Downloaded from www.genome.org on September 6, 2007

GENOME RESEARCH

Finding cis-regulatory elements using comparative genomics: Some lessons from ENCODE data

David C. King, James Taylor, Ying Zhang, Yong Cheng, Heather A. Lawson, Joel Martin, ENCODE groups for Transcriptional Regulation and Multispecies Sequence Analysis, Francesca Chiaromonte, Webb Miller and Ross C. Hardison

Genome Res. 2007 17: 775-786
Access the most recent version at doi:10.1101/gr.5592107

Supplementary data *"Supplemental Reseach Data"* **http://www.genome.org/cgi/content/full/17/6/775/DC1**

References This article cites 53 articles, 28 of which can be accessed free at: **http://www.genome.org/cgi/content/full/17/6/775#References**

Open Access Freely available online through the Genome Research Open Access option.

Email alerting service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here**

Notes

To subscribe to *Genome Research* go to:
http://www.genome.org/subscriptions/

© 2007 Cold Spring Harbor Laboratory Press

Downloaded from www.genome.org on September 6, 2007

Letter

Finding *cis*-regulatory elements using comparative genomics: Some lessons from ENCODE data

David C. King,[1,2,7] James Taylor,[1,3,7] Ying Zhang,[1,2] Yong Cheng,[1,2] Heather A. Lawson,[1,4] Joel Martin,[1,2] ENCODE groups for Transcriptional Regulation and Multispecies Sequence Analysis, Francesca Chiaromonte,[1,5] Webb Miller,[1,3,6] and Ross C. Hardison[1,2,8]

[1]Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [2]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [3]Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [4]Department of Anthropology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [5]Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [6]Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

Identification of functional genomic regions using interspecies comparison will be most effective when the full span of relationships between genomic function and evolutionary constraint are utilized. We find that sets of putative transcriptional regulatory sequences, defined by ENCODE experimental data, have a wide span of evolutionary histories, ranging from stringent constraint shown by deep phylogenetic comparisons to recent selection on lineage-specific elements. This diversity of evolutionary histories can be captured, at least in part, by the suite of available comparative genomics tools, especially after correction for regional differences in the neutral substitution rate. Putative transcriptional regulatory regions show alignability in different clades, and the genes associated with them are enriched for distinct functions. Some of the putative regulatory regions show evidence for recent selection, including a primate-specific, distal promoter that may play a novel role in regulation.

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

Deciphering the language and evolution of gene regulatory mechanisms is one of the challenging goals of genomics and systems biology. Even the most basic concepts about the relationship between function and evolution in noncoding DNA are still being refined (Miller et al. 2004; Dermitzakis et al. 2005). Conservation of noncoding sequences among divergent species, inferred from genomic sequence alignments, has been used widely as a predictor of *cis*-regulatory modules (CRMs) (Gumucio et al. 1996; Frazer et al. 2003). Notable success has been achieved with this approach (e.g., Elnitski et al. 1997; Loots et al. 2000; Nobrega et al. 2003). The underlying assumption is that orthologous DNA sequences serving a function common to the species under consideration have changed significantly less than neutral DNA over a sufficient phylogenetic distance. That decreased change, or higher similarity, is taken as a sign of evolutionary constraint, that is, that the DNA is subject to purifying selection. (In this paper, sequences found in common in two or more species by alignment algorithms will be called *conserved*. Those that show a signal for purifying selection will be called *constrained*.)

How often is that underlying assumption really true, and, when it is true, how strong is the signal for constraint in multispecies alignments? Certainly some stringently constrained noncoding sequences are functional. For instance, noncoding sequences conserved between mammals and fish serve as developmental enhancers in gain-of-function assays (Aparicio et al. 1995; Nobrega et al. 2003, 2004; Woolfe et al. 2005; Bejerano et al. 2006). In contrast, some apparently constrained noncoding DNA sequences have little or no obvious function. Some gene deserts contain large numbers of noncoding sequences apparently constrained in mammals, but deletion of two gene deserts from mice generated only mild phenotypes (Nobrega et al. 2004). This led the investigators to "question the functionality, if any, of many of the large number of noncoding sequences shared between mammals." Conversely, some nonconserved sequences are functional. For example, intensive studies from many laboratories have discovered numerous CRMs in globin gene complexes, but evaluation of multispecies sequence alignments shows that some of them are not conserved between human and mouse (Hughes et al. 2005; King et al. 2005). The diversity of results on the relationship between sequence constraint and function of regulatory regions ranges from studies indicating that almost all noncoding sequences in *Drosophila* are under constraint (Andolfatto 2005) to others concluding that promoters have been evolving with reduced constraint since the human–chimp divergence (Keightley et al. 2005).

Although some of the heterogeneity in conclusions may result from differences in methods of analysis, there is no reason to expect that all CRMs will be under the same level of constraint. Indeed, many genes show differences in expression patterns between human and mouse, and hence some sequences in the CRMs should have changed in these cases (e.g., Valverde-Garduno et al. 2004). Binding sites for some transcription factors change in orthologous CRMs, both in *Drosophila* (Ludwig et al. 1998) and in mammals (Dermitzakis and Clark 2002). Transcrip-

[7]These authors contributed equally to this paper.
[8]Corresponding author.
E-mail rch8@psu.edu; fax (814) 863-7024.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.5592107. Freely available through the *Genome Research* Open Access option.

Downloaded from www.genome.org on September 6, 2007

tion factor binding sites that have undergone this process of turnover may no longer align, which will decrease the inferred level of conservation.

The comprehensive pilot project data from the ENCODE Project Consortium (2007) provides the opportunity to evaluate more completely the relationship between function, conservation, and constraint in 1% of the human genome. We used the ENCODE protein occupancy and chromatin modification data to define a set of putative transcriptional regulatory regions (pTRRs). We then used the ENCODE sequence data and alignments to examine the variation in conservation and constraint among the pTRRs. Our analysis confirms wide variation in constraint for pTRRs. Moreover, this variation shows systematic patterns that provide biological insights and suggest improvements to computational predictions of functional elements.

Results

Identification of pTRRs

To define a set of pTRRs, we used the data from chromatin immunoprecipitated samples hybridized to high-density microarray chips (ChIP–chip; Ren et al. 2000) from the ENCODE Transcriptional Regulation Group (The ENCODE Project Consortium 2007). We restricted the protein occupancy data to sites bound by sequence-specific factors and identified using experimental platforms with high site resolution. We improved the specificity of this set by requiring support from at least one line of experimental evidence, including chromatin modifications associated with activation, DNase hypersensitivity (Sabo et al. 2006), and nucleosome depletion (FAIRE; Giresi et al. 2007), yielding a conservative high-resolution set of pTRRs. These and other data sets used in this paper are available at http://www.bx.psu.edu/projects/encode_pTRR.

For comparison we considered two other sets of regulation-associated elements derived from ENCODE data. First, promoter regions were generated from the results of Cooper et al. (2006), who tested 642 potential regions identified using 5′ ends of cDNA alignments in the ENCODE regions. Using the results of their assay, we considered two subsets determined by the range of activity: those validated in all 16 cell lines (ubiquitous promoters) and those validated in 1–5 cell lines (specific promoters). Second, we considered DNase hypersensitive sites (DHSs), as ascertained for ENCODE using quantitative chromatin profiling (Sabo et al. 2006), massively parallel signature sequencing (MPSS; Crawford et al. 2006b), and DNase-chip (Crawford et al. 2006a).

Approaches for measuring sequence-level constraint

A variety of approaches have been developed to measure evolutionary constraint using interspecies alignments. The ENCODE Multispecies Sequence Analysis (MSA) group (Margulies et al. 2007) focused on identifying discrete genomic regions under purifying selection. Using alignments of 23 mammalian species, they integrated three constraint-prediction methods to identify a set of multispecies constrained sequences (MCSs) covering ~4.9% of the human genome—corresponding with estimates that at least 5% of the human genome is under purifying selection (Waterston et al. 2002; Chiaromonte et al. 2003). They found that 40% of these MCSs overlapped coding exons, 20% overlapped other ENCODE functional element annotations, and the remaining 40% overlapped with no annotated functional element.

Each of the constraint prediction methods used for the identification of MCSs has a corresponding quantitative score that assigns a level of constraint to a genomic position or small window. Here we will consider the phastCons score (Siepel et al. 2005).

