LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-202112

rVISTA 2.0: Evolutionary Analysis of Transcription Factor Binding Sites

Gabriela G. Loots, Ivan Ovcharenko

January 30, 2004

Nucleic Acids Research

rVISTA 2.0: Evolutionary Analysis of Transcription Factor Binding Sites

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

[1]Genome Biology Division

[2]EEBI Computing Division

Lawrence Livermore National Laboratory

7000 East Avenue, L-441

Livermore, CA 94550

[*]For correspondence:

Phone: (925) 422-4723

Fax: (925) 422-2099

Email: loots1@llnl.gov, ovcharenko1@llnl.gov

ABSTRACT

Identifying and characterizing the patterns of DNA *cis*-regulatory modules represents a challenge that has the potential to reveal the regulatory language the genome uses to dictate transcriptional dynamics. Several studies have demonstrated that regulatory modules are under positive selection and therefore are often conserved between related species. Using this evolutionary principle we have created a comparative tool, rVISTA, for analyzing the regulatory potential of noncoding sequences. The rVISTA tool combines transcription factor binding site (TFBS) predictions, sequence comparisons and cluster analysis to identify noncoding DNA regions that are highly conserved and present in a specific configuration within an alignment. Here we present the newly developed version 2.0 of the rVISTA tool that can process alignments generated by both zPicture and PipMaker alignment programs or use pre-computed pairwise alignments of seven vertebrate genomes available from the ECR Browser. The rVISTA web server is closely interconnected with the TRANSFAC database, allowing users to either search for matrices present in the TRANSFAC library collection or search for user-defined consensus sequences. rVISTA tool is publicly available at http://rvista.dcode.org/.

INTRODUCTION

Unlike most prokaryotic genomes that are composed of tightly packed gene units with limited intergenic regions, eukaryotic genomes are rich in noncoding sequences of unknown functions. Extensive annotation of the human and mouse genomes has predicted in the vicinity of ~40,000 genes which account for less than 5% of the genome. An additional 40-45% of the mammalian genome is comprised of repetitive DNA, while the remaining 50% is noncoding in nature (Venter et al. 2001; Waterston et al. 2002). First glimpses at the human genome have revealed very few insights on new RNA coding genes, transcriptional regulatory elements or any other biologically relevant sequences present in noncoding regions. Although some parts of the noncoding genome will demonstrate no measurable biological functions, it is widely assumed that much of our genetic complexity is due to sophisticated regulatory noncoding signals that determine when, where and how much transcriptional activity each gene displays. Despite the importance of noncoding sequences in gene regulation, our ability to computationally identify and characterize these elements is very limited.

In multi-cellular organisms, modulation of gene expression is accomplished through the compound interaction of regulatory proteins (transcription factors) and specific DNA regions (*cis*-regulatory sequences or modules) they physically interact with. Numerous DNA footprinting studies carried out over the last decade have identified close to five hundred vertebrate specific transcription factors (TF), and the DNA sequences they recognize. The TRANSFAC database (http://www.biobase.de) (Matys et al. 2003; Wingender et al. 1996) represents the most comprehensive collection of TF binding specificities, summarized as position weight matrices (PWM). A major limitation for using PWMs to computationally identify functional transcription factor

binding sites (TFBSs) is that TFs bind to short degenerate sequence motifs (6-12 base pairs). These sequences occur very frequently in a genome, and experimentally it has been shown that only a very small fraction of these predicted TFBS are functionally relevant.

We have previously shown that the rVista tool combines pattern recognition with comparative sequence analysis to dramatically reduce the number of false positives TFBS matches (up to 95%) while the number of functional sites is minimally affected (decreased by less than 13%) (Loots et al. 2002). These results suggest an alternative strategy for sequence-based discovery of biologically relevant regulatory elements. To increase its versatility, and create a more efficient and user-friendly tool, here we present rVISTA 2.0, an improved web-based server that interconnects TFBS motif searches and cross-species sequence analysis with several comparative sequence analysis tools to significantly simplify and expedite its use. Originally, rVISTA required external alignment files to be submitted for analysis and was limited to only one alignment format. Also, we designed a new program for detecting TFBS that is significantly faster than the MATCH program originally accompanying the TRANSFAC database (Kel et al. 2003; Matys et al. 2003). This new development significantly decreases the processing time enabling the analysis of much larger genomic intervals.

