



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Mulan: Multiple-Sequence Local Alignment and Visualization for Studying Function and Evolution

I. Ovcharenko, G.G. Loots, B.M. Giardine, M. Hou, J.
Ma, R.C. Hardison, L.J. Stubbs, W. Miller

August 5, 2004

Genome Research

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Mulan: Multiple-Sequence Local Alignment and Visualization for Studying
Function and Evolution

UCRL: UCRL-JRNL-205739

IM: #309640

Mulan: Multiple-Sequence Local Alignment and Visualization for Studying Function and Evolution

Ivan Ovcharenko^{1,*}, Gabriela G. Loots², Belinda M. Giardine³, Minmei Hou⁴, Jian Ma⁴,
Ross C. Hardison³, Lisa Stubbs² and Webb Miller^{4,5}

¹Energy, Environment, Biology and Institutional Computing, Lawrence Livermore
National Laboratory, Livermore, CA 94550

²Genome Biology Division, Lawrence Livermore National Laboratory, Livermore, CA
94550

³Department of Biochemistry and Molecular Biology, The Pennsylvania State University,
University Park, PA 16802

⁴Department of Computer Science and Engineering, The Pennsylvania State University,
University Park, PA 16802

⁵Department of Biology, The Pennsylvania State University, University Park, PA 16802

*corresponding author

Phone: (925) 422-5035

Fax: (925) 422-2099

Email: ovcharenko1@llnl.gov

ABSTRACT

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 multi-species comparisons of the *GATA3* gene locus and the identification of elements that are conserved differently 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://bio.cse.psu.edu/>.

INTRODUCTION

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 (Gilligan et al. 2002; Pennacchio et al. 2001) and transcriptional regulatory elements (Elnitski et al. 2001; Loots et al. 2000). Several recent methods have emphasized the importance of multiple-sequence alignments (when 2 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; Ovcharenko et al. 2004b; Schwartz et al. 2000) or as a full multi-sequence global or pseudo-global alignment (Bray et al. 2003; Brudno et al. 2003; Ovcharenko et al. 2004a; Schwartz et al. 2003a; Thompson et al. 1994). 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 (MSLA) and present several examples for the application of this tool to study phenotypic differences in vertebrate species. *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 (ECR) 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 addition, *Mulan* is integrated with the *MultiTF* program that identifies evolutionarily conserved transcription factor binding sites (TFBS) 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 sequence analysis of multiple sequences that are present either as (1) draft or (2) finished configuration. First, the “draft” approach generates MSLAs, where a large fraction of sequences are broken up into separate contigs (such as draft multi-contig BAC sequences) and are compared to a reference sequence that is in “finished” format. This approach allows for the construction of an MSLA for multiple draft-quality sequences and subsequently for effective ordering-and-orientation (O&O) of unfinished sequences based on the reference genome. The second, “finished” approach operates with multiple high quality single-contig sequences, and is the main subject of this paper. This approach is restricted to one finished-quality sequence per species. Also, the “finished” approach is applicable to study multiple homologous (paralogous) regions from a single species, and in this case every sequence needs to be represented by a single contig.

Genomic sequences submitted to *Mulan* are aligned by the *tba* (threaded blockset aligner) program for “finished” sequences and by the *refine* program for “draft” sequences. A guiding phylogenetic tree is generated and requires user verification and confirmation for alignment visualization to proceed. The local alignment approach 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 a co-linearity of all input sequences (as in the case of a global multiple alignment), but instead in order to create textual alignment files it generates different projections of the “treaded block-set 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 one or more 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. Two different visualization schemes are implemented (1) pip-plot and (2) smooth-graph. 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 (Figure 1). Syntenic blocks are color-coded allowing for easy O&O of draft sequences by using the base sequence as a structural guide. The dot-plot option could also be instrumental for understanding the O&O structure of ‘draft’ sequences.

Visualization and data analysis strategies for multi-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 large number of sequences of varying lengths and different phylogenetic relationships, we provide a set of different visualization options applicable to any “finished” MSLA. The reference sequence selection is a dynamic process without any restrictions on the order of the input sequences or the species of origin, and can be interactively interchanged in the MSLA as desired. After indicating the base sequence from the list of sequences in the MSLA, the user selects the type of conservation profile

to be displayed. The standard option, ‘stacked pairwise’, is analogous to visualizing all pairwise alignments in a secondary *vs.* reference format used by the *zPicture* tool (Ovcharenko et al. 2004b). The stacking order is determined by the evolutionary relationship of each sequence to the reference sequence, with the most closely related species at the bottom and the most divergent comparison at the top.

‘Color density by interspecies conservation’ illustrates a relationship between the color-density of a conserved region and the number of species that share a particular region (Figure 2) such that, the more species share a region, the darker the conservation profile will be displayed. 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 (Figure 2). For example, while the pairwise human-mouse comparison for a 13kb long *GATA3* region identifies 34 ECRs (80bps/70% ECR parameters; Figure 2A), an 8-species scan of the multiple sequence conservation profile with a specified requirement that the human region is shared by at least 6 other species identifies the 4 most deeply conserved ECRs – 2 coding exons, an intronic and a promoter element located ~1kb upstream of the gene transcription start site (Figure 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 similarities from all the pairwise comparisons into a single conservation profile, the *phylogenetic shadowing* option effectively collects all the cumulative mismatches. The statistical post processing of the alignment data is an important part of the *phylogenetic shadowing* method and it has been previously described and implemented in the *eShadow* tool (<http://eshadow.dcode.org/>) (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 primate species (Figure 3A) and estimated the interspecies distances. ‘Stacked pairwise’ MSLA preferentially detects the human *ApoB* exon only in comparisons between the most distantly related species (Figure 3B) and does not allow dissecting this functional element selectively from the neutrally evolving background. In contrast, the *phylogenetic shadowing* visualization display accurately depicts the coding exon as the most highly conserved element in this region (Figure 3C), and 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 – two tools underlying the “draft” and “finished” *Mulan* alignment schema, we used a similar

