Opinion

Multi-species sequence comparison: the next frontier in genome annotation

Inna Dubchak* and Kelly Frazer†

Addresses: *Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. †Perlegen Sciences, 2021 Stierlin Ct., Mountain View, CA 94043, USA.

Correspondence: Inna Dubchak. E-mail: ildubchak@lbl.gov

Published: 27 November 2003

Genome ***Biology*** 2003, **4:**122

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/12/122

Abstract

Multi-species comparisons of DNA sequences are more powerful for discovering functional sequences than pairwise DNA sequence comparisons. Most current computational tools have been designed for pairwise comparisons, and efficient extension of these tools to multiple species will require knowledge of the ideal evolutionary distance to choose and the development of new algorithms for alignment, analysis of conservation, and visualization of results.

Comparison of DNA sequences from different species is an extremely efficient way to identify functional DNA elements - both coding regions and transcriptional control regions that lie beyond the coding sequences of genes. Several recent reviews of comparative sequence analysis [1-5] describe this fast-growing field and the computational resources that are currently available for a wide range of biological investigations. Most of the large-scale comparative studies completed to date have been based on pairwise comparison of sequences; such studies have resulted in the identification of new genes [6-9], and have proved efficient at discovering functional elements in non-coding genomic intervals [10-12]. Several groups have aligned the entire human and mouse genomes [13-15] and have presented comprehensive statistical data on the patterns of DNA conservation between the two species.

Recent comparative studies demonstrate that adding additional species to the analysis provides an even more powerful approach for detecting functionally important elements, because characteristic signatures - such as open reading frames and splice-site consensus sequences within genes, and motifs within regulatory elements - are easier to detect when they are conserved in multiple species [16]. For example, a recent large-scale study of over 12 megabases (Mb) of sequences from 12 species, derived from genomic regions orthologous to a 1.8 Mb region on human chromosome 7 that contains ten genes [17], demonstrated that some highly conserved elements revealed by multiple sequence alignments could not be reliably identified with any set of parameters in a pairwise human-mouse alignment.

As the number of available complete genome sequences increases, there is a clear need to understand what we can learn from multiple-species sequence comparisons. Studies of this type will require the development of new comparative algorithms and computational tools, such as multi-genome alignment techniques, analysis of conservation and visualization of comparative results. Developing easy-to-use and efficient techniques is not trivial, however, given that the algorithms should be capable of handling a whole range of evolutionary distances between multiple species and of providing new insights into biology.

Selecting multiple species for comparative analysis

The comparison of DNA sequences between evolutionarily distantly related species, such as humans and pufferfish, which diverged approximately 450 million years ago,

primarily identifies the coding sequences as conserved [18] - because transcribed protein coding sequences are highly functionally constrained, and thus change very slowly during evolution. The comparison of DNA sequences between species that diverged from a common ancestor around 40-80 million years ago - such as humans and mice, or two species of fruitflies (*Drosophila melanogaster* and *Drosophila pseudoobscura*), or two species of nematodes (*Caenorhabditis elegans* and *Caenorhabditis briggsae*) [19] - results in identifying as evolutionarily conserved both coding sequences and a significant number of noncoding sequences. Only a limited number of conserved noncoding sequences that have been identified through sequence comparisons between species at this evolutionary distance have been characterized functionally, however. Among those that have had their functions assigned are transcriptional regulatory elements of genes in close proximity [11] or genes as far away from each other as 200 kb [10]. Comparative analyses of genomic DNA from closely related species, such as humans and chimpanzees, on the other hand, identifies those sequences that have changed in recent evolutionary history [20,21]. Some of these sequence changes may have been partly responsible for the speciation of ancestral primates. Thus, comparison of a segment of DNA with the sequences of multiple species at different evolutionary distances allows one to identify coding sequences, conserved noncoding elements with regulatory functions, and those sequences that are unique for a given species. A recent report by Cooper *et al.* [22] proposed a method for quantitatively assessing the effectiveness of a comparative sequence analysis to identify new information in a genome: it uses the 'phylogenetic scope', representing the range of organisms that share a last common ancestor whose sequence can be inferred by adding each genome to the analysis. The comparative studies described below demonstrate that the evolutionary distance of the species in a sequence comparison analysis is critical for discovering potentially functional sequence elements.

