UCRL-JRNL-201727

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

ECR Browser: A Tool For Visualizing And Accessing Data From Comparisons Of Multiple Vertebrate Genomes

I. Ovcharenko, M. A. Nobrega, G. G. Loots, L. Stubbs

January 8, 2004

Nucleic Acids Research

Disclaimer

ECR Browser: A Tool For Visualizing And Accessing Data From Comparisons Of Multiple Vertebrate Genomes

Ivan Ovcharenko[1,2,*], Marcelo A. Nobrega[3], Gabriela G. Loots[1], and Lisa Stubbs[1,*]

[1]Genome Biology Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

[2]Energy, Environment, Biology and Institutional Computing, Lawrence Livermore National Laboratory, Livermore, CA 94550

[3]Department of Genome Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

* For correspondence

Email: ovcharenko1@llnl.gov, stubbs5@llnl.gov

Fax: (925) 422-2099

ABSTRACT

The increasing number of vertebrate genomes being sequenced in draft or finished form provide a unique opportunity to study and decode the language of DNA sequence through comparative genome alignments. However, novel tools and strategies are required to accommodate this increasing volume of genomic information and to facilitate experimental annotation of genome function. Here we present the ECR Browser, a tool that provides an easy and dynamic access to whole genome alignments of human, mouse, rat and fish sequences. This web-based tool (http://ecrbrowser.dcode.org) provides the starting point for discovery of novel genes, identification of distant gene regulatory elements and prediction of transcription factor binding sites. The genome alignment portal of the ECR Browser also permits fast and automated alignment of any user-submitted sequence to the genome of choice. The interconnection of the ECR browser with other DNA sequence analysis tools creates a unique portal for studying and exploring vertebrate genomes.

INTRODUCTION

The sequences of many vertebrate genomes including human, mouse, rat, and several fishes have recently been generated and assembled. With the exponential increase of sequencing performance and capabilities, the sequences of several other vertebrate genomes are expected to emerge in the near future. Several studies have underscored the value of alignments between orthologous sequences from different species, demonstrating clearly that conserved DNA segments provide a faithful guide in identification of functional sequence elements. This strategy has been validated both with the identification of novel genes and of functional noncoding elements (Elnitski et al. 2003; Loots et al. 2000; Pennacchio et al. 2001). While the comparative analysis of human and rodent sequences yeilds informative results in many cases (Elnitski et al. 2001; Hardison et al. 2003; Waterston et al. 2002) many genomic segments display either too little or too much conservation in such comparisons, due to the non-uniform structure and evolutionary rate across vertebrate genomes (Santini et al. 2003). Moreover, several cardinal features in the human genome are likely to have been acquired or shaped more recently than the human-mouse evolutionary separation (Eichler et al. 1998; Gardiner et al. 2003). These examples underscore the need of alternative comparative strategies that can accommodate the evolutionary asymmetries and architectural uniqueness of the human genome.

Several strategies have been devised recently to overcome these difficulties. In particular, the use of multiple species sequence comparisons has been proposed as an alternative to standard pairwise comparisons, aiming at the identification of a subset of sequences conserved in multiple species. Using this premise, a new method of multiple comparative sequence analysis was developed (Cooper et al. 2003), based on identification of an optimal dataset of species to compare that results in the best correlation between multiple conserved sequences (MCSes) and biologically relevant regions. A similar prioritization strategy, using comparisons between human sequence and that of a single distantly related species was recently shown to result in the enrichment of conserved elements corresponding to highly relevant functional sequences in megabase-long gene desert regions flanking human *DACH1* gene (Nobrega et al.

2003). Another strategy, called phylogenetic shadowing, has also been developed to detect and identify primate specific functional elements (Boffeli et al. 2003), which would not be detected in sequence comparisons of humans and rodents or more distant vertebrates. Finally, several other studies have emphasized the opposite category of elements, that is, those sequences that have arisen since the divergence of fish and primate lineages (Ghanem et al. 2003; Lettice et al. 2003; Nobrega et al. 2003). Taken together, these studies illustrate that no single comparative genomic strategy suffices for genome-wide comparative studies. Rather, there is a pressing need for the development of tools that can integrate sequences of multiple genomes in a custom-made fashion, allowing for a dynamic overlay of orthologous sequences from select numbers and types of species, as deemed necessary in a case-by case basis.

