Active Conservation of Noncoding Sequences Revealed by Three-Way Species Comparisons

Inna Dubchak,[1] Michael Brudno,[1] Gabriela G. Loots,[2] Lior Pachter,[3] Chris Mayor,[1] Edward M. Rubin,[2] and Kelly A. Frazer[2,4,5]

[1]*Center for Bioinformatics and Computational Genomics,* [2]*Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA;* [3]*Department of Mathematics, University of California at Berkeley, Berkeley, California 94720, USA*

Human and mouse genomic sequence comparisons are being increasingly used to search for evolutionarily conserved gene regulatory elements. Large-scale human–mouse DNA comparison studies have discovered numerous conserved noncoding sequences of which only a fraction has been functionally investigated A question therefore remains as to whether most of these noncoding sequences are conserved because of functional constraints or are the result of a lack of divergence time.

[The sequence data described in this paper have been submitted to the GenBank data library under accession nos. AF276990.]

Based on the supposition that actively conserved human–mouse noncoding sequences will be present in a third mammal, whereas noncoding regions that are similar because of an insufficient accumulation of random mutations will be absent, we sequenced ~200 kb of orthologous human (5q31), mouse (chromosome 11), and dog (chromosome 4) DNA. The functions of conserved noncoding sequences (syntenous gene regulatory elements) are unaffected by relatively small random insertions or deletions of base pairs, and therefore, standard local alignment algorithms that identify ungapped conserved regions are not ideally suited for their discovery. For this reason, comparative analysis was performed by generating pairwise global sequence alignments [human–dog (H/D), human–mouse (H/M), and mouse–dog (M/D)], and we developed an algorithm to search for blocks of similarity in the alignments. To view the conserved regions in the three pairwise sequence alignments simultaneously, we developed a new visualization tool, VISTA (visualization tool for alignment). Inspection of the graphical output of VISTA revealed that the H/D, H/M, and M/D alignments have almost identical patterns of noncoding sequence conservation (Fig. 1). The content and order of the six genes in this 200-kb region are the same for all three species; however, the coding regions of two genes, *Interleukin-4* and *Interleukin-13*, are only moderately conserved (~50% identity).

Previous H/M DNA comparison studies have used arbitrary cutoff criteria (≥*X*% identity over ≥*Y* bp) to define noncoding sequences as evolutionarily conserved (Loots et al. 2000). Here, we statistically determine cutoff criteria for defining conserved noncoding sequences by examining the three pairwise sequence alignments, H/D, H/M, and M/D, using intersection/union (I/U) analyses (Table 1). The cutoffs for which the sum of the three pairwise I/U values (largest number of overlapping, and least number of unique, conserved noncoding elements) was maximal were as follows: H/D, ≥92% identity over ≥120 bp; H/M, ≥80% identity over ≥120 bp; and M/D, ≥77% identity over ≥120 bp. These data indicate that the high percent identity noncoding sequences in the ~200-kb region examined are most similar in humans and dogs and therefore suggest that H/D DNA comparisons may be better than H/M DNA comparisons for detecting conserved noncoding elements.

At the optimal cutoffs, 16 H/D conserved noncoding sequences (CNSs) were identified of which 14 were present in all three pairwise sequence alignments. Two of the CNSs (at 97 kb and 108 kb) have been experimentally determined to be gene regulatory elements supporting the cutoff criteria obtained from the I/U analyses (Henkel et al. 1992; Loots et al. 2000). The two CNSs present in humans and dogs (at 2 kb and 98 kb) but not in mice (Fig. 1) may represent gene regulatory elements that either are not conserved at the sequence level between humans and mice or have been lost in mice during evolution. Using the statistically determined percent identity and length thresholds resulted in few putative false negatives (the CNSs are present in all three species but fall slightly below the cutoff value); however, a significant number of exons in the genes within the region do not meet these criteria

[4]Present address: Affymetrix, Santa Clara, California 95051 USA.
[5]Corresponding author.
E-MAIL kelly_frazer@affymetrix.com; FAX (408) 481-0422.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.142200.

