Available online at www.sciencedirect.com

SCIENCE DIRECT

GENOMICS

Genomics 84 (2004) 890–895

www.elsevier.com/locate/ygeno

Short Communication

Interpreting mammalian evolution using *Fugu* genome comparisons

Ivan Ovcharenko*, Lisa Stubbs, Gabriela G. Loots*

Genome Biology Division, Lawrence Livermore National Laboratory, 7000 East Avenue, L-441, Livermore, CA 94550, USA

Received 16 April 2004; accepted 18 July 2004
Available online 26 August 2004

Abstract

Recently, it has been shown that a significant number of evolutionarily conserved human–*Fugu* noncoding elements function as tissue-specific transcriptional enhancers in vivo, suggesting that distant comparisons are capable of identifying a particular class of regulatory elements. We therefore hypothesized that by juxtaposing human/*Fugu* and human/mouse conservation patterns we can define conservation criteria for discovering transcriptional regulatory elements specific to mammals. Genome-scale comparisons of noncoding human/*Fugu* evolutionary conserved elements (ECRs) and their humans/mouse counterparts revealed a particular signature common to human/mouse ECRs (≥350 bp long, ≥77% identity) that are also conserved in fishes. This newly defined threshold identifies 90% of all human/*Fugu* noncoding ECRs without the assistance of human–*Fugu* genome alignments and provides a very efficient filter for identifying functional human/mouse ECRs.
Published by Elsevier Inc.

Comparative sequence analysis of the human and the pufferfish *Fugu rubripes* genomes has revealed several novel functional coding and noncoding regions in the human genome [1,2]. In particular, the *Fugu* genome has been extremely valuable for identifying transcriptional regulatory elements in human loci harboring unusually high levels of evolutionary conservation to rodent genomes [3–5]. In such regions, the large evolutionary distance between humans and fishes provides an additional filter through which functional noncoding elements can be detected with high efficiency.

We have evaluated the noncoding conservation profile in human/*Fugu* genome alignments obtained from the ECR Browser [6] and generated by the blastz program [7]. Filtering of known and putative transcripts, pseudogenes, GenBank mRNAs, as well as proximal promoter sequences identified 2968 human/*Fugu* evolutionary conserved regions (ECRs) [≥70% identity (% ID) over ≥100 base-pairs (bp)] that are noncoding in nature and distantly positioned from the transcriptional start site of adjacent genes. These ECRs are predominantly clustered in discrete areas of the human genome, flanked by or inserted into the introns of 1026 human transcripts that together comprise only 5.6% of the 18,410 "known gene" loci (as annotated at UCSC Genome Browser [8] build 34 of the human genome). The transcripts bordering these ECR clusters were significantly enriched for genes involved in core biological processes such as development, transcription, morphogenesis, and neurogenesis, while also depleted in several species-specific functions such as immune response or cytokine activity (Fig. 1). This distribution suggests that human–*Fugu* sequence comparisons will be beneficial for identifying noncoding regulatory elements for only a small percentage of human genes. Moreover, the number of genes under the control of these putative regulatory elements could be even smaller if enhancers located between two genes influence gene expression of only one of the neighboring transcripts.

It has been estimated that ~5% of the human genome is under active selection, the majority of which will likely

* Corresponding authors. Fax: (925) 422-2099.
E-mail addresses: ovcharenko1@llnl.gov (I. Ovcharenko), loots1@llnl.gov (G.G. Loots).

0888-7543/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.ygeno.2004.07.011

