Mulan: Multiple-sequence local alignment and visualization for studying function and evolution

Ivan Ovcharenko, Gabriela G. Loots, Belinda M. Giardine, et al.

Genome Res. 2005 15: 184-194
Access the most recent version at doi:10.1101/gr.3007205

Mulan: Multiple-sequence local alignment and visualization for studying function and evolution

Ivan Ovcharenko,[1,6] Gabriela G. Loots,[2] Belinda M. Giardine,[3] Minmei Hou,[4] Jian Ma,[4] Ross C. Hardison,[3] Lisa Stubbs,[2] and Webb Miller[4,5]

[1]Energy, Environment, Biology and Institutional Computing, and [2]Genome Biology Division, Lawrence Livermore National Laboratory, Livermore, California 94550, USA; [3]Department of Biochemistry and Molecular Biology, [4]Department of Computer Science and Engineering, and [5]Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

Multiple-sequence alignment analysis is a powerful approach for understanding phylogenetic relationships, annotating genes, and detecting functional regulatory elements. With a growing number of partly or fully sequenced vertebrate genomes, effective tools for performing multiple comparisons are required to accurately and efficiently assist biological discoveries. Here we introduce Mulan (http://mulan.dcode.org/), a novel method and a network server for comparing multiple draft and finished-quality sequences to identify functional elements conserved over evolutionary time. Mulan brings together several novel algorithms: the TBA multi-aligner program for rapid identification of local sequence conservation, and the multiTF program for detecting evolutionarily conserved transcription factor binding sites in multiple alignments. In addition, Mulan supports two-way communication with the GALA database; alignments of multiple species dynamically generated in GALA can be viewed in Mulan, and conserved transcription factor binding sites identified with Mulan/multiTF can be integrated and overlaid with extensive genome annotation data using GALA. Local multiple alignments computed by Mulan ensure reliable representation of short- and large-scale genomic rearrangements in distant organisms. Mulan allows for interactive modification of critical conservation parameters to differentially predict conserved regions in comparisons of both closely and distantly related species. We illustrate the uses and applications of the Mulan tool through multispecies comparisons of the *GATA3* gene locus and the identification of elements that are conserved in a different way in avians than in other genomes, allowing speculation on the evolution of birds. Source code for the aligners and the aligner-evaluation software can be freely downloaded from http://www.bx.psu.edu/miller_lab/.

A significant growth in sequencing the genomes of complex organisms, including the recent completion of the chicken genome, opens new horizons in the field of comparative genomics and compels improvements on current tools and methodologies devoted to the identification of functional regions in multiple sequence alignments. It has now been well established that blocks of evolutionary conservation identified by cross-species comparative analysis correlate with functionally important DNA regions such as protein-coding genes (Pennacchio et al. 2001; Gilligan et al. 2002) and transcriptional regulatory elements (Loots et al. 2000; Elnitski et al. 2001). Several recent methods have emphasized the importance of multiple-sequence alignments (when two or more sequences are simultaneously aligned to each other) for comparative studies. It has been shown that comparisons of multiple closely related sequences through the phylogenetic shadowing approach are capable of identifying primate specific exons and enhancers (Boffelli et al. 2003; Ovcharenko et al. 2004a). In parallel, evolutionary comparisons of human, rodents, frog, and fish genomes identified more distantly related gene regulatory elements (Lettice et al. 2003; Nobrega et al. 2003).

Several available Web-based tools implement multiple-sequence analysis either as a series of pairwise alignments with a selected reference sequence (Mayor et al. 2000; Schwartz et al. 2000; Ovcharenko et al. 2004b) or as a full multiple-sequence global or pseudoglobal alignment (Thompson et al. 1994; Bray et al. 2003; Brudno et al. 2003; Schwartz et al. 2003a; Ovcharenko et al. 2004a). Applications of these tools differ by the type of sequences (nucleotide or amino acid) they are capable of processing, as well as by the maximum length and number of allowable input sequences. The primary drawback of the presently available tools is that none are capable of generating multiple-sequence local alignments that would accommodate evolutionary sequence reshuffling and/or inversions in a subset of sequences while also allowing for dynamic selection of the reference genome.

Here we report a new integrative comparative tool, Mulan, that dynamically and rapidly generates multiple-sequence local alignments (MSLAs), and we present several examples for the application of this tool to study phenotypic differences in vertebrate species. The Mulan alignment engine consists of several data analysis and visualization schemes for high-throughput identification of functional coding and noncoding elements conserved across large evolutionary distances. Mulan (1) determines phylogenetic relationships among the input sequences and generates phylogenetic trees, (2) constructs graphical and textual alignments, (3) dynamically detects evolutionary conserved regions (ECRs) in alignments, and (4) presents users with several visual display options for the generated conservation profiles. This tool is also able to implement the phylogenetic shadowing strategy for identifying slow-mutating elements in comparisons of multiple closely related species (Ovcharenko et al. 2004a). In

[6]Corresponding author.
E-mail ovcharenko1@llnl.gov; fax (925) 422-2099.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3007205. Article published online before print in December 2004.

15:184–194 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

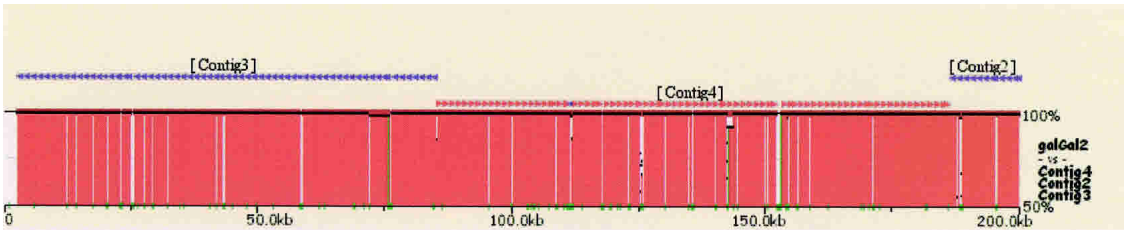

Figure 1. Mulan contig ordering based on homology to the reference sequence. The *top* layer of shaded lines indicates the location of contigs from a second sequence aligned to the base sequence, where red triangles pointing to the *right* specify forward-strand alignments, and purple triangles pointing to the *left* correspond to reverse-strand alignments. Contig names are indicated in square brackets. The JF2-73M16 chicken BAC clone (http://www.jgi.doe.gov/) consisting of three contigs was aligned to the chicken genome (chr28:4,000,000–4,200,000).

addition, Mulan is integrated with the MultiTF program that identifies evolutionarily conserved transcription factor binding sites (TFBSs) shared by all analyzed species, allowing for the decoding of the sequence structure of regulatory elements that are functionally conserved among different species. Mulan is publicly available at http://mulan.dcode.org.

Results

Alignment strategy

Mulan employs two alignment strategies that allow for comparative analysis of multiple sequences that are present either as (1) draft or (2) finished configuration. The first approach allows for the construction of an alignment for multiple draft-quality sequences and subsequently for effective order-and-orientation (O&O) of unfinished sequences based on the reference genome. The second approach operates with multiple high-quality single-contig (finished) sequences, and is the main subject of this paper.

Genomic sequences submitted to Mulan are aligned by the threaded blockset aligner (TBA) program for finished sequences and by the *refine* program for draft sequences (Blanchette et al. 2004). The local alignment approach utilized for both sequence types allows for reliable representation of inversions and genomic reshuffling events that have occurred in a subset of lineages since the last common ancestor. In doing so Mulan does not require colinearity between input sequences (as in the case of a global multiple alignment), but instead it generates different projections of the threaded blockset alignment to different reference sequences that are selected by the user. As a consequence, this approach ensures the detection of evolutionarily conserved elements throughout the alignment even if orthologous regions have been repositioned or inverted in only a subset of the input species.

