory elements, whereas many of their tests considered a mixture of coding and noncoding elements. In particular, the ClinVar database, which figured prominently in their experiments, includes very few noncoding variants (~5% of pathogenic variants). Second, when Kircher *et al.* did consider noncoding regions, they generally did not distinguish between *cis* regulatory elements and sequences that more directly influence the structure and content of protein-coding transcripts, such as splice sites. CADD has a natural advantage with these variants owing to its use of gene annotations, whereas the annotation-free fitCons scores may perform better in completely unannotated regions of the genome. Finally, the tests by Kircher *et al.* that explicitly considered putative *cis* regulatory elements were limited to a few loci and examined only correlations with saturation mutagenesis experiments, irrespective of a prediction threshold. We view our ROC-type comparisons based on multiple independent genome-wide sets of elements as a more direct and comprehensive demonstration of predictive power for *cis* regulatory elements. In any case, the comparison of these two closely related yet distinct approaches helps to identify strengths and weaknesses of each and may lead to new ideas for improved methodologies.

A side benefit of our model-based approach is that the base-wise probabilities of fitness consequences lead in a straightforward manner to an estimate of the SUS in the human genome. This estimate of the SUS reflects time scales since the divergence of humans and chimpanzees, about 4–6 million years ago, unlike conventional estimates based on tens or hundreds of millions of years of mammalian evolution. Nevertheless, our estimate of the SUS, at 4.2–7.5%, ends up being remarkably similar to those based on longer time scales, which have generally fallen between 3 and 8% (refs. 12,40–43,45,50). It also overlaps with a recent estimate of 7.1–9.2% based on patterns of insertion and deletion and an explicit model of evolutionary turnover[49]. We take the general concordance of these estimates, both with one another and with our fitCons- and fitConsD-based estimates, as a strong indication that the SUS has remained quite low (probably <10%) over various time scales in mammalian evolution. This finding stands in contrast to estimates that ~80% of nucleotides may be functional, based on measures of 'biochemical activity' (ref. 3). However, it is important to bear in mind that these evolutionary and biochemical estimates reflect somewhat different definitions of function, and this may explain some of the difference between them[16,18,19]. For example, the fitCons- and conservation-based estimates (excluding those based on indels) generally represent the fractions of positions at which point mutations will have fitness consequences, but they do not account for sequences (such as spacer elements) that would have fitness consequences if deleted but not mutated (see the **Supplementary Note** for discussion).

Apart from the absolute fraction of functional DNA in the human genome is the question of how much the functional content of the genome has changed over time through gains and losses of functional elements. Several studies have estimated that such turnover could allow the current SUS in the human genome to be ~2–3 times larger than estimated from comparisons across mammals[26,44,46–48]. Indeed, these findings have been proposed to explain, in part, the discordance between evolution-based and biochemical estimates of the functional fraction of the genome[26,51,52]. However, most of these analyses have accounted for turnover using relatively crude methods, for example, by relying on an apparently near-linear relationship between pairwise divergence and the estimated SUS[46,47] or by estimating functional content from mean SNP densities or derived allele frequencies in genomic regions not conserved across mammals[26] (but see ref. 49 for an improved model). Our analysis is more direct, by comparing analogous divergence-based and polymorphism-based estimates of the SUS calculated from exactly the same clusters of nucleotide positions. In addition, our analysis focuses on primate evolution, rather than attempting to account for turnover across mammals, where factors such as alignment error, orthology detection and genomic rearrangement can be problematic. The similarity between our estimates based on polymorphism (fitCons) and divergence (fitConsD) strongly suggests that evolutionary turnover has been modest during primate evolution, as massive turnover would be expected to lead to a substantial downward bias in the divergence-based estimates. Our power experiments indicate that this observation is not an artifact of reduced sensitivity in the fitCons scores. Nevertheless, we cannot rule out the possibility that compensating gains and losses on very recent time scales maintain a similar SUS while substantially altering the genomic composition of functional sequences.

