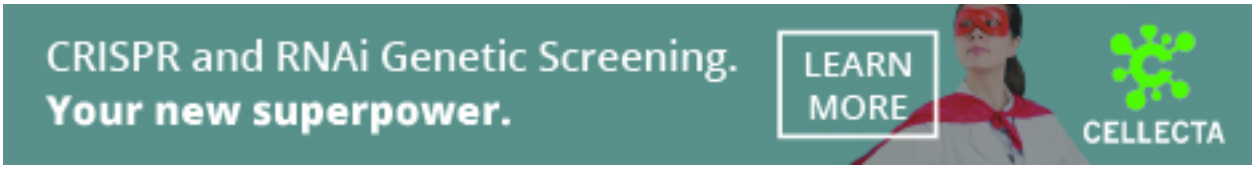
Distinguishing Regulatory DNA From Neutral Sites

Laura Elnitski, Ross C. Hardison, Jia Li, et al.

Genome Res. 2003 13: 64-72
Access the most recent version at doi:10.1101/gr.817703



Letter

Distinguishing Regulatory DNA From Neutral Sites

Laura Elnitski,[1,3] Ross C. Hardison,[1] Jia Li,[2] Shan Yang,[1] Diana Kolbe,[1,3] Pallavi Eswara,[3] Michael J. O'Connor,[3] Scott Schwartz,[3] Webb Miller,[3,4] and Francesca Chiaromonte[2,5,6]

[1]Departments of Biochemistry and Molecular Biology, [2]Statistics, [3]Computer Science and Engineering, [4]Biology, and [5]Health Evaluation Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

We explore several computational approaches to analyzing interspecies genomic sequence alignments, aiming to distinguish regulatory regions from neutrally evolving DNA. Human–mouse genomic alignments were collected for three sets of human regions: (1) experimentally defined gene regulatory regions, (2) well-characterized exons (coding sequences, as a positive control), and (3) interspersed repeats thought to have inserted before the human–mouse split (a good model for neutrally evolving DNA). Models that potentially could distinguish functional noncoding sequences from neutral DNA were evaluated on these three data sets, as well as bulk genome alignments. Our analyses show that discrimination based on frequencies of individual nucleotide pairs or gaps (i.e., of possible alignment columns) is only partially successful. In contrast, scoring procedures that include the alignment context, based on frequencies of short runs of alignment columns, dramatically improve separation between regulatory and neutral features. Such scoring functions should aid in the identification of putative regulatory regions throughout the human genome.

Because relatively few mammalian genes are well-characterized, annotation of genes in these large, complex genomes has largely relied on powerful ab initio programs such as GENSCAN (Burge and Karlin 1997) and on evidence-based methods such as EST database searches using local alignment tools like BLAST (Altschul 1997). Indeed, application of these two approaches, often using several distinct implementations of the ideas, has allowed dramatic, albeit imperfect, progress in identifying putative genes. Although the success of ab initio programs is dependent on a reasonably complete model of the structure of mammalian genes, their predictions can be overlapped with those of evidence-based methods to increase accuracy (e.g., GrailEXP; http://compbio.ornl.gov/grailexp/; Xu and Uberbacher 1997).

In principle, both ab initio and evidence-based approaches can be explored for identifying putative regulatory elements. An example of the latter is the search for clusters of particular transcription-factor binding sites (Berman et al. 2002; Jegga et al. 2002). As for the former, because a reliable model for regulatory elements has not yet been constructed, one must seek alternative strategies. In many studies of discrete loci, highly conserved noncoding sequences have proven to be good indicators of regulatory elements (e.g., Hardison et al. 1997b; Loots et al. 2000). However, not all regulatory elements are uniquely identified by human–mouse alignments (Flint et al. 2001), and the regional variation in evolutionary rates in mammals precludes finding a single criterion that distinguishes regulatory regions from neutral DNA genome-wide (Hardison 2000; Pennacchio et al. 2001; Hardison et al. 2003). Thus, scoring procedures that evaluate alignments for properties other than overall percent identity need to be developed to test the effectiveness of interspecies alignments as ab initio predictors of regulatory regions.

[6]Corresponding author.
E-MAIL chiaro@stat.psu.edu; FAX (814) 863-7114.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.817703.

A critical requirement for developing such procedures is a collection of well-characterized, experimentally determined regulatory regions. Alignments that overlap such a collection can then be evaluated quantitatively by a variety of scoring procedures, and their score values compared with those of coding sequences and neutrally evolving DNA. Here we report on the compilation of a regulatory region data set, and test the efficacy of simple and more sophisticated alignment-scoring schemes for distinguishing regulatory from neutral DNA. These studies were made possible by the availability of high-quality draft genome sequences of human (Lander et al. 2001) and mouse (Waterston et al. 2002), and of a collection of highly sensitive, specific alignments of the two genomes (Schwartz et al. 2003; Waterston et al. 2002).

We show that the power of high-scoring alignments as predictors of regulatory regions, previously demonstrated only for relatively small genomic loci, can be evaluated systematically by comparison with a good model for neutral DNA, that is, ancestral repeats that are relics of transposons active before the human–mouse split but defunct since the radiation (Lander et al. 2001; Waterston et al. 2002). The most successful analyses include the context of the alignment in the scoring procedure. This approach will be extended to whole-genome alignments between human and mouse, and the results will be made available as an online resource at http://bio.cse.psu.edu/.

RESULTS

Types of Data

In our analyses, we considered four classes of DNA segments aligned between human and mouse. The first is a collection of known regulatory regions, experimentally defined and trimmed to the smallest functional unit from which no further deletions can be made without reducing activity. The 95 sequences in this collection vary in length from 62–2973 bp. Only regulatory regions that aligned between human and

13:64–72 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

mouse were included (2 of the 95 regions were excluded because sequences were missing from the mouse assembly). However, no threshold for the amount of aligning DNA within a regulatory region was applied.

The second class is a large collection (1,400,000) of orthologous ancestral repeats that align between human and mouse. These ancestral repeats transposed prior to the rodent–primate split and are no longer active. Because no selective pressure (either positive or negative) is applied to these sites as they are accumulating mutations, they serve as a template for neutral evolution (Waterston et al. 2002).

The third class is a collection of coding regions derived from annotated human exons in RefSeq (Pruitt et al. 2000), as obtained from the Human Genome Browser (HGB; Kent et al. 2002). This serves as a positive control in our experiments, as it is a set of known functional elements whose alignments are clearly distinguishable from neutral DNA (Batzoglou et al. 2000). It should be noted that alignments located within the exons contribute to <2% of all alignments (Waterston et al. 2002).