Mulan alignment visualization is based on the zPicture display design (Ovcharenko et al. 2004b), where the reference sequence is linear along the horizontal axis and the percent identity is plotted along the vertical axis. In addition, Mulan contains a graphical annotation option for the alignment of draft sequences where contig names and alignment blocks can be visualized as tracks on top of the conservation profile (Fig. 1). Syntenic blocks are color-coded, allowing for easy O&O of draft sequences by using the base sequence as a structural guide.

Visualization and data analysis strategies for multiple-sequence local alignments

Multiple-sequence comparative analysis is a challenging task in terms of generating highly reliable alignments and graphically displaying the alignment results. To address the complexity stemming from user input sequence files that potentially consist of a large number of sequences of varying lengths and different

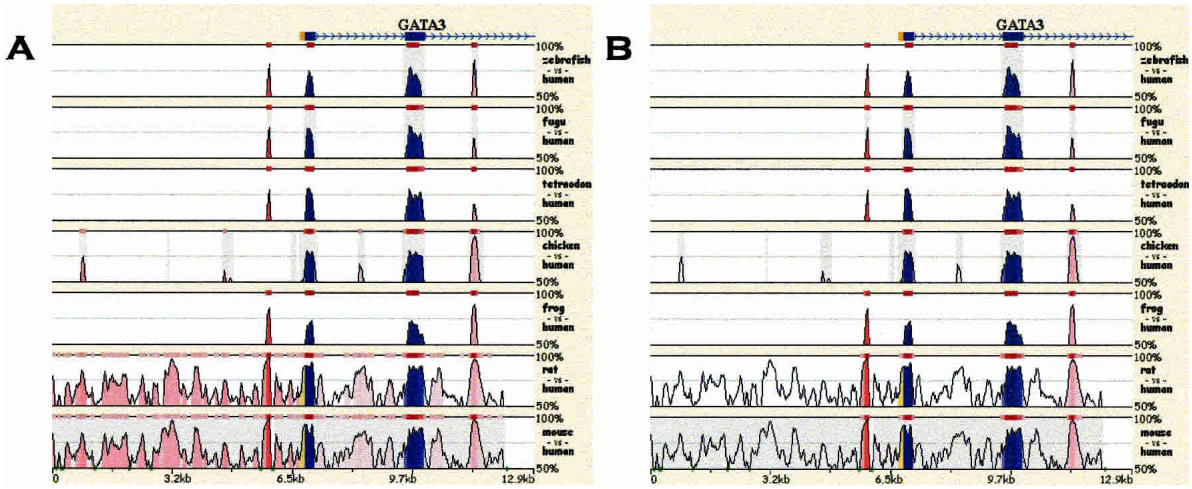

Figure 2. Stacked-pairwise conservation profile for a 13-kb region from the *GATA3* locus. Color-gradient visualization is implemented to differentially display regions that are differentially conserved in the input sequences (*A*). The color intensity of a conserved region depends on the number of different species that contain the region (the darker, the more conserved species). Only ECRs conserved in at least six out of seven total secondary species are highlighted in the alignment (*B*). Intergenic regions are in red, intronic in pink, coding exons in blue, and UTRs in yellow.

phylogenetic relationships, we provide a set of different visualization options applicable to any finished MSLA. For example, the reference sequence can be dynamically changed, and the new stacking order of conservation profiles with the rest of the species will be automatically determined using the evolutionary relationship of each sequence to the reference sequence, where more closely related species are at the bottom.

"Color density by interspecies conservation" illustrates a relationship between the color density of a conserved element and the number of species that share a particular region (Fig. 2) such that, the more species share a sequence, the darker the conservation profile will be displayed. (This analysis is performed for every pixel-wide region of the conservation plot. The number of ECRs from different species that overlap with a particular pixel count towards the number of species sharing this region.) In a recent study, it was observed that regions conserved in multiple species often correlate with functional elements (Frazer et al. 2004). Therefore, the color density of the plot can potentially highlight different DNA segments in the base sequence with unique evolutionary character. Similar to zPicture, Mulan allows for interactive and customized ECR analysis. Users can select the evolutionary criteria (length and percent identity) as well as indicate a specific requirement for an ECR element to be conserved in at least *n* number of species in the MSLA (Fig. 2). For example, while the pairwise human–mouse comparison for a 13-kb-long *GATA3* region identifies 34 ECRs (80 bp/70% ECR parameters; Fig. 2A), an eight-species scan of the multiple sequence conservation profile with a specified requirement that the human region is shared by at least six other species identifies the four most deeply conserved ECRs—two coding exons, an intronic and a promoter element located ~1 kb upstream of the gene transcription start site (Fig. 2B). This simple functional implementation of the Mulan tool immediately pinpoints key functional elements for the *GATA3* locus. Additionally, by clicking on an ECR the user can access the textual MSLA underlying the conserved region.

Two additional data representation modules are implemented in the Mulan tool: phylogenetic shadowing and "summary of conservation." While "summary of conservation" collects all the shared nucleotide similarities from all the pairwise comparisons into a single conservation profile, the phylogenetic shadowing option effectively collects all the cumulative nucleotide mismatches (Ovcharenko et al. 2004a). We tested the current implementation of the phylogenetic shadowing method on the *ApoB* locus, which has been sequenced and analyzed in comparisons of 14 different primate species (Boffelli et al. 2003). Mulan identified the correct phylogenetic relationship among the pri-