10:1304–1306 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

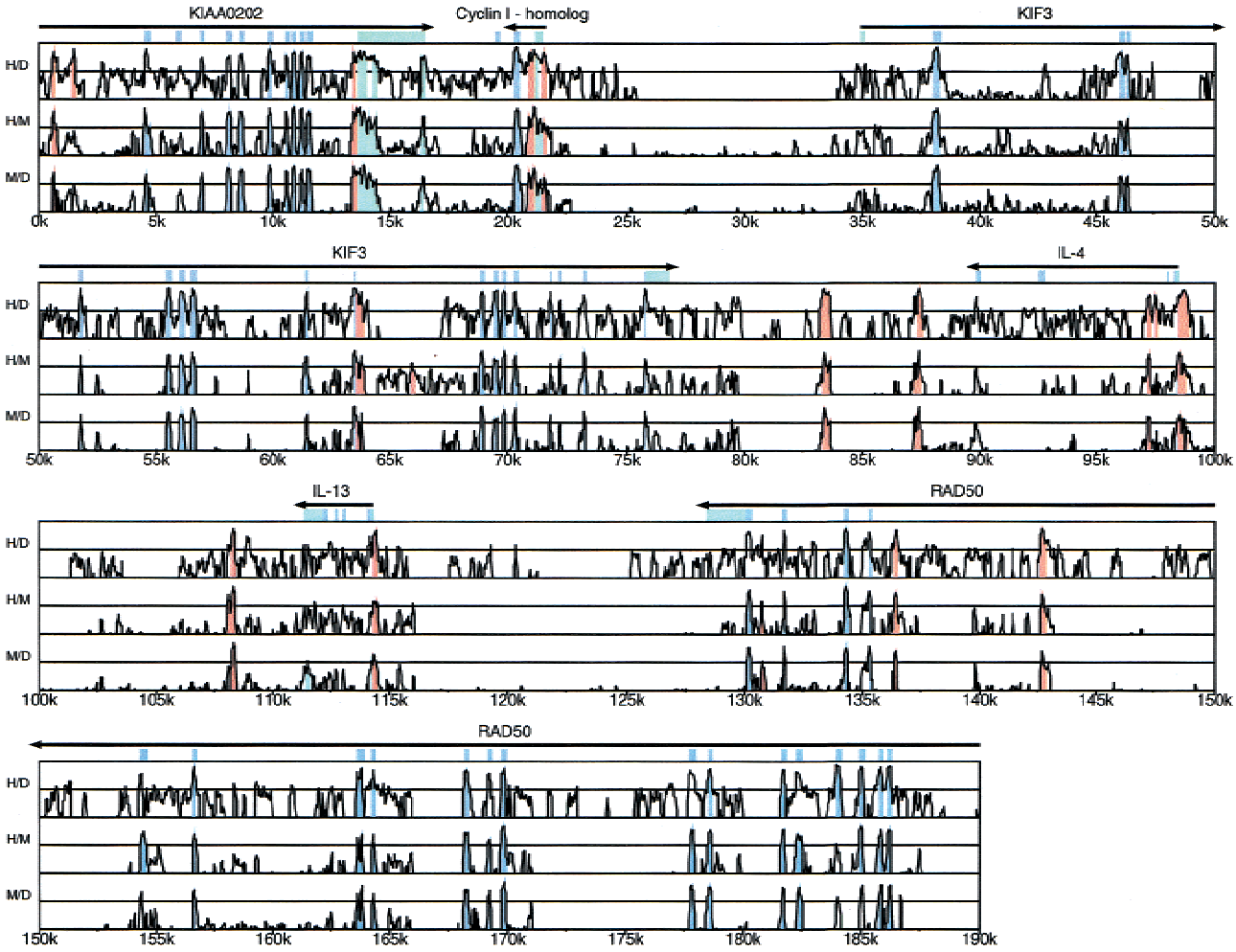

Figure 1 VISTA plot demonstrating peaks of similarity in the pairwise sequence alignments. Conserved sequences are shown relative to their positions in the human genome (horizontal axes), and their percent identities (50%–100%) are indicated on the vertical axes. The locations of coding exons (blue rectangles) and 5′- and 3′-untranslated regions (UTRs) (turquoise rectangles) are shown above the profile. Horizontal arrows indicate the direction of transcription for each gene. Peaks representing noncoding (red) and UTR (turquoise) sequences fitting the criteria for conserved elements as well as coding sequences (blue) meeting the percentage criteria over their entire length are indicated. The M/D alignment is mapped on the coordinates of the human genome sequence based on matching base pairs in the H/M alignment.

(Fig. 1). Less stringent cutoff criteria would have included these exons but resulted in the overprediction of noncoding sequences as conserved (false positives).

