

# Defining functional DNA elements in the human genome

Manolis Kellis<sup>a,b,1,2</sup>, Barbara Wold<sup>c,2</sup>, Michael P. Snyder<sup>d,2</sup>, Bradley E. Bernstein<sup>b,e,f,2</sup>, Anshul Kundaje<sup>a,b,3</sup>, Georgi K. Marinov<sup>c,3</sup>, Lucas D. Ward<sup>a,b,3</sup>, Ewan Birney<sup>g</sup>, Gregory E. Crawford<sup>h</sup>, Job Dekker<sup>i</sup>, Ian Dunham<sup>g</sup>, Laura L. Elnitskij<sup>j</sup>, Peggy J. Farnham<sup>k</sup>, Elise A. Feingold<sup>j</sup>, Mark Gerstein<sup>l</sup>, Morgan C. Giddings<sup>m</sup>, David M. Gilbert<sup>n</sup>, Thomas R. Gingeras<sup>o</sup>, Eric D. Green<sup>j</sup>, Roderic Guigo<sup>p</sup>, Tim Hubbard<sup>q</sup>, Jim Kent<sup>r</sup>, Jason D. Lieb<sup>s</sup>, Richard M. Myers<sup>t</sup>, Michael J. Pazin<sup>j</sup>, Bing Ren<sup>u</sup>, John A. Stamatoyannopoulos<sup>v</sup>, Zhiping Weng<sup>j</sup>, Kevin P. White<sup>w</sup>, and Ross C. Hardison<sup>x,1,2</sup>

<sup>a</sup>Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; <sup>b</sup>Broad Institute, Cambridge, MA 02139; <sup>c</sup>Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125; <sup>d</sup>Department of Genetics, Stanford University, Stanford, CA 94305; <sup>e</sup>Harvard Medical School and <sup>f</sup>Massachusetts General Hospital, Boston, MA 02114; <sup>g</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom; <sup>h</sup>Medical Genetics, Duke University, Durham, NC 27708; <sup>i</sup>Program in Systems Biology, University of Massachusetts Medical School, Worcester, MA 01605; <sup>j</sup>National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892; <sup>k</sup>Biochemistry and Molecular Biology, University of Southern California, Los Angeles, CA 90089; <sup>l</sup>Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520; <sup>m</sup>Marketing Your Science, LLC, Boise, ID 83702; <sup>n</sup>Department of Biological Science, Florida State University, Tallahassee, FL 32306; <sup>o</sup>Functional Genomics Group, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; <sup>p</sup>Bioinformatics and Genomics Program, Center for Genome Regulation, E-08003 Barcelona, Catalonia, Spain; <sup>q</sup>Medical and Molecular Genetics, King's College London and Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SD, United Kingdom; <sup>r</sup>Biomolecular Engineering, University of California, Santa Cruz, CA 95064; <sup>s</sup>Lewis Sigler Institute, Princeton University, Princeton, NJ 08544; <sup>t</sup>HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806; <sup>u</sup>Ludwig Institute for Cancer Research, University of California, San Diego, La Jolla, CA 92093; <sup>v</sup>Genome Sciences and Medicine, University of Washington, Seattle, WA 98195; <sup>w</sup>Human Genetics, University of Chicago, Chicago, IL 60637; and <sup>x</sup>Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved January 29, 2014 (received for review October 16, 2013)

With the completion of the human genome sequence, attention turned to identifying and annotating its functional DNA elements. As a complement to genetic and comparative genomics approaches, the Encyclopedia of DNA Elements Project was launched to contribute maps of RNA transcripts, transcriptional regulator binding sites, and chromatin states in many cell types. The resulting genome-wide data reveal sites of biochemical activity with high positional resolution and cell type specificity that facilitate studies of gene regulation and interpretation of noncoding variants associated with human disease. However, the biochemically active regions cover a much larger fraction of the genome than do evolutionarily conserved regions, raising the question of whether nonconserved but biochemically active regions are truly functional. Here, we review the strengths and limitations of biochemical, evolutionary, and genetic approaches for defining functional DNA segments, potential sources for the observed differences in estimated genomic coverage, and the biological implications of these discrepancies. We also analyze the relationship between signal intensity, genomic coverage, and evolutionary conservation. Our results reinforce the principle that each approach provides complementary information and that we need to use combinations of all three to elucidate genome function in human biology and disease.

## Quest to Identify Functional Elements in the Human Genome

Completing the human genome reference sequence was a milestone in modern biology. The considerable challenge that remained was to identify and delineate the structures of all genes and other functional elements. It was quickly recognized that nearly 99% of the ~3.3 billion nucleotides that constitute the human genome do not code for proteins (1). Comparative genomics studies revealed that the majority of mammalian-conserved and recently adapted regions consist of noncoding elements (2–10). More recently, genome-wide association studies have indicated that a majority of trait-associated loci, including ones that contribute to human diseases and susceptibility, also lie outside protein-coding regions (11–16). These findings suggest that the

noncoding regions of the human genome harbor a rich array of functionally significant elements with diverse gene regulatory and other functions.

Despite the pressing need to identify and characterize all functional elements in the human genome, it is important to recognize that there is no universal definition of what constitutes function, nor is there agreement on what sets the boundaries of an element. Both scientists and nonscientists have an intuitive definition of function, but each scientific discipline relies primarily on different lines of evidence indicative of function. Geneticists, evolutionary biologists, and molecular biologists apply distinct approaches, evaluating different and complementary lines of evidence. The genetic approach evaluates the phenotypic consequences of perturbations, the evolutionary

Author contributions: M.K., B.W., M.P.S., B.E.B., and R.C.H. designed research; M.K., B.W., M.P.S., B.E.B., A.K., G.K.M., L.D.W., and R.C.H. performed research; A.K., G.K.M., and L.D.W. contributed computational analysis and tools; M.K., B.W., M.P.S., B.E.B., E.B., G.E.C., J.D., I.D., L.L.E., P.J.F., E.A.F., M.G., M.C.G., D.M.G., T.R.G., E.D.G., R.G., T.H., J.K., J.D.L., R.M.M., M.J.P., B.R., J.A.S., Z.W., K.P.W., and R.C.H. contributed to manuscript discussions and ideas; and M.K., B.W., M.P.S., B.E.B., and R.C.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: In addition to data already released via the ENCODE Data Coordinating Center, the erythroblast DNase-seq data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, [www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo) (accession nos. GSE55579, GSM133959, and GSM1339560).

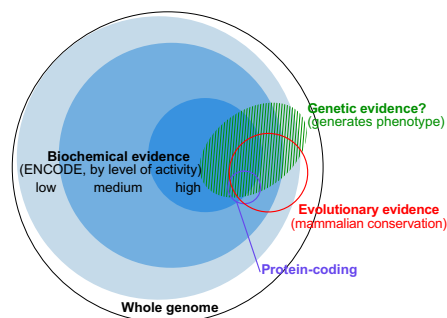
Authored by members of the ENCODE Consortium.

<sup>1</sup>To whom correspondence may be addressed. E-mail: manoli@mit.edu or rch8@psu.edu.

<sup>2</sup>M.K., B.W., M.P.S., B.E.B., and R.C.H. contributed equally to this work.

<sup>3</sup>A.K., G.K.M., and L.D.W. contributed equally to this work.

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1318948111/-/DCSupplemental](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1318948111/-/DCSupplemental).



**Fig. 1.** The complementary nature of evolutionary, biochemical, and genetic evidence. The outer circle represents the human genome. Blue discs represent DNA sequences acted upon biochemically and partitioned by their levels of signal [combined 10th percentiles of different ENCODE data types for high, combined 50th percentiles for medium, and all significant signals for low (see *Reconciling Genetic, Evolutionary, and Biochemical Estimates* and Fig. 2)]. The red circle represents, at the same scale, DNA with signatures of evolutionary constraint (GERP++ elements derived from 34 mammal alignments). Overlaps among the sequences having biochemical and evolutionary evidence were computed in this work (Fig. 3 and *SI Methods*). The small purple circle represents protein-coding nucleotides (Gencode). The green shaded domain conceptually represents DNA that produces a phenotype upon alteration, although we lack well-developed summary estimates for the amount of genetic evidence and its relationship with the other types. This summary of our understanding in early 2014 will likely evolve substantially with more data and more refined experimental and analytical methods.

approach quantifies selective constraint, and the biochemical approach measures evidence of molecular activity. All three approaches can be highly informative of the biological relevance of a genomic segment and groups of elements identified by each approach are often quantitatively enriched for each other. However, the methods vary considerably with respect to the specific elements they predict and the extent of the human genome annotated by each (Fig. 1).

