PUBLISHED IN ASSOCIATION WITH
COLD SPRING HARBOR LABORATORY PRESS

Computational screening of conserved genomic DNA in search of functional noncoding elements

Gill Bejerano[1], Adam C Siepel[1], W James Kent[1] & David Haussler[1,2]

[1]Center for Biomolecular Science and Engineering, [2]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, California 95064, USA. Correspondence should be addressed to G.B. (jill@soe.ucsc.edu).

The sequencing of the mouse genome allowed, for the first time, the large-scale estimation of the extent of sequence conservation within our own genome. In particular, it suggested that in mammals there is at least twice as much conserved genomic DNA as there is protein coding DNA[1]. The abundance of conserved noncoding regions holds even for so-called ultraconserved elements at the very tip of the mammalian conservation scale[2], as well as for sequences conserved between long-diverged vertebrates such as human and fish[3,4]. Much of the observed conservation appears to be the result of purifying selection, suggesting a wealth of uncharacterized functional elements and families[5], including transcriptional and post-transcriptional regulatory elements, chromatin structure–associated regions, noncoding RNAs, and perhaps altogether novel classes of functional elements[6]. Moreover, recent comparative sequencing efforts have revealed similarly rich sets of uncharacterized conserved noncoding sequences in other metazoans[7]. We outline here how to obtain sets of conserved regions from a wide range of model organisms. We then describe how to analyze the properties of these regions, filter out undesired ones, such as known and predicted coding regions, and rank the remainder for further computational and functional analysis. The method makes heavy use of the UCSC Genome Browser Database[8] and a suite of related web-accessible tools. The protocol is divided into four major steps: defining the genomic region of interest, based on the user's starting point (gene of interest, a region between two genetic markers, and other regions); selecting a subset of cross-species conserved elements within this region, based on a hidden Markov model that defines and scores genomic intervals for conservation; mapping the different properties of the interval set (such as transcript overlap and species coverage extent); and, finally, ranking the set for further analysis, based on a characteristic profile of the functional class of interest. The same protocol may be used to search for different functional classes of elements in all branches of the tree of life available in the UCSC Genome Browser, including vertebrate, insect, nematode and yeast. It can also easily incorporate custom types of information that the user has access to and allows for easy replacement of parts of the protocol, as our understanding of the relationship between function and sequence conservation, and of the different functional classes, improves. As an example, we present an informatic profile of vertebrate enhancer sequences and discuss a case for which such a method has led to the discovery of several functional enhancers. An accompanying protocol describes a complementary approach to identification of *cis*-regulatory DNA regions in complex genome assemblies by clustering of sequence motifs corresponding to known transcription factor binding sites[9].

MATERIALS

Any internet browser, running on a computer with access to the internet.

The UCSC Genome Browser

PROCEDURE

1| Use any internet browser to access the UCSC Genome Bioinformatics site (http://genome.ucsc.edu/).
➨ **TROUBLESHOOTING**

2| Select the "Genome Browser" option from the panel on the left.
➨ **TROUBLESHOOTING**

3| From the Gateway page, select the "clade", "genome" and "assembly" of interest from the pulldown menus. Entering a search term in the "position" box and selecting "submit" takes you to a results page that lists possible entries that contain your search term(s). If your search has a unique match, the search results page is skipped and you are taken directly to the Genome Browser (Step 5).
The genomic "position" box takes many different inputs. See the "Sample position queries" section of this web page for a comprehensive list, which includes explicit genomic coordinates, a region between two genetic markers, the name of a gene or protein product and additional options.

4| From the search results page, select the most appropriate link, or select your web browser 'back' button to return to Step 3 and attempt a different search.

Picking a genomic region of interest

5| The Genome Browser page displays the selected genomic region. This page contains several components, as described in the legend of **Figure 1**. To select the genomic region of interest, you may pan left or right and zoom in or out using the "move" and "zoom" control buttons at the top of the page. Clicking inside the "base position" track at the top of the main display will zoom in the display and center it on the clicked position. You may also change the content of the "position" box. All inputs described in Step 3 are allowed and will trigger the same search. You may display in different modes, or hide, each data track available for this assembly. Tweak the track controls via the "configure" button at the top of the page or use the rows of pulldown menus at the bottom part of the page (only the first row is shown in **Fig. 1**). Clicking a track title above a pulldown menu, or the button at the far left of each displayed track, opens a details page explaining the track content, often with additional track-specific display controls.