To fulfill this need, we have created a genome browser displaying multiple alignment of genomic sequence of various sequenced species including human, rodents and fish. This tool, called the ECR Browser, presents a dynamic representation of sequence comparisons, allowing for user-specified optimal analysis of genomic regions with differential divergence rates. Two main goals have driven the creation of the ECR Browser: (1) fast speed for user-specific genome alignments, and (2) flexibility to readily adjust alignment parameters including the number and type of genomes being compared, the genome to use as the “base” against which other genomes are compared, types of annotation to be displayed, thresholds for identification of significantly conserved sequence elements, and other features that permit the user to tailor comparisons specifically to regions characterized by different evolutionary rates. Furthermore, the

ECR browser is designed to permit incorporation of novel genomes immediately as their sequence becomes available in public databases.

ALIGNING GENOMES

Several strategies have recently been developed to analyze complete sequences of different genomes, from microbes to higher vertebrates (Couronne et al. 2003; Delcher et al. 2002; Karolchik et al. 2003; Schwartz et al. 2003). For the creation of the ECR Browser, we employed a strategy of genome alignment that is based on four consecutive sequence management steps. Briefly, after masking of repetitive elements, all the genomes were mapped pairwise, to establish large-scale syntenic relationships. Subsequently, each syntenically homologous pair of sequences was aligned. Finally, all the data was collected and stored in a central database that is then utilized by the ECR Browser to construct conservation profile graphs at the user's specification.

The main source of underlying genomic data utilized by the ECR Browser comes from the UCSC Genome Browser (Karolchik et al. 2003). In addition to the sequences of the human, mouse and rat genomes adopted from the UCSC Genome Browser we augmented the genome dataset with sequences of three fish genomes, namely *Fugu rubripes* (http://www.jgi.doe.gov/fugu/), *Tetraodon nigroviridis* (http://www.genoscope.cns.fr/externe/tetraodon/Ressource.html) and *Danio rerio* (http://www.sanger.ac.uk/Projects/D_rerio/). Repetitive elements in these genomes were identified and masked, marked either using a lower-case notation when available or by a local run of the RepeatMasker program (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker). Over millions of years of evolution since primate, rodent and fish speciation events, multiple large-scale rearrangements into the genomes of vertebrate organisms have been introduced. To identify related syntenic blocks in these divergent species, each pair of genomes was first mapped to each other. The dramatically different evolutionary history that separates primates and rodents from fishes, compared with the evolutionary separation between fishes or within the primate and rodent lineages, requires the application of different approaches to genome alignments of various types. For mapping syntenic homologies between more closely related species, such as humans and rodents, we used a locally installed version of the *BLAT* tool (Kent 2002). For comparative mapping of more distantly related species, such as humans and fishes (or rodents and fishes), the more sensitive but slower *blast* tool was employed (Altschul et al. 1990). At the final step of synteny mapping neighboring short hits of similarity were joined into large blocks of synteny (see Supplementary materials for details). Finally pairs of homologous sequences from each syntenic block were aligned with the use of the *blastz* alignment tool, with long alignments being cleaned from non-diagonal spurious hits [Suppl. Materials].

From a technical viewpoint, alignments of the human and mouse genomes (as an example) utilize 50Mbs of disk space (that is significantly less than the size of the original genome FASTA files) and require less than a week on a P4-processor machine to be created. This is significantly faster than any other genome alignment strategy previously reported (Couronne et al. 2003; Schwartz et al. 2003). This scale up in performance and significant savings of disk space permit us to have multiple genome alignments on hand with a relatively short response time to update the ECR Browser as new assemblies of genomes are released.

VISUALIZATION AND DATA BROWSING SCHEME

The conservation-profile visualization scheme of the ECR Browser tool is based on an idea originally implemented into the PipMaker tool (Schwartz et al. 2000) and later adopted by both Vista (Mayor et al. 2000) and zPicture tools (Ovcharenko et al. in press). In this model, the base genome sequence is schematically linearized as the horizontal axis, while the vertical axis represents the percentage identity between the base sequence and the sequence being compared (Figure 1). Evolutionary conserved regions (ECRs) are differentiated from the neutrally evolving background and are colorized depending on their identities, as protein coding exons, UTRs, introns, repetitive elements or conserved intergenic regions.