ALGORITHM

There are three major venues for entering the rVISTA tool: (1) submitting a PipMaker alignment file (http://bio.cse.psu.edu/pipmaker/) at the rVISTA homepage (http://rvista.dcode.org/), (2) dynamically generating and automatically forwarding (with a single mouse button click) zPicture alignments (http://zpicture.dcode.org/)

(Ovcharenko 2004) or (3) accessing pre-computed multiple genome alignment data available at the ECR Browser (http://ecrbrowser.dcode.org/) (Figure 1A). All these three tools providing alignments for the rVista 2.0 use the blastz program (Schwartz et al. 2003) to identify homologous regions and to produce local sequence alignments between the reference sequence and one or more other orthologous sequences. The local alignment method used by PipMaker, zPicture, and the ECR Browser tools provides a careful assessment of the evolutionary rearrangements ensuring the ability of rVista to detect TFBS that underwent evolutionary positional changes.

rVISTA analysis proceeds in four major steps: (1) detect TFBS matches in each individual sequence using PWM from TRANSFAC database, (2) identify pairs of locally aligned TFBS, (3) select TFBS present in regions of high DNA conservation and (4) create a graphical display that dynamically overlays individual or clustered TFBSs with the conservation profile of the genomic locus. Users have the option of either selecting matrices from the TRANSFAC library or inputting their own TFBS consensus sequences. The current TRANSFAC library utilized by rVista 2.0 contains representatives from ~500 vertebrate TF matrices that comprise ~400 TF families. Selected matrices from this library are additionally verified and improved. Users selecting TRANSFAC library have the option to specify the stringency to be used for the PWM identification.

We have replaced the MATCH (Kel et al. 2003) program accompanying the TRANSFAC (Matys et al. 2003; Wingender et al. 1996) database with a recently developed tfSearch tool for detecting TFBS [Ovcharenko I., unpublished]. tfSearch combines "suffix tree"-based fast substring searches (Delcher et al. 2002) with PWM scoring of substring similarities. Transforming the original sequence into the

suffix tree could use extensive memory (requiring ~100 times larger memory slot than the size of the sequence), but is highly efficient in localizing substrings. A substring of the size N will require N or less operations with the suffix tree in order to localize all the matches. PWM searches that use the suffix tree require a scan of the suffix tree at a depth less or equal to N and stop when the count at the node is below the PWM matrix similarity threshold selected by the user. Table 1 summarizes results of PWM detecting TFBS in two genomic loci, 100kb and 1Mb long, utilizing MATCH and tfSearch tools. The gain in speed obtained with the use of the tfSearch tool varies from 10- to 100-fold in comparison with the time required by the MATCH program. It is especially pronounced when a large number of PWMs is used. The speed improvement introduced into the rVista 2.0 tool significantly decreases the tool's response time due to the fact that detecting TFBSs in the sequence file is the performance bottleneck of this approach.

After localizing the TFBSs in both sequences, rVista proceeds with identifying pairs of ***aligned*** TFBS that are interconnected in the local blastz alignment. Genomic DNA insertions and deletions in either one of the sequences (that are identified as gaps in the alignment) that occur in the core part of a TFBS disqualify the prediction. Subsequently, rVISTA requires aligned TFBS predictions to be locally highly conserved. Local conservation of at least 80% sequence identity in a 20 bps sliding window spanning the binding site (that always include the core of the binding site) selects ***aligned-and-conserved*** TFBSs.

The final rVista web page contains detailed information on rVista processing results. This includes positional information on TFBSs prediction in both

sequences, distribution of *aligned* and *aligned-and-conserved* TFBSs. The report includes data on the location, percent identity and strand (Figure 1B) (reference sequence only). Conserved sites can also be visualized in the textual blast-like alignment, and are highlighted in blue. Finally, rVISTA results provide an interactive visualization module that allows overlaying positional information on TFBS predictions on top of a graphical conservation profile that includes annotation of protein coding features for the locus. Clustering analysis of TFBS permits for the search and subsequent visualization of complex TFBSs modules consisting of multiple different TFBSs (Figure 1C). For more informative analysis, users have the option to select for visualization only a subset of TF from the initial list provided. Several parameters can be adjusted: (1) alignment size (in bp) per layer, (2) window resolution, (3) types of sites to be displayed (all, aligned, conserved), and (4) the type of clustering analysis to be used. Two clustering options are available, individual and combinatorial. Individual clustering is used for identifying groups of TFBS belonging to the same family of TFs. Users have the option to indicate the number of sites and the size of the TFBS module they wish to identify. Combinatorial clustering is carried out for groups of TFBS belonging to two or more TF families. If the visualization module has been selected to display TFBS for 3 different TFs, and the users is interested in finding 100 bp TFBS modules with clusters of 5 sites, rVISTA will identify all ECRs with any combination of these sites. In the visual display rVISTA will present only sites that fit the selected criteria (Figure 1C).

