Probabilities of Fitness Consequences for Point Mutations Across the Human Genome

Brad Gulko[1], Ilan Gronau[2], Melissa J. Hubisz[2], Adam Siepel[2,*]

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

[2]Dept. of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA

[*]E-mail: acs4@cornell.edu

Abstract

The identification of noncoding functional elements based on high-throughput genomic data remains an important open problem. Here we describe a novel computational approach for estimating the probability that a point mutation at each nucleotide position in a genome will influence organismal fitness. These *fitness consequence* (fitCons) scores can be interpreted as an evolution-based measure of potential genomic function. We first partition the genome into clusters of positions having distinct functional genomic "fingerprints," based on cell-type-specific DNase-seq, RNA-seq, and histone modification data. Then we estimate the probability of fitness consequences for each cluster from associated patterns of genetic polymorphism and divergence using a recently developed probabilistic method called INSIGHT. We have generated fitCons scores for three human cell types based on publicly available genomic data and made them available as UCSC Genome Browser tracks. Like conventional evolutionary conservation scores, fitCons scores are clearly elevated in known coding and noncoding functional elements, but they show considerably better sensitivity than conservation scores for many noncoding elements. In addition, they perform exceptionally well in distinguishing ChIP-seq-supported transcription factor binding sites, expression quantitative trait loci, and predicted enhancers from putatively nonfunctional sequences. The fitCons scores indicate that 4.2–7.5% of nucleotide positions in the human genome have influenced fitness since the human-chimpanzee divergence. In contrast to several recent studies, they suggest that recent evolutionary turnover has had a relatively modest impact on the functional content of the genome. Our approach provides a unique new measure of genomic function that complements measures based on evolutionary conservation or functional genomics alone and is particularly well suited for characterizing turnover and evolutionary novelty.

Introduction

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 sets describing many different 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 important recent progress [3–6], remain poorly annotated and understood. The challenge of making full use of rich new data sets to characterize the functional roles of these sequences is central in many applications in genetics, ranging from uncovering the mechanisms behind the various processes by which gene expression is regulated, to identifying causal variants for phenotypes in genome-wide association (GWA) and expression quantitative trait locus (eQTL) studies, to predicting the consequences of individual mutations for personalized medicine [7].

One of the most powerful techniques available for gaining functional insights from sequence data makes use of the simple 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. Thus, sequences that are significantly more conserved between species than would be expected under neutral drift are likely to be functional. Many computational methods for functional element identification have been developed around this basic idea [7–15]. A major strength of this approach is that it sidesteps 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 conservation-based methods is the one conducted directly on genomes by nature over millenia, 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 discordances between sequence orthology and functional homology and changes in selective pressure that substantially complicate this type of analysis. As a consequence, methods of this kind tend to work well for relatively "stable" classes of elements such as protein-coding genes, microRNAs, and developmental enhancers [20–23], but less well for elements that appear to turn over more rapidly, such as long noncoding RNAs, tissue-specific enhancers, and many transcription factor binding sites [24–28].

This important limitation of conservation-based methods has led to two major alternative strategies for computational identification and characterization of functional elements, particularly in noncoding regions. The first strategy is to augment or replace information about evolutionary conservation from comparisons between species with information about genetic polymorphism within a target species of interest [29–37]. This *polymorphism-based approach* is more robust to turnover, which is much less prominent on the shorter evolutionary time scales associated with intraspecies variation. In addition, the challenges associated with genomic alignment and orthology identification are substantially reduced on these time scales. A major problem with this approach, however, is that polymorphic sites

tend to be sparse along the genome, so that some type of pooling of information—either across nucleotide sites within a genomic region, or across dispersed genomic elements of a particular type—is generally needed to support statistical inference. Furthermore, it is often not clear how to pool this genomic information in the absence of good-quality genomic annotations, as in most noncoding regions of eukaryotic genomes.

The second strategy is to forgo the use of evolutionary information and instead to predict functional roles from genomic data alone, typically by using machine-learning methods for supervised classification [38, 39] or clustering followed by labeling based on similar known examples [40–42]. This *functional genomic approach* has been useful in a number of cases, but it has the limitation that it depends strongly on previously characterized elements, which, in noncoding regions, are typically small in number and perhaps unrepresentative of the genome as a whole. Thus, this approach is more effective for identifying new examples of known types of elements than for discovering completely new classes of elements.

In this paper, we introduce a method for annotating both coding and noncoding regions of a genome that combines many of the strengths of the polymorphism-based and functional genomic approaches. Like functional genomic methods, our approach groups genomic regions by combinations of signals obtained from high-throughput functional genomic assays. However, instead of relying on known examples to classify the regions displaying each functional genomic "fingerprint," it characterizes these regions in terms of probabilities of mutational fitness consequences inferred from patterns of genetic variation. Thus, the method integrates both evolutionary and functional data in characterizing the potential functional importance of genomic regions.

For the evolutionary component of our approach, we make use of a recently developed method, called Inference of Natural Selection from Interspersed Genomically coHerent elemenTs (INSIGHT), that contrasts patterns of polymorphism and divergence in a collection of dispersed genomic elements with those in flanking neutral sites. INSIGHT accounts for both negative and positive selection and relies on a full probabilistic model of evolutionary change on both phylogenetic and population genetic time scales [43]. A key parameter from INSIGHT, called ρ, describes the fraction of nucleotides in a given class of genomic regions that are under the influence of natural selection. Our method assigns an estimate of ρ ($0 \leq \rho \leq 1$) to each nucleotide in the genome and treats this estimate as a *fitness consequence* (fitCons) score. In precise terms, the fitCons score at each genomic position i represents the probability that the nucleotide at i influences the fitness of the organism that carries it, as estimated from patterns of genetic variation in the set of all genomic positions that have the same functional genomic fingerprint as position i.

The paper is organized as follows. We first detail our strategy for generating genome-wide fitCons scores and then describe an analysis of scores based on data from the ENCODE project. We compare these fitCons scores with scores based on cross-species conservation, and evaluate their power for identifying likely regulatory sequences. We then use these scores to provide an independent estimate of the fraction of the human genome subject to direct influence by natural selection since the divergence of the human and chimpanzee lineages. We also examine evidence for evolutionary turnover by comparing the standard fitCons scores with similar scores based on divergence alone. Finally, we compare the fitCons scores for different cell types and show that they are consistent in many respects but do contain some useful cell-type-specific information.

Results

Conceptual Basis of Method

We begin by outlining several key concepts underlying the fitCons scores. Our first observation is that information about genetic variation can be used to estimate probabilities of fitness consequences for moderately large groups of genomic positions, but not for individual positions. We refer to this property as "groupwise predictivity" and contrast it with the more powerful "individual predictivity." Groupwise predictivity is an inevitable price of measuring the influence of evolution on short time scales, owing to the sparsity of informative sites along the genome.

Groupwise predictivity is a common feature of many problems in statistics and machine learning, with applications in diverse fields ranging from e-commerce to insurance to drug discovery, but often the key features influencing the outcome of interest are obvious and the groupings can be set *a priori*. Consider the problem of estimating the expected risk of an automobile accident in order to set an insurance premium. This is an example of a problem with groupwise predictivity, because there is limited information about accident risk from most individual drivers' records but accident rates for large groups of drivers of similar types can be estimated with high accuracy. In this case, however, the relevant features—such as the age, sex, and number of traffic violations of the driver—are more or less obvious. By contrast, the genomic "risk factors" for fitness-influencing mutations, particularly in noncoding regions, are much less clear. Thus, in our application, it is necessary not only to predict the fitness consequences of mutations within each group, but also to discover a grouping scheme that enables accurate predictions to be obtained.

A common assumption in sequence analysis is that assignments are grouped along the genome sequence, often via a first-order Markov process. This amounts to an assumption of spatial grouping in one dimension. Spatial grouping requires minimal, relatively noncontroversial, assumptions about group structure, and it works well in many applications. However, there is a major problem in the use of this approach with population genomic data in that the density of informative sites along the genome (usually on the order of one per hundred bases) is mismatched with the typical lengths of many noncoding elements of interest (a few bases to a few dozen bases). As a result, HMM-based or sliding-window methods can only be effectively applied on the scale of large genomic regions rather than individual elements. Alternative grouping schemes are needed to characterize fitness consequences in noncoding elements.

In addition, the problem of predicting the probability of mutational fitness consequences in a given group of sites has the complication that the outcome of interest is not directly observed. Unlike in the example of automobile insurance rates, where accident rates can be directly measured for each group of drivers, we cannot simply read off the fitness consequences of mutations within each group of genomic positions. Instead, we must infer these properties from patterns of polymorphism and divergence in these regions, using a modeling approach that allows for features such as variation in mutation rates and coalescent times along the genome, and influences from population structure and population bottlenecks. Thus, the mapping from groups to probabilities of fitness consequences is complex, highly nonlinear, and computationally expensive.

Calculation of FitCons Scores

We have addressed these various challenges using the following strategy (see Fig. 1 for an overview). We begin with a collection of genome-wide functional genomic data sets, including DNase-seq, RNA-seq, and ChIP-seq data for various histone modifications. As discussed later, these data sets may represent a single or multiple cell types, but for now let us assume they describe a single cell type of interest. These data sets can be pictured as a series of "tracks" along the reference genome sequence, each of which may be either continuous-valued or discrete (Fig. 1A).

The first step in our procedure is to cluster genomic positions by their joint patterns of functional genomic data, that is by the patterns observed in vertical cross-sections of these horizontal tracks (Fig. 1B). In particular, we partition genomic positions into K mutually exclusive and exhaustive classes, each of which has a characteristic functional genomic "fingerprint." For example, one class might contain genomic positions with a high DNase-seq signal, a moderate RNA-seq signal, and high signals for nearby monomethylation of lysine 4 in histone 3 (H3K4me1) and acetylation of lysine 27 in histone 3 (H3K27ac), suggesting transcribed enhancers. Another class might contain positions with a low DNase-seq signal, a high RNA-seq signal, and a signal for trimethylation of lysine 36 in histone 3 (H3K36me3), suggesting actively transcribed gene bodies. The number of classes, K, is assumed to be moderately large (typically in the hundreds) to allow rich and potentially unexpected combinations of functional genomic patterns to be captured while still allowing for large numbers of genomic positions per class. Notably, we do not assume that this clustering scheme tends to place nearby positions in the same class; in general, each cluster will contain genomic positions that are highly dispersed along the genome sequence.

In the second step, we use INSIGHT to separately estimate the probabilities of mutational fitness consequences based on patterns of polymorphism and divergence within each of the K classes (Fig. 1C). Importantly, INSIGHT does not attempt to model absolute allele frequencies, which is difficult because of the confounding effect of demography, but instead considers rates of both fixed differences and low and high derived allele frequencies in the positions of interest relative to those in nearby nucleotide sites believed to be neutrally evolving. These putatively neutral sites are identified by removing from the set of all genomic sites any positions likely to be influenced by natural selection, such as protein-coding genes, RNA genes, known transcription factor binding sites, evolutionarily conserved elements, and flanking sites of known elements (see Methods). The same set of neutral sites is used for all K classes. In the end, this step yields a single estimated value ρ for each of the K classes, which is interpreted as the fitCons score for that class.

Finally, we simply map these fitCons scores back to the genome sequence, by assigning to each nucleotide position the score estimated for the functional genomic class to which it is assigned (Fig. 1D). As noted above, each genomic position is thus assigned a value between 0 and 1, which represents 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 as that position. These fitCons scores uniquely integrate information from evolutionary and cell-type-specific functional genomic data.

