JCI The Journal of Clinical Investigation

Comparative genomic tools and databases: providing insights into the human genome

Len A. Pennacchio, Edward M. Rubin

J Clin Invest. 2003;111(8):1099-1106. https://doi.org/10.1172/JCI17842.

Spotlight

The increasing availability of genomic sequence from multiple organisms has provided biomedical scientists with a large dataset for orthologous-sequence comparisons. The rationale for using cross-species sequence comparisons to identify biologically active regions of a genome is based on the observation that sequences that perform important functions are frequently conserved between evolutionarily distant species, distinguishing them from nonfunctional surrounding sequences. This is most readily apparent for protein-encoding sequences but also holds true for the sequences involved in the regulation of gene expression. While these observations have frequently been made retrospectively following the analysis of previously discovered genes or gene-regulatory sequences, examination of orthologous genomic sequences from several vertebrates has shown that the inverse is also true. Specifically, studying evolutionarily conserved sequences is a reliable strategy to uncover regions of the human genome with biological activity. To assist biomedical investigators in taking advantage of this new paradigm, various comparative sequence-based visualization tools and databases have been developed. Already, these new publicly accessible resources have been successfully exploited by investigators for the discovery of biomedically important new genes and sequences involved in gene regulation. Comparative genomic visualization tools The two most commonly used comparative genomic tools are Visualization Tool for Alignment (VISTA) and Percent Identity Plot Maker (PipMaker) (1, 2). The primary goal of both programs is to turn raw orthologous-sequence data from […]

Find the latest version:

https://jci.me/17842/pdf

Comparative genomic tools and databases: providing insights into the human genome

Len A. Pennacchio and Edward M. Rubin

Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, California, USA
Joint Genome Institute, Walnut Creek, California, USA

J. Clin. Invest. **111**:1099–1106 (2003). doi:10.1172/JCI200317842.

The increasing availability of genomic sequence from multiple organisms has provided biomedical scientists with a large dataset for orthologous-sequence comparisons. The rationale for using cross-species sequence comparisons to identify biologically active regions of a genome is based on the observation that sequences that perform important functions are frequently conserved between evolutionarily distant species, distinguishing them from nonfunctional surrounding sequences. This is most readily apparent for protein-encoding sequences but also holds true for the sequences involved in the regulation of gene expression. While these observations have frequently been made retrospectively following the analysis of previously discovered genes or gene-regulatory sequences, examination of orthologous genomic sequences from several vertebrates has shown that the inverse is also true. Specifically, studying evolutionarily conserved sequences is a reliable strategy to uncover regions of the human genome with biological activity. To assist biomedical investigators in taking advantage of this new paradigm, various comparative sequence-based visualization tools and databases have been developed. Already, these new publicly accessible resources have been successfully exploited by investigators for the discovery of biomedically important new genes and sequences involved in gene regulation.

Address correspondence to: Len Pennacchio, Genome Sciences Department, MS 84-171, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, California 94720, USA.
Phone: (510) 486-7498; Fax: (510) 486-4229;
E-mail: LAPennacchio@lbl.gov.

Conflict of interest: The authors have declared that no conflict of interest exists.

Nonstandard abbreviations used: Visualization Tool for Alignment (VISTA); Percent Identity Plot Maker (PipMaker); National Center for Biotechnology Information (NCBI); University of California at Santa Cruz (UCSC); Sequence Search and Alignment by Hashing Algorithm (SSAHA); conserved noncoding sequence 1 (CNS1); stem cell leukemia (SCL).

Comparative genomic visualization tools

The two most commonly used comparative genomic tools are Visualization Tool for Alignment (VISTA) and Percent Identity Plot Maker (PipMaker) (1, 2). The primary goal of both programs is to turn raw orthologous-sequence data from multiple species into visually interpretable plots to drive biological experimentation. Some of their common features include the ability to compare multiple megabases of sequence simultaneously from two or more species, web accessibility, and the option to customize numerous features by the user. While each program uses different overall strategies, they both allow for the identification of conserved coding as well as noncoding sequences between species.

VISTA combines a global-alignment program (AVID) (3) with a running-plot graphical tool to display the alignment (1) (http://www-gsd.lbl.gov/vista/). Global alignments are produced when two DNA sequences are compared and an optimal similarity score is determined over the entire length of the two sequences (Figure 1). In contrast, PipMaker uses BLASTZ, a modified local-alignment program, and displays plots with solid horizontal lines to indicate ungapped regions of conserved sequence (i.e., blocks of alignments that lack insertions or deletions) (http://bio.cse.psu.edu/pipmaker/) (2). Local alignments are generated when two DNA sequences are compared and optimal similarity scores are determined over numerous subregions along the length of the two sequences (Figure 1).