APPLICATION

To illustrate the application of the rVISTA tool, we have carried out an unbiased analysis for the *NKX2.5* human locus with the intent to detect the regulatory element

known to play key role in cardiac development. The conservation profile available for this gene in the ECR Browser revealed several upstream and intronic noncoding elements in this locus (Figure 2A). rVISTA analysis of the ECR Browser alignment containing a ~7kb *NKX2.5* region was performed using Smad4 matrices. A TFBS search with a 0.85 PWM matrix cut-off identified 43 PWM matches across the locus, 4 of which are highly conserved in the human-mouse alignment (Figure 2B). All these 4 Smad4 TFBSs are localized inside of the single conserved element located ~2kb upsteam of the *NKX2.5* transcription start site. This highly conserved element has been previously shown to function as a cardiac enhancer in transient transgenics. In particular, one of these conserved *Smad4* TFBS coincides with the site mutated by Lien *et al* that is required for the proper activity of the *NKX2.5* cardiac enhancer (Figure 2C). It was also demonstrated that a two base pair mutation (from A to C) in the most highly conserved Smad4 TFBS was able to diminish the cardiac enhancer properties of this regulatory element (Figure 2C) (Lien et al. 2002).

CONCLUSIONS

Understanding the function of noncoding DNA, identifying and characterizing the structure of transcriptional regulatory elements embedded in the human genome creates a continuing challenge. We present a completely redeveloped rVISTA 2.0 web server, for high-throughput discovery of *cis*-regulatory elements. By combining interspecies sequence conservation, reliable TF matrices and combinatorial clustering of transcription factor binding sites (TFBSs), rvista 2.0 maximizes the probability of identifying functional TFBSs. The novel features and programs implemented into rvista 2.0 make this tool very powerful for identifying and analyzing TFBSs in long

genomic intervals. The interconnectivity with PipMaker, zPicture and the ECR Browser tools for genome comparative sequence analysis makes rVista 2.0 a valuable resource for establishing a direct link between the language of noncoding DNA and biological function of genomes.

ACKNOWLEDGEMENTS

We are grateful to Lisa Stubbs for constant support, critical comments and suggestion on the manuscript. 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 LEGENDS

Figure 1. rVISTA analysis data flow. rVISTA tool can be accessed via PipMaker (http://bio.cse.psu.edu/pipmaker/) and zPicture (http://zpicture.dcode.org/) through the local alignment of homologous sequences, or from the ECR Browser (http://ecrbrowser.dcode.org/) to utilize pre-computed alignments of seven vertebrate genomes that include human, mouse, rat, frog and three fish genomes (A). Users select the search criteria, and the results are returned in the same page as downloadable static data files and dynamic links to visual analysis of TFBS distribution (B). TFBS for pre-selected TFs can be visualized above the conservation profile as tick marks, and the clustering module can detect user-specified groups of TFBS (C).

Figure 2. TFBS analysis of *NKX2.5* genomic locus. *NKX2.5* genomic region was accessed in the ECR Browser (A). Human/Mouse and Human/Rat alignments are displayed (7kb in the window). Coding exons are in blue, untranslated regions (UTRs) are in yellow, conserved intronic noncoding ECRs are in pink and conserved intergenic ECRs are in red. The alignment was processed for Smad4 binding sites (B). Smad4 TFBS matches to the reference sequence (human) are in blue, aligned pairs - in red and aligned-and-conserved - in green. *NKX2.5* cardiac enhancer harbors 4 conserved Smad4 sites, one site corresponds with a previously functionally characterized Smad4 site (C).

REFERENCES

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.

Kel, A.E., E. Gossling, I. Reuter, E. Cheremushkin, O.V. Kel-Margoulis, and E. Wingender. 2003. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* **31:** 3576-3579.

Lien, C.L., J. McAnally, J.A. Richardson, and E.N. Olson. 2002. Cardiac-specific activity of an Nkx2-5 enhancer requires an evolutionarily conserved Smad binding site. *Dev Biol* **244:** 257-266.

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.