approach to the one previously described (Blanchette et al. 2004). Fifty sets of simulation sequences are produced, each containing 9 species. Guided by a phylogenetic tree, the simulation program starts with an ancestral sequence, and accurately simulates evolutionary processes to produce current sequences. Besides mutations such as substitution, deletion, and insertion, we also simulate inversions. Overlapped inversions result in transpositions. The length distribution and frequency of each kind of rearrangement are determined from empirical studies of mammalian sequences (Blanchette et al. 2004; Siepel and Haussler 2004; Thomas et al. 2003), using values for inversion that we estimated. Duplication is not simulated at this stage. Note that all rates are observed from neutral DNA.

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 case 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. This agreement score shows how reliable the predicted alignment is, including aligned positions and unaligned positions. Both *refine* and *tba* start with pairwise alignments produced by *blastz* program.

To be consistent in comparing aligners, the same *blastz* parameters (C=0, Y=3400, K=2000) are used for all datasets. *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. It should be noted that the graph shown here is for illustrating the difference between aligners, not for tuning the parameters, and the parameters for *blastz* are not optimal. There are 36 pairwise alignments for 9 species. Only representative pairs are shown to illustrate how performance of an aligner varies with evolutionary distance.

Several observations can be made about this graph. First, for sequences at very short evolutionary distance, such as human vs. chimp and human vs. 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 vs mouse and human vs dog. However, for sequences being pulled together by *refine* instead of direct pairwise alignment, the performance is worse, e.g. rat vs. mouse and cat vs. 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 or better than *refine*. *Tba* out-performs *refine* dramatically for cat vs. dog and especially rat vs. 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 5 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 9 species each with length of around 50kb, *tba* takes around 50 seconds on a modern workstation, while *refine* requires only around 7 seconds.

From Sequence Evolution to Genome Biology

We applied *Mulan* to the study of the evolutionary conservation of the human *GATA3* locus. *GATA3* a very important molecule shown to be involved in various biological processes throughout development, both in the early embryo and in adulthood (Lawoko-Kerali et al. 2002; Lim et al. 2000; Van Esch and Bilous 2001). 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 evolutionarily clades, spanning over 450MY of evolutionary time since the separation of mammals, amphibians, birds and fish (Figure 5).

Multi-sequence *Mulan* alignment identified all the coding exons of the *GATA3* gene as conserved segments in all the species highlighting the functional importance of this protein and suggesting that interspecies differences associated with the *GATA3* protein can originate from differences in noncoding, not in coding sequences. This is supported by noncoding conservation patterns significantly differed in comparison of the human sequence with different species (Figure 5B). Three main groups of conservation were identified: human/rodents, amphibian/fish, and chicken. Five ECRs (ECR1...ECR5) are shared by at least 4 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 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 an absence of them in the frog and fish genomes is responsible for hairless bodies of 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 the function driven by this element (that could be a *GATA3* enhancement at a particular stage in a particular tissue) is absent in birds. Could it be that this is a silencer element that blocks *GATA3* involvement in wing bones development and its absence is correlated 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 could provide answers to these questions.

It is interesting also to mention that the local alignment nature of the *tba* aligner (that constructs the core of the *Mulan* tool) enables a correct recapitulation of the conservation profile of 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 (Figure 5C). The artificial 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.

Multi-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 (Aerts et al. 2003; Lenhard et al. 2003; Loots and Ovcharenko 2004; Loots et al. 2002). *Mulan* incorporates a TFBS analysis tool, *multiTF* that is similar to pairwise-alignment-based *rVista* 2.0 (Loots and Ovcharenko 2004; Loots et al. 2002), 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 with the exception of chickens (Figure 6). This analysis was aimed at providing *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 excluding chicken (Figure 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 the *multiTF* in the scan of ECR3 multiple-sequence alignment to be shared by all the species using almost 400 other TFBS matrices (Figure 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* prediction for gene regulation for a set of homologous sequences. Such computational pre-screens 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 have selected three genomic regions ranging in size from 150kb to 230kb and corresponding to *PAX6*, *NKX2.5* and *NKX2.9/PAX9* genomic loci. It has been shown that *PAX6* has auto-regulatory activity mediated through a *PAX6* TFBS located in an intron (Kleinjan DA 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 CO 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 the previously specified TFBS 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 to 97% of the total number of predictions for the human sequence alone. The addition of chicken sequences further reduced the number of predictions 5 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. This data suggests that by analyzing TFBS patterns in multiple sequence alignments one can dramatically filter our 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 analysis 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*) allows users to find genomic intervals that meet defined conservation thresholds, alignment-based scores, and 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, 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 identification of the homologous DNA intervals in other species. This is complicated by the existence of paralogs of many sequences, generated by transposition or segmental duplications, and by chromosomal rearrangements. Thus, a given DNA interval, say in human, may match to multiple locations in mouse. 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 displayed as 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 and al. in press 2004). The predicted cis-regulatory modules are verified at a high rate (Hardison et al. 2003; Welch and al. in press 2004).

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.

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, mostly due to the importance of this organism in agriculture and as a model for non-mammalian 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 multi-species sequence analysis. Here, we described the *Mulan* tool that introduces a novel reliable approach to generate MSLA. The tool is capable of producing fast and accurate alignments for both distantly and closely related organisms, such as humans, primates, fish and/or chicken, properly taking into account the complexity of evolutionary sequence rearrangements such as inversions, transpositions, and sub-sequence 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 can be an important tool for generating hypotheses about the function of ECRs shared by a limited number of species (Frazer et al. 2004). The speed, on the order of minutes, with which *Mulan* is capable of handling Megabase-long genomic sequences, and the dynamic character of the user interface is remarkable. All the graphical representations are re-plotted in real time after visualization parameters are customized by user. Also, interactive conservation profiles allow user-selection of an ECR which displays the multiple sequence alignment for that element.