For visual comparison of the VISTA and PipMaker outputs, orthologous *ApoE* genomic sequence from humans and chimpanzees was independently examined by web-based versions of each program (Figure 2, a and b; and Table 1). In both cases, DNA sequences in FASTA format were submitted to web-based servers along with an annotation file of the location of exons and repeat sequences. In general, both programs provide similar interpretation of the input sequence files; namely, high levels of sequence homology are noted between both of these closely related primate species. In this example, known functional regions (exons and gene-regulatory elements) in the interval cannot be readily identified based on conservation because of lack of divergence time between humans and chimpanzees. As a second example, similar human versus mouse *ApoE* genomic-sequence comparisons were performed by both VISTA and PipMaker (Figure 2, c and d). Comparison of these more distantly related mammals revealed conserved sequences corresponding to previously defined functional elements. These include exonic sequences that display high levels of homology between humans and mice as well as two experimentally defined *ApoE* enhancers (Figure 2, c and d) (4–6). Additional conservation is noted upstream of exon 1 within the putative proximal promoter.

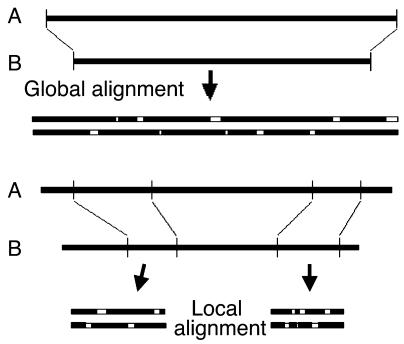

Figure 1
Comparison of local- and global-alignment algorithm strategies. Top: Global alignments are generated when two DNA sequences (A and B) are compared and an optimal similarity score is determined over the entire length of the two sequences. Bottom: Local alignments are produced when two DNA sequences (A and B) are compared and optimal similarity scores are determined over numerous subregions along the length of the two sequences. The local-alignment algorithm works by first finding very short common segments between the input sequences (A and B), and then expanding out the matching regions as far as possible.

These examples emphasize the importance of identifying the proper evolutionary distance for sequence comparisons to provide the correct window for identifying conserved sequences with functionality. For instance, human/chimpanzee comparison of the *ApoE* interval was not informative, while human/mouse comparison identified functional coding and noncoding sequences in this interval. While in this case primate/rodent comparison was informative, no two mammalian species provide the ideal distance for sequence comparison when the entire genome is examined, since different regions of the mammalian genome have evolved at significantly different rates (7–12). Thus, evolutionary distances must be varied depending on the genomic interval being studied and the biological question being investigated.

A useful characteristic of PipMaker is the linear contiguity of blocks (lines) that represent conserved elements with ungapped sequence alignments (i.e., blocks of alignments that lack insertions or deletions). This feature can aid in distinguishing coding sequence that is less flexible to insertions and/or deletions compared with functional noncoding DNA. In Figure 2c, note the linear blocks of alignments that appear beneath *ApoE* exons but not beneath regulatory sequences. A useful aspect of VISTA is the easily interpretable peaklike features depicting conserved DNA sequences. For instance, peaks of conservation are readily apparent beneath exons and gene-regulatory sequences (Figure 2d). While these peak features do not enable clear demarcation of exons boundaries, they allow the user to easily identify candidate gene-regulatory elements as well as evolutionarily conserved coding domains. Regardless of these differences in the alignment technique and display, both programs provide biomedical scientists with an easily accessible entry point to visualize comparative sequence data for regions of conservation (and putative function) surrounding a gene or genomic interval of interest. While VISTA and PipMaker are the most commonly used visualization packages, several additional tools for comparative genomic alignments with plotlike outputs are also available (13–16).

Whole-genome browsers

In the preceding section, computational tools for gene-by-gene (or region-by-region) analyses were described. These original tools sought to provide biologists with user-defined features for custom, small-scale analysis, frequently from sequence generated in individual laboratories that was manually input into the VISTA or PipMaker web server. The recent public availability of large amounts of whole-genome sequence for numerous organisms (human,

