7

Computational Prediction of *cis*-Regulatory Modules From Multispecies Alignments Using Galaxy, Table Browser, and GALA

Laura Elnitski, David King, and Ross C. Hardison

Summary

One major goal of genomics is to identify all the functional sequences in genomes, including sequences that regulate the expression of genes. Sequence conservation is a good, albeit imperfect, guide to these functional elements. We describe how to use publicly available servers (Galaxy, the UCSC Table Browser, and GALA) to find genomic sequences whose alignments (from blastZ and multiZ) show properties associated with *cis*-regulatory modules, such as high conservation score, high regulatory potential score, and conserved transcription factor binding sites. Links to these servers can be accessed at http://www.bx.psu.edu/ and http://genome.ucsc.edu/.

Key Words: Enhancers; promoters; gene regulation; multispecies sequence alignments; blastZ; multiZ; UCSC Genome Browser; GALA; Galaxy; human genome.

1. Introduction

With complete, or almost complete, genome sequences from a large number of species becoming available, the issue of assigning a function, if any, to each string of nucleotides has now moved to the forefront of activity in the human genome project ***(1)***. A string of nucleotides involved in a physiological process, such as encoding part of a protein (an exon) or specifying the spatiotemporal pattern of gene expression (e.g., a binding site for a transcription factor), is referred to here as a functional element in the genome. Much progress has been made in identifying genes using either *ab initio* predictions or evidence-based predictions, but a complete set of genes for most organisms cannot be unambiguously assigned ***(2)***. Computational detection of noncoding functional elements is even less well developed, mainly because of the limited understanding of the

From: *Methods in Molecular Biology, vol. 338: Gene Mapping, Discovery, and Expression: Methods and Protocols*
Edited by: M. Bina © Humana Press Inc., Totowa, NJ

role of DNA sequences in the molecular mechanisms of gene regulation or other noncoding functions ***(3–5)***. However, methods of comparative genomics succeed at a sufficiently high rate that they are commonly used to predict candidate *cis*-regulatory elements for experimental validation (e.g., ***6,7–10***).

cis-regulatory modules (CRMs) are sets of functional elements that are clustered to form a regulatory unit (such as a promoter or enhancer) that acts in *cis* to a gene to control its expression level, timing, or tissue specificity. A large number of bioinformatic approaches have been developed to help investigators predict CRMs. This chapter describes how to use publicly available, web-based bioinformatic servers developed in our research group and those of our collaborators to predict CRMs based on properties of vertebrate genomic sequence alignments. Additional excellent servers are described in other chapters in this book; some are listed in **Table 1**.

The Methods section (**Subheading 3.**) refers to several functions computed from genomic sequence alignments to bring out different features associated with regulatory functions. For instance, a fundamental observation is whether a sequence falls within an alignment. The methods discussed in this chapter utilize precomputed, whole-genome alignments of sequences from several species, generated with the programs blastZ ***(11)*** and/or multiZ ***(12)***. Several other alignment algorithms and servers have been developed, as described in a recent review ***(13)***. More recently servers with improved features have been developed, which provide enhanced abilities to align and analyze sequences provided by the user (**Table 1**).

Purifying (or negative) selection is one of the most general genomic features that indicate function. The precomputed, whole-genome alignments have been analyzed for evidence of purifying selection following their divergence from a common ancestor. This type of selection can be inferred using the phastCons program ***(14)***, which computes the likelihood that a given nucleotide in a sequence (represented as a column in the alignment) is in the 10% most slowly changing sequences in the genome. Scores associated with phastCons analyses are visualized in the "conservation" track on display at the UCSC Genome Browser ***(15)***. Presented as highly resolved scores with wide dynamic range, the scores increase with stronger evolutionary constraint. Higher scores are implicated in function, but they provide no insight into the nature of the function.