We applied *Mulan* to the conservation analysis of several species for the genomic locus of the *GATA3* gene, to a 200kb region of chicken chromosome 28, and to an *ApoB* exon in primates. The “draft” alignment option of the *Mulan* tool allows easy O&O of chicken BAC contigs using the WGS assembly as the reference sequence. Comparison of the chicken *GATA3* locus sequence with the counterparts from other species including human, rodents, frog, and fish identifies differential distribution of putative *GATA3* regulatory elements in different lineages and permits speculation about species-specific phenotypic differences. 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 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 sub-sequence 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*. Single coverage option is used at this stage to filter out low scoring alignments that overlap 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 post-processing 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* (threaded blockset aligner) program that has been previously described (Blanchette et al. 2004). The *tba* alignment tool generates MSLA separated into several gapped alignment blocks. Each alignment block represents a multiple-sequence alignment consisting of a subset from 1 to N sub-sequences (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 have to be represented by separate alignment blocks. All alignment blocks are collected into a complex multi-sequence local alignment with a single representation of each nucleotide from all species. If a particular sub-sequence 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 (for example, 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, limited length of the input sequence contracts information on the phylogenetic relationships between the input sequences that could bias the *tba* scheme of the optimal selection of partitioning the sequences into alignment blocks. Increasing the number of input sequences significantly complicates the task of determination of correct phylogenetic relationships among all the sequences by significantly increasing the number of variable parameters underlying this problem. 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 non-trivial 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 generated by the *refine* program. The user has an option to change the structure of the automatically generated phylogenetic tree by altering its textual representation in the case of an incorrect prediction. 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 evolutionarily 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 post-process *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 format:

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

where the optional numbers indicate branching distances from node to node as the number of single nucleotide mutations per kb of the sequence (this number could be converted to the branching distance in MYs if the divergence rate is known). It is postulated that every internal or top node of the phylogenetic tree branches to exactly two other nodes. Parentheses of the textual tree representation group represent a branching of an internal or top node. *Mulan* also generates graphical representation of the phylogenetic tree (see Figure 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 the 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 the *Mulan* is based on 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 100bps is utilized to scan through the array of *shaded* and fully conserved nucleotides and to plot the percentage of fully conserved nucleotides as a vertical coordinate. 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 multi-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 datasets of gene regulatory elements (Prince and Pickett 2002).

Multi-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 and Ovcharenko 2004; Loots et al. 2002) (that postulates that evolutionary conservation can be a very efficient filter for exclusion of the majority of false positive computational predictions of TFBS), the method of detection of TFBS that are shared among multiple species is different. *rVista 2.0*, which works with only pairwise sequence alignments, possesses several requirements on TFBS core alignments and requires the site to be present in a short island

of high sequence conservation. *MultiTF* does not rely on preferential local conservation of functional binding sites vs 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 TFBS 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 TFBS to the genomic sequences of different species (Loots and Ovcharenko 2004; Wingender et al. 1996). 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)) or ‘optimized for function’ method. Using the first approach, TFBS matrix similarity parameters can be fixed at the same level for all the transcription factor families, which is less or equal to 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 definition or for those that have relatively short binding sites (with a core of 6bps or less), but ensures a uniform level of sequence similarity between the consensus sequence and detected TFBS. The second approach, partially overcomes the problems associated with insufficiently well-defined PWMs and short binding sites, 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 3 or less sites per 10kb. This approximately defines a probability to encounter a pair of TFBS in 200bps region to be less than $1e^{-2}$, effectively decreasing the

number of false positive predictions of TFBS clusters and is called as ‘optimized for function’ method.

The second step excludes all the TFBS predictions overlapping with coding exons. Obviously, gene annotation of only one of the sequences (the reference sequence, for example) is sufficient at this step. The final step detects TFBS predictions that are shared by all the species and are located at the same position as defined by the alignment. In order to do so we scan through all the ‘anchor’ or fully conserved nucleotides (nucleotides that are identical in all the species in the multiple-sequence alignment; Figure 8). If a TFBS from the reference sequence is found to overlap with an “anchor” nucleotide we project this TFBS position to all the other species by using the alignment and excluding gaps (Figure 8). Starting and ending positions of the footprint of the reference sequence TFBS are compared to the starting and ending position for the same TFBS on the same strand as detected by the initial TFBS annotation. If corresponding TFBS can be identified in all 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 either a “draft” or “finished” multiple-sequence alignment, Mulan provides the user with the following list of options for sequence analysis and data download:

- Dynamic graphical visualization of conservation profiles with a number of options for data presentation.
- Pairwise dot-plots displaying alignment blocks and overview of sequence rearrangements.

- Dynamic batch detection of ECRs by varying ECR parameters for each of the pairwise alignments.
- *Refine* FASTA alignment file.
- Phylogenetic tree.
- Portal to the *multiTF* tool for detection of cross-species TFBS. (Available for “finished” *Mulan* alignments.)
- Opportunity to dynamically modify gene annotation.
- Download of input sequences in the original format and with repetitive elements masked by ‘N’ symbols as well as an annotation of types of repetitive elements.

ACKNOWLEDGMENTS

We are grateful to Collen 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 LLNL LDRD-04-ERD-052 grant; I.O. was in part supported by DOE SCW0345 grant. The work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory Contract No. W-7405-Eng-48.