Matys, V., E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A.E. Kel, O.V. Kel-Margoulis, D.U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31:** 374-378.

Ovcharenko, I., Loots, G.G., Hardison, R.C., Miller W., Stubbs L. 2004. zPicture: Dynamic Alignment and Visualization Tool for Analyzing Conservation Profiles. *Genome Res* **14:** 100.

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.

Venter, J.C. M.D. Adams E.W. Myers P.W. Li R.J. Mural G.G. Sutton H.O. Smith M. Yandell C.A. Evans R.A. Holt J.D. Gocayne P. Amanatides R.M. Ballew D.H. Huson J.R. Wortman Q. Zhang C.D. Kodira X.H. Zheng L. Chen M. Skupski G. Subramanian P.D. Thomas J. Zhang G.L. Gabor Miklos C. Nelson S. Broder A.G. Clark J. Nadeau V.A. McKusick N. Zinder A.J. Levine R.J. Roberts M. Simon C. Slayman M. Hunkapiller R. Bolanos A. Delcher I. Dew D. Fasulo M. Flanigan L. Florea A. Halpern S. Hannenhalli S. Kravitz S. Levy C. Mobarry K. Reinert K. Remington J. Abu-Threideh E. Beasley K. Biddick V. Bonazzi R. Brandon M. Cargill I. Chandramouliswaran R. Charlab K. Chaturvedi Z. Deng V. Di Francesco P. Dunn K. Eilbeck C. Evangelista A.E. Gabrielian W. Gan W. Ge F. Gong Z. Gu P. Guan T.J. Heiman M.E. Higgins R.R. Ji Z. Ke K.A. Ketchum Z. Lai Y. Lei Z. Li J. Li Y. Liang X. Lin F. Lu G.V. Merkulov N. Milshina H.M. Moore A.K. Naik V.A. Narayan B. Neelam D. Nusskern D.B. Rusch S. Salzberg W. Shao B. Shue J. Sun Z. Wang A. Wang X. Wang J. Wang M. Wei R. Wides C. Xiao C. Yan A. Yao J. Ye M. Zhan W. Zhang H. Zhang Q. Zhao L. Zheng F. Zhong W. Zhong S. Zhu S. Zhao D. Gilbert S. Baumhueter G. Spier C. Carter A. Cravchik T. Woodage F. Ali H. An A. Awe D. Baldwin H. Baden M. Barnstead I. Barrow K. Beeson D. Busam A. Carver A. Center M.L. Cheng L. Curry S. Danaher L. Davenport R. Desilets S. Dietz K. Dodson L. Doup S. Ferriera N. Garg A. Gluecksmann B. Hart J. Haynes C. Haynes C. Heiner S. Hladun D. Hostin J. Houck T. Howland C. Ibegwam J. Johnson F. Kalush L. Kline S. Koduru A. Love F. Mann D. May S. McCawley T. McIntosh I. McMullen M. Moy L. Moy B. Murphy K. Nelson C. Pfannkoch E. Pratts V. Puri H. Qureshi M.

Reardon R. Rodriguez Y.H. Rogers D. Romblad B. Ruhfel R. Scott C. Sitter M. Smallwood E. Stewart R. Strong E. Suh R. Thomas N.N. Tint S. Tse C. Vech G. Wang J. Wetter S. Williams M. Williams S. Windsor E. Winn-Deen K. Wolfe J. Zaveri K. Zaveri J.F. Abril R. Guigo M.J. Campbell K.V. Sjolander B. Karlak A. Kejariwal H. Mi B. Lazareva T. Hatton A. Narechania K. Diemer A. Muruganujan N. Guo S. Sato V. Bafna S. Istrail R. Lippert R. Schwartz B. Walenz S. Yooseph D. Allen A. Basu J. Baxendale L. Blick M. Caminha J. Carnes-Stine P. Caulk Y.H. Chiang M. Coyne C. Dahlke A. Mays M. Dombroski M. Donnelly D. Ely S. Esparham C. Fosler H. Gire S. Glanowski K. Glasser A. Glodek M. Gorokhov K. Graham B. Gropman M. Harris J. Heil S. Henderson J. Hoover D. Jennings C. Jordan J. Jordan J. Kasha L. Kagan C. Kraft A. Levitsky M. Lewis X. Liu J. Lopez D. Ma W. Majoros J. McDaniel S. Murphy M. Newman T. Nguyen N. Nguyen M. Nodell S. Pan J. Peck M. Peterson W. Rowe R. Sanders J. Scott M. Simpson T. Smith A. Sprague T. Stockwell R. Turner E. Venter M. Wang M. Wen D. Wu M. Wu A. Xia A. Zandieh and X. Zhu. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

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.