Another useful measure for identifying regions under evolutionary constraint is *alignability*, simply the fraction of an element that can be aligned between two species. Alignability reflects conservation of a region between the two species—existence of orthologous sequences—but it does not necessarily mean that the region is under constraint (purifying selection). Particular care must be taken when computing the alignability of noncoding features. Coding exons generally show much stronger sequence-level constraint than other classes of functional elements, and proximity to such highly constrained regions may allow the alignment of regions that would not match otherwise. To examine conservation and constraint in noncoding regions fairly, it is important to avoid this anchoring effect. Therefore, we have produced pairwise alignments between human and the 23 other ENCODE targeted species, using BLASTZ (Schwartz et al. 2003), after masking the coding portions of annotated exons in the human sequence (see Methods). The effect of this masking can be observed, for example, in human–mouse alignments. The amount of bases aligning in pTRRs and DHSs decreases by 5% and 8%, respectively (Table 1). The effect is more pronounced in the promoters, with a reduction of up to 16%, reflecting their proximity to the coding exons and sensitivity to the loss of alignment anchors in the coding regions. The computation of alignability score is designed to minimize the effect of unsequenced regions in the comparison species (see Methods).

Finally, the pairwise alignability scores, which are based on a specific pair of species, can be combined into a *composite alignability* score. This score is computed by taking the average of the pairwise alignabilities, weighted by the branch length from human to the comparison species.

Substitution rates vary across ENCODE regions and negatively correlate with estimates of constraint

The neutral rate at which nucleotide substitutions occur affects the ability to infer evolutionary constraint from sequence conservation. For example, a high level of sequence conservation might simply be due to a low neutral rate in that region, not true evolutionary constraint.

Neutral substitution rates vary substantially among ENCODE regions; estimates of human–mouse divergence produced by REV

Table 1. The fraction of the ENCODE regions aligned before and after hard-masking coding sequences for human–chimp and human–mouse alignments

	Chimp-aligned bases			Mouse-aligned bases		
Feature	Before mask	After mask	Difference	Before mask	After mask	Difference
DHS	0.93	0.91	3%	0.63	0.53	10%
pTRR	0.93	0.91	3%	0.78	0.70	8%
Specific promoters	0.89	0.85	4%	0.74	0.57	16%
Ubiquitous promoters	0.91	0.90	1%	0.73	0.63	11%
ARs	0.93	0.93	0%	0.34	0.27	7%

Downloaded from www.genome.org on September 6, 2007

models of substitutions in ancestral repeats (t_{AR}) range from 0.43 to 0.61 substitutions per site, consistent with the range observed in whole-genome studies on megabase sized intervals (Waterston et al. 2002; Hardison et al. 2003). Furthermore, t_{AR} correlates significantly with other measures of the neutral rate including divergence at fourfold degenerate codon positions (t_{4d}; $r = 0.51$) and the local density of single nucleotide polymorphisms ($r = 0.63$).

Consistent with the expectation that variation in the neutral rate affects constraint estimates, we find that the density of MCSs in each ENCODE region is negatively correlated with t_{AR} ($r = -0.57$, -0.50, and -0.48 for loose, moderate, and strict MCSs, respectively; Fig. 1). This is seen despite the fact that the MCS thresholds were determined based on randomly chosen alignments within ENCODE regions (Margulies et al. 2007). This normalization, however, apparently was not sufficient to completely eliminate the neutral rate effect between regions. The average phastCons score of each region is even more strongly correlated with t_{AR} ($r = -0.63$). This association indicates the importance of taking into account local variability, such as that of the neutral rate, when producing estimates of evolutionary constraint. Neither MCS nor phastCons computations completely correct the neutral rate effect across regions. Both theory (Eddy 2005) and empirical observations (Li and Miller 2003) show that constrained elements stand out with greater statistical power in regions that evolve faster. Normalization for local rate variation may improve the resolution of constrained elements in slower-evolving regions. The causes of the local rate variation are not fully understood, and several evolutionary processes have been discussed. For example, the regional variation may reflect areas that share adaptive trends, such as developmental genes in cold spots and immune-response genes in hot spots (Chuang and Li 2004).

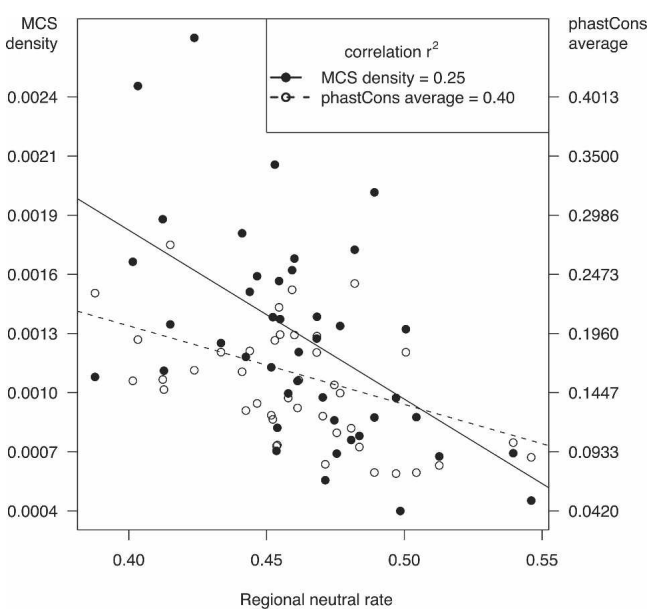

Figure 1. Negative correlation of neutral rate with measures of constraint. For each ENCODE region, the MCS density per nucleotide (black) and the phastCons average (red) are plotted against the human–mouse substitution rate in ARs (t_{AR}, an estimate of the neutral rate of substitution). The correlations with t_{AR} are -0.50 ($P = 0.0005$) and -0.64 ($P = 3 \times 10^{-6}$) for MCS density and phastCons, respectively; the inset gives the r^2 values.

Functional elements are better identified by alignment-based scores than by overlap with highly constrained regions

The ENCODE Consortium (2007) found that while most classes of noncoding functional elements are enriched for MCSs, many elements of every class considered do not overlap with them. This is consistent with the notion that 5% is a lower bound in evaluating what share of the genome is involved in function. Looking beyond the most highly constrained sequences of the genome by considering other quantities (or scores) calculated from genomic alignments can provide greater power in detecting functional elements. In addition to overlap with MCSs, here we consider phastCons scores and composite alignability. When comparing constraint scores of different elements from different regions of the genome, it is important to take regional variation into account. Here we performed *background correction* by scaling interval scores relative to the overall score of the containing ENCODE region (see Methods).

Figure 2 shows the distributions of an illustrative set of scores on the different classes of predicted functional elements. In general, we see that all of these measures have some ability to distinguish functional elements from the neutral background (non-MCS ancestral repeats [ARs]). Also, all are broadly distributed with a large number of high-end outliers, suggesting that in every class there is a subset of elements that is well characterized by each measure. However, we also see that all classes of elements, except promoters, have medians for MCS coverage at or near zero, indicating that at least half of the elements have no overlap with MCSs. Correcting for the background neutral rate improved the separation of the feature sets from ARs, both for phastCons and for composite alignability. In general, the DHS regions have the least separation from background. The background-corrected composite alignability gave the most consistent separation of the four functional classes from the neutral background.

The discriminatory power of each score can be evaluated by measuring sensitivity (the ability to identify the regulatory feature—pTRR, DHS, or promoter—at a given threshold) and specificity (the ability to exclude ARs at that threshold). However, it is impractical to compare MCS overlap with the other scores because of the limited range of possible specificities. MCS overlap presents a very high specificity even at the lowest threshold (no MCS overlap), excluding 97% of ARs. MCS overlap also has a very low maximum sensitivity for detecting regulatory elements that do not tend to be adjacent to exons, for example, ~0.26 for pTRRs versus ~0.63 for promoters. This low sensitivity could imply that most regulatory regions are not constrained. However, the score distributions suggest instead that MCSs select for only a very highly constrained subset of regulatory elements and miss many other regions that are under constraint.

Figure 3 compares performance (receiver operator characteristic, or ROC) curves for each of the scores (with the exception of MCS overlap) on selected classes of predicted functional elements. From these curves we can see that while phastCons performs best for classifying specific promoters, composite alignability gives better overall performance for finding pTRRs and ubiquitous promoters. Correction for regional background variation improved performance dramatically for composite alignability in all tests, whereas the improvement for phastCons was smaller but notable for promoters. None of the scores perform well for identifying DHSs. Many of these regions cannot be aligned at all, which affects all the scores, but particularly alignability (the jump in the ROC curve corresponds to zero alignability).