WEB SITE REFERENCES

<http://globin.cse.psu.edu/gala/>; GALA

<http://genome.ucsc.edu/>; UCSC Genome Browser

<http://mulan.dcode.org/>; *Mulan*

<http://zpicture.dcode.org/>; zPicture

<http://rvista.dcode.org/>; rVista 2.0

<http://eshadow.dcode.org/>; eShadow

Table 1.

| | 230kb PAX6 | 155kb NKX2.5 | 150kb 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

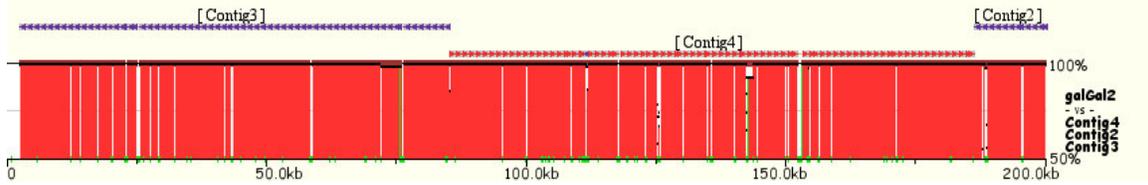


Figure 1. *Mulan* contig ordering based on homology to the reference sequence. The top layer of shaded lines indicates the location of contigs from a second sequence aligned to the base sequence where red, right-turned triangles specify forward strand alignments, and purple, left-turned triangles correspond to reverse strand alignments. Contig names are indicated in square brackets. The JF2-73M16 chicken BAC clone (<http://www.jgi.doe.gov/>) consisting of 3 contigs was aligned to the chicken genome (chr28:4,000,000-4,200,000).

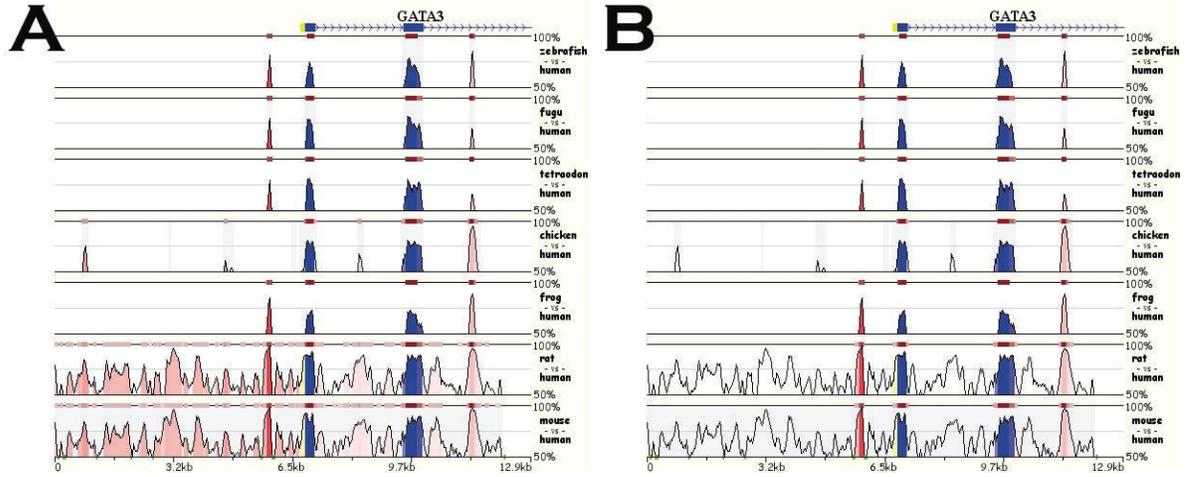


Figure 2. Stacked-pairwise conservation profile for 13kb region from the *GATA3* locus.

Color-gradient visualization is implemented to differentially display regions that are differently 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 6 out of 7 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.