We have focused on HUVECs in this report, but we also generated fitCons scores for two other cell types (H1 hESCs and GM12878 cells). A comparison across cell types (**Supplementary Note**) indicated that the genomic positions assigned to each functional class differed substantially across cell types, but equivalently defined clusters had concordant fitCons scores in the different cell types (**Supplementary Fig. 8**). When cell type–specific scores were examined, elements active in that cell type displayed significantly higher scores than inactive elements. Moreover, particular elements had higher scores in cell types for which they were active than in cell types for which they were inactive (**Supplementary Fig. 9**). Notably, we found that a set of integrated scores based on a simple, heuristic procedure (Online Methods) performed nearly as well as the cell type–specific scores in the target cell types but much better on elements from mismatched or pooled cell types (**Supplementary Fig. 10**). With more flexible and scalable clustering techniques, it may be possible to improve these methods by considering all cell types simultaneously, clustering sites by functional genomic fingerprints corresponding to multiple cell types and then producing a single set of scores reflecting these joint patterns. Such improvements, together with increases in the resolution and quality of the available functional genomic data, should result in improved power for the prediction of individual functional elements and refined estimates of the SUS.

**URLs.** Cold Spring Harbor Laboratory mirror of UCSC Genome Browser, http://genome-mirror.cshl.edu/; UCSC Genome Browser, http://genome.ucsc.edu/; INSIGHT, http://compgen.cshl.edu/INSIGHT/; GENCODE v15, ftp://ftp.sanger.ac.uk/pub/gencode/release_15/; GERP, http://mendel.stanford.edu/SidowLab/downloads/gerp/; CADD, http://cadd.gs.washington.edu/download; RegulomeDB, http://regulome.stanford.edu/downloads/; Gerstein laboratory ENCODE nets, http://encodenets.gersteinlab.org/; European Bioinformatics Institute's E-GEUV-1 data set, http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/analysis_results/.

## 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 L. Arbiza for helpful discussions and assistance with early analyses and G. Cooper for constructive criticism of our validation experiments and comparisons with CADD. This research was supported by US National Institutes of Health grant GM102192, a David and Lucile Packard Fellowship for Science and Engineering (to A.S.) and a postdoctoral fellowship from the Cornell Center for Comparative and Population Genomics (to I.G.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

### AUTHOR CONTRIBUTIONS

I.G. and A.S. conceived the study framework. B.G. and I.G. performed the experiments. All authors analyzed the data. B.G., M.J.H. and I.G. developed analysis tools. B.G., I.G. and A.S. wrote the manuscript. I.G. and A.S. supervised the research.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Mardis, E.R. A decade's perspective on DNA sequencing technology. *Nature* **470**, 198–203 (2011).
2. Wold, B. & Myers, R.M. Sequence census methods for functional genomics. *Nat. Methods* **5**, 19–21 (2008).
3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
4. Shen, Y. *et al.* A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).
5. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
6. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
7. 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).
8. Mayor, C. *et al.* VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046–1047 (2000).
9. Margulies, E.H., Blanchette, M., Program, N.C.S., Haussler, D. & Green, E.D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507–2518 (2003).
10. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
11. Ovcharenko, I., Boffelli, D. & Loots, G.G. eShadow: a tool for comparing closely related sequences. *Genome Res.* **14**, 1191–1198 (2004).
12. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
13. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
14. Asthana, S., Roytberg, M., Stamatoyannopoulos, J. & Sunyaev, S. Analysis of sequence conservation at nucleotide resolution. *PLOS Comput. Biol.* **3**, e254 (2007).
15. 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).
16. Graur, D. *et al.* On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* **5**, 578–590 (2013).
17. Niu, D.K. & Jiang, L. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem. Biophys. Res. Commun.* **430**, 1340–1343 (2013).
18. Doolittle, W.F. Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. USA* **110**, 5294–5300 (2013).
19. Eddy, S.R. The ENCODE project: missteps overshadowing a success. *Curr. Biol.* **23**, R259–R261 (2013).
20. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
21. Fay, J.C., Wyckoff, G.J. & Wu, C.I. Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234 (2001).
22. Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**, 1149–1152 (2005).
23. Eyre-Walker, A., Woolfit, M. & Phelps, T. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* **173**, 891–900 (2006).
24. Boyko, A.R. *et al.* Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* **4**, e1000083 (2008).
25. Wilson, D.J., Hernandez, R.D., Andolfatto, P. & Przeworski, M. A population genetics–phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet.* **7**, e1002395 (2011).
26. Ward, L.D. & Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337**, 1675–1678 (2012).
27. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
28. Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nat. Genet.* **45**, 723–729 (2013).
29. Narlikar, L. *et al.* Genome-wide discovery of human heart enhancers. *Genome Res.* **20**, 381–392 (2010).
30. Ritchie, G.R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).
31. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).
32. Hoffman, M.M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).
33. Hoffman, M.M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41**, 827–841 (2013).
34. Gronau, I., Arbiza, L., Mohammed, J. & Siepel, A. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. *Mol. Biol. Evol.* **30**, 1159–1171 (2013).
35. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
36. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using Regulome DB. *Genome Res.* **22**, 1790–1797 (2012).
37. Erwin, G.D. *et al.* Integrating diverse datasets improves developmental enhancer prediction. *PLOS Comput. Biol.* **10**, e1003677 (2014).
38. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
39. Core, L.J. *et al.* Analysis of nascent RNA identifies a unified architecture of transcription initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
40. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
41. Cooper, G.M. *et al.* Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* **14**, 539–548 (2004).
42. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
43. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
44. Ponting, C.P., Nellaker, C. & Meader, S. Rapid turnover of functional sequence in human and other genomes. *Annu. Rev. Genomics Hum. Genet.* **12**, 275–299 (2011).
45. Chiaromonte, F. *et al.* The share of human genomic DNA under selection estimated from human-mouse genomic alignments. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 245–254 (2003).
46. Meader, S., Ponting, C.P. & Lunter, G. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res.* **20**, 1335–1343 (2010).
47. Smith, N.G., Brandstrom, M. & Ellegren, H. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* **84**, 806–813 (2004).
48. Ponting, C.P. & Hardison, R.C. What fraction of the human genome is functional? *Genome Res.* **21**, 1769–1776 (2011).
49. Rands, C.M., Meader, S., Ponting, C.P. & Lunter, G. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet.* **10**, e1004525 (2014).
50. Lunter, G., Ponting, C.P. & Hein, J. Genome-wide identification of human functional DNA using a neutral indel model. *PLOS Comput. Biol.* **2**, e5 (2006).
51. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. USA* **111**, 6131–6138 (2014).
52. Pheasant, M. & Mattick, J.S. Raising the estimate of functional human sequences. *Genome Res.* **17**, 1245–1253 (2007).