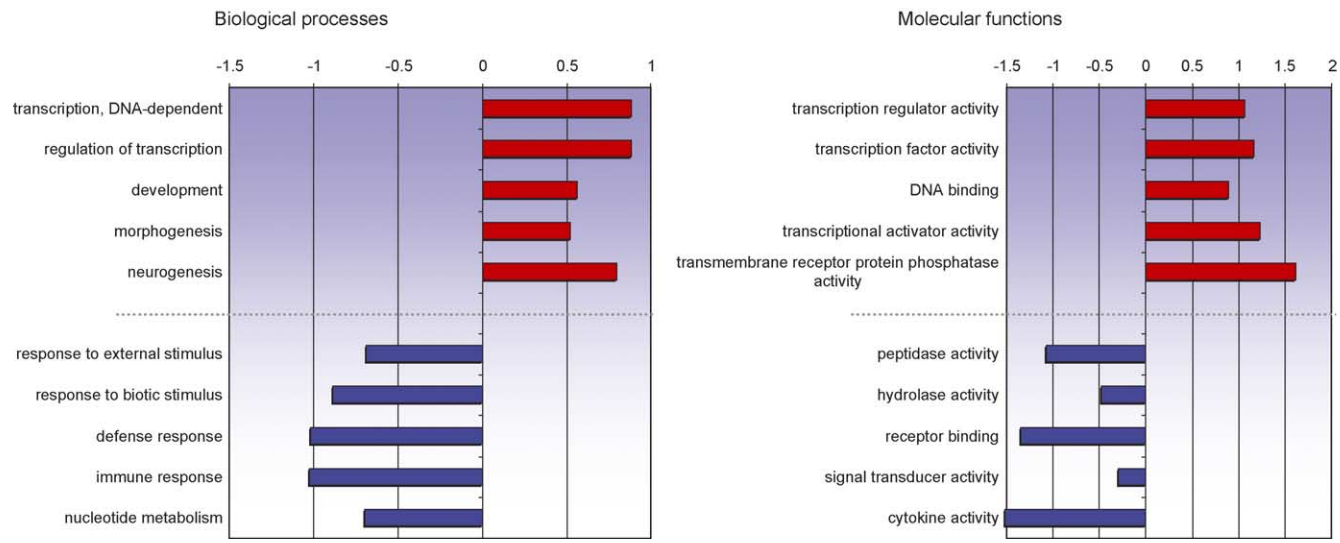

Fig. 1. Enriched and depleted GeneOntology categories in the dataset of genes flanked by the human/*Fugu* ECRs. Two (left and right) plots contain five of the most significantly enriched (in red) and depleted (in blue) gene categories as quantified by the z value (the difference between observed and expected number of genes divided by the standard deviation; reported results have P value < 0.01). Left and right plots separate gene categories into biological processes and molecular functions, respectively. Horizontal scale measures the natural logarithm of the ratio of observed-to-expected gene counts.

correspond to functional coding and noncoding sequences [9]. Human–rodent genome alignments [6] revealed 1.3 million noncoding ECRs with an average distribution of 68.8 ECRs per human gene locus, whereas the density varies according to the regional neutral substitution rates [10]. Assigning in vivo function to all these conserved elements is impossible with current technologies, and it is therefore critically important to identify ways to efficiently discriminate functional noncoding elements from neutrally evolving, but still highly conserved genomic DNA. This goal might be achieved if "fingerprints" unique to functional and nonfunctional noncoding conserved elements can be defined. Assuming that elements conserved between human and *Fugu* represent an incomplete yet highly enriched functional dataset, we approached this problem by studying signatures specific to human/mouse conserved noncoding elements that are also present in fishes.

We compared the distribution in size (bp) and percent identity (% ID) of human/rodent (h/r) and human/*Fugu* (h/f) noncoding ECRs (Fig. 2). In particular, we focused on a subset of h/r ECRs that are also represented in the *Fugu* genome (have h/f ECR counterparts), and quantified the h/r conservation parameters. This particular subset of h/r ECRs will be referred to as *core ECRs*. To create a comprehensive h/r ECR dataset we extracted all noncoding human/mouse ECRs from the genome alignments. Underrepresented regions in the mouse genome were extended by the available rat genomic sequences. The distribution in ECR length was strikingly similar between the human/mouse and the human/*Fugu* ECRs comparisons; 81% h/r and 86% h/f ECRs were shorter than 350 bp. In sharp contrast, the majority of the *core ECRs* were greater than 350 bp in length. Similar striking differences were observed for the level of sequence identity. While 82 and 71% of the h/r and h/f ECRs were found to range between 70 and 77% sequence identity, 90% of *core ECRs* showed greater than 77% ID. Therefore, our analysis suggests that a "mammalian evolutionary threshold" of ≥350 bp, ≥77% ID, conservation criteria recapitulates the majority of all conserved noncoding elements identified from distant h/f genome comparisons, and reduces the number of h/m conserved noncoding elements 10-fold, from 1.3 millions to 128 thousand ECRs, significantly simplifying the search for putative functional noncoding elements.