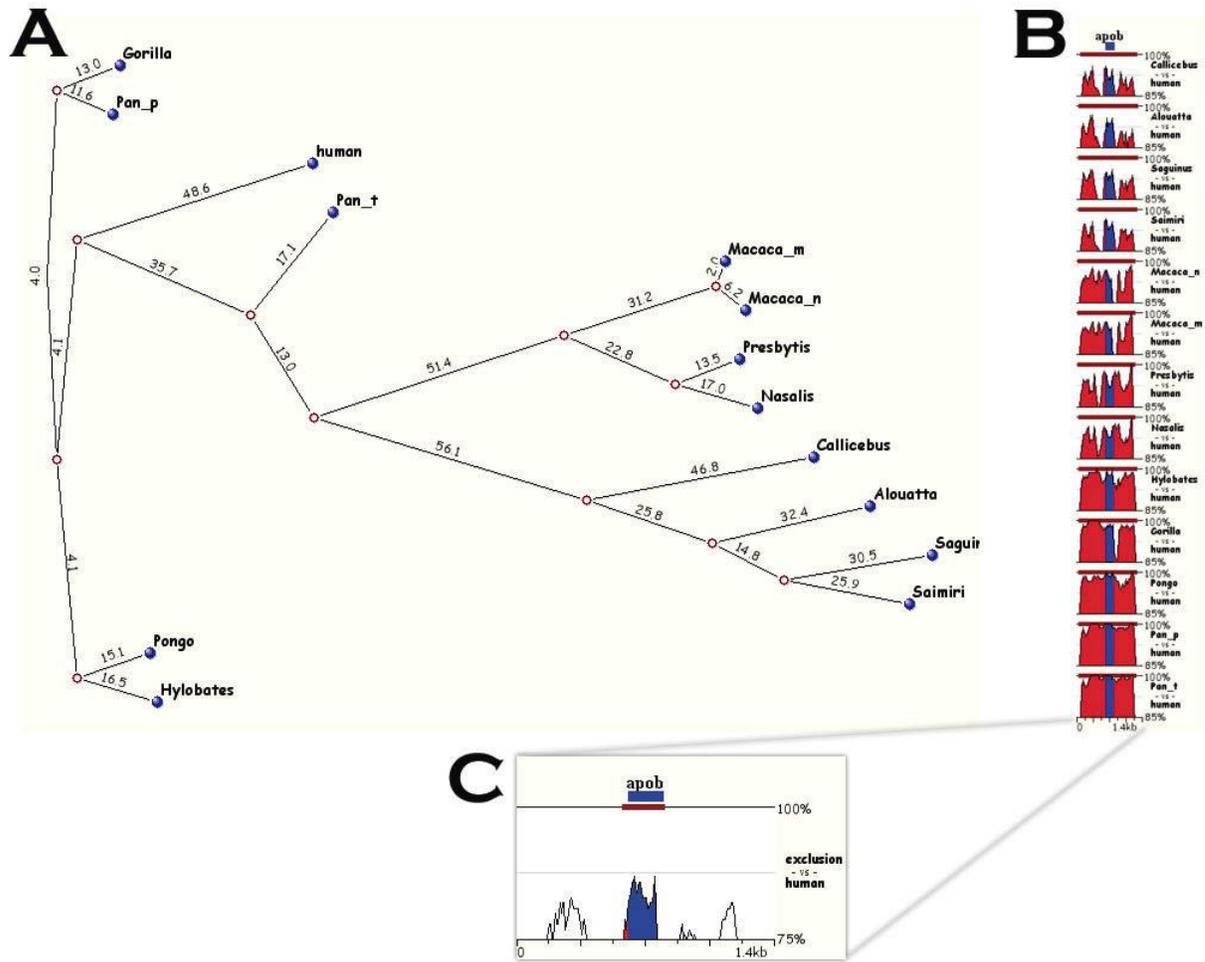


Figure 3. *Phylogenetic shadowing* option of the *Mulan* tool. *ApoB* region sequences from 14 primates were compared and phylogenetic relationships (A) and ‘stacked pairwise’ comparisons with the human reference sequence displayed (B). The *phylogenetic shadowing* conservation profile preferentially detects the *ApoB* coding exon from the neutrally evolving background (C). ECR parameters used for detecting exons: >85% identity; >100bp.

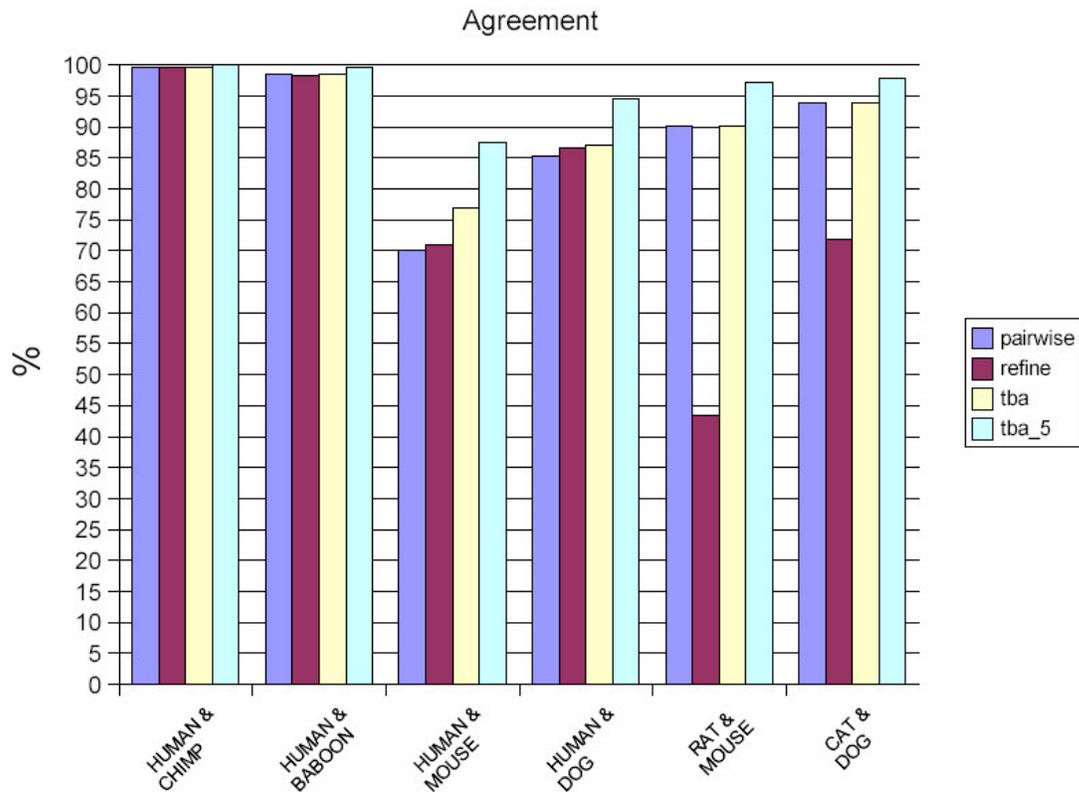


Figure 4. Agreement score of pairwise and multiple alignments produced by different aligners on a set of nine simulated mammalian sequences of length around 50kb. Pairwise results from *blastz* were post-processed to remove overlapped regions. Multiple aligners including *refine* and *tba* use the same pairwise alignments. *tba_5* refers to alignments from *tba*, but the agreement score allows mismatches within 5 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.