The fourth class is a collection of human–mouse alignments comprising all genomic DNA (omitting exons), which we refer to as bulk DNA.

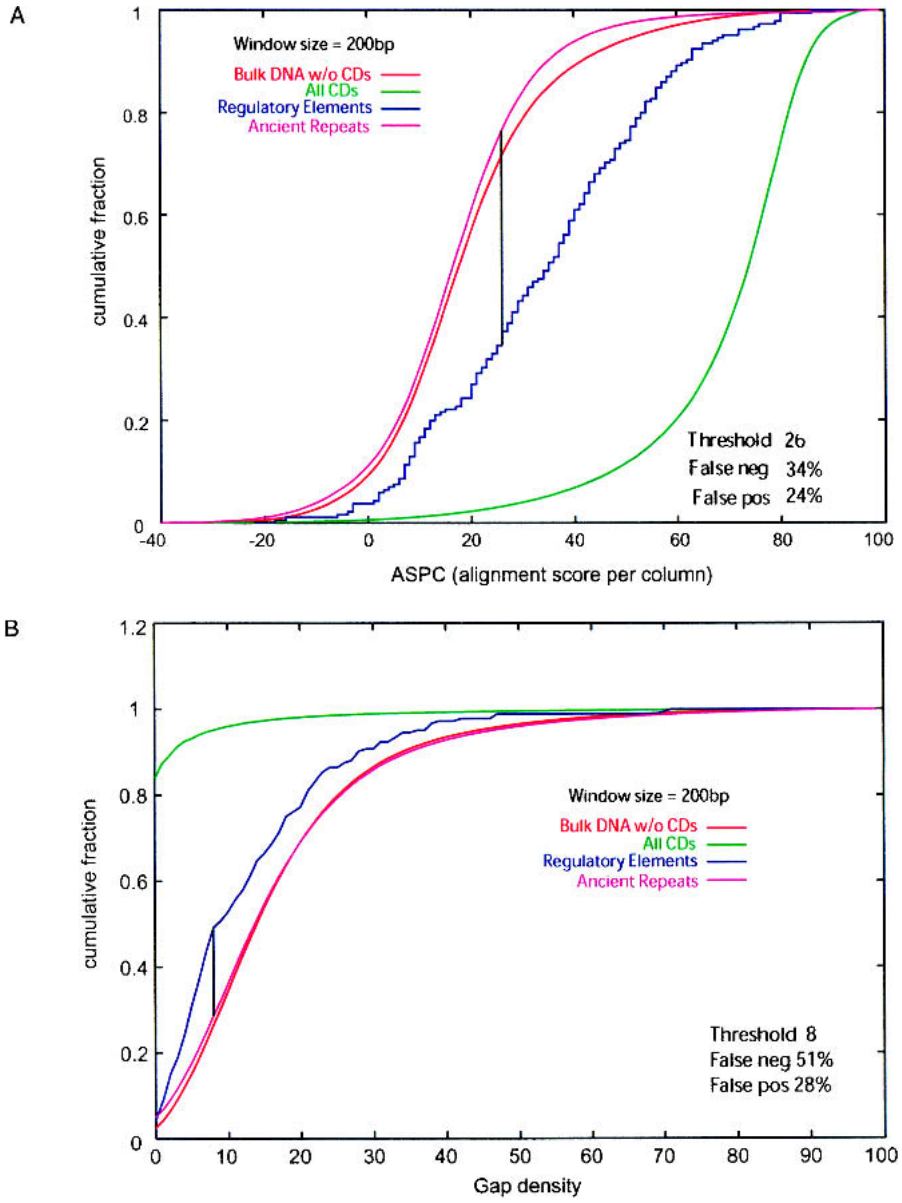


Figure 1 (*A*) Cumulative distributions of ASPC (alignment score per column) in 200-bp nonoverlapping windows from regulatory elements, ancient repeats, coding exons (cds), and bulk DNA alignments. The ASPC is calculated using the BLASTZ scoring scheme with a penalty for gaps. The vertical line represents the ASPC value at which regulatory element and ancient repeat distributions intersect (i.e., maximal distance between cumulative distributions). With this as a threshold, one obtains a certain percentage of false positives (ancient repeats above the threshold) and false negatives (regulatory elements above the threshold). (*B*) Cumulative distributions of gap density in 200-bp nonoverlapping windows from regulatory elements, ancient repeats, coding exons (cds), and bulk DNA alignments. The vertical line and percentages of false positives and false negatives are obtained as for ASPC in *A*, except that here false positives are ancient repeats below the threshold, and false negatives are regulatory elements above it.

Discrimination Based on Individual Pairing Frequencies

Human–mouse alignments (Schwartz et al. 2003; Waterston et al. 2002) in each of the four classes were partitioned into nonoverlapping windows of size 200 bp. Within each window, we computed the alignment score per column (ASPC), based on the BLASTZ scoring scheme (Schwartz et al. 2003), and the density of gaps.

The aligning program BLASTZ uses a scoring matrix (Chiaromonte et al. 2002) for the 16 possible pairings of *A*, *C*, *G*, and *T*, along with affine gap penalties (i.e., open gap and gap-extension penalties). The ASPC in a window is simply the sum of the frequencies of the 16-symbol alphabet comprising *A*, *C*, *G*, *T* pairings multiplied by the coefficients in the BLASTZ matrix, along with a special treatment for gaps (see Methods). A graph of the cumulative distribution of the ASPC for the four types of DNA shows that this score cleanly separates coding exons from ancestral repeats and bulk DNA, but fails to separate regulatory regions (Fig. 1A). In particular, the overlap between the ASPC distribution for the regulatory and ancestral repeat classes is substantial.

Generalizing the spirit of the ASPC, we consider linear combinations of frequencies of a 17-symbol alphabet comprising all *A*, *C*, *G*, *T* pairings plus an additional symbol for gaps. Gap density itself is one such combination, with the gap

frequency having a coefficient of 1, and each other symbol frequency a coefficient of 0. As seen in Figure 1B, gap density clearly separates coding exons, but leaves a substantial overlap between regulatory and ancestral repeats classes.

The effectiveness of these combinations was investigated systematically by constructing a data cloud of 693 points (from the four basic alignment collections) in 17 dimensions. The points correspond to the 93 alignments in the regulatory elements collection, plus 200 alignment segments of size 200 bp randomly selected from each of the ancestral repeats, exons, and bulk DNA collections. The dimensions correspond to 17-symbol frequencies computed on such alignments (see Methods section for details).