The ECR Browser dynamically constructs graphical conservation profiles for any region in the genome, which can be specified by either a gene name or by absolute genomic coordinates of the region of interest. Depending on user preferences the browser augments the conservation profile with an annotation of different genomic features, such as known genes, gene predictions, repetitive elements and single nucleotide polymorphisms, with annotations downloaded directly from UCSC Genome Browser. Other browser features such as zooming, shifting and re-centering allow for the rapid conversion of the genomic size and coordinates being analyzed (Figure 1).

To accommodate for the non-uniform evolutionary structure of the genomes of higher vertebrates, a flexible definition of ECR parameters was implemented in the browser. Display variability allows the user to require high stringency parameters in detecting ECRs in slowly evolving genomic regions or less stringent parameters to identify barely distinguishable, short ECRs in alignments, for example in rapidly diverging regions or in comparative analysis of distantly related species. Additional display customization is incorporated the permit selection of genomes to be added or removed from the comparative analysis display, so that, for example, only alignments of closely related species could be utilized in rapidly evolving genomic loci. Other custom features include the format of the displayed conservation plot (either a pip-plot or a smooth graph), a selection of different types of gene annotation, and selection of picture display parameters (Figure 2).

The ECR Browser is directly connected to genome sequence, readily providing user access to the underlying DNA sequence that corresponds to the genomic region being displayed. Also, the browser provides access to the sequence and a list of genomic positions of the ECRs detected in a region under a specified set of alignment conditions. To provide ready access to individual ECRs, we introduced the “Grab ECR” option to the browser which allows “one-click” access to any selected ECR in the conservation plot. This option connects to a detailed ECR description page containing ECR sequences from both species in any pairwise comparison, and a display of the underlying sequence alignment. In addition, sequence characteristics such as length, percent identity, G+C content and genomic coordinates are listed and are accompanied by links to the analysis of potential transcription factor binding sites inside the ECR, through the rVista program (Loots et al. 2002) (Figure 3).

SYNTENY

As a by-product of the synteny mapping that is required for accurate comparative alignements, the ECR browser is able to locate and reconstruct syntenic breakpoints, establishing relationships that can be later utilized to navigate between different genomes (Figure 4A). Using the syntenic alignments link, it is possible to jump from the display of a specific locus in one genome directly to the visualization of the syntenically homologous locus in another aligned genome (Figure 4B). This option permits users to compare size, organization, conserved features at the same locus in divergent genomes. It also permits users to compare ECRs arising in comparisons between human and mouse, for example, to those detected in the same genomic locus when rat and mouse genomes are compared.

The identity of the “base genome” in the display of a particular region can also be readily changed with the use of the ‘Base Genome’ feature. In contrast to the corresponding function available through the synteny links page, this option does not direct the user to a location in the newly selected genome, that is syntenically related to the original locus, and all the parameters including the genomic location and set of species involved in the alignment will be changed to the default values for the new base genome. This approach excludes uncertainties that arise due to multiple regions of paralogous homology between certain genomic regions; by contrast all significant regions of syntenic homology paralogy in each aligned genome are reported in the synteny links page and can be analyzed using that link.

GENOME ALIGNMENT

While pre-computed alignments of the genomes available in the ECR Browser are sufficient for multiple tasks, we also created custom-defined alignment options within the browser that allows the instantaneous alignment of any user-defined sequences. Such queries may be submitted either directly, in FASTA format, or automatically downloaded from GenBank using the accession number of the sequence to be aligned, which is then forwarded to the 'Genome Alignment' portal in the browser. Upon receiving the user-submitted sequence, the ECR Browser will rapidly map this sequence to the selected genome (either human, mouse, rat, or *Fugu* genome) using the *BLAT* tool. When the syntenically homologous region is identified in the selected "base" genome, the region is extracted along with the corresponding RefSeq gene annotation. *Blastz* alignments of the two sequences are made and a dynamic graphical visualization of the alignments is generated (Figure 5). A dynamic zPicture (Ovcharenko et al. in press)conservation plot, a portal to the rVista tool (Loots et al. 2002), an alignment dot-plot and a tool for dynamic annotation of ECRs in the alignment are also automatically provided.