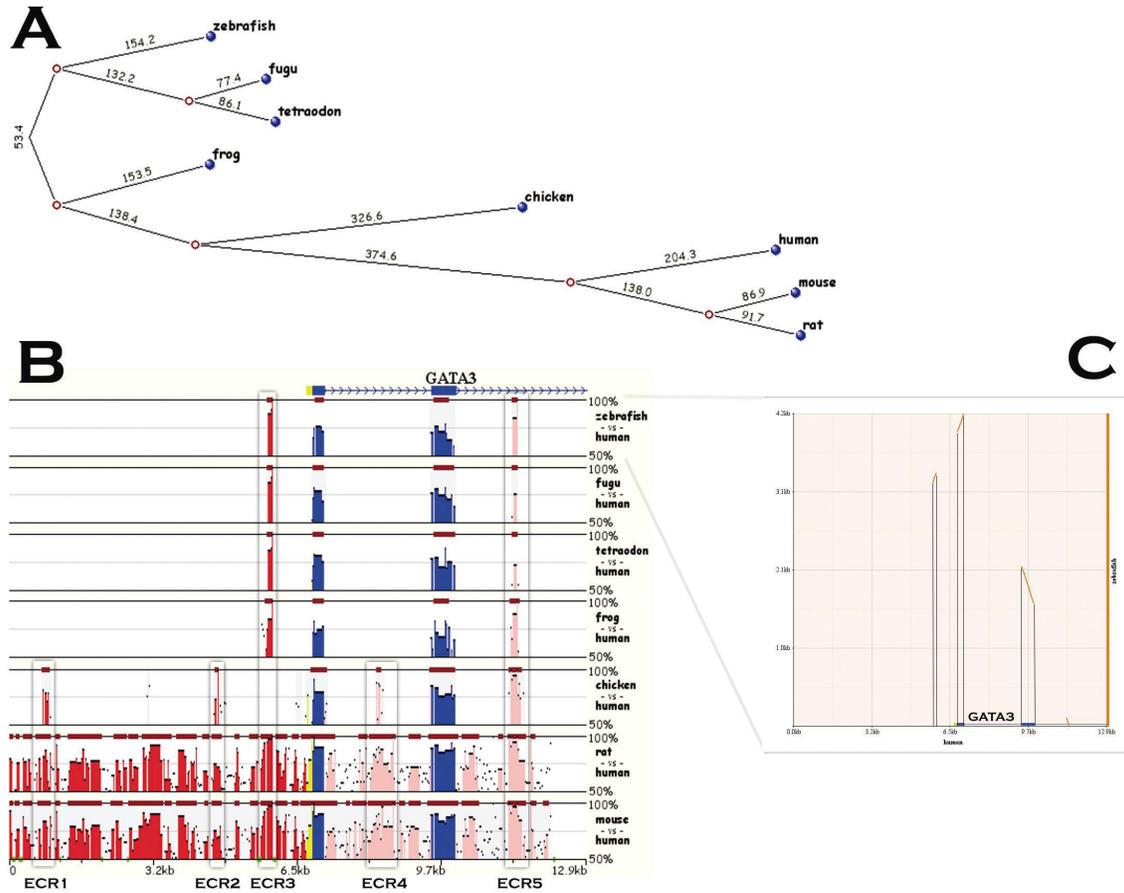


Figure 5. *Mulan* phylogenetic tree (A) and sequence conservation profile (B) for the *GATA3* gene locus from human, rat, mouse, chicken, frog, and three fish genomes. Each tree branch indicates the number of nucleotide substitutions from the closest node. Noncoding ECRs conserved (>70% identity;>80 bps) in at least 4 species (including human) are shaded and numbered as 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).

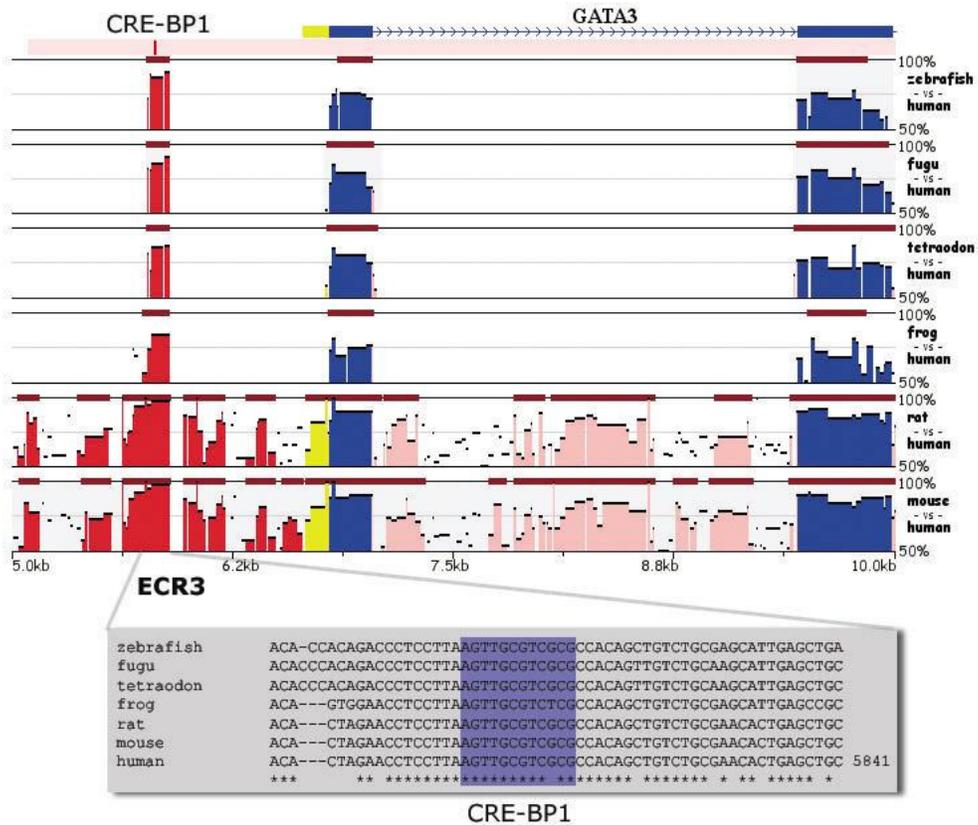


Figure 6. *multiTF* visualization of CRE-BP1 transcription factor binding site detected in the *GATA3* locus overlaid with the conservation profile of this locus as constructed with human, mouse, rat, frog, fugu, tetraodon, and zebrafish sequences. The bottom panel represents a 60 bps long alignment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

