LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-223593

# ECRbase: Database of Evolutionary Conserved Regions, Promoters, and Transcription Factor Binding Sites in Vertebrate Genomes

G.G. Loots, I. Ovcharenko

August 9, 2006

Bioinformatics

# *ECRbase*: Database of Evolutionary Conserved Regions, Promoters, and Transcription Factor Binding Sites in Vertebrate Genomes.

Gabriela Loots[1] and Ivan Ovcharenko[2,*]

[1]Biosciences and [2]Computational Directorates, Lawrence Livermore National Laboratory, 7000 East Avenue L-441, Livermore, CA. 94550, USA; *to whom correspondence should be addressed, tel. 925.422.5035; fax 925.422.2099; email, ovcharenko1@llnl.gov

**ABSTRACT**

Evolutionary conservation of DNA sequences provides a tool for the identification of functional elements in genomes. We have created a database of evolutionary conserved regions (ECRs) in vertebrate genomes entitled *ECRbase* that is constructed from a collection of pairwise vertebrate genome alignments produced by the ECR Browser database. *ECRbase* features a database of syntenic blocks that recapitulate the evolution of rearrangements in vertebrates and a collection of promoters in all vertebrate genomes presented in the database. The database also contains a collection of annotated transcription factor binding sites (TFBS) in all ECRs and promoter elements. *ECRbase* currently includes human, rhesus macaque, dog, opossum, rat, mouse, chicken, frog, zebrafish, and two pufferfish genomes. It is freely accessible at http://ECRbase.dcode.org.

## INTRODUCTION

Evolutionary conservation is a powerful method for identifying functional regions in a genome (1). In the recent years, genome comparisons have been efficiently applied to the discovery of novel genes (2) and regulatory elements (3,4). While sequences coding for proteins are strongly conserved across species, they encompass a small portion of a vertebrate genome. Some fraction of noncoding sequences is also conserved in the phylogeny of vertebrates, and increasing lines of evidence highlight the functional role of these evolutionary conserved regions (ECRs) in fundamental aspects of vertebrate biology. If ECRs are functionally important in vertebrate genomes, these regions should become a critical hunting ground for transcriptional regulatory signals that determine when, where, and in what quantities genes are expressed. In addition, genetic variation in these elements may be responsible for individual variability of gene expression that increases susceptibility to disease (5).

Contemporary genomics research is moving towards high-throughput and systematic whole-genome analysis that requires investigators to access comprehensive genomic data. Generating large datasets of computed alignments, ECRs and transcription factor binding site (TFBS) data on a genome scale require extensive computational resources that are not easily accessible by the average biologist. To facilitate genome-wide experimentation for investigators interested in pursuing global genomic analyses, we have created a portal to pre-computed, post-processed whole-genome alignment data that allows the extraction of ECRs, and promoter sequences as well as the TFBS associated with them, for all available vertebrate genomes.

## RESULTS

*ECRbase* includes ECRs identified in pairwise alignments of publicly available vertebrate genomes. The database is created on a platform that allows for constant growth to accommodate the dynamic nature of genome research where newly emerging genomes and improved releases of current genomes are constantly made available to the public. Currently, it includes data generated from 10 vertebrates: human, rhesus monkey, dog, opossum, rat, mouse, chicken, frog, pufferfish and zebrafish. In general, the number of ECRs in pairwise genome alignments reflects the evolutionary distance separating these genomes. For example, we observe 2.3 million (M) human/rhesus macaque ECRs and only 73 thousand (k) human/*Fugu* ECRs as the result of the larger evolutionary separation of humans and fish than humans and other primates. An exception to this trend is observed when species with dramatically different generation times are compared. For example, while humans and dogs are phylogenetically more distantly related than humans and rodents, human/dog comparisons reveal a greater degree of sequence conservation, due to the fact that rodents have a shorter generation time that have allowed for more opportunities to diverge their genomes (6-8). Correspondingly, the ECR coverage of the non-repetitive part of the human genome decreases 65-fold as we move from the most closely related genome to the most distantly genome in reference to the human genome, from 53.3% in the human-rhesus macaque to 0.8% in the human-fugu comparison (Figure 1). In contrast to the human genome, the variation in the number of ECRs and the genome coverage is relatively small for vertebrates occupying distant and distinct niches in the evolutionary tree. Consistent with this observation, the

number of ECRs in the *Fugu* genome slightly varies from 67k to 74k in comparison to six other vertebrate genomes (Table 1).