INTEGRATION WITH OTHER TOOLS

We intend to maintain the ECR Browser as a constantly updated tool that not only incorporates newly deposited and annotated sequences, but also provides direct connections to the growing set of publicly available external sequence analysis tools. Presently, an extensive annotation of known genes, gene predictions, experimental RNA evidence and many other features is available through the direct interface between the

ECR browser and the Genome Browser at UCSC (Karolchik et al. 2003) and the Ensembl Genome Browser (Birney et al. 2004; Hubbard et al. 2002). This portal permits the user to examine any non-genic conservation pattern against the UCSC evidence database on putative novel genes and noncoding RNAs. Also, the 'Synteny/Alignments' link of the ECR Browser brings user to the zPicture analysis web page, described above, offering an easy and fast way to distill a chosen pair-wise alignment out of the multiple genome alignments. Using the zPicture features various modifications can be applied to the alignment including the annotation feature which permits manual annotation for regions not annotated in the ECR Browser (for example, incorporating user-generated data), or editing of public annotation to add features retrieved from other experimental or computational sources.

As mentioned previously the ECR Browser is also interconnected with the rVista tool (Loots et al. 2002). rVista is capable of filtering out up to 95% false positive TRANSFAC (Wingender et al. 1996) predictions of transcription factors binding sites (TFBS) while preserving high sensitivity of the search. The rVista portal provides a unique opportunity to predict the function of a noncoding element. By identifying evolutionary conserved TFBS in an ECR, the rVista portal provides a basis for experimental testing and application of the known function of the conserved transcription factors towards understanding the function of a neighboring gene. Any pairwise alignment from the ECR Browser can be automatically submitted for rVista analysis via the 'Synteny/Alignment' link. Also, any ECR retrieved with the use of the 'Grab ECR' function can be submitted directly to rVista for binding-site analysis.

CONCLUSIONS

The ECR Browser tool is designed to highlight candidate functional noncoding elements and to visualize their genomic positions relative to the gene features in the genome. By using comparisons with multiple species from selected evolutionary clades, the ECR Browser provides flexibility in assessing evolutionary fates of noncoding sequences, allowing for comparisons that reflect sequence conservation over a range of time scales and in species with both shared and lineage-specific biological features. Comparisons with distant organisms, such as fish, will likely uncover the fundamental building blocks shared by all vertebrates, while the comparative sequence analysis with closer species such as rodents will highlight the functional structure of rapidly diverging genomic regions, including those that dictate lineage-specific traits, or sequences that are specific to mammals.

Because it is directly connected to other publicly available sequence analysis tools, the ECR Browser provides the user with an easy automated access to resources permitting a thorough annotation of functional elements in the genome (through the portal to the UCSC Genome Browser) and to the annotation of transcription factor binding sites (through the portal to the rVista tool). Because the underlying algorithms and tools that power the ECR browser are designed to permit rapid updates, the tool will be constantly updated with new sequence and new links to other relevant sequence analysis sites. These features, the ease with which conservation parameters and included datasets can be changed by the user, and the immediate dynamic display of alignment results, make the ECR browser a powerful new addition to the computational toolkit for annotating functional features in the human sequence and in other genomes sequenced now or in future years.

ACKNOWLEDGEMENTS

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.