Some of these differences stem from the fact that function in biochemical and genetic contexts is highly particular to cell type and condition, whereas for evolutionary measures, function is ascertained independently of cellular state but is dependent on environment and evolutionary niche. The methods also differ widely in their false-positive and false-negative rates, the resolution with which elements are defined, and the throughput with which they can be surveyed. Moreover, each approach remains incomplete, requiring continued method development (both experimental and analytical) and increasingly large datasets (additional species, assays, cell types, variants, and phenotypes). It is thus not surprising that the methods vary considerably with respect to the specific elements they identify. However, the extent of the difference is much larger than simply

technical limitations would suggest, challenging current views and definitions of genome function.

Many examples of elements that appear to have conflicting lines of functional evidence were described before the Encyclopedia of DNA Elements (ENCODE) Project, including elements with conserved phenotypes but lacking sequence-level conservation (17–20), conserved elements with no phenotype on deletion (21, 22), and elements able to drive tissue-specific expression but lacking evolutionary conservation (23, 24). However, the scale of the ENCODE Project survey of biochemical activity (across many more cell types and assays) led to a significant increase in genome coverage and thus accentuated the discrepancy between biochemical and evolutionary estimates. This discrepancy led to much debate both in the scientific literature (25–31) and in online forums, resulting in a renewed need to clarify the challenges of defining function in the human genome and to understand the sources of the discrepancy.

To address this need and provide a perspective by ENCODE scientists, we review genetic, evolutionary, and biochemical lines of evidence, discuss their strengths and limitations, and examine apparent discrepancies between the conclusions emanating from the different approaches.

**Genetic Approach.** Genetic approaches, which rely on sequence alterations to establish the biological relevance of a DNA segment, are often considered a gold standard for defining function. Mutations can be naturally occurring and identified by screening for phenotypes generated by sequence variants (13, 32) or produced experimentally by targeted genetic methods (33) or nongenetic interference (34). Transfection studies that use reporter assays in cell lines (35, 36) or embryos (37) can also be used to identify regulatory elements and measure their activities. Genetic approaches tend to be limited by modest throughput, although speed and efficiency is now increasing for some methods (36, 38–40). The approach may also miss elements whose phenotypes occur only in rare cells or specific environmental contexts, or whose effects are too subtle to detect with current assays. Loss-of-function tests can also be buffered by functional redundancy, such that double or triple disruptions are required for a phenotypic consequence. Consistent with redundant, contextual, or subtle functions, the deletion of large and highly conserved genomic segments sometimes has no discernible organismal phenotype (21, 22),

and seemingly debilitating mutations in genes thought to be indispensable have been found in the human population (41).

**Evolutionary Approach.** Comparative genomics provides a powerful approach for detecting noncoding functional elements that show preferential conservation across evolutionary time. A high level of sequence conservation between related species is indicative of purifying selection, whereby disruptive mutations are rejected, with the corresponding sequence deemed to be likely functional. Evidence of function can also come from accelerated evolution across species or within a particular lineage, revealing elements under positive selection for recently acquired changes that increase fitness; such an approach gains power by incorporating multiple closely related genomes because each species provides information about sequence constraint. Multispecies comparisons have been used in studies of diverse clades, ranging from yeast to mammals. Methods that detect sequences likely under selection have had success in recognizing protein-coding regions, structural RNAs, gene regulatory regions, regulatory motifs, and specific regulatory elements (3, 42–48). The comparative genomics approach can also incorporate information about mutational patterns that may be characteristic of different types of elements.

Although powerful, the evolutionary approach also has limitations. Identification of conserved regions depends on accurate multispecies sequence alignments, which remain a substantial challenge. Alignments are generally less effective for distal-acting regulatory regions, where they may be impeded by regulatory motif turnover, varying spacing constraints, and sequence composition biases (17, 49). Analyzing aligned regions for conservation can be similarly challenging. First, most transcription factor-binding sequences are short and highly degenerate, making them difficult to identify. Second, because detection of neutrally evolving elements requires sufficient phylogenetic distance, the approach is well suited for detecting mammalian-conserved elements, but it is less effective for primate-specific elements and essentially blind to human-specific elements. Third, certain types of functional elements such as immunity genes may be prone to rapid evolutionary turnover even among closely related species. More generally, alignment methods are not well suited to capture substitutions that preserve function, such as compensatory changes preserving RNA structure, affinity-preserving substitutions

within regulatory motifs, or mutations whose effect is buffered by redundancy or epistatic effects. Thus, absence of conservation cannot be interpreted as evidence for the lack of function.

Finally, although the evolutionary approach has the advantage that it does not require a priori knowledge of what a DNA element does or when it is used, it is unlikely to reveal the molecular mechanisms under selection or the relevant cell types or physiological processes. Thus, comparative genomics requires complementary studies.

**Biochemical Approach.** The biochemical approach for identifying candidate functional genomic elements complements the other approaches, as it is specific for cell type, condition, and molecular process. Decades of detailed studies of gene regulation and RNA metabolism have defined major classes of functional noncoding elements, including promoters, enhancers, silencers, insulators, and noncoding RNA genes such as microRNAs, piRNAs, structural RNAs, and regulatory RNAs (50–53). These noncoding functional elements are associated with distinctive chromatin structures that display signature patterns of histone modifications, DNA methylation, DNase accessibility, and transcription factor occupancy (37, 54–66). For example, active enhancers are marked by specific histone modifications and DNase-accessible chromatin and are occupied by sequence-specific transcription factors, coactivators such as EP300, and, often, RNA polymerase II. Although the extent to which individual features contribute to function remains to be determined, they provide a useful surrogate for annotating candidate enhancers and other types of functional elements.

The ENCODE Project was established with the goal of systematically mapping functional elements in the human genome at high resolution and providing this information as an open resource for the research community (67, 68). Most data acquisition in the project thus far has taken the biochemical approach, using evidence of cellular or enzymatic processes acting on a DNA segment to help predict different classes of functional elements. The recently completed phase of ENCODE applied a wide range of biochemical assays at a genome-wide scale to study multiple human cell types (69). These assays identified genomic sequences (i) from which short and long RNAs, both nuclear and cytoplasmic, are transcribed; (ii) occupied by sequence-specific transcription factors, cofactors, or chromatin

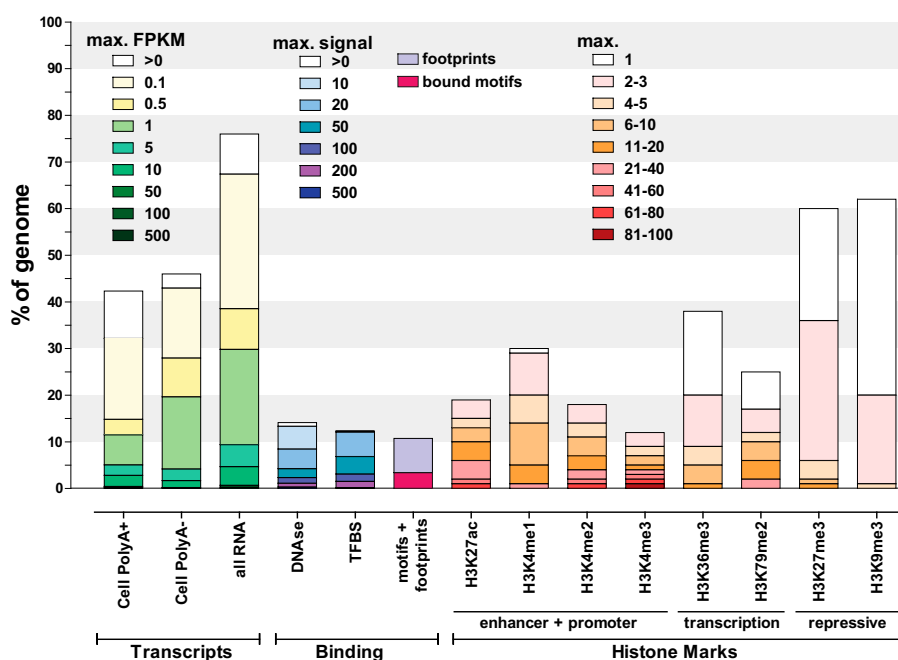
regulatory proteins; (iii) organized in accessible chromatin; (iv) marked by DNA methylation or specific histone modifications; and (v) physically brought together by long-range chromosomal interactions.

