Identification of higher-order functional domains in the human ENCODE regions

Robert E. Thurman, Nathan Day, William S. Noble and John A. Stamatoyannopoulos

Genome Res. 2007 17: 917-927
Access the most recent version at doi:10.1101/gr.6081407

Identification of higher-order functional domains in the human ENCODE regions

Robert E. Thurman,[1,2] Nathan Day,[3] William S. Noble,[2,3] and John A. Stamatoyannopoulos[2,4]

[1]*Division of Medical Genetics, University of Washington, Seattle, Washington 98195, USA;* [2]*Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA;* [3]*Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195, USA*

It has long been posited that human and other large genomes are organized into higher-order (i.e., greater than gene-sized) functional domains. We hypothesized that diverse experimental data types generated by The ENCODE Project Consortium could be combined to delineate active and quiescent or repressed functional domains and thereby illuminate the higher-order functional architecture of the genome. To address this, we coupled wavelet analysis with hidden Markov models for unbiased discovery of "domain-level" behavior in high-resolution functional genomic data, including activating and repressive histone modifications, RNA output, and DNA replication timing. We find that higher-order patterns in these data types are largely concordant and may be analyzed collectively in the context of HeLa cells to delineate 53 active and 62 repressed functional domains within the ENCODE regions. Active domains comprise ~44% of the ENCODE regions but contain ~75%–80% of annotated genes, transcripts, and CpG islands. Repressed domains are enriched in certain classes of repetitive elements and, surprisingly, in evolutionarily conserved nonexonic sequences. The functional domain structure of the ENCODE regions appears to be largely stable across different cell types. Taken together, our results suggest that higher-order functional domains represent a fundamental organizing principle of human genome architecture.

[Supplemental material is available online at www.genome.org.]

The concept that the chromosomes of higher eukaryotes are partitioned into discrete functional territories with distinct physical properties predates the genome era by over four decades (Cooper 1959). Such territories have classically been associated with the cytogenetic phenomena of heterochromatin and euchromatin, which are generally believed to represent repressed and active genomic regions, respectively (Lamond and Earnshaw 1998). However, the potential for translating cytogenetic observations to modern genome sequence annotations is severely limited by the low, multimegabase resolution of the former, the lack of an automated approach for such an endeavor, and by the finding that euchromatic and heterochromatic territories may themselves contain subregions with unique functional properties governing the regulation of their constituent genes (Gilbert and Bickmore 2006).

A variety of genetic, biochemical, and cytological data now support the existence of multigene functional domains spanning up to hundreds of kilobases within human and other vertebrate genomes (Dillon 2003). Control of gene expression by long-range elements may be confined within discrete regulatory domains demarcated by insulator elements that function, at least in part, by physically segregating chromatin loops at the level of the nuclear matrix (Felsenfeld et al. 2004). For example, studies of the human and chicken beta-globin loci indicate that in erythroid cells, linked developmental and differentiation stage-specific genes lie within a domain of open chromatin, share common *cis*-regulatory sequences, have similar patterns of histone modification, and replicate early during S phase (Felsenfeld 1996; Felsenfeld et al. 2004). In nonerythroid cells, these genes are repressed, lie within compacted chromatin, and replicate late during S phase. Active and repressed states are also associated with differential localization of the globin gene region of Chr11 within the cell nucleus (Osborne et al. 2004). As such, the ~100-kb regions surrounding the beta-like globin genes on Chr.11 and the alpha-like globin genes on Chr.16 are commonly referred to as the beta- and alpha-globin domains (Felsenfeld et al. 2004; Dean 2006). Analogously, the Th2 cytokine genes share *cis*-regulatory sequences that coordinate the expression of genes within a ~250-kb active chromatin domain (Lee et al. 2005); similar findings have been described in the context of numerous other loci (Li et al. 2002). Recently, many developmentally regulated genes were found to reside in large genomic territories highlighted by high levels of repressive histone modifications (Bernstein et al. 2005; Hochedlinger and Jaenisch 2006). In summary, the term "domain" is frequently used to denote a genomic territory that may encompass multiple colocated genes sharing common transcriptional regulatory properties such as tissue specificity of expression. Although there is no universally agreed upon definition of a genomic domain, the literature is nearly unanimous in considering such regions to be of at least gene size, and preferably larger.

A major outstanding question is, therefore, to what degree do large-scale functional genomic studies support the existence of such multigene domains as a general phenomenon of human genome organization. Systematic delineation of higher-order functional domains within the human genome is of great interest for several reasons. First, a "domain map" may highlight groups of genes that occupy a common *cis*-regulatory environment. Second, transitions between active and repressed domains

[4]Corresponding author.
E-mail jstam@u.washington.edu; fax (206) 267-1094.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.6081407. Freely available online through the *Genome Research* Open Access option.

may mark the position of *cis*-regulatory elements such as insulators and boundary elements important for domain control. Third, since active and repressed domains may occupy distinct nuclear neighborhoods, knowledge of domain structure is expected to provide a missing link between the genomic sequence and nuclear architecture.

Although critical for our understanding of genome structure and function, the identification and study of higher-order functional domains has heretofore been hindered by the lack of large-scale data sets that provide a continuous picture of multiple functional genomic parameters across sufficiently large stretches of human genome sequence. Under the ENCODE Project Consortium (2004), a variety of functional experimental data types have now been assayed across a selected 1% of the human genome. These include activating and repressive histone modifications, RNA transcription levels, and DNA replication timing (The ENCODE Project Consortium 2004). Many of these data types have been collected from multiple cell types. However, they are all present in the context of an ENCODE Consortium-designated common cell line, HeLaS3. The ENCODE experimental data sets are unprecedented not only in terms of scale, but also by the fact that multiple distinct functional features have been measured simultaneously over the same genomic regions as a continuous function of genomic position. These data provide the first opportunity to address systematically the delineation of functional domains in human chromosomes based on multiple independently ascertained functional features.

Here we develop a computational approach for discrimination of domain-level features in ENCODE data types. We find that higher-order patterns in histone modifications, transcription, and DNA replication timing are generally concordant and can be used to define discrete active and repressed functional domains ranging from ~20 kb to 1 Mb in size. Active and repressed domains differ markedly from one another with respect to annotated genomic features including gene content, CpG islands, the spectrum of repetitive elements, and the density of conserved nonexonic sequences. The overall active/repressed domain structure appears to be largely stable across two ENCODE common (and unrelated) tissue types, suggesting that it represents a fundamental organizing principle of human genome structure.

Results

Data types and scale

We focused our analyses on four ENCODE experimental data types, including bulk RNA/transcriptional output, DNA replication time, histone acetylation (H3), and histone H3k27 trimethylation. Prior data suggest that the average values of all of these data types vary over regions that are considerably larger than the average gene (Bernstein et al. 2005; Azuara et al. 2006; Hochedlinger and Jaenisch 2006). One of the key challenges in comparing one or more experimental data type sampled in a continuous, high-resolution fashion across the genome is to account for scale of effect, i.e., at what genomic resolution (e.g., 25, 50, 100 kb, etc.) does a particular behavior become manifest for a given ENCODE data type?