In general, the decrease in the number of ECRs observed as the evolutionary distance increases is different for coding and noncoding regions. For example, while over 80% of ECRs shared among mammals are noncoding in nature, over 75% of ECRs shared between humans and either fish or amphibians are coding (Figure 1). It has been previously reported that noncoding elements that are deeply conserved throughout the evolution of vertebrates have particular DNA signatures (4,9,10) and are tightly linked to developmental and transcription factor genes (4). To account for variation in divergence rates, the analysis of noncoding ECRs that flank genes from different functional categories requires the ability to dynamically select the species to be compared in loci evolving at different rates. Therefore, the availability of multiple genome comparisons provided by the *ECRbase* comes with an additional value by allowing the selection of the most informative species in comparisons for any locus in the human or any other available vertebrate genome.

While transcription is known to depend on promoter function, a paradigm that has long been established (11), increasing lines of evidence also highlight the importance of long-range/distant regulatory elements that are embodied by conserved elements present in the vicinity of genes of interest (1,3,12). To generate a resource that is all inclusive, *ECRbase* is not restricted to the analysis of promoter sequences, but instead comprises all conserved noncoding elements in any available genome. All ECR annotations in the *ECRbase* include length and percent identity demarcations that allow for the subsequent

selection of the most conserved noncoding ECRs in a locus of interest. Also, automatically pre-computing lists of coreECRs (9) [identified using 350 basepairs (bp)/77% percent identity (ID) threshold] are made available that can be used as candidate regulatory elements in loci of well conserved genes. We and others, have previously shown that coreECRs in comparisons of closely related vertebrates (between different mammals, for example) selectively identify elements that have a high probability of being conserved across large evolutionary distances (9,10).

Sequence analysis of noncoding ECRs and promoter elements is essential for searching for gene regulatory elements. Since the understanding of gene regulatory mechanisms requires the identification of transcription factors binding and acting on transcriptional regulatory elements, *ECRbase* provides detailed annotation of TFBS across all ECRs and promoter elements stored in the database. TFBS are identified using available libraries of transcription factor binding motifs or position weight matrices (PWM) from the most recent version of the TRANSFAC database (currently, version 9.4; http://biobase.de) (13) in combination with the previously described *tfsearch* TFBS mapping algorithm (14). Transcription factors tends to recognize and bind to short DNA motifs that usually range from 6 to 12 bp in length (15). Because of the highly degenerate nature of TFBS, it has been shown that computational annotation of TFBS can results in a large number of false positive predictions. To partially overcome this problem we are using a previously published method to decrease the number of false positive predictions by increasing the thresholds of TFBS mapping such that the number of TFBS annotations is minimized (14). Although the application of these thresholds decreases the number of false positive

predictions by an order of magnitude, still its application to entire genome datasets results in the identification of 4.8M and 73.5M TFBS in human promoters and human-mouse ECRs, correspondingly. Therefore, a statistical post-processing may be required to select TFBS that have a high likelihood of being functional. One post-processing strategy is to focus on associations of TFBS that are enriched in regions flanking co-functional or co-expressed genes (16-18). The *ECRbase* provides ECR information for both sequences being compared, therefore, the overlap of TFBS cohorts in orthologous ECRs could allow for the identification of actively conserved TFBS using phylogeny as a filter.

## DATABASE ORGANIZATION AND METHODS

The schematic structure of the *ECRbase* data analysis is presented in Figure 2. The database first processes whole genome pairwise alignments of multiple vertebrate genomes available from the ECR Browser database (19) to identify evolutionary conserved regions (ECRs). Currently there are over 26M ECRs available in the *ECRbase* that correspond to regions shared by all pairwise comparisons of all the available species which currently include: human, rhesus macaque, mouse, rat, dog, opossum, chicken, frog, zebrafish, and/or *Fugu* genomes. Next, these ECRs are used to determine synteny blocks that interconnect these genomes. Due to the fact that the identified synteny blocks are based on nucleotide alignments, not on protein similarity, and thus are capable of precisely demarcating synteny breakpoints in long intergenic regions, they can potentially provide more accurate synteny maps with longer syntenic stretches for closely related vertebrates (such as human and mouse, for example) then those that are restricted to gene comparisons. In parallel to the ECR identification we've implemented the extraction of