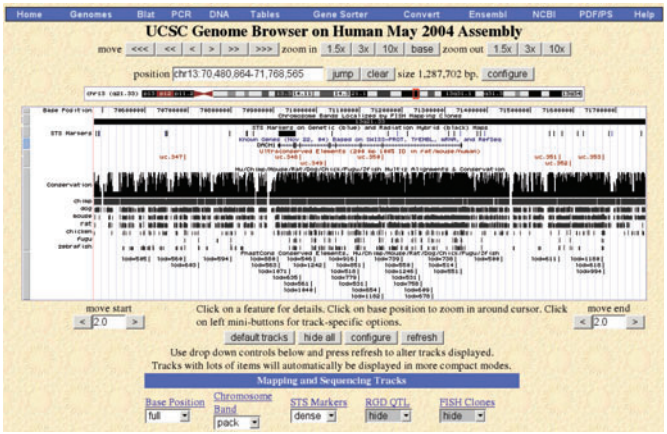

Figure 1 | The main display of the UCSC Genome Browser[8]. The page has several components. They are, from top to bottom: a list of links to complementary tools and resources; a panel for display control; a chromosomal ideogram; the main display, comprising several tracks showing different types of information about the genomic region on display; a second panel of display control; and several rows of pulldown menus (of which only the first is shown), each controlling the display status of an individual track in the main window. Clicking the title of a track pulldown menu, or the button to the left of the track display, leads to a details page explaining the track and often offering additional track-specific display controls.

When you are performing chromosome- or genome-wide surveys, the genomic region of interest is predefined. This is also true when you have focused, via preliminary analysis, on the region spanned between two genetic markers. In other cases, such as screening the genomic vicinity of a gene for cis-regulatory elements, the Genome Browser provides a richly annotated interactive viewer to help determine the region of interest.
▲ **CRITICAL STEP**

6| From a comparative point of view, two tracks are particularly useful. The first is the "Conservation" track (shown in **Fig. 1**). Switch the visibility of this track, found in the "Comparative Genomics" section, to "full" using the pulldown menu, and click one of the "refresh" buttons.
The bottom half of this track displays pairwise alignments between the reference sequence and the best matching genomic region (often the orthologous one) from other available genomes. Values are only plotted when genomic DNA could be aligned with confidence to the reference sequence. They peak only where an aligned base pair is identical to the reference species. When zoomed in sufficiently, the display switches to show the actual aligned bases.
The top half of this track holds a single plot that gives a combined measure of conservation of the reference genomic location. This is determined by examining the observed conservation among all aligned species with respect to the underlying phylogeny and branch lengths.
This track offers many additional display configuration options, accessible via its details page, including "Vertical viewing range", which lets one plot values only for the more conserved regions (for example, by setting the range to 0.99–1).
➨ **TROUBLESHOOTING**

7| The "Most Conserved" track (bottom track in **Fig. 1**) defines discrete conserved elements in the reference genome, based on the same multiple alignment and underlying phylogeny and branch lengths. It does so by utilizing a two-state hidden Markov model to label each base of the reference genome as either conserved or non-conserved. An uninterrupted run of bases labeled conserved is defined as a discrete conserved element and is scored for level of conservation against a null model of neutral evolution. The higher the score, the more conserved the element (see the "Most Conserved" track details page).

(**i**) Turn the track to "pack" or "full" mode to see the different conserved elements within the displayed genomic region. In these modes, a "Most Conserved" label does not appear to the left of the track display. Instead it is recognized by the centered label "PhastCons Conserved Elements". Note, however, that any track in "dense" mode will display the track control label to the left of the track display.

(**ii**) Click on an individual element, or its label, to see an element's details page. The page contains the label of the element, which gives its actual log odds (lod) score. For display purposes these scores are transformed monotonically to a scale of 0–1,000. The "Score" field contains this display score for each conserved interval. The page also offers a "Position" link that leads to a browser display zoomed on the element.

(**iii**) Upon returning to the general details page of the "Most Conserved" track you may select the "Show only items with unnormalized score at or above (range: 0 to 1000)" box to make the track show only elements with a display score at or above the threshold you have set.