Clustering Scheme

In our initial approach, we have employed a simple exhaustive partitioning scheme for the clustering step. We began by reasoning that three highly informative functional genomic data types—DNase-seq, RNA-seq, and ChIP-seq data describing histone modifications—represent largely orthogonal sources of information, describing DNA accessibility, transcription, and chromatin states, respectively. We then partitioned the data into small numbers of classes along each of these three axes by a simple scheme, and considered all possible combinations (the Cartesian product) of these DNase-seq, RNA-seq, and ChIP-seq class assignments. For the DNase-seq data, we divided genomic positions into three levels of "signal," depending on whether they fell outside of DNase-seq peaks (0), in broad peaks only (1), or in narrow peaks (2) based on the peak calls from ENCODE (see Methods). Similarly, for the RNA-seq data, we divided genomic positions into four groups by separating positions with no aligned reads (0) from those with a low (1), medium (2) or high (3) mean read depth. The thresholds differentiating "low", "medium", and "high" were determined by a heuristic scheme designed to maximize the available information (Methods). For the chromatin modifications, we simply re-used the assignments to 26 distinct categories based on the ChromHMM method [40, 42]. Finally, we distinguished between sites that fell outside (0) and within (1) annotated protein-coding sequences (CDSs), which we expect to show fairly pronounced differences in selective pressure. By considering all possible combinations of these assignments, we obtain $3 \times 4 \times 26 \times 2 = 624$ distinct functional genomic signatures. Using this scheme, we separately partitioned the genome into these 624 categories for each of three different karyotypically normal cell types: human umbilical vein epithelial cells (HUVECs), H1 human embryonic stem cells (H1 hESC), and lymphoblastoid cells (GM12878). In practice, there was sufficient data to run INSIGHT for 443–447 classes (across the three cell types), with median numbers of genomic positions per class of 165 to 224 thousand (see Supplementary Table S1).

In principle, the steps of clustering genomic positions and estimating probabilities of fitness consequences could instead be addressed in combination, perhaps by a strategy that iteratively refined both the clusters and the fitCons scores (dashed line in Fig. 1). This approach might enable finer-grained partitioning of classes and improved power for functional elements, but it adds considerable complexity to the problem and is left for future work (see Discussion).

Genomic Distribution of FitCons Scores

To obtain a general overview of the genomic distribution of fitCons scores, we considered the *composition* and *coverage* of nucleotide sites of various annotation types as the fitCons score threshold S was varied. We partitioned nucleotide positions in the human genome into GENCODE-annotated [44] coding sequences (CDS), $5'$ and $3'$ untranslated regions (UTRs), proximal promoters (1000 bp upstream of annotated transcription start sites), introns, intergenic regions, small noncoding RNAs (sncRNAs), and long noncoding RNAs (lincRNAs), as well as transcription factor binding sites (TFBS) supported by ChIP-seq data from ENCODE [3, 37] (see Methods). In this section we will focus on results for HUVECs, but a comparison of cell types will be considered later in the paper.

When the score threshold $S = 0$, all nucleotide sites in the genome are included and the composition of annotations reflects the overall genomic distribution, with large fractions of intergenic and intronic sites, much smaller fractions of

CDS, $5'$ and $3'$ UTRs, and promoter regions, and negligible fractions devoted to sncRNAs, lincRNAs, and TFBSs (Fig. 2A). As S increases, however, the sites in known functional classes become strongly enriched relative to the intergenic and intronic sites. Regions such as $3'$ UTRs, and to a lesser extent, $5'$ UTRs, promoters, and introns, are most enriched at intermediate scores, reflecting moderate levels of natural selection in these regions. Coding sequences, by contrast, dominate at the highest scores. Note that these enrichments for high scores in functionally annotated genomic regions occur despite no use of genomic annotations in the scoring scheme (with the single exception of 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.

If we instead consider the coverage of various types of functional elements at various score thresholds S, we can more clearly observe the behavior of the fitCons scores for various classes of functional noncoding sites that account for only small fraction of genome, such as TFBSs, sncRNAs, and promoters (Fig. 2B). Not surprisingly, the coverage properties are best for CDSs, $3'$ UTRs, and $5'$ UTRs (in that order), but they are also considerably elevated above the intergenic background for promoters, TFBSs, lincRNAs, and sncRNAs. In this plot, the enrichment for functional noncoding annotations in the moderately high scoring range (between roughly 0.08 and 0.40) can again be seen clearly, while scores in CDSs almost all exceed 0.5. For example, a score threshold of $S = 0.08$ identifies 6.7% of intergenic regions, but 23–48% of regulatory and ncRNA sites, 67–75% of UTR sites, and 99.99% of CDS sites.

We have developed UCSC Genome Browser tracks that display fitCons scores across the genome for each cell type of interest (see http://genome-mirror.bscb.cornell.edu, hg19 assembly). As compared with existing evolutionary conservation scores, the fitCons track provides similar, but complementary, information about potential genomic function. FitCons scores are clearly elevated in CDS and UTR exons, like conservation scores, but they often do a better job of highlighting candidate regulatory elements, particularly ones outside the core promoter (see Fig. 3). Many of these regulatory elements may not be significantly conserved across species because they have arisen recently in evolutionary time, while others may have spuriously low conservation scores because of errors in orthology detection or alignment. Another important difference from conservation scores is that the fitCons scores also often differ significantly across cell types, as will be discussed later in the article.

An examination of the marginal dependency of the fitCons scores on the strengths of the DNase-seq and RNA-seq signals (here, focusing on non-CDS regions) indicates that the scores tend to increase with increasing signal, as might be expected (Fig. 4). The increase in fitCons score with DNase-seq signal (Fig. 4A) probably reflects an increasing density of *cis*-regulatory elements, whereas the increase with RNA-seq signal (Fig. 4B) most likely primarily reflects changes in the type of transcription unit identified (e.g., regions with low RNA-seq signal are enriched for various noncoding RNAs, while regions with high RNA-seq signal are enriched for more conserved $3'$ UTRs). Importantly, these monotonically increasing relationships are not assumed by the model but arise naturally from patterns of polymorphism and divergence associated with differential signals.

When joint patterns of DNase-seq and RNA-seq are considered, however, a more complex pattern emerges: fitCons scores increase with DNase-seq intensity at low RNA-seq intensities, as they do overall, but this trend is reversed at high RNA-seq intensities (Fig. 4C). A closer examination reveals that this pattern is explained by the enrichment for

DNase-I-hypersensitivity near the $5'$ ends of genes. In particular, conditional on a high RNA-seq signal, a high DNase-seq signal tends to be associated with $5'$ UTRs and upstream regions of genes, which are under somewhat weaker selection than the $3'$ UTRs associated with lower DNase-seq signals. This example demonstrates that the fitCons scores can capture non-additive and potentially surprising relationships between functional genomic covariates and natural selection.

Predictive Power for *Cis*-Regulatory Loci

While the fitCons scores appear to behave in a reasonable manner in relation to known genomic annotations, it is unclear from the previous analysis how well they will perform as predictors of regulatory function. Therefore, we evaluated the predictive power of these scores for known cell-type-specific regulatory elements in comparison with two widely used phylogenetic conservation scoring methods, the phyloP [15] and Genomic Evolutionary Rate Profiling (GERP) [13] programs. We also considered a new program, called Combined Annotation Dependent Depletion (CADD) [45], that measures divergence on the branch from the human/chimpanzee most recent common ancestor to present-day humans, using a combination of a support vector machine (SVM) and simulations of nucleotide divergence rates. We evaluated the performance of all four scoring methods in predicting three types of functional elements: (1) binding sites for various transcription factors supported by chromatin immunoprecipitation and sequencing (ChIP-seq) data from the ENCODE project [3, 37]; (2) high-resolution expression quantitative trait loci (eQTL) identified in a recent large-scale study [6]; and (3) enhancers based on chromatin marks [46] (see Methods for details). These are among the best available large-scale surrogates for direct evidence of functional roles in transcriptional regulation, and, importantly, annotation sets (1) and (2) are based on quite different functional genomic data from that used by fitCons, although there is some overlap between set (3) and the ChromHMM data.

To place the different scoring methods on equal footing, we first plotted the coverage of each class of regulatory element at the nucleotide level as a function of the total fraction of the noncoding genome that exceeded each score threshold, varying the thresholds to include 0–20% of noncoding positions in the genome (Fig. 5). This strategy allows us to measure the extent to which the elements of interest display "signals" that rise above the "noise" 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 all of the other methods considered. For example, at a total noncoding coverage of 2.5%, fitCons scores achieve nearly 70% coverage of TFBSs, whereas the other methods all have less than 15% coverage. Similarly, the coverage of enhancers was about 40% at 2.5% noncoding coverage, while the other scoring methods showed almost no signal above background. The excellent performance of the fitCons scores on these tests is enabled by the use of markers for regulatory function, such as DNase-I hypersensitivity and promoter- or enhancer-associated chromatin modifications, in pooling data across genomic positions.

We also performed a more formal evaluation of the trade-off between sensitivity and specificity for the same scoring methods and sets of noncoding elements using receiver operating characteristic (ROC) curves. These tests are related to the ones above, but plot true positive rates (here, equal to coverage) against a direct measure of false positive rates based on the fraction of nucleotides in a matched null set of sequences that had scores exceeding each threshold

S (see Methods). We summarized the overall performance of each scoring method using the area under the ROC curve (AUC), which ranges from 0.5 for random guessing on a matched set of positive and negative examples to 1.0 for a perfect classifier. In the case of eQTLs, we also considered RegulomeDB, a database that scores nucleotide variants for regulatory potential based on a combination of experimental data, manual annotations, and computational predictions [47]. We found that, in all cases, fitCons scores were considerably better predictors of regulatory function than phyloP, GERP, CADD, or RegulomeDB scores (Supplementary Fig. S1), with AUCs of 0.92, 0.75, and 0.88 on TFBSs, eQTLs, and enhancers, respectively, compared with second-best AUCs of 0.68, 0.60, and 0.61. Notably, the divergence-based scores (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. The CADD method generally performs more poorly on these noncoding elements than previously reported [45], possibly because it is implicitly optimized for coding regions (see Discussion). These results demonstrate that, by making use of cell-type-specific functional genomic data and considering recent evolutionary time periods, fitCons scores exhibit substantially improved sensitivity for *cis*-regulatory elements.

The tests above were based on regulatory elements that are putatively active in the cell type for which the scores were produced, to exploit fitCons's use of cell-type-specific functional data. To evaluate how well these advantages carried across cell types, we created an "integrated" fitCons score by combining information from three cell types (see 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. S2).

Proportion of the Human Genome Under Selection

In addition to being useful as predictors of function, the fitCons scores could also shed light on the important question of what fraction of nucleotides in the human genome have direct fitness-influencing functions. Estimation of this "share under selection" (SUS) has primarily been addressed using methods that consider longer evolutionary time scales, during which turnover of functional elements may be an important confounding factor [23, 48–51]. Our methods should be less sensitive to turnover because they measure the influence of natural selection on 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, their average can be interpreted as the expected fraction of nucleotides in the genome that have fitness-influencing functions—the same quantity estimated by conservation-based methods. This approach yields an estimate of 7.5% ($\pm$0.1%) for HUVEC. The estimates ranged from 7.5–7.8% for the three analyzed cell types, and changed little when we used the cell-type integrated scores instead. Among sites under selection, we estimate that 9% are in CDS, 2.2% in $3'$ UTRs, 0.9% in promoters (with 1kb of the TSS), 35% in introns, 52% in intergenic regions, and $<$0.5% in each of several other noncoding annotation classes. These estimates are generally consistent with, but on the high end, of those based on divergence patterns between species, which have indicated that 3–8% of the genome is under selection,

with noncoding bases outnumbering coding bases by factors between about 2.5 and 8 [12, 23, 48, 52, 53]. Interestingly, our estimates are somewhat lower than estimates that have attempted to allow for evolutionary turnover, which have been 2–3 times higher than the pan-mammalian estimates of $\sim$5% [23, 35, 53, 54] (see Discussion).