We applied Principal Component Analysis (PCA) to ascertain the variability structure of alignments in these dimensions. The first principal component is the direction of maximal variability of the data, to which each original coordinate (17-symbol frequencies) participates via linear combination coefficients, and it explains a certain share of the overall variance. The second principal component is the direction, orthogonal to the first, which has maximal variability after that, and explains a certain share of the overall variance, and so on (the coefficients for the first and second principal components are reported in Table 1 and Fig. 2C). The first principal plane is the span of the first two principal components, and it captures a share of the overall variance equal to the sum of the two component shares. If this sum is high, the orthogonal projection of the data cloud onto the first principal plane provides a good low-dimensional approximation of the data structure.

In our case, the first principal component plane explains 82% of the overall cloud variability. Projecting the data onto this plane gives a boomerang-like shape (Fig. 2A). The first principal component captures a tradeoff between gaps and matches, with *C* and *G* matches weighing more than *A* and *T* matches. The second principal component captures a tradeoff between *C* and *G* matches on one side, and *A* and *T* matches on the other, with gaps playing a nonnegligible but much smaller role (see Table 1 and Fig. 2C). From the vantage point of the principal plane, simple linear combinations of individual paring frequencies seem to not provide good discrimination of genomic features: One arm of the boomerang cloud is dominated by neutral and bulk DNA, whereas the other arm is dominated by exons and regulatory regions, but nonfunctional and functional features have a sizeable overlap at the convergence of the two arms. Next, we applied Sliced Inverse Regression (SIR) to identify linear combinations of high discriminatory power. This analysis uses the same data cloud used in PCA plus the known classification of each point (regulatory region, coding region, ancestral repeat, or bulk DNA), which plays the role of a categorical response variable. When used for categorical responses, SIR (Li 1991; Cook 1998 and references therein) is a close relative of traditional discriminant analysis. The first SIR direction aims at maximal relevance to the classification, with each original coordinate participating via linear combination coefficients. Subsequent orthogonal SIR directions are progressively less relevant for the classification (coefficients for the first SIR direction are reported in Table 1 and Fig. 2C). The first combination identified by SIR (see Fig. 2B) exhibits a relatively small variation range on the data (standard deviation 0.038, whereas that for PCA1 is 0.157 and for PCA2 0.088). In fact, it is well outside the first principal plane (maximal data variability). A projection of SIR1 on the first principal plane is shown in Figure 2A. However, as can be seen in Figure 2B, SIR1 succeeds in separating both regulatory elements and exons from ancestral repeats and bulk DNA. The overlap between the distributions for the regulatory and ancestral repeat classes is still nonnegligible, but much reduced with respect to the overlaps presented by ASPC and gap density. The coefficients for the 17-symbol frequencies (Table 1, Fig. 2C) suggest a tradeoff, having matches (in particular, *C* and *G* matches) and some mismatches (in particular, *CA*, *CG*, and *GC*) on one side, and other mismatches (in particular, *AG*, *GT*, and *TG*) on the other side (subsequent SIR directions are not discussed here).

Interestingly, the broad features revealed by PCA and SIR are robust when the 17-symbol alphabet is collapsed into a smaller set. For instance, the boomerang shape in the first principal plane and the degree of class separation along SIR1 remain similar when these linear analyses are applied to the 4-dimensional data cloud obtained by collapsing the alphabet into *M* (matches), *T* (*C* to/from *T* and *G* to/from *A*), *V* (*A* to/from *C*, *A* to/from *T*, *G* to/from *C*, and *G* to/from *T*), and gaps (results not shown). Thus it is reasonable to examine smaller alphabets, but now considering scores taking into account the context of the alignments.

Table 1. Coefficients of the Linear Combinations Expressing First and Second Principal Components, and First SIR Direction

Pairing_freq	PCA1	PCA2	SIR1
faa_200	**−0.154958**	**−0.475658**	0.111993
fac_200	0.000931	−0.033166	−0.113294
fag_200	−0.006815	−0.080282	−0.441445
fat_200	0.002420	−0.053753	−0.088708
fca_200	−0.000446	−0.004373	0.239844
fcc_200	**−0.265820**	**0.577022**	0.238753
fcg_200	−0.004137	0.018679	0.269244
fct_200	−0.012399	−0.009786	−0.096109
fga_200	−0.014155	−0.014012	−0.034745
fgc_200	−0.006835	0.021381	0.332223
fgg_200	**−0.268567**	**0.476319**	0.301653
fgt_200	−0.003292	−0.005356	−0.480595
fta_200	0.008541	−0.051466	−0.043727
ftc_200	−0.002311	−0.071793	−0.012283
ftg_200	−0.000159	−0.035737	−0.348887
ftt_200	**−0.168724**	**−0.411842**	0.133115
fgap_200	**0.896756**	**0.153753**	0.032474

Columns in the table are eigenvectors from spectral decompositions of appropriate variance/covariance matrices. Thus, each has norm 1 (the squares of the coefficients add up to 1), and PCA1 and PCA2, which come from the same decomposition, are orthogonal (the cross products add up to 0).

Discrimination Based on Frequencies of Short Pairing Patterns

The first such score was motivated by the observation that many transcription factors bind to a cognate