Note that the "Conservation" and "Most Conserved" tracks may update frequently, as more species and newer assemblies are added.
To select a handful of the most conserved elements, you may fine-tune the display threshold in Step 7 and manually inspect each region that scores above it. We outline below a structured way to perform the same screen en masse.

Obtaining a set of highly conserved elements

8| Click the "Tables" link in the blue bar at the top of the page. If you are screening a restricted genomic region, you may want to have that exact region in view within the Genome Browser when clicking the link (this sets the 'region' default in the next step). This takes you to the UCSC Table Browser[10] (**Fig. 2**), a powerful tool for the examination and analysis of the data underlying the Genome Browser. The bottom part of its main page has a description of the Table Browser controls.

9| Make sure that the desired values are selected in the "clade", "genome" and "assembly" pulldown menus in the Table Browser main page (**Fig. 2**). To obtain the top-scoring elements in the interval of your choice, select "Comparative Genomics" in the "group" pulldown menu and "Most Conserved" as the track of interest. Only one table, "PhastConsElements", is associated with this track.

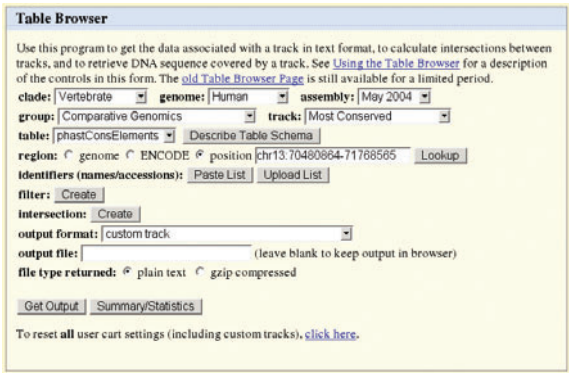

Figure 2 | The UCSC Table Browser[10] main page. The Table Browser is a powerful tool for analyzing the data underlying the Genome Browser. The bottom part of this page (not shown) explains each of the available controls.

(**i**) The "region" radio button menu lets you select the region of choice. If you have just clicked through from the Genome Browser, it defaults to the "position" option, filled with the last-viewed coordinates. You may edit these coordinates or enter any of the search terms described in Step 3 and press "lookup". To perform a chromosome-wide screen, enter the name of the chromosome in the position box, and to perform a genome-wide survey, change the radio button selection to "genome".

(**ii**) Pressing "summary/statistics" at this point will show summary statistics of all "Most Conserved" intervals within the selected region (regardless of the display threshold you may have set in Step 7). Press the 'back' button of your web browser to return to the main Table Browser page.

10| To focus on the top-scoring elements within your region of choice, click the "create" button next to the "filter" label. This takes you to the filter definition page. You may filter by any field or combination of fields from the table that defines this track. To filter by score, change the pulldown menu next to "score" to the greater-or-equal-to symbol and type a threshold in the adjoining input box. You may use the same value you found useful in Step 7. Hit "submit" and you are returned to the main Table Browser page.

▲ **CRITICAL STEP**

11| To see that the filter has indeed been set, note that the options next to the "filter" label have now changed from "create" to "edit" and "cancel".

(**i**) Clicking the "summary/statistics" button will now report statistics corresponding only to the subset of elements that the filter allows.

(**ii**) To change the filter, simply click the "edit" button next to the "filter", label and return to Step 10.

12| The "output format" pulldown menu offers several ways to obtain the selected elements, including download options (see the bottom of that page for details). To continue working with this subset, in the form of a user-defined browser track, choose the "custom track" option and click the "get output" button.

The 'Output as Custom Track' page offers several additional options. A link is available to a help page with extensive details about the many features of custom tracks. For our purposes, the defaults suffice. You may want to change the track name (default is "tb_PhastConsElements") and the description line. Click "get custom track in genome browser" and you are taken back to the Genome Browser. Note that your custom track is near the top of the display and a pulldown menu has been created for it in the controls section. You can now explore the different elements in the Genome Browser, as you would for any other track.

13| If you wish, you may return to Step 8 and define a second custom track or redefine the same custom track using a refined set of criteria. To selectively remove one or more custom tracks, return to the Table Browser, select "All Tracks" in the "group" menu. Your custom-defined tracks will now be listed in the "track" pulldown menu. Select the track you wish to remove. A new button labeled "Remove Custom Track" will appear; click it. To return to the Genome Browser display, use the "Genomes" link in the top panel.