One possible problem with the INSIGHT-based estimates is that they will tend to be biased upward for classes in which the true fraction is close to zero, because violations of modeling assumptions will tend to push these estimates to small positive values (estimates of less than zero are not permitted; see Supplementary Material). To address this problem, we performed the same calculation for designated "neutral" sites that intersect with the large class of genomic positions having a "null" functional genomic fingerprint (i.e., no DNase-seq, RNA-seq, or histone modification signal), in order to estimate the contribution of model misspecification to our estimate of the SUS. This estimate, which came out 3.3%, can be considered an upper bound on the error contribution because these putatively neutral sites undoubtedly include some sites under selection. We then subtracted this 3.3% from our naive estimate of 7.5% to obtain 4.2% as an estimated lower bound for the SUS. (These estimates are for HUVEC, but the results for the other cell types were very similar.) This correction disproportionally reduces the fractions of selected sites in the noncoding classes, causing the CDS fraction to increase to 15.3%, the fraction in $3'$ UTRs to increase to 3.3%, and the fractions in all other annotation classes to remain about the same or decrease slightly (Supplementary Table S2). Overall, this analysis 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, estimates that are remarkably consistent with those based on measures of divergence between mammalian genomes (see Discussion).

Implications for Evolutionary Turnover of Functional Elements

To better understand the differences between our new scores and conventional phylogenetic conservation scores, we compared fitCons scores with mean mammalian phyloP scores [15] for each of our 624 clusters. Given the substantial differences in modeling assumptions, methods for inference, and data underlying these two sets of scores, they are remarkably well correlated overall ($R^2 = 0.87$; Supplementary Fig. S3). This general correlation of scores based on human population genetic data and divergence patterns across placental mammals suggests that the signal for natural selection is driven primarily by purifying selection maintained over tens of millions of years.

To make this comparison more direct, we devised a set of alternative fitCons scores (denoted "fitConsD") based on the same clusters of sites but an alternative estimator of the fraction of nucleotides under natural selection that considers nucleotide divergence across humans and nonhuman primates rather than patterns of human polymorphism and human-specific divergence (see Methods). Thus, the fitCons and fitConsD scores indicate probabilities of fitness consequences per nucleotide defined in analogous ways but over two different evolutionary time scales. Before comparing these two sets of scores, we corrected both of them for model misspecification by subtracting the estimates obtained for neutral sites in the null class, as described in the previous section.

After these adjustments, we observe that both measures of selection are well correlated, with $R^2 = 0.89$ (Fig. 6A). As in the phyloP comparison, we observe clear differences between the coding and noncoding sites, with the scores for coding sequences being generally higher, and those for noncoding sequences generally lower, under both scoring

methods. Interestingly, we also observe an excess of variance in the noncoding sites ($R^2 = 0.51$ vs. $R^2 = 0.68$ for coding), with a slight excess of clusters having higher fitCons than fitConsD scores. These observations further suggest that the bulk of the signal for selection has been maintained over long evolutionary time periods, but that there are some classes that show stronger recent than ancient natural selection.

We can more directly assess the impact of turnover by subtracting the fitConsD score from the fitCons score for various annotation classes of interest. This difference between the (adjusted) scores can be interpreted as the "net gain" in functional nucleotides on population genetic time scales relative to primate-divergence time scales, with a negative value representing a net loss (see Fig. 6B). Furthermore, this net gain can be divided by the fitCons score for each annotation class of interest to obtain a net gain rate per functional nucleotide (diagonal lines in Fig. 6B). We observe relatively low amounts of turnover overall, with net gains for our standard set of annotation classes ranging between -0.7% and 1.7% and net gain rates roughly between -10% and 10%. These observations suggest that the net gain or loss of functional nucleotides in both coding and noncoding functional elements accounts for no more than about 10% of all functional sites, although these estimates do not exclude the possibility that offsetting gains and losses produce considerable evolutionary flux in the functional composition of the genome (see Discussion).

We did observe a few clusters with net gain rates as high as 20%. These clusters were associated with an absence of DNase-seq or RNA-seq signal and with chromatin modifications suggesting transcriptional elongation, which were strongly enriched in the introns of expressed genes (see Fig. 6B), suggesting the possibility of evolutionary innovation associated with mRNA splicing or post-transcriptional regulation.

Differences between Cell Types

We have focused so far on results for HUVECs, but, as noted above, we also generated fitCons scores for H1 hESC and GM12878. To compare the scores for different cell types, we began by examining the 624 functional genomic classes across the three cell types, in terms of both the genomic positions assigned to each class, and the fitCons scores estimated for those positions. (Note that our partitioning scheme ensures that the the same 624 class definitions are used for each cell type.) Not surprisingly, we found that the genomic positions assigned to each class differed substantially across cell types, with fewer than 10% of positions being assigned to the same functional class across all three cell types, and roughly half or more being assigned to different functional classes in all three cell types. Despite their association with different genomic positions, however, equivalently defined clusters exhibited highly similar fitCons scores across cell types (Pearson correlation > 0.94 for all pairs; Supplementary Fig. S4). Thus, while the patterns of activity differ substantially across cell types, the evolutionary signatures associated with genomic positions that display particular patterns of activity are remarkably consistent across cell types.

To examine the degree to which the scores convey cell-type-specific information, we next examined fitCons scores for elements that are active in one cell type an inactive in another. In particular, we examined subsets of eQTL and proximal promoters (within 100bp of the annotated transcription start site) that appear to be active in H1 hESC but inactive in HUVEC (H1 HESC+/HUVEC−) or inactive in H1 hESC and active in HUVEC (H1 HESC−/HUVEC+) based on RNA-seq data for the same cell types (see Methods). For each of these groups of elements, we compared

mean fitCons scores computed for each of the two cell types (H1 hESC and HUVEC). We found that, based on the scores computed for each cell type, the active elements in that cell type had significantly higher scores than the inactive elements (compare the two gold bars and the two purple bars in each panel in Fig. 7). In addition, the same sets of functional elements have significantly higher fitCons scores for the cell type in which they are active than for the one in which they are inactive (compare adjacent gold and purple bars in Fig. 7). Similar patterns were observed for comparisons involving GM12878 (results not shown). These findings demonstrate that, while the fitCons scores for all cell types are based on the same polymorphism and divergence data, they nevertheless convey cell-type-specific information through the use of cell-type-specific functional data for clustering.

Despite the advantages of the cell-type-specific scores, it is sometimes desirable to have single set of scores that integrate information from multiple cell types. We created an integrated set of scores by assigning each genomic position to the functional class for which it received the highest score across the three analyzed cell types, and then recomputing the fitCons scores based on these new genome-wide assignments (see Methods for details). This simple approach produces scores that have nearly equal predictive power for elements active in all three cell types, and much better power than the cell-type-specific scores have when they are applied to a mismatched cell type. The integrated scores are publicly available along with the cell-type-specific scores in our genome browser tracks (http://genome-mirror.bscb.cornell.edu, hg19 assembly). We recommend the use of these scores when scores for a matching or similar cell type to the one of interest are not available.

Discussion

We have presented a new approach that combines functional genomic data with data describing variation within and between species. The essential idea of the approach is to use the 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 those classes based on 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.

Based on our experiments, the 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. In part, this improvement in performance reflects the use of cell-type-specific data and, indeed, it is most pronounced when considering elements active in the cell-type used to compute the scores (Fig. 5 and Supplementary Fig. S1). Nevertheless, the fitCons scores also show a clear performance advantage when considering all annotated elements rather than just active ones (Supplementary Fig. S2). The approach of grouping genomic sites by functional genomic signatures, and then measuring groupwise fitness consequences based on patterns of genetic variation, appears to offer real benefits for prediction of regulatory function, as compared with

methods that consider either genetic divergence or functional genomic data alone. In addition, in at least some cases, the use of population genomic data helps to identify elements that appear to have emerged recently in evolutionary time and therefore display no significant reductions in interspecies divergence.

Interestingly, the recently published CADD method performed no better on these tests than conventional conservation scores, despite reports by the authors of significant advantages over GERP, phastCons, phyloP, and other methods [45]. It is likely that this difference in CADD's performance reflects a focus on *cis*-regulatory elements in our study versus a focus on coding regions in Kircher et al.'s validation experiments. Because the SVM-based training procedure use by CADD is executed globally for the entire genome and is based on differences between observed variants and those that appear in neutral simulations, which will be especially pronounced in coding regions, and because many of the hundreds of features it considers are relevant in coding regions only, it seems likely that the method is implicitly optimized for coding sequences at the expense of noncoding regulatory sequences. It may be interesting to experiment with an alternative version of CADD that is trained and tested on noncoding variants only, and perhaps makes use of cell-type-specific functional genomic covariates.

It is worth noting that the fitCons framework, by being based on a full probabilistic model of evolution, has some important advantages over generalized methods for regression and classification such as SVMs, neural networks, or generalized linear models. While they are powerful predictors in the presence of high-quality training data, methods of this kind are often "black boxes" in the sense that they provide good approximations of prediction functions without yielding much insight into the precise nature of the relationships among independent and dependent variables. Our approach, by contrast, provides interpretable information about the selective effects of mutations in each of the classes of sites considered. The probabilistic approach also allows missing data to be handled simply and naturally, it can provide information about features such as adaptive substitutions and genetic load, and it supports likelihood-based model comparisons (see [37, 43]).

A side benefit of our model-based approach is that the basewise probabilities of fitness consequences lead in a straightforward manner to an estimate of the "share under selection" (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, and therefore differs substantially from 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 mostly fallen between 3 and 8%. For example, the first estimate of this kind based on whole-genome sequencing data indicated that about 5% of short (50 bp) windows were under selection (4.7–7.9% for windows ranging from 30-200 bp) [48, 52]. Subsequent analyses of constrained elements in three or more mammalian genomes yielded estimates of the SUS of 3–8% [12, 49, 50], while an estimate based on insertions and deletions in human, mouse, and dog alignments was slightly lower, at 2.6–3.3% [55]. More recently, an analysis based on deep alignments of 29 mammals predicted that 5.4% of 12-bp windows were under selection [51]. We take the general concordance of these estimates, both with one another and with our fitCons- and fitConsD-based estimates, as an indication that the SUS has been fairly stable at roughly 5% over various time-scales in mammalian evolution.

Following the publications by the ENCODE Consortium in 2012, there has been a great deal of discussion in the

scientific literature, the scientific press, and social media about the discordance between evolution-based estimates of the SUS and estimates of the functional content of the genome based on high-throughput measures of biochemical activity, which were reported to be as high as 80% [3, 56]. For various reasons, the ENCODE-based claims do appear to require an excessively generous definition of "function" (see [16–19]). Nevertheless, it is worth emphasizing that the question of the functional content of the genome is inevitably dependent on how function is defined. Consider two possible definitions of "functional" DNA: sequences that (1) when mutated (by point mutations), or (2) when deleted, produce a phenotype. Under the first definition, genomic positions such as fourfold degenerate sites in coding regions or degenerate positions in TFBSs will generally not be functional, whereas under the second definition they will be functional, because their presence is required to maintain the functional coherence of a larger element (they are both examples of "spacer" elements). Other examples of functional sequences whose function does not depend on the precise identity of each nucleotide at each position include sequences separating binding sites for interacting TFs, sequences in short introns, and sequences that maintain the spacing properties of *cis*-regulatory elements relative to target genes. Importantly, most estimates of the SUS, including ours, have made use of definition (1), whereas measures of biochemical activity are more consistent with definition (2) in some respects (although not all spacer elements will be biochemically active). It is hard to imagine that this distinction could account for the difference between estimated genomic fractions of $\sim$80% and $\sim$5%, but it is nevertheless worth bearing in mind that our estimate of the SUS and those from comparative genomics are based on a fairly restrictive definition of function. Indeed, our methods indicate that the SUS in annotated coding regions is only about 60%, a fraction that would undoubtedly rise under definition (2).