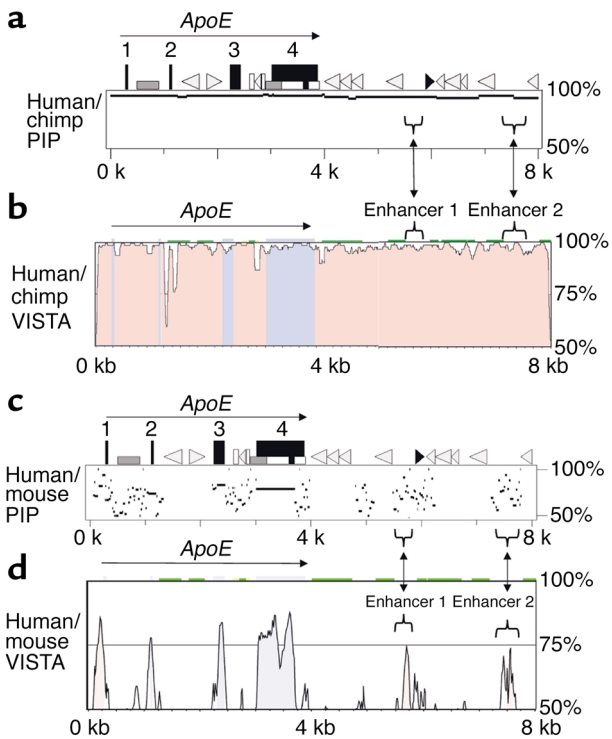

Figure 2
Human/chimpanzee and human/mouse *ApoE* genomic-sequence comparisons. (**a**) PipMaker analysis with human sequence depicted on the horizontal axis and percentage similarity to chimpanzee on the vertical axis. Exons are indicated by black boxes and repetitive elements by triangles above the plot. Each PIP horizontal bar indicates regions of similarity based on the percent identity of each gap-free segment in the alignment. Once a gap (insertion or deletion) is found within the alignment, a new bar is created to display the adjacent correspondent gap-free segment. (**b**) VISTA analysis with human sequence shown on the *x* axis and percentage similarity to chimpanzee on the *y* axis. The graphical plot is based on sliding-window analysis of the underlying genomic alignment. In this illustration, a 100-bp window is used that slides at 40-bp nucleotide increments. Blue and pink shading indicate conserved coding and noncoding DNA, respectively. Green and yellow bars immediately above the VISTA plot correspond to various repetitive DNA elements. (**c**) PipMaker analysis with human sequence depicted on the horizontal axis and percentage similarity to mouse on the vertical axis. (**d**) VISTA analysis with human sequence shown on the *x* axis and percentage similarity to mouse on the *y* axis. Two experimentally defined enhancers are indicated on each of the plots (4–6).

Table 1
Comparative genomic websites for various computational tools and databases

Comparative genomic visualization tools	Websites
VISTA	http://www-gsd.lbl.gov/vista/
PipMaker	http://bio.cse.psu.edu/pipmaker/
Whole-genome annotation browsers	
NCBI Map Viewer	http://www.ncbi.nlm.nih.gov
UCSC Genome Browser	http://genome.ucsc.edu/
Ensembl	http://www.ensembl.org/
Whole-genome comparative genomic browsers	
UCSC Genome Browser	http://genome.ucsc.edu/
VISTA Genome Browser	http://pipeline.lbl.gov/
PipMaker	http://bio.cse.psu.edu/genome/hummus/
Custom comparisons to whole genomes	
GenomeVista (AVID)	http://pipeline.lbl.gov/cgi-bin/GenomeVista
UCSC Genome Browser (BLAT)	http://genome.ucsc.edu/
ENSEMBL (SSAHA)	http://www.ensembl.org/
NCBI (BLAST)	http://www.ncbi.nlm.nih.gov/blast/

mouse, rat, fugu, tetraodon, ciona, etc.) has enabled large-scale analysis of individual genomes as well as genome-to-genome comparisons. These whole-genome analyses, accessible through web-based browsers, provide preprocessed databases for the scientific community (17–21).

Annotation browsers

The completion of a draft sequence and assembly of the human genome was an enormous accomplishment and provided a vast sequence dataset readily accessible to biomedical investigators. While these sequence data were initially useful for researchers seeking additional genomic sequence for individual genes of interest based on homology searches, the original assembly was simply a large database composed of strings of A's, C's, T's, and G's that lacked reference to and descriptions of key landmarks. Fortunately, this void has rapidly been filled by the success of large computational projects focused on the detailed annotation of the human genome. Today, three large centers provide human-genome annotation: the National Center for Biotechnology Information (NCBI), the University of California at Santa Cruz (UCSC), and the Sanger Center. These annotation outputs are all web-accessible and are known as NCBI Map Viewer, UCSC Genome Browser (22), and Ensembl (23), respectively (Table 1). In addition to exon annotation across the entire genome, these browsers contain a tremendous amount of additional annotation for features such as repetitive DNA, expressed-sequence tags, CpG islands, and single-nucleotide polymorphisms.

Comparative genomic browsers

In addition to gene annotation for the entire human genome, online resources have also recently become available for whole-human/whole-mouse comparative sequence data. Several important advances have made whole-genome comparisons possible. Whole-genome assemblies, in addition to satisfying the obvious need for sequence data for a given genome, have provided the substrates for genome-to-genome comparisons. Furthermore, the successful whole-genome annotation of genes, including their chromosomal location, serves as a reference for the position of a given alignment in the genome; previous gene-by-gene comparisons required the user to painstakingly input these annotation features. For mammals, this gene annotation is most detailed for the human genome, though progress is being made in annotating the puffer fish, mouse, and rat genomes. As a consequence, current whole-genome comparisons primarily use the human genome as the base reference sequence. Three major resources are currently available for preprocessed human/mouse whole-genome comparisons: UCSC Genome Browser, VISTA Genome Browser, and PipMaker (Table 1).

The UCSC Genome Browser has recently integrated comparative sequence information for annotation of the human genome. Similar to this browser's other annotation fields, comparative genomic information is presented as "tracks." To illustrate the UCSC Genome Browser's comparative genomic analysis, several tracks for the human/mouse *ApoE* interval are shown (Figure 3). These comparative data are presented in two formats.