The precomputed, whole-genome alignments have also been analyzed for the likelihood of involvement as a CRM, computed as a regulatory potential (RP) score ***(16,17)***. Considered as short runs of columns (containing from two to five aligned positions), regions are analyzed for their frequency of appearance in a training set of known regulatory elements vs a training set of ancestrally derived neutral DNA. This function is influenced by the degree of evolutionary constraint, as is phastCons, but it also incorporates information about patterns in

Table 1
URLs for Servers Used to Predict CRMs

Property	Server	URL
Genome sequences, alignments, and annotations	UCSC Genome Browser and Table Browser	http://genome.ucsc.edu/
	GALA	http://www.bx.psu.edu/
	Galaxy	http://www.bx.psu.edu/
	ECR Browser	http://www.dcode.org/
Aligners	zPicture, Mulan, eShadow	http://www.dcode.org/
	PipMaker, MultiPipMaker	http://www.bx.psu.edu/
	VISTA	http://genome.lbl.gov/vista/index.shtml
	MAVID	http://baboon.math.berkeley.edu/mavid/
	LAGAN	http://lagan.stanford.edu/lagan_web/index.shtml
Phylogenetic footprints	FootPrinter2.0	http://wingless.cs.washington.edu/htbin-post/unrestricted/FootPrinterWeb/FootPrinterInput2.pl
	ConSite	http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite
	rVista 2.0, multiTF	http://www.dcode.org/
Gene expression data	Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/geo/
	ArrayExpress	http://www.ebi.ac.uk/arrayexpress/
Motif discovery	Meme	http://meme.sdsc.edu/meme/website/intro.html
	MotifSampler	http://www.esat.kuleuven.ac.be/~thijs/Work/MotifSampler.html
	Weeder	http://159.149.109.16:8080/weederWeb/
	AlignAce	http://atlas.med.harvard.edu/
	Crème 2.0	http://www.dcode.org/

the alignments ***(16)***. Empirical evaluations of the effectiveness of this approach for finding regulatory regions of proven function show that both RP and PhastCons work well with some highly conserved datasets, such as enhancers of developmental genes ***(18)***. RP performs better than phastCons on a very difficult reference set containing all the CRMs in the human *HBB* gene complex.

None of the alignment-derived scores, including phastCons and RP scores, are sufficiently specific for highly reliable predictions of CRMs ***(19)***. Therefore, it is prudent to combine these with other features commonly found in CRMs, such as binding sites for transcription factors. Many binding site motifs have been discovered and are recorded in resources such as TRANSFAC ***(20)*** and JASPAR ***(21)***. Tools to identify motifs, based on overrepresentation of sequence strings in a given set of sequences, are also widely used ***(5,22)***. In general, any approach to find motifs in one single sequence returns an excess of false positives. Requiring strict conservation in alignments of human, mouse, and rat sequences reduces the number of hits to binding sites for transcription factors by a factor of about 40 ***(23)***. This chapter describes how to access matches to conserved transcription factor binding sites (cTFBS) computed by the program *tffind* ***(24)***.

Precomputed binding sites allow a user to look for sites of interest that fall within a neighborhood of a genomic locus, without setting strict limitations on the amount of sequence being submitted in the search. In contrast, someone using a server to find matches to TFBS in a sequence will typically extract a few kilobases upstream and downstream of a gene to submit. The limitation of the analysis to a certain distance around a gene may inadvertently exclude important regions. The use of precomputed binding sites allows a user to select a larger region and subsequently reduce it through queries of a more refined region.

The data discussed in this chapter are stored in databases at the University of California at Santa Cruz (UCSC) Genome Browser ***(15)*** and GALA, a database of genome sequence alignments and annotations ***(25,26)*** (**Table 1**). A recently released metaserver, Galaxy, provides a platform for integrative analyses of genomic sequences and annotations ***(27)***. The metaserver uses the query engines from remote databases such as the UCSC Table Browser ***(28)*** and other resources to retrieve primary data, and it provides operations and tools to filter, combine, and analyze the data. The Galaxy metaserver project is new and should grow to connect to many data repositories and provide a large suite of operations and tools. GALA is a more mature database project that also provides access to alignment and annotation results. GALA follows the traditional approach of recording all the data in a database on one large machine, whereas Galaxy accesses data from remote sites. Instructions for acquiring and analyzing data to predict CRMs using both Galaxy (in conjunction with the UCSC Table Browser) and GALA are presented in the Methods section (**Subheading 3.**).

The basic method described in this chapter is to retrieve candidate CRMs in erythroid cells as noncoding DNA segments with a high phastCons score or high RP score and a conserved match to a GATA-1 binding site. GATA-1 is a transcription factor that is essential for proper gene expression during late erythroid maturation ***(29)***. A description is given of how to obtain noncoding genomic

DNA segments with the desired phastCons or RP scores, how to obtain conserved GATA-1 binding sites, and how to identify all the conserved or high-RP intervals with a conserved GATA-1 binding site in close proximity. A similar approach could be followed for any binding site of interest, when some information is known regarding preferential tissue specificity of the factor.