Mapping properties of highly conserved elements

14| An advantage of having a custom track holding only the elements of interest is that you can now intersect this track with any other track available for the same assembly. These intersections help prioritize and classify the different conserved elements for further scrutiny—conserved regions upstream of genes may be *cis*-regulatory, elements in 3′ untranslated region may be involved in post-transcriptional regulation, intronic elements may be involved in splicing regulation, among others.

(i) For example, to see which of the highly conserved elements overlap exons of protein coding genes, click on the "Tables" link at the top of the browser display to return to the Table Browser. Make sure the "clade", "genome" and "assembly" entries are appropriate. For "group" select "All Tracks" and in the "track" pulldown menu find the name of your newly minted track. Next, make sure that the entire region of interest is selected (and not just a smaller region from which you may have clicked the "Tables" link above). Click the "create" button next to "intersection". Select "Gene and Gene Prediction Tracks" and "Known Genes" from the pulldown menu.

(ii) The page offers several different ways to compare and contrast any two tracks using complementation, union and intersection of the genomic regions they cover. To see which elements overlap a known gene you may use the default setting. Click "submit" to return to the main Table Browser display.

15| As was the case in setting up a filter (Steps 10 and 11), you may now use the "summary/statistics" and "edit" or "cancel" buttons near the "intersection" label to see how many elements match the intersection and then refine it. You can download this shorter list and later organize a set of these intersections into a table. To do this, select the appropriate "output format", type in the desired output file name, set the "file type" and click "get output". You can also scrutinize this list in the form of a second custom track, which will be displayed below your first custom track. To do this, repeat the instructions in Step 12. You can also continue working only with those elements that do not intersect a coding gene. To do so, edit the intersection appropriately and then load the results as a custom track.

16| The UCSC "Known Genes" track used in Step 14 is an attempt to collect and map the structures of many protein-coding genes (see the track details page) and is available only for selected genome assemblies. We recommend, however, that you also perform similar intersections with other protein-coding genes, pseudogenes, noncoding genes and gene prediction tracks that appear to be reliable in the region of interest. The details page of individual genes or predictions, accessible by clicking a gene structure when the track is in "pack" or "full" mode, has additional information about the item. This information is often useful in determining the value of a specific annotation.

17| To assess the transcriptional potential of the elements, intersect them, as demonstrated in Step 14, with selected tracks in the "mRNA and EST tracks" group.

Transcription data can be noisy. The details page of individual transcripts shows how well the transcript matches the genomic DNA. It often has information about the library from which the transcript was sequenced, as well as a link to its GenBank record and more information. If you suspect a transcript to be triggered by genomic priming, you can configure the "Short Match" track to highlight poly(A) tracts and examine its 3′-end vicinity.

18| The "Conservation" track (Step 6) shows in which other species a region of interest is found. To perform this screen for several elements simultaneously, you may intersect your custom track with the chain or net track of each species of interest. The chain track is created by stringing together pairwise alignment blocks using a permissive gap scoring scheme that allows for local inversions and overlapping deletion events. The chains thus attempt to span genomic regions shared by the two species, owing to both speciation and duplication events. The net track picks the best-scoring chain for every region of the reference genome, attempting to capture orthologous regions[11].

(**i**) Note that through these tracks you can also examine the genomic context of the interval of interest in the aligned species. Turn the "chain/net" track to "pack" or "full", zoom to the region of interest, click a filled bar within the track and in the resulting details page you can select to see either the entire corresponding chain/net in the other species, or just the region corresponding to the browser window.

(**ii**) To examine the orthologous gene in more distant species, click the respective gene in the "Known Genes" track, where available, and follow the "Other Species" link in the gene details page.

▲ **CRITICAL STEP**

19| It is often valuable to look for paralogous elements within the reference genome itself[5]. For that, you may use the "Self Chain" track, where available, to examine paralogous locations both manually and via a Table Browser intersection.

Where the self-chains are not available, you can manually search for additional occurrences of a few elements by obtaining the raw sequence via the "DNA" link at the top of the Genome Browser page and searching for it via the "Blat"[12] link in the same panel. You can also examine genes similar to the gene in the vicinity of which the conserved elements lie, by following the "Gene Sorter"[13] link in the top panel. Several gene similarity measures are available, including sequence homology, expression homology and functional annotation homology.