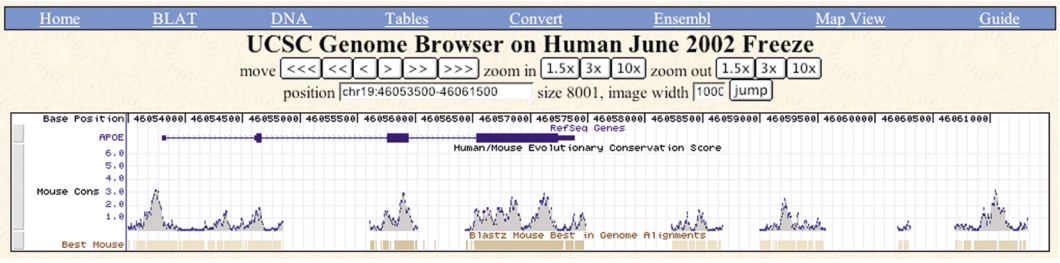

Figure 3
UCSC Genome Browser output for human/mouse sequence comparison of the *ApoE* gene (22). Human sequence is depicted on the *x* axis, and the numbering corresponds to the position of human chromosome 19 based on the UCSC June 2002 freeze (22). Note the different scoring system in contrast to percent identity, with peaks representing L-scores that take into account the context of the level of conservation. Conservation in relatively nonconserved regions receives higher L-scores than similar conservation in relatively highly conserved regions. As a second display of conservation, the "best mouse" track uses blocks whose length and shading represent the conservation.

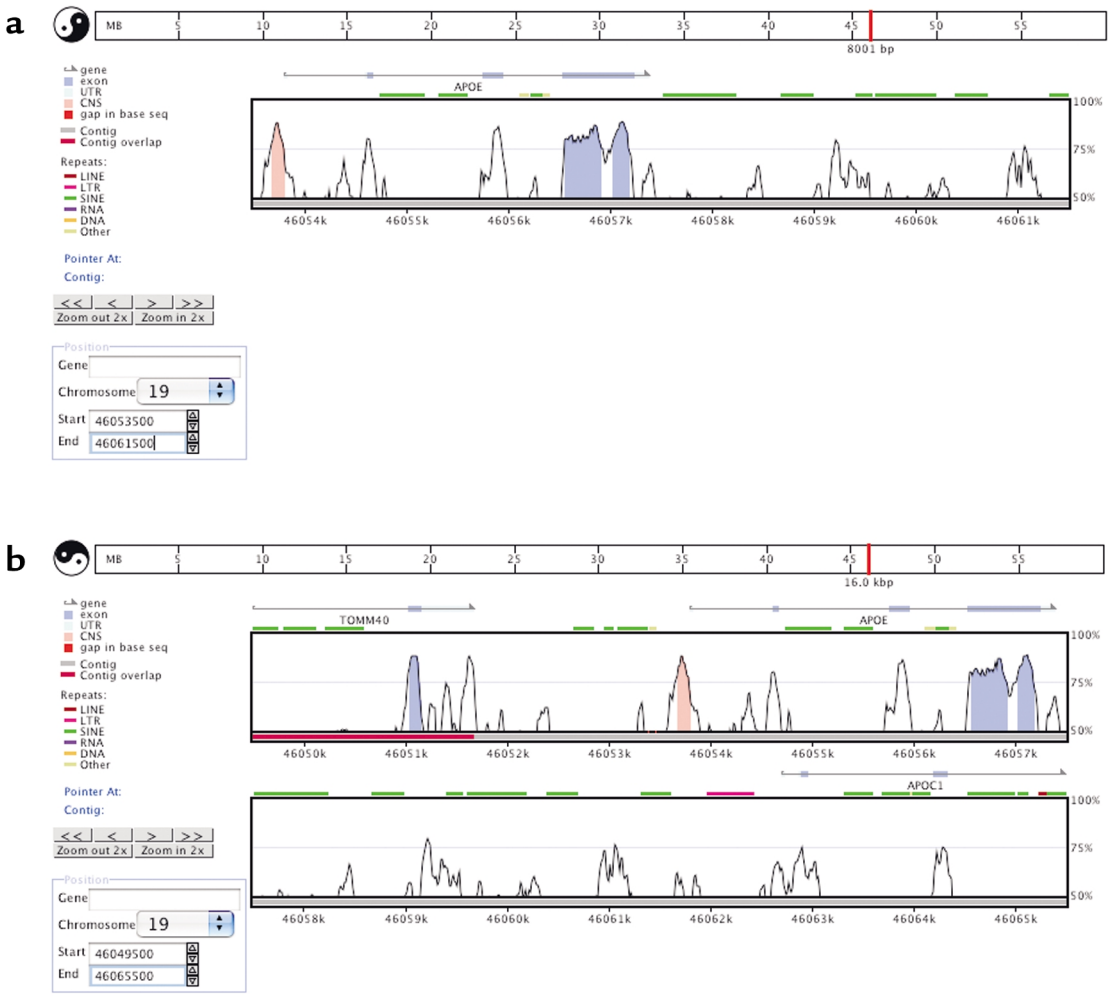

Figure 4
VISTA Genome Browser output for human/mouse sequence comparison of the *ApoE* gene (1). (**a**) The same genomic interval found in Figure 3 was examined. (**b**) A twofold "zoom out" was performed on the interval found in **a**, allowing the neighboring *ApoE* genes to be determined. Colored bars immediately above the VISTA plot correspond to various repetitive DNA elements.