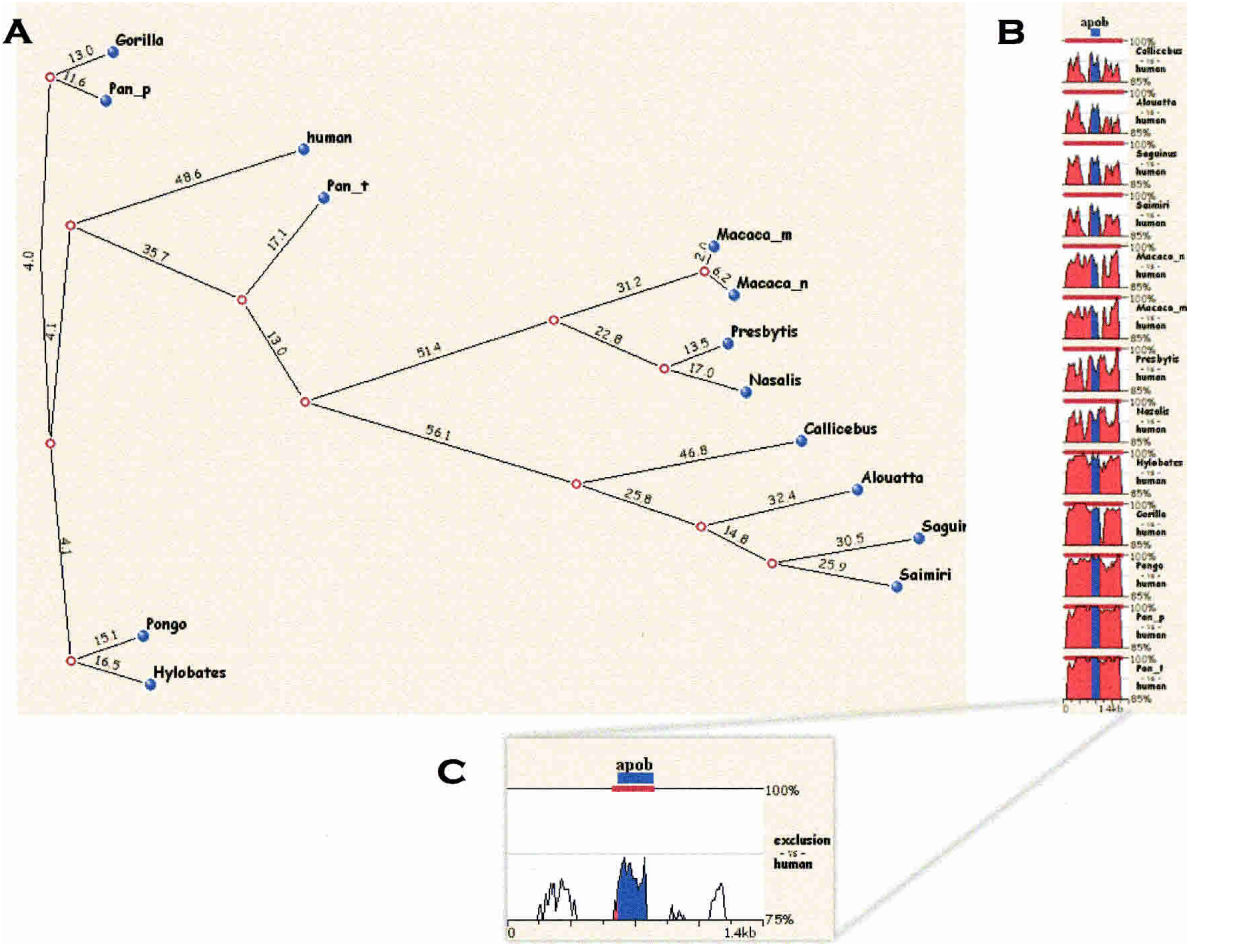

Figure 3. Phylogenetic shadowing option of the Mulan tool. *ApoB* region sequences from 14 primates were compared to determine the phylogenetic relationship (*A*) and visualize the conservation by stacked pairwise display (*B*). The phylogenetic shadowing conservation profile preferentially detects the *ApoB* coding exon from the neutrally evolving background (*C*). ECR parameters used for detecting exons: >85% identity; >100 bp.

mate species (Fig. 3A), and while the "stacked pairwise" approach preferentially detects the human *ApoB* exon only in comparisons between the most distantly related species (Fig. 3B), the phylogenetic shadowing visualization display accurately depicts the coding exon as the most highly conserved element in this region (Fig. 3C). In addition, the identified ECR sharply defines the exon boundaries without any a priori knowledge of its location.

Evaluation of alignment tools

To evaluate and compare the performance of the *refine* and TBA programs—the tools underlying the draft and finished Mulan alignment scheme, respectively—we followed the approach of Blanchette et al. (2004); that is, simulated neutral evolutionary processes were applied to a hypothetical ancestral sequence, where the frequency and length distribution of inversions were estimated from available genome alignments. The simulation program records the true relationship among the generated sequences, which can be regarded as the true alignment. The extent of agreement for alignments produced by aligners is then determined. There are several methods to measure this similarity. The agreement score defined in the original TBA paper is used here (Blanchette et al. 2004). In cases of alignments with inversions, there is an agreement if the i^{th} base position of sequence A is aligned to the j^{th} position of sequence B in both the predicted and the true alignments, under the condition that the two alignments have the same orientation, or the i^{th} position of sequence A is aligned to a gap in both the computed and the true alignments. The agreement score is determined from the fraction of positions of the predicted alignment that agree with the true alignment.

To be consistent in comparing aligners, the same BLASTZ parameters (C = 0, Y = 3400, K = 2000) were used for all data sets. TBA uses the guiding tree of *(((human chimp) baboon)(rat mouse)) ((cow pig)(cat dog)))*. *Refine* uses human as the reference sequence. The performance of aligners with respect to the agreement score is illustrated in Figure 4. Only representative pairs are shown, to illustrate how performance of an aligner varies with evolutionary distance.

Several observations can be made about the graph in Figure 4. First, for sequences at very short evolutionary distance, such as human versus chimp and human versus baboon, all methods work well. Second, *refine* performs as well as or a little better than BLASTZ alone for pairs containing the reference sequence, for example human versus mouse and human versus dog. However, for sequences being pulled together by *refine* instead of direct pairwise alignment, the performance is worse, for example, rat versus mouse and cat versus dog.

Third, TBA performs as well as or better than BLASTZ alone for all comparisons. For closely related species, TBA does not lose accuracy, while for distantly related species, TBA significantly improves accuracy (e.g., human vs. mouse). At the same time, TBA performs as well as or better than *refine*. TBA outperforms *refine* dramatically for cat versus dog and especially rat versus mouse. TBA builds alignments starting from leaves of the phylogenetic tree, utilizing the fact that pairwise alignment between two species with closer evolutionary relationship is more reliable than with distantly related species. For instance, TBA directly uses the rat–mouse alignment, whereas *refine* aligns rat to mouse based on information about how the two align to a distant intermediary. For instance, a human region might align to mouse but not to rat (rat is evolving slightly faster than mouse), though the corresponding mouse and rat regions are easily aligned to each other; TBA will correctly match the human, mouse, and rat regions, but *refine* will match only human and mouse.

Fourth, the regions of disagreement in an alignment are composed of mismatches, unidentified alignments, and false alignments. By regarding mismatches within five base positions as correct matches, TBA_5 shows a substantial increase on agreement score. In other words, mismatches in an alignment produced by TBA are frequently very close to their correct match positions. For some analyses, close agreement with the true aligned position is adequate. Although the performance of TBA is better than *refine* for certain cases, the running time for TBA is much longer than *refine*. For aligning nine species each with a length of ~50 kb, TBA takes ~50 sec on a modern workstation, while *refine* requires only ~7 sec.

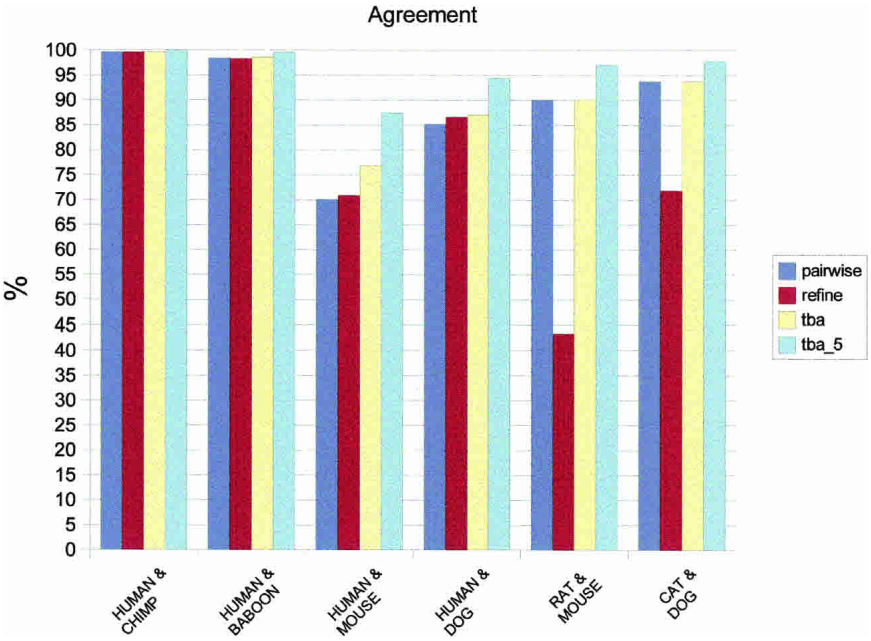

Figure 4. Agreement score of pairwise and multiple alignments produced by different aligners on a set of nine simulated mammalian sequences of length ~50 kb. Pairwise results from BLASTZ were postprocessed to remove overlapped regions. Multiple aligners including *refine* and TBA use the same pairwise alignments. TBA_5 refers to alignments from TBA, but the agreement score allows mismatches within five base positions. Agreement scores of multiple aligners are measured from the pairwise alignments induced by pairs of species. All values are averaged over 50 sets of simulation sequences. Parameters used in the simulation and alignment programs are described in the text.

From sequence evolution to genome biology

We applied Mulan to study the evolutionary conservation of the human *GATA3* locus. GATA3 is a very important molecule shown to be involved in various biological processes throughout development, in both the early embryo and adulthood (Lim et al.