The stem cell leukemia genomic interval

For many years the mouse and human genomes, which diverged from a common ancestor about 65-75 million years ago, have been extensively used for comparative studies [9-12]; but it is still an open question as to which species should be added to this comparative analysis to derive the most information content. Among several recent studies providing guidance for selecting additional species is the investigation of the stem cell leukemia (*SCL*) genomic interval, originally based on a human-mouse sequence comparison [23], and later expanded to include three additional species: chicken, pufferfish and zebrafish [12]. This analysis demonstrates that mouse-human alignments show high levels of sequence similarity for all coding exons and for all eight known murine regulatory regions of the *SCL* locus. Human-mouse-chicken alignments identified the similarity of all coding exons and also discovered protein-binding motifs in five of the known regulatory regions. Thus, inclusion of the chicken DNA sequences allowed for superior functional annotation of a subset of the regulatory regions that had already been identified by the human-mouse comparison.

Pairwise mouse-pufferfish and mouse-zebrafish sequence alignments identified only some of the coding exons, and found similarity for only two of the eight known regulatory regions in the pufferfish comparison; and no significant similarity was found for any known regulatory region in the zebrafish comparison [12]. This analysis suggests that comparative analysis of zebrafish and mammalian genomic sequences might be of limited value for the identification of functionally significant noncoding sequences in the *SCL* region; and these results are consistent with what is expected on the basis of the evolutionary distance of the species analyzed.

Drosophila melanogaster compared with other species

The analysis of conservation between *Drosophila melanogaster* and four other *Drosophila* species (*D. erecta, D. pseudoobscura, D. willistoni*, and *D. littoralis*) that have different divergence times (6-15, 46, 53 and 61-65 million years, respectively) [24] has generated several important conclusions to guide further functional studies of these species [25]. One conclusion is that the addition of a third species could reveal functional constraints in otherwise non-significant pairwise exon comparisons. All *D. melanogaster* genes identified in divergent species show evidence of functional constraint; and including more distantly related species defines the exact position of short regulatory elements that are hard to find in the long regions of non-coding sequence conservation observed in closely related species. Non-coding conserved sequences have also been found to be spatially clustered, and these clusters can be used to predict enhancer sequences [25]. This work provided a solid basis for choosing species whose genome sequences would be most useful in aiding the functional annotation of coding and *cis*-regulatory sequences in *D. melanogaster: D. pseudoobscura,* which has recently been sequenced, was recognized as the best for discovery of functional genomic features, and adding *D. willistoni* to the comparison allows the dissection of regions of the *Drosophila* genome under different levels of functional constraint.

Saccharomyces cerevisiae

Using multiple alignments in the *S. cerevisiae* and related fungal genome annotation projects [26] provided a powerful demonstration of functional analysis, yielding results that would be difficult to obtain by other computational and experimental methods. A comprehensive comparison of the genome of the yeast *S. cerevisiae* with those of three related *Saccharomyces* species (*S. paradoxus, S. mikatae* and *S. bayanus*) [26] yielded a major revision to the yeast gene catalog, reducing the total count by about 500 genes. In addition, motif analysis automatically identified a

number of genome-wide elements, including most known regulatory motifs and numerous new motifs suitable for biological study.

Multiple primate analysis

Another approach to multiple species sequence analysis, 'phylogenetic shadowing' [21], is used for comparison of evolutionarily closely related species. It demonstrates the utility of sequence comparisons within the primate group for discovering common mammalian, as well as primate-specific, functional elements in the human genome, which could not be achieved by comparison of more evolutionarily distant species. Rubin and colleagues [21] showed that the high information content of comprehensive primate sequence comparisons could be captured with a small subset of phylogenetically close primates, such that sequence from as few as four or six primate species compared with humans might be sufficient for the identification of a large fraction of functional elements in the human genome, many of which are likely to be missed by human-mouse comparisons. While the number of multi-species comparative studies grows, it is becoming clear that reasonable selection of species for comparison of a particular genomic interval is still to a large extent an intuitive process, with some guidance from previous successful comparative studies.