Although the approach described using premapped matches to binding sites for the entire genome is useful, other computational tools are being developed to discover motifs (short nucleotide strings). These extensions of basic pattern matching require a given motif to be enriched in, for example, sequences immediately upstream from a set of coexpressed genes. Thus, they are frequently used to find candidates for common regulatory elements controlling similarly expressed genes. Clusters of coexpressed genes are commonly deduced from transcriptional profiles based on microarray or other experiments measuring expression. Two large public databases of gene expression data are located at the Gene Expression Omnibus and ArrayExpress (**Table 1**). A sample of motif-finding servers is listed in **Table 1**. These simply require that users submit a list of sequences, such as the promoters (known or predicted) for a set of coexpressed genes. The servers use different methods ***(5)*** for motif discovery. An evaluation of the performance of these methods was recently published and provides further information on the subject ***(22)***.

2. Materials

The only material required is a computer connected to an Internet service provider and running an Internet browser (such as Internet Explorer, Safari, Mozzilla, or Netscape).

3. Methods

3.1. Retrieving Strongly Conserved, Noncoding Genomic Intervals With Galaxy/UCSC Table Browser

1. Enter the Galaxy portal by pointing your Internet browser to the URL http://www.bx.psu.edu/ (**Table 1**) and clicking on "Galaxy."
2. At the Galaxy portal, you are presented with a few options. The first is to go to the UCSC Table Browser to retrieve any of the rich variety of data recorded there and automatically upload it to Galaxy. However, for phastCons and RP scores, it is more efficient to choose "Galaxy featured datasets" (*see* **Note 1**). On the new page, select the genome of the species of interest (e.g., Human) and the desired sequence assembly (e.g., hg17: May 2004) (*see* **Note 2**). The available options are specific to the genome assembly; for example, hg17 currently offers:
   a. Known regulatory regions [93 regions].
   b. phastCons (stringent, top approx 5%) [1,313,584 regions].

c. phastCons (sensitive, ≥0.2) [26,277,600 regions].
d. Regulatory potential (3way, human-mouse-dog, >0) [5,800,931 regions].

3. Regarding phastCons scores, select option b for regions under intense constraint or option c for increased sensitivity (*see* **Note 3**). Then click on the button labeled "Go." The results are added to your history page, which is displayed on your computer.
4. The next step is to retrieve the locations of all exons so that they can be removed from the high phastCons intervals. Users should return to the Galaxy Portal by clicking on "Portal" on the top row of the window in your Internet browser. At the Portal, click on the link to the UCSC Table Browser.
5. To retrieve exons, use the Table Browser pull-down menus to select "Genes and Gene Prediction Tracks" under the category of "group" and "Known Genes," found under "track" (*see* **Note 4**). If desired, the query can be limited to a particular genomic interval using the window labeled "position" (*see* **Note 5**). Because you entered the Table Browser via Galaxy, the default for "output format" is "send data to Galaxy." Now click on "get output."
6. A window appears that gives you the option to select whole genes, exons, coding exons, and so on. Select "Exons," and click on "Send query to Galaxy" (*see* **Note 6**).
7. This returns the user automatically to the Galaxy History Page (*see* **Note 7**), where each query appears as a short description (*see* **Note 8**) followed by the number of results retrieved.
8. In preparation for performing an operation, you need to select the desired datasets. Select the boxes for the queries of high phastCons scores and exons. Now select "Perform operations like intersection, etc." and click on "Go."
9. On the Query Operations page, the two queries now appear, and you should click on the box next to the operation "Subtraction." The screen automatically refreshes. Use the pull-down menus to determine the order and type of subtraction. In this case, it should be the query for phastCons intervals minus the query for the KnownGenes exons, removing "only overlapping segments." Click on "Go" (*see* **Note 9**).
10. The user is returned automatically to the History Page, which will show the number of results when the operation has completed. If the operation is listed as "running," the user should click "Refresh" periodically until the operation is finished. The resulting genomic intervals are noncoding, highly conserved DNA segments, which is one class of candidates for CRMs.
11. Galaxy provides several forms of output, which are accessed by clicking "Get output" followed by "Go." At the Display Options page, select "Genome Browser" to view each of the returned intervals in the UCSC Genome Browser (*see* **Note 10**), or "Raw result file" to obtain a file with the desired genomic intervals. Other options include viewing the results in the Ensembl browser (*see* **Note 11**).