First, a highly conserved sequence track is displayed as blocks whose length and shading indicate the size and level of homology between humans and mice (Figure 3, best mouse track). Second, human/mouse conservation data are depicted as a track with running plots displaying "L-scores" to indicate the level of conservation (Figure 3, mouse cons track). The power of this latter scoring system is that conservation is examined in the context of the genomic interval (rather than its strict percent identity for a given interval). Regions of high conservation in otherwise nonconserved intervals receive higher L-scores than regions of conservation in relatively highly conserved intervals. The rationale for such a strategy is based on the fact that neutral rates of DNA sequence change are highly variable in the mammalian genome (20). Thus, conservation in regions with faster neutral rates of change is more likely to be functional than conservation in slowly evolving intervals.

The VISTA Genome Browser is a complementary web-based browser for interactive visualization of comparative sequence data using a VISTA plot format (Table 1). Features include customized definition of the window size of a region under investigation (zoom), tools for extracting DNA sequence from a region of interest, and tables of highly conserved DNA within an interval. The website is also integrated with the UCSC Genome Browser, allowing for a portal to immediately jump from comparative sequence data to more detailed annotation of the human genome.

As an example of the VISTA Genome Browser output, the human/mouse *ApoE* genomic interval was examined (Figure 4a). This plot was obtained by submission of the gene symbol *ApoE* at the VISTA Genome Browser website (Table 1). Note the similarity between the human/mouse VISTA plot obtained through genome-to-genome comparison and the gene-by-gene analysis shown in Figure 2d. This resource instantaneously provides precomputed human/mouse data, in contrast to the detailed custom input files required by the standard VISTA analysis program. Furthermore, this resource allows for immediate "zoom-in" and "zoom-out" options to

characterize the interval in more detail. For instance, by zooming out, one can readily identify neighboring genes, as well as candidate conserved noncoding sequences that may be important in gene regulation of *ApoE* (Figure 4b). While these preprocessed datasets appear to have wide-ranging biomedical value, they have not made the traditional VISTA program obsolete. The traditional VISTA program remains well suited for custom genome annotation beyond what is publicly available, for sequence comparisons besides human/mouse comparisons, and for specialized user-defined VISTA plots containing nonstandard features.

A third set of preprocessed genome data is available through PipMaker (24) (Table 1). In this analysis, human/mouse genomic-alignment plots are provided in a nonbrowser format and are retrievable as a PDF file for a gene or region of interest.

Efforts are being made to provide preprocessed comparative data beyond human and mouse. For instance, the VISTA and UCSC Genome Browsers have recently added rat genomic sequence. This allows the examination of human/mouse, human/rat, and mouse/rat comparative data, providing the opportunity to determine what is shared and what is unique to each species. In the near future, additional vertebrate genome assemblies will become available, and it is expected that they will be integrated into a similar framework. While significant computational challenges exist with such a complex dataset, more efficient algorithms are being developed, and the insights gained from multiple, simultaneous genome comparisons are likely to be significant.

Custom comparison to whole genomes

In addition to preprocessed whole-genome comparative data, several additional tools allow for any sequence from any organism to be compared with previously assembled and annotated genomes. They include GenomeVista and a server available through UCSC Genome Browser (Table 1).

GenomeVista uses the same data sources and algorithmic methods as are used to generate the alignments for the VISTA Genome Browser, but it allows users to input their own sequence of interest for direct comparison with the human, mouse, or rat genome. One can acquire these sequence files from in-house sequencing projects, or automatically retrieve them from sequence databases such as GenBank by simply inputting the accession number for the desired sequence at the GenomeVista website. The GenomeVista data output is similar to that of the VISTA Genome Browser but allows species other than those available in the current alignment to be examined in the context of the annotated human or mouse genome.

Similar to GenomeVista, the UCSC Genome Browser also allows custom sequence comparison with the human, mouse, or rat genome assembly (Table 1). This comparison uses BLAT, a modified BLAST alignment program, and provides an extremely fast homology search (25, 26). This tool is useful to quickly determine the mapping location for a sequence of interest and the annotation within that interval. The tool's speed, however, comes at the cost of reduced alignment sensitivity, and the complementary use of alternative comparative genomic tools such as VISTA or PipMaker is warranted. Similar fast homology searches against genomes are available at Ensembl and NCBI using the Sequence Search and Alignment by Hashing Algorithm (SSAHA) (27) and BLAST (25) alignment tools, respectively.

General insights from genomic-sequence comparisons of humans and mice

With these computational tools and databases, what early comparative genomic insights have been obtained about the human genome? The recent completion of the mouse genome draft sequence led to the surprising result that approximately 40% of the human genome's 3 billion base pairs could be aligned to the mouse genome at the nucleotide level (20). Using a separate conservation criterion of human/mouse sequences with ≥70% identity over ≥100 bp, more than 1 million independent human/mouse conserved elements could be defined (26). An obvious question arising from the identification of all this conservation is what (if any) is the functional significance of these conserved sequences?