20| You can also intersect with your own custom tracks: click the "Genomes" link at the top of the Genome or Table Browser page and click the "add your own custom track" button. The custom track upload page contains a link to detailed help about custom tracks, as well as a link to a growing repository of custom tracks that researchers worldwide have contributed.

Once a custom track is loaded in the Genome Browser, it also becomes available for intersection in the Table Browser and is found under the "All Tracks" group.

21| Apart from scrutinizing genome drafts within the UCSC Genome Browser, you should also consider searching one of the large sequence repositories such as GenBank[14] or EMBL[15]. This search is especially useful for detecting hits from species that have been only partially sequenced.

To facilitate external sequence analysis, note that the Table Browser has a "sequence" option in the "output format" menu that exports the DNA sequences underlying your selection.

22| When a conserved element overlaps a transcript not tagged as coding or has a sequence match to a species more distantly related than expected (see **Box 1**), you may wish to assess whether it overlaps an uncharacterized coding exon, or a pseudogene. In such cases, it is advisable to use the Genome Browser to examine whether the conserved element is included reliably in any gene prediction. To view the six reading frames of the element, turn the "Base Position" track to "full". When you zoom in sufficiently, three reading frames appear. Click the arrow at the left of the base position track to alternate between the three Watson and three Crick strand frames.

Predicting signatures of noncoding function

23| It is valuable to attempt to characterize different types of functional signatures within your set of highly conserved elements, even if you are interested in one particular type of function. Consider the following possibilities and recommended tools for evaluating each.

(i) Our current understanding dictates that enhancer sequences contain binding sites for one or more enhancing or repressing transcription factors. Unfortunately, reliable binding site prediction is extremely difficult and current tools, including comparative ones, predict many false positive matches. For ways to improve prediction accuracy, see reference 9. You may also want to consult a recent evaluation study of many of the available tools[16], before deciding which tool to use.

(ii) We have a better understanding of the sequence constraints imposed on noncoding RNAs. Tools for comparative prediction of folding potential for small RNAs have matured recently, and we expect to add a reliable prediction track to the Genome Browser in the near future. Meanwhile, one may select tools using a recent evaluation study of several of these[17].

(iii) Unfortunately, our current understanding of other classes of functional noncoding elements, such as insulators and chromatin structure–associated sequences, is very limited. The few tools that do exist[18] should thus be treated with due caution.

BOX 1 EXAMPLE: VERTEBRATE ENHANCERS

An enhancer is a regulatory DNA sequence, often hundreds of bases long, which can exert its effect over large genomic distances[25]. It can be found as far as 1 Mb away[20,21], upstream or downstream[21,26] of the gene it regulates. It can reside in an intron of the regulated gene[25], or even in an intron of a neighboring gene[20,21]. It is thought that each enhancer is responsible for a subset of the total expression pattern of the gene it regulates[25], usually conferring to it tissue or cell type specificity[21,25]. Indeed, the known enhancers are often found in large tracts of the human genome devoid of protein-coding genes known as gene deserts, affecting genes with complex expression patterns and key developmental genes[21]. Structurally, enhancers are thought to mediate gene expression by containing several binding sites for different sequence-specific transcription factors, most often a combination of activators and repressors[25]. They seem, both from experiments and established protocols, to work regardless of their orientation in DNA[26]. Little else is understood of the constraints that dictate their actual primary sequence.

From a comparative perspective, strong conservation across many lineages seems to often correlate with enhancer activity[2–4]. But mammalian enhancers have also been characterized that cannot be found in the current genome drafts of ray finned fish[3]. By contrast, it seems that sequence conservation between vertebrates and sea squirt (or beyond) is often indicative of coding sequence from uncharacterized genes or transposon open reading frames (G.B., unpublished observations). Perhaps surprisingly, most conserved elements, including characterized enhancers, are found to be unique within their respective genome, using standard sequence comparison tools[5]. Nonetheless, some of these elements do have paralogs. Paralogous elements found near paralogous genes suggest a strong functional relationship between the gene and the element[5]. Note that although paralogous instances in general serve as confidence boosters for functional importance, some functional assays (such as knockout, knockdown and mutation) may be hindered by phenotypic rescue from the surrogate elements.