2000; Van Esch and Bilous 2001; Lawoko-Kerali et al. 2002). In particular, it was recently shown that GATA3 is one of the key players involved in bone formation, differentiation of hair follicles, and tooth development (Andl et al. 2004). There are known to be distinct differences in these processes between humans, rodents, avians, amphibians, and fish. Therefore, we anticipated that we should observe some subtle genomic differences in a multiple sequence comparison at this locus between representative genomes from different evolutionary clades, spanning over 450 million years (Myr) of evolutionary time since the separation of mammals, amphibians, birds, and fish (Fig. 5).

Multiple-sequence Mulan alignments identified all the coding exons of the *GATA3* gene as conserved segments in all of the species, highlighting the functional importance of this protein and suggesting that interspecies differences associated with the GATA3 protein most likely originate from differences in noncoding sequences. This is supported by noncoding conservation patterns that significantly differ in comparison of the human sequence with different species (Fig. 5B). Three main groups of conservation were identified: human/rodents, amphibian/fish, and chicken. Five ECRs (ECR1–ECR5) are shared by at least four different species (including human). One of them, the intronic ECR5, was present in all species, suggesting a key role of this element for the *GATA3* locus. For example, it could be a general enhancer element responsible for the expression of this gene. Three other ECRs, upstream ECR1 and ECR2 and intronic ECR4, are shared only by humans, rodents, and chicken and are not detected in either frog or fish lineages (there are no remains of sequence conservation of these ECRs in indicated species that would be displayed as short lines on Fig. 5B otherwise), suggesting a putative differential expression of the *GATA3* gene in these two groups of genomes as regulated by this subset of three ECRs. One could speculate that the key involvement of the *GATA3* gene in the hair/feathers growth regulation pathway could be indeed regulated by one of these three ECRs, and that their absence from the frog and fish genomes may be responsible for the lack of hair in these species. More interesting is the conservation of the ECR3 element across multiple species. This element is present in all but the chicken genome. While the conservation with fish suggests functionality of this element (Ghanem et al. 2003; Lettice et al. 2003; Nobrega et al. 2003), an absence of this element in the chicken lineage would suggest that the function driven by this element (which could be a *GATA3* enhancement at a particular stage in a particular tissue) is absent

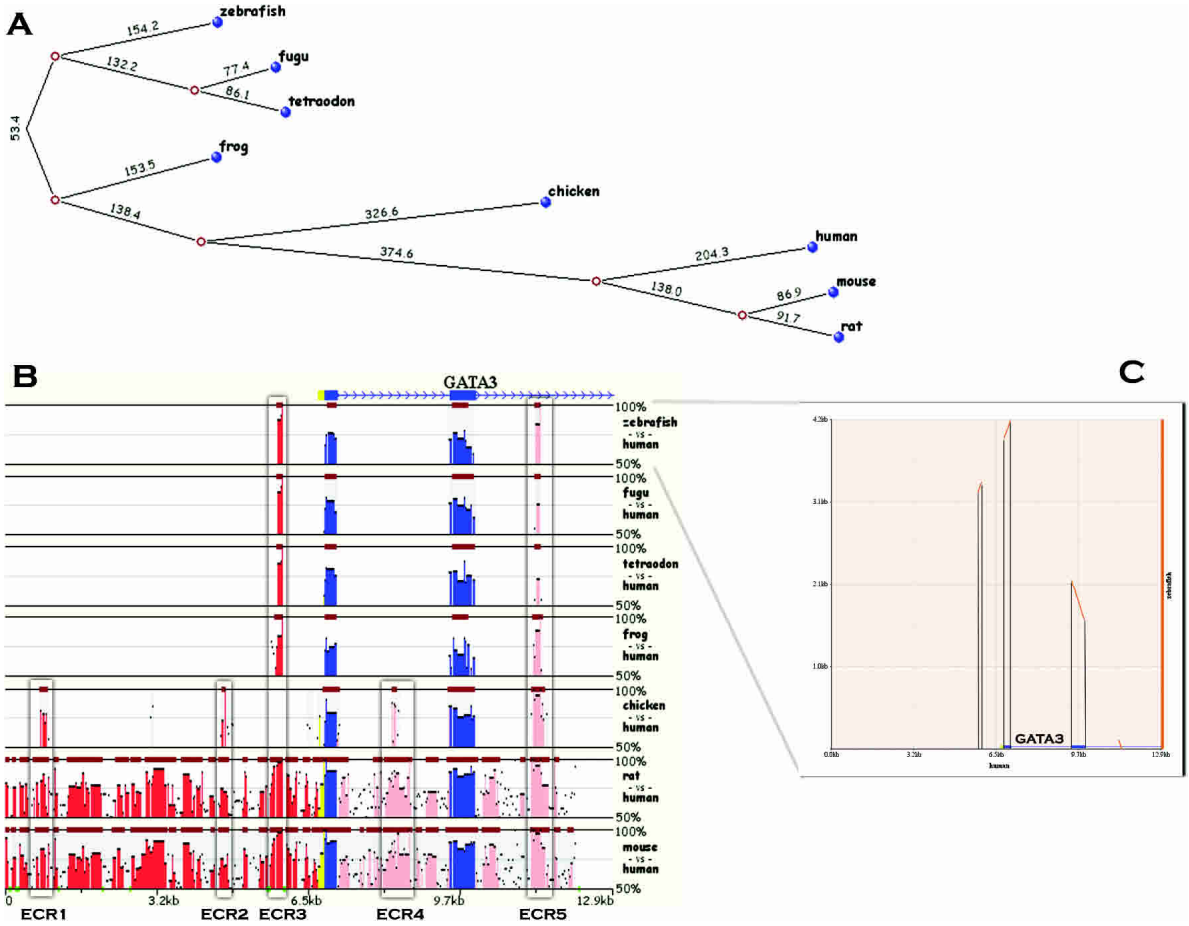

Figure 5. Mulan phylogenetic tree (*A*) and sequence conservation profile (*B*) for the *GATA3* gene locus from human, rat, mouse, chicken, frog, and three fish genomes. Each tree branch indicates the number of nucleotide substitutions from the closest node. Noncoding ECRs conserved (>70% identity; >80 bp) in at least four species (including human) are shaded and numbered ECR1–5. Coding exons are in blue, UTRs in yellow, intergenic elements in red, and intronic in pink. ECRs are depicted as dark red bars above each pairwise alignment. Repetitive elements are depicted as green boxes on the *bottom* axis. Alignments resulting from the reverse strand are shaded in gray, and blocks on the forward and reverse strands can be visualized in a dot-plot between the zebrafish and the human local alignment (*C*).

in birds. Could it be that this is a silencer element that blocks GATA3 involvement in wing bone development, and its absence correlates with the development of wings in birds? While practical application of the Mulan tool for comparative analysis of multiple species can generate different hypotheses similar to the ones described, only follow-up experimental biology can provide concrete answers to these questions.

It is also interesting to mention that the local alignment nature of the TBA aligner (which constitutes the core of the Mulan tool) enables the correct recapitulation of the conservation profile for the *GATA3* locus with all the species. In particular, the draft quality of the zebrafish genome represents this locus as a combination of forward- and reverse-strand sequences joined together (Fig. 5C). The synteny breakpoint appearing after the first *GATA3* exon is probably just an artifact of the assembly of this locus. Otherwise it would destroy the integrity of the *GATA3* ORF in zebrafish.

Multiple-sequence conservation of transcription factor binding sites

The ability to accurately predict functional transcription factor binding sites (TFBS) is a powerful approach for sequence-based discovery of gene regulatory sequences and for elucidating gene regulation networks and mechanisms. To combat the overabundance of false-positive computational predictions stemming predominantly from the small size of TFBS footprints and from poorly defined position weight matrices (PWM), evolutionary sequence analysis has been proposed as a robust strategy for filtering out false-positive sites (Loots et al. 2002; Aerts et al. 2003; Lenhard et al. 2003; Loots and Ovcharenko 2004). Mulan incorporates a TFBS analysis tool, multiTF that is similar to pairwise-alignment-based rVista 2.0 (Loots et al. 2002; Loots and Ovcharenko 2004; http://rvista.dcode.org), but implements a different method of detecting TFBS present in all the sequences included in the multiple alignment.