## ONLINE METHODS

**Functional genomic data.** RNA-seq and DNase-seq data for HUVECs, H1 hESCs and GM12878 cells were downloaded from the UCSC Genome Browser. Chromatin states for the same three cell types were downloaded from the European Bioinformatics Institute's FTP site (see **Supplementary Table 3**). For DNase-seq, we considered two replicate experiments from University of Washington (UW) data for each cell type. However, only one UW replicate was available for H1 hESCs, so additional DNase-seq data for this cell line was obtained from Duke University. For each replicate DNase-seq experiment, we downloaded broad and narrow peak calls. For RNA-seq, we selected a single replicate from the Caltech poly(A)$^+$ 75-bp paired-end read data, after examining several alternative data sets. For chromatin states, we used the 25-state ChromHMM segmentation generated in December 2012 (ref. 33).

**Clustering approach.** We produced a separate partitioning for each cell type on the basis of the functional genomic data. The broad and narrow DNase-seq peaks were used to partition sites in the genome into three mutually exclusive classes: sites that fell in a narrow peak in both replicate experiments (class 2); sites that fell in a broad peak in at least one replicate and did not fall in a narrow peak in both replicates (class 1); and sites that fell outside of all called peaks (class 0). This three-level scheme allowed for both high sensitivity (class 1) and high specificity (class 2). For H1 hESCs, only one set of broad peak calls was available to define class 1. For the RNA-seq data, we partitioned sites in the genome into four mutually exclusive classes (0–3) on the basis of the number of reads aligned at each position. Read depth thresholds were set separately for each cell type through a process that aims to minimize the conditional entropy of concentrations of predicted sites under selection (**Supplementary Note**). Chromatin states were defined directly from the 25 states in ChromHMM, with a 26th state containing sites not assigned to any chromatin class. The Cartesian product of these partitions, together with the partition into coding and noncoding sequences, resulted in $3 \times 4 \times 26 \times 2 = 624$ distinct functional classes.

**Running INSIGHT.** The INSIGHT method infers the fraction of nucleotide sites under selection ($\rho$) for a given collection of sites by comparing patterns of within-species polymorphism and between-species divergence within these sites and within putatively neutrally evolving sites nearby. A detailed description of the method, sequence data and data quality filters is given in ref. 34. For each non-empty fitCons site cluster, INSIGHT was used to estimate $\rho$, which was used as the fitCons score of all sites in that cluster. To reduce sensitivity to estimates with high uncertainty, we filtered out clusters for which the estimated standard error was greater than 40% of the estimated value of $\rho$. To increase computational efficiency, clusters larger than 20 Mb in size were partitioned into smaller subclasses, and estimates of $\rho$ were computed as weighted averages (weighted by the number of informative sites) across subclasses.