To address this, we aimed to normalize the diverse data types to a common scale using wavelet analysis (Percival and Walden 2000), which provides a framework for multiscale analysis. By decomposing a given data type into increasingly coarse scales, wavelet analysis allows broader and broader trends in the data to reveal themselves. Wavelets are distinguished from Fourier analysis (which also provides a decomposition of a given signal in terms of multiple scales/frequencies) by the ability to localize signal behavior in both frequency and "time" (in this context, genomic position). As such, wavelets are far better suited to detect domain-type behaviors that span discrete genomic intervals. Wavelets have been used for the analysis of genomic data to uncover local periodic patterns in DNA-bending profiles (Audit et al. 2004) and gene-expression data (Allen et al. 2003; Jeong et al. 2004) to predict protein structures (Lio and Vannucci 2000) and to correlate a variety of genomic data on multiple scales in microbial genomes (Allen et al. 2006), and a variety of other applications (Lio 2003).

Overview of "wavelet segmentation" approach

Our approach is summarized in Figure 1 and consists of two key steps. First is the representation of raw continuous genomic data at a high scale via wavelet smoothing. The second is the application of hidden Markov models (HMMs) to partition, in an unbiased fashion, the ENCODE regions at any given scale into two states based on the signal strength of the wavelet smoothed data as a function of genomic position. Following this partitioning, we assign labels of "active" and "repressed" to the states based on their agreement with received notions of genomic functionality. For example, active chromatin regions are believed to be characterized by high levels of transcription, the presence of activating histone modifications, and replication during the early S phase (Cremer et al. 2000; Dillon 2006). It should be emphasized that although this latter assignment follows a loose heuristic, the HMM segmentation is executed without supervision. Segmentation into more than two states is easily accomplished under this framework, and two states were chosen (1) to match the preconceived notions of active and repressed domains, and (2) in order to achieve the simplest possible model for study. All results presented here, including segmentation data, are publicly available at http://noble.gs.washington.edu/proj/domain.

Transformation and segmentation of ENCODE functional data

We began by computing for each ENCODE data type the wavelet smooth over a range of scales. We did not perform smoothing on the TR50 data for this analysis because it is already highly smoothed by construction (Karnani et al. 2007). Figure 1B shows an example wavelet transformation of the H3ac signal across a 1.7-Mb region (ENm005; Chr21, 32, 668, 497-34, 363, 496) into a continuous range of scales, from 2 to 200 kb. We then segmented the data from each scale using a hidden Markov model. An HMM is a statistical model that assumes observable data are generated from a predetermined number of hidden background probability distributions called "states" (see Methods). For this application we employed a simple two-state model. The crucial output from the model is the state assignment, which gives a state label ("0" or "1") to each observation. For segmentations based on a single data type, these states typically correspond to relatively low and high values, respectively. An exemplary segmentation is shown in Figure 1 for the 64-kb scale in Figure 1B (dashed line). Importantly, however, the inputs to the HMM can be multivariate, which enables the computation of two-state segmentations simultaneously based on multiple data types.

Selection of scale for domain-level analyses

For the purposes of our analyses, we wanted a single common scale at which all data types could be analyzed. As wavelet scale

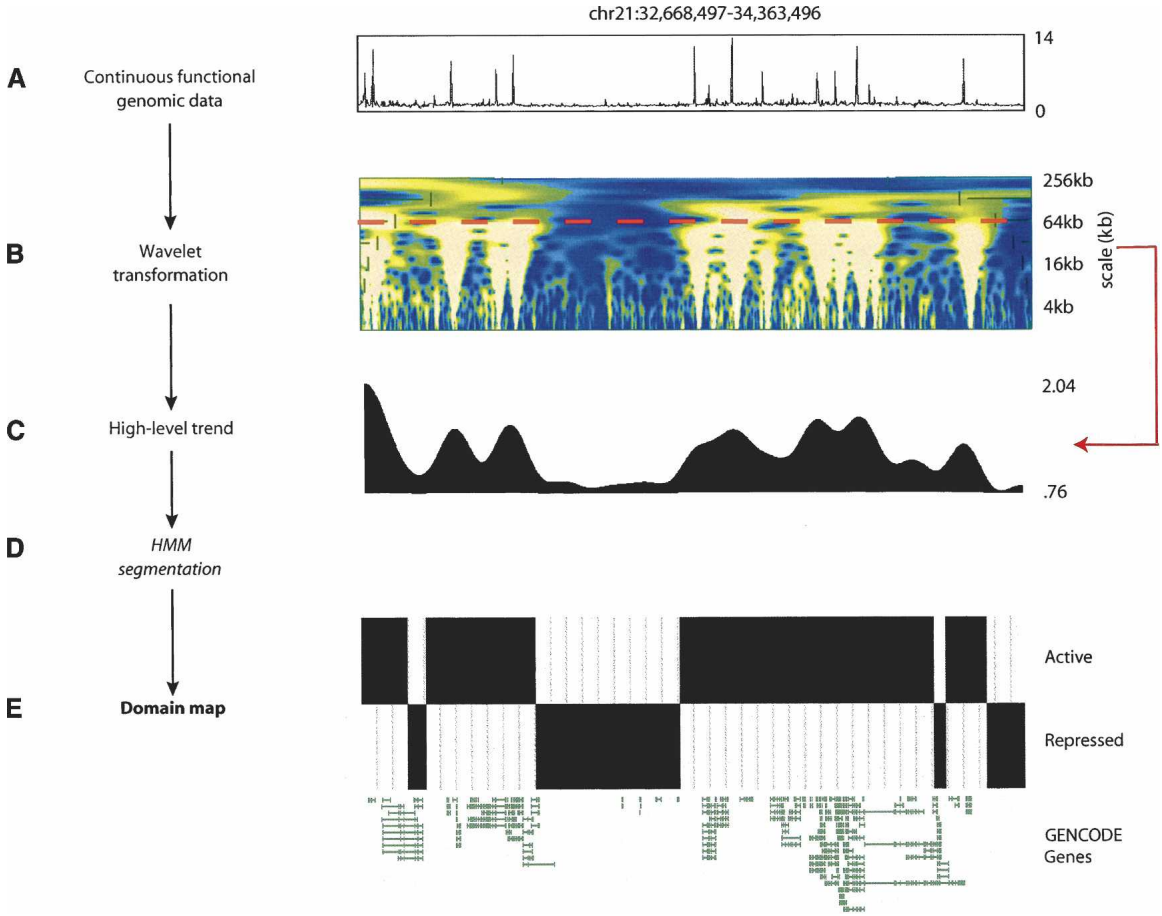

Figure 1. Wavelet segmentation approach for functional domain mapping. (*A*) Exemplary continuous functional data type (H3 acetylation) for ENCODE region ENm005. (*B*) Continuous wavelet transform heatmap ("scalogram") of H3 acetylation data. In the heatmap, the horizontal axis represents genomic position, while the vertical axis represents wavelet scale. Each color in the scalogram represents the magnitude of the wavelet coefficient at that genomic position and scale, ranging from blue (small magnitude) to white (large magnitude). Larger magnitude wavelet coefficients imply a strong trend in the original data at that position and scale. The 64-kb scale is marked with a dashed red line. (*C*) Wavelet smoothed data at the 64-kb scale obtained using MODWT. Horizontal axis: genomic position. Vertical axis: wavelet coefficient at that position at the 64-kb scale. (*D*) Results from two HMM state segmentation of data from *C*, based on fitting HMM to H3ac data over all ENCODE regions. The *top* row indicates state 1 regions in black, while the *bottom* row indicates state 0 regions in black. The high state (state 1) is taken to represent active domains based on the assumption that H3ac is an activating mark. (*E*) GENCODE gene annotations for ENm005. Note the correspondence between state 1/active and GENCODE gene and density.

increases, the individual segment lengths resulting from an HMM segmentation will generally increase, with a concomitant decrease in the number of segments. In accordance with the concept of a domain, we desired both that the median segment length be larger than the average gene size (~25) (Lander et al. 2001) and that the minimum segment length exceed a value representing the lower possible size limit on domains. We selected this lower bound to be 10 kb, reasoning (1) that while smaller than the average gene size, this interval was larger than ~50% of human genes, and (2) that regions smaller than 10 kb may begin to converge on the signals resulting from prominent individual regulatory elements (e.g., some promoters and other functional elements, where histone modification signals may extend for ~2–5 kb). To visualize this dependence, we computed wavelet transformations into a discrete range of scales and then performed HMM segmentations for each scale (Supplemental Fig. S1). This revealed that to achieve a minimum segment length of 10 kb required that we analyze a minimum wavelet scale of ~64 kb for most data types. That scale, in turn, yielded median segment lengths of 80–350 kb (see below).