Downloaded from www.genome.org on September 6, 2007

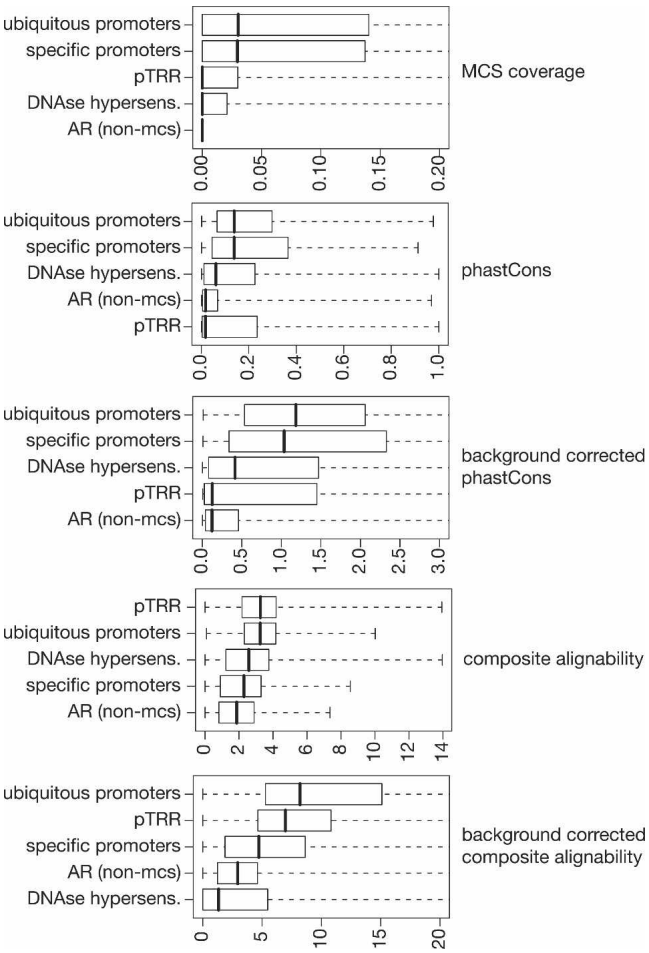

Figure 2. Distributions of scores in regulatory regions for alignment-based measures. Each panel shows the score distributions as box plots, with the box extending from the 25th to 75th percentiles and the vertical line giving the median. The distribution boxes are ordered by the medians of each data set. The graph for MCS coverage shows the fraction covered by MCSs (Margulies et al. 2007). The phastCons score (Siepel et al. 2005) is a likelihood of being in the slowly changing class of genomic sequences; the graph shows the average phastCons value per interval. For the background corrected phastCons score, the graph shows the interval average divided by the regional average. The composite alignability score, before and after background correction is in arbitrary units. In all cases, larger numbers reflect greater constraint (MCS coverage and phastCons) or greater conservation (composite alignability).

To more directly compare the performance of each constraint score (again with the exception of MCS overlap) on each feature set, we examined the sensitivity when specificity was fixed at 0.75 (for each score this is the threshold at which 75% of ARs were excluded; Table 2). Composite alignability performs best for ubiquitous promoters, while the phastCons score performs best for discriminating specific promoters. Background-corrected composite alignability achieves excellent specificity for identifying pTRRs. The correction contributes substantially to this performance, increasing the specificity from ~0.60 to ~0.76. In general, correcting for regional variation yields a (sometimes substantial) improvement in performance, regardless of whether phastCons or alignability is used. The DHS data set is not discriminated with good specificity by any of these quantities, again suggesting that constraint is of limited utility for identifying these elements.

These results show that while each set of regulatory regions shows evidence for constraint, individual elements differ widely by any of the measures applied. Some are changing faster than presumptively neutral DNA, others are constrained in all species examined, and the rest fall into a level of constraint between these extremes. In the next section, we turn to functional inferences that can be drawn from the phylogenetic extent of conservation.

Functional elements are conserved at varying phylogenetic distances

The most distant species to which a human region aligns can be used to estimate the clade in which that DNA region has a common function, that is, it captures clade specificity. Therefore, we examined the exon-masked alignments to find the most distant species that still aligns with human for each member of the feature data sets. We only required that a single base pair of human sequence align with the comparison species, but in practice, we found that all alignments covered at least 10% of each human region. After masking exons, the vast majority of pTRRs and DHSs aligned to placental mammals (70%–71%) or to mammals including the marsupial monodelphis and/or the monotreme platypus (14%–21%, Table 3). A small but notable fraction of pTRRs (3%) aligns only in primates; this fraction is greater for DHSs (11%). A similar fraction shows the opposite behavior, aligning over the larger phylogenetic distance to other tetrapods or other vertebrates. This trend is also seen for the promoters, with a slightly greater fraction (4%) of those expressed in a subset

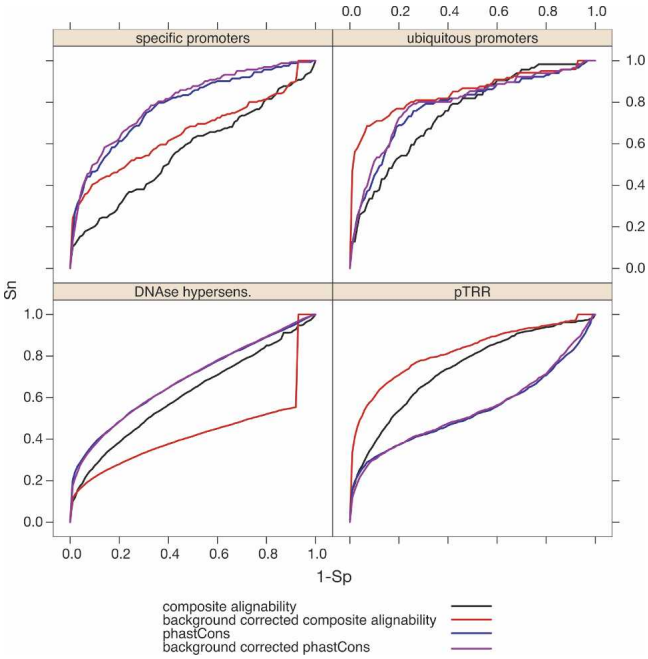

Figure 3. Receiver operator characteristic (ROC) graphs showing the performance of alignment-based scores to discriminate regulatory regions from neutral DNA. The ROC graphs show the sensitivity (Sn) and 1 − specificity (1 − Sp) as the alignability and phastCons thresholds are increased. Clear discrimination leads to curves deflected into the upper *left* quadrant of the graph. Sensitivity is measured as the ability to capture members of the four indicated sets of regulatory regions. Specificity is measured as the ability to exclude ARs, a model of neutral DNA. The results are given for each score, before and after regional background correction.

Downloaded from www.genome.org on September 6, 2007

Table 2. Sensitivity of different scores when specificity is fixed at 0.75

Feature	Score	Correction type	
		None	Background
DHS	Alignability	0.4322	0.3075
	phastCons	**0.5346**	0.5316
pTRR	Alignability	0.6033	**0.7552**
	phastCons	0.3989	0.4053
Specific promoters	Alignability	0.3681	0.5205
	phastCons	0.6687	**0.6871**
Ubiquitous promoters	Alignability	0.5948	**0.8017**
	phastCons	0.7328	0.7845

The highest performing score for each feature is shown in boldface type.

of cell lines aligning out to fish. The biochemical support for function of these pTRRs and DHSs is equally strong. Thus, to the extent that the alignments are reflections of true evolutionary relationships, these results are most easily interpreted as indicating the clades in which an ancestral functional element remains active in extant species.