3.2. Retrieving High-RP, Noncoding Genomic Intervals With Galaxy/UCSC Table Browser

The procedure for finding high-RP intervals via Galaxy is the same as outlined in **Subheading 3.1.**, except that when using the "Galaxy featured data-

sets" (accessed through the Galaxy portal), the user should choose option D "Regulatory potential (3way, human-mouse-dog, >0) [5,800,931 regions]" (*see* **Note 12**).

3.3. Retrieving Conserved Matches to Transcription Factor Binding Sites (cTFBS) Using Galaxy/UCSC Table Browser

Conserved matches to binding sites for transcription factors with weight matrices in TRANSFAC ***(20)*** can be obtained via the UCSC Table Browser and retrieved into Galaxy. The software used is an update of the *tffind* program ***(24)***.

1. From the Galaxy Portal page, take the link to the UCSC Table Browser, where the user should select the group "Expression and Regulation" and the track "TFBS Conserved." Under region, select position and enter the chromosome name and coordinates of interest. Otherwise, select "genome."
2. To restrict the search to a single binding site, you should filter by name. Next to "filter," click on "create." This brings up a new page, on which you should select "does" match, next to "name." The name of the binding site matrix should be entered after "match." For instance, using the term "V$GATA*" will return conserved matches to a set of weight matrices for GATA-1 and GATA-3 binding sites (*see* **Note 13**).
3. Press "submit" to upload the filter to the Table Browser query page, and then click on "get output." Results will appear on the Galaxy history page.

3.4. Integrating the Conservation or RP Data With cTFBS at Galaxy

1. At the Galaxy history page, the user now has the noncoding intervals with high phastCons scores, the noncoding intervals with high RP scores, and intervals with conserved GATA-1 binding sites. Select two of the results to combine, e.g., noncoding high-RP intervals and conserved GATA-1 binding sites, by clicking on the buttons next to each query.
2. Under "Action to Perform," click on the button for "Perform operations like intersection, etc." and click "Go." This takes the user to the Query Operations page. Only the queries selected from the history page are transferred to the operations page. For a given number of queries, only a certain set of operations is allowed. Those that are not allowed are dimmed.
3. To find all the noncoding, high-RP intervals that have a conserved GATA-1 binding site in proximity to them, under "Operation," click on the button next to "Proximity" (*see* **Note 14**). After the screen refreshes, use the pull-down options to return regions from the noncoding, high-RP query results that lie less than 50 bp in either direction from a region in the query for conserved GATA-1 binding sites. Click on "Go," which returns you to the history page. The page initially returned frequently shows the new query as "running." Again, periodically click "Refresh" to obtain the results.
4. The results are the predicted CRMs, based on three criteria—they have a high RP score, they are not exons, and they are close to or encompass a conserved match to

a GATA-1 binding site. To retrieve the results of a selected query, select "Get output" from the list of "Actions to Perform" and hit "Go." For viewing the results, select "UCSC Browser custom track" or "Ensembl Genome Browser custom track." For a plain text file, select "Raw result file (bed)." The desired action is taken when you click "Go." Other features can be combined, such as high phastCons scores, and other operations can be performed on the data, using the utilities at Galaxy.