We also used these experiments to test the sensitivity of domain boundaries to the choice of wavelet scale, and found them to be robust. We performed four-track segmentations (TR50, RNA, H3ac, and H3K27me3) on wavelet-smoothed data at the 32-, 64-, and 128-kb scales. Generally, one expects fewer, larger segments as the scale increases. So, in comparisons of the 64-kb results with the other two, we looked at each segment boundary for the coarser scale, and located the closest segment boundary to it from the finer scale segmentation. For the 64- versus 32-kb comparison, 45 of the 115 64-kb segment boundaries coincided exactly with boundaries from the 32-kb segmentation. The median distance between boundary elements was 5 kb (five observations at the common scale of 1 kb for all datasets), and 80/115 boundaries fell within 10 kb. The maximum distance was 110 kb, followed by 84 kb. Those two boundaries corresponded to one segment in a region that had more segments in the 64-kb segmentation than in the 32-kb segmentation, breaking the general rule. For the 64-kb versus 128-kb comparison, exactly half of the 86 boundaries in the 128-kb segmentation coincided precisely with boundaries in the 64-kb

segmentation, and more than two-thirds (60/85) fell within 10 kb.

Domains based on segmentation of individual ENCODE data types

We performed wavelet transformations into a 64-kb scale for the following experimental data types measured in the ENCODE common cell type HelaS3: activating histone modifications (H3k4me1, H3k4me2, H3k4me3, H3ac, H4ac) (Koch et al. 2007), repressive histone mark H3k27me3 (The ENCODE Project Consortium 2007), and bulk RNA output (The ENCODE Project Consortium 2007). Replication timing has been measured across ENCODE regions (Karnani et al. 2007), and is expressed as a continuous function of genomic position as the TR50 curve, which represents the time from the start of the S phase (in hours), at which 50% of a given genomic interval has replicated. The expected relationships between these data are that active chromatin should exhibit higher levels of histone acetylation and H3K4 methylation, higher average levels of bulk RNA transcription, and earlier replication time. Conversely, repressed chromatin is expected to exhibit the inverse, with late replication time, loss of activating histone modifications, low RNA output, and increased levels of histone modifications such as H3K27Me3 associated with gene silencing.

We computed single-variable segmentations based on an approximate wavelet scale of ~64 kb for each of the data types, except for TR50, which is already highly smoothed. The resulting number of segments falling into each state for each data type are shown in Table 1. We then measured the degree of concordance between each pair of data types by counting the percentage of bases for which the state assignments agreed or disagreed (Table 2). H3ac and H4ac displayed the highest degree of concordance with the other activating histone modifications and the other data types. This accords well with the proposal that histone acetylation is an important marker of active and repressed chromatin domains (Struhl 1998).

Identification of domains based on multiple experimental data types

Next, we asked whether we could exploit the rich ENCODE experimental resource to segment the ENCODE regions into active and repressed domains using multiple data types simultaneously. We computed a simultaneous two-state HMM segmentation (see Methods) on the following four experimental data types: H3ac, H3k27me3, RNA, and TR50. Only one representative activating histone mark (H3ac) was chosen in order to consider the simplest possible model that still encompassed the different functional types of ENCODE data. Exemplary results are shown in Figure 2. As evidence for the robustness of this multivariate segmentation in summarizing the information contained in the individual data types, this segmentation was found to be highly concordant with each of the constituent single-variable segmentations: 89% concordant with the H3ac segmentation, 80% with RNA, 62% with H3K27me3, and 76% with TR50. The HMM delineated a total of 115 domains, 62 in state 0 and 53 in state 1, comprising 56% and 44% of the ENCODE regions, respectively (Table 3). The median size of domains in state 1 was considerably smaller than state 0 (131 versus 189 kb). The fitted HMM probability distributions for each state indicated that domains in state 1 regularly displayed higher levels of H3ac and RNA output, combined with lower levels of H3k27me3 and earlier replication time (lower TR50); domains in state 0 generally had the inverse. It should be emphasized that the HMM recognized the aforementioned associations between functional genomic data types in an unsupervised fashion.

In accordance with received concepts, we considered state 1 to represent active and state 0 to represent repressed domains. An exemplary repressed domain is shown in Figure 2. ENCODE region ENm005 contains a subregion encompassing the *OLIG1* and *OLIG2* genes. These genes encode developmental stage-specific transcription factors necessary for formation of the cerebellum (Ligon et al. 2006), and are embedded within a ~400-kb region that contains a high density of evolutionarily conserved noncoding sequences; indeed, this entire domain displays conserved synteny with the chicken *OLIG1* and *OLIG2* locus. This region was readily delineated by the HMM and assigned to state 0. The *OLIG1/2* domain replicates late during the S phase, has low levels of H3 acetylation and transcriptional activity, and displays high levels of repressive histone modification H3k27me3. High levels of H3k27me3 are found across numerous other developmental-specific gene loci (Bernstein et al. 2005; Hochedlinger and Jaenisch 2006), the complete silencing of which is critical, since overexpression in mature tissues may produce a malignant phenotype (Nichols and Nimer 1992).

Table 1. Summary of segmentations based on individual data types

	Number of segments		Total size of state (Mb)	
	Active	Repressed	Active	Repressed
H3ac	57	69	12.5	17.3
H4ac	55	61	13.6	16.3
H3K4me1	75	88	12.8	17.0
H3K4me2	77	95	10.9	18.9
H3K4me3	82	101	9.4	20.5
H3K27me3	59	50	18.9	11.0
RNA	86	95	12.3	17.1
TR50	48	46	16.7	12.2

Shown for each data type are the number of segments in each state and the cumulative size of each state (Mb) derived by summing the lengths of the individual segments. Note that the total size (active + repressed) of both states varies slightly between individual data types because of missing data in some data sets versus others.

High-confidence active and repressed regions from individual segmentations

We next asked to what degree the domain map created using simultaneous segmentation of four data types recapitulated or extended the picture resulting from segmentations produced using each of the constituent data types individually. To address this, we identified regions that were designated active or repressed in all individual data-type segmentations. Using this approach, we defined 51 "high-confidence" active domains comprising 6.1 Mb, and 43 high-confidence repressed domains comprising 4.2 Mb. Of these 10.3 Mb, or more than one-third of the ENCODE regions, all of the repressed regions and all but 1.5 kb of the active regions were concordant with the simultaneous four-data-type segmentation. This result provides a level of confidence that the multiple-data-type segmentation is effective at capturing consensus features of the constituent data types.