Currently, the most obvious human genomic functional elements that display high levels of conservation across species are exons. This is not unexpected based on the known functional importance of the proteins that they encode. In one recent study, initial comparative data analyses indicate that greater than 90% of known human exons are conserved within the mouse (20, 28). Thus, we might expect that a subset of the approximately 1 million conserved human/mouse elements coincide with exons. As an exercise, we can roughly estimate the number of exons in the human genome. Current data suggest that there are about 30,000 human genes with an average of about 8 exons per gene, which indicates approximately 240,000 human exons (the average exon size is 150 bp). With a small number of exons not displaying conservation because of either their fast evolution or lack of an orthologous counterpart, this suggests that approximately 20% (200,000/1,000,000) of conserved human/mouse DNA elements are accounted for by coding sequence.

What can be said for the remaining approximately 800,000 roughly exon-sized conserved human/mouse sequences? It appears that a large portion of human/mouse conserved DNA occupies noncoding regions of the mammalian genome, although, in contrast to exons, we have very few clues as to their immediate functional significance. One of our biggest current genomic challenges is to determine how many of these noncoding conserved sequences are functional, and their precise biological role(s).

One category of functional noncoding DNA is sequences that participate in the regulation of neighboring genes. On a small scale, comparative genomics has proven its ability to uncover important gene-regulatory

elements based solely on conservation (8–10, 29–31). This is despite the fact that most transcription factor–binding sites are on the order of 6–12 bp in length. It appears that many gene-regulatory elements are frequently found within much larger blocks of conservation (80–500 bp), most likely because regulatory elements are a composite of numerous transcription factor–binding sites that direct gene expression. Unfortunately, to date we have only catalogued a small number of gene-regulatory elements outside of proximal promoters, and it is difficult to estimate how many of the 800,000 human/mouse exon-sized conserved noncoding sequences serve gene-regulatory (or other biological) functions. Similar to current successful exon prediction programs, future computational exploration of such datasets may reveal common features among various conserved noncoding sequence subclasses that allow for future predictions of sequences with similar biological activity. With human/mouse conservation serving as a filter for prioritizing human sequences likely to have biological activity, we predict that hypotheses based purely on comparative sequence data should increasingly lead to biological insights. In the next section, we focus on a limited number of recent examples in which comparative genomics has led to biological discoveries.

Gene identification

One of the clear utilities of comparative sequence analysis is for exon and gene identification. As stated previously, of the approximately 1 million human/mouse conserved elements, about one-fifth are probably due to conserved exons. Thus, while a significant fraction of the genes in the human genome have likely already been identified, genome-wide scans for conserved human/mouse sequences should aid in the identification of genes missed in the initial annotation of human sequence alone. Indeed, there have been several recent examples in which comparative sequence data have led to the discovery and functional understanding of previously undefined genes.

The complete human/mouse orthologous-sequence dataset proved particularly valuable in the characterization of gene families in humans and mice (32). For instance, by comparing olfactory receptor gene families on human chromosome 19, computational analysis indicated that humans have approximately 49 olfactory receptor genes, but only 22 had maintained an open reading frame and appeared functional. This contrasts with the vast majority of the homologous mouse genes that have retained an open reading frame. This finding of reduced olfactory receptor diversity in humans is consistent with the reduced olfactory needs and capabilities of humans relative to rodents. As a second example, pheromone receptor genes were also examined. In humans, 19 pheromone receptor genes were identified, but only one appeared functional. In contrast, homologous mouse sequences revealed 36 pheromone receptor genes, and at least 17 had maintained a complete open reading frame. Again, these data are consistent with the reduced pheromone response in humans relative to mice. This subset of examples highlights the use of comparative genomics to inventory gene content and correlate the differences to species-related biology.

Human/mouse comparative data have also led to the discovery of previously undetected biomedically important genes. Of particular relevance to cardiovascular disease was the discovery of *APOA5* in the chromosome 11 apolipoprotein gene cluster (33). While the human sequence for the genomic interval containing the intensively studied *APOA1/C3/A4* gene cluster had been available for many years, it was only comparison of the recently available orthologous mouse sequence that alerted investigators to the presence of *APOA5*. Through this comparative genomic entry point, functional studies were performed in mice that indicated that alteration in the level of *APOA5* significantly impacted plasma triglyceride concentrations. Mice overexpressing human *APOA5* displayed significantly reduced triglycerides, while mice lacking *ApoA5* had a large increase in this lipid parameter. In addition, multiple studies in humans have also supported a role for common *APOA5* genetic variation in influencing plasma triglyceride concentrations (33–37). To date, consistent and strong genetic associations have been established between minor *APOA5* alleles and increased triglycerides in Caucasian, African-American, Hispanic, and Asian populations (33–37).

Thus, even in well-studied genomic intervals such as the chromosome 11 apolipoprotein gene cluster, significant discoveries are possible through the exploitation of comparative sequence data. Though whole-genome annotation efforts are providing the location for the majority of genes in the human genome, undefined genes still exist. The above examples provide strong evidence for the utility of comparative genomic data to facilitate the identification of coding sequences based on conservation. An important follow-up question is, how well does this strategy apply to the identification of sequences encoding other important biological activities embedded in the human genome?

