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.gencodegenes.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/.

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/.

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/gencode_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.

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.

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/. Detailed step-by-step procedures to reproduce the results are provided at 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/ (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/. 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. 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 and corresponding scores from http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP_scores.tar.gz. The genomic domain was defined by all positions with both a nonzero rate score and nonzero rejected substitution (RS) score.

PhastCons9Primates, PhastCons32PlacentalMammals, PhastCons46Vertebrates. 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, Tfbs, 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/.

To produce Fig. 3 and Fig. S3*A*, 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. S3*B*, 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, Tolstorukov 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.

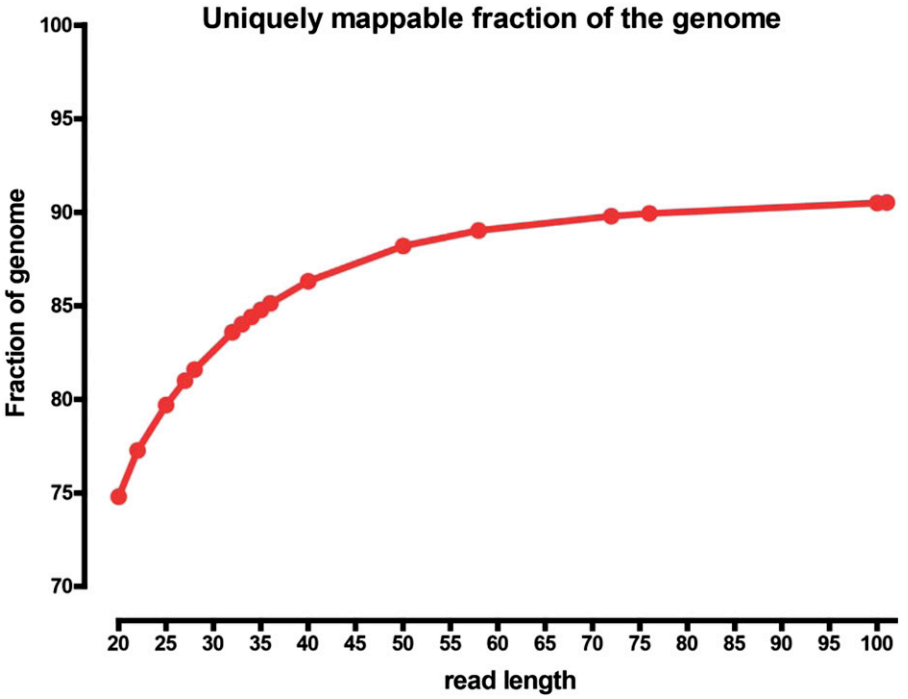

Fig. S1. Uniquely mappable fraction of the human genome at various sequencing read lengths.

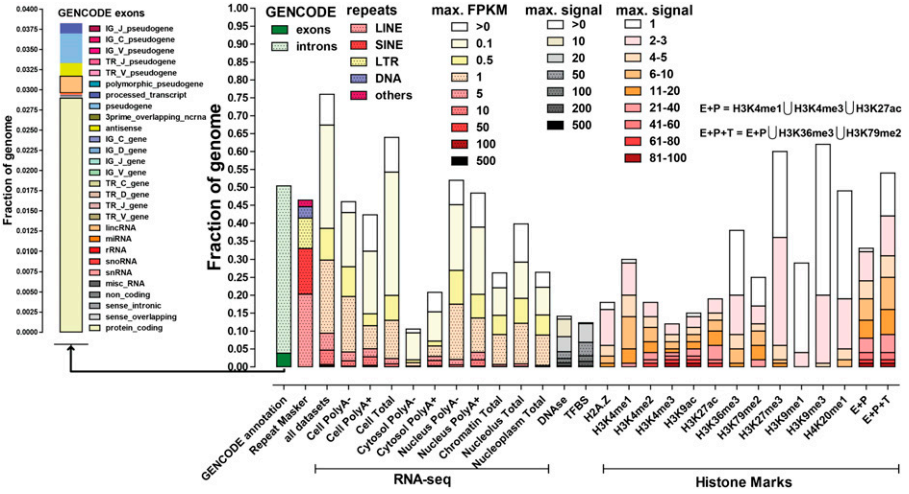

Fig. S2. Summary of coverage of the human genome by encyclopedia of DNA elements (ENCODE) data. Shown is the fraction of the human genome covered by ENCODE elements in at least one cell line per tissue for each assay, as well as genomic coverage by annotated genes and repetitive elements. Version 16 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 –log10 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. Djebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414):101–108.

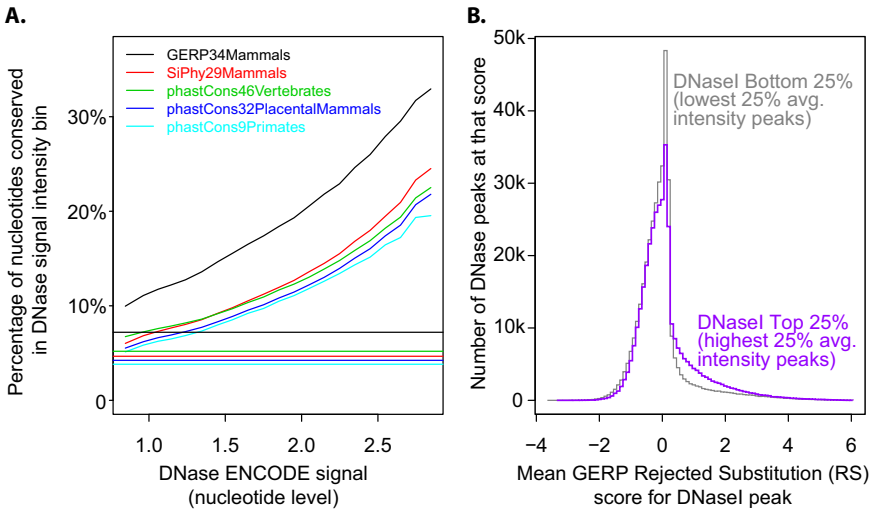

Fig. S3. Relationship between DNase signal intensity score and conservation by five different metrics. (*A*) Nucleotides annotated by DNase I hypersensitive regions were binned by log10 signal enrichment score, and the fraction conserved by five conservation metrics was plotted (1–3). (*B*) DNase I peaks were sorted by their signal enrichment score, and the mean basewise GERP RS (rejected substitution) score was calculated for each peak. The distribution of conservation scores for the top and bottom quartile of peaks by signal enrichment score were plotted in purple and gray, respectively.

1. Lindblad-Toh K, et al.; Broad Institute Sequencing Platform and Whole Genome Assembly Team; Baylor College of Medicine Human Genome Sequencing Center Sequencing Team; Genome Institute at Washington University (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478(7370):476–482.
2. 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.
3. Siepel A, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15(8):1034–1050.