Table 2. Concordance between segmentations of individual data types

	H3ac	H4ac	H3K4me1	H3K4me2	H3K4me3	H3K27me3	RNA
H4ac	85%						
H3K4me1	73%	80%					
H3K4me2	86%	80%	77%				
H3K4me3	80%	73%	68%	87%			
H3K27me3	59%	63%	62%	55%	52%		
RNA	75%	74%	67%	72%	68%	49%	
TR50	70%	75%	74%	69%	60%	70%	61%

Data in each cell represents the concordance (%) between segmentations of the two corresponding (row, column) data types. Concordance is measured by counting the percent of genomic bases for which the state assignments agree in both segmentations.

Distribution of genes and transcripts between active and repressed domains

We next considered to what degree annotated genomic features were enriched (or depleted) in either the active or the repressed state. To address this issue, we computed the overlap of each state with a variety of annotated features including the transcriptional start and termination sites of known genes and annotated mRNA and spliced EST transcripts, CpG islands, various classes of transposable (repetitive) elements, and evolutionarily conserved noncoding sequences (Figure 3; Table 4). For each feature, we computed enrichment/depletion percentages relative to random expectation and derived an empirical *P* value for the significance of the observation via permutation testing (see Methods). We also computed adjusted *P*-values using the Benjamini–Hochberg correction (Benjamini and Hochberg 1995) for controlling the false discovery rate (FDR) in a multiple testing scenario (see Methods).

We observed significant enrichment in state 1 for many elements expected to be associated with active chromatin territories, including transcriptional start and stop sites for genes, mRNAs, spliced ESTs, and CpG islands. The marked disparity in gene and transcript content between active and repressed domains prompted us to ask whether particular classes of genes were enriched in active versus repressed domains, and vice versa. To address this question, we examined available gene ontology (GO) annotations (Ashburner et al. 2000) for RefSeq genes and mapped these to GENCODE (Harrow et al. 2006) genes. This analysis revealed marked overrepresentation ($P < 1.44 \times 10^{-13}$) within repressed domains of genes involved in signal transduction, including a significant complement of olfactory G-protein-coupled receptors (28/61 genes; Supplemental Tables S1, S1b). This result accords well with the general observation that many such genes are only active within a limited range of tissues. To confirm that the resulting enrichment for genes in active segments was not driven primarily by the inclusion of RNA transcription in the model, we compared the results with a three-track segmentation comprising H3Ac, H3K27Me3, and TR50 only. This resulted in a domain map 78% concordant with the four-state RNA-containing model, and exhibited significant enrichment of transcriptional start sites (known genes, 40%; mRNAs, 45%; and spliced ESTs, 67%).

Distribution of transposable elements between active and repressed domains

Viewed in aggregate, transposable (repetitive) elements identified by RepeatMasker (Jurka et al. 2005) exhibit a balanced distribution between active and repressed domains. However, we found marked disparities for individual classes of repetitive elements. L1 LINES, DNA repeats, and LTR elements are all enriched in state 0 regions (repressed). However, *Alu* SINES are enriched in state 1 regions, and to a significantly higher degree than the other elements. *Alu*s are known to be enriched in gene-rich regions (Batzer and Deininger 2002), and are believed to insert preferentially into open or active chromatin territories.

Domain architecture and noncoding evolutionary conservation

Next, we considered to what degree the four-data-type domain map captured the distribution of noncoding evolutionary conservation, and also to what degree the domain map could be refined by incorporating evolutionary conservation as an independent data type.

Evolutionarily conserved noncoding sequences have been proposed to mark the locations of functional elements important for gene and genomic regulation (Hardison 2000). Although *cis*-regulatory elements such as enhancers can operate at considerable distances (Li et al. 2002; Nobrega et al. 2003; Spitz et al. 2003), there is a clear deterioration of efficacy with increasing distance from target promoters (Harju et al. 2005). Therefore, if conserved noncoding sequences harbor *cis*-regulatory elements important for gene regulation, the a priori expectation is that they should exhibit increased density in active regions of the genome. To explore this hypothesis, we examined the distribution of nonexonic (i.e., noncoding and non-UTR) conserved sequences identified by the ENCODE Multi-Species Alignments Analysis Group (MSA) (The ENCODE Project Consortium 2007). Contrary to expectation, we found active domains to be depleted in conserved nonexonic sequences, with 34% of all CNEs in active domains versus 66% in repressed; see Table 4. This corresponded to an active domain depletion of 18% over random expectation.

We then asked to what degree incorporation of noncoding evolutionary conservation as an independent data type would impact the domain map. We transformed the MSA CNE data into a continuous density function by calculating the fraction of CNE bases in a 3-kb sliding window (where 3 kb represents the 99.9th percentile of the length of elements in the MSA CNE data set). We then performed wavelet analysis on this continuous noncoding conservation density function as described above and computed a new segmentation based on the individual CNE density data type, as well as a simultaneous five-data-type segmentation (H3ac, RNA, H3k27me3, TR50, and noncoding conservation).

The individual CNE segmentation was poorly concordant with each of the other four individual-data-type segmentations, and only 62% concordant with the five-data-type segmentation. The domain map based on five data types, however, was highly concordant (98%) with the four-data-type segmentation described above. The HMM delineated 53 active and 61 repressed

