Resource

Visualization of Multiple Genome Annotations and Alignments With the K-BROWSER

Kushal Chakrabarti[1] and Lior Pachter[2,3]

[1]*Department of Computer Science,* [2]*Department of Mathematics, University of California, Berkeley, Berkeley, California 94720, USA*

We introduce a novel genome browser application, the K-BROWSER, that allows intuitive visualization of biological information across an arbitrary number of multiply aligned genomes. In particular, the K-BROWSER simultaneously displays an arbitrary number of genomes both through overlaid annotations and predictions that describe their respective characteristics, and through the multiple alignment that describes their global relationship to one another. The browsing environment has been designed to allow users seamless access to information available in every genome and, furthermore, to allow easy navigation within and between genomes. As of the date of publication, the K-BROWSER has been set up on the human, mouse, and rat genomes.

Genome browsers at present (Kent et al. 2002; Clamp et al. 2003; Couronne et al. 2003, Hubbard et al. 2002) allow scientists to succinctly represent and visualize vast amounts of biological, genome-related information. This ability to efficiently represent and rapidly visualize such information has led to many important discoveries and otherwise proven generally invaluable. For instance, the UCSC genome browser serves more than 50,000 pages per day, most of which are for biomedical researchers studying regions surrounding various genes of interest. Although some of these requests are for genomic sequence, the overwhelming majority are used to display genomic annotations and predictions. These annotations and predictions, more commonly known as tracks, have been traditionally used to specify the locations of gene annotations and predictions, sequenced mRNAs, SNPs, and ESTs, and so on. More recently, tracks have also been used to specify the level of transcription in genomic regions and display cross-species conservation.

These tracks, in conjunction with present genome browser technology, have greatly contributed to recent breakthroughs by allowing rapid cross-referencing of diverse types of information. For instance, a scientist arguing for the recognition of a putative gene or exon would probably find important the ability to simultaneously use gene predictions, known mRNAs, and exons—even though information from one type of resource may not be sufficient, combined evidence might be persuasive. With the quality and diversity of these tracks quickly increasing, it is expected that they will greatly expand the power and scope of genome browsers.

In addition to the ability to cross-reference tracks within a particular genome, the ability to cross-reference tracks across genomes has also shown itself to be invaluable. Known as the comparative genomics method, this approach exploits the observation that functionally important regions in related organisms will, as a result of selective pressure, exhibit well-defined and interesting patterns of sequence conservation. This approach has been particularly useful and has led to significant improvements in cross-species gene finding (Korf et al. 2001; Alexandersson et al. 2003; Parra et al. 2003), structural RNA gene prediction (McCutcheon and Eddy 2003), and protein-binding site prediction (Boffelli et al. 2003; Cliften et al. 2003; Kellis et al. 2003). For instance, with regard to the previous example, if one needed to persuasively argue the existence of a gene in a particular organism, prior knowledge of an experimentally verified homolog in a syntenic region of a related organism could prove invaluable.

Present genome browsers, however, lack the ability to clearly represent information across genomes. To the extent that these browsers were originally designed to either display information on a single genome or relationships between multiple genomes—but not both—they all suffer from the same shortcoming; that is, they lack the ability to simultaneously display multiple alignment information and single genome tracks. The VISTA browser (http://pipeline.lbl.gov/vistabrowser), for instance, displays the degree of conservation between genomes but can only display the RefSeq gene track, and only for one genome at a time. On the other hand, the UCSC Genome Browser can simultaneously display both conservation and multiple tracks, but is restricted to a single genome. Despite somewhat successful efforts to develop solutions, for example, cross-species tracks, the UCSC Genome Browser does not scale well to the simultaneous visualization of any more than two genomes. Furthermore, it is difficult for a user to execute even simple queries on three genomes, for example, identify genes conserved between human and mouse but not in the rat genome. Although the Ensembl Genome Browser (Clamp et al. 2003; http://www.ensembl.org) does not at present have native support for comparative genomics information, it would face many of the same problems and limitations as the UCSC Genome Browser if it were to start. Similarly, the Ensembl Synteny Viewer (Clamp et al. 2003), despite being able to meaningfully represent synteny relationships between two genomes, is limited to two genomes, and seems further limited by its interface to a very small number of genomes.