To correlate our findings with the conservation profiles of known regulatory elements we analyzed a 2.6 Mb region from the human *DACH* gene locus where recently seven human enhancers have been mapped [3]. Of the 1367 h/r noncoding ECRs (>100 bp/>70% ID), 34 are also present in *Fugu*. The majority of these *Fugu* elements conserved in humans, rodents, and other species progressively increase in length as the phylogenetic distance decreases (Fig. 3). A conservation criterion of ≥350 bp/≥77% ID identified 302 h/r ECRs and recapitulated 33/34 of the h/f conserved elements, while excluding 78% of the original h/m ECRs and maintaining 100% of the experimentally validated regulatory elements. Other known distant regulatory elements, including SHH and DLX1-specific developmental enhancers exceeded this conservation threshold (≥350 bp/≥77% ID) in h/r genomic alignments, independent of their presence in the *Fugu* genome (Table 1) [11,12]. We also applied these newly defined parameters on human–chicken and human–frog whole genome alignments available from the ECR Browser [6]. Over 72% of ~7500 human–frog and 55.4% of ~71,200 human–chicken noncoding ECRs that are also present in rodents obey this "stringent evolutionary

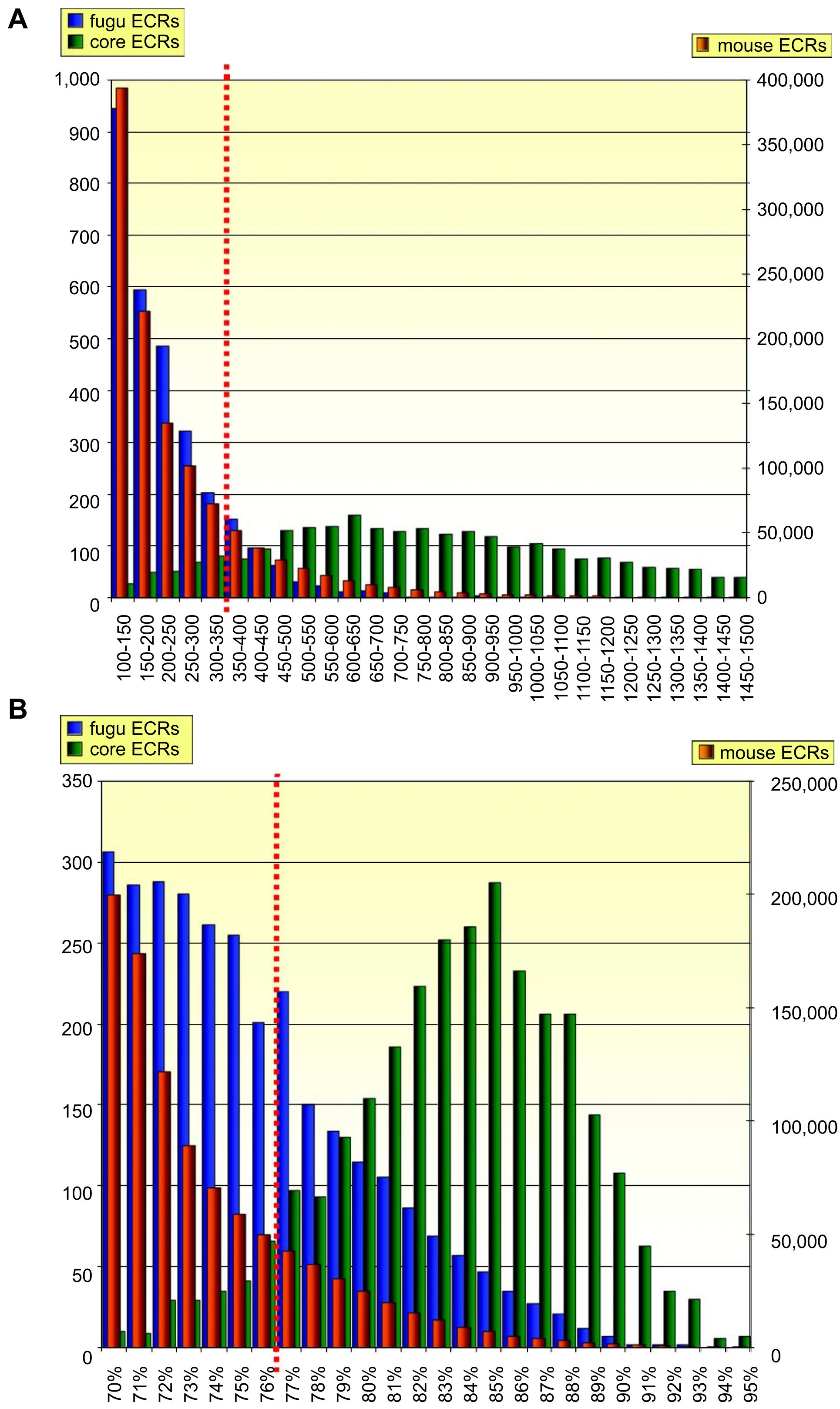

Fig. 2. Genome scan of ECR length (A) and percentage identity (B). Human/*Fugu* ECRs are in blue, human/rodent ECRs are in orange, and human/rodent *core ECRs* are in green. *x* axis, size in bp (A) and percentage identity (B); *y* axis, number of ECRs per given category. Please note that the number of human/rodent ECRs is scaled with the *right y* axis, while two other categories are scaled with the *left y* axis.