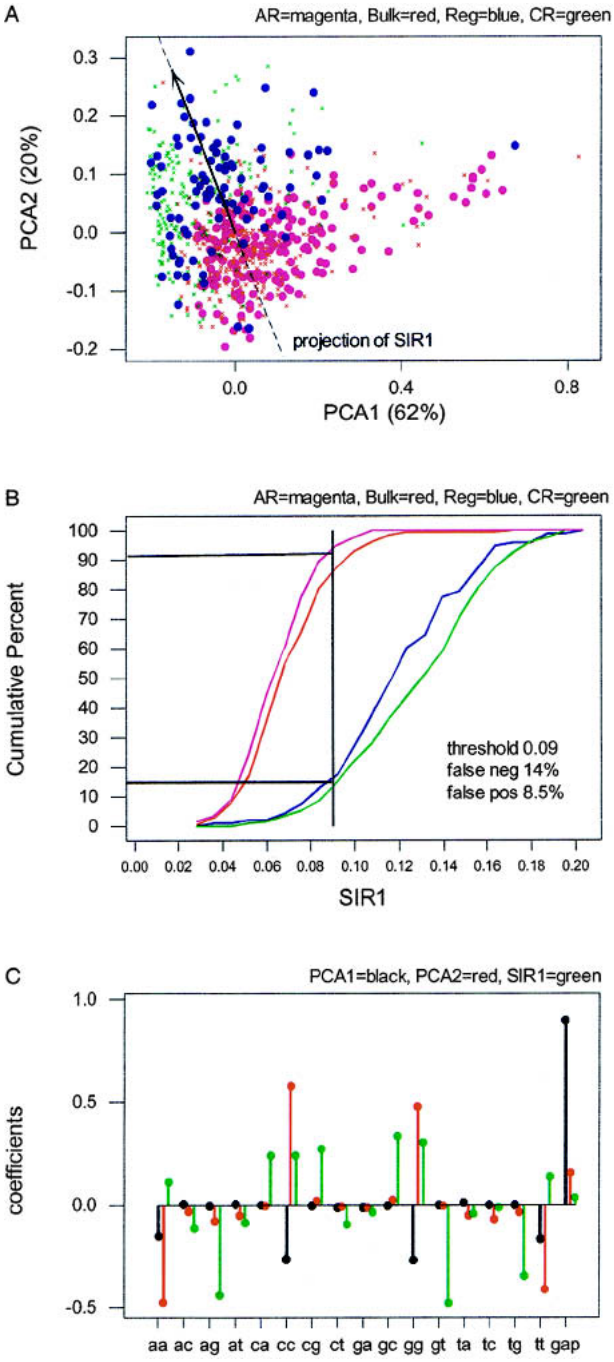


Figure 2 First principal plane projection for frequencies on a 17-symbol alphabet comprising all *A, C, G, T* pairings plus an additional symbol for gaps. The data cloud contains 93 regulatory elements (Reg), plus 200 alignment segments of size 200 bp randomly selected from each of ancient repeats (AR), coding exons (CR), and bulk DNA (shown as different marks). Percentages of explained variability are reported for the first and second principal component (total for the plane, 82%). The black line is a projection of SIR1 (see *B*) on the first principal plane. (*B*) Cumulative distributions of SIR1 (first Sliced Inverse Regression linear combination) for frequencies on the 17-symbol alphabet. The distributions concern 93 regulatory elements (Reg), plus 200 alignment segments of size 200 bp randomly selected from each of ancient repeats (AR), coding exons (CR), and bulk DNA. The vertical line and percentages of false positives (ancient repeats above the threshold) and false negatives (regulatory elements below the threshold) are obtained as for ASPC in Figure 1A. (*C*) Coefficients of the linear combinations expressing first (black) and second (red) principal components, and first SIR direction (green). These are eigenvectors from spectral decompositions of appropriate variance/covariance matrices (see Methods and Table 1). Thus, each has norm 1 vector (the squares of the coefficients add up to 1), and PCA1 and PCA2, which come from the same decomposition, are orthogonal (the cross products add up to 0).

DNA sequence of ~6 bp. Clusters of transcription-factor binding sites tend to comprise known enhancers (e.g., Berman et al. 2002; Jegga et al. 2002), with up to 68% of the sites being conserved between humans and rodents (Dermitzakis and Clark 2002). We identified occurrences of exact hexamer matches by sliding a 6-bp window along alignments, 1 bp at a time. Then, partitioning whole-genome alignments into 200-bp nonoverlapping windows, we measured the density of exact hexamer matches in the four classes of DNA examined. As seen in Figure 3A, this score does better than ASPC and gap density, but worse than SIR1, in regard to the overlap between the distributions for the regulatory and ancestral repeat classes. Like SIR1, the density of exact hexamer matches does not distinguish regulatory from coding regions, but this is to be expected because the latter generally align without gaps (Makalowski et al. 1996).

Next, we compute more complex context-embedding scores as (properly normalized) log-odds ratios from Markov models. These models are formulated for increasingly finer underlying alphabets of matches, mismatches, gaps, and higher orders. The order expresses how many preceding contiguous positions are considered in modeling the probability of each symbol at a given location. For this analysis, we restrict attention to the regulatory versus neutral discrimination, and use the 93 regulatory segments plus 200 ancestral repeat segments of size 200 bp already used for the PCA and SIR analyses.

Table 2 reports percentages of overlap for various alphabets and orders. In accordance with our observations of the collapsibility of the 17-symbol alphabet, and the importance of length-6 patterns for regulatory sequence, we reach an excellent discriminatory performance when using a 5-symbol alphabet comprising M_{AT} (matches of *A*s and *T*s), M_{GC} (matches of *G*s and *C*s), *V*, *T* and gaps, and fifth order. This score outperforms our best score based on frequencies of individual pairings (SIR1), and it completely eliminates the overlap between the distributions for the regulatory and ancestral repeat classes (dark blue and magenta cumulative distribution functions in Fig. 3B).

In Figures 1 through 3A and Table 2 we provide false-negative (regulatory elements that can be mistaken for ancestral repeats) and false-positive (ancestral repeats that can be mistaken for regulatory elements) percentages associated with the various scores we considered. These percentages concern the regulatory and ancestral repeats segments used in training, and exploit the natural thresholds defined by the score value at which regulatory element and ancestral repeat distributions intersect. The 5-symbol fifth-order Markov model log-odds score, as well as log-odds scores derived from larger alphabets, present 0% false negatives and 0% false positives because the two distributions do not overlap. However, because of the fairly limited collection of regulatory elements

presently available for training, we went a step further and assessed performance robustness for our 5-symbol fifth-order log-odds score using two different cross-validation schemes. The results reported in Table 3 are very encouraging given the small size of our training data: Erroneous classification percentages (false positive and false negatives) are ~6%, and ambiguous classification percentages (score range between the two nonoverlapping distributions) remain about or well below 20%.

Although our 5-symbol fifth-order log-odds score expression is derived based only on regulatory and ancestral repeat data, it can be used to score any alignment segment. As additional controls, we scored the sets of 200 randomly selected segments of length 200 bp from coding regions and bulk DNA previously used in PCA and SIR, as well as sets of 200 randomly selected segments of length 200 bp from 3′- and 5′-UTRs. The corresponding cumulative distribution functions are shown in green, bright blue, orange, and purple in Figure 3B. Interestingly, the regulatory distribution is similar in shape, but shifted to the right (higher values) of that for coding regions. Thus, unlike other scores considered in our study, the 5-symbol fifth-order log-odds score may be successfully capturing something other than simple conservation, such as pairing patterns more common in regulatory than in coding DNA.