There is, however, a more fundamental problem that precludes most present approaches from implementing a convenient multiple genome browser system. Because the genomic alignments that must underlie such browsers are not necessarily one-to-one, that is, are not necessarily between orthologous pairs, it is difficult for present genome browsers to represent many-to-many alignments as an alignment between genomes. In other words, it would be difficult for them to develop a representation that natively compared different genomes and not simply different regions.

[3]Corresponding author.
E-MAIL lpachter@math.berkeley.edu; FAX (510) 642-8204.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1957004.

K-BROWSER

The K-BROWSER has been specifically built to allow intuitive visualization of biological information both within and across an

14:716–720 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

arbitrary number of multiply aligned genomes. It has been designed around two principles.

Genome Symmetry

Every genome contains useful information, and a browsing solution should not limit the ability to navigate within or across genomes.

Genome Homology

Related genomes have evolved from a common ancestor, and these evolutionary relationships should be accurately reflected in both the representation and visualization of information.

The major component of the K-BROWSER, the image generation engine, takes as input a specific region in a specific genome, and produces a set of images that succinctly represents the requested region and all orthologous regions. In particular, the K-BROWSER generates a single image for each supported genome, displaying the corresponding orthologous region through (1) the tracks supported on that particular genome, (2) the multiple alignment that underlies the entire set of orthologous regions, and (3) the degree of conservation thereof. In addition, the K-BROWSER also provides the option to immediately and easily download the underlying multiple alignment.

Implementation

The K-BROWSER consists of several logical components, as shown in Figure 1, built on the foundation of the UCSC Genome Browser (Kent et al. 2002). Roughly speaking, the K-BROWSER requires an orthology map defining sets of orthologous regions in the input genomes, global alignments of the same orthologous regions, and track databases. It builds, based on the global alignments and the orthology map, a specialized database of segment tables, which is essentially an efficient representation of the alignment. It subsequently preprocesses the input track databases according to the segment tables and homology map, producing databases of realigned tracks (which are described in further detail below). Upon a Web request, the K-BROWSER front-end and