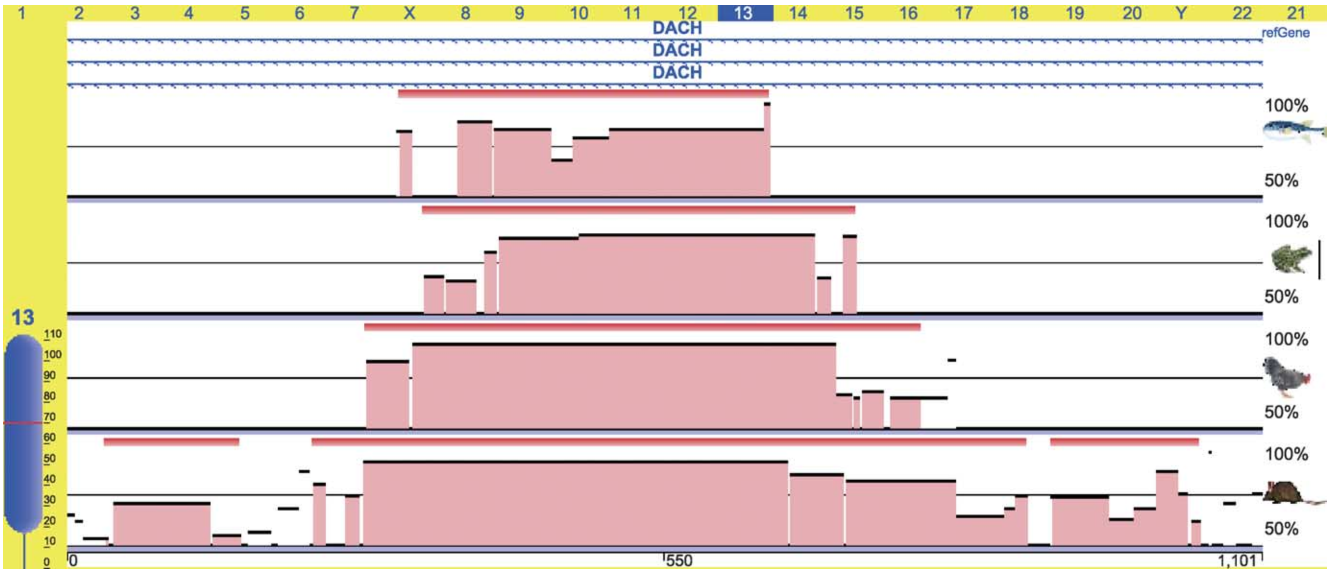

Fig. 3. ECR Browser conservation profile of a DACH gene intronic region (chr13:70,169,300–70,170,400; NCBI human genome build 34) that is present in humans, rodents, chickens, frogs, and fishes demonstrates a gradual increase in length while approaching mammals from fishes. There are also two flanking human/mouse ECRs that are longer than 100 bp, but are shorter than 350 bp that are not conserved in any other species studied except humans and mice.

threshold" rule of conservation in the analysis of human–rodent counterpart ECRs. As we move closer in evolution within the vertebrate radiation, more stringent conservation criteria are required to amplify the signal to noise ratio to allow the visualization of functional regions in alignments that lack sufficient evolutionary time to diverge in neutral regions.

Concluding, we suggest a novel approach for analyzing human/rodent conservation profiles that is capable of reconstructing more ancestral evolutionary relationships and distinguishing functional conserved elements from the neutrally evolving genomic background. By applying a ">350 bps/>77%" ID threshold to the analysis of human/rodent conservation profiles we were able to recapitulate the majority of human/fish conserved elements and to generate a small set of elements that have a high probability of being functional noncoding domains. Similar statistical approaches will be critical for understanding phylogenetic relationships through systematic pairwise genomic comparisons, and has the potential to facilitate the identification of regulatory elements specific to recently evolved species such as humans and their primate relatives.