Finally, there are known cases of RNA transcripts originating from the location of an active enhancer, suggesting a possible connection between the act of transcription and chromatin accessibility of actively transcribed regions[21]. Although an overlap between a mature transcript and a conserved sequence is not evidence enough to exclude enhancer activity, it does seem that highly conserved elements overlapping characterized exons are most often involved in RNA editing or alternative splicing of the overlapped exon[2]. Likewise, to the best of our knowledge, no characterized enhancer is known to overlap a noncoding RNA. All the properties we have just enumerated can be combined using our protocol into a screening process that yields a high quality list of putative novel enhancers.

When intending to perform functional assays, you should also attempt to predict in which tissues or cell types, and at which developmental stage, a putative enhancer may affect gene regulation. Our best guess at the moment combines knowledge of the expression repertoire of the putative target gene, of the transcription factors predicted to bind to the enhancer and of expression patterns of orthologous, paralogous and nearby enhancers. As we improve our understanding of enhancers and other conserved noncoding sequences, our ability to rank candidate lists is also bound to improve. Such future improvements can be easily integrated into the protocol outlined here.

Ranking the elements for functional analysis

24| Once you have listed the different properties of the highly conserved elements, you may want to pick a subset and prioritize it for functional characterization. To do so, it is useful to draw an informatic profile of the functional class you are interested in and match the elements against it. The example described in **Box 1** illustrates one such class, that of vertebrate enhancers.

TROUBLESHOOTING TABLE

PROBLEM	SOLUTION
Steps 1–24 Something went wrong. How do I contact you?	Click the "Home" link at the top panel and then "Contact Us" at the bottom of the left tab in our front page. The resulting page explains when and how to get in touch with us (if you cannot reach this page, e-mail genome@soe.ucsc.edu). At the top of this page you will find "Help" and "FAQ" links. You can also use this page to subscribe, browse or search our mailing lists, which have announcements and user questions and answers.
What data or tools are available for download?	Consult our downloads page, reachable by clicking the "Home" link in the top panel and then "downloads" on the left tab in our front page.
Steps 2–24 The actual UCSC Genome Browser display differs from the description and/or figures in this document.	The browser evolves continuously. Examine the actual display; while appearances may change, we attempt to improve the interface incrementally and coherently. If you fail to perform the desired action, contact us (see above).
Step 6–24 A track I am interested in, or that is mentioned in this document, is not available for the assembly I am in.	Not all tracks are available for all assemblies. Look for a similar track in the appropriate track group. To check whether your track of interest is available in another assembly use the "Genomes" link in the top panel to return to the gateway page (see Step 3). You may contact us (see above) to inquire about upcoming availability of certain tracks.

CRITICAL STEPS

Step 5 There is now no clear way to define the extent of the region where *cis*-regulatory elements affecting any given gene may lie. This region seems to depend on the proximity and expression profile of nearby protein-coding genes. Where available, the UCSC "Known Genes" gene details pages can provide valuable information about the functional annotation and expression patterns of these genes. They can also be used to hypothesize alternative targets for enhancers characterized out of genomic context. See **Box 1** for location preferences of known enhancer sequences.

Step 10 Clearly, no simple correlation exists between sequence conservation and function. Thus, no thresholding will, in general, yield an exhaustive set of functional elements within a region. Genomic elements may be functionally important without showing cross-species conservation, and vice versa. Nonetheless, the more conserved an element is between farther diverged species, the more likely it is that this is due to strong negative selection for function. Barring additional knowledge of your region of interest, it is recommended to set the conservation threshold according to the number of elements you are interested in screening and their genomic distribution.

Step 18 The fewer assemblies your multiple alignment holds (pairwise in particular) and the lower you set the conservation score cutoff threshold in Step 10, the more likely it is for contamination-based hits to enter your list. Contaminants are regions that are foreign to the assembled genome. They are present owing to some sequencing or assembly mishap and are eventually weeded out in later assemblies. A contaminating human DNA sequence present in a fish assembly, for example, will generate a near-perfect match between the false fish region and the human one. Such hits will spuriously boost the interval conservation score. The chains and nets, described in the body of this step, are valuable in spotting and weeding out contaminants. When examining an

alignment between two species that is missing from a related genome assembly (for example, a human-fish match that is missing in mouse), remember to make the "gap" track visible. Holes in a chromosomal assembly, or a partial set of scaffolds, can also explain genomic omissions. But be most suspicious when a whole scaffold in one species matches near-perfectly another species, but does not assemble with any other region in its own genome draft. Chain and nets are also valuable for spotting unflagged protein coding and noncoding pseudogenes, including ribosomal ones. These would often appear as highly conserved between the different species, but matching copies would be nonsyntenic, as a result of independent retroposition events in the different lineages.