vertebrate promoters using RefSeq, knownGene, and "Other species RefSeq" gene annotations available from the UCSC Genome browser database (20,21). At the final step, DNA sequences of the identified ECRs and promoters undergo annotation of TFBS. All the processed data is collected, binned according to the corresponding genome, and distributed through the central *ECRbase* interface available at http://*ECRbase*.dcode.org. Large ECRs and TFBS files are compressed (using the 'gzip' utility) to facilitate data downloads. Despite the compression, some of the files are relatively large and, therefore, some users may find it helpful to use automated file download utilities for fetching data from the *ECRbase*. Below we summarize the details of methods employed for data extraction and generation.

***Evolutionary Conserved Regions.*** ECRs are computed as regions greater than 100bps in length and greater than 70% nucleotide sequence identity (Table 1). For a region to be classified as an ECR, it is required to be present in both species. There are cases when a conserved region in one species has accumulated significant insertions in the second species and, thus, its second species conservation falls below the threshold. Elements that exhibit this conservation pattern are excluded from the database. Stricter thresholds, of a minimum length of 350bps and conservation level of 77% ID are used for identifying conserved elements termed *coreECRs* – regions that are implied to have a higher probability of being functional than regular ECRs (9,10). *ECRbase* reports genome positional information of ECRs (and coreECRs), their length and percent identity as well as the corresponding parameters for their orthologues in other genomes.

***Synteny.*** Synteny between vertebrate genomes was determined as previously described (14). Briefly, we used sets of 3 consecutive ECRs (two neighboring ECRs were selected as 'consecutive' if they were separated by <100kb in both genomes) to define anchors of inter-genome synteny. These synteny anchors were used to construct larger synteny blocks by clustering ECR triplets from matching chromosomes using the same maximum 100kb separation threshold (Table 2). Since a great number of genomes are available in draft sequence format (in a multi-scaffold configuration), several artificial synteny breakpoints originate simply from the scaffold edges prematurely disrupting the synteny structure. Short scaffolds can also potentially prevent the identification of the 3-ECR synteny anchors thus also leading to the elimination of some synteny relationships and/or generation of incomplete syntenic blocks. Therefore, synteny assignments originating from unfinished genomes should be treated with caution.

***Promoters.*** *ECRbase* utilizes RefSeq and knownGene gene annotation available at the UCSC Genome browser database (21) to localize the genomic position and the strand of gene transcripts in vertebrate genomes. Overlapping transcripts are combined into unique genes and the outermost 5' end representing the most probable transcription start site (TSS) of the gene is identified. Next, the data extraction utility selects ≤1.5kb region upstream of the gene TSS, annotates it as the promoter element and automatically fetches the corresponding DNA sequence (repetitive elements are indicated by lower-case letters consistent with data representation in the UCSC Genome browser). Promoter elements are limited to intergenic spaces and are dependent on the location of neighboring genes. In cases where the intergenic region is significantly shorter than 1.5kb, the identified promoters span the entire intergenic space between the two

transcripts and are therefore less than 1.5kb. *ECRbase* reports positional and directional information of promoters as well as it provides the name of the gene the promoter is associated with. Bi-directional promoters (promoters shared by two genes transcribed in a head-to-head manner) are reported twice – once for each transcript.

***Transcription factor binding sites.*** We utilize TRANSFAC Professional database of position weight matrices or PWM (version 9.4) (13) to map candidate TFBS in genomic sequences. TFBS are mapped as previously reported, using the *tfSearch* (14) utility that employs a suffix tree technique to rapidly identify motifs in DNA sequences. In an effort to limit the number of false positive TFBS predictions we avoid using default PWM sequence similarity parameters, but instead perform an independent optimization of thresholds for different TFBS that warrants 5 or less TFBS predictions per 10kb of random sequence. Each ECR and promoter element undergoes a TFBS mapping, and positional and directional information of each TFBS inside these elements is collected afterwards and distributed through the corresponding portal of the *ECRbase*.