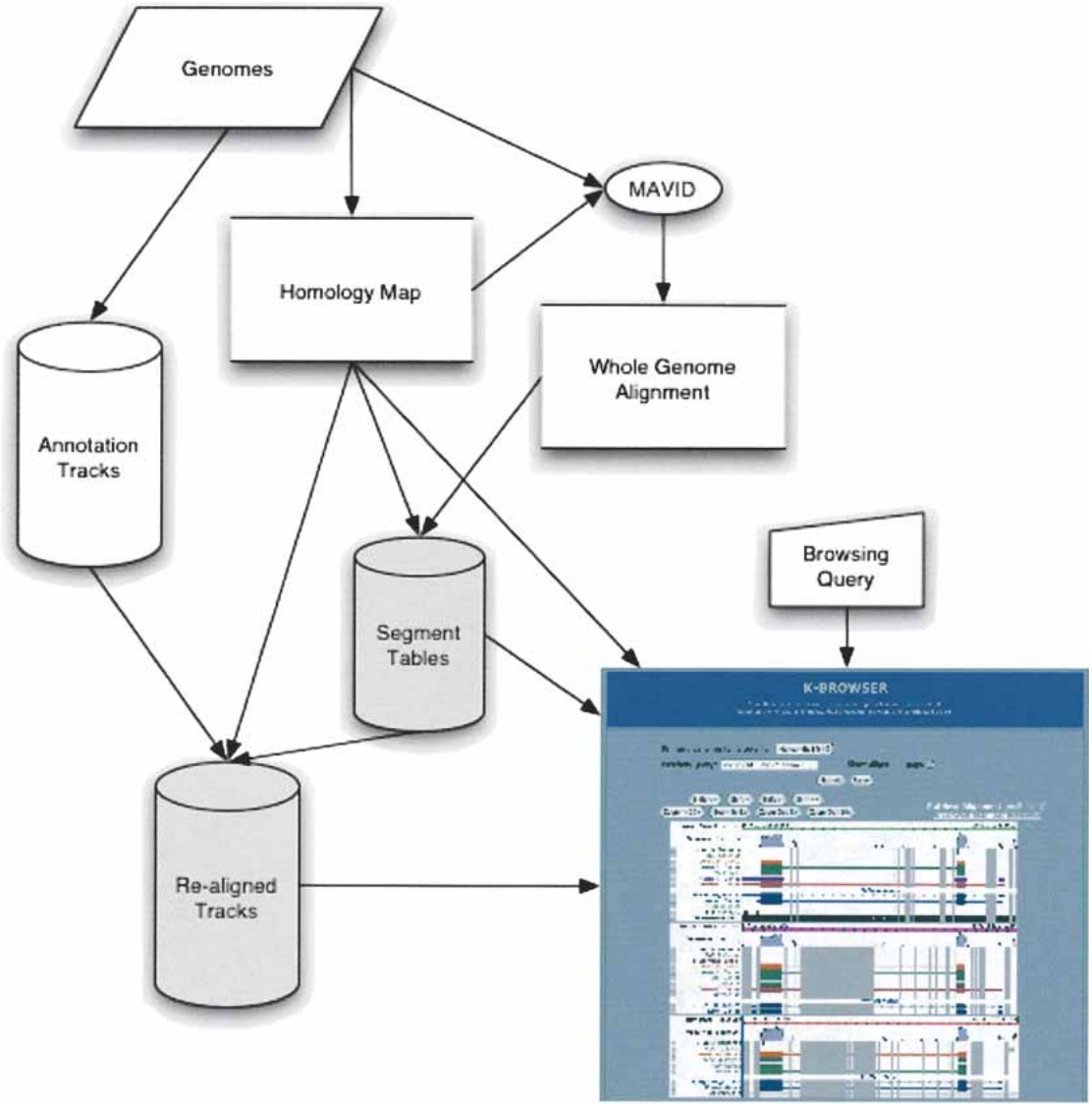

Figure 1 Schematic diagram of the K-BROWSER infrastructure. K-BROWSER builds two new databases (shaded) in addition to the standard annotation track database obtained from UCSC. These databases are based on a whole-genome alignment constructed using a homology map (in our case the blocks were aligned with the MAVID program). Using the segment tables and realigned tracks, a user query can be rapidly converted into a K-BROWSER figure.

image generation components use these databases to produce the downloaded images.

Of these components, the two responsible for the production of realigned track databases and image generation are most important, and are described in more detail below.

Track Realignment

This component is responsible for the necessary adjustment of the track database to make it consistent with the multiple alignment. With regard to Figure 1, it handles the preprocessing of the segment table and the realigned track databases, which correspond exactly to the two databases that the final user interface uses.

Intuitively, because the underlying sequence has been aligned, and one wants to visualize the tracks with respect to aligned sequence, the track must also be "aligned." As such, we extend the well-understood paradigm of sequence alignment, which requires that any two homologous bases occupy the same column of an (ideal) multiple alignment, and define the analogous idealization for visualized tracks. In particular, for an ideal visualization, we require that any two homologous entries (e.g., orthologous exons in mouse and rat) be visualized at the same position (e.g., in the same range of pixels on the horizontal axis). To this extent, realignment refers to the process of adjusting the original tracks with respect to a multiple alignment, so that this constraint is met.

Another interpretation of this realignment component is that it converts positions in genomic coordinates to alignment coordinates. Such an interpretation is both meaningful and useful, as alignment coordinates allow one to succinctly refer to positions in an arbitrary number of genomes with a single position. In fact, this interpretation is critical for a number of purposes, such as in determining which regions are orthologous to a requested region. It is interesting to note that, given this interpretation, the previously described readjustment reduces to the problem of converting each genomic-coordinate position in each track to the corresponding alignment-coordinate position.