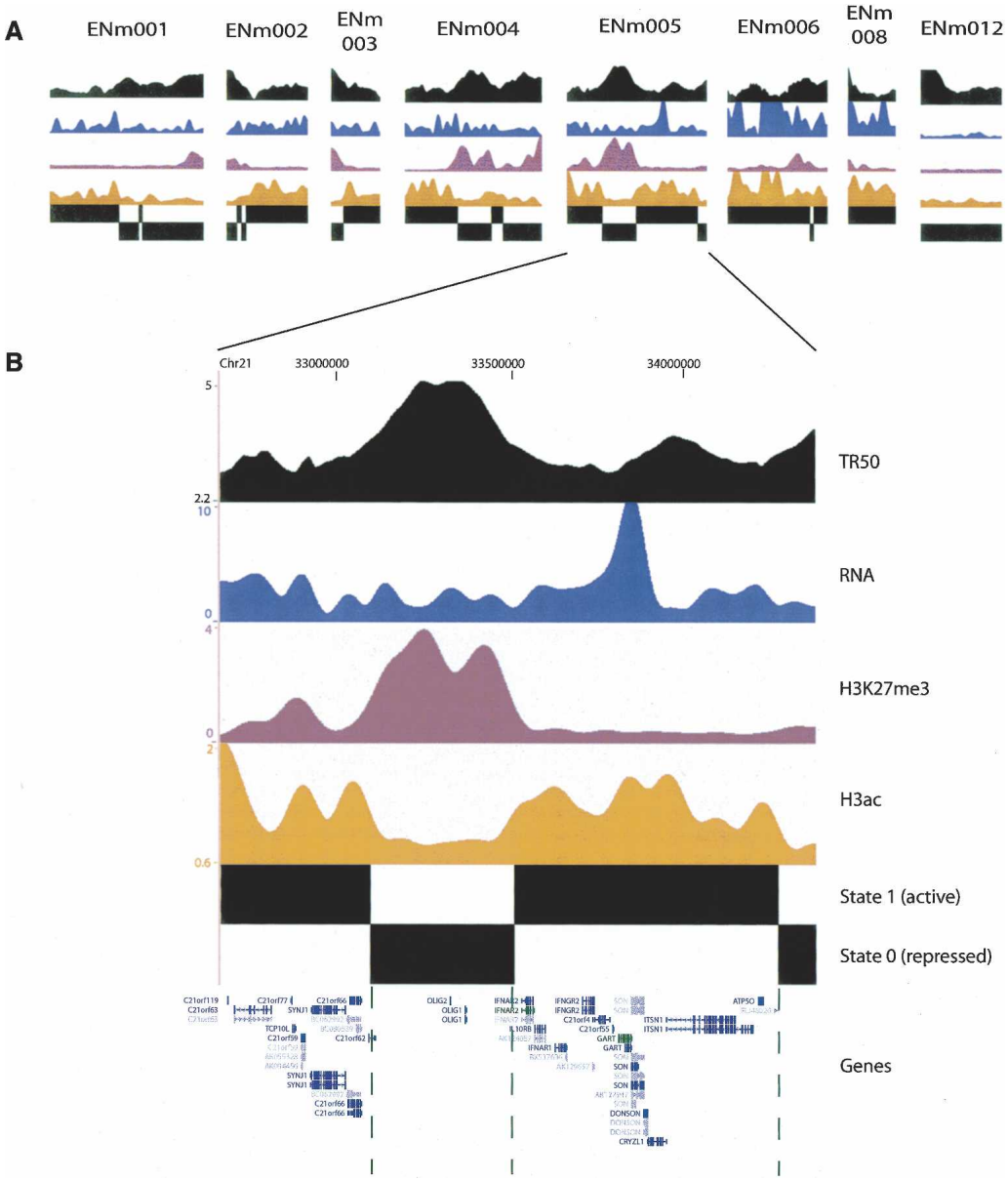

Figure 2. Simultaneous segmentation of four ENCODE functional data types. (*A*) Exemplary results from eight ENCODE regions ENm001 (1.8 Mb), ENm002 (1 Mb), ENm003 (600 kb), ENm004 (1.7 Mb), ENm005 (1.6 Mkb), ENm006 (1 Mb), ENm008 (1 Mb), and ENm012 (1.2 Mb). For each ENCODE region subpanel, wavelet smoothed data are displayed as tracks ordered *top*-to-*bottom* as follows: TR50 (black), RNA (blue), H3K27me3 (purple), and H3ac (orange). State assignments (domains) resulting from simultaneous HMM segmentation are shown at *bottom* as black rectangles; see Fig. 1 for additional description. (*B*) Close-up of ENCODE region ENm005, with bracketed intervals indicating exemplary domains in states 0 and 1. State 1 generally corresponds to higher levels of RNA and H3ac and lower levels of TR50 and H3K27me3, and is therefore assigned the active label, while state 0 is correspondingly assigned repressed. The latter contains the oligodendrocyte-specific *OLIG1* and *OLIG2* genes, which are repressed in the tissues studied under ENCODE.

domains, totaling 44% (12.9 Mb) and 56% (16.3 Mb) of the ENCODE regions, respectively.

Next, we examined the distribution of genes and other annotated sequence features in each state (Supplemental Fig. S2; Supplemental Table S2). This analysis revealed that incorporation of noncoding conservation had almost no effect on the distribution of the annotated features we considered between repressed and active domains. For example, the fraction of GENCODE (Harrow et al. 2006) annotated transcription start sites in active domains was 78.3% for the four-data-type segmentation and 78.5% for the five-data-type segmentation. The enrichment of CNEs themselves did change, although not dramatically, with the active domains in the five-data-type segmentation depleted in CNEs by 23% over random expectation (uncorrected *P*-value of 0.02).

Evidence for stability of domain architecture across unrelated tissue types

The distribution of genes and transcripts between active and repressed domains suggested that these domain definitions should

Table 3. Summary of segmentations based on four data types (H3ac, H3K27me3, RNA, TR50) segmentation results

	Active	Repressed
Number of segments	53	62
Total size (Mb)	12.9	16.3
Average segment length (kb)	244	263
Median segment length (kb)	131	189
Minimum segment length (kb)	28	17

Shown are the number and size of segments (as in Table 1), together with the average, median, and minimal segment lengths.

be persistent across tissue types. In the domain maps based on either four- or five-data types, ~78% of transcriptional start sites of GENCODE genes fall into the active state (P <0.0001). This figure is surprisingly high, given that only a single tissue type was sampled and that, a priori, only a fraction of genes (~30%) is expected to be actively transcribed in the context of a given tissue. Similarly, ~88% of spliced EST 5′ ends in public databases are found within active domains, yet these transcripts represent aggregate analyses of dozens of different human cell types. These data suggest that at any given time, a significant majority of genes in the genome lie within active domains that are permissive for transcriptional activity. As such, restructuring of higher-order functional domains may not be a prerequisite for activation and regulation of most genes.

Of the data types studied above, H3k27me3 and replication timing data were available only for HeLa cells. However, activating histone modifications (H3ac, H4Ac, H3k4me1, H3k4me2, H3k4me3) and RNA data were available for the second ENCODE

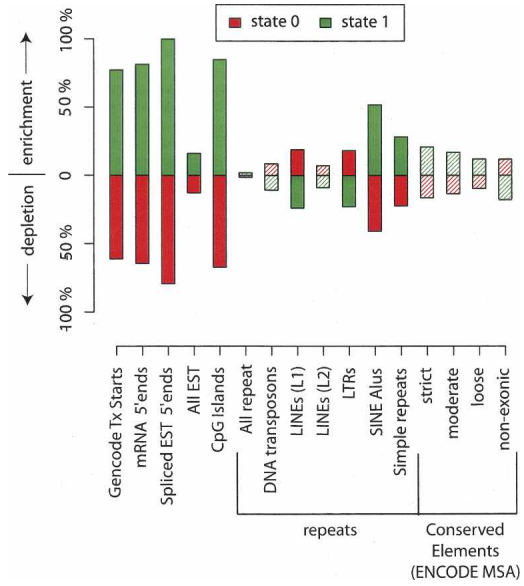

Figure 3. Enrichment and depletion of annotated genomic features in active and repressed domains. Data are based on simultaneous segmentation of four data types (H3ac, H3K27me3, RNA, TR50). Green bars correspond to active state regions, red bars to repressed regions. Values (Y axis) indicate percentage enrichment or depletion over random expectation. For example, GENCODE TxStarts are ~71% enriched over expectation in active regions and ~61% depleted under expectation in repressed regions (see Table 4 for corresponding data). Shaded bars reflect enrichment or depletion that is not significant at the 0.01 level based on the label permutation test (see Methods).

Consortium common cell type, EBV-transformed primary lymphoblastoid cells (line GM06990, Coriell). We therefore asked to what degree the domain segmentations produced for these individual data types would be concordant between two unrelated tissues. Overall, we found these figures to be quite high, with activating histone modifications averaging 74% concordance, and 81% RNA. Taken together, the results from individual data types and the distribution of annotated genomic features between active and repressed domains delineated on the basis of four or five data types strongly suggest that the functional domains we delineated are likely to be persistent across tissue types.

Discussion

The concept that the human genome in vivo is partitioned into a series of functional domains is widely espoused in the literature. Heretofore, however, this model has been based largely on extrapolation from limited studies of specific developmentally regulated mammalian multigene loci under control of distal regulatory elements (Dean 2006), and from analogous studies in avians (Felsenfeld et al. 2004) and flies (Drewell et al. 2002). Such domains exhibit early replication and high levels activating histone modification when their constituent genes are transcriptionally active or committed, and later replication time, depletion of activating marks, and higher average levels of repressive histone modifications when transcriptionally silenced. Expanding on these characteristics are observations that coregulated genes tend to cluster along human chromosomes (Dolganov et al. 1996; Su et al. 2004), and that silenced developmental regulators acquire repressive histone modifications in a domain-like pattern (Bernstein et al. 2006). Prior studies have therefore collectively created the expectation that human gene loci in an active versus repressed functional state should be differentiable on the basis of histone modification patterns, transcriptional permissivity, and timing of replication during the S phase. Previously, however, neither the requisite data sets nor the analytical tools have been available to evaluate the generality of this concept.