3.5. Retrieving High-RP or High-phastCons Intervals in Noncoding Sequences Using GALA

1. The GALA database is accessed at http://www.bx.psu.edu/ (**Table 1**) by selecting the link for "GALA." On the home page, the user finds links to GALA databases built for genomes of five different species (human, mouse, rat, chimp, and chicken), with up to three assemblies for each (*see* **Note 15**). Click on "Query page" under the appropriate species and assembly (e.g., Human July 2003 data release).
2. The query page is presented as an expandable selection of choices for categories, i.e., genes and gene predictions, expressed sequence tags and mRNA, comparative genomics, variation and repeats, expression and regulation, and mapping and sequencing, which are compatible with groups on the UCSC Genome Browser (*see* **Note 16**). Halfway down the page, you will find the query boxes for "Regulatory potential scores based on multiple alignments," with options for filtering the results by a minimum and maximum score. A good score for the minimal threshold is 0.001; leave the "less than or equal to" box blank. Alternatively, you may wish to query on the next item, PhyloHMM Cons (an earlier name for phastCons). A good score for the minimal threshold is 0.4 ***(18)*** (*see* **Note 17**).
3. Users wishing to investigate only a small genomic locus can choose the button to "Restrict search to interval" (near the bottom of the form). Otherwise, proceed to select the choice of output. "Text list" is the preferred choice when preparing datasets for use with subsequent operations.
4. Click "run query in background," so the server will save the results for 48 h. The results are returned on the GALA history page, where they can be combined with other queries (*see* **step 6**).
5. To collect exons, return to the GALA query page, and for the category "Genes and gene models," click on "Show the fields for this category" and then "Refresh" (toward the bottom of the page). The new page has many options for obtaining genes or parts of genes. Under "Protein Coding Genes, GALA's default set of genes," go to "Other gene fields" and click the box for "exons." Scroll to the bottom of the page, restrict the query to a chromosomal interval if desired, choose "text file" under "Output," and click on "run query in background."
6. Use the GALA history page to remove the exons from the high-RP intervals. Click the box next to each query on which you want to perform an operation (such as subtraction). Under "Compound queries," choose "SUBTRACTION." If you follow the steps in the order covered in here, choose the option to subtract "earlier minus later query" to subtract exonic intervals from the high-RP intervals. Using the pull-down menu, specify that "only overlapping segments" should be removed.

Click on "Run compound query in the background," located almost at the bottom of the page. The results are noncoding, high-RP intervals.

3.6. Retrieving Conserved Matches to TFBS Using the GALA Server

1. On the GALA query page, under the category of "Expression and Regulation," go to "Transcription factor binding sites" and choose, e.g., "only binding sites conserved in hg16Mm3Rn3, cutoff used was 0.85" (*see* **Note 18**).
2. Click on the button after "To select/add factor names," which opens a new page with all the choices. Select those of interest, and press the button "add selections to main form," which is at the bottom of the selection page (*see* **Note 19**).
3. The user is returned to the GALA query page. As before, users can limit the query by entering a restricted genomic interval, or they can query the entire genome. After selecting the desired output (e.g., "text list"), the user should click on "Run query in the background." A results page appears, after which the user can go to the history page.

3.7. Integrating the Conservation or RP Data With cTFBS Data at GALA

1. The GALA history page lists the queries that have been run, such as noncoding high-RP intervals and conserved GATA-1 binding sites, along with the number of results obtained for each. To find features that are in proximity to others, scroll down the page under "Compound queries" to "Proximity."
2. Enter the appropriate query numbers in the boxes under "Proximity," specifying that the noncoding, high-RP intervals "lie within 50bp" of regions in the conserved GATA-1 binding site query (*see* **Note 20**).
3. Select the type of output (such as "text list"), and then click "Run compound query in the background."
4. The results returned are the CRMs predicted by having a high-RP score, not being exons, and being close to a GATA-1 binding site that is conserved among human, mouse, and rat. Other criteria can be applied, and other operations (such as intersections or clustering) can be used for alternative predictions.

4. Notes

1. Instead of using the "Galaxy featured datasets," the user can follow the link to the Table Browser and retrieve genomic intervals whose phastCons scores exceed a desired threshold. However, this step takes a rather long time for the entire genome (searching through about 800 million records), and it is likely to time-out. Thus a user should limit this search to a specific interval (megabases should be no problem), or one can use the preselected intervals deposited in the "featured datasets." A similar logic holds for the RP scores.
2. It is often the case that the most recent assembly is the more complete and better annotated. However, it takes some time for annotations to be "lifted" onto new

assemblies, and thus for some time after a new assembly is released, more information will be available on the previous assembly. As of this writing, the very extensive data on the ENCODE regions are available only for hg16, the July 2003 assembly of human.