Pairwise alignments between the reference human genome and the genomes of mouse, rat, and *Fugu* were generated as previously described [6]. Briefly, we first created synteny maps using the BLAT tool [13] for human–rodent comparisons and the more sensitive blastn program for human/*Fugu* comparisons [14]. Next, all homologous regions were aligned by the blastz local aligner tool [7]. The main goal of the alignment strategy has been to generate a single all-inclusive ECR coverage detected between a pair of compared genomes independent of the evolutionary history of the organism of origin.

All pairwise alignments were scanned using a sliding window (≥100 bp/≥70% ID) to identify ECRs with these minimum criteria [15]. Overlapping ECRs originating from paralogous or nonspecific alignments were filtered out and the longest representative ECRs for each region were reported. Thus 1,267,379 human/mouse and 65,949 human/*Fugu* ECRs were identified by this strategy. The majority of these human/*Fugu* ECRs corresponded to protein coding exons of known annotated genes and pseudogenes. To define a dataset of noncoding human/*Fugu* ECRs we excluded all the putative coding ECRs. First, we filtered out the exons of RefSeq, Ensembl, known genes, human, and nonhuman mRNAs mapped to the human genome [as annotated at the UCSC genome browser [8]. Next, we excluded unannotated genes and pseudogenes identified either by non-RefSeq mRNAs or sequence similarity to proteins from different species. All ECRs carrying significant sequence similarity to the NCBI nonredundant protein database (derived by blastx homology search; *e* value ≤1*e*–5) were identified and filtered out. This process reduced the size of the noncoding human/*Fugu* dataset to 2968 ECRs. Also, human genomic contaminations incorporated into the *Fugu rubripes* v3.0 genome assembly were initially detected using a criteria of ≥200 bp/≥95% ID. Significant matches were manually curated to identify contaminations, which were consequently excluded from the analysis (for example, *Fugu* scaffold_1388 matching to the HSA2 sequence with the 99% sequence similarity over 19 kb was removed from the analysis). In total, 28 putatively contaminated *Fugu* scaffolds were removed from the analysis.