Apart from the question of 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 attempted to address this "turnover" question in various ways, and have concluded that the current SUS in the human genome could be $\sim$2–3 times larger than estimates based on comparisons across mammals [23, 35, 53, 54]. Indeed, this observation has been offered as a possible explanation for the discordance between evolution-based and biochemical estimates of the functional fraction of the genome [35, 56] (see also [57]). However, these studies 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 [53, 54] or by estimating functional content from mean SNP densities, heterozygosity levels, or frequencies of derived alleles in genomic regions not conserved across mammals [35] (see [58]). Our comparison of the fitCons and fitConsD scores is more direct, by making use of analogous divergence-based and polymophism-based estimators for the SUS and applying them to exactly the same clusters of nucleotide positions. In addition, our analysis focuses on the more restricted question of evidence for selection on the time scales of primate divergence and human-specific polymorphism and divergence, rather than attempting to account for turnover across the mammalian phylogeny, where factors such as alignment error, orthology detection, and genomic rearrangements can be problematic. The high degree of concordance between our estimates based on polymorphism (fitCons) and divergence (fitConsD) strongly suggests that evolutionary turnover has played at most a modest role in genome evolution, at least on the time scale of primate evolution, because if massive turnover had

occurred during this time period, the divergence-based estimates would be expected to show a substantial downward bias. Furthermore, our power experiments suggest that this observation is not an artifact of reduced sensitivity in the fitCons scores. Nevertheless, we cannot rule out the possibility of abundant compensating gains and losses on very recent time scales, resulting in a similar SUS but different genomic composition of functional sequences.

An obvious area for improvement for our methods is in the handling of multiple cell types. Our current approach is simply to apply the method separately to each cell type for which functional genomic data is available, and then produce a separate, integrated set of scores in a heuristic post-processing step (see Methods). This approach does reveal differences between cell types, but it does not make use of the full joint information in the functional genomic data for multiple cell types. In reality, the observed patterns of polymorphism and divergence reflect some average across all of the natural cell types relevant to fitness. With improved clustering techniques, it may be possible to consider all cell types simultaneously, cluster sites by multi-cell-type functional genomic fingerprints, and then produce a single set of scores reflecting these joint patterns. In this way, nucleotide sites that not only exhibit similar functional genomic fingerprints, but that are active in similar cell types, would be grouped together, potentially resulting in improved predictive power.

Even aside from the issue of integrating data across cell types, it will be useful to explore more sophisticated approaches to clustering. Our current strategy of considering the Cartesian product of discrete covariate categories will not scale to large numbers of covariates. In addition, in some cases it results in overly coarse-grained clusters, whereas in others it fails to group together small clusters with similar fitness consequences. An iterative approach to clustering, as shown in Fig. 1, could enable the estimated fitCons scores to help guide the clustering scheme. This approach will be complex and computationally intensive because of the expense of estimating model parameters from genome-wide data and because of the need for some type of regularization method to constrain the coherence of functional classes while the variance in the classwise fitCons scores is maximized, but it could support the use of much larger, richer sets of covariates. Larger numbers of covariates, potentially from multiple cell types, together with steady improvements in the resolution and quality of the available functional genomic data, should result in improved power for individual functional elements and refined estimates of the share under selection.

Methods

Functional Genomic Data

RNA-seq and DNase-seq data for HUVEC, H1 hESC, and GM12878 were downloaded from the University of California, Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu). Chromatin states for the same three cell types were downloaded from the European Bioinformatics Institute ftp site (see Supplementary Table S3). For chromatin states, we used the 25-state ChromHMM segmentation generated in December, 2012 [42]. For RNA-seq, we selected Caltech poly-A-selected 75 bp paired end read data after examining several data sets and considering trade-offs among data quality, coverage and read depth. A single replicate experiment was used per cell type. 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 hESC so additional DNase-seq data for this cell line was obtained from Duke University. For each replicate experiment, we downloaded broad and narrow peak calls. Coding sequences (CDS) were identified uniformly across cell types using GENCODE data (see below).

Genome Partitioning

We produced a separate partitioning for each cell type based on the functional genomic data. Chromatin states were defined directly from the 25 states in ChromHMM. A 26th state was assigned to sites not assigned a chromatin class. The broad and narrow DNase-seq peaks were used to partition sites in the genome into the following three mutually exclusive classes: sites that fall in a narrow peak in both replicate experiments (2); sites that fall in a broad peak in at least one of the two replicates and do not fall in a narrow peak in both replicates (1); and sites that fall outside of all called peak (0). This three-level scheme was designed to allow for both high sensitivity (class 1) and high specificity (class 2). For H1 hESC, 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) based on numbers 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 (see **Supplementary Methods**). The thresholds chosen were (1, 7, 36) for HUVEC, (1, 5, 20) for H1 hESC and (1, 11, 38) for GM12878. The Cartesian product of these partitions, together with the partition into coding and noncoding sequences (see below), resulted in $26 \times 3 \times 4 \times 2 = 624$ distinct combinations that defined the functional categories, or clusters, for the computation of the fitCons scores.

Running INSIGHT

INSIGHT was used to compute the fitCons score for each non-empty site cluster (see Table S1). The INSIGHT method infers the fraction of nucleotide sites under selection (ρ) for a given collection of sites by comparing patterns of within-species polymorphism and between-species divergence for that collection with the patterns observed in nearby putative neutral sites. A detailed description of the method and its application to human data is given in [37, 43]. Briefly, for human polymorphism, we used high-coverage genome sequences for 54 unrelated human individuals from the 69 sequences released by Complete Genomics (http://www.completegenomics.com/public-data/69-Genomes/) [59], and for outgroup genomes, we used the chimpanzee (panTro2), orangutan (ponAbe2), and rhesus Macaque (rheMac2) reference genomes. Various filters were applied to eliminate repetitive sequences, recent duplications, CpG sites, and regions not showing conserved synteny with outgroup genomes. Putatively neutral sites were identified by excluding exons of known protein-coding and RNA genes plus 1 kb of flanking sites on each side, and previously predicted conserved noncoding elements plus flanking regions of 100 bp. Matching of the analyzed sites with putatively neutral sites was done using a 10 kb sliding window, taking care to avoid matching sites on opposite sides of a recombination hotspot. All data and filters were as described by Gronau et al. [43] except that the small collection of sites in identified TFBSs was not removed from the neutral sites in this case.

For each site cluster, INSIGHT was used to compute a maximum likelihood estimate of ρ, as well as a standard error approximated using the curvature method (see [43]). The estimate of ρ was used as the fitCons score associated with all sites in that cluster. To ensure our results were not influenced by estimates with high uncertainty, we filtered out clusters for which the estimated standard error was larger than 40% of the estimated value of ρ. To increase computation efficiency, clusters larger than 20 Mb were partitioned into smaller subclasses, and INSIGHT was executed separately on each of these subclasses. An estimate of ρ for the entire cluster was obtained by taking the average of subclass estimates of ρ weighted according to the number of unfiltered sites in each sub cluster.

GENCODE Annotations

Transcript annotations generated by GENCODE v15 [44] were downloaded from the Wellcome Trust Sanger Institute ftp site (ftp://ftp.sanger.ac.uk/pub/gencode/release_15/gencode.v15.annotation.gtf.gz), and used to define eight site classes: coding sequences (CDS), 5′ untranslated regions (UTRs), 3′ UTRs, promoters, introns, long intergenic noncoding RNAs (lincRNA), short noncoding RNAs (sncRNA), 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 the fitCons partitioning scheme. A slightly more conservative set was used in our subsequent analyses; this one was 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 of protein coding genes were identified by transcripts with feature_ type="UTR" and gene_ type="protein_coding". Each UTR was designated as 5′ or 3′ based on whether it was immediately upstream of the start codon or immediately downstream of the stop codon, respectively. Introns were defined by positions that fall within a transcript for a protein-coding gene but outside of the CDS and UTR transcripts of that gene. Promoters were defined as the 1000 bp immediately upstream of the first (i.e., most upstream) transcription start site for each protein-coding gene. In addition, a similarly defined alternative set of 100 bp promoter regions was used in assessing differences between cell types (see Fig. 7). LincRNAs were identified using transcripts with feature_type="exon" and gene_type="lincRNA". Similarly, sncRNA consisted of transcripts with feature_type="exon" and gene_type $\in$ {"miRNA", "snRNA", "snoRNA"}. Positions in the more inclusive CDS set were removed from all noncoding classes to avoid overlap with coding sequences. When computing the composition of sites exceeding various fitCons score thresholds by annotation type (Fig. 2A), if multiple annotations applied to a nucleotide position, it was assigned to a single category in the following order of priority: CDS, TFBS, promoter, sncRNA, lincRNA, 5′ UTR, 3′ UTR, intron, and intergenic.

Cis-Regulatory Elements

Transcription factor binding sites (TFBSs) were drawn from a set prepared by Arbiza et al. [37] for 78 transcription factors, based on chromatin immunoprecipitation and sequencing (ChIP-seq) data from ENCODE (downloadable at (http://genome-mirror.bscb.cornell.edu/). This set contains roughly 1.4 million binding sites of mean length 11 bp, each of which is associated with the cell types in which it was detected. For enhancers, we used the distal regulatory

modules predicted by Gerstein et al. [46]. We downloaded the file enets4.Distal_cell_line.txt from http://encodenets.gersteinlab.org/ 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 from which each enhancer was derived. Expression quantitative trait loci (eQTL) identified by Lappalainen et al. [6] were downloaded from www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/analysis_results/. We used the four files {EUR373,YRI89}.{exon,gene}.cis.FDR5.best.rs137.txt.gz to identify 6,760 distinct autosomal positions and the transcripts associated with each. As with other noncoding classes, we removed CDS positions from all regulatory elements.

Identifying Active Elements per Cell Type

For some of our analyses we considered the subset of elements in each annotation class for which we had evidence of activity in a given cell type. In the cases of TFBS and enhancers, we simply used the cell type designations provided in their annotation files (see above). For other classes of elements, e.g., eQTLs and promoters, we defined the active subset using a set of GENCODE transcripts and genes that showed significantly high levels of RNA transcription in the Caltech RNA-seq data. For this purpose, we downloaded from the UCSC Genome Browser files containing normalized read counts in reads per kilobase per million at the levels of both transcripts and genes for HUVEC, H1 hESC, and GM12878 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/). A transcript (or a gene) was defined as being active in a certain cell type if the 95% confidence interval if its normalized read count in that cell type fell within the top one third of normalized read counts for transcripts (or genes). The threshold determining the top one third (1.477 for transcripts and 4.966 for genes) was computed by aggregating information from all three cell types. We then determined the set of active eQTLs in each cell type as the ones associated with an active gene, using the GENCODE gene identifier specified for each eQTL in the data file. Similarly, we defined elements in our collections of promoters, UTRs, CDS, and introns, as active if they were associated with an active transcript. For the comparison between cell types (Fig. 7) we also used collections of eQTLs and promoters found to be inactive in a given cell type. Those 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 Genomic Evolutionary Rate Profiling (GERP) [13] method were downloaded from http://mendel.stanford.edu/SidowLab/downloads/gerp/ (file hg19.GERP_scores.tar.gz, generated in August, 2010). Scores from phastCons [12] and phyloP [15] for 46 placental mammals were downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/; subdirectories phastCons46way/placentalMammals/ and phyloP-46way/placentalMammals/). The Combined Annotation Dependent Depletion (CADD) scores [45] were downloaded from http://cadd.gs.washington.edu/download/ (file whole_genome_SNVs.tsv.gz, downloaded in September, 2013). This file specifies for each genomic position a separate score for each of the three possible variant bases in that site. We took the maximum of these three scores, which yielded the best performance for CADD in our comparisons.