Here we have shown that the ENCODE regions are readily divisible into extended domains with common characteristics measured by a combination of functional genomic assays. Although the ENCODE territories comprise only 1% of the genome, they were selected to capture the diversity of human genomic domains, including extremes of gene density and noncoding evolutionary conservation. As such, we anticipate that our basic conclusions should be extensible to the genome at large.

We have developed an analytical paradigm combining wavelet analysis with hidden Markov model segmentation that should likewise be extensible horizontally across the genome, and also vertically into different cell and tissue types as more data become available in each dimension. This paradigm may be particularly useful in the context of model organisms, where the acquisition of genome-wide data is more readily accomplished.

It is perhaps surprising that the human genome is so readily partitioned into regions with distinct functional phenotypes on the basis of relatively few experimental data types. This suggests that, for a given tissue type, additional functional experimental data types may not alter the basic domain map dramatically. Rather, it is likely that data such as additional repressive histone modifications (e.g., H3k9 methylation) or additional markers of active chromatin such as DNaseI sensitivity will help to refine the

Table 4. Enrichment and depletion of annotated genomic features in active and repressed domains based on the four-track simultaneous segmentation

Element	State 0 repressed	State 1 active	State 0 enriched	State 0 *P*-value	State 0 adjusted *P*-value	State 1 enriched	State 1 *P*-value	State 1 adjusted *P*-value
GENCODE TxStarts	588 (0.217)	2121 (0.783)	−0.611	0	0	0.771	0	0
mRNA TxStarts	917 (0.199)	3698 (0.801)	−0.644	0	0	0.812	0	0
Spliced EST TxStarts	15,189 (0.116)	115,531 (0.884)	−0.792	0.0001	0.000246	0.999	0.0013	0.002773
EST Overlap	9,250,097 (0.486)	9,776,837 (0.514)	−0.128	0.0024	0.0048	0.162	0.0005	0.001143
CpG Island Overlap	69,104 (0.183)	308,325 (0.817)	−0.672	0	0	0.848	0	0
All repeats (RepeatMasker)	7,266,816 (0.550)	5,952,407 (0.450)	−0.015	0.1915	0.1915	0.018	0.1738	0.1794
DNA transposons	491,817 (0.605)	320,946 (0.395)	0.085	0.0262	0.034933	−0.107	0.0145	0.0232
LINEs (L1)	2,718,900 (0.663)	1,379,214 (0.337)	0.189	0	0	−0.239	0	0
LINEs (L2)	596,075 (0.597)	402,723 (0.403)	0.07	0.0559	0.06389	−0.088	0.0438	0.05191
LTRs	1,354,800 (0.659)	700,604 (0.341)	0.182	0.0001	0.000246	−0.229	0.0001	0.000246
SINE *Alus*	1247416 (0.329)	2538678 (0.671)	−0.409	0	0	0.516	0	0
Simple repeats	298,957 (0.433)	391,776 (0.567)	−0.224	0.006	0.01129	0.283	0.0076	0.01351
Conserved elements (ENCODE MSA), strict	332,652 (0.466)	381,347 (0.534)	−0.165	0.0312	0.03840	0.208	0.0226	0.03144
Conserved elements (ENCODE MSA), moderate	706,557 (0.483)	756,471 (0.517)	−0.134	0.0298	0.03814	0.169	0.0181	0.02633
Conserved elements (ENCODE MSA), loose	1,757,850 (0.504)	1,726,910 (0.496)	−0.096	0.0174	0.02633	0.121	0.0105	0.01768
Conserved elements (ENCODE MSA), nonexonic	589,993 (0.660)	303,737 (0.340)	0.131	0.0961	0.1025	−0.184	0.0590	0.06510

Column 1 indicates genomic annotations based on genes and transcripts, CpG islands, repetitive elements, and evolutionary conservation. Rows 1–3: Total number of GENCODE transcriptional start sites, mRNA 5′ ends, and 5′ ends of spliced ESTs. Row 4: Total percent of bases overlapped by all ESTs (spliced and unspliced). Total ESTs. Row 5: Total percent of bases overlapped by CpG islands. Rows 6–12: Total percent of bases overlapped by all repetitive elements; DNA repeats; L1 LINEs; L2 LINEs; LTR elements; *Alu* SINEs; and simple repetitive sequences. Rows 13–15: Total percent of bases overlapped by different stringencies (strict, moderate, loose) of conserved sequence elements identified by the ENCODE MSA group. Row 16: Percent bases overlapped by nonexonic noncoding MSA conserved sequences at moderate stringency (=MSA moderate minus GENCODE exons). Columns 2–3 contain the number of elements (or the number of bases of each element) in each state, with the fraction of the total in parentheses. Columns 4–9 show the relative proportion of enrichment or depletion of the element within each state over the expected value, the corresponding *P*-values, and an adjusted *P*-value. *P*-values and adjusted *P*-values are computed via permutation of the state labels using 10,000 iterations (see Methods).

borders of domains. These borders may be particularly important for the identification of regulatory elements that exert insulator or domain boundary function (Bell et al. 2001).

It is even more surprising that the domain architecture remains largely intact across two unrelated tissues, cervical carcinoma cells (HeLa) and B-lymphoblastoid cells (GM06990), at least at the level of the individual ENCODE data types that we studied. Although this observation awaits validation in further tissue environments and the inclusion of additional experimental data types, it suggests a model in which the majority of the genome (~75%) remains in a predictable functional state, where additional large-scale remodeling of domain architecture is not required for gene activity or repression. It also raises the possibility of a universal functional "skeleton" that is common between cell types, and may in turn reflect some basic constraints of nuclear organization. On the other hand, the fact that up to 25% of the genome may be contained in large-scale functional domains that do vary between tissues suggests that the regulatory mechanisms underlying such remodeling must be pervasive. It will be of considerable interest to define such domain-stable and domain-variable regions in the genome through the study of multiple cell types, as this may point the way to heretofore unappreciated regulatory connections between diverse *cis*-linked genes.

One intriguing feature of our domain map is the degree to which highlighting genomic territories by integrating multiple functional data types exposes an organization that cannot be readily predicted from analysis of large-scale patterns of evolutionary conservation alone. Although it is commonly assumed that conserved noncoding sequences may be involved principally in gene activation (e.g., as enhancers), repression of gene activity is an equally important regulatory faculty; indeed, in the case of certain transcriptional factors involved in development and differentiation, repression is equally critical because incomplete silencing may potentiate malignancy.

In summary, our results collectively provide important new insights into the functional organization of the human genome and suggest a general analytical framework for approaching higher-level functional domains in complex genomes.

Methods

Wavelet analyses

The basis for understanding wavelet transforms is the continuous wavelet transform (CWT) (Torrence and Compo 1998; Percival and Walden 2000). Mathematically, for a given time series $x(t)$, the CWT $W(a, t)$ for given scale a and time t is given by

$$W(a,t) \equiv \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(u)\psi\left(\frac{u-t}{a}\right) du,$$

where $\psi(s)$ is the wavelet function of choice, satisfying the basic properties $\int_{-\infty}^{\infty} \psi(u)\, du = 0$ and $\int_{-\infty}^{\infty} \psi^2(u)\, du = 1$. Simple examples of $\psi(s)$ include the Haar wavelet and the so-called "Mexican hat" wavelet. The wavelet coefficient $W(a, t)$ captures information about the local behavior of x at scale a near time (genomic position, in our case) t.