The distribution of log-odds scores for alignments in bulk DNA (noncoding, aligned DNA) is similar in shape to that of ancestral repeats, and although shifted to the right, still has minimal overlap with the regulatory elements distribution (Fig. 3B, shown in bright blue). Thus, our score distinguishes regulatory regions not only from ancestral repeats but also from bulk DNA.

The 3′- and 5′-UTR distributions are similar to one another and cover a range in values similar to that for the regulatory distribution. The distributions of UTR scores are more concentrated (steeper cumulative distribution functions) and slightly shifted to the right of the regulatory region distribution. Thus, our score also detects UTRs. Indeed, regulatory elements can reside in or overlap with untranslated regions. The fact that the UTR distributions overlap that of regulatory elements and both score higher than do the coding regions shows that sequence conservation is not the only nor the prevailing determinant for our score.

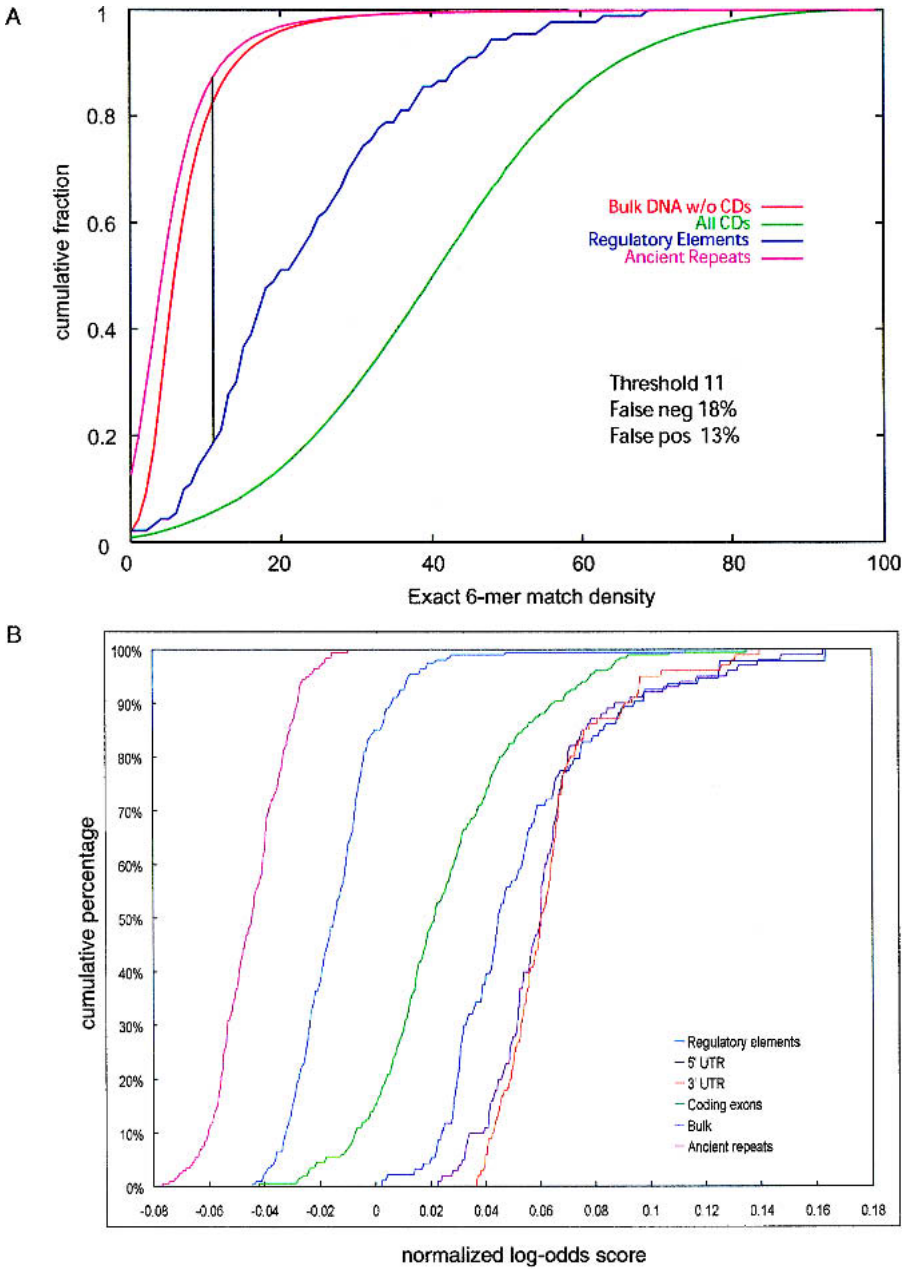


Figure 3 (*A*) Cumulative distributions of exact hexamer matches density in 200-bp nonoverlapping windows from regulatory elements, ancient repeats, exons, and bulk DNA alignments. The density is calculated by scrolling over 6-nt sequences with no gaps in each window. Vertical line and percentages of false positives (ancient repeats above the threshold) and false negatives (regulatory elements below the threshold) are obtained as for ASPC in Figure 1A. (*B*) Cumulative distributions of (normalized) log-odds score from fifth-order 5-symbol alphabet Markov Models. The score expression is derived based on 93 regulatory elements and 200 alignment segments of size 200 bp randomly selected from ancient repeats. The cumulative distributions for these are shown in dark blue and magenta, respectively. Because the distributions do not intersect, any threshold between the maximum score value for ancient repeats and the minimum score value for regulatory elements guarantees 0% false positives and 0% false negatives. The green, orange, purple, and bright blue cumulative distributions are obtained applying the score expression to segments from coding regions, UTRs, and bulk DNA.

Table 2. False-Negative and False-Positive Percentages for (Normalized) Log-Odds Scores From Markov Models, for Various Orders and Alphabets

	Order							
	1		2		3		5	
Alphabet	FN	FP	FN	FP	FN	FP	FN	FP
2	20.4	13	19.4	9.5	24.7	5.0	14.0	12.5
3	23.7	9.5	11.8	18.5	16.1	10.5	9.7	6.0
4	21.5	11.0	19.4	9.0	12.9	10.0	1.1	3.0
5a	9.7	12.5	9.7	7.0	10.8	4.0	0	0
5b	23.7	6.5	21.5	5.0	8.6	7.0	0	0
7a	11.8	12.5	8.6	8.0	4.3	3.5	0	0
7b	16.1	7.0	8.6	6.0	3.2	0.5	0	0

Percentages are obtained as for ASPC in Figure 1A. Alphabets considered here are: (2) match, other; (3) match, mismatch, gap; (4) match, transition, transversion, gap, (5a) match (*A* or *T*), match (*C* or *G*), transition, transversion, gap; (5b) match (*A* or *G*), match (*C* or *T*), transition, transversion, gap; (7a) match (*A* or *T*), match (*C* or *G*), transition, transversion (*A*–*T*), transversion (*C*–*G*), transversion (other), gap; (7b) match (*A*), match (*C*), match (*G*), match (*T*), transition, transversion, gap.
Fifth order models on alphabets of 5 or more symbols give scores for which regulatory elements and ancestral repeats distributions do not intersect (any threshold between the maximum score value for ancestral repeats and the minimum score value for regulatory elements guarantees 0% false positive and 0% false negatives).