Figure 1. ECR Browser visualization of the Lim Domain Only 1 (LMO1) gene 46kb loci in the human genome located at chr11:8209000-8255200 (UCSC freeze 16; NCBI Build 34). Conservation profile of the human region in a comparison with the mouse, rat, fugu, tetraodon and zebrafish genomes. Five genomes that were compared to the human original region are plotted as 5 horizontal layers of conservation diagrams and the small image icon at the right sided of the plot represents a specie corresponding to the alignment. Each layer contains a pip-type plot tha consists of a multiple short horizontal black lines. Each of the lines represents an ungapped alignment with the vertical height of the line describing the nucleotide identity underlying an alignment. LMO1 gene is depicted in blue and yellow colors with the blue bars depicting the exons that are involved into the protein coding and yellow bars describing the UTRs. Direction of the gene is given by arrow lines. Dark red bar on top of every layer provides an overview of the distribution of ECRs and is used to colorize underlying alignments. A conserved alignment is blue if it overlaps with a coding exon, yellow – UTR, pink – intron, red – intergenic region. The green bars at the bottom that are shaded to the top in gray indicate repetitive elements in the base sequence. The top bar of the browser provides with the quick-links to different chromosomes while the left bar represents a dynamic chromosomal map. Mouse-click on the top bar will result in the change of the chromosome and the click on the chromosome image in the left bar will shift the browser window to that locus on the chromosome. Right <UCSC Browser> button will direct the user to the visualization of the same genomic interval in the UCSC Genome Browser (Karolchik et al. 2003). The bottom set of multiple buttons provides with zooming and shifting capabilities accompanied with the 'Grab ECR' function that will be described later. The dynamics of the browser is also achieved through the browser plot being recentered at the position of the mouse click on the plot.

Figure 2. ECR Browser settings allow flexible approach towards analysis of differently evolving genomic regions and dynamic choice of the plot and alignment settings.

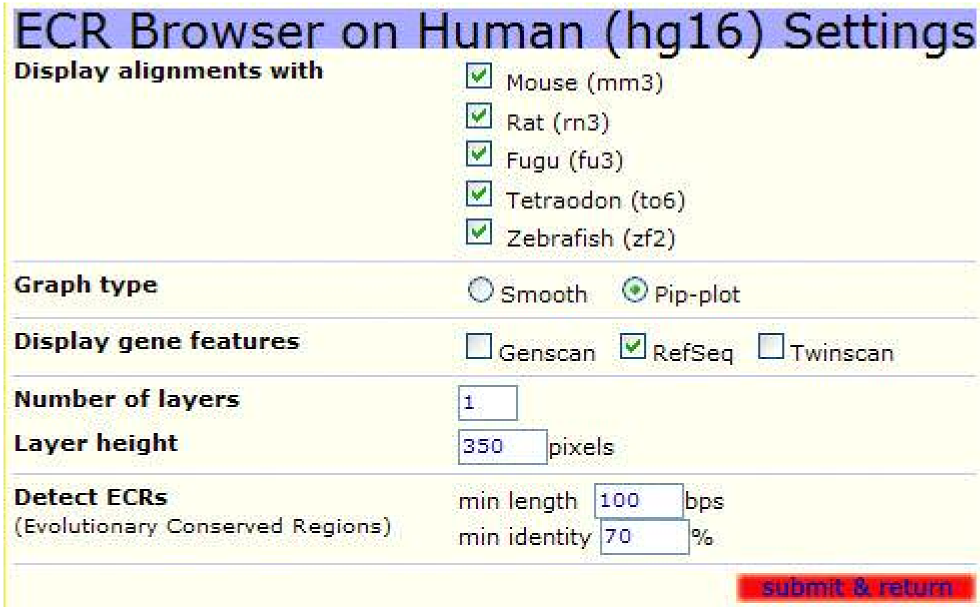


Figure 3. 'Grab ECR' feature – an access to the sequence, alignment and sequence analysis tools for a single ECR.

ECR :: Evolutionary Conserved Region	**Human**[hg16] - **Mouse**[mm3] **ECR**
Type	**516** bps, **72.3%** identity
Location	**chr10:8117671-8118186**[hg16]
Transcription factor binding sites	rVista
Oligo/primers design	Human Mouse
GC count	Human .. GC: 45.54%, AT: 54.46%, OTHER: 0.00% Mouse .. GC: 47.13%, AT: 52.87%, OTHER: 0.00%

Alignment

```
                  10        20        30        40        50        60
Human     GTGTATTgAACAaTaaGAGATAATAATCTAttaaCATTgTCaTcaCGTtGcGtTtTgCTC
          ||||||| |||| |  ||||||||||||||    |||| || |  ||| | | | | |||
Mouse     GTGTATTtAACAcT--GAGATAATAATCTAaggcCATTtTCtTggCGT-GtGaTgTcCTC
     3050       3040        3030      3020      3010       3000

                  70        80        90       100        110
Human     TGCCCtTcCaGacaTCTctACATGgAtGCCATaaGCTCtT-TCtTCTTATCTAGGTGTTg
          ||||| | | |   |||  ||||| | |||||  |||| | || |||||||||||||||
Mouse     TGCCCaTaCtG-tgTCTgcACATGtAaGCCATggGCTCcTgTCcTCTTATCTAGGTGTTt
          2990      2980      2970      2960      2950      2940
```