We also considered RegulomeDB [47] as a method for ranking single nucleotide polymorphisms (SNPs), such as eQTLs, according to evidence from functional genomic data. RegulomeDB classifies each common SNP into one of 13 categories (1a–f, 2a–b, 3–7) ranked from strongest (1a) to weakest (7) evidence for function. We downloaded the thirteen categories from http://regulome.stanford.edu/downloads/ in January, 2013.

Receiver Operating Characteristic (ROC) Curves

We used ROC curves to describe the ability of each scoring schemes to discriminate between functional and nonfunctional TFBSs, enhancers, and eQTLs. For TFBSs and enhancers, we used annotations to indicate the set of functional elements (see above) and used as a non-functional control the filtered, putatively neutral sites identified for INSIGHT (see above). For eQTLs, the set of common SNPs in dbSNP v137 was used as the control set. In all three cases, we removed any sites in our functional set from the negative control set. For each scoring scheme and annotation type, a point on a ROC plot indicates the fraction of 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). In computing the fractional coverage for each scoring scheme, we ignore positions that are not scored, so as not to penalize methods such as RegulomeDB that provide partial coverage. Note that the other four scoring schemes had similar overall coverage.

Integrating fitCons Scores Across Cell Types

The main challenge in generating fitCons scores that integrate functional genomic data across cell types, within the context of our simple partitioning scheme, is to avoid a combinatorial explosion in the number of functional genomic "fingerprints" and corresponding site clusters. We addressed this problem by fixing the partitioning scheme to the original $26\times3\times4\times2$=624 fingerprints (see above), but altering the rule by which nucleotide sites are assigned to clusters to reflect information from multiple cell types. Our goal was to select for each nucleotide site the cluster assignment, from its three cell-type-specific assignments, that was most informative about its function. To achieve this goal, we first computed a cell-type aggregated estimate of ρ for each of the 624 possible fingerprints by running INSIGHT on the collection of *all* sites associated with that fingerprint in any of the three cell types. Note that, unlike in the standard fitCons pipeline (see Fig. 1), these collections of sites overlap with one another. In a second step, we then partitioned the sites into non-overlapping clusters by choosing, for each genomic position, the fingerprint (out of the three) that had the highest cell-type aggregated ρ. Finally, INSIGHT was executed once more on each of these disjoint clusters to obtain cell-type integrated fitCons scores.

Share Under Selection

Assuming a given partitioning of the genome into K mutually exclusive and exhaustive clusters, $C_1, \ldots, C_K$, and fitCons scores computed for these K clusters, $\rho(C_1), \ldots, \rho(C_K)$, the number of genomic positions under selection in cluster C_i is estimated by $\rho(C_i)|C_i|$ (since ρ is an estimate of the fraction of sites under selection under the INSIGHT

model). Thus, for an arbitrary collection of sites, S, the number of sites under selection is estimated using the formula $\text{sel}(S) = \sum_i \rho(C_i)|C_i \cap S|$, and the average fitCons score for S is given simply by $\rho(S) = \text{sel}(S)/|S|$. To avoid under-estimation of $\rho(S)$, we do not filter out fitCons scores with high uncertainty in these calculations, as we do for other analyses (see above). To account for possible over-estimation of ρ in very large clusters with low fractions of sites under selection, we ran INSIGHT on the large collection of sites (790 Mb) obtained by intersecting the collection of putative neutral sites used to fit the neutral model in INSIGHT (see above) with the cluster, C_{null}, consisting of noncoding sites in the 'Quiescent' chromatin state with no DNase-seq and RNA-seq signal (0 levels defined above). The estimated value of ρ, denoted ρ_{neut}, was then subtracted from each fitCons score to obtain a conservative lower bound, $\rho(S) - \rho_{\text{neut}}$, for the fraction of sites under selection in each collection of interest.

FitConsD and Evolutionary Turnover

Our measures of turnover are based on comparisons between fitCons scores, which estimate the fraction of sites under selection since divergence from chimpanzee, with scores obtained using a parallel method, fitConsD, which measures natural selection in longer evolutionary timescales, namely, since divergence of human, chimpanzee, orangutan, and rhesus macaque. To make this comparison as direct as possible, fitConsD scores were computed using the same pipeline we developed for fitCons (see Fig. 1), except that in step C we replaced the INSIGHT model with an evolutionary model that considers sequence divergence between the four primate genomes.

Briefly, we obtained these divergence-based estimates as follows. We downloaded the multiple genome alignment for 46 placental mammals from the UCSC Genome Browser (http://genome.ucsc.edu), and extracted from it the subalignment for the four primates. In each of the three non-human genomes, we filtered out nonsyntenic regions and positions with genotype quality below 20. Additionally, we masked out sites filtered in the INSIGHT analysis to eliminate repetitive sequences, recent duplications, and CpG sites (see above). We assumed a fixed branch-weighted phylogeny T for the four-species tree, which was obtained by fitting a phylogenetic substitution model to fourfold degenerate sites in coding sequences, and we used the partitioning of the genome into 624 clusters $\{C_i\}$ defined using functional data from the HUVEC cell line (see above).

With these preparations, we estimated a divergence-based fraction of sites under selection, $\rho_{\text{div}}(C_i)$ for each cluster C_i, as follows. First, we created a pseudo-alignment consisting of columns from the original four-species alignment corresponding to positions in C_i. We then used the *phyloFit* procedure from RPHAST [60] to estimate a maximum-likelihood scaling factor s_i for the tree T for this pseudoalignment. This scaling factor s_i is an estimate of the relative evolutionary rate in cluster C_i, compared with the pre-estimated neutral model, but it does not yet consider variation in the neutral substitution rate along the genome. Therefore, we additionally computed similar scale factors for 10 kb blocks of neutral sites across the genome, using the same neutral sites and windowing scheme as for the fitCons scores (see above). Specifically, for each 10 kb window w, we computed a maximum-likelihood neutral scaling factor s_w^{neut} for T. We then defined the neutral scale s_i^{neut} for a cluster C_i as the weighted average of neutral scales $\{s_w^{\text{neut}}\}$ in the associated neutral blocks (i.e., the average is weighted by the size of the intersection of cluster C_i and each window w). Now the relative rate of substitution in C_i compared to the expectation under neutrality could be computed as,

s_i/s_i^{neut}. Under the assumption that negative selection dominates (see, e.g., [61]), an estimate of the fraction of sites under selection in cluster C_i is therefore given by $\rho_{\text{div}}(C_i) = 1 - s_i/s_i^{\text{neut}}$.

References

1. Mardis ER (2011) A decade's perspective on DNA sequencing technology. Nature 470: 198–203.
2. Wold B, Myers RM (2008) Sequence census methods for functional genomics. Nat Methods 5: 19–21.
3. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.
4. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, et al. (2012) A map of the cis-regulatory sequences in the mouse genome. Nature 488: 116–120.
5. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, et al. (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489: 83–90.
6. Lappalainen T, Sammeth M, Friedlander MR, 't Hoen PA, Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506–511.
7. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12: 628–640.
8. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, et al. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16: 1046-1047.
9. Margulies EH, Blanchette M, NISC Comparative Sequencing Program, Haussler D, Green ED (2003) Identification and characterization of multi-species conserved sequences. Genome Res 13: 2507-2518.
10. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, et al. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299: 1391-1394.
11. Ovcharenko I, Boffelli D, Loots GG (2004) eShadow: a tool for comparing closely related sequences. Genome Res 14: 1191-1198.
12. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050.
13. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901–913.
14. Asthana S, Roytberg M, Stamatoyannopoulos JA, Sunyaev S (2007) Analysis of sequence conservation at nucleotide resolution. PLoS Comput Biol 3: e254.
15. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20: 110–121.
16. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, et al. (2013) 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.
17. Niu DK, Jiang L (2013) Can ENCODE tell us how much junk DNA we carry in our genome? Biochem Biophys Res Commun 430: 1340–1343.
18. Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci USA 110: 5294–5300.
19. Eddy SR (2013) The ENCODE project: missteps overshadowing a success. Curr Biol 23: R259–261.
20. Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, et al. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. Cell 120: 21–24.
21. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499–502.

22. Siepel A, Diekhans M, Brejova B, Langton L, Stevens M, et al. (2007) Targeted discovery of novel human exons by comparative genomics. Genome Res 17: 1763–1773.

23. Ponting CP, Nellaker C, Meader S (2011) Rapid turnover of functional sequence in human and other genomes. Annu Rev Genomics Hum Genet 12: 275–299.

24. Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol 19: 1114–1121.

25. Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, et al. (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol 2: e130.

26. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, et al. (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328: 1036–1040.

27. McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, et al. (2011) Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471: 216–219.

28. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, et al. (2012) Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet 8: e1002841.

29. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654.

30. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158: 1227–1234.

31. Andolfatto P (2005) Adaptive evolution of non-coding DNA in Drosophila. Nature 437: 1149–1152.

32. Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173: 891–900.

33. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4: e1000083.

34. Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M (2011) A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet 7: e1002395.

35. Ward LD, Kellis M (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 337: 1675–1678.

36. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342: 1235587.

37. Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, et al. (2013) Genome-wide inference of natural selection on human transcription factor binding sites. Nat Genet 45: 723–729.

38. Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, et al. (2010) Genome-wide discovery of human heart enhancers. Genome Res 20: 381–392.

39. Ritchie GR, Dunham I, Zeggini E, Flicek P (2014) Functional annotation of noncoding sequence variants. Nat Methods 11: 294–296.

40. Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28: 817–825.

41. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, et al. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods 9: 473–476.

42. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, et al. (2013) Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res 41: 827–841.

43. Gronau I, Arbiza L, Mohammed J, Siepel A (2013) Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. Mol Biol Evol 30: 1159–1171.

44. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22: 1760–1774.

45. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46: 310–315.

46. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, et al. (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489: 91–100.

47. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22: 1790–1797.

48. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.

49. Cooper GM, Brudno M, Stone EA, Dubchak I, Batzoglou S, et al. (2004) Characterization of evolutionary rates and constraints in three mammalian genomes. Genome Res 14: 539-548.

50. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803–819.

51. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482.

52. Chiaromonte F, Weber RJ, Roskin KM, Diekhans M, Kent WJ, et al. (2003) The share of human genomic DNA under selection estimated from human-mouse genomic alignments. In: Cold Spring Harbor Symp Quant Biol. volume 68, pp. 245-254.

53. Meader S, Ponting CP, Lunter G (2010) Massive turnover of functional sequence in human and other mammalian genomes. Genome Res 20: 1335–1343.

54. Smith NGC, Brandstrom M, Ellegren H (2004) Evidence for turnover of functional noncoding DNA in mammalian genome evolution. Genomics 84: 806-813.

55. Lunter G, Ponting CP, Hein J (2006) Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput Biol 2: e5.

56. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, et al. (2014) Defining functional DNA elements in the human genome. Proc Natl Acad Sci USA 111: 6131–6138.

57. Pheasant M, Mattick JS (2007) Raising the estimate of functional human sequences. Genome Res 17: 1245–1253.

58. Green P, Ewing B (2013) Comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions". Science 340: 682.

59. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327: 78–81.

60. Hubisz MJ, Pollard KS, Siepel A (2011) PHAST and RPHAST: Phylogenetic analysis with space/time models. Briefings in Bioinformatics 12: 41–51.

61. Kondrashov AS, Crow JF (1993) A molecular approach to estimating the human deleterious mutation rate. Hum Mutat 2: 229–234.

62. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, et al. (2014) The UCSC Genome Browser database: 2014 update. Nucleic Acids Res 42: D764–770.