An advantage of such functional genomics evidence is that it reveals the biochemical processes involved at each site in a given cell type and activity state. However, biochemical signatures are often a consequence of function, rather than causal. They are also not always deterministic evidence of function, but can occur stochastically. For example, GATA1, whose binding at some erythroid-specific enhancers is critical for function, occupies many other genomic sites that lack detectable enhancer activity or other evidence of biological function (70). Likewise, although enhancers are strongly associated with characteristic histone modifications, the functional significance of such modifications remains unclear, and the mere presence of an enhancer-like signature does not necessarily indicate that

a sequence serves a specific function (71, 72). In short, although biochemical signatures are valuable for identifying candidate regulatory elements in the biological context of the cell type examined, they cannot be interpreted as definitive proof of function on their own.

### What Fraction of the Human Genome Is Functional?

Limitations of the genetic, evolutionary, and biochemical approaches conspire to make this seemingly simple question difficult to answer. In general, each approach can be used to lend support to candidate elements identified by other methods, although focusing exclusively on the simple intersection set would be much too restrictive to capture all functional elements. However, by probing quantitative relationships in data from the different approaches, we can begin to gain a more sophisticated picture of the nature, identity, and extent of functional elements in the human genome.



**Fig. 2.** Summary of the coverage of the human genome by ENCODE data. The fraction of the human genome covered by ENCODE-detected elements in at least one cell line or tissue for each assay is shown as a bar graph. All percentages are calculated against the whole genome, including the portion that is not uniquely mappable with short reads and thus is invisible to the analysis presented here (see Fig. S1). A more detailed summary can be found in Fig. S2. For transcripts, coverage was calculated from RNA-seq-derived contigs (104) using the count of read fragments per kilobase of exon per million reads (FPKM) and separated into abundance classes by FPKM values. Note that FPKMs are not directly comparable among different subcellular fractions, as they reflect relative abundances within a fraction rather than average absolute transcript copy numbers per cell. Depending on the total amount of RNA in a cell, one transcript copy per cell corresponds to between 0.5 and 5 FPKM in PolyA+ whole-cell samples according to current estimates (with the upper end of that range corresponding to small cells with little RNA and vice versa). “All RNA” refers to all RNA-seq experiments, including all subcellular fractions (Fig. S2). DNase hypersensitivity and transcription-factor (TFBS) and histone-mark ChIP-seq coverage was calculated similarly but divided according to signal strength. “Motifs+footprints” refers to the union of occupied sequence recognition motifs for transcription factors as determined by ChIP-seq and as measured by digital genomic footprinting, with the fuchsia portion of the bar representing the genomic space covered by bound motifs in ChIP-seq. Signal strength for ChIP-seq data for histone marks was determined based on the *P* value of each enriched region (the  $-\log_{10}$  of the *P* value is shown), using peak-calling procedures tailored to the broadness of occupancy of each modification (SI Methods).



**Case for Abundant Junk DNA.** The possibility that much of a complex genome could be nonfunctional was raised decades ago. The C-value paradox (27, 73, 74) refers to the observation that genome size does not correlate with perceived organismal complexity and that even closely related species can have vastly different genome sizes. The estimated mutation rate in protein-coding genes suggested that only up to ~20% of the nucleotides in the human genome can be selectively maintained, as the mutational burden would be otherwise too large (75). The term “junk DNA” was coined to refer to the majority of the rest of the genome, which represent segments of neutrally evolving DNA (76, 77). More recent work in population genetics has further developed this idea by emphasizing how the low effective population size of large-bodied eukaryotes leads to less efficient natural selection, permitting proliferation of transposable elements and other neutrally evolving DNA (78). If repetitive DNA elements could be equated with nonfunctional DNA, then one would surmise that the human genome contains vast nonfunctional regions because nearly 50% of nucleotides in the human genome are readily recognizable as repeat elements, often of high degeneracy. Moreover, comparative genomics studies have found that only 5% of mammalian genomes are under strong evolutionary constraint across multiple species (e.g., human, mouse, and dog) (2, 3).

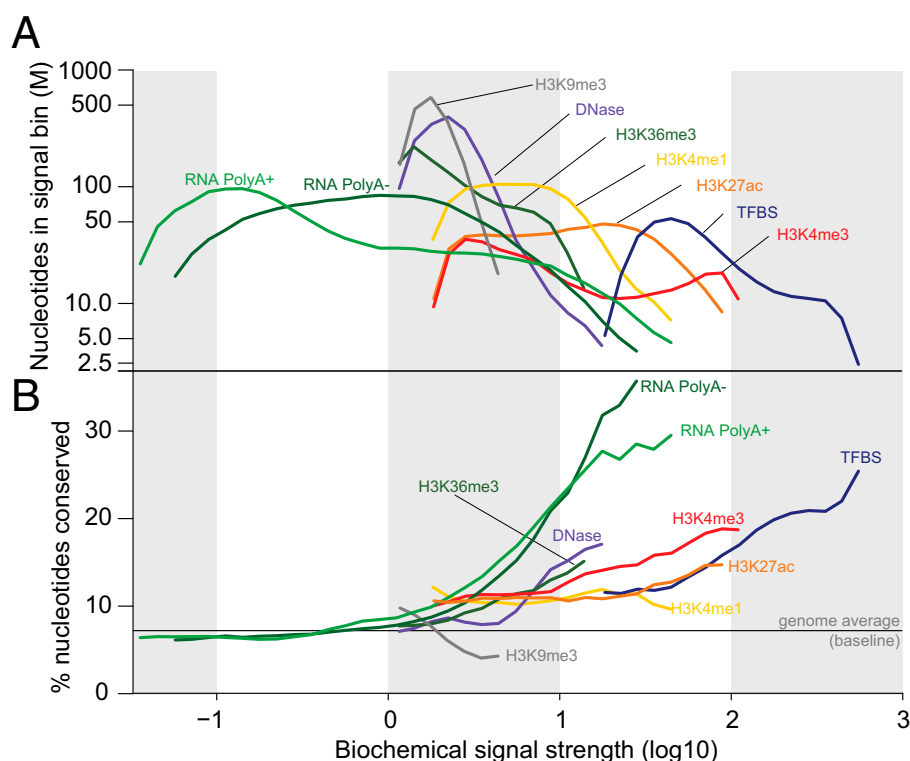
**Case for Abundant Functional Genomic Elements.** Genome-wide biochemical studies, including recent reports from ENCODE, have revealed pervasive activity over an unexpectedly large fraction of the genome, including noncoding and nonconserved regions and repeat elements (58–60). Such results greatly increase upper bound estimates of candidate functional sequences (Fig. 2 and Fig. S2). Many human genomic regions previously assumed to be nonfunctional have recently been found to be teeming with biochemical activity, including portions of repeat elements, which can be bound by transcription factors and transcribed (79, 80), and are thought to sometimes be exapted into novel regulatory regions (81–84). Outside the 1.5% of the genome covered by protein-coding sequence, 11% of the genome is associated with motifs in transcription factor-bound regions or high-resolution DNase footprints in one or more cell types (Fig. 2), indicative of direct contact by regulatory proteins. Transcription factor occupancy and nucleosome-resolution DNase hypersensitivity maps overlap greatly and each cover approximately

15% of the genome. In aggregate, histone modifications associated with promoters or enhancers mark ~20% of the genome, whereas a third of the genome is marked by modifications associated with transcriptional elongation. Over half of the genome has at least one repressive histone mark. In agreement with prior findings of pervasive transcription (85, 86), ENCODE maps of polyadenylated and total RNA cover in total more than 75% of the genome. These already large fractions may be underestimates, as only a subset of cell states have been assayed. However, for multiple reasons discussed below, it remains unclear what proportion of these biochemically annotated regions serve specific functions.

The lower bound estimate that 5% of the human genome has been under evolutionary constraint was based on the excess conservation observed in mammalian alignments (2, 3, 87) relative to a neutral reference (typically ancestral repeats, small introns, or fourfold degenerate codon positions). However, estimates that incorporate alternate references, shape-based constraint (88), evolutionary turnover (89), or lineage-specific constraint (90) each suggests roughly two to three times more constraint than pre-

viously (12–15%), and their union might be even larger as they each correct different aspects of alignment-based excess constraint. Moreover, the mutation rate estimates of the human genome are still uncertain and surprisingly low (91) and not inconsistent with a larger fraction of the genome under relatively weaker constraint (92). Although still weakly powered, human population studies suggest that an additional 4–11% of the genome may be under lineage-specific constraint after specifically excluding protein-coding regions (90, 92, 93), and these numbers may also increase as our ability to detect human constraint increases with additional human genomes. Thus, revised models, lineage-specific constraint, and additional datasets may further increase evolution-based estimates.

