doi:10.1006/geno.2002.7018

PipTools: A Computational Toolkit to Annotate and Analyze Pairwise Comparisons of Genomic Sequences

Laura Elnitski,[1,2,]* Cathy Riemer,[1] Hanna Petrykowska,[2] Liliana Florea,[3] Scott Schwartz,[1] Webb Miller,[1,4] and Ross Hardison[2]

[1]*Department of Computer Science and Engineering,*
[2]*Department of Biochemistry and Molecular Biology,*
[3]*Celera Genomics, Rockville, Maryland 20850 and*
[4]*Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802*

**To whom correspondence and reprint requests should be addressed at the Department of Computer Science and Engineering, 326 Pond Laboratory, The Pennsylvania State University, University Park, PA 16802. Fax: 814-865-3176. E-mail: elnitski@bio.cse.psu.edu.*

Sequence conservation between species is useful both for locating coding regions of genes and for identifying functional noncoding segments. Hence interspecies alignment of genomic sequences is an important computational technique. However, its utility is limited without extensive annotation. We describe a suite of software tools, PipTools, and related programs that facilitate the annotation of genes and putative regulatory elements in pairwise alignments. The alignment server *PipMaker* uses the output of these tools to display detailed information needed to interpret alignments. These programs are provided in a portable format for use on common desktop computers and both the toolkit and the *PipMaker* server can be found at our Web site (http://bio.cse.psu.edu/). We illustrate the utility of the toolkit using annotation of a pairwise comparison of the mouse MHC class II and class III regions with orthologous human sequences and subsequently identify conserved, noncoding sequences that are DNase I hypersensitive sites in chromatin of mouse cells.

Key Words: sequence alignment; computational biology; sequence analysis; DNA.

Introduction

Typically, an individual genomic sequence is annotated using *ab initio* gene prediction programs or by comparison of the sequence with databases of protein and EST sequences. Several programs or workbenches for assimilating these various annotations were reviewed and evaluated by Fortna and Gardiner [1]. In a complementary approach, functional regions such as genes and regulatory elements are identified using comparative alignments of either genomic DNA or translated genomic sequences from two or more species. This approach takes advantage of the fact that most protein-coding exons have a gap-free alignment exceeding 75% identity at both the nucleotide and the amino acid levels in human/mouse comparisons [2,3]. A high-scoring alignment that is not a known or predicted gene indicates that a noncoding feature of functional importance may be present. Criteria for selecting likely candidates for regulatory regions (such as those of [4,5]) thus far required a minimum percentage identity and length of sequence match. These thresholds successfully identified functional regulatory elements; however, strict adherence to these criteria will miss some important regulatory sites with lower levels of sequence conservation between humans and mice (e.g., hypersensitive sites 3 and 4 of the β-globin locus control region [6]). Therefore, effective identification of putative regulatory regions via sequence alignment requires a stringent nucleotide level alignment, complete annotation of coding regions of genes, and tools to identify and evaluate conserved regions.

Servers such as *PipMaker* [7] and *VISTA* [8] are available for aligning long genomic DNA sequences. The *PipMaker* server aligns genomic sequences rapidly and with high sensitivity and returns percentage identity plots (pips), other displays, and nucleotide level alignments. This paper describes a complementary toolkit called PipTools that facilitates the annotation of genomic sequences used in alignments, which is essential for the effective interpretation of those alignments. As an illustration, these tools were used to compile a pairwise alignment of fully annotated genes in the 1.5-Mb region of the mouse class II and III major histocompatibility (MHC) locus with an orthologous region in hu-

0888-7543/02 $35.00

TABLE 1: Tools for generating annotation files

Program	From	To	Comments
genbank2exons	GenBank entry	*exons* file	Converts GenBank annotations to an *exons* file
exons2mrna	*exons* file	mRNA sequence	Creates an mRNA sequence based on exon coordinates and a GenBank sequence
*sim4**	mRNA sequence	*exons* file	Aligns an mRNA sequence to genomic DNA and returns gene and exon coordinates
genbank2repeats	GenBank entry	*repeats* file	Returns coordinates of repeats in the GenBank file
exons2underlays	*exons* file	*underlay* file	Produces *underlays* for genes, exons, and UTRs using adjustable colors
rmask2repeats	RepeatMasker output	*repeats* file	Converts list of repeats to a simpler format
genscan2exons	Genscan output	*exons* file	Converts Genscan output to an *exons* file
genscan2underlays	Genscan output	*underlay* file	Converts Genscan output to an *underlay* file
shift-pos	Coordinate list	Coordinate list	Shifts all coordinates in a file by a given offset, or converts them to the coordinates of the reverse complement sequence
find-cpg	Sequence file	Coordinate list	Finds CpG islands in a sequence
reverse-comp	Sequence file	Sequence file	Returns the reverse complement of a sequence
mask-seq	Sequence file	Sequence file	Substitutes characters in a sequence file with lowercase letters or ambiguous characters such as N or X
*blest**	EST database	Coordinate list	Aligns a genomic sequence to an EST database and returns the coordinates of matching sequences
blest2exons	*blest* output	*exons* file	Converts the EST coordinates provided by *blest* to an ordered *exons* file
sort-exons	*exons* file	*exons* file	Sorts the genes and exons for use by *PipMaker*

* Available separately.