We used the Mulan/multiTF combination to analyze the distribution of TFBS in ECR3 from the *GATA3* locus that is shared by all vertebrate species but chicken (Fig. 6). This analysis was aimed at providing an in silico evidence for the bone-specific function of this element to support the hypothesis that the absence of this element could possibly be related to the process of wing formation in birds. PWM matrices for 399 vertebrate TFBS families available from the TRANSFAC Professional 7.3 library (http://www.biobase.de/) were used to scan for binding sites that are shared among all species except chicken (Fig. 6). We used the "optimized for function" search option of the multiTF that weights PWM matrices differently by minimizing and balancing out the abundance of false-negative hits from different matrices.

Interestingly, only one putative TFBS corresponding to the CRE-BP1 regulatory protein was detected by multiTF in the scan of the ECR3 multiple-sequence alignment to be shared by all the species using almost 400 other TFBS matrices (Fig. 6). CRE-BP1, also known as ATF2, has been shown to trigger the development of primary fibrosarcomas in the chicken wing (van Dam and Castellazzi 2001). The interconnection between the role of this regulatory protein in the chicken wing, the detection of CRE-BP1 TFBS in an unbiased screen of the multiple-sequence alignment for the ECR3 element, and the absence of this conserved element only from the chicken genome supports the idea that this element functions as a regulator of *GATA3* transcriptional activity in bone development and possibly participates in specification of wings in avians. One can speculate that if the deletion of this element was one of the factors that resulted in hollow bones in birds, then it could be that this deletion was an early step towards the evolutionarily development of wings. While the in vivo function of the ECR3 element, the direct regulation of *GATA3* transcription by CRE-BP1 protein, and the disruption of this pathway in birds is speculative, this example illustrates how the Mulan/multiTF theoretical approach can be efficiently applied to generate and refine in silico gene regulation predictions for a set of homologous sequences. Such computational prescreens prioritize targets to be used in subsequent in vivo experiments, as well as establish new potential molecular links that have not yet been defined experimentally.

To demonstrate the cumulative effect of searching for TFBS in multiple sequence alignments and the dramatic functional enrichment resulting from each additional sequence incorporated into the comparison, we analyzed several regions encompassing known functional sites (Table 1). We selected three genomic regions ranging in size

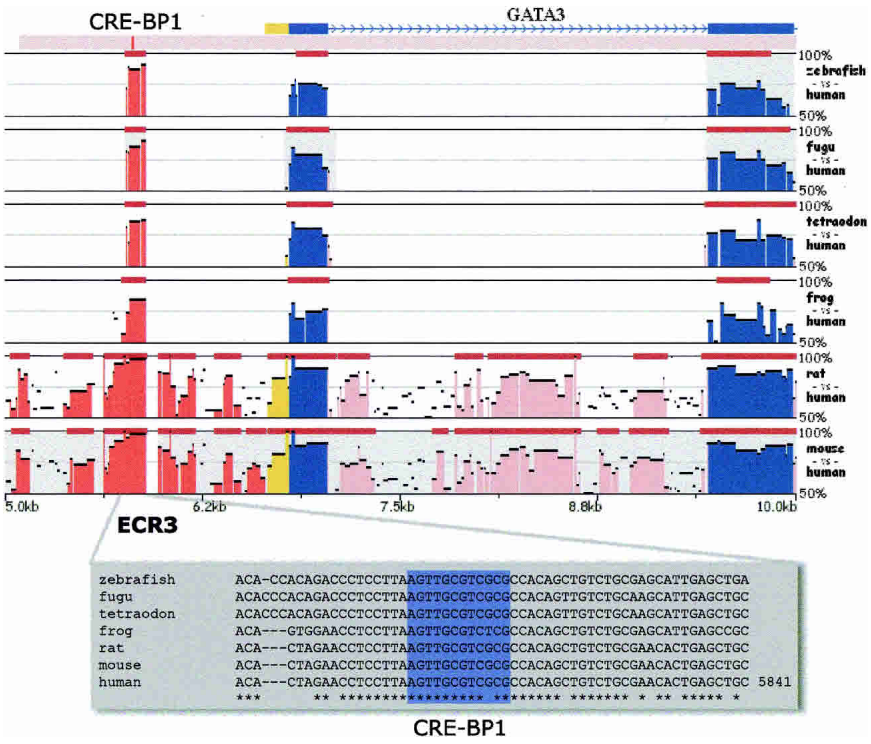

Figure 6. multiTF visualization of CRE-BP1 transcription factor binding site detected in the *GATA3* locus overlaid with the conservation profile of this locus as constructed with human, mouse, rat, frog, *Fugu*, tetraodon, and zebrafish sequences. The *bottom* panel represents a 60-bp-long alignment for the ECR3 core region that contains the CRE-BP1 binding site (blue) shared by all the species.

from 150 kb to 230 kb and corresponding to *PAX6*, *NKX2.5*, and *NKX2.9*/*PAX9* genomic loci. It has been shown that *PAX6* has autoregulatory activity mediated though a *PAX6* TFBS located in an intron (Kleinjan et al. 2004). *NKX2.5* is tightly regulated by SMAD and GATA proteins, and several such sites have been mapped to promoter proximal regions (Brown et al. 2004). PAX9/NKX2.9 expression is controlled by the zinc finger transcription factor Gli, and a functional site has been mapped distal to the *PAX9* and proximal to the *NKX2.9* promoter (Santagati et al. 2003). Using Mulan/multiTF we searched for these previously specified TFBSs, first in human/mouse alignments, and then we systematically added rat, chicken, frog, and fish sequences and analyzed the sites preserved in each multiple alignment. In general, the most dramatic reduction in the number of predicted sites was observed in comparisons with rodents, eliminating 90%–97% of the total number of predictions for the human sequence alone. The addition of chicken sequences further reduced the number of predictions by five- to 20-fold, and the addition of frog usually preserved 2–5 final conserved sites. In all cases, the known functional sites were present among the conserved sites in the human/mouse/rat/chicken/frog alignments. The addition of fish sequences was informative ~50% of the time, but it is worth noting that in these cases only the known functional sites were preserved, suggesting that distant comparisons can be extremely useful when clear homology can be established. These data suggest that by analyzing TFBS patterns in multiple-sequence alignments, one can dramatically filter out sites that have diverged throughout evolution, and select for sites that are most likely functional. This analysis does not establish which set of organisms are ideal for these types of studies, since such comparisons will in general be region- and gene-specific.

Mulan-GALA interconnection and finding orthologous regions

The database of genomic DNA sequence alignments and annotations (GALA; http://globin.cse.psu.edu/gala/) allows users to find genomic intervals that meet defined conservation thresholds, alignment-based scores, gene annotation criteria, transcription factor bindings site patterns, expression profiles, and other features (Giardine et al. 2003). Once a region of interest has been found in GALA, a user may wish to examine it using the Mulan tool. Likewise, once an ECR element has been identified by using Mulan, users have the option to utilize GALA to find additional information about the region containing it. Thus, two-way data flow has been established between the GALA database and the Mulan server.

The interconnection link of GALA to Mulan is established through forwarding a list of homologous regions from different species from GALA to Mulan. One of the critical steps in generating a multiple alignment in a locus is identifying the homologous DNA intervals in the other species. This is complicated by the existence of paralogs of many sequences, generated by transposition, segmental duplications, and chromosomal rearrangements. Thus, a given DNA interval, say in human, may match to multiple locations in the mouse genome. Furthermore, a long DNA segment in human may match to several orthologous regions in mouse that could have a different order and orientation than the human sequence (Kent et al. 2003). The problem of automatically determining best orthologous regions is an open problem in comparative genome informatics.