The 5-symbol fifth-order log-odds score can be computed on any segment of human genomic DNA aligned with mouse to produce an index of regulatory potential in comparison to neutral behavior. We are presently implementing software that will provide a regulatory potential index at each site in the human genome (aligned with mouse) as the log-odds score for the 200-bp window centered at the site (available at http://bio.cse.psu.edu/).

DISCUSSION

A lower bound on the fraction of the human genome likely under selection is ~5%–6% (Waterston et al. 2002; F. Chiaromonte, R. Weber, K.M. Roskin, M. Diekhans, J. Kent, and D. Haussler, in prep.). Only ~1.5% of the genome codes for protein, leaving at least 3.5% under selection for some other function. One of the critical types of noncoding but functional sequence that one wishes to find are gene regulatory elements, such as promoters, enhancers, silencers, and boundary elements. Some of these are clearly conserved among mammalian species (Hardison 2000; Pennacchio et al. 2001), but methods for cleanly distinguishing them from other genomic sequence alignments have not been systematically examined.

Some previous criteria applied for identifying potential regulatory elements have been noncoding alignments in otherwise divergent regions (Hardison et al. 1997a; Hardison et al. 2000) or a combination of some minimal length and percent identity, such as at least 100 bp of gap-free alignment and >75% identity (Loots et al. 2000). These approaches have been successful for certain genomic regions (Jackson et al. 1996; Elnitski et al. 1997; Loots et al. 2000), but not all known regulatory elements fit those criteria (e.g., Flint et al. 2001). One of the major obstacles to applying a single criterion for potential regulatory regions genome-wide is the substantial variation in the underlying mutation rates from region to region (Hardison et al. 2002; Waterston et al. 2002). Conservation scores that incorporate the local neutral substitution rate can be used to compute a likelihood of a particular sequence being under selection, and these are now available genome-wide (Waterston et al. 2002). Thus one can determine if a particular sequence is likely to be functional, but these conservation scores do not address the type of function for each sequence.

We have performed several discrimination analyses, comparing the behavior of two types of functional sequences (regulatory and coding regions) with neutral DNA (ancestral repeats) and bulk DNA in alignments between human and

Table 3. Results of Two Cross-Validation Schemes for the Log-Odds Score Obtained From 5-Symbol (Match of *A* or *T*, Match of *G* or *C*, Transition, Transversion, Gap) Fifth Order Markov Models

Cross-validation scheme	Reg. correct	Reg. ambiguous	Reg. erroneous	Anc. rep. correct	Anc. rep. ambiguous	Anc. rep. erroneous
5–5 (100)	81.4	12.2	6.4	73	21	6
Leave-one-out	78.49	15.05	6.452	72.5	21	6.5

Because the regulatory and ancestral repeats distributions do not overlap, instead of reporting correct and erroneous classification rates relative to an arbitrary threshold, we list three percentages; namely, correct classifications, ambiguous cases (falling between the two nonoverlapping distributions), and erroneous classifications. In the first cross-validation scheme, we withhold from training and then classify five regulatory elements and five ancestral repeat segments selected at random. This procedure is repeated 100 independent times, and correct, ambiguous, and erroneous classification percentages are obtained averaging over these replications. The second scheme is a leave-one-out cross-validation, in which each data point in turn is withheld from training and then classified.

mouse. Scores based on individual pairing frequencies, such as ASPC, gap density, and leading PCA and SIR linear combinations, do separate different classes of aligned segments. However, even when selected for maximal discrimination (SIR analysis), these scores leave a substantial overlap between regulatory and neutral DNA, and remain higher for coding than for regulatory DNA.

More complex procedures, which exploit short pairing patterns characteristics of regulatory DNA through the use of Markov models, allow a much higher score separation between regulatory elements and ancestral repeats, and eventually score regulatory elements higher than coding regions. Generally, score separation between regulatory elements and ancestral repeats improves as the order of the Markov models and the size of the underlying alphabet increase. We obtain excellent results when using fifth-order (which captures length-6 patterns), and a 5-symbol alphabet that distinguishes between *A*/*T* and *G*/*C* matches.

Functional regions, and among them known regulatory regions, tend to be higher in GC content than other DNA (Waterston et al. 2002). Also our SIR analysis provides some evidence that alignment pairs preserving GC content are favored in functional regions. SIR1, in which both regulatory and coding regions score higher than neutral and bulk DNA, has positive coefficients for *CC* and *GG* matches as well as *CG* and *GC* mismatches. However, GC content alone does not fully explain the effects seen in SIR1. For instance CA mismatches also have positive coefficients, and some substitutions that would increase GC content (e.g., *AG* and *TG*) have negative coefficients.

The role of gaps in these discriminatory analyses is subtle, but apparent. In the distribution of gap density, coding regions stand out clearly as gap-free elements, whereas regulatory regions behave in a way very similar to ancestral repeats and bulk DNA. Also, the frequency of gaps plays a role in first and second principal components, and the presence of gaps is indirectly measured by the density of matching hexamers (sequences with gaps are automatically discarded from the density computation). From a functional viewpoint, there is no reason to expect regulatory regions to be gap-poor. In fact, gaps may be of some benefit in relaxing a stringent requirement for exact spacing of upstream elements.

The discriminatory power of the log-odds score based on 5-symbol fifth-order Markov models is impressive in the present analysis, suggesting that it will be helpful in finding candidate regulatory regions genome-wide. However, our training data comprised only 93 regulatory regions. This is a small fraction of the total regulatory elements for as many as 30,000 genes. Also, many of the 93 regulatory elements we used are tissue-specific. Thus, the training set is not necessarily a representative sampling of all regulatory elements, and it may be biased toward the types of elements most frequently studied to date, such as those for tissue-specific or inducible genes. It follows that some types of regulatory elements are likely underrepresented in the regulatory potentials we are in the process of computing for genome-wide alignments. Future applications of these methods should incorporate larger training sets as more regulatory elements are described in detail. Also, it is possible that more refined methods could distinguish different types of regulatory elements, for example, constitutive versus tissue-specific. As more members of other classes of regulators are characterized, such as silencers, insulators, and boundary elements, development of discrimination scores for these classes may be feasible. These should be fertile grounds for future studies.