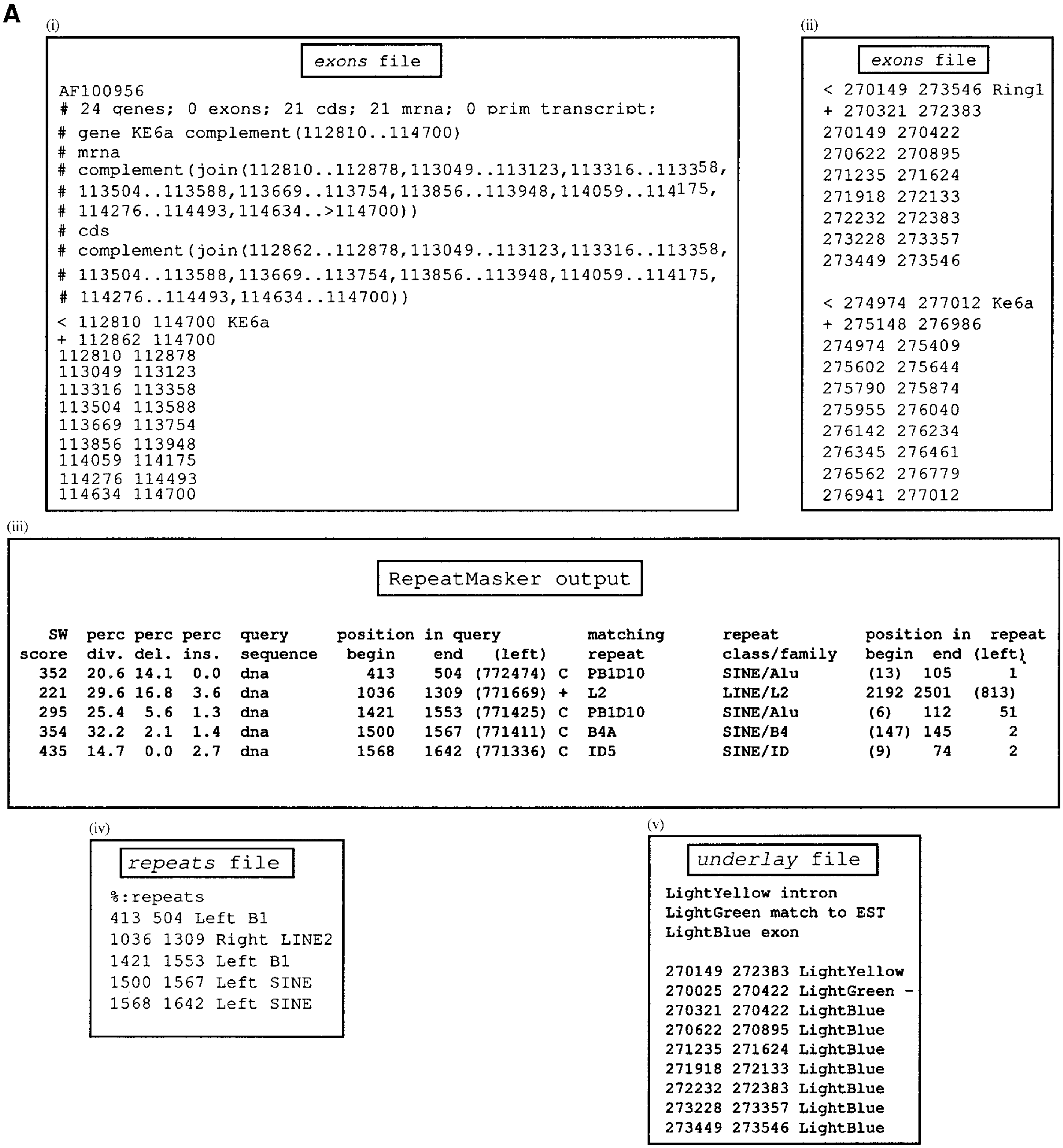
man, enabling a comprehensive search for putative regulatory elements that stand out above the background level of sequence conservation. We show that one highly conserved MHC region found between the mouse *H2-Ke4* and *Rxrb* genes is hypersensitive to DNase I in chromatin.