Genes associated with clade-specific pTRRs show distinct functional enrichments

The functional regions found in specific clades may share particular properties. Here we focus our attention on pTRRs to investigate whether the elements conserved in each clade tend to regulate distinctive functional classes of genes, as described by Gene Ontology (GO) terms (Ashburner et al. 2000). The coding regions of virtually all genes in the ENCODE regions are conserved in the species examined from primates to fish, but a subset of pTRRs associated with some of these genes is clade specific. Our study was designed to test whether ENCODE genes associated with clade-specific pTRRs are enriched in particular functional categories. The gene nearest each element was inferred to be its target of regulation. The GO terms associated with the inferred target genes were analyzed to find the ones significantly enriched for each clade (<5% false discovery rate, or FDR; see Methods). Four of the five clades show a substantial number of GO term enrichments: primate, placental mammals, mammals (including marsupial and monotreme), and tetrapods. These significant terms were then filtered to find the terms distinctively enriched for a clade, for example, significantly enriched in that clade, but not in any other clade. Selected frequently occurring GO terms in the distinctively enriched sets for each clade are shown in Table 4.

Some of the distinctive GO categories are consonant with known lineage-specific features and others reveal novel insights. The pTRRs conserved in primates (but no further) are enriched for immune-related receptor function. This is consistent with reports of immune-related adaptations at the sequence level of genes (Altschuler et al. 2005) and has also been seen in recent gene duplications and copy number variation in human (Aldred et al. 2005). An example is a set of pTRRs in the 5′ flanking region of *LILRA4*, which encodes a member of the leukocyte immunoglobulin-like receptor subfamily (Fig. 4A). The ChIP–chip data from the ENCODE Consortium indicate that this DNA is occupied by CEBPE, PU.1 (SPI1), and the retinoic acid receptor in HL-60 cells, but this sequence is found only in primates. On the other extreme, pTRRs aligning in tetrapods (from humans to chicken or *Xenopus*) are enriched in GO terms for transcription factors (Table 4).

Many of the GO terms enriched in genes associated with pTRRs found only in placental mammals are related to inhibitors of proteases. Several genes contribute to this group, including the serpins and *TIMP3* (Table 4). Other terms found with multiple genes relate to ion transport.

Several genes associated with pTRRs conserved in all mammals (including marsupials), but not birds and fish, have a role in cell cycle control. One is *STAG2*, encoding stromal antigen 2, which plays a role in chromosome dissociation during mitosis (Hauf et al. 2005). Many pTRRs are found in the first intron, supported by binding by SP1 and MYC (Fig. 4B); this region is strongly conserved in mammals including monodelphis but no further. A homolog to the gene *STAG2* is found in all clades examined, but our analysis suggests that while the basic dissociation process is present in all species, some aspect of its regulation differs between mammals and other vertebrates.

pTRRs in candidate regions for recent selection

The observation that constraint varies broadly within regulatory elements could be explained by some subset of them being either under positive selection or degrading because of relaxation of selection (Keightley et al. 2005). Here, we use human polymorphism and interspecies divergence to assess whether an element or class of elements shows evidence of selection, and to distinguish between negative selection (constraint) and positive selection (adaptation). A significant excess of polymorphism relative to divergence is consistent with negative selection, and a significant excess of divergence relative to polymorphism is consistent with positive selection, although other factors such as changes in population size can also explain the results. We applied the McDonald-Kreitman test (McDonald and Kreitman 1991) in 10-kb windows across the ENCODE regions (H. Lawson, J. Martin, D.C. King, B. Giardine, W. Miller, and R.C. Hardison, in prep.), using the ratio of polymorphisms to divergence in ARs within each window to estimate the local rate of mutation and fixation of changes in likely neutral sites (Waterston et al. 2002; Hardison et al. 2003). The ratio of polymorphism to divergence for all noncoding, non-AR sites was compared with the ratio for AR sites in each window. Neutral theory predicts that the two ratios will be

Table 3. Partitioning of putative regulatory regions by phylogenetic clade

Data set	Total	Primate	Placental mammal	Mammal	Tetrapod	Vertebrate
pTRR	1369	43 (0.03)	971 (0.71)	290 (0.21)	57 (0.04)	8 (0.01)
DHS	8099	931 (0.11)	5635 (0.70)	1123 (0.14)	313 (0.04)	97 (0.01)
Specific promoters	158	20 (0.13)	100 (0.63)	26 (0.16)	6 (0.04)	6 (0.04)
Ubiquitous promoters	116	1 (0.01)	73 (0.63)	32 (0.28)	8 (0.07)	2 (0.02)
ARs	23,840	3697 (0.16)	19,857 (0.83)	260 (0.01)	19 (0.00)	4 (0.00)

For each clade, the table lists the number (fraction) of members of a class of predicted functional elements whose most distant aligning species is in that clade. Three ARs found only in human are not included in this table.

Downloaded from www.genome.org on September 6, 2007

Table 4. Selected GO terms distinctly enriched within clades

Clade	Process	GO:ID	Population	Sample	*q*-value	Term description	Genes
Primates	Immune response	GO:0006955	38/1369	10/43	0.0006	Immune response	*KIR2DS2, IL5, LILRA4*
		GO:0004872	98/1369	12/43	0.0009	Receptor activity	*IFNGR2, KIR2DS2, LAIR1, LENG9, LILRA4*
		GO:0004871	150/1369	13/43	0.0083	Signal transducer activity	*IFNGR2, LAIR1, IL5, LENG9, KIR2DS2, LILRA4*
	Phosphatase	GO:0003993	11/1369	5/43	0.0071	Acid phosphatase activity	*HISPPD2A*
Placental mammals	Protease inhibition	GO:0004867	31/1369	31/971	0.0005	Serine-type endopeptidase inhibitor activity	*SERPINB2, SERPINB3, SERPINB7, SERPINB8, SERPINB10*
		GO:0004866	70/1369	62/971	0.0025	Endopeptidase inhibitor activity	*BIRC4, SERPINB2, SERPINB3, SERPINB7, SERPINB8, SERPINB10, SPP2, RENBP, TIMP3*
	Ion transport	GO:0015075	57/1369	50/971	0.0117	Ion transporter activity	*SLC22A5, ATP11A, SLC4A3, CATSPER2, CFTR, CACNG8, SLC22A4, SLC5A4, ATP11A, ATP5O, CACNG6, SLC10A3*
Mammals	Mitosis and cell cycle	GO:0007059	14/1369	10/290	0.0005	Chromosome segregation	*STAG2*
		GO:0051301	14/1369	10/290	0.0005	Cell division	*STAG2*
		GO:0007067	16/1369	11/290	0.0003	Mitosis	*STAG2, YWHAH*
	Aminoglycan synthesis	GO:0006024	27/1369	13/290	0.006	Glycosaminoglycan biosynthesis	*EXT1*
Tetrapods	Transcriptional regulation	GO:0043565	39/1468	10/57	0.0040	Sequence-specific DNA binding	*HOXA11, FOXP4, HOXA7, HOXA13, HOXA4, HOXA3, EVX1*
		GO:0003700	58/1468	10/57	0.0155	Transcription factor activity	*HOXA11, FOXP4, HOXA7, HOXA13, HOXA4, HOXA3, EVX1*

The *q*-value indicates false discovery rate, distinct terms have a *q*-value < 0.05 in a specific clade and not in any other.