We have implemented a partial, but quite useful solution, by using the chains and nets (Kent et al. 2003) of whole-genome BLASTZ alignments (Schwartz et al. 2003a). The program *liftOver* reads a chain of alignments and finds corresponding positions between those specified in a first species and their homologs in a second. Once a DNA interval is specified in GALA, the user can easily access a page to find estimated orthologous positions in other species. Currently, chains are available to convert among human, mouse, rat, and chicken. We have limited the search to chains that are on the top level of nets. This does miss some regions, because some DNA that is rearranged between species is on lower levels of the nets. Automatically finding a more comprehensive set of orthologous sequences is a goal of future work.

As an example, the *ZFPM1* gene, which encodes a multiple Zn-finger protein called Friend of GATA1 (FOG1), was identified in GALA and the orthologous regions were found in mouse, rat, and chicken. These were automatically transferred to Mulan, which also picked up the annotation from the *knownGenes* track at the UCSC Genome Browser (Kent et al. 2002). The alignments were computed by TBA and are displayed in Figure 7. Note that several intronic regions are highly conserved. We separately used GALA to find intervals with both a high regulatory potential score in human–mouse–rat multiple alignments (Kolbe et al. 2004) and a conserved match to weight matrices for GATA1 [computed genome-wide using *tffind* (Schwartz et al. 2003b) and recorded in GALA]. These intervals were added to the annotation file at Mulan, using the dynamic editor and displaying them as arrows labeled with the fold-enhancement in separate assays (Welch et al. 2004). The predicted *cis*-regulatory modules are verified at a high rate (Hardison et al. 2003; Welch et al. 2004).

Table 1. Multiple alignments in combination with multiTF were used to analyze the transcription factor binding site distribution of functionally characterized sites

	230 kb PAX6	155 kb NKX2.5		150 kb Pax9/Nkx2.9
	Kleinjan DA, 2004	Brown CO, 2004		Santagati F, 2003
TFBS	PAX6	SMAD	GATA	Gli
human	209	2027	1834	652
+ mouse	15	60	82	84
+ rat	15	59	63	61
+ chicken	3	3	10	5
+ frog	**1** + 1	**1** + 2	**2** + 3	**1** + 1
+ fish*	**1**	0	0	**1**

**Fugu rupripes*, *Fugu tetraodon* or zebrafish sequences were used interchangeably based on coverage.
With each additional species, the number of predicted sites was reduced. The addition of frog and fish mostly identified only the functionally known sites (in bold).

Interconnection with the UCSC genome browser database

Mulan is dynamically linked to the UCSC genome browser database (Karolchik et al. 2003). It is possible to automatically fetch sequence and gene annotation files for the human, mouse, rat, chicken, or *Fugu* genomes by transferring the UCSC genome browser positional address to the Mulan submission page. This process significantly facilitates the management of sequence dataflow and minimizes the time required for a user to prepare their own sequence files.

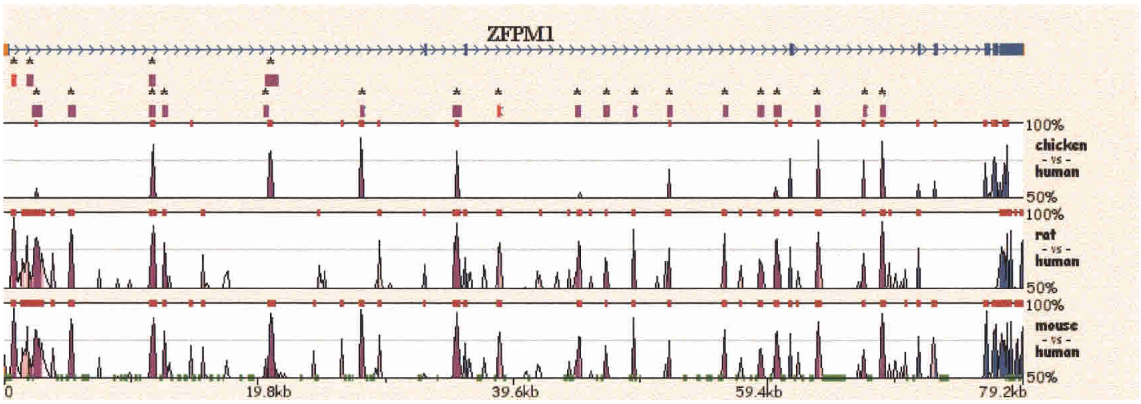

Figure 7. Conservation of *ZFPM1* among human, mouse, rat, and mouse, using TBA at the Mulan server. The large introns have several highly conserved regions. Those with conserved GATA-1 binding sites and high regulatory potential (predicted CRMs) are indicated by a set of purple and red blocks under the gene demarcated by star symbols. Red color of two block elements means that they are positive for binding GATA-1 in erythroid cells, as assayed by chromatin immunoprecipitation (Welch et al. 2004).

Discussion

The exponential growth of available DNA sequences produced by international genome-sequencing cohorts is creating an invaluable, enormous collection of genomic sequences from different eukaryotic and prokaryotic organisms. Particularly the addition of the chicken genome, *Gallus gallus*, marks a multifaceted advance in biology, largely due to the importance of this organism in agriculture and as a model for nonmammalian vertebrate development, but equally importantly due to its strategic evolutionary position in the tree of life between mammals and fish. The chicken genome provides a priceless substrate for genomic comparisons, and will allow us to better understand the overall genomic structure and evolution of vertebrates. To fully capitalize on this information-rich genome, we require innovative methods and tools for conducting creative comparative multispecies sequence analysis. Here, we described the Mulan tool that introduces a novel reliable approach to generate MSLAs. The tool is capable of producing fast and accurate alignments for both distantly and closely related organisms, such as humans, primates, fish, and chicken, properly taking into account the complexity of evolutionary sequence rearrangements such as inversions, transpositions, and subsequence reshuffling.

Mulan introduces several novel options for users to manipulate both the textual alignments and the graphical conservation displays to differently address the conservation structure of either closely or distantly related species. In particular, the option of coloring conserved regions using a gradient based on the number of species in which the region is conserved, coupled with a module that filters out ECRs that are shared by fewer than a requested number of species, permits straightforward identification of elements that are shared by a subset of species. This is illustrated through comparisons of the chicken *GATA3* locus to the orthologous regions from humans, rodents, frog, and fish. This type of analysis can be important for generating hypotheses about the function of ECRs shared by a limited number of species (Frazer et al. 2004).

The speed with which Mulan is capable of handling Megabase-long genomic sequences (on the order of minutes) and the dynamic character of the user interface are remarkable. Interactive conservation profiles allow user-selection of an ECR that displays the multiple-sequence alignment for that element. The dynamic interconnection between Mulan and the multiTF tool presents an effective way to identify transcription factor binding sites shared by multiple species. These tools can be used to predict the function of anonymous noncoding ECRs and to approach the description of gene regulation methods and networks. In addition, the draft alignment option of the Mulan tool allows easy O&O of chicken BAC contigs using the WGS assembly as the reference sequence.

In sharp contrast to several other available global multiple-sequence alignment tools, the threaded blockset alignment strategy implemented by Mulan detects and properly processes DNA rearrangements often characteristic of synteny among distantly related genomes. Also, it highlights subsequent reshufflings in order to restore all the changes responsible for the evolutionary history of multiple related sequences. Because of these features, Mulan permits the dynamic interchange of reference sequences and will accordingly generate textual (and graphical) MSLAs interactively, and very rapidly.

Methods

Generating alignments

Mulan aligns draft and finished sequences using different alignment strategies. The draft approach employs a combination of BLASTZ and *refine* programs (Schwartz et al. 2003b). Pairwise alignments between each secondary sequence and the reference sequence are done initially by BLASTZ. The single-coverage option is used at this stage to filter out low-scoring alignments that overlap with high-scoring ones. Effectively, this allows each reference sequence nucleotide to be covered by either one or no alignment block from one of the secondary sequence contigs in each set of pairwise alignments. Alignment postprocessing is carried out by the *refine* program, which collects all the pairwise alignments into a single FASTA-formatted gapped alignment file that is available for the user to download from the results Web page.