COMMENTS

We have outlined a screen to identify and rank genomic noncoding elements for functional characterization. The screen relies on conservation across sufficient evolutionary time as an indication of strong negative selection owing to functional constraints. Function, however, does not necessarily entail a strong cross-species conservation signature, and vice versa. Functional elements can be gained or lost in different lineages. They can also diverge in terms of primary sequence while maintaining their function. By contrast, local mutational cold spots or hyper-repair mechanisms may account for conserved regions that have no function[2]. Nonetheless, conservation remains the single most powerful 'universal' indicator of function, made apparent by the vast majority of different functional regions that have already been characterized. Even the use of several closely related species has recently been shown to hold potential for functional element detection, owing to the fact that each species diverges independently[19]. Without the comparative method, there remains the much harder task of detecting functional elements based solely on characteristics of their primary sequences. Having invoked the conservation paradigm, we are left still with the challenge of classifying the putative functional elements into different classes. But combining conservation with a partial profile, as that of enhancer sequences presented in **Box 1**, seems to yield viable candidates with encouraging success rates[20,21]. Technically, the extent of a conserved element, as well as its conservation score, can be defined in several different ways. One can compare an alignment block of two or more species and demand a minimal number of columns within it to be identical[2,3]. More sophisticated methods that take into account phylogenetic distances and neutral substitution rates include two related binomial-based and parsimony-based scoring methods[22] and a method that looks for nucleotide substitution deficits compared to the observed neutral rate[23]. The method we rely on estimates two substitution rate trees that differ by a multiplicative factor, one for conserved and another for neutrally evolving positions (expanded on in the "Most Conserved" track details page). It is beyond the scope of this work to compare the different methods. Also note that when delimiting genomic regions based on conservation alone, several nearby functional units may end up in a single conserved region. Independent matches of subintervals[5], distinct clusters of binding sites and other interval properties may help parse correctly these 'multidomain' genomic regions. To conclude, we have described a general method to obtain highly conserved genomic elements from the diverse set of species represented in the UCSC Genome Browser, to characterize them and to rank them for further scrutiny. Such screens serve as the first step when we turn to tackle the grand challenge of deciphering the full repertoire of heritable function.

EXAMPLE OF APPLICATION: HUMAN *DACH1* GENE ENHANCERS

In a recent functional study Nobrega and colleagues[24] scanned the genomic region flanking the human *DACH1* gene for enhancer sequences. This gene was chosen because it is known to be involved in early development, it is surrounded by two large gene deserts, and its promoter sequence shows evidence of paucity of regulatory sequences. Using a method similar to ours, the authors picked nine human genomic regions, all highly conserved in vertebrates as diverged as fish, and tested them for enhancer activity using reporter gene constructs in transgenic mouse embryos at day 12.5–13.5. As shown in **Table 1**, seven of the nine constructs (including only one of three ultraconserved elements[2] present in the screen) drove the expression of the reporter gene in reproducible patterns. These were found to be subsets of the *DACH1* gene expression repertoire.

Table 1 Properties of human *DACH1* elements screened for enhancer activity in transgenic mouse embryos[24].

Construct (result)	Conservation display score	Distance (kb) to *DACH1*	cDNA/EST overlap	Species coverage	Human paralogs	Comment
Dc1 (–)	715	+646 (3′)	none	core in fish	not of core	includes uc.347
Dc2 (+)	750	+139 (intron)	fish EST	core in fish	+336 kb *DACH2* intron	
Dc3 (+)	750	+14 (intron)	none for core	core in fish	+30 kb *DACH2* intron	
Dc4 (+)	726	–227 (5′)	none	core in fish	not of core	includes uc.351
Dc5 (–)	685	–253 (5′)	none	part to chicken, part to fish	none	includes uc.352
Dc6 (+)	733	–560 (5′)	fish EST	core in fish	none	gap in rat assembly rn3
Dc7 (+)	813	–641 (5′)	none	core in fish	in CHM intron, –220 kb of *DACH2*	
Dc8 (+)	759	–646 (5′)	none	core in fish	in CHM intron, –237 kb of *DACH2*	
Dc9 (+)	718	–783 (5′)	none	core in fish	not of core	