Downloaded from www.genome.org on September 6, 2007

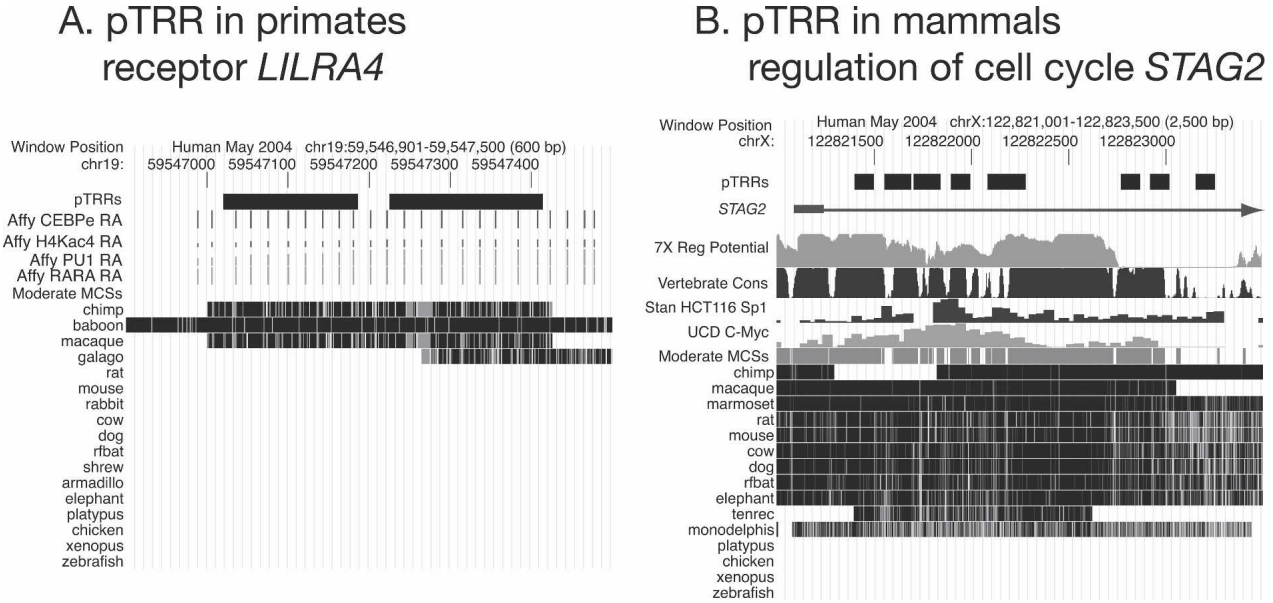

Figure 4. Examples of clade-specific pTRRs. The panels show views from the UCSC Genome Browser (Thomas et al. 2007), focused on pTRRs found only in primates (*A*, close to the *LILRA4* gene) or in mammals including marsupials (*B*, within the *STAG2* gene). The tracks from *top* to *bottom* show the pTRRs (black rectangles), the gene if the pTRR is in a gene, the ENCODE transcription-related data (The ENCODE Project Consortium 2007) that led to the identification of a pTRR, the moderate MCSs (Margulies et al. 2007), and the positions of segments aligned with the indicated comparison species using TBA (Blanchette et al. 2004). The transcription-related ENCODE data tracks are ChIP–chip data from Affymetrix on occupancy in HL60 cells by CEBPE, PU.1 (SPI1), and the retinoic acid receptor as well as hyperacetylation of histone H4 (*A*), and ChIP–chip data from Stanford on occupancy in HCT116 cells by SP1 and from the University of California at Davis on occupancy in HeLa cells by MYC (*B*). Panel *B* also shows the scores for regulatory potential (Taylor et al. 2006) and phastCons (Vertebrate Cons; Siepel et al. 2005).

the same for DNA that is not under selection, and this hypothesis was evaluated with a χ^2 test. Of the 33 windows in the ENCODE regions that show the strongest deviations from neutrality, we found that 16 contained pTRRs. Three of the 16 windows showed an excess of divergence consistent with positive selection while 13 showed a deficit of divergence consistent with negative selection. The limited overlap of pTRRs with windows that deviate from neutrality does not suggest enrichment for pTRRs. In fact, all regulatory data sets examined showed less overlap than expected by chance. However, we do not expect that most regulatory regions would be implicated in this analysis, because only a subset of pTRRs is are likely to show recent selection.

One example of pTRRs in a window with a signature for recent positive selection is near *PDLIM4* (Fig. 5). The encoded protein has PDZ and LIM domains, and it regulates the association of actin stress fibers with actinin. Variants in the noncoding portion of this gene are associated with osteoporosis (Omasu et al. 2003). The noncoding sequences in a 10-kb window encompassing the 5′ flank and first two introns deviate significantly from neutrality, based on divergence from chimpanzee, and the low neutrality index suggests positive selection. The pTRRs in the introns reflect binding of MYC and SP1 as well as chromatin modifications (FAIRE and DNase HSs, Fig. 5). Another striking example is the *SPAG4* gene, defects in which are associated with reduced sperm mobility and infertility.

Recent selection supports a novel function for a primate-specific, distal promoter

Three pTRRs in a window showing a signature of recent purifying selection provide evidence for the importance of distal transcription in the regulation of human beta-globin genes. The pTRRs are located close to the *UBQLN3* gene, about 250 kb from the *HBB* gene complex (Fig. 6). They are in a window that deviates significantly from neutrality in comparison with chimpanzee (and rhesus, not shown), with a deficit in divergence consistent with recent constraint. The pTRRs are close to a promoter for a set of long transcripts that can extend into the embryonic *HBE1* and fetal *HBG2* genes; other spliced products of the transcripts are noncoding. These transcripts are present in erythroid K562 cells, as shown by RT-PCR assays (Fig. 6). The major promoters for production of globin mRNA are proximal to the genes, and the role of these transcripts that initiate distally is unclear. The promoter is in an endogenous LTR-containing retrovirus that is found only in primates (humans, apes, and simians), thus precluding functional tests in mice. The fact that the promoter resides in a window significantly deviating from neutrality is consistent with an important biological role for this activity in higher primates. If indeed the explanation for the deviation from neutrality is selection, the excess polymorphism in non-AR sites suggests that the region is under recent purifying selection, that is, to maintain a primate-specific function. The resolution of the test is not sufficient to directly implicate this distal promoter as the target of selection, but it does provide an intriguing candidate.

Although this promoter is distal to the *HBB* complex along the linear chromosome, it is close to the locus control region of the *HBB* complex in the nucleus of K562 cells, as revealed by chromosome conformation capture (3C; Dekker et al. 2002). The interaction frequency measured by 3C (Fig. 6) is determined by cross-linking DNA to proteins in cells, isolating the cross-linked DNA, digesting with a restriction enzyme, and ligation under conditions that favor rejoining ends within a DNA molecule. DNA segments that are far apart in the linear sequence but close

Downloaded from www.genome.org on September 6, 2007

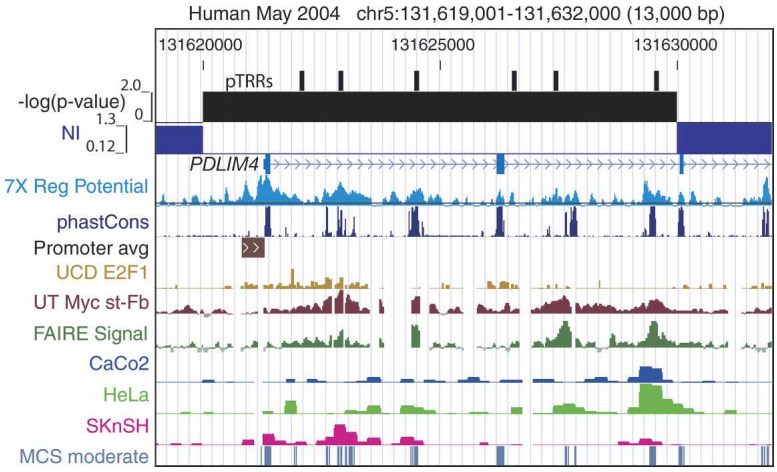

Figure 5. Recent positive selection in the *PDLIM4* gene. The customized view from the UCSC Genome Browser covers about 13 kb of ENCODE region ENm002 centered at gene *PDLIM4*. It shows the locations of pTRRs, the *P*-value for deviation from neutrality for 10-kb windows (H. Lawson, J. Martin, D.C. King, B. Giardine, W. Miller, and R.C. Hardison, in prep.), the neutrality index (Rand and Kann 1996; with values >1 implying negative selection and values <1 implying positive selection), positions of known genes, regulatory potential scores (Taylor et al. 2006), phastCons scores (Siepel et al. 2005), active promoters (Cooper et al. 2006), occupancy by E2F1 (Bieda et al. 2006), occupancy by c-Myc (Kim et al. 2005), formaldehyde-assisted isolation of regulatory elements (FAIRE) (Giresi et al. 2007), DNaseI hypersensitive sites measured in $CaCo_2$, HeLa, SKnSH cell lines/phenotypes (Sabo et al. 2006), and moderate MCSs generated by the ENCODE Multispecies Analysis group (Margulies et al. 2007).

in the nucleus will form novel junctions between restriction fragments. The frequency of detecting novel junctions, as assayed by PCR, is normalized to the frequency observed when uncrosslinked genomic DNA from this region (in a BAC clone) is analyzed in the same way. This BAC DNA control adjusts for preferential ligation between some restriction fragments. The interaction between HS2 of the locus control region and the active globin genes such as *HBG1* gene has been previously documented (e.g., Carter et al. 2002; Tolhuis et al. 2002; Vakoc et al. 2005; Dostie et al. 2006), and represents a stable interaction between the HS2 enhancer and a highly transcribed gene. The interaction frequency between the distal promoter and HS2 is lower but substantially above that of several other DNA fragments closer to HS2. This result is supported by data in a recent report (Dostie et al. 2006). Thus, the results indicate significant interactions between the distal promoter and the LCR, along with the conventional promoters for the globin genes. This proximity, combined with the observations that noncoding transcripts from the distal promoter extend into the globin genes and that the region containing the distal promoter shows evolutionary signals consistent with recent selection, suggest that the distal promoter could play a role in regulation of globin gene expression.