Multi-species sequence alignment and analysis of conservation

As well as selecting a set of species that provide maximum functional content, the quality of the sequence alignment must also be sufficient to the task in hand. Single pairwise comparisons of sequences do not allow for the detection of conserved sequence strings with high precision, given that functional elements - such as transcriptional-regulator binding sites - are quite short compared to the surrounding nonfunctional sequence. Thus, functional signals can sometimes be indistinguishable from the 'noise' that results from aligning divergent nonfunctional sequences. The hope is that multiple sequence alignment provides a way to increase the sensitivity of the search for regulatory signals.

The area of sequence alignment is well developed, but many of its problems are far from being completely resolved, especially for multiple species [27]. Alignment methods can be roughly divided into local alignments, which produce optimal similarity scores between subregions of the two sequences, and global alignments, which generate optimal similarity scores over the entire length of the two sequences. Global alignments attempt to find a monotonically increasing map between the letters of each sequence, in the process rejecting alignments that overlap or cross over. A recently published review on comparative genomics gives more details of the various kinds of alignment [1]. Unfortunately, a comprehensive study of the strengths and weaknesses of different alignments algorithms applied to different biological problems has yet to appear.

The local and global alignment methods that generate pairwise comparisons can also be used for multiple species, but multiple alignments are considerably more difficult to compute because of statistical complexity and the difficulties of scoring the results. Progressive multiple alignment is a heuristic technique that uses successive applications of a pairwise alignment algorithm. The best-known progressive alignment program, CLULSTALW [28], is very efficient in aligning proteins and short nucleotide sequences, but it is not suitable for long genomic regions [29]. Below we describe new alignment techniques that can handle long DNA sequences efficiently.

Global alignments

Two recently developed algorithms, MLAGAN [16,30] and MAVID [31,32], are designed for global alignment of both evolutionarily close and distant megabase-length genomic sequences. The MLAGAN [16,30] algorithm assumes that the phylogenetic tree is known, as is usually the case for large vertebrate genomes. The program is based on progressive alignment: a multiple alignment of K sequences is constructed in $K-1$ pairwise alignment steps, such that in each step two sequences, or intermediate multiple alignments, are aligned.

MLAGAN uses LAGAN as the global pairwise-alignment subroutine, and introduces new methods for scoring and refining a multiple alignment. It also aligns the sequences in the order of the given phylogenetic tree. For example, MLAGAN aligns sequences from human, chimpanzee, mouse, rat, and chicken, in the following order: first, human-chimpanzee; second, mouse-rat; third, human-chimpanzee to mouse-rat; fourth, human-chimpanzee-mouse-rat to chicken. Each alignment step merges two sequences or alignments into a larger alignment, effectively building a profile of all the sequences. The results obtained with MLAGAN on the cystic fibrosis (*CFTR*) genomic region [16], suggest that multiple alignments are better than pairwise alignments at aligning conserved exons between distant species: it was precise enough to refine mis-annotated splicing sites.

MAVID [31,32] is a progressive global alignment program that works by recursively aligning the 'alignments' at ancestral nodes of the guide phylogenetic tree. At each internal node, ancestral sequences are inferred from the existing alignments using maximum likelihood, and these alignments are then aligned using the global aligner AVID [33]. The multiple alignment is used to build a phylogenetic tree for the sequences, which is subsequently used as a basis for identifying conserved regions in the alignment.

Local alignments

MultiPipMaker [34,35] uses multiple pairwise local alignments of secondary sequences against a reference sequence

to create a crude multiple alignment that is subsequently refined to generate a true multiple alignment. Analysis of multiple alignments generated by MultiPipMaker [35] allowed for discovery of regulatory elements in the mammalian *WNT2* genomic region, and confirmed the phylogenetic inference that horses are evolutionarily more closely related to cats than to cows [17]. Alignments between the human sequence and the sequence of each of the other 12 species used in the analysis [17] showed, as expected, that the fraction of sequence that can be aligned generally decreases with increasing evolutionary distance from humans (except for mouse and rat).