Identification of regulatory sequences

One of the first studies to use solely human/mouse comparative genomics as an approach to identify gene-regulatory elements was the examination of a cytokine gene cluster (including five ILs and 18 other genes) on human chromosome 5q31 (38). In this work, human/mouse comparative analysis was performed on a 1-Mb region, and 90 conserved noncoding sequences (≥70% identity over ≥100 bp) were identified. Of these elements, several corresponded to previously known gene-regulatory elements. One previously undefined conserved noncoding element was explored in finer detail based exclusively on its human/mouse sequence conservation (400 bp at 87% identity between human and mouse). This element was named conserved noncoding sequence 1 (CNS1) and was localized to the 15-kb interval between IL-4 and IL-13. To characterize the function of CNS1,

transgenic and knockout mouse studies were performed (38–40). Through these studies it was shown that CNS1 dramatically impacted the expression of three human cytokine (IL-4, IL-5, and IL-13) genes separated by more than 120 kb of sequence. Thus, from a purely comparative sequence-based starting point, conservation of sequence alone led to the identification of a novel gene-regulatory element that acts over long distances to modulate genes important in the inflammatory response. Follow-up studies to the initial discovery of CNS1 further support that this 400-bp element contains transcription factor–binding sites that coactivate IL-4, IL-5, and IL-13 (39, 40). The role of these ILs in a variety of common conditions such as asthma and inflammatory bowel disease has focused attention on CNS1.

A second example of comparative sequence analysis identifying gene-regulatory sequences prior to functional studies is the examination of a genomic interval containing the stem cell leukemia (SCL) gene (10, 14, 41). In these studies, the orthologous SCL genomic interval was examined in human, mouse, chicken, fugu, and zebrafish. All of the exons and eight known gene-regulatory elements in the interval were conserved between humans and mice, though only a subset were conserved between humans and chickens or between humans and fish. These data question the utility of sequence comparisons beyond mammals in thoroughly identifying gene-regulatory elements. However, in this study, power was obtained by the use of simultaneous deep sequence comparison across all five species of the highly conserved SCL intervals, including the promoter, exon 1, and the 3′ untranslated poly(A) region. Through phylogenetic footprinting (42), two highly conserved promoter sequences were shown to be necessary for full SCL expression in erythroid cells. This study showed that pairwise sequence comparisons had variable utility for identifying previously defined functional elements, and that deep sequence alignments could reveal highly conserved functional motifs.

While these examples are limited because large stretches of human and mouse orthologous genomic sequence have only recently become accessible, they highlight the power of comparative sequence analysis in discovering various functional regions of the human genome. Based on the evolutionary relationship among vertebrates, conservation provides a blueprint to our shared genomic machinery. While evolutionary conservation of DNA sequence alone cannot indicate function, its identification provides a strategy to reveal and prioritize otherwise unrecognizable sequences for further biological experimentation. Though most current comparative genomic insights have been derived from human/mouse sequence comparisons, more distant evolutionary groups (such as fish, birds, amphibians, and reptiles) will also contribute to the further annotation and understanding of the human genome. Since an undefined fraction of human/mouse conservation is likely to be nonfunctional, the analysis of sequences conserved between humans and mice as well as nonmammalian species will further enrich for biologically active sequences.

Conclusions

The flood of genomic-sequence data from a wide variety of animal species has only just begun. While databases, algorithms, and strategies for simultaneously examining sequence from evolutionarily related species already exist, large computational and experimental challenges lie ahead as sequence data exponentially increase. A field likely to expand significantly with the increasing availability of genomic sequence from multiple species is the computational identification of gene-regulatory and other noncoding functional DNA elements. Though we can currently make reasonable predictions for coding sequences embedded in the mammalian genome, only a limited number of functional elements have been identified in the more than 97% of the genome that is noncoding. The generation of a large dataset of conserved noncoding sequences coupled with other high-throughput genomic information such as gene expression data should contribute to the development of a vocabulary of DNA sequence that dictates gene expression and other noncoding functions embedded within the human genome. In the future, the annotation of the human genome that can be obtained through the various genome browsers will likely include sequences involved in gene regulation in addition to the already existing annotation of exons.

The recent availability and analysis of human and mouse genomic sequence have provided strong support for the future value of sequence information in biomedicine. We are approaching an era in which sequence data no longer limit us but, rather, accumulate rapidly with functional studies lagging behind. Intriguingly, though we are challenged by this glut of sequence information, additional genome sequences from mammalian and nonmammalian species will further help us to even better prioritize regions of the human genome for functional studies.

Acknowledgments

This work was supported in part by the NIH–National Heart, Lung, and Blood Institute Programs for Genomic Application grant HL-66681 (to E.M. Rubin) through the US Department of Energy under contract no. DE-AC03-76SF00098.