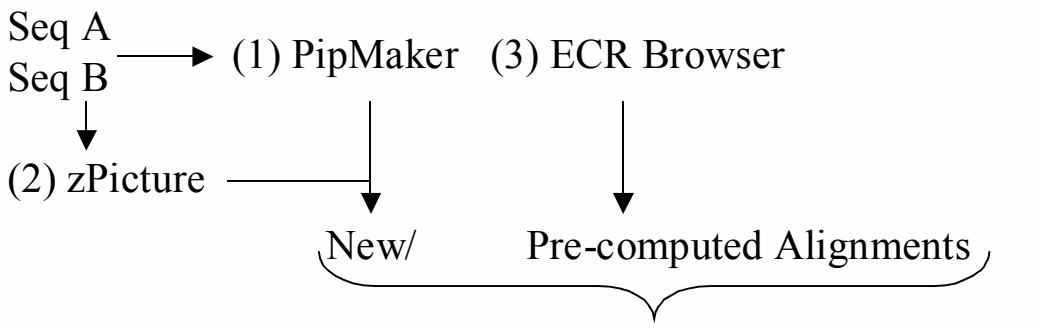
Wingender, E., P. Dietze, H. Karas, and R. Knuppel. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24:** 238-241.

Table 1. Comparative detection of PWMs in long genomic intervals performed by MATCH (Kel et al. 2003) and tfSearch programs. Two different PWM matrix cut-offs (with equivalent core cut-offs in case of MATCH tool), 0.75 and 0.85 were analyzed. Analysis for all the 491 available TRANSFAC (Matys et al. 2003) PWMs is compared to the analysis performed with a single GATA3 PWM. Test was performed on a 2.2GHz Dell PC running RedHat Linux 7.3. Two loci, 1Mb at chr20:10,000,000-11,000,000 (human genome, NCBI Build 34) containing ANKDR5, SNAP25, MKKS, and JAG1 genes and 100kb at chr20:10,000,000-10,100,000 (human genome, NCBI Build 34) containing ANKDR5 gene were analyzed.

Region/PWMs	MATCH 0.75	tfSearch 0.75	Speed increase	MATCH 0.85	tfSearch 0.85	Speed increase
1Mb / 491 PWMs	12243.0s	708.4s	17x	4029.5s	54.6s	74x
100kb / 491 PWMs	1235.5s	15.3s	81x	405.1s	3.9s	105x
1Mb / GATA3	40.1s	4.4s	9x	39.9s	3.2s	13x
100kb / GATA3	4.0s	0.2s	20x	4.0s	0.2s	20x