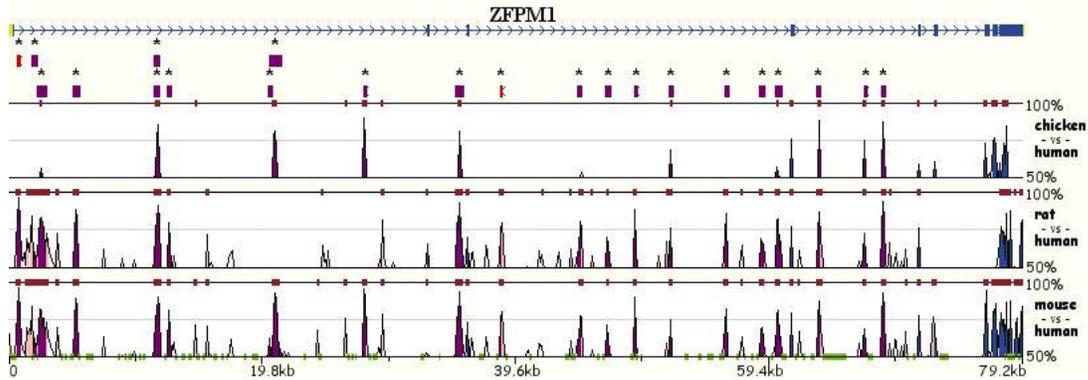


Figure 7. Conservation of *ZFPM1* among human, mouse, rat and mouse, using *tba* at the Mulan server. The large introns have several highly conserved regions. Those with conserved GATA-1 binding sites and high regulatory potential (predicted CRMs) are indicated as a set of purple and red blocks under the gene demarcated by apostrophe 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 and al. in press 2004).

| | |
|-----------|--------------------------------------|
| fugu | TCCTGCCAGCTCTCTGG-GCTGTGTCGCCC-CGTTT |
| tetraodon | ----GCCAGCTCTCTGG-GCTGTGTCGCCC-CGTTT |
| chicken | GTCTGTCTGAGCA--GGGACTGTCTCTATTAGCTG |
| frog | GCCTGTCTGAGCT--GGGACTGTCTCTATTAGCTG |
| mouse | GACTGCCTGAGCA--GG-ACTGTCTCTATTAGTTG |
| human | GACTGCCTGAGCA--GG-ACTGTCTCTATTAGTTG |
| | * * * * * ** * * * * * * * * * * * * |

Figure 8. Schematic visualization of the *multiTF* scheme of identification of transcription factor binding sites that are shared by multiple species. Blue font color indicates a transcription factor binding site 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.

d

REFERENCES

- Aerts, S., G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. De Moor. 2003. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res* **31**: 1753-1764.
- Andl, T., K. Ahn, A. Kairo, E.Y. Chu, L. Wine-Lee, S.T. Reddy, N.J. Croft, J.A. Cebra-Thomas, D. Metzger, P. Chambon, K.M. Lyons, Y. Mishina, J.T. Seykora, I.E. Crenshaw, and S.E. Millar. 2004. Epithelial *Bmpr1a* regulates differentiation and proliferation in postnatal hair follicles and is essential for tooth development. *Development*.
- Blanchette, M., W.J. Kent, C. Riemer, L. Elnitski, A.F. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green, D. Haussler, and W. Miller. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**: 708-715.
- Boffelli, D., J. McAuliffe, D. Ovcharenko, K.D. Lewis, I. Ovcharenko, L. Pachter, and E.M. Rubin. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**: 1391-1394.
- Bray, N., I. Dubchak, and L. Pachter. 2003. AVID: A global alignment program. *Genome Res* **13**: 97-102.
- Brudno, M., C.B. Do, G.M. Cooper, M.F. Kim, E. Davydov, E.D. Green, A. Sidow, and S. Batzoglou. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**: 721-731.
- Elnitski, L., J. Li, C.T. Noguchi, W. Miller, and R. Hardison. 2001. A negative cis-element regulates the level of enhancement by hypersensitive site 2 of the beta-globin locus control region. *J Biol Chem* **276**: 6289-6298.
- Frazer, K.A., H. Tao, K. Osoegawa, P.J. de Jong, X. Chen, M.F. Doherty, and D.R. Cox. 2004. Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res* **14**: 367-372.
- Ghanem, N., O. Jarinova, A. Amores, Q. Long, G. Hatch, B.K. Park, J.L. Rubenstein, and M. Ekker. 2003. Regulatory roles of conserved intergenic domains in vertebrate *Dlx* bigene clusters. *Genome Res* **13**: 533-543.
- Giardine, B., L. Elnitski, C. Riemer, I. Makalowska, S. Schwartz, W. Miller, and R.C. Hardison. 2003. GALA, a database for genomic sequence alignments and annotations. *Genome Res* **13**: 732-741.
- Gilligan, P., S. Brenner, and B. Venkatesh. 2002. Fugu and human sequence comparison identifies novel human genes and conserved non-coding sequences. *Gene* **294**: 35-44.
- Hardison, R.C., F. Chiaromonte, D. Kolbe, H. Wang, H. Petrykowska, L. Elnitski, S. Yang, B. Giardine, Y. Zhang, C. Riemer, S. Schwartz, D. Haussler, K.M. Roskin, R.J. Weber, M. Diekhans, W.J. Kent, M.J. Weiss, J. Welch, and W. Miller. 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., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31**: 51-54.