Table 1
Experimentally characterized distant enhancer elements in the mouse

ECR gene	Enhancer	Size (bp)	H/M % ID	*Fugu* cons
Dachhund	Nobrega et al., 2003 [3]			
Dc1	Negative	630	89%	Yes
Dc2	Hindbrain	1405	89%	Yes
Dc3	For-, hindbrain spinal cord, retina	2458	88%	Yes
Dc4	Retina	1132	83%	Yes
Dc5	Negative	730	88%	Yes
Dc6	Midbrain, redina, drg	891	89%	Yes
Dc7	Limb bud	1401	88%	Yes
Dc8	Forbrain, neural tube	1023	87%	Yes
Dc9	Hindbrain, neural tube, genitalia	2247	82%	Yes
Dlx1-2	Ghanem et al., 2003 [11]			
I12a	Mesenchyme cells, branchyal arch	1784	84%	Yes
I12b	Telencephalon, diencephalon	864	92%	Yes
Dlx5-6	Ghanem et al., 2003 [11]			
mI56i	Telencephalon	1477	88%	Yes
mI56ii	Forbrain	830	88%	Yes
SHH	Lettice et al., 2003 [12]	1205	83%	Yes
Hoxc8	Anand et al., 2003 [16]	583	82%	Yes
IL4/IL13	Loots et al., 2000 [15]	472	79%	No
FGF4	Luster et al., 2003 [17]	566	81%	No
pax6/nkx2.8	Santagati et al., 2003 [4]			
cns6		500	83%	Yes
cns+2		1600	82%	Yes
pax7	Lang et al., 2003 [18]			
intron1		608	85%	No
ApoE	Zheng et al., 2004 [19]			
Brain		420	75%	No

Human/*Fugu* noncoding ECRs were used to detect overlying human/rodent ECRs (http://ecrbrowser.dcode.org/). Due to the draft status of the mouse genome some human/*Fugu* elements were absent from the mouse genome. In such cases, the missing human/mouse ECRs were augmented by human/rat ECRs, when available. The length and level of sequence identity were calculated for each ECR (Fig. 2).

Acknowledgments

This work was supported by LLNL-LDRD 04-ERD-052. The work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory Contract No. W-7405-Eng-48.

References

[1] B. Venkatesh, P. Gilligan, S. Brenner, Fugu: a compact vertebrate reference genome, FEBS Lett. 476 (2000) 3–7.

[2] S. Aparicio, J. Chapman, E. Stupka, N. Putnam, J.M. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, M.D. Gelpke, J. Roach, T. Oh, I.Y. Ho, M. Wong, C. Detter, F. Verhoef, P. Predki, A. Tay, S. Lucas, P. Richadson, S.F. Smith, M.S. Clark, Y.J. Edwards, N. Doggett, A. Zharkikh, S.V. Tavtigian, D. Pruss, M. Barnstead, C. Evans, H. Baden, J. Powell, G. Glusman, L. Rowen, L. Hood, Y.H. Tan, G. Edgar, T. Hawkins, B. Venkatesh, D. Rokhsar, S. Brenner, Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes, Science 297 (2002) 1283.

[3] M.A. Nobrega, I. Ovcharenko, V. Afzal, E.M. Rubin, Scanning human gene deserts for long-range enhancers, Science 302 (2003) 413.

[4] F. Santagati, K. Abe, V. Schmidt, T. Schmitt-John, M. Suzuki, K. Yamamura, K. Imai, Identification of Cis-regulatory elements in the mouse Pax9/Nkx2-9 genomic region: implication for evolutionary conserved synteny, Genetics 165 (2003) 235–242.

[5] F. Spitz, F. Gonzalez, D. Duboule, A global control region defines a chromosomal regulatory landscape containing the HoxD cluster, Cell 113 (2003) 405–417.

[6] I. Ovcharenko, M.A. Nobrega, G.G. Loots, L. Stubbs, ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes, Nucleic Acids Res. 32 (2004) W280–W286.

[7] S. Schwartz, W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler, W. Miller, Human–mouse alignments with BLASTZ, Genome Res 13 (2003) 103–107.

[8] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006.