In addition to gene regulatory elements, the set of DNA under selection that does not code for protein should contain RNA-coding genes and possibly determinants of chromosome structure and function. Such nonregulatory, noncoding DNA may have alignment characteristics distinctive from those of the regulatory regions. Hence some noncoding sequences with high locally adjusted conservation scores, reflecting a strong likelihood of selection (Roskin et al. 2002; Waterston et al. 2002), may have low regulatory potential, using the scores described here. These are candidates for functional sequences not involved in regulation (or more precisely, not promoters and enhancers with properties similar to the ones in our training set). Further analysis of such sequences could be informative.

As genomic alignments from multiple species become available, the ability to distinguish functional regions should improve (Flint et al. 2001; Botcherby 2002). Applying high-order Markov models to multiple alignments will be challenging, because of the exponential explosion in the state space size to be counteracted through meaningful collapses in the alphabet of all multiple nucleotide combinations. Again, solutions to such issues should be sought in future studies.

The ab initio approaches described here are designed to improve the reliability of predictions of regulatory elements based on alignments of genomic DNA. Such predictions have been helpful but not infallible in previous studies. The critical test of these tools and resources will be their application in experimental analyses. We look forward to seeing how novel potential regulatory elements identified by these and other tools behave when tested for activity in appropriate biological systems.

METHODS

All analyses use the build 30 assembly of the human genome aligned to the February 2002 freeze from the Mouse Genome Sequencing Consortium; available at the Human Genome Browser Web site (http://genome-test.cse.ucsc.edu/). We used the axtBEST alignments, which are based on the BLASTZ alignments and give the best mouse hit to each human position. The regulatory region collection is available online at http://bio.cse.psu.edu/mousegroup/Reg_annotations/. It was compiled from collections described in Wasserman and Fickett (1998), Krivan and Wasserman (2001), Dermitzakis and Clark (2002), and references therein. Each entry was trimmed to the smallest functional unit described in the literature, below which further deletions caused a loss of activity. Ancestral repeats are those present in the last common ancestor to mouse and human, as determined by the amount of divergence from the consensus sequence (Hardison et al. 2003; Waterston et al. 2002). Alignments in bulk DNA were masked at the position of exons as listed in the HGB RefSeq track (http://genome-test.cse.ucsc.edu/; http://www.ncbi.nlm.nih.gov/LocusLink/).

Alignment Score per Column

The ASPC is computed using coefficients in the BLASTZ scoring matrix (Schwartz et al. 2003):

	A	*C*	*G*	*T*
A	91	−114	−31	−123
C	−114	100	−125	−31
G	−31	−125	100	−114
T	−123	−31	−114	91

As for gap positions, $O = -400$ is the coefficient for open gap penalty, and $E = -30$ that for extension gap penalty. Also, N or n in the human sequence gets an $N = -100$ coefficient, and X or x in the human sequence gets an $X = -1000$ coefficient.

To compute the ASPC for nonoverlapping 200-bp windows, we used the following protocol: Alignments of <200 bp are ignored. When parsing an alignment in 200-bp windows, if the last segment is less than half of the required size (200/2 = 100), it is merged into the previous window to form a window larger than 200 but smaller than 300 bp. For each window, the ASPC is the sum of individual position scores divided by the window size (alignment gaps in the human sequence induce windows that actually contain <200 positions, but the denominator is always taken to be 200). For alignments in bulk DNA, coding regions were masked prior to calculating a score.

Principal Component Analysis and Sliced Inverse Regression

We used our four basic alignment collections to construct a data cloud of 693 points in 17 dimensions as follows: 93 of the 95 alignments in the regulatory elements collection were retained as such, and 17-symbol frequencies (A, C, G, T pairings and gaps) were computed on each. Then, 200 alignments of length greater than or equal to 200 bp were selected at random from each of the ancestral repeats, coding regions, and bulk DNA collections. The 17-symbol frequencies were computed on the first 200 bp only of each of these alignments. The 200-bp size was selected for compatibility with the size of nonoverlapping windows used for ASPC and gap density (in relation to the length of regulatory elements in our collection, about a third of them have lengths smaller than or equal to 200 bp).

PCA (see, e.g., Gnanadesikan 1997) is based on the spectral decomposition of the overall variance/covariance matrix. The eigenvectors represent orthogonal directions (i.e., linear combinations of the 17 frequencies) ranked in decreasing order of data variability. The corresponding eigenvalues (in nondecreasing order) quantify this variability.

For SIR (Li 1991; Cook 1998 and references therein), we use the same data cloud used in PCA, plus the information relative to the points known classification (a categorical response variable). When used for categorical responses, SIR is a close relative of Fisher linear discriminant analysis (FLDA). Estimation is based on the spectral decomposition of the between classes variance/covariance matrix, standardized by the overall variance/covariance matrix (as opposed to the within variance/covariance matrix, used for standardization in FLDA). The eigenvectors represent orthogonal directions (i.e., linear combinations of the 17 frequencies) ranked in decreasing order of relevance to the classification. The corresponding eigenvalues (in nondecreasing order) quantify this relevance.

Density of Exact Hexamer Matches

Exact hexamer matches are counted by sliding a 6-bp window across aligned sequence 1 bp at a time (whenever the sliding window contains six matches, the count is increased by 1). For each nonoverlapping 200-bp window, the density is computed by dividing the count relative to the window by the overall number of sliding 6-bp windows contained in the window (namely, $200 - 6 + 1$). Alignments shorter than 200 bp are discarded, and segments shorter than 200 bp at the end of alignments are merged with the previous window or discarded, as illustrated above for the ASPC calculation. Bulk DNA was masked for coding exons prior to analysis.