1. Mayor, C., et al. 2000. VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics.* **16**:1046–1047.
2. Schwartz, S., et al. 2000. PipMaker: a web server for aligning two genomic DNA sequences. *Genome Res.* **10**:577–586.
3. Bray, N., Dubchak, I., and Pachter, L. 2003. AVID: a global alignment program. *Genome Res.* **13**:97–102.
4. Grehan, S., Allan, C., Tse, E., Walker, D., and Taylor, J.M. 2001. Expression of the apolipoprotein E gene in the skin is controlled by a unique downstream enhancer. *J. Invest. Dermatol.* **116**:77–84.
5. Grehan, S., Tse, E., and Taylor, J.M. 2001. Two distal downstream enhancers direct expression of the human apolipoprotein E gene to astrocytes in the brain. *J. Neurosci.* **21**:812–822.
6. Shih, S.J., et al. 2000. Duplicated downstream enhancers control expression of the human apolipoprotein E gene in macrophages and adipose tissue. *J. Biol. Chem.* **275**:31567–31572.
7. Pennacchio, L.A., and Rubin, E.M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* **2**:100–109.
8. Jimenez, G., Gale, K.B., and Enver, T. 1992. The mouse beta-globin locus control region: hypersensitive sites 3 and 4. *Nucleic Acids Res.* **20**:5797–5803.
9. Dubchak, I., et al. 2000. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10**:1304–1306.

10. Gottgens, B., et al. 2000. Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat. Biotechnol.* **18**:181–186.
11. Hood, L., Rowen, L., and Koop, B.F. 1995. Human and mouse T-cell receptor loci: genomics, evolution, diversity, and serendipity. *Ann. N. Y. Acad. Sci.* **758**:390–412.
12. Koop, B.F., and Hood, L. 1994. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat. Genet.* **7**:48–53.
13. Margot, J.B., Demers, G.W., and Hardison, R.C. 1989. Complete nucleotide sequence of the rabbit beta-like globin gene cluster. Analysis of intergenic sequences and comparison with the human beta-like globin gene cluster. *J. Mol. Biol.* **205**:15–40.
14. Gottgens, B., et al. 2001. 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.* **11**:87–97.
15. Jareborg, N., and Durbin, R. 2000. Alfresco: a workbench for comparative genomic sequence analysis. *Genome Res.* **10**:1148–1157.
16. Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**:2478–2483.
17. Aparicio, S., et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science.* **297**:1301–1310.
18. Lander, E.S., et al. 2001. Initial sequencing and analysis of the human genome. *Nature.* **409**:860–921.
19. Venter, J.C., et al. 2001. The sequence of the human genome. *Science.* **291**:1304–1351.
20. Waterston, R.H., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature.* **420**:520–562.
21. Dehal, P., et al. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science.* **298**:2157–2167.
22. Kent, W.J., et al. 2002. The human genome browser at UCSC. *Genome Res.* **12**:996–1006.
23. Hubbard, T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30**:38–41.
24. Schwartz, S., et al. 2003. Human-mouse alignments with BLASTZ. *Genome Res.* **13**:103–107.
25. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410.
26. Kent, W.J. 2002. BLAT: the BLAST-like alignment tool. *Genome Res.* **12**:656–664.
27. Ning, Z., Cox, A.J., and Mullikin, J.C. 2001. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**:1725–1729.
28. Couronne, O., et al. 2003. Strategies and tools for whole-genome alignments. *Genome Res.* **13**:73–80.
29. Duret, L., and Bucher, P. 1997. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7**:399–406.
30. Hardison, R.C., Oeltjen, J., and Miller, W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* **7**:959–966.
31. Hardison, R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16**:369–372.
32. Dehal, P., et al. 2001. Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science.* **293**:104–111.
33. Pennacchio, L.A., et al. 2001. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science.* **294**:169–173.
34. Endo, K., et al. 2002. Association found between the promoter region polymorphism in the apolipoprotein A-V gene and the serum triglyceride level in Japanese schoolchildren. *Hum. Genet.* **111**:570–572.
35. Nabika, T., Nasreen, S., Kobayashi, S., and Masuda, J. 2002. The genetic effect of the apoprotein AV gene on the serum triglyceride level in Japanese. *Atherosclerosis.* **165**:201–204.
36. Pennacchio, L.A., et al. 2002. Two independent apolipoprotein A5 haplotypes influence human plasma triglyceride levels. *Hum. Mol. Genet.* **11**:3031–3038.
37. Talmud, P.J., et al. 2002. Relative contribution of variation within the APOC3/A4/A5 gene cluster in determining plasma triglycerides. *Hum. Mol. Genet.* **11**:3039–3046.
38. Loots, G.G., et al. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science.* **288**:136–140.
39. Mohrs, M., et al. 2001. Deletion of a coordinate regulator of type 2 cytokine expression in mice. *Nat. Immunol.* **2**:842–847.
40. Lee, G.R., Fields, P.E., and Flavell, R.A. 2001. Regulation of IL-4 gene expression by distal regulatory elements and GATA-3 at the chromatin level. *Immunity.* **14**:447–459.
41. Gottgens, B., et al. 2002. Transcriptional regulation of the stem cell leukemia gene (SCL): comparative analysis of five vertebrate SCL loci. *Genome Res.* **12**:749–759.
42. Gumucio, D.L., et al. 1996. Evolutionary strategies for the elucidation of cis and trans factors that regulate the developmental switching programs of the beta-like globin genes. *Mol. Phylogenet. Evol.* **5**:18–32.