Summary of the Tools

An annotation file for a sequence can be derived from several starting points, and we created tools for converting existing annotations and the output of various programs for genomic sequence analysis into annotations for percentage identity plots (Table 1). For example, one set of tools can produce annotation files from a GenBank entry that contains coordinates for mRNA, coding sequences (CDSs), or repeats. Other tools convert the output from programs such as *Genscan* [9] for gene prediction, *sim4* [10] for matches to cDNAs, or *blest* [11] for matches to ESTs into the appropriate format and coordinate system for use with *PipMaker.* The PipTools are designed for sequential use, e.g., beginning with a GenBank file of sequence features and ending with annotation files used by *PipMaker.* These files are *exons,* for listing the name, directionality, and coordinates of genes with their individual exons and CDSs; *repeats,* containing the families and positions of interspersed repeats; and *underlays,* for color representation of sequence features in the alignment. Sample file formats are shown in Fig. 1A and described in detail at the *PipMaker* link on our Web site (http://bio.cse.psu.edu/) and in Elnitski *et al.* [12].

Tools for Annotating Genomic DNA Sequences Prior to Alignment

Annotation of a 1.5-Mb sequence contig from mouse MHC class II and III genomic DNA found on chromosome 17 illustrates the use of the toolkit prior to the sequence alignment step. This first step of annotating features in the mouse genomic sequence will be followed by subsequent steps of aligning the mouse sequence with the orthologous regions of human chromosome 6 using *PipMaker* and identifying conserved noncoding sequences in the alignment. These regions are found using postalignment tools for analyzing genomic comparisons. The mouse sequence is the reference (or first) sequence submitted to *PipMaker* with annotations for genes, repetitive elements, and color underlays. The human sequence is submitted as the second sequence without annotations. The purpose of each tool is described in Table 1, relationships between them are shown in Fig. 1B,


A
(i)
(ii)
(iii)
(iv)
(v)

```
exons file

AF100956
# 24 genes; 0 exons; 21 cds; 21 mrna; 0 prim transcript;
# gene KE6a complement(112810..114700)
# mrna
# complement(join(112810..112878,113049..113123,113316..113358,
# 113504..113588,113669..113754,113856..113948,114059..114175,
# 114276..114493,114634..>114700))
# cds
# complement(join(112862..112878,113049..113123,113316..113358,
# 113504..113588,113669..113754,113856..113948,114059..114175,
# 114276..114493,114634..114700))
< 112810 114700 KE6a
+ 112862 114700
112810 112878
113049 113123
113316 113358
113504 113588
113669 113754
113856 113948
114059 114175
114276 114493
114634 114700
```

```
exons file

< 270149 273546 Ring1
+ 270321 272383
270149 270422
270622 270895
271235 271624
271918 272133
272232 272383
273228 273357
273449 273546

< 274974 277012 Ke6a
+ 275148 276986
274974 275409
275602 275644
275790 275874
275955 276040
276142 276234
276345 276461
276562 276779
276941 277012
```

```
RepeatMasker output

   SW  perc perc perc  query     position in query        matching     repeat          position in  repeat
score  div. del. ins.  sequence   begin    end   (left)   repeat       class/family    begin  end (left)
  352  20.6 14.1  0.0  dna          413    504 (772474) C PB1D10       SINE/Alu        (13)  105      1
  221  29.6 16.8  3.6  dna         1036   1309 (771669) + L2           LINE/L2         2192 2501  (813)
  295  25.4  5.6  1.3  dna         1421   1553 (771425) C PB1D10       SINE/Alu        (6)   112     51
  354  32.2  2.1  1.4  dna         1500   1567 (771411) C B4A          SINE/B4         (147) 145      2
  435  14.7  0.0  2.7  dna         1568   1642 (771336) C ID5          SINE/ID         (9)    74      2
```

```
repeats file

%:repeats
413 504 Left B1
1036 1309 Right LINE2
1421 1553 Left B1
1500 1567 Left SINE
1568 1642 Left SINE
```

```
underlay file

LightYellow intron
LightGreen match to EST
LightBlue exon

270149 272383 LightYellow
270025 270422 LightGreen -
270321 270422 LightBlue
270622 270895 LightBlue
271235 271624 LightBlue
271918 272133 LightBlue
272232 272383 LightBlue
273228 273357 LightBlue
273449 273546 LightBlue
```