3. Selecting the more sensitive threshold for phastCons score (≥0.2) returns a large set of intervals that does the best job of finding known CRMs in the *HBB* gene complex ***(18)***. However, it almost certainly returns many false positives, and for some purposes, the more stringent threshold may be more appropriate.
4. The choice of the collection of genes used is, of course, up to the user. The Known Genes track is very extensive and quite reliable, but it misses some genes. Users may prefer RefSeq, Ensembl, or other sets. Users should be aware that despite the considerable overlap in these gene sets, there are many differences, and these will affect the results of subtracting them from a set of intervals to find noncoding conserved sequences.
5. In this step, and in all steps in which the user has an option to limit a query to a particular interval, it is important to realize that the larger the interval examined, the more time it takes for the database to complete the query. Thus, searching the entire genome (approx 3000 Mb) takes considerably longer than searching the ENCODE regions (approx 30 Mb), which will take longer than a given locus (perhaps 0.3 Mb). Likewise, the number of features in the intervals searched is a major determinant of time to complete the query. phastCons and RP scores are given for every aligning nucleotide, and thus there are almost 800 million of these records to search. In contrast, the number of exons in the KnownGenes set is about 400,000, and thus a query to retrieve them takes less time. For full data on dense features like phastCons or RP, downloading files is much more efficient.
6. Users may instead wish to choose exons with an additional short interval, e.g., 10 bases, at each end. By doing so, the user will include regions that may be indirectly under selection because of their proximity to exons.
7. The Galaxy history page will load immediately, even if the query has not finished running at the Table Browser. In this case, at the end of the query, the notation "running" appears. The user should periodically click on "refresh" to see when the query has been completed and the results sent to Galaxy.
8. In this step, or any time the user is on the history page, one of the options is to edit the descriptions. Select a query, and click on "More." The screen refreshes, and now the option to "Edit query descriptions" is displayed. This editing is particularly helpful for the results of operations, for which Galaxy simply refers to the queries by number, not by content. A similar feature is implemented in GALA.
9. The time it takes for an operation to complete is determined primarily by the number of intervals that are in each query.
10. All the returned intervals can easily be viewed in the UCSC Genome Browser. On the left of the Genome Browser display is a list of all the returned intervals, which are hyperlinks to new views that show each region. The text file that can be returned is in BED format, in which the first three columns are chromosome, start position, and stop position for each interval.

11. After seeing the results, if the user decides that the genomic regions selected for the queries requires optimization, e.g., it was too small or too large, return to **step 5** and enlarge or reduce the coordinate distance.
12. Selecting "Regulatory potential (3way, human-mouse-dog, >0)" returns a set of 5.8 million intervals that does the best job of finding known CRMs in the *HBB* gene complex ***(18)***. It probably also returns some false positives. To increase the stringency of the search, users can go the UCSC Table Browser, and select "Expression and Regulation" as the group and "3x Reg Potential" as the track (currently only available on the human July 2003 assembly). By clicking on the "create" button for "filter," the user gets to a page at which the threshold can be set higher, e.g., dataValue is ≥0.001. By clicking on "submit," this filter will be applied to the query when it is run.
13. The first set of filters is for the table of conserved binding sites, and the "name" refers to the name (or ID) of the weight matrix for a binding site. Thus one could enter a TRANSFAC ID for a particular weight matrix, such as "V$GATA1_02." Of course, this requires that the user know these IDs, which can be obtained from TRANSFAC. In the example given here, a wild card character ("*") was used to filter on "V$GATA*," which will include multiple binding sites for GATA-1 and GATA-3 (which have very similar binding sites). In order to filter based on the name of the transcription factor (not the binding site), users can take advantage of the ability of the Table Browser to filter on fields in related tables. On the filter page, choose the option to allow filtering on hg17.tfbsConsFactors, and choose "factor does match GATA-1" (or the name of the desired factor).
14. Users can find features in proximity to other features, such as described here, and the distance between them is set by the user. Alternatively, users may elect to perform a simple intersection. Note that the screen refreshes for each newly selected operation, because the parameters and choices relevant to each operation differ. In our research, we have found that using proximity has predicted some active CRMs that were missed by the intersection operation, but this is not frequent.
15. On the GALA home page users may want to access "Annotation statistics" to see the all the different types of data recorded, the number of records in each, and a partial list of fields in each table. Users can also go directly to their history page.
16. The default GALA query page lists only minimal or no choices for categories such as genes and gene models. Users who want to query on information within these should click on "Show the fields for this category" and then "Refresh" (toward the bottom of the page).
17. Users may wish to choose alignments computed between different species or filtered in various ways. These options are all under the comparative genomics section of the query page.
18. The options available for binding sites in GALA differ by the species and genome assembly. Here we selected binding sites conserved in human-mouse-rat alignments (hg16Mm3Rn3), but users can select other alignments, such as a pairwise human-chicken (hg16Gg2) or five-way human-chimp-mouse-rat-chicken (hg16Pt1Mm3

Rn3Gg2). The threshold scores ("cutoff") for the matches to the weight matrices are adjusted in each case.