Figure 1

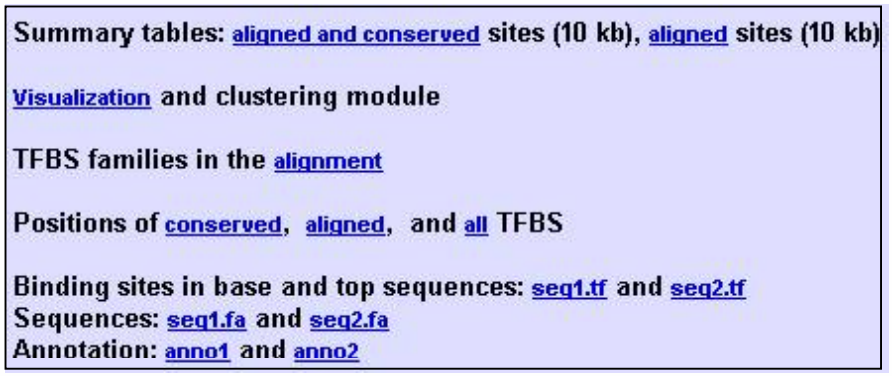
A



Select Transcription Factors/ Matrix Similarity

•Biobase matrices •User defined consensus sequences

B



C

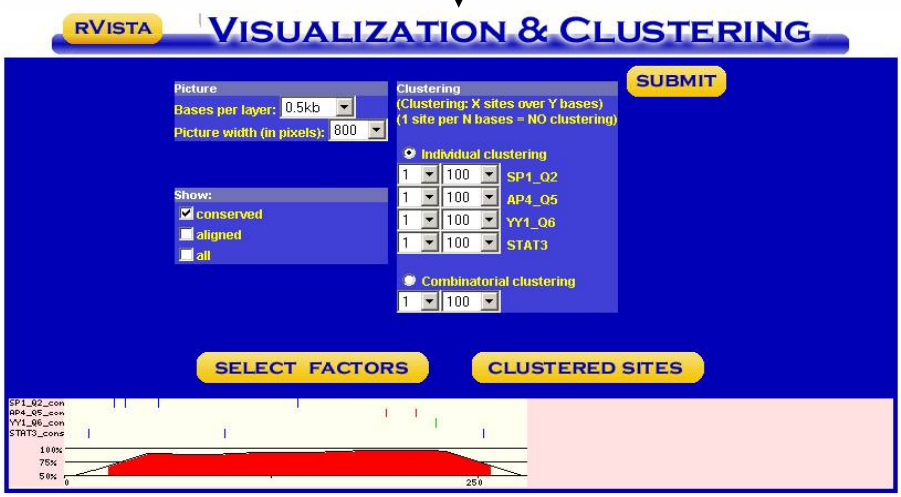
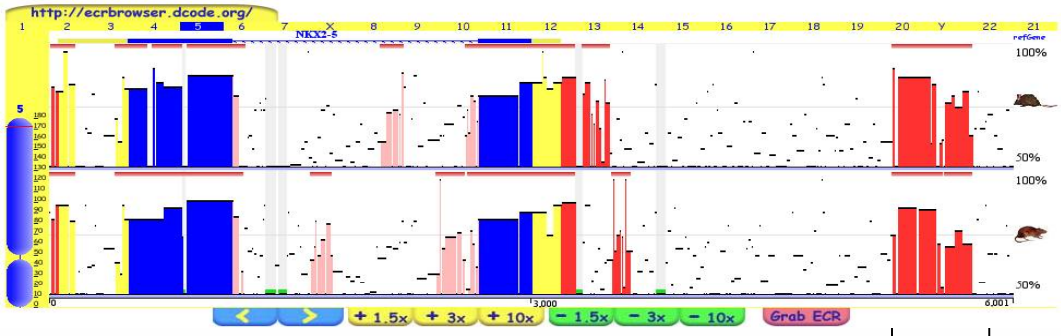
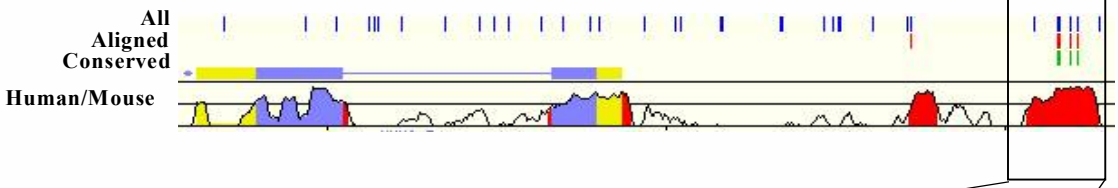


Figure 2

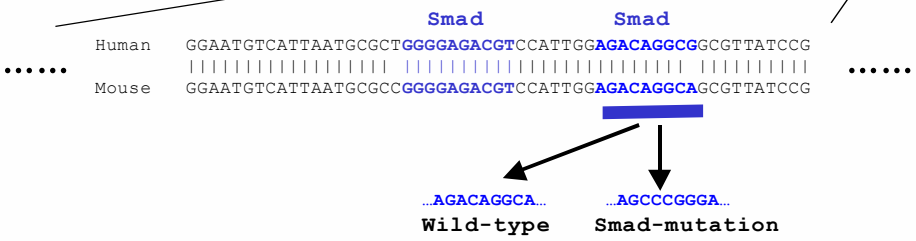
A


http://ecrbrowser.dcode.org/
NKX2-5
100%
50%
100%
50%
+ 1.5x
+ 3x
+ 10x
- 1.5x
- 3x
- 10x
Grab ECR

B


All
Aligned
Conserved
Human/Mouse

C


Smad
Smad
Human GGAATGTCATTAATGCGCTGGGGAGACGTCCATTGGAGACAGGCGGCGTTATCCG
Mouse GGAATGTCATTAATGCGCCGGGGAGACGTCCATTGGAGACAGGCAGCGTTATCCG
....AGACAGGCA...
....AGCCCGGGA...
Wild-type
Smad-mutation