Discussion

The ENCODE pilot project (The ENCODE Project Consortium 2007) has produced excellent resources both for defining putative regulatory elements (through extensive protein binding and chromatin accessibility data) and for evaluating the extent of interspecies conservation of these sequences. The Multispecies Sequence Analysis group of the ENCODE pilot project (Margulies et al. 2007) focused on identifying the most highly constrained regions of the human genome and produced a set of MCSs that cover ~5% of the bases in the ENCODE regions, consistent with estimates that at least 5% of the genome is under constraint between human and mouse (Waterston et al. 2002; Chiaromonte et al. 2003). However, with the exception of protein coding exons, few classes of functional elements are well predicted by these highly constrained regions. We find this to be particularly true for gene regulatory elements.

Though these regions lack the level of deep evolutionary constraint required for MCS annotation, they still exhibit detectable evidence for constraint. We find that quantities based on interspecies comparisons can discriminate many of these regulatory regions from neutral DNA. Further, correcting for regional background variation increases this discrimination ability, sometimes substantially. This less stringent view of evolutionary constraint allows the identification of a wider range of potentially important sequences. While every class of elements may contain some subset that exhibits deep conservation—such as the regulatory elements associated with developmentally important genes (Woolfe et al. 2005)—deep conservation is the exception rather than the rule. Not only is 5% only a lower bound for constrained DNA in eutherian mammals, but it is perhaps a vast underestimate of the amount of functional sequence that can be detected using the right interspecies comparisons. Relaxing the requirement for deep constraint, and instead examining conservation and constraint at varying distances, reveals more functional elements.

Different classes of elements may tend to be constrained over different phylogenetic spans, and even within a class of elements the depth of constraint may vary. We find this to be the case for the various types of regulatory elements considered here. Many elements not identified as MCSs nonetheless show constraint within some subset of the mammalian phylogeny. In addition, we find that elements conserved within different clades are associated with genes that are significantly and distinctly enriched for particular functional categories. We stress that these functional enrichments were obtained by examining only the genes in the ENCODE regions. As more data become available, a similar analysis should be done on clade-specific pTRRs in all genes. This may reveal additional, and possibly stronger, enrichments for functional categories.

Analysis of within-species variation, combined with interspecies comparison, has the power to detect regions that are subject to positive selection, as well as regions that have only recently become subject to negative selection (constraint). By combining human polymorphism data with sequences of primates closely related to humans, we have found putative regulatory elements of both types. We have identified a primate-specific distal promoter within a 10-kb region showing evidence for recent selection. The noncoding transcripts from this promoter extend into the beta-globin gene locus. If indeed the distal promoter is a target of selection within the window, then this deviation from neutrality suggests that the promoter and its transcripts are playing an important role. Active genes are associated

Downloaded from www.genome.org on September 6, 2007

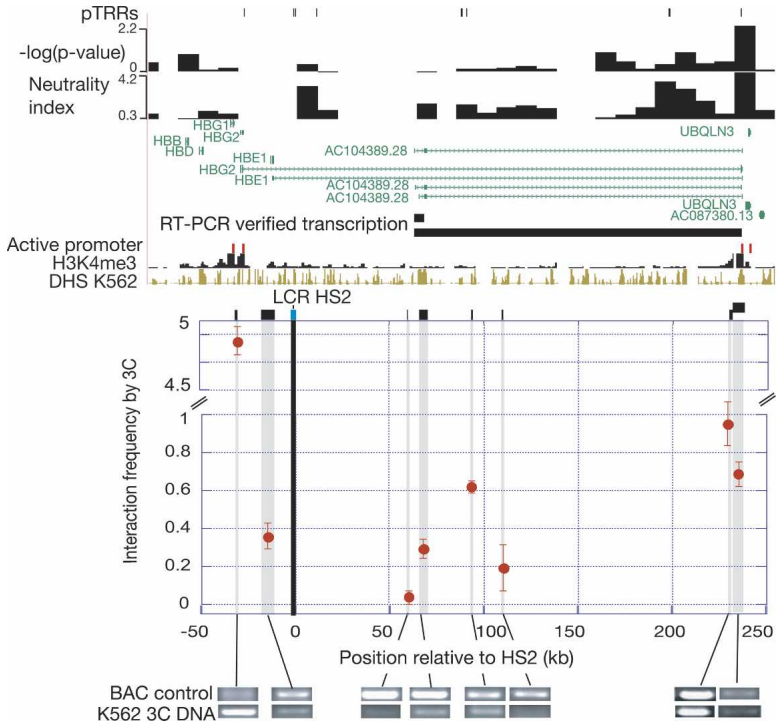

Figure 6. Recent purifying selection in a distal promoter for a noncoding transcript. The customized view from the UCSC Genome Browser (*top*) covers 315 kb of ENCODE region ENm009 extending from the *HBB* gene complex to the distal gene *UBQLN3* (chr11:5,185,001–5,500,000 in the May 2004 assembly of the human genome). Many tracks are the same as in Fig. 5. In addition, the figure shows trimethylation on lysine 4 of histone H3 and DHSs in the cell line K562 (The ENCODE Project Consortium 2007), along with maps of transcripts independently confirmed by RT-PCR in K562 cells in this study. The graph in the *middle*, which is aligned with the Browser view, shows the frequencies of interaction among the HindIII DNA fragments indicated by the black rectangles above the graph, using chromosome conformation capture (Dekker et al. 2002). Data shown are the average of two independent experiments, with each measurement normalized to the BAC DNA control (a BAC containing these segments of human chromosome 11 that is digested and ligated by the same procedure as the chromosomal DNA in K562 cells). Images of electrophoretic gels are below the graph; they indicate the abundance of the PCR products for each 3C interaction, both for the BAC control (*top* row) and the interactions in K562 cells (*bottom* row).

with transcription factories, and loci that produce more transcripts tend to spend more time in the factory (Osborne et al. 2004). One interesting possibility, suggested by the proximity of this promoter to the locus control region, is that transcription from the distal promoter may be part of the process that keeps the beta-globin gene locus strongly associated with transcription factories in erythroid nuclei.

Our analysis of comprehensive functional data in combination with multiple species alignments over the 1% of the human genome covered by the ENCODE pilot project has led to several lessons for practical application. First, it is unlikely that sequence comparisons alone, in the absence of high-throughput biochemical data, will identify gene regulatory regions comprehensively. The continuation of the ENCODE project and other efforts for genome-wide data on protein occupancy and chromatin modifications will provide much valuable information on gene regulatory regions. Second, comparative sequence analysis can help in interpreting these comprehensive new functional data, but a variety of approaches should be used. Overlap with MCSs indicates a stringent constraint on function. Quantitative constraint scores, and less stringent measures like composite alignability, are useful to capture the range of constraint levels seen in noncoding functional elements. Indeed, a measure such as alignability, which can have relatively weak requirements in terms of conservation of individual bases, is likely to be relatively robust to some types of changes that do not disrupt function, such as turnover of transcription factor binding sites and nonconsequential rearrangements. Third, the phylogenetic extent of conservation of a regulatory region may be related to the physiological role of the target gene. Another intriguing possibility is that the extent of conservation may relate to particular mechanistic properties of the regulatory regions. Both these avenues for interpreting the clade-specificity of regulatory regions should be pursued in the future. Fourth, intraspecies polymorphisms and divergence from closely related species should be examined for evidence of recent selection. It is possible that a substantial fraction of the regulatory regions in humans (or any species) have been active only recently on an evolutionary time scale. We have used one approach based on the McDonald-Kreitman test. Much effort is being devoted to developing better tools for interpreting these data, and important progress is expected in the future.