This conversion process requires three inputs: (1) an orthology map that defines sets of orthologous regions between genomes, (2) a multiple alignment of every set of regions in the orthology map, and (3) databases of tracks for each of the genomes. These inputs are, respectively, derived from ongoing work (C. Dewey, pers. comm.) in homology maps, MAVID multiple alignments (Bray et al. 2003; Bray and Pachter 2004; http://baboon.math.berkeley.edu/mavid/), and the UCSC Genome Browser Database (Karolchik et al. 2003; http://genome.ucsc.edu). As output, the component produces a new database for each genome, and each database includes tables containing the realigned tracks and a new type of table known as a segment table.

Segment tables are useful in that they simply, efficiently, and uniquely represent an alignment. A segment table is then simply a table that contains certain information about every segment in the aligned sequence, with a segment defined to be a maximal ungapped sequence in the aligned sequence. To characterize an aligned sequence uniquely, a segment table need only include the length, the unaligned start position, and the aligned start position of each segment. Because an aligned sequence is simply the original sequence interspersed with gaps, and the segment table indicates all gaps in the alignment, that is, all positions that are not in any segment, the aligned sequence is uniquely characterized.

To the extent that segment tables efficiently represent the underlying alignment, they are used throughout the K-BROWSER. For instance, they allow one to efficiently translate genomic-coordinate positions to and from alignment-coordinate positions. In addition, they can be used to rapidly determine the exact positions and lengths of gaps in a particular region of the alignment. As it turns out, the latter is important for image generation, whereas the former is critical to both image generation and track realignment.

Although the tracks can, in principle, be realigned in the obvious way using segment tables, this approach requires substantial computational overhead because of repetitive database accesses. As a result, it does not scale with respect to the track database, and we instead use another approach that requires only a constant number of database accesses, whose cost is amortized over the different tracks. In particular, we iterate through each orthology set in the homology map, retrieve the appropriate segment table subset, and build an array that maps each genomic-coordinate index position to the corresponding alignment-coordinate position. Track realignment then simply reduces to the problem of extracting the positions of each entry in the track, looking them up in the array in constant time, and saving the results. This conversion has the intuitive effect of "stretching" out tracks over gaps in the alignment. However, because it is easy to determine the gaps in the region, it is also easy to determine the actual positions that the realigned track actually covers.

Despite being conceptually straightforward, track realignment is complicated by at least two factors: the diverse and dynamic nature of the track databases, and the existence of large-scale evolutionary events, for example, inversions. The K-BROWSER presently implements simple but sufficient solutions to both problems.

Because the diversity of tracks is constantly increasing as a result of new evidence, alignments, predictions, and annotations, it is impossible to know a priori exactly which fields of track databases need to be adjusted. Even though nearly all of the different tracks follow a small number of standard schema, it is quite possible that schema could either change or that new schema will be introduced. To this end, one must require either manual intervention or relatively complicated automatic (and error-prone) inference to ensure that the proper fields are adjusted. The present track realignment implementation relies on a precomputed, human-verified set of appropriate fields.

Large-scale duplication and deletion events are easily handled by the K-BROWSER, but inversions impose certain additional requirements on track realignment. In particular, because the track realignment phase requires an orthology map, at most one duplication is categorized as an ortholog and hence placed in the map. A deletion, similarly, does not violate the invariant that every region map to its ortholog; in the worst case, some regions will not have any orthologs, which is perfectly sensible. It is worth noting, however, that regions without orthologs are assumed to have the trivial alignment, that is, the original, ungapped sequence, and are realigned accordingly. Given all of this, inversions also do not break any orthology map invariant. The peculiar characteristic of inversions, however, is that they require that at least one genome be represented on the negative strand. We found this particular method of display to be somewhat unintuitive, especially when it was the original region requested by the user for visualization, and therefore require that the user-requested region always be displayed on the positive strand. It turns out that it is impractical to dynamically flip strands during image generation, and hence, the K-BROWSER requires precomputed track database realignments on both strands.