Another program, Multiz, developed for large-scale comparison of multiple sequences, takes BLASTZ/axtBest [35] as the pairwise input. This program has been used for the alignment of the mouse and the rat draft assemblies to the human genome [36].

Motif finding

'Phylogenetic footprinting' [37] aims to discover specific protein-binding sites within regulatory regions of multiple sequences on the basis of phylogenetic relationships. It is a method that is mostly applied to promoter regions of orthologous genes. Sumiyama with coauthors [38] attained good results by using multiple sequence comparison combined with a small window size (where the window is the region analyzed in each sub-comparison). This high-resolution procedure can predict the binding sites of transcription factors and reveal polymorphisms in control elements between phylogenetic clades. Phylogenetic footprinting was applied to the *Hoxc8* early enhancer region, where it successfully identified a known protein-binding *cis*-regulatory motif that had previously been analyzed in depth by functional methods [38]. The authors demonstrated that an eight-species analysis is clearly superior to the conventional two-species methodology for this type of study.

Another group of specialized phylogenetic footprinting algorithms finds the most conserved motifs among the input sequences, as measured by a parsimony score of the underlying phylogenetic tree [39,40]. These algorithms have been used successfully to identify a variety of regulatory elements, some known and some novel, in sets of diverse vertebrate DNA sequences. Although phylogenetic footprinting methods show a lot of promising results, their use requires prior information about the location of orthologous regions in genomic intervals of interest. Multiple sequence alignments can help in defining these regions by finding longer conserved regions that can serve as guides to functionally important elements [10].

Analysis of conservation

The most obvious but difficult question to ask in comparative studies is how to define a functionally significant level of sequence conservation between species. Although two-way comparison is effective for discovery of evolutionarily constrained elements, distinguishing them from conserved sequences that are present due to lack of sufficient divergence time is not straightforward and requires knowledge of the neutral substitution rate [13,41]. In the majority of comparative genomics studies the definition of a significant level of conservation between two species has been intuitive, or based on biological experience. For example, aligning sequences in divergent noncoding regions proved useful in analyzing the enhancer in the β-globin locus-control region [42]. The conventional cutoff of 70-75% conservation over 100 base-pairs for the human-mouse comparison has produced discoveries of several important biologically functional elements [10,43]. One of the major obstacles to applying a single universal conservation criterion for potential regulatory regions is the substantial variation in the underlying mutation rates from region to region [13,41]. Conservation scores that incorporate the local neutral substitution rate are now available for the human and mouse genomes [13], and they can help to determine if a particular sequence is likely to be functional.

A more detailed analysis of interspecies pairwise genomic sequence alignments, aiming to distinguish regulatory regions from neutrally evolving DNA, has appeared recently [44]. This study proposed scoring procedures that evaluate alignments for properties other than overall percentage identity, although highly conserved noncoding sequences have proven to be good indicators of regulatory elements; among these procedures are discrimination on the basis of frequencies of nucleotide pairs or gaps, in combination with scoring procedures that include the alignment context, using frequencies of short runs of alignment columns. This study [44] thus gave a good start for extensive testing of these measures.

Adding genomic sequences from multiple vertebrates to the analysis makes the problem of estimating conservation even less trivial. Expanding pairwise analyses of conservation to multiple sequence alignments would require calculation of the neutral substitution rate between all pairs of sequences. That would give a weighted contribution of each sequence in the multiple alignment, but would also requires much more detailed evolutionary information than is available now. A three-way comparison makes it possible to enrich a pairwise alignment, and a simplified method for calculating a level of active non-coding conservation in such a comparison [45] is based on the supposition that actively conserved human-mouse noncoding sequences are likely to be present in additional mammals, whereas noncoding regions that are similar because of an insufficient accumulation of random mutations will not be present in other mammals.