Results of genome-wide association studies might also be interpreted as support for more pervasive genome function. At present, significantly associated loci explain only a small fraction of the estimated trait heritability, suggesting that a vast number of additional loci with smaller effects remain to be discovered. Furthermore, quantitative trait locus (QTL) studies have revealed thousands of genetic variants that influence gene



**Fig. 3.** Relationship between ENCODE signals and conservation. Signal strength of ENCODE functional annotations were defined as follows: log10 of signal intensity for DNase and TFBS, log10 of RPKM for RNA, and log10 of  $-\log_{10} P$  value for histone modifications. Annotated regions were binned by 0.1 units of signal strength. (A) The number of nucleotides in each signal bin was plotted. (B) The fraction of the genome in each signal bin covered by conserved elements (by genomic evolutionary rate profiling) (115) was plotted.

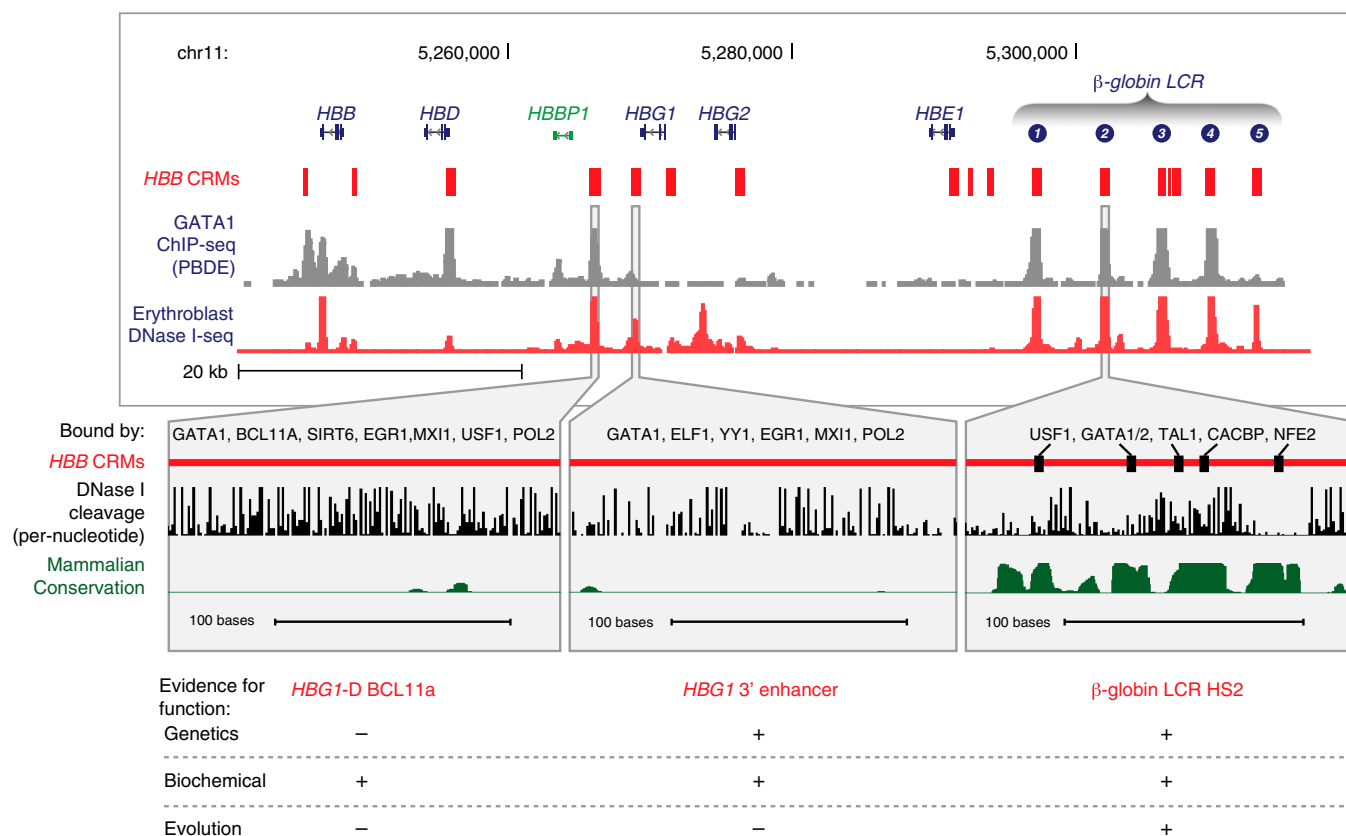
expression and regulatory activity (94–98). These observations raise the possibility that functional sequences encompass a larger proportion of the human genome than previously thought.

### Reconciling Genetic, Evolutionary, and Biochemical Estimates

The proportion of the human genome assigned to candidate functions varies markedly among the different approaches, with estimates from biochemical approaches being considerably larger than those of genetic and evolutionary approaches (Fig. 1). These differences have stimulated scientific debate regarding the interpretation and relative merits of the various approaches (26–29). We highlight below caveats of each approach and emphasize the importance of integration and new high-throughput technologies for refining estimates and better understanding the functional segments in the human genome.

Although ENCODE has expended considerable effort to ensure the reproducibility of detecting biochemical activity (99), it is not at all simple to establish what fraction of the biochemically annotated genome should be regarded as functional. The dynamic range of biochemical signals differs by one or more orders of magnitude for many assays, and the significance of the differing levels is not yet clear, particularly for lower levels. For example, RNA transcripts of some kind can be detected from ~75% of the genome, but a significant portion of these are of low abundance (Fig. 2 and Fig. S2). For polyadenylated RNA, where it is possible to estimate abundance levels, 70% of the documented coverage is below approximately one transcript per cell (100–103). The abundance of complex nonpolyadenylated RNAs and RNAs from subcellular fractions, which account for half of the total RNA coverage of the genome, is likely to be even lower, although their absolute quantification is not

yet achieved. Some RNAs, such as lncRNAs, might be active at very low levels. Others might be expressed stochastically at higher levels in a small fraction of the cell population (104), have hitherto unappreciated architectural or regulatory functions, or simply be biological noise of various kinds. At present, we cannot distinguish which low-abundance transcripts are functional, especially for RNAs that lack the defining characteristics of known protein coding, structural, or regulatory RNAs. A priori, we should not expect the transcriptome to consist exclusively of functional RNAs. Zero tolerance for errant transcripts would come at high cost in the proofreading machinery needed to perfectly gate RNA polymerase and splicing activities, or to instantly eliminate spurious transcripts. In general, sequences encoding RNAs transcribed by noisy transcriptional machinery are expected to be less constrained, which is consistent with data shown here for very low abundance



**Fig. 4.** Epigenetic and evolutionary signals in *cis*-regulatory modules (CRMs) of the *HBB* complex. (Upper) Many CRMs (red rectangles) (106) have been mapped within the cluster of genes encoding  $\beta$ -like globins expressed in embryonic (*HBE1*), fetal (*HBG1* and *HBG2*), and adult (*HBB* and *HBD*) erythroid cells. All are marked by DNase hypersensitive sites and footprints (Gene Expression Omnibus accession nos. GSE55579, GSM1339559, and GSM1339560), and many are bound by GATA1 in peripheral blood derived erythroblasts (PBDEs). (Lower, Left) A DNA segment located between the *HBG1* and *HBD* genes is one of the DNA segments bound by BCL11A (109, 110) and several other proteins (ENCODE uniformly processed data) to negatively regulate *HBG1* and *HBG2*. It is sensitive to DNase I but is not conserved across mammals. (Center) An enhancer located 3' of the *HBG1* gene (red line) (108) is bound by several proteins in PBDEs and K562 cells (from the ENCODE uniformly processed data) and is sensitive to DNase I, but shows almost no signal for mammalian constraint. (Right) The enhancer at hypersensitive site (HS)2 of the locus control region (LCR) (red line) (107) is bound by the designated proteins at the motifs indicated by black rectangles. High-resolution DNase footprinting data (116) show cleavage concentrated between the bound motifs, which are strongly constrained during mammalian evolution, as shown on the mammalian phastCons track (48).

RNA (Fig. 3). Similarly, a majority of the genome shows reproducible evidence of one or more chromatin marks, but some marks are in much lower abundance, are preferentially associated with nonconserved heterochromatin regions (e.g., H3K9me<sub>3</sub>; Fig. 3B), or are known to act at a distance by spreading (105). Indeed, for any given biochemical assay, the proportion of the genome covered is highly dependent on the signal threshold set for the analysis (Fig. 2 and Fig. S2). Regions with higher signals generally exhibit higher levels of evolutionarily conservation (Fig. 3 and Fig. S3). Thus, one should have high confidence that the subset of the genome with large signals for RNA or chromatin signatures coupled with strong conservation is functional and will be supported by appropriate genetic tests. In contrast, the larger proportion of genome with reproducible but low biochemical signal strength and less evolutionary conservation is challenging to parse between specific functions and biological noise.