Log-Odds Ratios From Markov Models

This analysis uses the 93 regulatory segments plus 200 ancestral repeat segments of size 200 bp already used for the PCA and SIR analyses. A separate K-th order Markov model on an alphabet of J symbols is estimated using each of the two sets. Estimation is based on empirical frequencies as follows: For a given data set, the occurrences of all $J^{(K+1)}$ possible $(K + 1)$-symbol combinations are counted (a 1 is added to each count to prevent problems with outcomes that apparently had 0 probability because of our relatively small data sets). Then the counts associated with hexamer strings that share the same first 5 symbols are pooled and normalized to create transition probabilities of the Markov chain; for instance, transition probability(MMMMMG) = occurrences(MMMMMG)/occurrences(MMMMM), where occurrences(MMMMM) is the sum of the counts for MMMMM*****, as ***** ranges on all alphabet symbols. These transition probabilities can be arranged in a matrix of size $J^K \times J$ so that each row corresponds to a combination of the first K symbols and each column corresponds to an additional symbol following them.

Based on the two transition probability matrices resulting from this process, a score matrix is formed by taking log-odds ratios; that is, the natural logarithm of the ratio of the probability for a regulatory region to the probability for a neutral region (ln(prob_reg/prob_neutral)). To sum up, each possible string of contiguous $(K + 1)$ symbols is assigned with a log-odds score. To measure how much more likely an alignment to be analyzed is regulatory as compared with neutral, we compute the ratio between the probabilities of this alignment being generated by the Markov model of the regulatory region and by the model of the neutral region. The log of this ratio (raw score) is simply the summation over the entire length of the alignment of the log-odds ratios for each contiguous $(K + 1)$ symbols string in it (given position, and previous K ones).

Because the raw score is strongly dependent on the length of an alignment (this changes over a wide range for our regulatory set), normalization is needed to allow the use of a single threshold in deciding whether an alignment is regulatory or neutral. The final calculated score is given by (raw score − expectation * length)/(length$^{1.25}$). The expectation is the average score of a contiguous $(K + 1)$ symbol string using the equilibrium distribution of the ancestral repeats Markov chain.

ACKNOWLEDGMENTS

We thank Manolis Dermitzakis (University of Geneva) for help with the known regulatory regions as well as members of the Mouse Sequencing Consortium for support and data sharing during this study, and two referees for suggestions that led to substantial improvements in this paper. This work is supported by NHGRI grants HG-02325 (L.E.) and HG-02238 (R.H., F.C., and W.M.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.* **10:** 950–958.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the

Drosophila gene. *Proc. Natl. Acad. Sci.* **99:** 757–762.
Botcherby, M. 2002. Harvesting the mouse genome. *Comp. Funct. Genom.* **3:** 319–324.
Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.
Chiaromonte, F., Yap, V.B., and Miller, W. 2002. Scoring pairwise genomic sequence alignments. *Pac. Symp. Biocomput.* 115–126.
Cook, R.D. 1998. *Regression graphics*. Wiley, New York.
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.
Elnitski, L., Miller, W., and Hardison, R. 1997. Conserved E boxes function as part of the enhancer in hypersensitive site 2 of the β-globin locus control region. Role of basic helix–loop–helix proteins. *J. Biol. Chem.* **272:** 369–378.
Flint, J., Tufarelli, C., Peden, J., Clark, K., Daniels, R.J., Hardison, R., Miller, W., Philipsen, S., Tan-Un, K.C., McMorrow, T., et al. 2001. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the α globin cluster. *Hum. Mol. Genet.* **10:** 371–382.
Gnanadesikan, R. 1997. *Methods for statistical data analysis of multivariate observations*, 2nd ed. Wiley, New York.
Hardison, R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16:** 369–372.
Hardison, R., Slightom, J.L., Gumucio, D.L., Goodman, M., Stojanovic, N., and Miller, W. 1997a. Locus control regions of mammalian β-globin gene clusters: Combining phylogenetic analyses and experimental results to gain functional insights. *Gene* **205:** 73–94.
Hardison, R.C., Oeltjen, J., and Miller, W. 1997b. Long human–mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7:** 959–966.
Hardison, R., Roskin, M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Covariation infrequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* (this issue).
Jackson, J.D., Petrykowska, H., Philipsen, S., Miller, W., and Hardison, R. 1996. Role of DNA sequences outside the cores of DNase hypersensitive sites (HSs) in functions of the β-globin locus control region. Domain opening and synergism between HS2 and HS3. *J. Biol. Chem.* **271:** 11871–11878.
Jegga, A.G., Sherwood, S.P., Carman, J.W., Pinski, A.T., Phillips, J.L., Pestian, J.P., and Aronow, B.J. 2002. Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res.* **12:** 1408–1417.
Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.
Krivan, W. and Wasserman, W.W. 2001. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* **11:** 1559–1566.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Li, K.R. 1991. Sliced Inverse Regression for dimension reduction. *J. Am. Stat. Assoc.* **86:** 316–342.
Loots, G.G., Locksley, R.M., Blankespoor, C.M., 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.
Makalowski, W., Zhang, J., and Boguski, M.S. 1996. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* **6:** 846–857.
Pennacchio, L.A., Olivier, M., Hubacek, J.A., Cohen, J.C., Cox, D.R., Fruchart, J.C., Krauss, R.M., and Rubin, E.M. 2001. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294:** 169–173.
Pruitt, K.D., Katz, K.S., Sicotte, H., and Maglott, D.R. 2000. Introducing RefSeq and LocusLink: Curated human genome resources at the NCBI. *Trends Genet.* **16:** 44–47.
Roskin, K.H., Diekhans, H., Kent, W.J., and Haussler, D. 2002. Score functions for assessing conservation in locally aligned regions of DNA from two species. Tech. Rep. UCSC-CRL-02-03. Center of Biomolecular Science and Engineering, Baskin Engineering, University of California, Santa Cruz, CA.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R., Haussler, D., and Miller, W. 2003. Human–mouse alignments with BLASTZ. *Genome Res.* (this issue).
Wasserman, W.W. and Fickett, J.W. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278:** 167–181.
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.
Xu, Y. and Uberbacher, E.C. 1997. Automated gene identification in large-scale genomic sequences. *J. Comp. Biol.* **4:** 325–338.

WEB SITE REFERENCES

http://bio.cse.psu.edu/; Penn State Bioinformatics Group.
http://bio.cse.psu.edu/mousegroup/Reg_annotations/; collection of regulatory elements.
http://compbio.ornl.gov/grailexp/; Grail Experimental Gene Discovery Suite.
http://genome-test.cse.ucsc.edu/; Human, mouse and rat genome browsers at UCSC Genome Bioinformatics.
http://www.ncbi.nlm.nih.gov/LocusLink/; Locus Link at NCBI.

Received September 16, 2002; accepted in revised form November 14, 2002.