The discrete wavelet transform (DWT) can be thought of as

a discretization of the CWT across evenly spaced values of t and dyadic scales $a_j = 2^j\delta$, where δ is the resolution of $x(t)$. For fixed level J (scale $2^J\delta$), the DWT allows for a decomposition

$$x = \sum_{j=1}^{J} D_j + S_J,$$

of x into a sum of the wavelet *smooth* S_J and wavelet *details* D_j, each of which is a time series the same length as x. This decomposition is called a multiresolution analysis. Each D_j represents the local variation in x at scale $2^j\delta$, while $S_J = x - \sum_{j=1}^{J} D_j$ can be thought of as a smoothed version of x, with the details at lower scales removed. The maximal overlap DWT (MODWT) (Percival and Walden 2000) is a modification of the DWT that also gives rise to the multiresolution analysis above, but which, among other things, allows input sequences x of arbitrary length (the DWT requires the length of x to be a power of 2), at the cost of requiring more, redundant, intermediary wavelet coefficients.

For this analysis we use the Daubechies "least asymmetric" LA(8) wavelet filter (Percival and Walden 2000) (discrete analog of the wavelet function $\psi[s]$). This is a general purpose wavelet, whose "width" (8) strikes a balance between good smoothing properties and lack of significant boundary effects. We use the R package waveslim for computing multiresolution analyses, using reflection boundary conditions.

Wavelet analyses require equally spaced data. Each ENCODE data type has a nominal spacing interval based on the assay used to collect the data (~50 base pairs [bp] for RNA transcription, ~1 kb for Sanger histone modifications, for instance). However, all data types have gaps beyond the nominal spacing, due to assay-specific issues such as repeats in the original genomic sequence. As a preprocessing step to wavelet analysis, we thus interpolate through the gaps using two methods. For gaps <2 kb, we linearly interpolate the available flanking data. For gaps >2 kb, we fill using an interpolated loess curve, where the width of the loess window is chosen to be 50 times the gap length (R function loess, default weights).

Hidden Markov models

A Hidden Markov model (HMM) is a statistical model for systems assumed to be governed by a stochastic process defined by a predetermined finite number of hidden states (Rabiner 1995). Each state is associated with a probability distribution (the emission probabilities) from which the observed outputs are generated. The transition from one state to another is also defined by a random process, so there is a probability associated with each pair of states; collectively these are the transition probabilities.

In our application, the problem is to simultaneously learn the model parameters defining the emission and transition probabilities and the most likely states corresponding to each observed value. Our outputs (RNA transcription levels, DNA replication timing, etc.) are continuous variables, so we require continuous emission probability distributions. We assume independent Gaussian distributions for all emission probabilities.

We used the HMMSeg software package (see also http://noble.gs.washington.edu/proj/hmmseg; Day et al. 2007) for computations. Each model is trained using expectation maximization, repeated 10 times from random initial parameters. The trained model with the highest total probability is used to find the single state path with the highest probability using the Viterbi algorithm (Rabiner 1995). In order to segment multiple tracks at once, HMMSeg allows for multivariate emission probabilities, defined by multivariate, independent Gaussian distributions (diagonal covariance structure).

Data processing

Raw data

All data except for the MSA conserved sequences are downloadable from the UCSC browser, under the "ENCODE" section, using hg17 coordinates. Specific track names are as follows: bulk RNA/transcriptional output: encodeAffyRnaHeLaSignal; DNA replication time: encodeUvaDnaRepTr50; Histone H3 acetylation (H3): encodeSangerChipH3acHela; and Histone H3k27 trimethylation: encodeUcsdChipH3K27me3. We constructed a continuous track representing the density of conserved nonexonic sequences identified by the ENCODE MSA group by computing the fractional occupancy in a sliding 3-kb window, stepping at 1-kb intervals throughout the ENCODE regions. This track is available at http://noble.gs.washington.edu/proj/domain/.

Preprocessing

Prior to wavelet processing, data are preprocessed using the linear and loess interpolation strategy described in the main text, producing equally spaced data at intervals consistent with the original resolution: 50 bp for RNA and DNA replication timing, 500 bp for H3K27me3, and 1000 bp for H3ac. To reduce noise, the interpolated RNA data are thresholded by replacing all negative values with zeros. The conservation density data set is equally spaced by construction, at a resolution of 1000 bp.

Wavelet smoothing

Each data set except for TR50 is smoothed out to a common scale of 64 kb using MODWT wavelet smoothing, and using R function mra from the wavelslim package: parameter settings are method = "modwt," wf = "la8," boundary = "reflection," and n.levels = 10 for RNA and replication timing, seven for H3K27me3, and six for H3ac and conservation density. For the multitrack segmentations, the resulting smoothed data sets are then each interpolated at the 1000-bp resolution interpolated coordinates for H3ac.

Data were not available for DNA replication timing in one of the 44 ENCODE regions (ENm011), because this region is not represented on the Affy ENCODE 1.0 tiling DNA microarray used to generate data. As such, only data for the remaining 43 ENCODE regions are used in the subsequent analysis.

HMM segmentation

We used the HMMSeg software package (Day et al. 2007) to perform individual and simultaneous multitrack segmentations of the processed datasets (43 files per data set), with the following parameter settings: num-states = 2, num-starts = 10, and max-iter = 100. Default values are used for the rest of the parameters. HMMSeg is freely available at http://noble.gs.washington.edu/proj/hmmseg/.

Enrichment of annotated elements

All annotated elements are available publicly through the UCSC Genome Browser. The expected overlap of any element with a specified state in an arbitrary segmentation is just Nd, where N is the total number of elements (or total number of base pairs occupied by the elements), and d is the fraction of the segmented area covered by the specified state. If N_s is the observed number of elements in the given state, then the relative enrichment over the expected value is $(N_s - Nd)/Nd$. In the special case of the conserved noncoding sequence elements, d is corrected to exclude coding exons and UTRs from the total segmented area.

Significance testing

Enrichment *P*-values are computed as follows. A segmentation divides the ENCODE regions into alternating segments with labels 0 and 1. We randomly shuffled the state labels on each segment (in particular, they no longer need to alternate) and recomputed the enrichment/depletion percentage in each of the two new states. We repeated this procedure 10,000 times. The reported *P*-value is the fraction of times that the shuffled enrichment values are as, or more, extreme than the observed enrichment. We also report adjusted *P*-values using the Benjamini–Hochberg correction (Benjamini and Hochberg 1995) for controlling the false discovery rate in a multiple-testing scenario, computed for each segmentation using all permutation-computed *P*-values for both states using the R function *p.adjust* from the *stats* package.

GO annotations

We used the GO::TermFinder software (Boyle et al. 2004), which assesses the statistical signficance of a given subset of genes annotated using the hierarchical annotation set forth by the Gene Ontology (GO) Consortium (Ashburner et al. 2000).

We applied the software to our segmentations, using the set of Ensembl (Hubbard et al. 2005) genes (v37), many of which include annotations using the GO standard.