Another major variable underlying the differences in genome coverage is assay resolution. Biochemical methods, such as ChIP or DNase hypersensitivity assays, capture extended regions of several hundred bases, whereas the underlying transcription factor-binding elements are typically only 6–15 bp in length. Regulatory motifs and DNase footprints within bound regions show much stronger evidence of constraint than surrounding nucleotides that nevertheless fall within the region. Functional elements predicted from chromatin-state annotations tend to span even larger regions (e.g., the median length of enhancer states is ~600 bp), although the driver nucleotides can be similarly few. Biochemical activity may also spread from neighboring regions, in genomic coordinates or 3D genome organization, making it even more difficult to establish the potential nucleotide drivers. Nonetheless, immediately consigning a biochemically marked region to the nonfunctional bin for lack of a driver motif would be premature. Genetic tests by deletion or sequence substitution are needed to resolve the question of their functional significance.

Thus, unanswered questions related to biological noise, along with differences in the resolution, sensitivity, and activity level of the corresponding assays, help to explain divergent estimates of the portion of the human genome encoding functional elements. Nevertheless, they do not account for the entire gulf between constrained regions and biochemical activity. Our analysis revealed a vast portion of the genome that appears to be evolving neutrally according to our

metrics, even though it shows reproducible biochemical activity, which we previously referred to as “biochemically active but selectively neutral” (68). It could be argued that some of these regions are unlikely to serve critical functions, especially those with lower-level biochemical signal. However, we also acknowledge substantial limitations in our current detection of constraint, given that some human-specific functions are essential but not conserved and that disease-relevant regions need not be selectively constrained to be functional. Despite these limitations, all three approaches are needed to complete the unfinished process of inferring functional DNA elements, specifying their boundaries, and defining what functions they serve at molecular, cellular, and organismal levels.

### Functional Genomic Elements and Human Disease

Presently, ~4,000 genes have been associated with human disease, a likely underestimate given that the majority of disease-associated mutations have yet to be mapped. There is overwhelming evidence that variants in the regulatory sequences associated with such genes can lead to disease-relevant phenotypes. Biochemical approaches provide a rich resource for understanding disease-relevant functional elements, but they are most powerful as part of a multifaceted body of evidence for establishing function. Three specific examples from the  $\beta$ -globin locus illustrate how biochemical data can be integrated with evolutionary constraint and genetic assays of function (Fig. 4). The expression of globin genes at progressive stages of development is controlled by transcription factors binding at multiple *cis*-regulatory modules (CRMs) (106), but these CRMs differ dramatically in epigenetic signals and evolutionary history. For example, the independently acting enhancer LCR hypersensitive site 2 (HS2) (107) shows strong constraint on the motifs bound by transcription factors and strong DNase footprints. A second CRM, *HBG1* 3' enhancer (108), is also bound *in vivo* by GATA1 (and other proteins) and is active as an enhancer, but shows almost no constraint over mammalian evolution. Last, a third location, *HBG1*-D (109, 110), shows DNase hypersensitivity but lacks

biological activity in enhancer assays. Rather, binding of this and other CRMs in the locus by BCL11A leads to a reorganization of the chromatin interactions and repression of genes encoding the fetally expressed  $\gamma$ -globins in adult erythroid cells. This CRM is virtually devoid of evidence of mammalian constraint, at least in part because the adult-stage silencing of  $\gamma$ -globin genes is specific to primates. These vignettes illustrate the complementary nature of genetic, evolutionary, and biochemical approaches for understanding disease-relevant genomic elements and also the importance of data integration, as no single assay identifies all functional elements.

### Conclusion

In contrast to evolutionary and genetic evidence, biochemical data offer clues about both the molecular function served by underlying DNA elements and the cell types in which they act, thus providing a launching point to study differentiation and development, cellular circuitry, and human disease (14, 35, 69, 111, 112). The major contribution of ENCODE to date has been high-resolution, highly-reproducible maps of DNA segments with biochemical signatures associated with diverse molecular functions. We believe that this public resource is far more important than any interim estimate of the fraction of the human genome that is functional.

By identifying candidate genomic elements and placing them into classes with shared molecular characteristics, the biochemical maps provide a starting point for testing how these signatures relate to molecular, cellular, and organismal function. The data identify very large numbers of sequence elements of differing sizes and signal strengths. Emerging genome-editing methods (113, 114) should considerably increase the throughput and resolution with which these candidate elements can be evaluated by genetic criteria. Given the limitations of our current understanding of genome function, future work should seek to better define genome elements by integrating all three methods to gain insight into the roles they play in human biology and disease.

- 1 Lander ES, et al.; International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.
- 2 Waterston RH, et al.; Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.
- 3 Lindblad-Toh K, et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478(7370):476–482.

- 4 Ponting CP, Hardison RC (2011) What fraction of the human genome is functional? *Genome Res* 21(11):1769–1776.
- 5 Jones FC, et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484(7392):55–61.
- 6 Grossman SR, et al.; 1000 Genomes Project (2013) Identifying recent adaptations in large-scale genomic data. *Cell* 152(4):703–713.
- 7 Fraser HB (2013) Gene expression drives local adaptation in humans. *Genome Res* 23(7):1089–1096.