## SUMMARY AND FUTURE DIRECTIONS

*ECRbase* provides uniform access to evolutionary conserved regions extracted from pairwise comparisons of multiple vertebrate genomes. It also distributes information on synteny blocks that link vertebrate genomes along with genome-wide annotation of promoters and TFBS in promoters and ECRs. *ECRbase* is a resource that can facilitate studies of gene regulation and evolution on a multi-genome scale. Inter-species ECRs (especially those mapped to the human genome) can be utilized to prioritize the selection of functional elements for disease linkage studies as well as for primary targets of patient

and model organism re-sequencing projects. Pre-computed annotations of TFBS in ECRs and promoter elements provide a platform for studies of gene regulatory pathways and identification of *cis*-regulatory modules of TFBS that are linked to co-regulated genes. Datasets of ECRs and TFBS can be interchangeably coupled to simultaneously identify matching TFBS in two species and thus to identify TFBS that are phylogenetically conserved in different genome comparisons. As ECR Browser alignments follow the most current availability of genomic data, constant updates of sequenced vertebrate genomes at the UCSC Genome browser database propagate the generation of new ECR Browser alignments, and consequently lead to the follow up expansion and/or updating of the *ECRbase* database. In the future, we plan to expand the set of *ECRbase* features. Specifically, these developments will include the generation of single nucleotide polymorphisms (SNPs) dataset in human and other species ECRs, automated searches across all *ECRbase* data (including cross-species searches) by gene name or accession number, and implementation of improvements and other new features suggested by *ECRbase* users. This database is designed to serve as a community resource, therefore user input on ease of navigation and data retrieval and overall usefulness are vital for its future evolution.

## AVAILABILITY

*ECRbase* is publicly available at http://*ECRbase*.dcode.org for both academia and private sector. There are no limits on data downloads. This article should be cited in research projects that utilize *ECRbase* data.

**ACKNOWLEDGEMENTS**

G.G.L. and I.O were supported by LLNL LDRD-04-ERD-052 grant; and I.O. was in part supported by LLNL LDRD-06-ERD-004 grant. The work was performed under the auspices of the United States Department of Energy by the University of California, Lawrence Livermore National Laboratory Contract W-7405-Eng-48.

*Conflict of interest statement.* None declared.

**TABLES**

**Table 1.** Number of ECRs in inter-species alignments for the human (hg18), mouse (mm8), rat (rn4), dog (canFam2), chicken (gg2), frog (xt4), and fugu (fu4) genomes (in thousands).

| | **Dog** | **Mouse** | **Rat** | **Chicken** | **Frog** | **Fugu** |
|---|---|---|---|---|---|---|
| **Human** | 2,521 | 1,289 | 1,189 | 200 | 120 | 73 |
| **Dog** | | 1,042 | 972 | 178 | 115 | 71 |
| **Mouse** | | | 2,311 | 169 | 109 | 74 |
| **Rat** | | | | 162 | 107 | 70 |
| **Chicken** | | | | | 117 | 67 |
| **Frog** | | | | | | 73 |

**Table 2.** Longest synteny block size from inter-sepecies comparison of the human (hg18), mouse (mm8), chicken (gg2), frog (xt4), and fugu (fu4) genomes (in thousand basepairs, kb).

| **Second** / **Base** | **Human** | **Mouse** | **Chicken** | **Frog** | **Fugu** |
|---|---|---|---|---|---|
| **Human** | - | 56,101 | 48,475 | 9,233 | 9,656 |
| **Mouse** | 54,507 | - | 38,134 | 7,888 | 8,783 |
| **Chicken** | 19,105 | 16,106 | - | 7,952 | 4,704 |
| **Frog** | 5,479 | 5,333 | 6,529 | - | 3,529 |
| **Fugu** | 1,990 | 1,723 | 1,539 | 1,397 | - |

## FIGURE LEGENDS

**Figure 1.** Coverage of the human genome by ECRs (in Mb) from different species comparisons – rhesus macaque, dog, mouse, rat, opossum, chicken, frog, fugu, and zebrafish. Pie-charts of ECR binning into different gene features (coding, UTR, putatively coding – those that overlap only with an mRNA exon, or noncoding) accompany each interspecies comparison. Annotation of coding exons and UTRs is made using RefSeq and UCSC knownGene annotations (20,21).