Figure 4. Syneny links in the ECR Browser - GATA3 human gene synteny in mouse, rat and fugu genomes (plot A). Human (hg16, chr10:8101472-8120859; plot B) region is linked to the orthologous regions in the mouse (mm3, chr2:9837229-9857435; plot C) and rat (rn3, chr17:80002273-80023251; plot D) genomes. Four genome comparisons (human, mouse, rat, and fugu) are present in all three cases with a difference in the base genome.

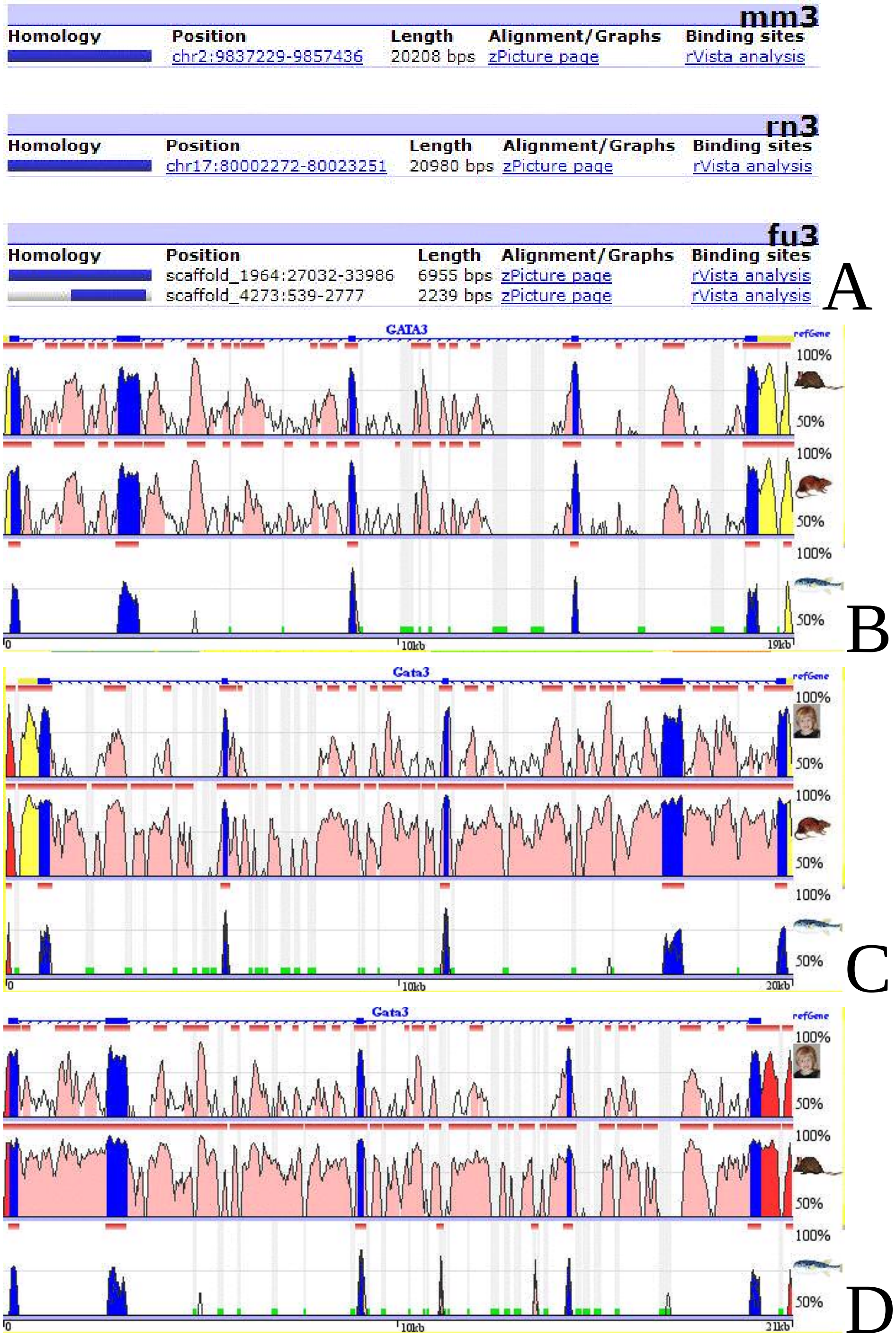
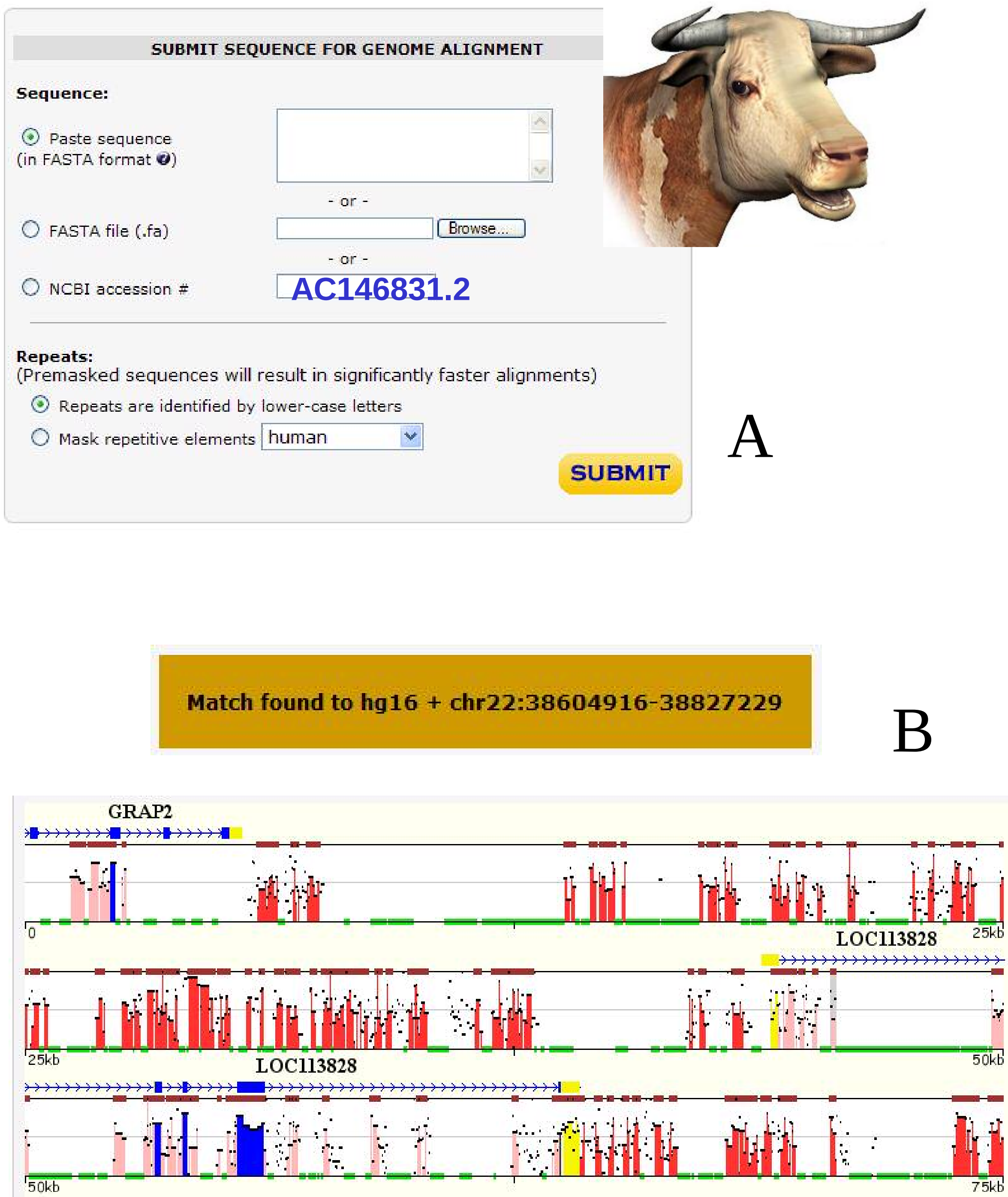