- 8 Jeong S, et al. (2008) The evolution of gene regulation underlies a morphological difference between two *Drosophila* sister species. *Cell* 132(5):783–793.
- 9 Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* 134(1):25–36.
- 10 Chan YF, et al. (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. *Science* 327(5963):302–305.
- 11 Kleinjan DA, van Heyningen V (2005) Long-range control of gene expression: Emerging mechanisms and disruption in disease. *Am J Hum Genet* 76(1):8–32.
- 12 Kleinjan DA, Lettice LA (2008) Long-range gene control and genetic disease. *Adv Genet* 61:339–388.
- 13 Hindorf LA, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106(23):9362–9367.
- 14 Maurano MT, et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337(6099):1190–1195.
- 15 Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M (2012) Linking disease associations with regulatory information in the human genome. *Genome Res* 22(9):1748–1759.
- 16 Ward LD, Kellis M (2012) HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 40(Database issue):D930–D934.
- 17 Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: Conservation and turnover. *Mol Biol Evol* 19(7):1114–1121.
- 18 Costas J, Casares F, Vieira J (2003) Turnover of binding sites for transcription factors involved in early *Drosophila* development. *Gene* 310:215–220.
- 19 Moses AM, et al. (2006) Large-scale turnover of functional transcription factor binding sites in *Drosophila*. *PLoS Comput Biol* 2(10):e130.
- 20 Ludwig MZ, Patel NH, Kreitman M (1998) Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: Rules governing conservation and change. *Development* 125(5):949–958.
- 21 Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. *Science* 302(5644):413.
- 22 Ahituv N, et al. (2007) Deletion of ultraconserved elements yields viable mice. *PLoS Biol* 5(9):e234.
- 23 McGaughey DM, et al. (2008) Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at *phox2b*. *Genome Res* 18(2):252–260.
- 24 Vakhrusheva OA, Bazykin GA, Kondrashov AS (2013) Genome-Level Analysis of Selective Constraint without Apparent Sequence Conservation. *Genome Biol Evol* 5(3):532–541.
- 25 Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci USA* 110(14):5294–5300.
- 26 Graur D, 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(3):578–590.
- 27 Eddy SR (2012) The C-value paradox, junk DNA and ENCODE. *Curr Biol* 22(21):R898–R899.
- 28 Eddy SR (2013) The ENCODE project: Missteps overshadowing a success. *Curr Biol* 23(7):R259–R261.
- 29 Mattick JS, et al. (2013) The extent of functionality in the human genome. *HUGO J* 7(1):2.
- 30 Niu DK, Jiang L (2012) Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem Biophys Res Commun* 430(4):1340–1343.
- 31 Germain PL, Ratti E, Boem F (2014) Junk or functional DNA?: ENCODE and the function controversy. *Biology & Philosophy*, 10.1007/s10539-014-9441-3.
- 32 Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33(Database issue):D514–D517.
- 33 Amsterdam A, et al. (1999) A large-scale insertional mutagenesis screen in zebrafish. *Genes Dev* 13(20):2713–2724.
- 34 Berns K, et al. (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* 428(6981):431–437.
- 35 Ernst J, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473(7345):43–49.
- 36 Kheradpour P, et al. (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* 23(5):800–811.
- 37 Visel A, et al. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457(7231):854–858.
- 38 Patwardhan RP, et al. (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* 30(3):265–270.
- 39 Melnikov A, et al. (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 30(3):271–277.
- 40 Pfeiffer BD, et al. (2008) Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proc Natl Acad Sci USA* 105(28):9715–9720.
- 41 MacArthur DG, et al.; 1000 Genomes Project Consortium (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335(6070):823–828.
- 42 Stark A, et al.; Harvard FlyBase curators; Berkeley *Drosophila* Genome Project (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* 450(7167):219–232.
- 43 Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937):241–254.
- 44 Xie X, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* 434(7031):338–345.
- 45 Thomas JW, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424(6950):788–793.
- 46 Cliften P, et al. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301(5629):71–76.
- 47 Boffelli D, et al. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299(5611):1391–1394.
- 48 Siepel A, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15(8):1034–1050.
- 49 Elitski L, et al. (2003) Distinguishing regulatory DNA from neutral sites. *Genome Res* 13(1):64–72.
- 50 Bartel DP (2009) MicroRNAs: Target recognition and regulatory functions. *Cell* 136(2):215–233.
- 51 Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 81:145–166.
- 52 Aravin AA, Hannon GJ, Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* 318(5851):761–764.
- 53 Olovnikov I, Aravin AA, Fejes Toth K (2012) Small RNA in the nucleus: The RNA-chromatin ping-pong. *Curr Opin Genet Dev* 22(2):164–171.
- 54 Grosfeld F, van Assendelft GB, Greaves DR, Kollias G (1987) Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* 51(6):975–985.
- 55 Agarwal S, Rao A (1998) Long-range transcriptional regulation of cytokine gene expression. *Curr Opin Immunol* 10(3):345–352.
- 56 Lakshmanan G, Lieuw KH, Grosfeld F, Engel JD (1998) Partial rescue of GATA-3 by yeast artificial chromosome transgenes. *Dev Biol* 204(2):451–463.
- 57 Noonan JP, McCallion AS (2010) Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet* 11:1–23.
- 58 Nardone J, Lee DU, Ansel KM, Rao A (2004) Bioinformatics for the ‘bench biologist’: How to find regulatory regions in genomic DNA. *Nat Immunol* 5(8):768–774.
- 59 Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* 57:159–197.
- 60 Li CC, Ramirez-Carrozzi VR, Smale ST (2006) Pursuing gene regulation ‘logic’ via RNA interference and chromatin immunoprecipitation. *Nat Immunol* 7(7):692–697.
- 61 Weinmann AS, Farnham PJ (2002) Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods* 26(1):37–47.
- 62 Johnson KD, Bresnick EH (2002) Dissecting long-range transcriptional mechanisms by chromatin immunoprecipitation. *Methods* 26(1):27–36.
- 63 Rada-Iglesias A, et al. (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470(7333):279–283.
- 64 Creighton MP, et al. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 107(50):21931–21936.
- 65 Ozsolak F, et al. (2008) Chromatin structure analyses identify miRNA promoters. *Genes Dev* 22(22):3172–3183.
- 66 Horak EC, Snyder M (2002) Global analysis of gene expression in yeast. *Func Integ Genomics* 2(4–5):171–180.
- 67 ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* 306(5696):636–640.
- 68 Birney E, et al.; ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146):799–816.
- 69 ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
- 70 Cheng Y, et al. (2009) Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* 19(12):2172–2184.
- 71 Henikoff S, Shilatifard A (2011) Histone modification: Cause or cog? *Trends Genet* 27(10):389–396.
- 72 Weiner A, et al. (2012) Systematic dissection of roles for chromatin regulators in a yeast stress response. *PLoS Biol* 10(7):e1001369.
- 73 Thomas CA, Jr. (1971) The genetic organization of chromosomes. *Annu Rev Genet* 5:237–256.
- 74 Gregory TR (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol Rev Camb Philos Soc* 76(1):65–101.
- 75 Keightley PD (2012) Rates and fitness consequences of new mutations in humans. *Genetics* 190(2):295–304.
- 76 Ehret CF, De Haller G (1963) Origin, development and maturation of organelles and organelle systems of the cell surface in *Paramecium*. *J Ultrastruct Res* 23(Suppl 6):1–42.
- 77 Ohno S (1972) So much “junk” DNA in our genome. *Brookhaven Symp Biol* 23:366–370.
- 78 Lynch M (2007) *The Origins of Genome Architecture* (Sinauer Associates, Sunderland, MA).
- 79 Kamal M, Xie X, Lander ES (2006) A large family of ancient repeat elements in the human genome is under strong selection. *Proc Natl Acad Sci USA* 103(8):2740–2745.
- 80 Lowe CB, Bejerano G, Haussler D (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci USA* 104(19):8005–8010.
- 81 Lowe CB, et al. (2011) Three periods of regulatory innovation during vertebrate evolution. *Science* 333(6045):1019–1024.
- 82 McClintock B (1956) Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol* 21:197–216.
- 83 de Souza FS, Franchini LF, Rubinstein M (2013) Exaptation of transposable elements into novel cis-regulatory elements: Is the evidence always strong? *Mol Biol Evol* 30(6):1239–1251.
- 84 Nishihara H, Smit AF, Okada N (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res* 16(7):864–874.
- 85 Clark MB, et al. (2011) The reality of pervasive transcription. *PLoS Biol*, 9(7):e1000625, discussion e1001102.
- 86 Jacquier A (2009) The complex eukaryotic transcriptome: Unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet* 10(12):833–844.
- 87 Lindblad-Toh K, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069):803–819.
- 88 Parker SC, Hansen L, Aaba HO, Tullius TD, Margulies EH (2009) Local DNA topography correlates with functional noncoding regions of the human genome. *Science* 324(5925):389–392.
- 89 Meader S, Ponting CP, Lunter G (2010) Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res* 20(10):1335–1343.
- 90 Ward LD, Kellis M (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337(6102):1675–1678.
- 91 Scally A, Durbin R (2012) Revising the human mutation rate: Implications for understanding human evolution. *Nat Rev Genet* 13(10):745–753.
- 92 Lohmueller KE, et al. (2011) Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet* 7(10):e1002326.
- 93 Ward LD, Kellis M (2013) Response to comment on “Evidence of abundant purifying selection in humans for recently acquired regulatory functions” *Science* 340(6133):682.
- 94 Dimas AS, et al. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325(5945):1246–1250.
- 95 Montgomery SB, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464(7289):773–777.
- 96 Battle A, et al. (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* 24(1):14–24.
- 97 Degner JF, et al. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482(7385):390–394.
- 98 Pickrell JK, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464(7289):768–772.
- 99 Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* 5(3):27.
- 100 Lovén J, et al. (2012) Revisiting global gene expression analysis. *Cell* 151(3):476–482.
- 101 Islam S, et al. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* 21(7):1160–1167.

- 102** Marinov GK, et al. (2014) From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.*
- 103** Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621–628.
- 104** Djebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414):101–108.
- 105** Talbert PB, Henikoff S (2006) Spreading of silent chromatin: Inaction at a distance. *Nat Rev Genet* 7(10):793–803.
- 106** King DC, et al. (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res* 15(8): 1051–1060.
- 107** Tuan DY, Solomon WB, London IM, Lee DP (1989) An erythroid-specific, developmental-stage-independent enhancer far upstream of the human “beta-like globin” genes. *Proc Natl Acad Sci USA* 86(8):2554–2558.
- 108** Bodine DM, Ley TJ (1987) An enhancer element lies 3’ to the human A gamma globin gene. *EMBO J* 6(10):2997–3004.
- 109** Xu J, et al. (2010) Transcriptional silencing of gamma-globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes Dev* 24(8):783–798.
- 110** Sankaran VG, et al. (2011) A functional element necessary for fetal hemoglobin silencing. *N Engl J Med* 365(9):807–814.
- 111** Gerstein MB, et al. (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489(7414):91–100.
- 112** Trynka G, et al. (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* 45(2): 124–130.
- 113** Ran FA, et al. (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* 154(6): 1380–1389.
- 114** Carr PA, Church GM (2009) Genome engineering. *Nat Biotechnol* 27(12):1151–1162.
- 115** Davydov EV, et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLOS Comput Biol* 6(12):e1001025.
- 116** Hesselberth JR, et al. (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* 6(4):283–289.



# Supporting Information

Kellis et al. 10.1073/pnas.1318948111

## SI Methods

**Data Processing and Element Identification.** For all analyses, we used encyclopedia of DNA elements (ENCODE) datasets present at the ENCODE Data coordination center up to an including the June 2012 freeze, unless explicitly stated otherwise.