Image Generation

The second and most critical component is image generation. This component takes as input a genomic region query, and produces an image for every corresponding region in the multiple alignment. The set of these images, as a whole, forms an intuitive

representation of the multiple alignment; as individual images, they are representations of the tracks on the corresponding genomes. For instance, one can see in Figure 1 that the K-BROWSER has produced three images for a query on human Chromosome 12.

The K-BROWSER image generation component builds an entirely new framework based on subroutines borrowed from the UCSC Genome Browser. In regard to the latter, the K-BROWSER borrows low-level database- and track-processing subroutines provided with UCSC Genome Browser (Kent et al. 2002). This is important as it essentially creates a logical layer of abstraction between the K-BROWSER and the track databases, which is made especially critical in light of the aforementioned heterogeneous and dynamic nature of the databases. As a result, the K-BROWSER image generation component requires almost no knowledge of the underlying tracks, and can hence provide support for almost any track also supported by the UCSC Genome Browser.

Our original contributions to the K-BROWSER image generation component consist of (1) a new high-level framework that extends UCSC Genome Browser functionality, and (2) special methods to efficiently represent and visualize multiple alignments.

The former extends the UCSC Genome Browser code to handle image production of regions that cross orthology set boundaries. Ultimately, this code appears to let UCSC build an independent image for each orthologous set and proportionately "stitch" them together afterward. In practice, however, the K-BROWSER does not generate independent images because it would be grossly inefficient and inelegant. Instead, it seamlessly integrates into lower-level data-processing UCSC code and is responsible for allocating the proportionate amount of space within an image and appropriately organizing UCSC data-processing function calls.

The necessary, complement functionality—the ability to produce images within an orthologous set—is supported by the original UCSC Genome Browser code and aforementioned special methods. This functionality is implemented in two phases: we first use UCSC Genome Browser code to paint in the realigned tracks, and then walk through the region again to paint in gaps. In particular, because the produced images represent alignment-coordinate regions and the realigned tracks are already in alignment coordinates, the unmodified UCSC Genome Browser can be naively used to produce track visualizations. Recalling, however, that the realignment phase "stretched" tracks over multiple alignment gaps, it is clear that the UCSC Genome Browser code will entirely ignore gaps and paint arbitrary features in their place. To this end, we implement special methods that iterate through the appropriate subset of the segment table and over-