The vast majority of the H/M CNSs identified in the 200-kb region examined are also present in dog. This is an important finding as it suggests that a large fraction of the high percent identity noncoding elements identified through H/M DNA comparison studies are conserved because of functional constraints. A problem with two-species sequence comparison studies is that cutoff values for defining noncoding elements as conserved are based on biologists' intuition for what constitutes a biologically significant threshold. Our simultaneous comparison of orthologous sequences from three mammals allowed us statistically to determine percent identity and length thresholds to define actively CNSs. These cutoff values may be useful guidelines for identifying CNSs in genomic regions for which only human and mouse DNA sequences are available.

METHODS

Genomic Sequences

Human 5q31 (NT 000170) and mouse chromosome 11 (AC005742) sequences were obtained as described (Loots et al. 2000). A dog chromosome 4 bacterial artificial chromosome (BAC) was isolated from BACPAC resources library (RPCI-81), sequenced in draft format, and the contigs were ordered and oriented (AF276990).

Sequence Alignments and Visualization

Sequences were globally aligned using GLASS (global align-

Table 1. Intersection/Union Analysis of H/M vs M/D Conserved Noncoding Sequences

		M/D				
	%ID	67%	72%	77%	82%	87%
H/M	70%	37%	31%	21%	17%	10%
	75%	38%	46%	41%	32%	20%
	80%	31%	58%	94%	63%	36%
	85%	25%	47%	68%	76%	63%
	90%	13%	26%	37%	50%	62%

A subset of the I/U analysis for the H/M vs M/D sequence alignments at varying % identity (ID) over a fixed length (120bp) is shown. Each entry is the I/U percentage, which indicates the proportion of CNSs that were present in either the H/M or M/D alignments (at ≥ the given % ID criteria) that were present in both. The % ID optimal cutoffs for CNSs is the H/M (≥80% ID) and M/D (≥77% ID) values correspond to the largest I/U percentage (red).

ment system) (Batzoglou et al. 2000), and conserved regions were identified by calculating the percent of identical nucleotides within a 100-nucleotide window moved in 25-nucleotide increments across the alignments. The source code of VISTA, the Java program for visualization of alignments, is available upon request, and a VISTA server can be accessed at http://www-gsd.lbl.gov/vista.

I/U Analysis

Conserved segments with percent identity X and length Y were defined to be regions in which every contiguous subsegment of length Y was at least X% identical to its paired sequence. These segments were then merged to define the conserved regions. The I/U analyses were performed to define length and identity cutoffs as follows: The set of conserved regions between the H/M (denoted by A) and the M/D (denoted by B) were identified. Regions $a \in A$ and $b \in B$ were considered equal if they overlapped in the mouse sequence. I/U was then obtained by computing $|A \cap B|/|A \cup B|$ where $|A \cap B| = \min(|A \subset B|,|B \subset A|)$ and $|A \cup B| = |A| + |B| - \max(|A \subset B|,|B \subset A|)$. $|A \subset B|$ is the number of regions in A that are equal to regions in B. This number might be different from $|B \subset A|$ because it is possible that multiple regions in one alignment are equal to one region in the other.

ACKNOWLEDGMENTS

We thank Keith Lewis, Willow Dean, and Cathy Blankespoor for DNA sequencing and Nila Patil for valuable remarks on the manuscript. This work was supported by the following grants: U.S. Department of Energy contract DE-AC376SF00098 and NIH GM-5748202 (K.A.F.)

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

Ansari-Lari, M.A., Oeltjen, J.C., Schwartz, S., Zhang, Z., Muzny, D.M., Lu, J., Gorrell, J.H., Chinault, A.C., Belmont, J.W., Miller, W., et al. 1998. Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res.* **8:** 29–40.

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.* (in press).

Gottgens, B., Barton, L.M., Gilbert, J.G., Bench, A.J., Sanchez, M.J., Bahn, S., Mistry, S., Grafham, D., McMurray, A., Vaudin, M., et al. 2000. Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat. Biotechnol.* **18:** 181–186.

Hardison, R.C., Oeltjen, J., and Miller, W. 1997. Long human–mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7:** 959–966.

Henkel, G., Weiss, D.L., McCoy, R., Deloughery, T., Tara, D., and Brown, M.A. 1992. A DNase I-hypersensitive site in the second intron of the murine IL-4 gene defines a mast cell-specific enhancer. *Immunology* **149:** 3239–3246.

Koop, B.F. and Hood, L. 1994. Striking sequence similarity over almost 100 kilobases of human and mouse t-cell receptor DNA. *Nat. Genet.* **7:** 48–53.

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.

Oeltjen, J.C., Malley, T.M., Muzny, D.M., Miller, W., Gibbs, R.A., and Belmont, J.W. 1997. Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* **7:** 315–329.

Received March 28, 2000; accepted in revised form July 12, 2000.