Visualization of results

Visualization of results is a critical aspect of comparative sequence analysis, since manual examination of alignment

on the scale of long genomic regions presents a significant challenge and is not efficient. Alignment-browsing systems should identify regions that exhibit properties suggestive of a particular biological function, for example well-conserved segments within an alignment, or matching the consensus sequence for a specific transcription-factor-binding site [27].

There are several publicly available visualization tools for long pairwise DNA alignments. PipMaker [15,34] represents the level of conservation in ungapped regions of a BLASTZ local alignment as horizontal dashes. VISTA [45-47] displays comparative data in the form of a curve, where conservation is calculated in a sliding window of a gapped global alignment. SynPlot [23] also generates a curve plot calculated from a global alignment, but displays it slightly differently. All three tools can also be used to visualize multiple pairwise alignments [1,12,23,34], but one of the sequences needs to be selected as a reference, and the level of conservation is displayed on its scale. The same principle of selecting a reference sequence is utilized for whole genomes in the UCSC genome browser [36,48] and the VISTA browser [14,49].

Figure 1 shows a multiple pairwise VISTA display of a 5 kilobase fragment of the *CFTR* region aligned by MLAGAN [16]. This view is based on the coordinates of the human sequence and displays the level of conservation between human and all other sequences in the multiple alignment. The first exon of the *CAV1* gene is clearly well conserved across all 11 species, including the pufferfish *Fugu*. The upstream region of the *CAV1* gene (at 183 kb) has a distinct area of non-coding conservation across most of the pairwise comparisons, ranging from human/mouse to human/chicken. On the other hand, there are some peaks of non-coding conservation (at 181 kb) that are found in some mammalian species, but not others.

Knowing the phylogenetic relationship among species is important for building and analyzing multiple alignments, so visualizing sequence alignment data while taking phylogenetic trees into account presents a significant advance. A recently developed new program from the VISTA family, Phylo-VISTA (short for Phylogenetic VISTA) [50], uses phylogenetic relationships as a guide to display and analyze the level of conservation across internal nodes of the phylogenetic tree. Using the entire multiple alignment, not a reference sequence, as a base in the *x* axis of the visualization allows for additional options, such as presentation of comparative data together with available annotations for all sequences, and computation of a measure of similarity for any node of the phylogenetic tree.

In conclusion, pairwise sequence comparisons of the complete genomes of human and mouse have brought the revolutionary discovery that more than half of the functionally conserved sequences in the human genome are not protein-encoding [13]. Unfortunately, pairwise studies also make it

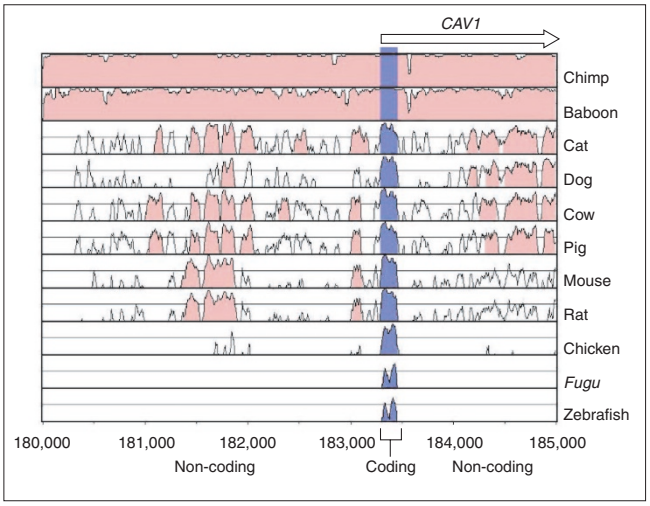

Figure 1
Multi-VISTA display of a 5 kilobase fragment of the MLAGAN alignment of the *CFTR* region [16]. This view is based on the coordinates of the human sequence and displays the level of conservation between human and all other sequences in the multiple alignment. A fragment of the *CAV1* gene is shown as an arrow above the plots. The following cutoffs were used to show the conserved regions: above 80% over 100 bp for chimpanzee, baboon, cat, dog, cow, pig, mouse and rat; above 65% over 100 bp for chicken; above 50% over 100 bp for *Fugu* and zebrafish.