19. Users can select by ID for weight matrices for factors instead of by name of the factor. Queries of all binding sites (not just the conserved ones) must be a limited to a chromosomal interval because of the very large number of sites in the entire genome (about 212 million for the human genome).
20. Users can elect to do intersections or other operations. Clustering is also supported, e.g., requiring that each high-RP interval have at least two conserved factor-binding sites within it. This set of operations is supported in both Galaxy and GALA.

Acknowledgments

This work was supported by NIH grants DK65806 (to R.H.) and HG02325 (to L.E.).

References

1. Collins, F. S., Green, E. D., Guttmacher, A. E., and Guyer, M. S. (2003) A vision for the future of genomics research. *Nature* **422,** 835–847.
2. Parra, G., Agarwal, P., Abril, J. F., Wiehe, T., Fickett, J. W., and Guigo, R. (2003) Comparative gene prediction in human and mouse. *Genome Res.* **13,** 108–117.
3. Pennacchio, L. A. and Rubin, E. M. (2001) Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* **2,** 100–109.
4. Hardison, R. C. (2003) Primer on comparative genomics. *Public Library of Science, Biology* **1,** 156–160.
5. Wasserman, W. W. and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5,** 276–287.
6. Gumucio, D., Shelton, D., Zhu, W., et al. (1996) Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the beta-like globin genes. *Mol. Phylog. Evol.* **5,** 18–32.
7. Loots, G. G., Locksley, R. M., Blankespoor, C. M., et al. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288,** 136–140.
8. Hardison, R. C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16,** 369–372.
9. Nobrega, M. A., Ovcharenko, I., Afzal, V., and Rubin, E. M. (2003) Scanning human gene deserts for long-range enhancers. *Science* **302,** 413.
10. Miller, W., Makova, K. D., Nekrutenko, A., and Hardison, R. C. (2004) Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5,** 15–56.
11. Schwartz, S., Kent, W. J., Smit, A., et al. (2003) Human-mouse alignments with *Blastz*. *Genome Res.* **13,** 103–105.
12. Blanchette, M., Kent, W. J., Riemer, C., et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–715.

13. Frazer, K. A., Elnitski, L., Church, D., Dubchak, I., and Hardison, R. C. (2003) Cross-species sequence comparisons: a review of methods and available resources. *Genome Res.* **13,** 1–12.
14. Siepel, A., Bejerano, G., Pedersen, J. S., et al. (2005) Evolutionarily conserved elements in vertebrate, fly, worm and yeast genomes. *Genome Res.* **15,** 1034–1050.
15. Kent, W. J., Sugnet, C. W., Furey, T. S., et al. (2002) The human genome browser at UCSC. *Genome Res.* **12,** 996–1006.
16. Elnitski, L., Hardison, R. C., Li, J., et al. (2003) Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13,** 64–72.
17. Kolbe, D., Taylor, J., Elnitski, L., et al. (2004) Regulatory potential scores from genome-wide three-way alignments of human, mouse and rat. *Genome Res.* **14,** 700–707.
18. King, D. C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W., and Hardison, R. C. (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res.* **15,** 1051–1060.
19. Berman, B. P., Pfeiffer, B. D., Laverty, T. R., et al. (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol.* **5,** R61.
20. Wingender, E., Chen, X., Fricke, E., et al. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* **29,** 281–283.
21. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32,** (Database issue) D91–D94.
22. Tompa, M., Li, N., Bailey, T. L., et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23,** 137–144.
23. Gibbs, R. A., Weinstock, G. M., Metzker, M. L., et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521.
24. Schwartz, S., Elnitski, L., Li, M., et al. (2003) MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.* **31,** 3518–3524.
25. Giardine, B. M., Elnitski, L., Riemer, C., et al. (2003) GALA, a database for genomic sequence alignments and annotations. *Genome Res.* **13,** 732–741.
26. Elnitski, L., Giardine, B., Shah, P., et al. (2005) Improvements to GALA and dbERGEII: Databases featuring genomic sequence alignment, annotation and experimental results. *Nucleic Acids Res.* **32,** (Database issue) D466–D447.
27. Giardine, B., Riemer, C., Hardison, R. C., et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15,** 1451–1455.
28. Karolchik, D., Hinrichs, A. S., Furey, T. S., et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32,** D493–D496.
29. Weiss, M. J. and Orkin, S. H. (1995) GATA transcription factors: key regulators of hematopoiesis. *Exp. Hematol.* **23,** 99–107.