Figure 5. Genome alignment portal of the ECR Browser tool. Genomic sequence from any specie that is submitted either as a FASTA file or automatically downloaded from GenBank by an accession number can be mapped and aligned to either human, mouse, rat or fugu genomes (plot A). A genome alignment of a cow sequence identified by AC14683 accession number with the human genome (hg16 freeze; plot B)



REFERENCES.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Birney, E., D. Andrews, P. Bevan, M. Caccamo, G. Cameron, Y. Chen, L. Clarke, G. Coates, T. Cox, J. Cuff, V. Curwen, T. Cutts, T. Down, R. Durbin, E. Eyras, X.M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. Hotz, V. Iyer, A. Kahari, K. Jekosch, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, C. Woodwark, M. Clamp, and T. Hubbard. 2004. Ensembl 2004. *Nucleic Acids Res* **32:** D468-470.

Couronne, O., A. Poliakov, N. Bray, T. Ishkhanov, D. Ryaboy, E. Rubin, L. Pachter, and I. Dubchak. 2003. Strategies and tools for whole-genome alignments. *Genome Res* **13:** 73-80.

Delcher, A.L., A. Phillippy, J. Carlton, and S.L. Salzberg. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30:** 2478-2483.

Eichler, E.E., S.M. Hoffman, A.A. Adamson, L.A. Gordon, P. McCready, J.E. Lamerdin, and H.W. Mohrenweiser. 1998. Complex beta-satellite repeat structures and the expansion of the zinc finger gene cluster in 19p12. *Genome Res* **8:** 791-808.

Elnitski, L., R.C. Hardison, J. Li, S. Yang, D. Kolbe, P. Eswara, M.J. O'Connor, S. Schwartz, W. Miller, and F. Chiaromonte. 2003. Distinguishing regulatory DNA from neutral sites. *Genome Res* **13:**64 -72.

Elnitski, L., J. Li, C.T. Noguchi, W. Miller, and R. Hardison. 2001. A negative cis-element regulates the level of enhancement by hypersensitive site 2 of the beta-globin locus control region. *J Biol Chem* **276:** 6289-6298.

Gardiner, K., A. Fortna, L. Bechtel, and M.T. Davisson. 2003. Mouse models of Down syndrome: how useful can they be? Comparison of the gene content of human chromosome 21 with orthologous mouse genomic regions. *Gene* **318:** 137-147.

Ghanem, N., O. Jarinova, A. Amores, Q. Long, G. Hatch, B.K. Park, J.L. Rubenstein, and M. Ekker. 2003. Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters. *Genome Res* **13:**533 -543.

Hardison, R.C., 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, and D. Haussler. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* **13:**13 -26.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. 2002. The Ensembl genome database project. *Nucleic Acids Res* **30:** 38-41.

Karolchik, D., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31:** 51-54.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

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

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

Loots, G.G., I. Ovcharenko, L. Pachter, I. Dubchak, and E.M. Rubin. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res* **12:**832 -839.

Mayor, C., M. Brudno, J.R. Schwartz, A. Poliakov, E.M. Rubin, K.A. Frazer, L.S. Pachter, and I. Dubchak. 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046-1047.

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

Ovcharenko, I., G.G. Loots, R.C. Hardison, W. Miller, and L. Stubbs. in press. zPicture: Dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res*.

Pennacchio, L.A., M. Olivier, J.A. Hubacek, J.C. Cohen, D.R. Cox, J.C. Fruchart, R.M. Krauss, and E.M. Rubin. 2001. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294:** 169-173.

Santini, S., J.L. Boore, and A. Meyer. 2003. Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. *Genome Res* **13:** 1111-1122.

Schwartz, S., W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler, and W. Miller. 2003. Human-mouse alignments with BLASTZ. *Genome Res* **13:** 103-107.

Schwartz, S., Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. 2000. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* **10:**577 -586.

Waterston, R.H. 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.K. Karlsson D. Karolchik A. 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 and E.S. Lander. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Wingender, E., P. Dietze, H. Karas, and R. Knuppel. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24:** 238-241.