clear that functional noncoding sequences are not easily distinguished from non-functional segments that happen to have accumulated very few mutations since the last common ancestor. Initial reports suggest that multi-species DNA comparisons have greater potential for filtering out evolutionarily neutral regions, and should therefore provide a more reliable basis for decoding and annotating genomic sequences at high resolution. This would improve our ability to discover non-protein-coding functional elements, which are currently poorly understood in comparison to their coding counterparts. Thus, we face the exciting prospect of discovering which species are the most informative in comparative studies, developing sophisticated algorithms for multi-sequence alignment and analysis of conservation, and building new effective visualization techniques for comparative data.

Acknowledgements

The authors are grateful to Nameeta Shah, Michael Brudno, Olivier Couronne, Shyam Prabhakar, Len Pennacchio and Alexander Poliakov for help and discussion. ID was partially supported by the Programs for Genomic Applications grant from the NHLBI/NIH.

References

1. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC: **Cross-species sequence comparisons: a review of methods and available resources.** *Genome Res* 2003, **13:**1-12.
2. Pennacchio LA, Rubin EM: **Genomic strategies to identify mammalian regulatory sequences.** *Nat Rev Genet* 2001, **2:**100-109.
3. Sidow A: **Sequence first. Ask questions later.** *Cell* 2002, **111:**13-16.