Construct name follows that given in reference 24. Result indicates whether the construct drove expression reproducibly in mouse embryonic day 12.5–13.5, or not (labeled + or –, respectively). The conservation display scores (see Step 7) are from the current eight-way multiple alignment in the UCSC May 2004 (hg17) human genome freeze and are given for the top scoring region, or core, of the construct. Distances to *DACH1* (Hsa 13) and *DACH2* (Hsa X) are measured from their documented transcriptional start site. The uc entries (ultraconserved elements) nomenclature follows that given in reference 2. These can be loaded as a browser custom track (Step 20; http://www.soe.ucsc.edu/~jill/ultra.html). Note that some of these properties may change as genome drafts improve and as more cDNA and EST libraries are sequenced.

ACKNOWLEDGMENTS

The authors are grateful to the UCSC Genome Browser staff for continuously developing and maintaining the infrastructure on which this protocol relies. We also thank N. Ahituv for critically reading the manuscript and M. Nobrega for advice. A.C.S. is supported by the Graduate Research and Education in Adaptive bio-Technology (GREAT) Training Program of the UC Systemwide Biotechnology Research and Education Program, grant #2004-33. D.H. and G.B. are supported by a US National Human Genome Research Institute grant P41 HG002371. D.H. is also supported by a Howard Hughes Medical Institute grant.

SOURCE

This protocol was contributed directly by the authors listed on the first page. For additional information describing the use of databases and search programs, please see Mount, D.W. *Bioinformatics: Sequence and Genome Analysis* 2nd edn. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2004).

1. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
2. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
3. Ovcharenko, I., Stubbs, L. & Loots, G.G. Interpreting mammalian evolution using Fugu genome comparisons. *Genomics* **84**, 890–895 (2004).
4. Woolfe, A. et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2004).
5. Bejerano, G., Haussler, D. & Blanchette, M. Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics* **20** (suppl. 1), I40–I48 (2004).
6. Dermitzakis, E.T., Reymond, A. & Antonarakis, S.E. Conserved non-genic sequences—an unexpected feature of mammalian genomes. *Nat. Rev. Genet.* **6**, 151–157 (2005).
7. Glazov, E.A., Pheasant, M., McGraw, E.A., Bejerano, G. & Mattick, J.S. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res.* (in the press).
8. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
9. Papatsenko, D. & Levine, M. Computational identification of regulatory DNAs underlying animal development. *Nat. Methods* **2**, 529–534 (2005).
10. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32,** 493–496 (2004).
11. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci.* **100**, 11484–11489 (2003).
12. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
13. Kent, W.J. *et al.* Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.* **15**, 737–741 (2005).
14. McGinnis, S. & Madden, T.L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*

32, 20–25 (2004).
15. Kanz, C. *et al.* The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **33**, D29–D33 (2005).
16. Tompa, M. *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23**, 137–144 (2005).
17. Gardner, P.P. & Giegerich, R. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* **5**, 140 (2004).
18. Glazko, G.V., Koonin, E.V., Rogozin, I.B. & Shabalina, S.A. A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genet.* **19**, 119–124 (2003).
19. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
20. Ahituv, N., Rubin, E.M. & Nobrega, M.A. Exploiting human–fish genome comparisons for deciphering gene regulation. *Hum. Mol. Genet.* **13** (special issue 2), 261–266 (2004).
21. Kleinjan, D.A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32 (2005).
22. Margulies, E.H., Blanchette, M., Haussler, D. & Green, E.D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507–2518 (2003).
23. Cooper, G.M. *et al.* Characterization of evolutionary rates and constraints in three Mammalian genomes. *Genome Res.* **14**, 539–548 (2004).
24. Nobrega, M.A., Ovcharenko, I., Afzal, V. & Rubin, E.M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
25. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).
26. Bondarenko, V.A., Liu, Y.V., Jiang, Y.I. & Studitsky, V.M. Communication over a large distance: enhancers and insulators. *Biochem. Cell Biol.* **81**, 241–251 (2003).