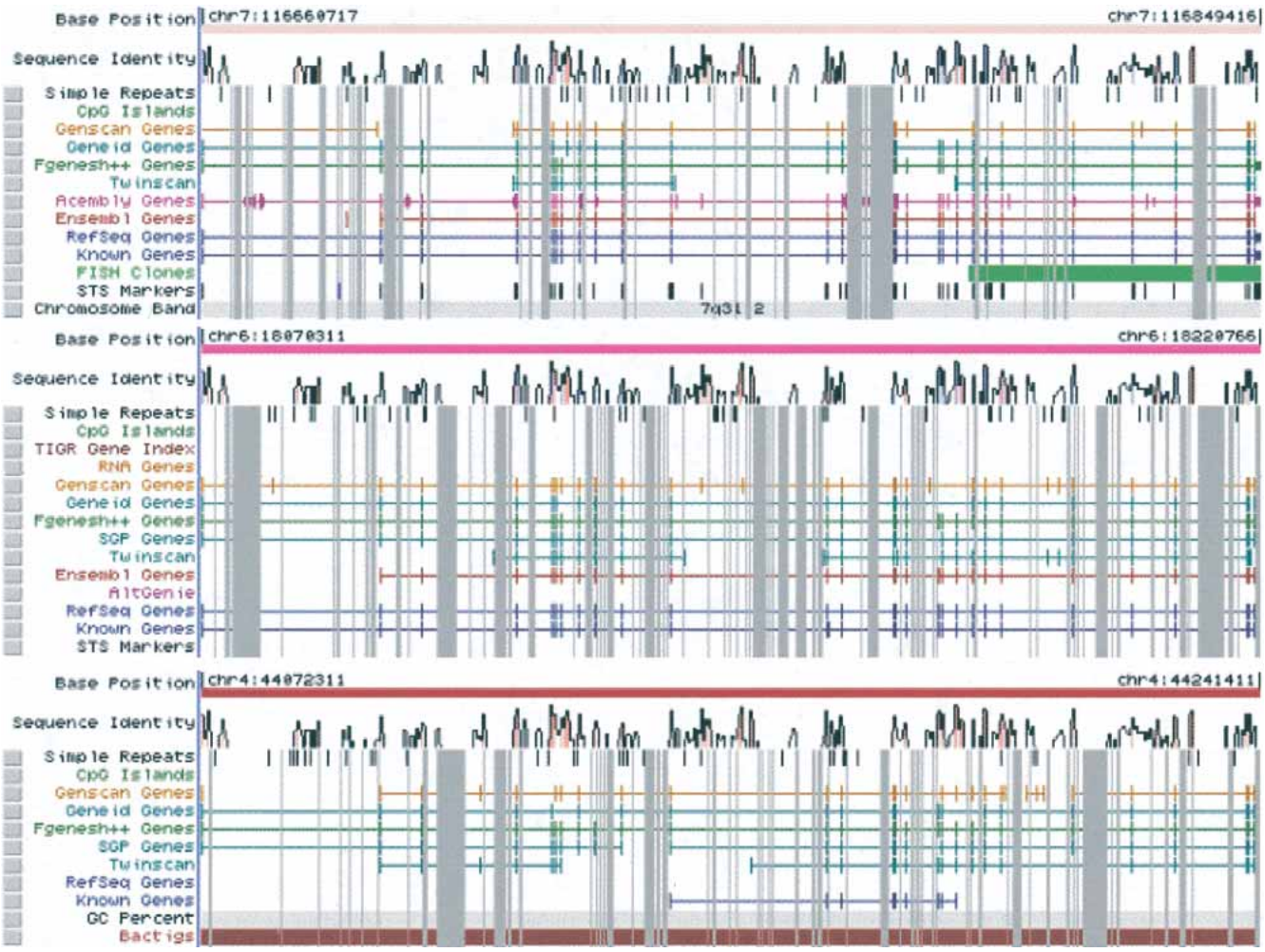

Figure 2 A K-BROWSER screenshot of the cystic fibrosis gene region (CFTR). (*Top* panel) Human annotations (April 2003); (*middle* panel) mouse annotations (February 2003); and (*bottom* panel) rat annotations (June 2003). The gray bars indicate gaps (arising from insertions or deletions) in the sequences. The alignment for this region can be downloaded in PHYLIP or multiFasta format, and it is possible to navigate using zoom buttons, gene name searching, and position jumping (buttons not shown).

write the regions of the image that correspond to multiple alignment gaps.

Conservation

The K-BROWSER has the ability to display a conservation plot above the tracks. With respect to Figure 1, this simply produces another track in the realigned track database. We have implemented two such tracks and can display (1) the percent identity in sliding windows along the sequences, or (2) a phylogenetic distance between a specific sequence and the root of the tree that relates the aligned sequences.

With regard to the identity plot, the K-BROWSER scores each position in the multiple alignment as the fraction of completely conserved columns in a window centered about that position. In addition, it allows the user to select a track according to which the conservation plot is to be colored, that is, blue for exonic regions, red for conserved noncoding regions, and so on. To this extent, it extends the useful identity plots on the Vista Genome Browser (Couronne et al. 2003).

Furthermore, given a phylogenetic tree and an evolutionary model, the K-BROWSER can also compute the average probability that the root sequence is different from the leaf sequence in a window centered about a specified position. This metric is meaningful as it allows one to not just determine if a genomic region is conserved with other genomes, but, in fact, to infer the rate at which it is evolving from the root. This score can be computed by inferring, for the aforementioned window, the distances from the root to each of the observed nucleotides. Roughly, these distances can be interpreted as the average number of mutations between the root and the leaf in a continuous-time Markov chain. As such, they can be exponentiated to determine the av-erage probability of mutation between the root and leaf nucleotides.