4. Ureta-Vidal A, Ettwiller L, Birney E: **Comparative genomics: genome-wide analysis in metazoan eukaryotes.** *Nat Rev Genet* 2003, **4:**251-262.
5. Wei L, Liu Y, Dubchak I, Shon J, Park J: **Comparative genomics approaches to study organism similarities and differences.** *J Biomed Inform* 2002, **35:**142-150.
6. Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES: **Human and mouse gene structure: comparative analysis and application to exon prediction.** *Genome Res* 2000, **10:**950-958.
7. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction.** *Bioinformatics* 2001, **17 Suppl 1:**S140-S148.
8. Guigo R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C, Antonarakis SE, Brent MR: **Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes.** *Proc Natl Acad Sci USA* 2003, **100:**1140-1145.
9. Pennacchio LA, Olivier M, Hubacek JA, Cohen JC, Cox DR, Fruchart JC, Krauss RM, Rubin EM: **An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing.** *Science* 2001, **294:**169-173.
10. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA: **Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.** *Science* 2000, **288:**136-140.
11. Hardison RC, Oeltjen J, Miller W: **Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome.** *Genome Res* 1997, **7:**959-966.
12. Gottgens B, Barton LM, Chapman MA, Sinclair AM, Knudsen B, Grafham D, Gilbert JG, Rogers J, Bentley DR, Green AR: **Transcriptional regulation of the stem cell leukemia gene (SCL) - comparative analysis of five vertebrate SCL loci.** *Genome Res* 2002, **12:**749-759.
13. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril, JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al.*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
14. Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, Rubin EM, Pachter L, Dubchak I: **Strategies and tools for whole genome alignments.** *Genome Res*, 2003, **13:**73-80.
15. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ.** *Genome Res* 2003, **13:**103-107.
16. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S, NISC Comparative Sequencing Program: **LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.** *Genome Res* 2003, **13:**721-731
17. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC *et al.*: **Comparative analyses of multi-species sequences from targeted genomic regions.** *Nature* 2003, **424:**788-793.
18. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, *et al.*: **Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*.** *Science* 2002, **297:**1301-1310.
19. Kent WJ, Zahler AM: **Conservation, regulation, synteny, and introns in a large-scale *C. briggsae* - *C. elegans* genomic alignment.** *Genome Res* 2000, **10**:1115-1125.
20. Frazer KA, Chen X, Hinds DA, Pant PV, Patil N, Cox DR: **Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates.** *Genome Res* 2003, **13:**341-346.
21. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM: **Phylogenetic shadowing of primate sequences to find functional regions of the human genome.** *Science* 2003, **299:**1391-1394.
22. Cooper GM, Brudno M, Green ED, Batzoglou S, Sidow A, NISC Comparative Sequencing Program: **Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes.** *Genome Res* 2003, **13:**813-820.
23. Gottgens B, Gilbert JG, Barton LM, Grafham D, Rogers J, Bentley DR, Green AR: **Long-range comparison of human and mouse *SCL* loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences.** *Genome Res* 2001, **11:**87-97.
24. Powell JR: *Progress and Prospects in Evolutionary Biology: The* Drosophila *Model.* Oxford: Oxford University Press; 1997.
25. Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, *et al.*: **Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome.** *Genome Biol* 2002, **3:**research0086.1-0086.20.
26. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423:**241-254.
27. Miller W: **Comparison of genomic DNA sequences: solved and unsolved problems.** *Bioinformatics* 2001, **17:**391-397.
28. Thompson JD, Higgins DG, Gibson TJ: **CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
29. Thompson JD, Plewniak F, Poch O: **A comprehensive comparison of multiple sequence alignment programs.** *Nucleic Acids Res* 1999, **27:**2682-2690.
30. **MLAGAN** [http://lagan.stanford.edu]
31. Bray N, Pachter L: **The MAVID multiple alignment server.** *Nucleic Acids Res* 2003, **31:**3525-3526.
32. **MAVID** [http://baboon.math.berkeley.edu/mavid]
33. Bray N, Dubchak I, Pachter L: **AVID: a global alignment program.** *Genome Res* 2003, **13:**97-102.
34. **PipMaker and MultiPipMaker** [http://bio.cse.psu.edu/pipmaker/]
35. Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, Smit A, Green ED, Hardison RC, Miller W, NISC Comparative Sequencing Program: **MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences.** *Nucleic Acids Res* 2003, **31:**3518-3524.
36. **Genome browser at UCSC** [http://genome.ucsc.edu]
37. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT: **Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints.** *J Mol Biol* 1988, **203:**439-455.
38. Sumiyama K, Kim CB, Ruddle FH: **An efficient *cis*-element discovery method using multiple sequence comparisons based on evolutionary relationships.** *Genomics* 2001, **71:**260-262.
39. Tompa M: **Identifying functional elements by comparative DNA sequence analysis.** *Genome Res* 2001, **11:**1143-1144.
40. Blanchette M, Tompa M: **Discovery of regulatory elements by a computational method for phylogenetic footprinting.** *Genome Res* 2002, **12:**739-748.
41. Hardison RC: **Conserved noncoding sequences are reliable guides to regulatory elements.** *Trends Genet* 2000, **16:**369-372.
42. Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, *et al.*: **Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution.** *Genome Res* 2003, **13:**13-26.
43. Elnitski L, Miller W, Hardison R: **Conserved E boxes function as part of the enhancer in hypersensitive site 2 of the β-globin locus control region. Role of basic helix-loop-helix proteins.** *J Biol Chem* 1997, **272:**369-378.
44. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, Eswara P, O'Connor MJ, Schwartz S, Miller W, Chiaromonte F: **Distinguishing regulatory DNA from neutral sites.** *Genome Res* 2003, **13:**64-72.
45. Dubchak I, Brudno M, Pachter LS, Loots GG, Mayor C, Rubin EM, Frazer KA: **Active conservation of noncoding sequences revealed by three-way species comparisons.** *Genome Res* 2000, **10:**1304-1306.
46. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I: **VISTA: visualizing global DNA sequence alignments of arbitrary length.** *Bioinformatics* 2000, **16:**1046-1047.
47. **VISTA** [http://www-gsd.lbl.gov/vista]
48. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006.
49. **VISTA Genome Browser** [http://pipeline.lbl.gov]
50. Shah N, Couronne O, Pennacchio LA, Brudno M, Batzoglou S, Bethel E, Rubin EM, Hamann B, Dubchak I: **Phylo-VISTA: an interactive visualization tool for multiple DNA sequence alignments.** *Bioinformatics,* in press.