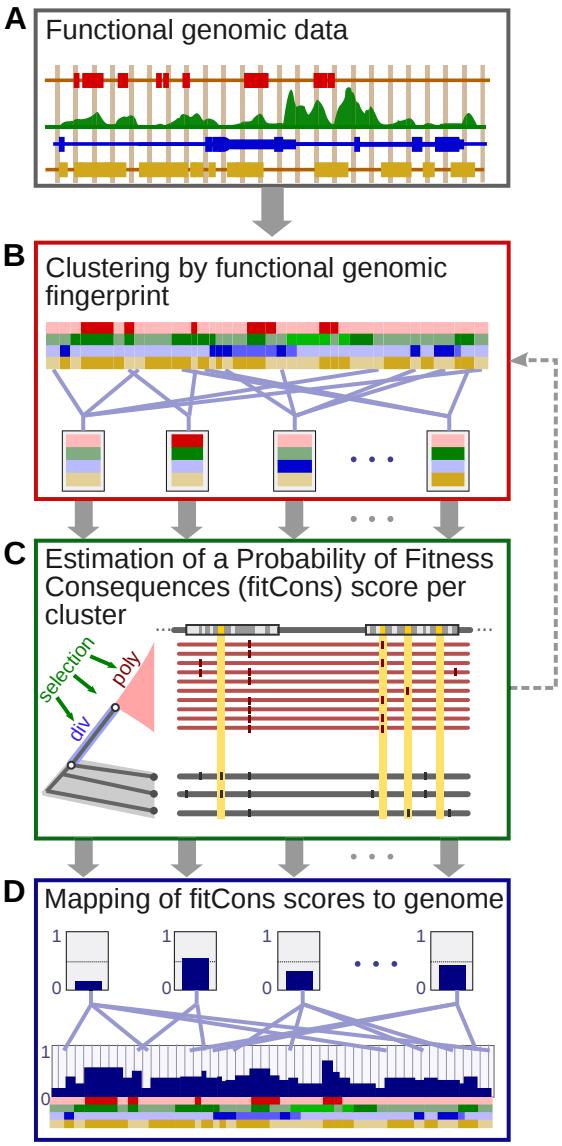


Figure 1: **Illustration of procedure for calculating fitCons scores.** (A) Functional genomic data, such as DNase-seq, RNA-seq and histone modification data, are arranged along the genome sequence in tracks. (B) Nucleotide positions in the genome are clustered by joint patterns across these functional genomic tracks. (C) For each cluster, patterns of polymorphism and divergence are analyzed using INSIGHT [43] to obtain an estimate of the fraction of nucleotides under natural selection (ρ) in that cluster. This quantity is interpreted as a probability that each nucleotide position influences the fitness of the organism that carries it, or a fitness consequence (fitCons) score. (D) The fitCons score for each cluster is assigned to all genomic positions that were included in the cluster. 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, the clustering is 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 line).

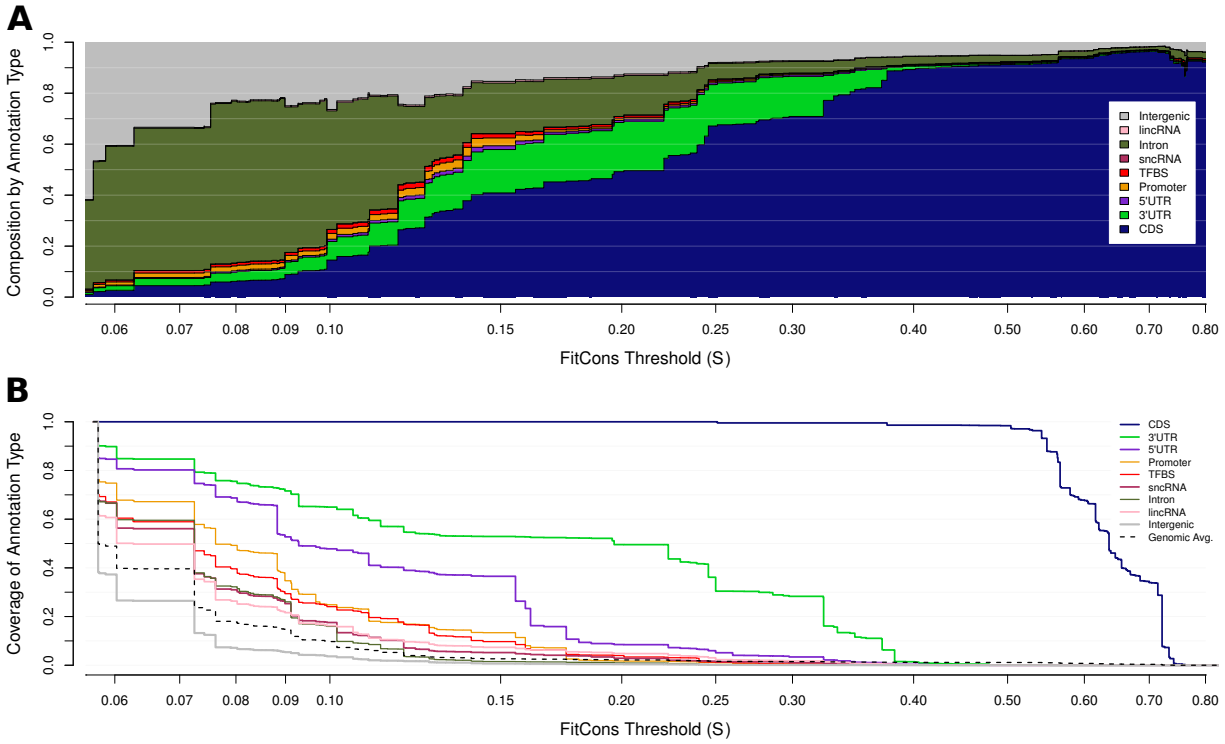


Figure 2: **Composition and coverage of high-scoring genomic regions according to fitCons.** (A) Composition by annotation type in regions that exceed a fitCons score threshold of S, as S is varied across the range of possible scores. Each vertical cross-section of the plot can be thought of as a narrow "stacked bar" representation of the composition by annotation type of all genomic positions at which the fitCons score $> S$. 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, coding sequences (CDS; navy) are increasingly enriched and intergenic (gray) sequences are increasingly depleted. Regions experiencing moderate levels of selection, such as untranslated regions (UTRs; purple and green), promoters (orange), small noncoding RNAs (sncRNAs; eggplant), and introns (olive), are most enriched at intermediate scores. A logarithmic scale is used for the x-axis to enable the differences at low score thresholds to be observed more clearly. (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.

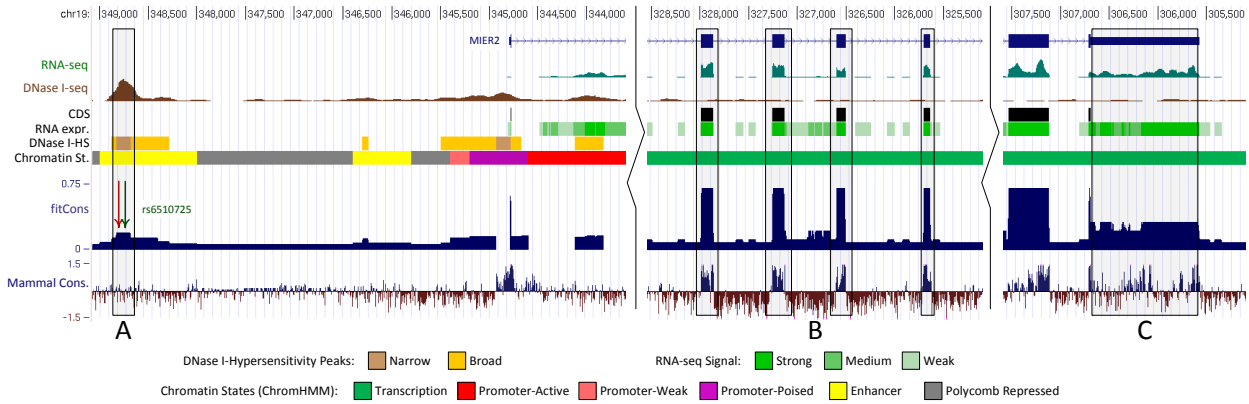


Figure 3: **Genome browser display showing functional genomic fingerprints and fitCons scores.** Shown, from top to bottom, are the exons of the *MIER2* gene (blue); the raw RNA-seq (turquoise) and DNase-seq (brown) signals; the four discretized tracks used to define the 624 functional genomic fingerprints, including annotation-based CDSs (black), RNA-seq signal (green), DNase-seq signal (yellow and brown), and chromatin modifications (multiple colors; see key at bottom); the fitCons scores based on those fingerprints; 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. A ChIP-seq-supported TFBS 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 absense of DNase-I hypersensitivity or CDS annotations, displays moderately elevated fitCons scores and patches of evolutionary conservation. Browser tracks are publicly available at http://genome-mirror.bscb.cornell.edu (hg19 assembly).

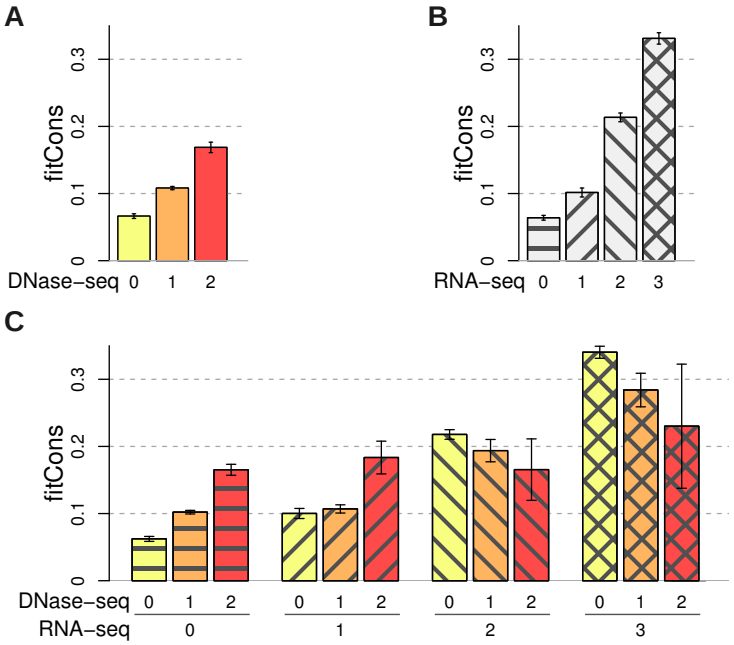


Figure 4: **Average fitCons scores as a function of DNase-seq and RNA-seq intensity.** Results represent averages across all non-CDS clusters having the marginal or joint property of interest. (A) FitCons scores increase with DNase-seq intensity. 0: No DNase-seq signal; 1: broad peaks; 2: narrow peaks. (B) FitCons scores increase with RNA-seq intensity. 0: no RNA-seq reads; 1-3: weak to strong RNA-seq signal (see Methods). (C) FitCons scores behave in a non-additive manner as joint combinations of DNase- 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.