**Protein coding and noncoding genes.** We used version 16 of the GENCODE annotation (1), which can be downloaded from [www.encodegenes.org/releases/16.html](http://www.encodegenes.org/releases/16.html).

**Transcript segments.** We used RNA-seq–derived contigs from Djebali et al. (2) (January 2011 freeze). Specifically, the \*Contigs.bedRNAElements.gz files were downloaded from <http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq/>.

For the coverage analysis, we used the  $\log_{10}$  of the fragments per kilobase of exon per million reads (FPKM) values in column 7 of the browser extensible data (BED) files as scores for each contig.

**DNase-hypersensitive peaks.** DNase-seq datasets from the University of Washington production center were uniformly processed to identify hypersensitive peaks. The HotSpot peak caller was used to call peaks passing a false discovery rate (FDR) of 1%. Full details of peak calling procedures are provided at <http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeUwDnase>. The peaks can be downloaded from <http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/release6/>. Narrow peak calls corresponding to replicate 1 from each of the cell types were used (these files are named \*Rep1.narrowPeak.gz). Signal enrichment values corresponding to column 7 in the narrowPeak files were used as scores for the peaks.

**Transcription factor ChIP-seq peaks.** Transcription factor (TF) ChIP-seq datasets were processed to identify reproducible peaks of ChIP enrichment relative to corresponding sequenced input-DNA controls. The peak calls can be downloaded from <http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/>.

The read alignment files were filtered to discard multimapping reads and duplicates. The SPP peak caller (3) was used to call peaks on replicate datasets and subsampled pseudoreplicates (obtained by pooling reads from all replicates and randomly subsampling without replacement two pseudoreplicates with half the total number of pooled reads). The irreproducible discovery rate (IDR) framework (IDR threshold of 2%) was used to identify reproducible and rank-consistent peaks by comparing identifications across replicates and pseudoreplicates. Full details are provided at <http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeAwgTfbsUniform>. Step-by-step instructions are provided at <https://sites.google.com/site/anshulkundaje/projects/idr>. The SPP signal scores corresponding to column 7 in the narrowPeak files were used as scores for the peaks.

**Histone modification ChIP-seq regions of enrichment.** Histone modification ChIP-seq datasets were processed to identify regions of ChIP enrichment relative to corresponding sequenced input-DNA controls. Read alignment files were filtered to discard multimapping reads and duplicates.

We used the MACS2 peak caller (version 2.0.10.20130712) to identify regions of enrichment over a wide range of signal strength. Enriched regions were scored on individual replicates, pooled data (reads pooled across replicates), and subsampled pseudoreplicates (obtained by pooling reads from all replicates and randomly subsampling, without replacement, two pseudoreplicates with half the total number of pooled reads).

We used MACS2 to identify three types of regions of enrichment: (i) narrow peaks of contiguous enrichment (narrowPeaks) that pass a Poisson  $P$  value threshold of 0.01; (ii) broader

regions of enrichment (broadPeaks) that pass a Poisson  $P$  value threshold of 0.1 (using MACS2's broad peak mode); and (iii) gapped/chained regions of enrichment (gappedPeaks) defined as broadPeaks that contain at least one strong narrowPeak.

To obtain reliable regions of enrichment, we restricted our analysis to enriched regions identified using pooled data that were also independently identified in both pseudoreplicates. The coverage and conservation analysis only used histone modification datasets from the Broad Institute Production group. We used the gappedPeak representation for the histone marks with relatively compact enrichment patterns. These include H3K4me3, H3K4me2, H3K4me1, H3K9ac, H3K27ac, and H2A.Z.

For the diffused histone marks, H3K36me3, H3K79me2, H3K27me3, H3K9me3, and H3K9me1, we used the broadPeak representation. These peak calls were not optimally thresholded by design to allow for analysis of genomic coverage over a wide range of signal enrichment.

Additional details and step-by-step instructions are provided at <https://sites.google.com/site/anshulkundaje/projects/encodehistonemods>.

The gappedPeak and broadPeak files can be downloaded from [www.broadinstitute.org/~anshul/projects/encode/rawdata/peaks\\_histone/mar2012/broad/combrep\\_and\\_ppr/](http://www.broadinstitute.org/~anshul/projects/encode/rawdata/peaks_histone/mar2012/broad/combrep_and_ppr/).

The narrowPeak files (not used in any of the analyses) can be downloaded from [www.broadinstitute.org/~anshul/projects/encode/rawdata/peaks\\_histone/mar2012/narrow/combrep\\_and\\_ppr/](http://www.broadinstitute.org/~anshul/projects/encode/rawdata/peaks_histone/mar2012/narrow/combrep_and_ppr/).

The negative  $\log_{10}$  of Poisson  $P$  values of enrichment present in column 8 of the peak files was used as scores for the peaks in the coverage analysis.

**DNase-I high-resolution footprints.** High-resolution footprints from deep DNase-seq data (January 2011 freeze) were previously identified in ENCODE Project Consortium 2012. These can be downloaded from [http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration\\_data\\_jan2011/byDataType/footprints/jan2011/encode\\_TF\\_footprints.out](http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/footprints/jan2011/encode_TF_footprints.out).

**Bound TF motifs.** TF binding site motif instances present within ChIP-seq peaks of the corresponding TFs were previously identified in ENCODE Project Consortium 2012 (January 2011 freeze). These can be downloaded from [http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration\\_data\\_jan2011/byDataType/motifs/jan2011/bound\\_motifs.bed](http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/motifs/jan2011/bound_motifs.bed).

**Repeat elements.** Repeat Master annotations were downloaded from the University of California, Santa Cruz (UCSC) genome browser (April 2011). The file that was used can be downloaded from [http://woldlab.caltech.edu/~georgi/ENCODE-Function-2014\\_public/repeatMasker/hg19-repeats](http://woldlab.caltech.edu/~georgi/ENCODE-Function-2014_public/repeatMasker/hg19-repeats).

**Calculation of Genomic Coverage by Different Data Types.** The fraction of the genome covered by each data type was evaluated as follows. For each element (RNA contigs or ChIP-seq/DNase-seq–enriched regions), a scoring metric (FPKM for RNA-seq or a measure of signal strength for other data types as specified in the previous section) was calculated when the elements were originally identified.

Each position in the genome was then assigned the maximum score across all elements that cover it from all experiments in a given group (e.g., the maximum FPKM of all RNA-seq contig covering a given base pair in all Cell PolyA+ RNA-seq experiments).

The fraction of the genome with maximum scores between specific ranges of scores was then calculated to produce the coverage histogram plots shown.

The exact set of files used for each analysis and code is available at [http://woldlab.caltech.edu/~georgi/ENCODE-Function-2014\\_public/](http://woldlab.caltech.edu/~georgi/ENCODE-Function-2014_public/). Detailed step-by-step procedures to reproduce the results are provided at [http://woldlab.caltech.edu/~georgi/ENCODE-Function-2014\\_public/processing\\_documentation.pdf](http://woldlab.caltech.edu/~georgi/ENCODE-Function-2014_public/processing_documentation.pdf).

**Conservation vs. Coverage Analysis. Coverage scores.** The maximum scores, as described in the previous section and available from [www.broadinstitute.org/~lward/Kellis2014\\_DefiningFunctionalDNA/score\\_tracks/](http://www.broadinstitute.org/~lward/Kellis2014_DefiningFunctionalDNA/score_tracks/) (in BED format with scores, split by chromosome), were then used to bin the data tracks into regions by score (Fig. 3). We used the following scores: (i) for DNase peaks,  $\log_{10}$  of signal enrichment scores; (ii) DNase hypersensitivity and transcription factor (TFBS) ChIP-seq peaks,  $\log_{10}$  of signal enrichment scores; (iii) RNA,  $\log_{10}$  of FPKM; and (iv) ChIP-Seq of histone modifications,  $\log_{10}[-\log_{10}(P \text{ value})]$ .

Annotated regions were binned by 0.1 units of these transformed scores.

**Conserved elements definition.** For each of the conservation definitions, two sets of genomic intervals were defined: (i) conserved elements called by the algorithm and (ii) a genomic domain within which that algorithm had provided base-level scores. Elements were intersected with the domain before further analysis. All resulting elements and domains are in [www.broadinstitute.org/~lward/Kellis2014\\_DefiningFunctionalDNA/cons\\_definitions/](http://www.broadinstitute.org/~lward/Kellis2014_DefiningFunctionalDNA/cons_definitions/). Only the autosomal genome was considered for this analysis.