- Kent, W.J., R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* **100**: 11484-11489.
- Kent, W.J., C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, and D. Haussler. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996-1006.
- Kolbe, D., J. Taylor, L. Elnitski, P. Eswara, J. Li, W. Miller, R. Hardison, and F. Chiaromonte. 2004. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res* **14**: 700-707.
- Lawoko-Kerali, G., M.N. Rivolta, and M. Holley. 2002. Expression of the transcription factors GATA3 and Pax2 during development of the mammalian inner ear. *J Comp Neurol* **442**: 378-391.
- Lenhard, B., A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg, and W.W. Wasserman. 2003. Identification of conserved regulatory elements by comparative genome analysis. *J Biol* **2**: 13.
- Lettice, L.A., S.J. Heaney, L.A. Purdie, L. Li, P. de Beer, B.A. Oostra, D. Goode, G. Elgar, R.E. Hill, and E. de Graaff. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**: 1725-1735.
- Lim, K.C., G. Lakshmanan, S.E. Crawford, Y. Gu, F. Grosveld, and J.D. Engel. 2000. Gata3 loss leads to embryonic lethality due to noradrenaline deficiency of the sympathetic nervous system. *Nat Genet* **25**: 209-212.
- Loots, G.G., R.M. Locksley, C.M. Blankespoor, Z.E. Wang, W. Miller, E.M. Rubin, and K.A. Frazer. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**: 136-140.
- Loots, G.G. and I. Ovcharenko. 2004. rVISTA 2.0: Evolutionary Analysis of Transcription Factor Binding Sites. *Nucleic Acids Res.*
- Loots, G.G., I. Ovcharenko, L. Pachter, I. Dubchak, and E.M. Rubin. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res* **12**: 832-839.
- Mayor, C., M. Brudno, J.R. Schwartz, A. Poliakov, E.M. Rubin, K.A. Frazer, L.S. Pachter, and I. Dubchak. 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**: 1046-1047.
- Nobrega, M.A., I. Ovcharenko, V. Afzal, and E.M. Rubin. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302**: 413.
- Ovcharenko, I., D. Boffelli, and G.G. Loots. 2004a. eShadow: a tool for comparing closely related sequences. *Genome Res* **14**: 1191-1198.
- Ovcharenko, I., G.G. Loots, R.C. Hardison, W. Miller, and L. Stubbs. 2004b. zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res* **14**: 472-477.
- Pennacchio, L.A., M. Olivier, J.A. Hubacek, J.C. Cohen, D.R. Cox, J.C. Fruchart, R.M. Krauss, and E.M. Rubin. 2001. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294**: 169-173.
- Prince, V.E. and F.B. Pickett. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet* **3**: 827-837.

- Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406-425.
- Santagati, F., K. Abe, V. Schmidt, T. Schmitt-John, M. Suzuki, K. Yamamura, and K. Imai. 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., L. Elnitski, M. Li, M. Weirauch, C. Riemer, A. Smit, E.D. Green, R.C. Hardison, and W. Miller. 2003a. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res* **31**: 3518-3524.
- Schwartz, S., W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler, and W. Miller. 2003b. Human-mouse alignments with BLASTZ. *Genome Res* **13**: 103-107.
- Schwartz, S., Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. 2000. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* **10**: 577-586.
- Siepel, A. and D. Haussler. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* **21**: 468-488.
- Thomas, J.W., J.W. Touchman, R.W. Blakesley, G.G. Bouffard, S.M. Beckstrom-Sternberg, E.H. Margulies, M. Blanchette, A.C. Siepel, P.J. Thomas, J.C. McDowell, B. Maskeri, N.F. Hansen, M.S. Schwartz, R.J. Weber, W.J. Kent, D. Karolchik, T.C. Bruen, R. Bevan, D.J. Cutler, S. Schwartz, L. Elnitski, J.R. Idol, A.B. Prasad, S.Q. Lee-Lin, V.V. Maduro, T.J. Summers, M.E. Portnoy, N.L. Dietrich, N. Akhter, K. Ayele, B. Benjamin, K. Cariaga, C.P. Brinkley, S.Y. Brooks, S. Granite, X. Guan, J. Gupta, P. Haghghi, S.L. Ho, M.C. Huang, E. Karlins, P.L. Laric, R. Legaspi, M.J. Lim, Q.L. Maduro, C.A. Masiello, S.D. Mastrian, J.C. McCloskey, R. Pearson, S. Stantripop, E.E. Tiongson, J.T. Tran, C. Tsurgeon, J.L. Vogt, M.A. Walker, K.D. Wetherby, L.S. Wiggins, A.C. Young, L.H. Zhang, K. Osoegawa, B. Zhu, B. Zhao, C.L. Shu, P.J. De Jong, C.E. Lawrence, A.F. Smit, A. Chakravarti, D. Haussler, P. Green, W. Miller, and E.D. Green. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**: 788-793.
- Thompson, J.D., D.G. Higgins, and T.J. Gibson. 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 M. Castellazzi. 2001. Distinct roles of Jun : Fos and Jun : ATF dimers in oncogenesis. *Oncogene* **20**: 2453-2464.
- Van Esch, H. and R.W. Bilous. 2001. GATA3 and kidney development: why case reports are still important. *Nephrol Dial Transplant* **16**: 2130-2132.
- Welch, J. and e. al. in press 2004.
- Wingender, E., P. Dietze, H. Karas, and R. Knuppel. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**: 238-241.