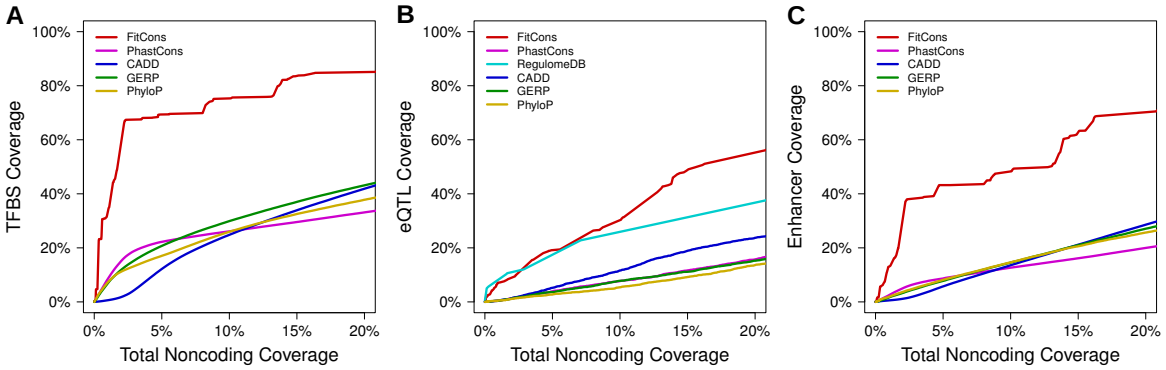


Figure 5: **Coverage of active *cis*-regulatory elements as a function of total coverage of the noncoding genome.** Coverage of each type of element is shown as the score threshold is adjusted to alter the total coverage of noncoding sequences in the genome, excluding sites annotated as CDS or UTR. FitCons is compared with scores from the Combined Annotation Dependent Depletion (CADD) [45], Genomic Evolutionary Rate Profiling (GERP) [13], phastCons [12], and phyloP [15] programs (see Methods). (A) Coverage of transcription factor binding sites detected by ChIP-seq in HUVECs [37]. (B) Coverage of high-resolution eQTLs identified in a recent large-scale study [6], restricted to eQTLs associated with genes transcribed in HUVECs. Coverage of eQTLs is also computed for classification of single nucleotide variants by RegulomeDB [47]. (C) Coverage of enhancers identified by characteristic chromatin marks [46] assayed in HUVECs.

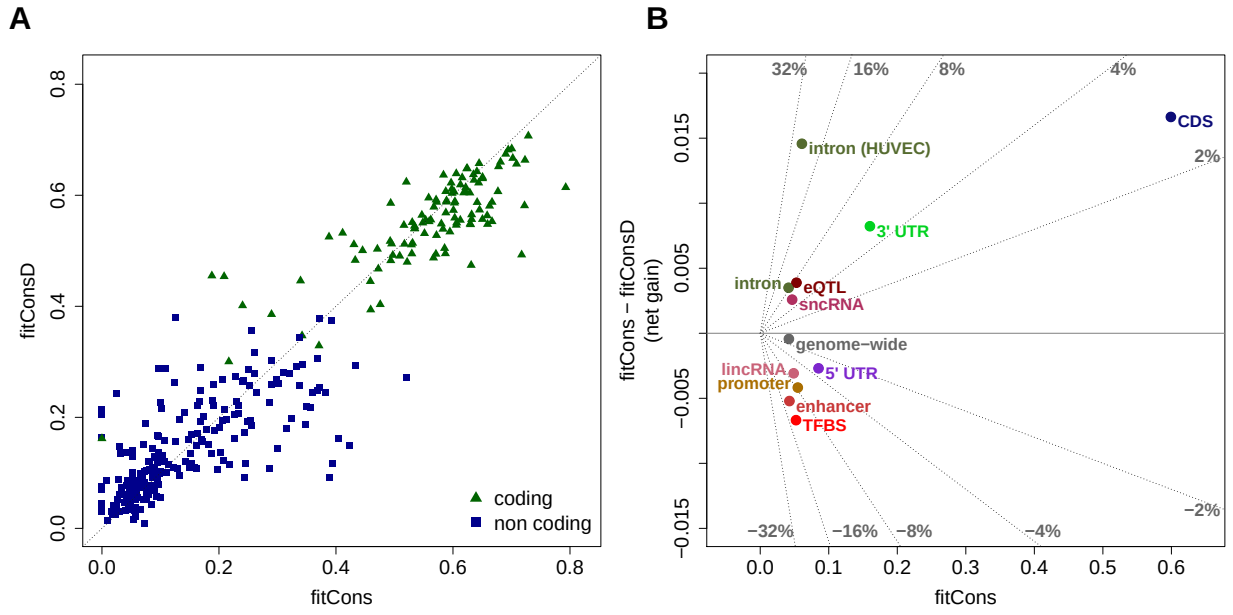


Figure 6: **Comparison between fitCons and fitConsD scores.** FitConsD is an alternative estimate of fitness consequences, computed using a similar computational pipeline developed for fitCons, but with a different evolutionary model that considers divergence between four primate genomes (see Methods). (A) FitCons and FitConsD scores are shown for the clusters defined using functional genomic data from HUVECs. Scores are shown for the 325 clusters of size 10 kb or larger, distinguishing between coding clusters (green triangles) and non-coding clusters (blue squares). Both sets of scores are corrected by subtracting the possible contribution from model misspecification (see Methods). Correlation between the two sets of scores is high ($R^2 = 0.89$), with slightly higher correlation for coding ($R^2 = 0.68$) than for noncoding ($R^2 = 0.51$). (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. 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," computed as NGR = (fitCons − fitConsD)/fitCons. The NGR can be interpreted as an estimate of the net gain per functional site. Notice that the NGR is fairly small for all annotation classes considered, with the exception of the introns of genes transcribed in the cell type considered (HUVEC) (see text).

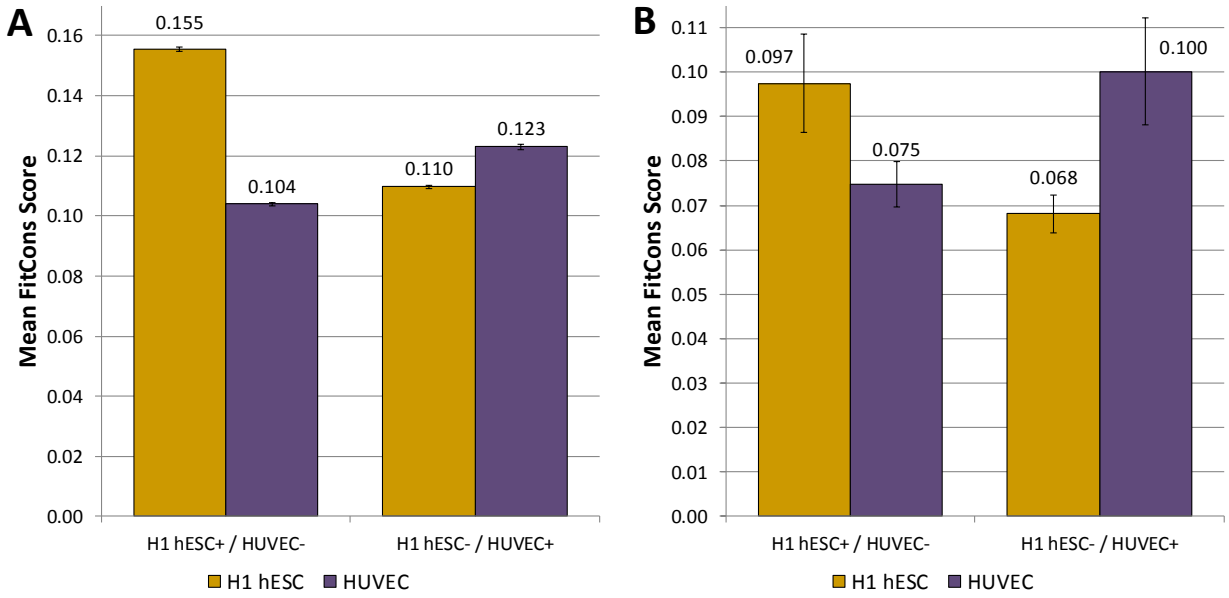
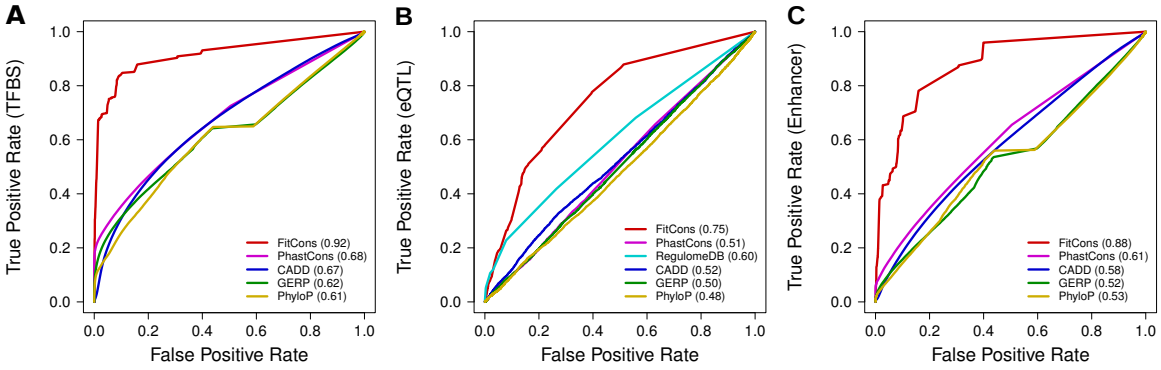
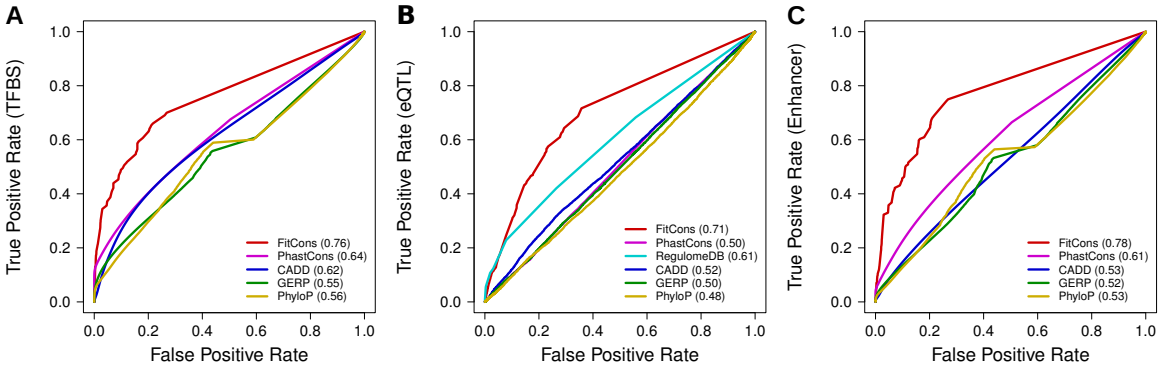


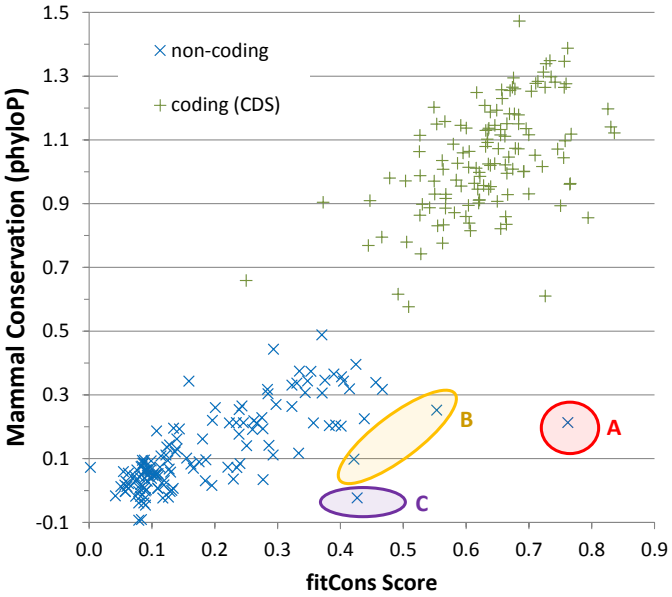
Figure 7: **FitCons scores reflect cell-type specific activity.** Mean fitCons score for (A) eQTL and (B) 100 bp promoters that are active in one cell type and inactive in another, based on RNA-seq data for the associated gene (see Methods). FitCons scores computed using functional genomic data from H1 hESC cells (gold bars), elements active in H1 hESC and inactive in HUVEC (H1 hESC+/HUVEC−) are significantly higher than those for elements inactive in H1 hESC and active in HUVEC (H1 hESC−/HUVEC+). The opposite pattern is observed for fitCons scores computed using functional genomic data from HUVEC (purple bars).