**Neutral sites.** The collection of sites predicted to be free from the influence of natural selection (neutral sites) was derived from a set identified previously[28,34,53]. Briefly, this set was obtained by eliminating from all genomic sites those likely to be under direct natural selection, including (i) exons of annotated protein-coding genes and the 1,000 bp flanking them on either side; (ii) RNA genes from GENCODE v11 and the 1,000 bp flanking them; and (iii) conserved noncoding elements (identified by phastCons) and the 100 bp flanking them. This set was used in both the INSIGHT analysis and the power analysis.

**GENCODE annotations.** Transcript annotations from GENCODE v15 (ref. 54) were downloaded from the Sanger Institute's FTP server and used to define eight site classes: CDSs, 5′ UTRs, 3′ UTRs, promoters, introns, lincRNAs, sncRNAs and intergenic (sites not falling within any protein-coding transcription unit). Transcripts annotated with feature type = "CDS" and gene type = "protein coding" were used to define the CDS set for fitCons. For subsequent analysis, we used a slightly more conservative set, obtained by additionally requiring feature type = "gene," gene status = "KNOWN," transcript status = "KNOWN" and the identification of both start and stop codons within the transcript. UTRs were defined from transcripts having feature type = "UTR" and gene type = "protein coding" and were designated as 5′ or 3′. Introns were defined by positions that fell within a protein-coding transcript but outside of the CDS and UTRs. Promoters were defined as the 1,000 bp immediately upstream of the first (most upstream) transcription start site for each protein-coding gene. A similarly defined alternative set of 100-bp promoter regions was used in assessing differences between cell types (**Supplementary Fig. 9**). lincRNAs were identified by transcripts with feature type = "exon" and gene type = "lincRNA." Similarly, sncRNAs consisted of transcripts with feature type = "exon" and gene type ∈ {"miRNA," "snRNA," "snoRNA"}. Positions in the more inclusive CDS set were removed from all noncoding classes.

***Cis* regulatory elements.** Transcription factor binding sites were drawn from a set for 78 transcription factors, based on ChIP-seq data from ENCODE[28] downloadable from our UCSC Genome Browser mirror. This set contained roughly 1.4 million binding sites of a mean length of 11 bp, each of which was associated with the cell types in which it was detected. For some tests, we considered only the subset of nucleotide positions inside these transcription factor binding sites that corresponded to motif positions with strong base preferences, defined as those positions at which the consensus allele appeared in at least 90% of all binding sites (according to the inferred motif model). For enhancers, we used the distal regulatory modules described in ref. 38. We downloaded the file enets4.Distal_cell_line.txt from the Gerstein laboratory ENCODE nets and extracted from it a total of 19,005 enhancer-transcript associations, covering 5,834 unique autosomal loci with a mean length of 888 bp, along with the cell types associated with each predicted enhancer. The eQTLs described in ref. 6 were downloaded from the European Bioinformatics Institute's E-GEUV-1 data set. We used the 4 files {EUR373,YRI89}.{exon,gene}.cis.FDR5.best.rs137.txt.gz to identify 6,760 distinct autosomal positions and the associated transcripts, removing all positions overlapping CDSs.

**Identifying active elements per cell type.** In several analyses, we considered the subset of elements in each annotation class for which we had evidence of activity in a given cell type. To identify the cell types in which transcription factor binding sites and enhancers were active, we used the cell type designations provided in the corresponding annotation files. For other classes of elements, we defined the active elements using a set of GENCODE transcripts and genes that showed significantly elevated levels of RNA transcription in the Caltech RNA-seq data. These were transcripts (or genes) for which the 95% confidence interval of the normalized read count in a given cell type fell within the top one-third of the normalized read counts for transcripts (or genes) across all three cell types (with thresholds of 1.477 for transcripts and 4.966 for genes). Active eQTLs were identified via associated active genes using the GENCODE gene identifier specified for each eQTL. Active promoters, UTRs, CDSs and introns were identified via associated active transcripts. For the comparison between cell types (**Supplementary Fig. 9**), we also used collections of eQTLs and promoters found to be inactive in a given cell type. These were defined in a similar way, by using transcripts and genes falling in the bottom third of the distribution of normalized read counts.