Acknowledgments

We thank members of the ENCODE Consortium who generated data that were analyzed under this project, including principal investigators Ian Dunham, Tom Gingeras, Anindya Dutta, and Bing Ren. We thank members of the ENCODE Chromatin and Replication Analysis Group for their many helpful suggestions and insights during the genesis of this work, and Elliott Margulies of the ENCODE Multi-Species Alignment Analysis Group for providing data on conserved elements. This work was supported by the NHGRI ENCODE Project.

References

Allen, T.E., Herrgrd, M.J., Liu, M., Qiu, Y., Glasner, J.D., Blattner, F.R., and Palsson, B.O. 2003. Genome-scale analysis of the uses of the *Escherichia coli* genome: Model-driven analysis of heterogeneous data sets. *J. Bacteriol.* **185:** 6392–6399.

Allen, T.E., Price, N.D., Joyce, A.R., and Palsson, B.O. 2006. Long-range periodic patterns in microbial genomes indicate significant multiscale chromosomal organization. *PLoS Comput. Biol.* **2:** 13–21.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Audit, B., Vaillant, C., Arnéodo, A. and Thermes, C. 2004. Wavelet analysis of DNA bending profiles reveals structural constraints on the evolution of genomic sequences. *J. Biol. Phys.* **30:** 33–81.

Azuara, V. , Perry, P., Sauer, S., Spivakov, M., Jorgensen, H.F., John, R.M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., et al. 2006. Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.* **8:** 532–538.

Batzer, M.A. and Deininger, P.L. 2002. *Alu* repeats and human genomic diversity. *Nat. Rev. Genet.* **3:** 370–379.

Bell, A.C., West, A.G., and Felsenfeld, G. 2001. Insulators and boundaries: Versatile regulatory elements in the eukaryotic. *Science* **291:** 447–450.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B.* **57:** 289–300.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas 3rd, E.J., Gingeras, T.R., et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120:** 169–181.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315–326.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botsetin, D., Cherry, J.M., and Sherlock, G. 2004. GO::TermFinder–Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20:** 3710–3715.

Cremer, T., Kreth, G., Koester, H., Fink, R.H., Heintzmann, R., Cremer, R., Solovei, I., Zink, G., and Cremer, C. 2000. Chromosome territories, interchromatin domain compartment, and nuclear matrix: An integrated view of the functional nuclear architecture. *Crit. Rev. Eukaryot. Gene Expr.* **10:** 179–212.

Cooper, K.W. 1959. Cytogenetic analysis of major heterochromatic elements (especially Xh and Y) in *Drosophila melanogaster*, and the theory of "heterochromatin." *Chromosoma* **10:** 535–588.

Day, N., Hemmaplardh, A., Thurman, R.E., Stamatoyannopoulos, J.A., and Noble, W.S. 2007. Unsupervised segmentation of continuous genomic data. *Bioinformatics* ePUB 17384021.

Dean, A. 2006. On a chromosome far, far away: LCRs and gene expression. *Trends Genet.* **22:** 38–45.

Dillon, N. 2003. Gene autonomy: Positions, please. *Nature* **425:** 457.

Dillon, N. 2006. Gene regulation and large-scale chromatin organization in the nucleus. *Chromosome Res.* **14:** 117–126.

Dolganov, G., Bort, S., Lovett, M., Burr, J., Schubert, L., Short, D., McGurn, M., Gibson, C., and Lewis, D.B. 1996. Coexpression of the interleukin-13 and interleukin-4 genes correlates with their physical linkage in the cytokine gene cluster on human chromosome 5q23-31. *Blood* **87:** 3316–3326.

Drewell, R.A., Bae, E., Burr, J., and Lewis, E.B. 2002. Transcription defines the embryonic domains of *cis*-regulatory activity at the *Drosophila* bithorax complex. *Proc. Natl. Acad. Sci.* **99:** 16853–16858.

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* (in press).

Felsenfeld, G. 1996. Chromatin unfolds. *Cell* **86:** 13–19.

Felsenfeld, G., Burgess-Beusse, B., Farrell, C., Gaszner, M., Ghirlando, R., Huang, S., Jin, C., Litt, M., Magdinier, F., Mutskov, V., et al. 2004. Chromatin boundaries and chromatin domains. *Cold Spring Harb. Symp. Quant. Biol.* **69:** 245–250.

Gilbert, N. and Bickmore, W.A. 2006. The relationship between higher-order chromatin structure and transcription. *Biochem. Soc. Symp.* **73:** 59–66.

Hardison, R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16:** 369–372.

Harju, S., Navas, P.A., Stamatoyannopoulos, G., and Peterson, K.R. 2005. Genome architecture of the human β-globin locus affects developmental regulation of gene expression. *Mol. Cell. Biol.* **25:** 8765–8778.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Legarde , J., Gilbert, J.G., Storey, R., Swarbreck, D., et al. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol.* **7:** 1–9.

Hochedlinger, K. and Jaenisch, R. 2006. Nuclear reprogramming and pluripotency. *Nature* **441:** 1061–1067.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., et al. 2005. Ensembl 2005. *Nucleic Acids Res.* 33: D447–D453.

Jeong, K.S. Ahn, J., and Khodursky, A.B. 2004. Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biol.* **5:** R86.

Jurka, J., Kapitonov, V.V., Pavvlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110:** 462–467.

Karnani, N., Taylor, C., Malyhotra, A., and Dutta, A. 2007. Pan-S replication patterns and chromosomal domains defined by genome tiling arrays of ENCODE genomic areas. *Genome Res.* (this issue).

Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karaoz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., et al. 2007. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.* (this issue) doi:10.1101/gr.5704207.

Lamond, A.I. and Earnshaw, W.C. 1998. Structure and function in the nucleus. *Science* **280:** 547–553.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., and FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lee, G.R., Spilianakis, C.G., and Flavell, R.A. 2005. Hypersensitive site 7

of the TH2 locus control region is essential for expressing TH2 cytokine genes and for long-range intrachromosomal interactions. *Nat. Immunol.* **6:** 42–48.
Li, Q. , Peterson, K.R., Frang, X., and Stamatoyannopoulos, G. 2002. Locus control regions. *Blood* **100:** 3077–3086.
Ligon, K.L., Fancy, S.P., Franklin, R.J., Rowitch, D.H. 2006. Olig gene function in CNS development and disease. *Glia* **54:** 1–10.
Lio, P. 2003. Wavelets in bioinformatics and computational biology: State of art and perspectives. *Bioinformatics* **19:** 2–9.
Lio, P. and Vannucci, M. 2000. Wavelet change-point prediction of transmembrane proteins. *Bioinformatics* **16:** 376–382.
Nichols, J. and Nimer, S.D. 1992. Transcription factors, translocations, and leukemia. *Blood* **80:** 2953–2963.
Nobrega, M.A., Ovcharenko, I., Afzal, V., Rubin, E.M. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302:** 413.
Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W., et al. 2004. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.* **36:** 1065–1071.
Percival, D.B. and Walden, A.T. 2000. *Wavelet methods for time series analysis*. Cambridge University Press, Cambridge, UK.
Rabiner, L.R. 1995. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77:** 257–286.
Spitz, F., Gonzalez, F., and Duboule, D. 2003. A global control region defines a chromosomal regulatory landscape containing the *HoxD* cluster. *Cell* **113:** 405–417.
Struhl, K. 1998. Histone acetylation and transcriptional regulatory mechanisms. *Genes & Dev.* **12:** 599–606.
Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, C., et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci.* **101:** 6062–6067.
Torrence, C. and Compo, G.P. 1998. A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79:** 61–78.

Received October 29, 2006; accepted in revised form March 27, 2007.