**Figure 2.** Schematic pipeline of the *ECRbase* data analysis.

**Figure 1.**

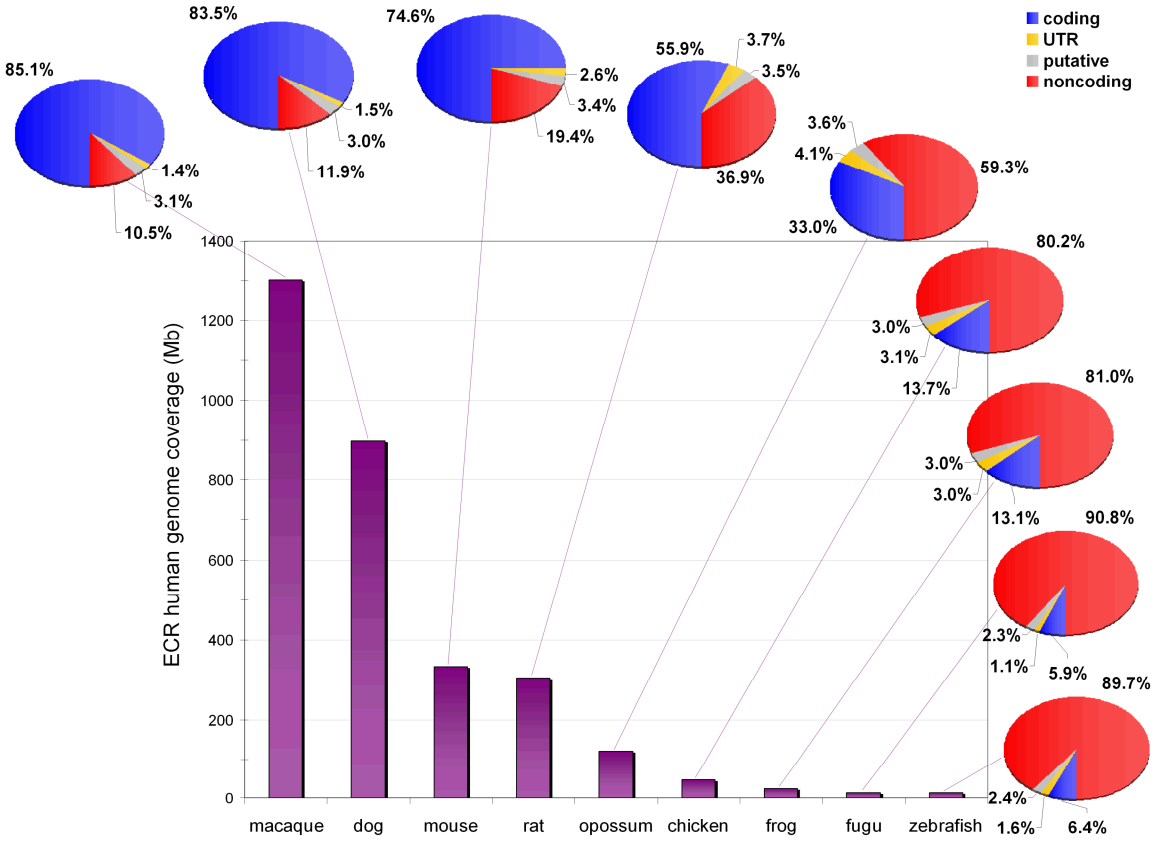
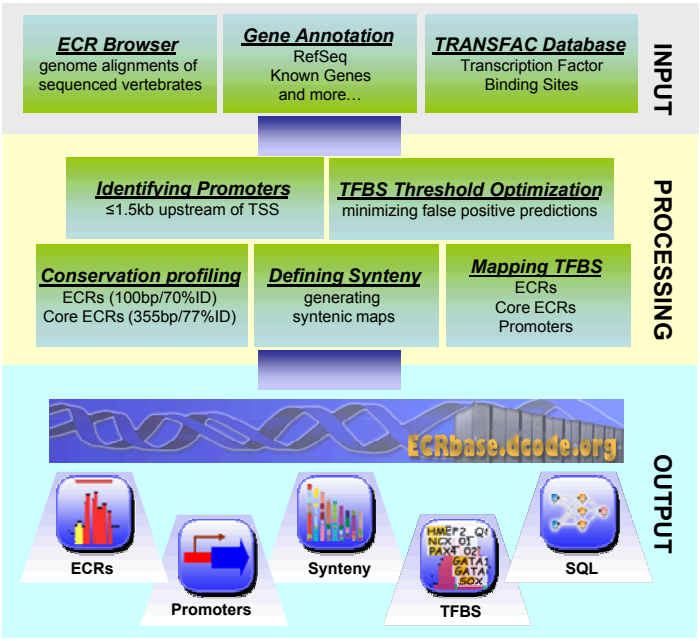