**SiPhy29Mammals.** Constrained elements were obtained from [www.broadinstitute.org/~orzuk/data/elements/hg19\\_29way\\_omega\\_lods\\_elements\\_12mers.chr\\_specific.fdr\\_0.1\\_with\\_scores.txt.gz](http://www.broadinstitute.org/~orzuk/data/elements/hg19_29way_omega_lods_elements_12mers.chr_specific.fdr_0.1_with_scores.txt.gz). The genomic domain was considered as all regions with non-N nucleotides in the hg19 reference genome.

**GERP34Mammals.** Constrained elements were obtained from [http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP\\_elements.tar.gz](http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP_elements.tar.gz) and corresponding scores from [http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP\\_scores.tar.gz](http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP_scores.tar.gz). The genomic domain was defined by all positions with both a non-zero rate score and nonzero rejected substitution (RS) score.

**PhastCons9Primates, PhastCons32PlacentalMammals, PhastCons46-Vertebrates.** Elements were obtained from the UCSC Genome Browser, using the Table Browser function to obtain primate, placental mammals, and vertebrate elements. The genomic domain was obtained using wigFix files from <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/> and delineating only those regions with scores defined in the wigFix file.

**Coverage vs. Conservation Analysis.** To produce Fig. 1, evolutionary evidence was defined using GERP elements described above, protein-coding elements were defined by regions annotated as “CDS” in genes labeled as “protein\_coding” in Gencode v16, and the following ENCODE tracks were used to define levels of activity: H3K27ac, H3K36me3, H3K4me1, H3K4me3, Tfb, LongRnaSeq.all, and UwDnase. To define “high” activity, we used the portion of each of these tracks exceeding the top 10th percentile of signal (for each track), and took their union (across tracks). For “medium” activity, we used the same procedure, taking the union of all elements in the top 50% of each track. The resulting intersections are reported in [www.broadinstitute.org/~lward/Kellis2014\\_DefiningFunctionalDNA/venn/](http://www.broadinstitute.org/~lward/Kellis2014_DefiningFunctionalDNA/venn/).

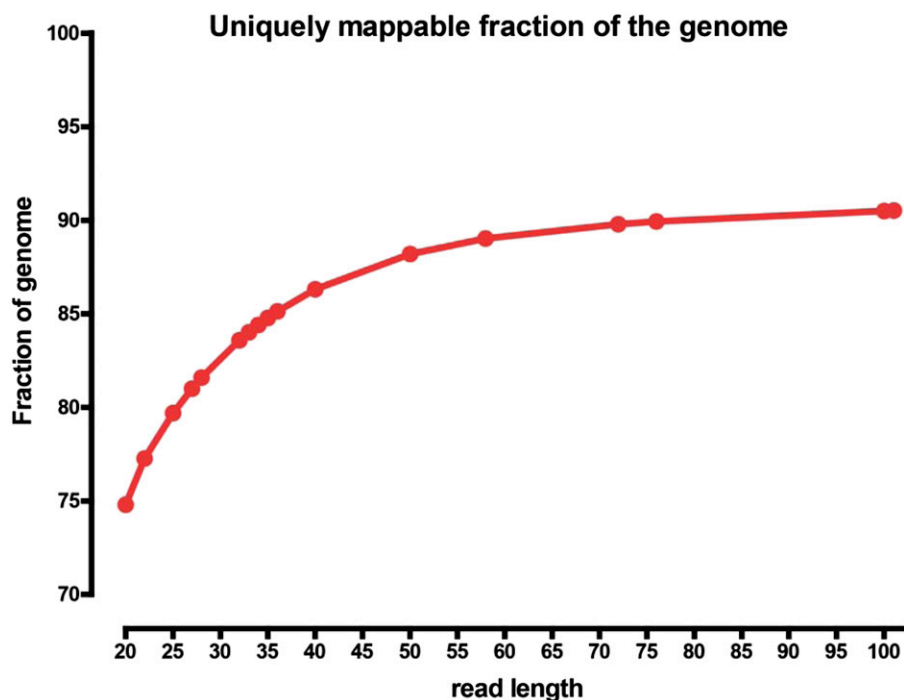
To produce Fig. 3 and Fig. S3.4, for each bin of functional data, the overlap with both conserved elements and the domain for each conservation metric was calculated using BEDTools (4).

The fraction of bases conserved in each bin of functional data was defined as the fraction of bases in conserved elements divided by the fraction of bases in the domain. For plotting clarity, bins containing the top and bottom one percentile of scores for functional data were excluded, as well as bins containing fewer than 10 kb covered by the intersection of the functional elements and the domain.

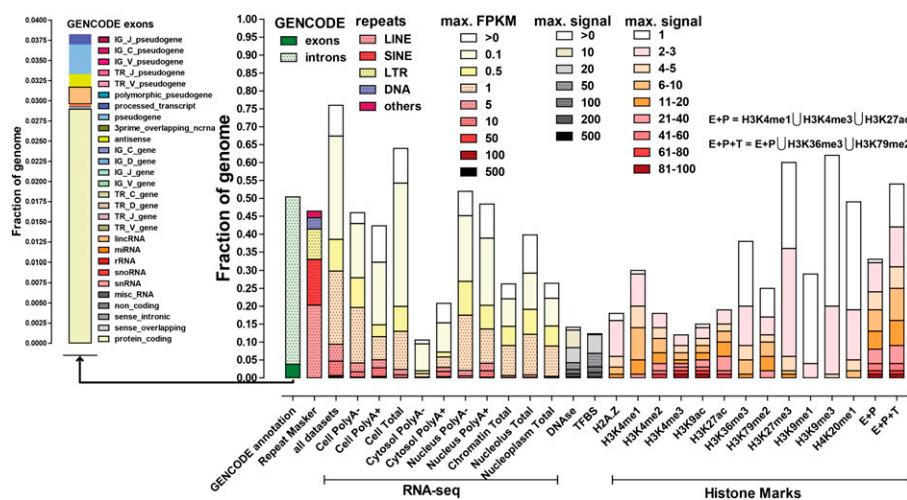
To produce Fig. S3B, genomic evolutionary rate profiling (GERP) RS scores obtained as described above were used, and for each DNase peak (as described above, taking the union across cell types of UW DNase peaks), the coverage score (as described above) and mean basewise GERP RS score were calculated.

1. Harrow J, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* 22(9):1760–1774.
2. Djebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414):101–108.

3. Kharchenko PV, Tolstourov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 26(12):1351–1359.
4. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.



**Fig. S1.** Uniquely mappable fraction of the human genome at various sequencing read lengths.



**Fig. S2.** Summary of coverage of the human genome by encyclopedia of DNA elements (ENCODE) data. Shown is the fraction of the human genome covered by ENCODE elements in at least one cell line per tissue for each assay, as well as genomic coverage by annotated genes and repetitive elements. Version 16 of the GENCODE annotation (1) was used to calculate coverage by annotated genes. Detailed breakdown of the coverage of the genome by the exons of protein coding genes and various noncoding transcripts and pseudogenes is shown separately. The Repeat Masker annotation downloaded from the UCSC Genome Browser was used to calculate coverage of the genome by repetitive elements. For transcripts, coverage was calculated from RNA-seq-derived contigs (2) separated into abundance classes by fragments per kilobase of exon per million reads (FPKM) values. Note that FPKMs are not directly comparable between different subcellular fractions as they reflect relative abundances within a fraction rather than average absolute transcript copy numbers per cell. Depending on the total amount of RNA in a cell, one transcript copy per cell corresponds to between 0.5 and 5 FPKM in PolyA+ whole cell samples according to current estimates (with the upper end of that range corresponding to small cells with little RNA and vice versa). “All RNA” refers to all RNA-seq experiments, including all subcellular fractions. DNase hypersensitivity and transcription factor (TFBS) and histone mark ChIP-seq coverage was calculated similarly but divided according to signal strength. “Motifs+footprints” refers to the union of occupied sequence recognition motifs for transcription factors as determined by ChIP-seq and as measured by digital genomic footprinting, with the purple portion of the bar representing the genomic space covered by bound motifs in ChIP-seq. Signal strength for ChIP-seq data for histone marks was determined based on the *P* value of each enriched region (the  $-\log_{10}$  of the *P* value is shown), using peak calling procedures tailored to the broadness of occupancy of each modification (*SI Methods*). “E+P” and “E+P+T” refer to the union of coverage by histone marks associated with enhancers and promoters (E+P) or enhancers, promoters, and transcriptional activity (E+P+T).

1. Harrow J, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* 22(9):1760–1774.
2. Diebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414):101–108.