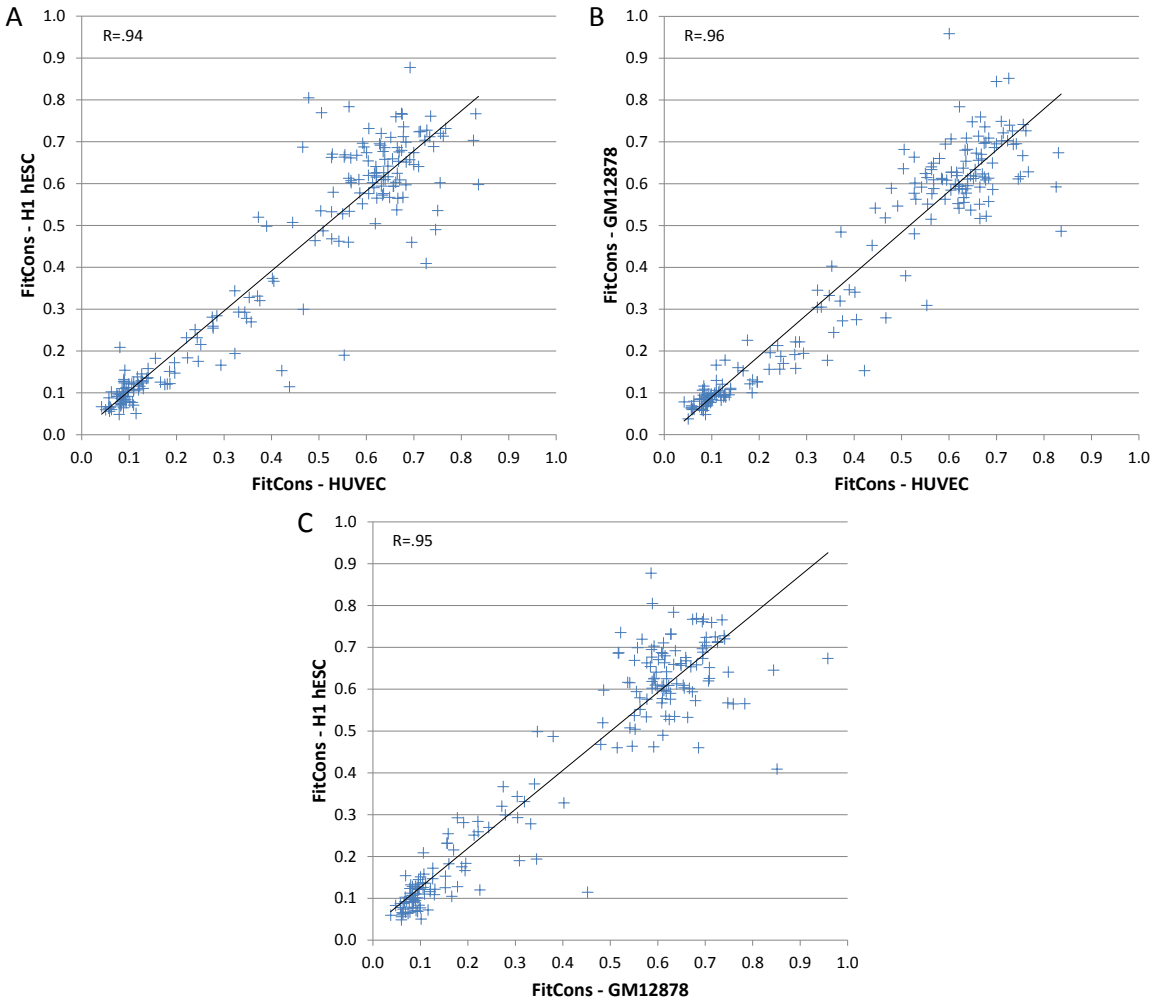
Supplementary Figure S1: **Receiver operating characteristic (ROC) curves for cell-type-specific regulatory elements.** Three types of regulatory elements were considered: (A) transcription factor binding sites (TFBSs), (B) expression QTLs (eQTLs), and (C) enhancers identified by chromatin marks. Separate curves are shown for fitCons, CADD [45], GERP [13], and phyloP [15] scores. True positive rates were estimated by the fraction of nucleotides in annotated elements having scores that exceed a given score threshold and false positive rates were estimated by the fraction of nucleotides in a matched set of "null" elements having scores that exceed the same threshold (see Methods for details). Each curve is generated by varying this threshold across the full range of scores for the corresponding method. In this case, only elements "active" in the cell-type for which the fitCons scores were produced (HUVEC) were considered; see Supplementary Figure S2 for results for a pooled set of elements across cell types. AUC values, shown in parentheses, represent areas under the ROC curve and provide an overall measure of predictive value.



Supplementary Figure S2: **Receiver operating characteristic (ROC) curves for regulatory elements pooled across cell types.** Three types of regulatory elements were considered: (A) transcription factor binding sites (TFBSs), (B) expression QTLs (eQTLs), and (C) enhancers identified by chromatin marks. Separate curves are shown for fitCons, CADD [45], GERP [13], and phyloP [15] scores. The fitCons scores used here are computed by aggregating functional information across HUVEC, H1 hESC, and GM12878 cells (see Methods). Note that some regulatory elements might not be active in any of the three cell types.



Supplementary Figure S3: **Comparison of fitCons scores and phylogenetic conservation scores.** Each of the 624 clusters is represented by single point, with its x coordinate given by the mean placental mammalian phyloP score for the associated genomic positions [15, 62] and its y coordinate given by the fitCons score calculated as shown in Fig. 1. The clusters naturally fall in two groups, corresponding to coding sequences (CDS) with lower scores (green crosses) and noncoding sequences with higher scores (blue Xs). Three groups of outliers are shown, representing clusters with elevated fitCons scores relative to their phyloP scores. Cluster (A) consists of 1200 genomic positions in narrow DNase-seq peaks with no RNA-Seq signal, yet with chromatin modifications indicating transcription activity. These sites are strongly enriched for ChIP-seq-supported TFBSs, and may contain enhancers with lowly expressed eRNAs not detectable from the available RNA-seq data. The two clusters in (B) contain 92.8 kb of sequence defined by high RNA-seq signals, broad DNase-seq peaks, and Pol II binding, and are strongly enriched for $3'$ UTR and ncRNA annotations. Cluster (C) contains 52.7 kb of sequence with no DNase-seq but some RNA-seq signal, along with insulator-associated chromatin modifications. This class is strongly enriched for eQTLs and CTCF binding sites, suggesting transcriptional silencing activity. Thus, all four of these clusters appear to be rich in regulatory sequences that could plausibly have experienced weak natural selection during most of mammalian evolution, but come under stronger selection recently on the human lineage.



Supplementary Figure S4: FitCons scores for all functional classes for (A) HUVEC vs. H1 hESC, (B) HUVEC vs. GM12878 and (C) GM12878 vs. H1 hESC cells. While the individual positions assigned to each class vary widely according to cell type, the fitCons scores remain relatively constant, with Pearson correlations > 0.94 and Spearman correlations > 0.87 between pairs of cell types.

Supplementary Table S1: **FitCons Site Clusters**

Cell type	Number of clusters			Coverage[d]	Median size[e]
	Non-empty [a]	Viable [b]	Confident [c]		
HUVEC	496	447	269	99.5%	165 kb
H1 hESC	491	444	271	99.1%	174 kb
GM12878	493	443	258	99.4%	224 kb

[a]Number of non-empty clusters in partitioning.
[b]Number of non-empty clusters after applying INSIGHT filters.
[c]Number of clusters for which the ratio between the estimated standard error for ρ and the estimated value of ρ is at most 0.4.
[d]Percentage of bases in confident clusters out of 2,881,033,286 bp in human autosomes.
[e]Median size (in kilobases) of confident clusters.

Supplementary Table S2: **Share Under Selection for Various Annotation Classes**

Class	Size (Mbp)	% of genome	% of sites in class under selection[a]	% of selected sites in class[b]	Adjusted % of selected sites in class[c]
genome	2,881.0	100.0%	7.5%±0.1%	100.0%	100.0%
coding	30.7	1.1%	63.2%±0.4%	9.0%	15.3%
3' UTR	24.7	0.9%	19.5%±0.3%	2.2%	3.3%
5' UTR	4.1	0.1%	13.4%±0.5%	0.3%	0.3%
promoter	21.5	0.7%	9.0%±0.3%	0.9%	1.0%
ncRNA[d]	8.0	0.3%	8.1%±0.1%	0.3%	0.3%
intron	1,008.8	35.4%	7.5%±0.1%	35.2%	35.1%
intergenic	1,768.2	61.5%	6.3%±0.1%	51.7%	44.1%

[a]Fraction of sites under selection in a class computed using HUVEC fitCons scores (see Methods).
[b]Fraction of total number of sites under selection in the genome that is estimated to fall in class.
[c]Fraction of total number of sites under selection in the genome that is estimated to fall in class, corrected by considering the lower bound on ρ obtained by subtracting $\rho_{\text{neut}} = 3.3\%$ (see Methods).
[d]Union of lincRNA set and sncRNA set.

Supplementary Table S3: **Sources of functional genomic data**

Data Type	Cell Type	Source	URL and File
DNase-seq data	——-	University of Washington	http://hgdownload-test.cse.ucsc.edu/goldenPath/currentGenomes/Homo_sapiens/encodeDCC/wgEncodeUwDnase/
DNase-seq broad peaks	HUVEC		wgEncodeUwDnaseHuvecHotspotsRep{1,2}.broadPeak.gz
DNase-seq narrow peaks	HUVEC		wgEncodeUwDnaseHuvecPkRep{1V2,2}.narrowPeak.gz
DNase-seq broad peaks (rep 1&2)	GM12878		wgEncodeUwDnaseGm12878HotspotsRep{1,2}.broadPeak.gz
DNase-seq narrow peaks (rep 1&2)	GM12878		wgEncodeUwDnaseGm12878PkRep{1,2}.narrowPeak.gz
DNase-seq broad peaks (rep 1)	H1 hESC		wgEncodeUwDnaseH1hescHotspotsRep1.broadPeak.gz
DNase-seq narrow peaks (rep 1)	H1 hESC		wgEncodeUwDnaseH1hescPkRep1.narrowPeak.gz
DNase-seq narrow peaks (rep 2)	H1 hESC	Duke University	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeOpenChromDnase/ wgEncodeOpenChromDnaseH1hescPk.narrowPeak.gz
RNA-seq	——-	Caltech	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/
RNA-seq	HUVEC		wgEncodeCaltechRnaSeqHuvecR2x75Th1014Il200SigRep1V4.bigWig
RNA-seq	GM12878		wgEncodeCaltechRnaSeqGm12878R2x75Il200SigRep2V2.bigWig
RNA-seq	H1 hESC		wgEncodeCaltechRnaSeqH1hescR2x75Th1014Il200SigRep2V4.bigWig
Chromatin states	——-	Broad Institute	http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/segmentations/jan2011/hub/
Chromatin states	GM12878		gm12878.ChromHMM.bb
Chromatin states	H1 hESC		h1hesc.ChromHMM.bb
Chromatin states	HUVEC		huvec.ChromHMM.bb
Annotated transcripts	——-	GENCODE v15	ftp://ftp.sanger.ac.uk/pub/gencode/release_15/gencode.v15.annotation.gtf.gz