Methods

Sequence alignments

For phastCons calculations, the multiple-species alignments of ENCODE regions (including coding exons) generated using TBA by the ENCODE Multispecies Sequence Analysis group were used (Blanchette et al. 2004; Margulies et al. 2007). For computing alignability and maximal phylogenetic extent of analysis, alignments were computed between ENCODE sequences (The ENCODE Project Consortium 2007), in which human sequences were hard-masked for coding exons. BLASTZ (Schwartz et al. 2003) was run with modified parameters to increase sensitivity, because one of the major sources of alignments seeds (coding exons) was masked. In particular, the threshold for MSPs (K) was set at 1800 and the threshold for gapped alignments (L) was set at 2300. Alignments were filtered for single coverage with respect to the human sequence. The software for producing and processing alignments is available (http://www.bx.psu.edu/miller_lab/).

ENCODE data sets and sequence

Annotations of coding sequence were taken from the ENCODE Consortium (2007), as were AR regions—these are defined as older than the common ancestor of human and dog. ENCODE promoter regions were taken from Cooper et al. (2006), who identified 921 potential promoters based on full-length cDNA libraries. Of these they tested all those associated with multiexon genes (528) and a sample of those associated with single-exon genes (114) in 16 diverse cell lines using transient transfection reporter assays, declaring a DNA fragment as functional in a

Downloaded from www.genome.org on September 6, 2007

given cell line if it showed significant activity relative to negative controls.

Preparation of pTRRs

A subset of the ChIP–chip identified binding sites produced by the ENCODE transcriptional regulation consortium (The ENCODE Project Consortium 2007) was selected emphasizing (1) high-resolution site identification and (2) sequence-specific binding not exclusively associated with transcription start sites. To achieve high resolution only experiments performed on the NimbleGen or Affymetrix platforms were used. The 5% FDR identified sites were used; however, the hits identified using the NimbleGen platform were not post-processed to eliminate multiple sites within 1 kb. All sites were expanded to a representative genomic region covering at least 100 bp. In defining this set only experiments for the following factors were included: SP1, SP3, E2F1, E2F4, MYC, STAT1, JUN, CEBPE, PU.1 (SPI1), RARecA. All of these factors bind to DNA with sequence specificity and are not known to be exclusively associated with 5′ ends of genes. Thus, the resulting set contains high-resolution binding sites, which may contain both proximal and distal regulatory elements. We eliminated all sites overlapping repetitive regions (due to limitations of array hybridization) or coding exons (though sequence-specific binding in coding exons is interesting, signals in these regions are dominated by the constraints of protein coding function).

To refine this set further we identified subsets supported by additional experimental evidence suggestive of regulatory function. For each site we determined whether it was supported by additional ChIP–chip evidence for certain histone modifications associated with activation (H3K4me2, H3K4me3, H3K4ac) or factors associated with general chromatin modification (SMARCC1/2, P300 [EP300], Brg1 [SMARCA4]), as well as DNaseI hypersensitivity and nucleosome depletion (Crawford et al. 2006a,b; Sabo et al. 2006; Giresi et al. 2007). For all analysis here, we required pTRRs to have at least one such line of support.

These and other data sets used in this paper are available at http://www.bx.psu.edu/projects/encode_pTRR.

Alignability, background correction, and score comparison

Alignability was computed relative to human coordinates as the fraction of human bases aligning with another species—any position covered by a local alignment is considered aligned, regardless or whether that position is a match, mismatch, or gap. Some of the comparison species are not sequenced completely. To help minimize the effect of unsequenced regions on the alignability calculation, the positions of the aligned sequence blocks were compared with the boundaries of the sequence contigs in the comparison species, and cases of nonaligning segments associated with ambiguous sequence coverage were discarded from the analysis (contiguity was determined by the mafAddIRows program; B. Raney, pers. comm.). For example, if a nonaligning block is flanked by aligning blocks that are adjacent in a contiguous sequence, this case is regarded as a valid, nonaligning segment; however, if a nonaligning block is flanked by ends of separate contig sequences, then it is possible that no sequence is available for the (potential) homolog to the nonaligning block in the comparison species, and the unaligned segment in human is ignored. In addition, blocks spanning poor-quality sequence are also ignored. All other cases were treated as nonaligning blocks (Supplemental Figure 1). Clade assignments resulted from the deepest species with a positive alignability score per region. The species that defined each clade are as follows: vertebrates: zebra fish, *Fugu*, or tetraodon; tetrapods: *Xenopus* or chicken; mammals: platypus or monodelphis; placental mammals: armadillo, cow, dog, elephant, hedgehog, mouse, rabbit, rat, rfbat, shrew, or tenrec; *primates*: chimp, baboon, macaque, marmoset, or galago. Composite alignability was computed as the average of the pairwise alignabilities, weighted by branch length to human.

Correction for constraint scores and alignability (*background correction*) was performed by normalizing the score computed for an interval based on the score computed for the ENCODE that contains it. In the case of constraint scores, where each interval score is actually an average over positions, we divide the interval average by the region average. For pairwise alignability, the total alignability of the ENCODE region is used in the denominator. Background corrected composite alignability was computed as the average of the background corrected pairwise alignabilities was taken, weighted as described above.

Score comparisons and ROC results were performed by calibrating sensitivity and specificity of feature scores versus neutral interval scores. ARs were used to represent neutral intervals. For increased stringency a small number of ARs overlapping MCSs were excluded from this set. Here, we defined sensitivity as the fraction of the feature data set scoring greater than or equal to any given threshold. To evaluate specificity, we defined the fraction of the neutral data set scoring less than any given threshold as the specificity at that threshold. To summarize performance results, a threshold was chosen to equalize the sensitivity and specificity.

GO enrichments

Each pTRR was associated with its inferred target gene from the known genes defined by the UCSC Genome Browser Database (Hinrichs et al. 2006), which was then used to extract the associated gene ontology terms. Enrichment of GO terms associated with elements conserved in a given clade was evaluated under a hypergeometric distribution, using all pTRR elements as the population. Hypergeometric *P*-values were then corrected for multiple testing using the method of Storey and Tibshirani (2003), except that rather than implementing the correction with a postulated null distribution for *P*-values (π_0), a simulation using 1000 random samples was used. The resulting "*q*-values" measure significance in terms of false discovery rate. For example, declaring terms positive if their *q*-value is ≤ 0.05 has an FDR of 5%. Within each clade, distinctly significant terms were defined as those significant in that clade and not in any other clade.

Chromosome conformation capture (3C)

The 3C assay (Dekker et al. 2002) was performed essentially as described by Vakoc et al. (2005). K562 cells were treated with formaldehyde to cross-link proteins to DNA. The cross-linked chromatin was isolated, digested with the restriction endonuclease HindIII, and ligated. Novel ligation junctions, indicative of proximity in chromatin in the cell, were detected by PCR, using one primer at the reference locus (HS2 of the *HBB* locus control region) and second primers near the termini of the fragments indicated in Figure 6. The relative proximity was determined by comparing the results from cellular DNA with control BACs in vitro. The BACs (RP11–910p5 and RP11–680G13) encompass the region of chromosome 11 interrogated in the experiment. The BAC DNA was digested with HindIII and ligated to form a template for PCR that reflects the ligation frequency of the HindIII fragments in free solution. Comparison of the PCR results detecting novel junctions between the cross-linked cells and the DNA in solution gives an enrichment in ligation efficiency that reflects proximity in the nucleus. Band intensities of the PCR product

Downloaded from www.genome.org on September 6, 2007

were quantified with ImageJ software. The primers used were: HS2: GTTTGCTTAGAAGGTTACAGAACCAGAAGG; HBE: CCATTGTATCTGTCCCCTTGAATCATCATCC; HBG1: AAGCCTGCACCTCAGGGGTGAATTCTTTG; 67 kb: CATGGTTCAGAGAAAAATCCATAACAACATCAAG; 60 kb: GTTCCTTCTCAACATCTGTGAAGAGAAGCA; 93 kb: TTTCAGTTTTATCTGTCAAGAGCAAAATTTGAG; 110 kb: GATTTTCGCTCACTACCAGGCCTTGGGATG; 229 kb: TGCAAACAAGGATCTAGTCTGAGATCCCAAG; 231 kb: TCTTCATGCATCATGAAATAATCTTGGAGCCAG.

Acknowledgments

This work was supported by NIH grants from NHGRI (HG002238, W.M.) and NIDDK (DK65806, R.H.), by Tobacco Settlement Funds from the Pennsylvania Department of Health, and by the Huck Institutes of Life Sciences, The Pennsylvania State University.