coding
UTR
putative
noncoding
ECR human genome coverage (Mb)
0
200
400
600
800
1000
1200
1400
macaque
dog
mouse
rat
opossum
chicken
frog
fugu
zebrafish
85.1%
1.4%
3.1%
10.5%
83.5%
1.5%
3.0%
11.9%
74.6%
2.6%
3.4%
19.4%
55.9%
3.7%
3.5%
36.9%
3.6%
4.1%
59.3%
33.0%
80.2%
3.0%
3.1%
13.7%
81.0%
3.0%
3.0%
13.1%
90.8%
2.3%
1.1%
5.9%
89.7%
2.4%
1.6%
6.4%

**Figure 2**


INPUT
ECR Browser
genome alignments of sequenced vertebrates
Gene Annotation
RefSeq
Known Genes
and more…
TRANSFAC Database
Transcription Factor Binding Sites
PROCESSING
Identifying Promoters
≤1.5kb upstream of TSS
TFBS Threshold Optimization
minimizing false positive predictions
Conservation profiling
ECRs (100bp/70%ID)
Core ECRs (355bp/77%ID)
Defining Synteny
generating syntenic maps
Mapping TFBS
ECRs
Core ECRs
Promoters
OUTPUT
ECRbase.dcode.org
ECRs
Promoters
Synteny
TFBS
SQL

## REFERENCES


1. Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M. and Frazer, K.A. (2000) *Science*, **288,** 136-140.
2. 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) *Science*, **294,** 169-173.
3. Nobrega, M.A., Ovcharenko, I., Afzal, V. and Rubin, E.M. (2003) *Science*, **302,** 413.
4. 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) *PLoS Biol*, **3,** e7.
5. Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavare, S. *et al.* (2005) *PLoS Genet*, **1,** e78.
6. Kirkness, E.F., Bafna, V., Halpern, A.L., Levy, S., Remington, K., Rusch, D.B., Delcher, A.L., Pop, M., Wang, W., Fraser, C.M. *et al.* (2003) *Science*, **301,** 1898-1903.
7. Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E. *et al.* (2004) *Nature*, **428,** 493-521.
8. 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) *Nature*, **420,** 520-562.
9. Ovcharenko, I., Stubbs, L. and Loots, G.G. (2004) *Genomics*, **84,** 890-895.
10. Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E.M., Couronne, O. and Pennacchio, L.A. (2006) *Genome Res*, **16,** 855-863.
11. Thomas, M.C. and Chiang, C.M. (2006) *Crit Rev Biochem Mol Biol*, **41,** 105-178.
12. Ghanem, N., Jarinova, O., Amores, A., Long, Q., Hatch, G., Park, B.K., Rubenstein, J.L. and Ekker, M. (2003) *Genome Res*, **13,** 533-543.
13. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K. *et al.* (2006) *Nucleic Acids Res*, **34,** D108-110.
14. Ovcharenko, I., Loots, G.G., Giardine, B.M., Hou, M., Ma, J., Hardison, R.C., Stubbs, L. and Miller, W. (2005) *Genome Res*, **15,** 184-194.
15. Loots, G.G. and Ovcharenko, I. (2004) *Nucleic Acids Res*, **32,** W217-221.
16. Qin, Z.S., McCue, L.A., Thompson, W., Mayerhofer, L., Lawrence, C.E. and Liu, J.S. (2003) *Nat Biotechnol*, **21,** 435-439.
17. Sharan, R., Ben-Hur, A., Loots, G.G. and Ovcharenko, I. (2004) *Nucleic Acids Res*, **32,** W253-256.
18. Perco, P., Kainz, A., Mayer, G., Lukas, A., Oberbauer, R. and Mayer, B. (2005) *Biosystems*, **82,** 235-247.
19. Ovcharenko, I., Nobrega, M.A., Loots, G.G. and Stubbs, L. (2004) *Nucleic Acid Research*.
20. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) *Nucleic Acids Res*, **33,** D501-504.

21. 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) *Nucleic Acids Res*, **34,** D590-598.