[9] R.H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S.E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M.R. Brent, D.G. Brown, S.D. Brown, C. Bult, J. Burton, J. Butler, R.D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A.T. Chinwalla, D.M. Church, M. Clamp, C. Clee, F.S. Collins, L.L. Cook, R.R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K.D. Delehaunty, J. Deri, E.T. Dermitzakis, C. Dewey, N.J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D.M. Dunn, S.R. Eddy, L. Elnitski, R.D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G.A. Fewell, P. Flicek, K. Foley, W.N. Frankel, L.A. Fulton, R.S. Fulton, T.S. Furey, D. Gage, R.A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T.A. Graves, E.D. Green, S. Gregory, R. Guigo, M. Guyer, R.C. Hardison, D. Haussler, Y. Hayashizaki, L.W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D.B. Jaffe, L.S. Johnson, M. Jones, T.A. Jones, A. Joy, M. Kamal, E. Karlsson, D. Karolchik, E. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W.J. Kent, A. Kirby, D.L. Kolbe, I. Korf, R.S. Kucherlapati, E.J Kulbokas, D. Kulp, T. Landers, J.P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D.R. Maglott, E.R. Mardis, L. Matthews, E. Mauceli, J.H. Mayer, M. McCarthy, W.R. McCombie, S. McLaren, K. McLay, J.D. McPherson, J. Meldrim, B. Meredith, J.P. Mesirov, W. Miller, T.L. Miner, E. Mongin, K.T. Montgomery, M. Morgan, R.

Mott, J.C. Mullikin, D.M. Muzny, W.E. Nash, J.O. Nelson, M.N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M.J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K.H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C.S. Pohl, A. Poliakov, T.C. Ponce, C.P. Ponting, S. Potter, M. Quail, A. Reymond, B.A. Roe, K.M. Roskin, E.M. Rubin, A.G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M.S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J.B. Singer, G. Slater, A. Smit, D.R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J.P. Vinson, A.C. Von Niederhausern, C.M. Wade, M. Wall, R.J. Weber, R.B. Weiss, M.C. Wendl, A.P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R.K. Wilson, E. Winter, K.C. Worley, D. Wyman, S. Yang, S.P. Yang, E.M. Zdobnov, M.C. Zody, E.S. Lander, Initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520–562.

[10] R.C. Hardison, K.M. Roskin, S. Yang, M. Diekhans, W.J. Kent, R. Weber, L. Elnitski, J. Li, M. O'Connor, D. Kolbe, S. Schwartz, T.S. Furey, S. Whelan, N. Goldman, A. Smit, W. Miller, F. Chiaromonte, D. Haussler, Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution, Genome Res. 13 (2003) 13–26.

[11] N. Ghanem, O. Jarinova, A. Amores, Q. Long, G. Hatch, B.K. Park, J.L. Rubenstein, M. Ekker, Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters, Genome Res. 13 (2003) 533–543.

[12] L.A. Lettice, S.J. Heaney, L.A. Purdie, L. Li, P. de Beer, B.A. Oostra, D. Goode, G. Elgar, R.E. Hill, E. de Graaff, A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly, Hum. Mol. Genet. 12 (2003) 1725–1735.

[13] W.J. Kent, BLAT–the BLAST-like alignment tool, Genome Res. 12 (2002) 656–664.

[14] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

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

[16] S. Anand, W.C. Wang, D.R. Powell, S.A. Bolanowski, J. Zhang, C. Ledje, A.B. Pawashe, C.T. Amemiya, C.S. Shashikant, Divergence of Hoxc8 early enhancer parallels diverged axial morphologies between mammals and fishes, PNAS 100 (2003) 15666–15669.

[17] T.A. Luster, A. Rizzino, Regulation of the FGF-4 gene by a complex distal enhancer that functions in part as an enhanceosome, Gene 323 (2003) 163–172.

[18] D. Lang, C.B. Brown, R. Milewski, Y.Q. Jiang, M.M. Lu, J.A. Epstein, Distinct enhancers regulate neural expression of Pax7, Genomics 82 (2003) 553–560.

[19] P. Zheng, L.A. Pennacchio, W. Le Goff, E.M. Rubin, J.D. Smith, Identification of a novel enhancer of brain expression near the apoE gene cluster by comparative genomics, Biochim. Biophys. Acta 1676 (2004) 41–50.