References

Aldred, P.M., Hollox, E.J., and Armour, J.A. 2005. Copy number polymorphism and expression level variation of the human alpha-defensin genes *DEFA1* and *DEFA3*. *Hum. Mol. Genet.* **14:** 2045–2052.

Altschuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J., and Donnelly, P. 2005. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Andolfatto, P. 2005. Adaptive evolution of noncoding DNA in *Drosophila*. *Nature* **437:** 1149–1152.

Aparicio, S., Morrison, A., Gould, A., Gilthorpe, J., Chaudhuri, C., Rigby, P., Krumlauf, R., and Brenner, S. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc. Natl. Acad. Sci.* **92:** 1684–1688.

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. *Nat. Genet.* **25:** 25–29.

Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, W.J., and Haussler, D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441:** 87–90.

Bieda, M., Xu, X., Singer, M.A., Green, R., and Farnham, P.J. 2006. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res.* **16:** 595–605.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.

Carter, D., Chakalova, L., Osborne, C.S., Dai, Y.F., and Fraser, P. 2002. Long-range chromatin regulatory interactions in vivo. *Nat. Genet.* **32:** 623–626.

Chiaromonte, F., Weber, R.J., Roskin, K.M., Diekhans, M., Kent, W.J., and Haussler, D. 2003. The share of the human genome under selection estimated from human-mouse genomic alignments. *Cold Spring Harbor Symp. Quant. Biol.* **68:** 245–254.

Chuang, J.H. and Li, H. 2004. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol.* **2:** doi: 10.1371/journal.pbio.0020029.

Cooper, S.J., Trinklein, N.D., Anton, E.D., Nguyen, L., and Myers, R.M. 2006. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* **16:** 1–10.

Crawford, G.E., Davis, S., Scacheri, P.C., Renaud, G., Halawi, M.J., Erdos, M.R., Green, R., Meltzer, P.S., Wolfsberg, T.G., and Collins, F.S. 2006a. DNase-chip: A high-resolution method to identify DNaseI hypersensitive sites using tiled microarrays. *Nat. Methods* **3:** 503–509.

Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., Davis, S., Margulies, E.H., Chen, Y., Bernat, J.A., Ginsburg, D., et al. 2006b. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16:** 123–131.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. 2002. Capturing chromosome conformation. *Science* **295:** 1306–1311.

Dermitzakis, E.T. and Clark, A.G. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. *Mol. Biol. Evol.* **19:** 1114–1121.

Dermitzakis, E.T., Reymond, A., and Antonarakis, S.E. 2005. Conserved non-genetic sequences—an unexpected feature of mammalian genomes. *Nat. Rev. Genet.* **6:** 151–157.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. 2006. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16:** 1299–1309.

Eddy, S.R. 2005. A model of the statistical power of comparative genome analysis. *PLoS Biol.* **3:** e10.

Elnitski, L., Miller, W., and Hardison, R.C. 1997. Conserved E-boxes function as part of the enhancer in hypersensitive site 2 of the beta-globin locus control region: role of basic helix-loop-helix proteins. *J. Biol. Chem.* **272:** 369–378.

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

Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I., and Hardison, R.C. 2003. Cross-species sequence comparisons: A review of methods and available resources. *Genome Res.* **13:** 1–12.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. 2007. FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* (this issue) doi: 10.1101/gr.5533506.

Gumucio, D., Shelton, D., Zhu, W., Millinoff, D., Gray, T., Bock, J., Slightom, J., and Goodman, M. 1996. Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching program of the beta-like globin genes. *Mol. Phylog. Evol.* **5:** 18–32.

Hauf, S., Roitinger, E., Koch, B., Dittrich, C.M., Mechtler, K., and Peters, J.M. 2005. Dissociation of cohesin from chromosome arms and loss of arm cohesion during early mitosis depends on phosphorylation of SA2. *PLoS Biol.* **3:** doi: 10.1371/journal.pbio.0030069.

Hardison, R.C., Roskin, K., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Co-variation in divergence by substitution, deletion, transposition and recombination during mammalian evolution. *Genome Res.* **13:** 13–26.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. 2006. The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Res.* **34:** D590–D598.

Hughes, J.R., Cheng, J.-F., Ventress, N., Prabhakar, S., Clark, K., Anguita, E., De Gobbi, M., de Jong, P., Rubin, E., and Higgs, D.R. 2005. Annotation of *cis*-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proc. Natl. Acad. Sci.* **102:** 9830–9835.

Keightley, P.D., Lercher, M.J., and Eye-Walker, A. 2005. Evidence or widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3:** e42.

Kim, J., Bhinge, A.A., Morgan, X.C., and Iyer, V.R. 2005. Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat. Methods* **2:** 47–53.

King, D., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W., and Hardison, R.C. 2005. Evaluation and comparison of conservation and regulatory potential scores for detecting *cis*-regulatory modules in aligned mammalian genome sequences. *Genome Res.* **15:** 1051–1060.

Li, J. and Miller, W. 2003. Significance of interspecies matches when evolutionary rate varies. *J. Comput. Biol.* **10:** 537–554.

Loots, G.G., Locksley, R., Blankespoor, C., Wang, Z.-E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.

Ludwig, M.Z., Patel, N.H., and Kreitman, M. 1998. Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: Rules governing conservation and change. *Development* **125:** 949–958.

Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., Hou, M., et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* (this issue) doi: 10.1101/gr.6034307.

McDonald, J.H. and Kreitman, M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351:** 652–654.

Miller, W., Makova, K., Nekrutenko, A., and Hardison, R.C. 2004. Comparative genomics. *Ann. Rev. Genomics Human Genet.* **5:** 15–56.

Nobrega, M.A., Ovcharenko, I., Afzal, V., and Rubin, E.M. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302:** 413.

Nobrega, M.A., Zhu, Y., Plajzer-Frick, I., Afzal, V., and Rubin, E.M. 2004.

Downloaded from www.genome.org on September 6, 2007

Megabase deletions of gene deserts result in viable mice. *Nature* **431:** 988–993.
Omasu, F., Ezura, Y., Kajita, M., Ishida, R., Kodaira, M., Yoshida, H., Suzuki, T., Hosoi, T., Inoue, S., Shiraki, M., et al. 2003. Association of genetic variation of the *RIL* gene, encoding a PDZ-LIM domain protein and localized in 5q31.1, with low bone mineral density in adult Japanese women. *J. Hum. Genet.* **48:** 342–345.
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.
Rand, D.M. and Kann, L.M. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol. Biol. Evol.* **13:** 735–748.
Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309.
Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A., et al. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods* **3:** 511–518.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human-mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.
Siepel, A., Bejerano, G., Pederson, J.S., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, J., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
Storey, J.D. and Tibshirani, R. 2003. Statistical significance for genome wide studies. *Proc. Natl. Acad. Sci.* **100:** 9440–9445.
Taylor, J., Tyekucheva, S., King, D.C., Hardison, R.C., Miller, W., and Chiaromonte, F. 2006. ESPERR: learning strong and weak signals in functional elements. *Genome Res.* **16:** 1596–1604.
Thomas, D.J., Rosenbloom, K.R., Clawson, H., Hinrichs, A.S., Trumbower, H., Raney, B.J., Karolchik, D., Barber, G.P., Harte, R.A., Hillman-Jackson, J., et al. 2007. The ENCODE Project at UC Santa Cruz. *Nucleic Acids Res.* **35:** D663–D667.
Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol. Cell* **10:** 1453–1465.
Vakoc, C.R., Letting, D.L., Gheldof, N., Sawado, T., Bender, M.A., Groudine, M., Weiss, M.J., Dekker, J., and Blobel, G.A. 2005. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Mol. Cell* **17:** 453–462.
Valverde-Garduno, V., Guyot, B., Anguita, E., Hamlett, I., Porcher, C., and Vyas, P. 2004. Differences in the chromatin structure and *cis*-element organization of the human and mouse GATA1 loci: Implications for *cis*-element identification. *Blood* **104:** 3106–3116.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al. 2005. Highly conserved noncoding sequences are associated with vertebrate development. *PLoS Biol.* **3:** e7. doi: 10.1371/journal.pbio.0030007.

Received June 2, 2006; accepted in revised form March 7, 2007.