High-quality finished sequences (contiguous single-sequence FASTA files) are aligned using a modified version of the TBA program previously described (Blanchette et al. 2004). The TBA alignment tool generates MSLAs separated into several gapped alignment blocks. Each alignment block represents a multiple-sequence alignment consisting of a subset from 1 to

N subsequences (where N is the total number of input sequences). The orientation of subsequences is variable, but no sequence reshufflings are acceptable inside an alignment block and must be represented by separate alignment blocks. All alignment blocks are collected into a complex multiple-sequence local alignment with a single representation of each nucleotide from all species. If a particular subsequence is not reliably aligned to any other sequence, it will be represented in the alignment block by itself.

Phylogenetic tree guidance for TBA alignments

TBA dynamic programming alignment maximizes the total alignment score combined from the scores corresponding to each of the alignment blocks with no user-specified limitations on locations of the alignment blocks or the specification of sequences constructing a particular alignment block (e.g., one alignment block can include human–mouse–rat–*Fugu* and another one just human–*Fugu* if the rodent counterpart for this alignment block is missing). At the same time, short or numerous sequences limit the content information for the phylogenetic relationships between the input sequences, potentially biasing the TBA scheme for the optimal selection of partitioning the sequences into alignment blocks. A known phylogeny of the input sequences provided to the TBA aligner would greatly improve the reliability of the final alignments.

The phylogenetic relationship of the input sequences is essential; however, we do not require the user to manually input this information, as this could be a nontrivial task. Instead, Mulan predicts a phylogenetic tree describing the evolutionary history of the input species and just asks the user to verify the correctness of it prior to the final step of the TBA alignment. Phylogenetic tree prediction is generated using an intermediate limited multiple-sequence local alignment, which is produced by the *refine* program. The user has the option to change the structure of the automatically generated phylogenetic tree by altering its textual representation. No corrections were necessary while testing Mulan on several input examples with significantly diverged sequences.

Neighbor-joining method for phylogenetic tree construction

The neighbor-joining (NJ) method provides a computationally efficient approach for constructing phylogenetic trees from information on evolutionary sequence divergence (Saitou and Nei 1987). NJ very efficiently generates a topology of a phylogenetic tree, and calculates branch lengths by minimizing the total evolutionary change (the total length of the tree branches). We apply the NJ method to postprocessing *refine* multiple-sequence alignments generated at the intermediate stage of finished Mulan alignments or at the final stage of draft Mulan alignments. Starting with the matrix of pairwise distances between each pair of the sequences, our implementation of the NJ method generates a textual representation of the phylogenetic tree in the format acceptable by the TBA program. For example, human, mouse, rat, and *Fugu GATA3* locus comparison converted into the textual tree structure can be presented in the following New Hampshire-like format:

((human:12006.2 fugu:15716.8):908.1(mouse:3852.2 rat:3889.8): 908.1),

where the branching distances are in the number of single nucleotide mutations per kb. Mulan also generates a graphical representation of the phylogenetic tree (see Fig. 5A, for example). At the intermediate step of the optional manual curation of the phylogenetic relationships among the input species, the user is not required to indicate branching distances, but just to regroup the nodes by altering the textual representation of the phylogenetic tree.

Phylogenetic shadowing and summary conservation profiles

Phylogenetic shadowing is based on the assumption that closely related lineages (such as different primate or rodent phylogenetic clades) accumulate mutations independently from each other after the speciation event (Boffelli et al. 2003). By comparing several closely related sequences one can consider a nucleotide from the reference sequence to be diverged (or shaded) in the set of input sequences if this nucleotide does not match the same nucleotide from any other species included in the multiple-sequence alignment. The density of shaded nucleotides should be lower in the slow-mutating functional regions that are distinguished by the selection pressure applied to them.

Practical implementation of phylogenetic shadowing in Mulan is based on the differentiation of shaded and fully conserved nucleotides (that are exactly the same in all sequences in the alignment) and treating them as a set of simple matches and mismatches projected to the reference sequence (Ovcharenko et al. 2004a). A sliding window of 100 base pairs is used to scan through the array of shaded and fully conserved nucleotides, and to plot the percentage of fully conserved nucleotides on the vertical axis. After scanning all the positions in the reference sequence, a smooth-type conservation profile is created. ECRs are detected with the percent identity parameter used as a threshold for the percentage of fully conserved nucleotides in the sliding window. The user can adjust the length of the sliding window for the detection of ECRs that will effectively define the minimal length of an ECR.

The "summary conservation" option of the Mulan tool is very similar in implementation to the phylogenetic shadowing option, but differs in the underlying assumption and the produced graphical visualization profile. Instead of identifying fully conserved nucleotides, Mulan identifies nucleotides from the reference sequence that have matches with at least one other species. Basically, a nucleotide is called conserved in this method if it is conserved in any of the pairwise comparisons. (One can refer to the phylogenetic shadowing and "summary conservation" methods as AND and OR logical operators applied to a multiple-sequence alignment). Application of the "summary conservation" option will be beneficial in the cases of divergent degeneration and complementation of duplicated genes when different gene duplicates can display different data sets of gene regulatory elements (Prince and Pickett 2002).

Multiple-sequence conservation of transcription factor binding sites

Mulan utilizes the multiTF tool to identify transcription factor binding sites (TFBS) that are shared among all the sequences involved in the alignment. While multiTF is based on the same principle as rVista 2.0 (Loots et al. 2002; Loots and Ovcharenko 2004) (that postulates that evolutionary conservation can be a very efficient filter for exclusion of the majority of false-positive computational predictions of TFBSs), the method of detecting TFBSs shared among multiple species is different. rVista 2.0 works only with pairwise sequence alignments, possesses several requirements on TFBS core alignments and requires sites to be present in a short island of high sequence conservation. MultiTF does not rely on preferential local conservation of functional

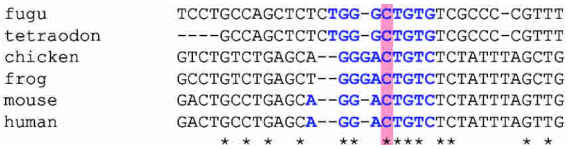

Figure 8. Schematic visualization of the multiTF method of identifying TFBSs shared by multiple species. Blue font color indicates a TFBS with the consensus sequence of [t/g/a]GG[g/a]CTGT[g/c] that would be detected by multiTF. Light-red shading highlights one of the anchor nucleotides for this binding site detection.

binding sites versus the neutrally evolving background as rVista does; instead it requires a binding site just to be present in all the species at the same position as dictated by the Mulan alignment.

In the first step, putative TFBSs are identified in all the original sequences by using TRANSFAC PWM matrices to define consensus sequences and the *tfSearch* utility to map consensus sequences of TFBSs to the genomic sequences of different species (Wingender et al. 1996; Loots and Ovcharenko 2004). Two approaches for the differential TFBS identification are available for the user at this stage—explicit selection of TFBS matrix similarity parameters (the TFBS matrix similarity parameter defines the level of identity required between a consensus sequence and the genomic sequence; Wingender et al. 1996) and the "optimized for function" method. Using the first approach, TFBS matrix similarity parameters can be fixed at the same level for all of the transcription factor families, which is ≤1.0. This usually results in extremely high levels of false-positive TFBS predictions for TF matrices with insufficient experimental evidence for the consensus sequence or for those that have relatively short binding sites (with a core of six base pairs or less), but ensures a uniform level of sequence similarity between the consensus sequence and detected TFBSs. The second approach, the "optimized for function" method, partially overcomes these problems by using optimized matrix similarity parameters. The optimization is performed independently for each of the TFBSs to limit the density of a TFBS in a random sequence by three or less sites per 10 kb. This approximately defines a probability to encounter a pair of TFBS in a 200-bp region to be less than 1e-2, effectively decreasing the number of false-positive predictions of TFBS clusters.