**Comparison with other scores.** Base-wise scores from the GERP[13] method, the CADD[35] method, and the phastCons[12] and phyloP[15] methods were downloaded from the respective websites (see URLs; file hg19.GERP_scores.tar.gz generated in August 2010 for GERP, file whole_genome_SNVs.tsv.gz downloaded in September 2013 for CADD and the UCSC Genome Browser 46 placental mammal conservation tracks for phastCons and phyloP). CADD scores are specified for each genomic position and each of the three possible variant bases at that position. We took the maximum of these three scores, which yielded the best performance for the CADD method in our comparisons. We also used RegulomeDB[36] (downloaded in January 2013) to rank SNPs, such as eQTLs, into 1 of 13 categories according to evidence from functional genomic data. Finally, we obtained EnhancerFinder scores[37] for 1,500-bp windows tiled across the genome directly from the authors. We used the general, non-tissue-specific scores and averaged them at positions contained in multiple overlapping windows.

**Receiver operating characteristic curves.** We used ROC curves to measure the ability of each scoring scheme to discriminate between functional and nonfunctional regulatory elements. For transcription factor binding sites and enhancers, we used the annotations described above as true positives and

doi:10.1038/ng.3196

defined true negatives from our filtered, putatively neutral sites. For eQTLs, our negative set consisted of all 9.8 million variants tested in ref. 6, excluding indels, non-simple variants and positions that showed possible associations at a threshold of nominal $P < 0.05$ (7.6 million SNPs remained). In all three cases, we additionally removed any sites in the positive set from the negative set. A point on a ROC plot indicates the fraction of the annotated genomic positions with scores higher than a given score (true positive rate) versus the fraction of control genomic positions with scores higher than that score (false positive rate). Positions with no scores were ignored when computing fractional coverage.

**Integrating fitCons scores across cell types.** We generated a series of fitCons scores that integrate functional genomic data across the 3 cell types by using the original 624 fingerprints and altering the rule by which sites are assigned to clusters to reflect information from multiple cell types. Our approach attempts to select a fingerprint for each site that is likely to be most informative about the site's function, while avoiding a bias toward higher scores with an increasing number of cell types. See the **Supplementary Note** for details.

**Share under selection.** Assume a partitioning of the genome into $K$ mutually exclusive and exhaustive clusters, $C_1$, $C_2$, …, $C_K$, and a corresponding set of fitCons scores, $\rho(C_1)$, $\rho(C_2)$, …, $\rho(C_K)$. Note that the expected number of genomic positions under selection in cluster $C_i$ is given by $\rho(C_i)|C_i|$ because $\rho$ is an estimate of the fraction of sites under selection. For an arbitrary collection of sites $S$, the expected number of sites in $S$ that are under selection is given by $\text{sel}(S) = \Sigma_i \rho(C_i)|C_i \cap S|$, and the average fitCons score for $S$ is given by $\rho(S) = \text{sel}(S)/|S|$. To avoid underestimation of $\rho(S)$, we do not filter out fitCons scores with high uncertainty in these calculations, as we do for other analyses. In addition, to account for possible overestimation of $\rho$ in very large clusters having low fractions of sites under selection, we ran INSIGHT on the intersection of our neutral sites and all noncoding sites in a 'quiescent' chromatin state with no DNase-seq or RNA-seq signal. We then subtracted the estimated value of $\rho$, denoted $\rho_{\text{neut}}$, from the raw fitCons score to obtain a conservative lower bound, $\rho(S) - \rho_{\text{neut}}$, for the fraction of sites under selection in $S$.

**fitConsD and evolutionary turnover.** To make the comparison between fitCons and fitConsD as direct as possible, fitConsD scores were computed using the same pipeline we developed for fitCons (**Fig. 1**), except that in step 3 we replaced the INSIGHT model with an evolutionary model that considers sequence divergence between the human, chimpanzee, orangutan and rhesus macaque genomes. fitConsD scores are based on an estimate $s_i$ for the relative evolutionary rate of each cluster $C_i$ in comparison with a neutral model globally estimated for the four-primate phylogeny. This relative rate is then compared with the relative rate $s_{i\,\text{neut}}$ estimated for the putative neutral regions flanking sites in $C_i$, and a divergence-based estimate of the fraction of sites under selection in cluster $C_i$ is given by $\rho_{\text{div}}(C_i) = 1 - s_i/s_{i\,\text{neut}}$ (**Supplementary Note**).

[53] Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034 (2011).

[54] Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

doi:10.1038/ng.3196