Applications

We conclude with a brief review of some of the possible applications of the K-BROWSER. Figure 2 shows a screenshot of the K-BROWSER in the cystic fibrosis (CFTR) gene region (Thomas et al. 2003). To find this gene in the browser, it suffices to type "cystic fibrosis" in the query box. The K-BROWSER returns two links, one for the human gene and the other for mouse (there is only a partial mRNA alignment in rat that is not detected as CFTR). Clicking on either link loads up the K-BROWSER. At this stage there is very little online computation thanks to the precomputed databases—only the regions and their tracks are pulled out from the database.

The K-BROWSER picture immediately reveals the complex insertion and deletion patterns in the region. It is obvious, for example, just by visual inspection, that there has been a large rodent insertion in the middle of the region. A closer inspection reveals that this large insertion is accompanied by several significant, simultaneous deletions and a very large number of small deletions in the rodents. Indeed, there are >3000 gaps in the rodent alignments, of which approximately two-thirds are <10 bp in length; in contrast, there are 2000 gaps in the human sequence, half of which are <10 bp.

The incomplete rat mRNA is also immediately obvious, thanks to the aligned ab initio predictions between the genomes. It is interesting to note that GENSCAN annotated the initial exon correctly only in mouse, and not a single ab initio method correctly annotated the gene in human (although SGP annotated the gene correctly in mouse). The conservation plot above the sequences instantly reveals conserved noncoding sequences (in red), and for viewing the mouse it is useful to switch the track base for the coloring. Zooming out (3×) reveals the larger-scale synteny in this region of Chromosome 7 in the human, and zooming out another 10× reveals that the synteny among Chromosomes 7 (human), 6 (mouse), and 4 (rat) is preserved throughout the entire region. The alignments for all of these regions can be retrieved, and are conveniently compressed for large regions.

Availability

The K-BROWSER can be accessed at http://hanuman.math.berkeley.edu/kbrowser/. The source code is available upon request.

ACKNOWLEDGMENTS

We thank Nicolas Bray and Colin Dewey for suggestions and help with the alignments. Yin Lau helped with the Web site design. L.P. was partially supported by the NIH (R02-HG02362-01), and K.C. was partially supported by a COR grant from UC Berkeley.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Alexandersson, M., Cawley, S., and Pachter, L. 2003. SLAM: Cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res.* **13:** 496–502.

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

Bray, N. and Pachter, L. 2004. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.* (this issue).

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

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen Y., Clark, L., Cox, T., Cuff, J., Curwen, V., et al. 2003. Ensembl 2002: Accommodating comparative genomics. *Nucleic Acids Res.* **31:** 38–42.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71–76.

Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L., and Dubchak, I. 2003. Strategies and tools for whole-genome alignments. *Genome Res.* **13:** 73–80.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., and Down, T. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30:** 38–41.

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

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–253.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The Human Genome Browser at UCSC. *Genome Res.* **12:** 996–1006.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17:** s140–s148.

McCutcheon, J.P. and Eddy, S.R. 2003. Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Res.* **31:** 4119–4128.

Parra, G., Agarwal, P., Abril, J.F., Wiehe, T., Fickett, J.W., and Guigó, R. 2003. Comparative gene prediction in human and mouse. *Genome Res.* **13:** 108–117. gene-centered resources. *Nucleic Acids Res.* **29:** 137–140.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., and McDowell, J.C. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.

WEB SITE REFERENCES

http://baboon.math.berkeley.edu/mavid/; MAVID multiple alignment server.
http://genome.ucsc.edu/; UCSC genome browser.
http://hanuman.math.berkeley.edu/kbrowser/; K-BROWSER home page.
http://pipeline.lbl.gov/vistabrowser/; VISTA genome browser.
http://www.ensembl.org/; ENSEMBL genome browser home page.

Received September 10, 2003; accepted in revised form November 17, 2003.

Visualization of Multiple Genome Annotations and Alignments With the K-BROWSER

Kushal Chakrabarti and Lior Pachter

Genome Res. 2004 14: 716-720
Access the most recent version at doi:10.1101/gr.1957004