The second step excludes all of the TFBS predictions overlapping with coding exons. Obviously, gene annotation of only one of the sequences (e.g., the reference sequence) is sufficient at this step. The final step detects TFBS predictions that are shared by all of the species and are located at the same position as defined by the alignment. In order to do so we scan through all of the anchor or fully conserved nucleotides (nucleotides that are identical in all of the species in the multiple-sequence alignment; Fig. 8). If a TFBS from the reference sequence is found to overlap with an anchor nucleotide, we project this TFBS position to all of the other species by using the alignment and excluding gaps (Fig. 8). Starting and ending positions of the footprint of the reference sequence TFBS are compared to the starting and ending positions for the same TFBS on the same strand as detected by the initial TFBS annotation. If corresponding TFBS can be identified in all of the species in the alignment, this is reported by the multiTF.

List of options provided from the Mulan results Web page

In summary, upon generating an alignment, Mulan provides the user with the following list of options:

- Dynamic graphical visualization of conservation profiles
- Pairwise dot-plots as an overview of sequence rearrangements
- Dynamic batch detection of ECRs with varying ECR parameters
- Phylogenetic tree
- Portal to the multiTF tool for detection of cross-species TFBSs
- Dynamic modification of gene annotation

Acknowledgments

We thank Colleen Elso for her critical suggestions on the manuscript. W.M and R.H. were supported by NHGRI grant HG02238; G.G.L was supported by an LLNL LDRD-04-ERD-052 grant; I.O. was in part supported by a DOE SCW0345 grant. The work was performed under the auspices of the U.S. Department of Energy by the Univ. of California, Lawrence Livermore National Laboratory Contract #W-7405-Eng-48.

References

Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y., and De Moor, B. 2003. Toucan: Deciphering the *cis*-regulatory logic of coregulated genes. *Nucleic Acids Res.* **31:** 1753–1764.

Andl, T., Ahn, K., Kairo, A., Chu, E.Y., Wine-Lee, L., Reddy, S.T., Croft, N.J., Cebra-Thomas, J.A., Metzger, D., Chambon, P., et al. 2004. Epithelial Bmpr1a regulates differentiation and proliferation in postnatal hair follicles and is essential for tooth development. *Development* **131:** 2257–2268.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299:** 1391–1394.

Bray, N., Dubchak, I., and Pachter, L. 2003. AVID: A global alignment program. *Genome Res.* **13:** 97–102.

Brown III, C.O., Chi, X., Garcia-Gras, E., Shirai, M., Feng, X.H., and Schwartz, R.J. 2004. The cardiac determination factor, Nkx2-5, is activated by mutual cofactors GATA-4 and Smad1/4 via a novel upstream enhancer. *J. Biol. Chem.* **279:** 10659–10669.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. 2003. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13:** 721–731.

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

Frazer, K.A., Tao, H., Osoegawa, K., de Jong, P.J., Chen, X., Doherty, M.F., and Cox, D.R. 2004. Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res.* **14:** 367–372.

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

Giardine, B., Elnitski, L., Riemer, C., Makalowska, I., Schwartz, S., Miller, W., and Hardison, R.C. 2003. GALA, a database for genomic sequence alignments and annotations. *Genome Res.* **13:** 732–741.

Gilligan, P., Brenner, S., and Venkatesh, B. 2002. *Fugu* and human sequence comparison identifies novel human genes and conserved non-coding sequences. *Gene* **294:** 35–44.

Hardison, R.C., Chiaromonte, F., Kolbe, D., Wang, H., Petrykowska, H., Elnitski, L., Yang, S., Giardine, B., Zhang, Y., Riemer, C., et al. 2003. Global predictions and tests of erythroid regulatory regions. In *Genome of homo sapiens*, pp. 335–344. Cold Spring Harbor Press, Cold Spring Harbor, NY.

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

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res*. **12:** 996–1006.
Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci*. **100:** 11484–11489.
Kleinjan, D.A., Seawright, A., Childs, A.J., and van Heyningen, V. 2004. Conserved elements in Pax6 intron 7 involved in (auto)regulation and alternative transcription. *Dev. Biol*. **265:** 462–477.
Kolbe, D., Taylor, J., Elnitski, L., Eswara, P., Li, J., Miller, W., Hardison, R., and Chiaromonte, F. 2004. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res*. **14:** 700–707.
Lawoko-Kerali, G., Rivolta, M.N., and Holley, M. 2002. Expression of the transcription factors GATA3 and Pax2 during development of the mammalian inner ear. *J. Comp Neurol*. **442:** 378–391.
Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N., and Wasserman, W.W. 2003. Identification of conserved regulatory elements by comparative genome analysis. *J. Biol*. **2:** 13.
Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet*. **12:** 1725–1735.
Lim, K.C., Lakshmanan, G., Crawford, S.E., Gu, Y., Grosveld, F., and Engel, J.D. 2000. Gata3 loss leads to embryonic lethality due to noradrenaline deficiency of the sympathetic nervous system. *Nat. Genet*. **25:** 209–212.
Loots, G.G. and Ovcharenko, I. 2004. rVISTA 2.0: Evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*. **32:** W217–W221.
Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.
Loots, G.G., Ovcharenko, I., Pachter, L., Dubchak, I., and Rubin, E.M. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res*. **12:** 832–839.
Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. 2000. VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046–1047.
Nobrega, M.A., Ovcharenko, I., Afzal, V., and Rubin, E.M. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302:** 413.
Ovcharenko, I., Boffelli, D., and Loots, G.G. 2004a. eShadow: A tool for comparing closely related sequences. *Genome Res*. **14:** 1191–1198.
Ovcharenko, I., Loots, G.G., Hardison, R.C., Miller, W., and Stubbs, L. 2004b. zPicture: Dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res*. **14:** 472–477.
Pennacchio, L.A., Olivier, M., Hubacek, J.A., Cohen, J.C., Cox, D.R., Fruchart, J.C., Krauss, R.M., and Rubin, E.M. 2001. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294:** 169–173.
Prince, V.E. and Pickett, F.B. 2002. Splitting pairs: The diverging fates of duplicated genes. *Nat. Rev. Genet*. **3:** 827–837.
Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. **4:** 406–425.
Santagati, F., Abe, K., Schmidt, V., Schmitt-John, T., Suzuki, M., Yamamura, K., and Imai, K. 2003. Identification of *cis*-regulatory elements in the mouse Pax9/Nkx2–9 genomic region: Implication for evolutionary conserved synteny. *Genetics* **165:** 235–242.
Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. 2000. PipMaker—A web server for aligning two genomic DNA sequences. *Genome Res*. **10:** 577–586.
Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E.D., Hardison, R.C., and Miller, W. 2003a. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res*. **31:** 3518–3524.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003b. Human–mouse alignments with BLASTZ. *Genome Res*. **13:** 103–107.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22:** 4673–4680.
van Dam, H. and Castellazzi, M. 2001. Distinct roles of Jun : Fos and Jun : ATF dimers in oncogenesis. *Oncogene* **20:** 2453–2464.
Van Esch, H. and Bilous, R.W. 2001. GATA3 and kidney development: Why case reports are still important. *Nephrol. Dial. Transplant* **16:** 2130–2132.
Welch, J.J., Watts, J.A., Vakoc, C.R., Yao, Y., Wang, H., Hardison, R.C., Blobel, G.A., Chodosh, L.A., and Weiss, M.J. 2004. Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* (in press).
Wingender, E., Dietze, P., Karas, H., and Knuppel, R. 1996. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res*. **24:** 238–241.

Web site references

http://www.bx.psu.edu/miller_lab/; Source code for the aligners and the aligner-evaluation software.
http://globin.cse.psu.edu/gala/; GALA.
http://mulan.dcode.org; Mulan.
http://rvista.dcode.org/; rVista 2.0.
http://www.jgi.doe.gov/; Joint Genome Institute sequencing facility.
http://www.biobase.de/; TRANSFAC.

Received July 14, 2004; accepted in revised form August 31, 2004.