FIG. 1. Illustration of sample file formats and the relationships between complementary PipTools programs. (A) Sample files for *exons, repeats,* and *underlays* that are used in concert with the *PipMaker* alignment program. (*i*) An *exons* file generated by the *genbank2exons* program with embedded information from the GenBank file is illustrated. (*ii*) An *exons* file that results from a *sim4* comparison of mRNA to genomic DNA is shown. Note that the *H2-Ke6a* coordinates in these two panels differ because they are specific for the BAC sequence (*i*) and the full-length genomic contig (*ii*). (*iii* and *iv*) Acceptable *repeats* file formats generated by the RepeatMasker server and the *rmask2repeats* program, respectively, are shown. An *underlay* file is shown in (*v*). (B) Programs in the PipTools package that can be used to generate the three types of annotation files depicted in A: *repeats, exons,* and *underlay* files. Tools used in series are shown as intermediate stages along the path to the final annotation file. Sequences or file types that serve as starting points in any path are shown on the left in gray.

doi:10.1006/geno.2002.7018

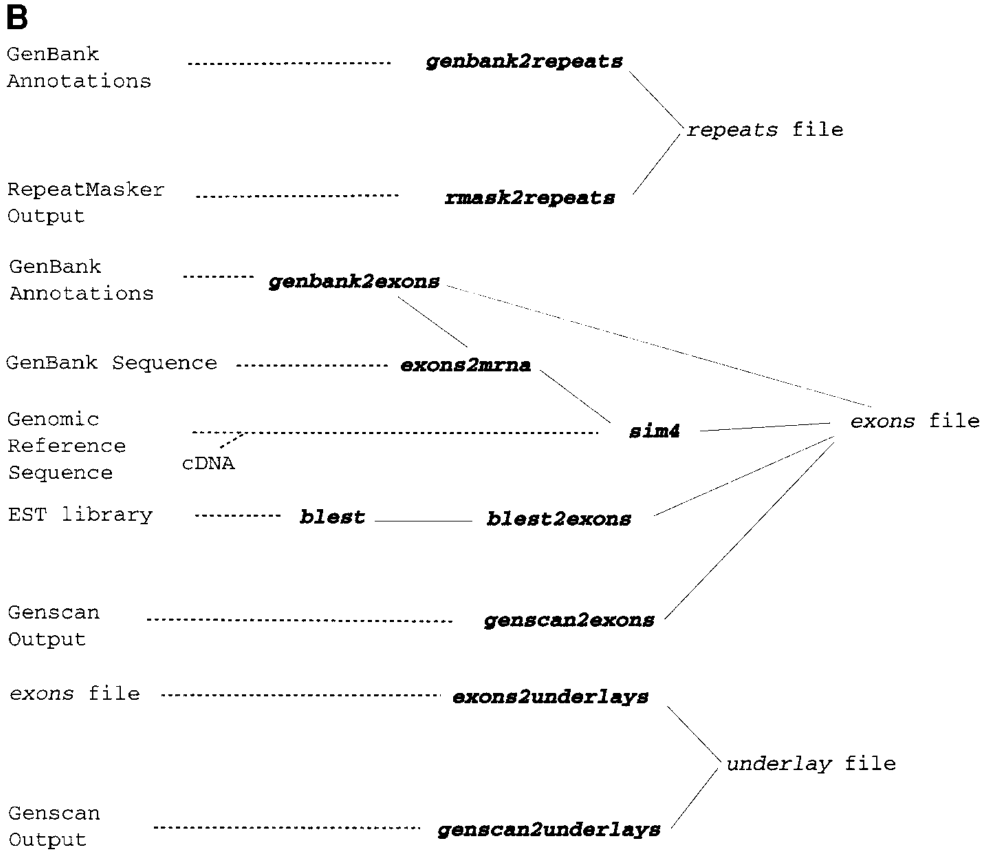


FIG. 1. Continued

and both are illustrated in more detail at the supplemental Web site (available through the *PipMaker* link at http://bio.cse.psu.edu/) and in Elnitski *et al.* [12].

Annotating Exons in the Mouse Contig Sequence

The full-length mouse genomic DNA contig NT_002588 (version 2, International Mouse Genome Project) is available at the NCBI Entrez server (http://www.ncbi.nlm.nih.gov/). Annotations for this 1.5-Mb sequence originally consisted of the names of sequence-tagged sites and links to other databases such as LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/) and the Mouse Genome Database at The Jackson Laboratory (http://www.informatics.jax.org/), but omitted essential information about the coordinates and names of transcription units in this region. More extensive annotations were found in a series of smaller, overlapping BAC clones of roughly 100–200 kb each (AF110520, AF100956, AF027865, AF050157, AF030001, AF049850, AF109906, AF109905, AF109719, and AC007080; unpublished entries submitted by L. Rowen *et al.*).[5] All of these BAC sequences were annotated with the names and coordinates of genes, mRNAs, CDSs, and repeats. Thus curated annotations of these sequence features already existed for transfer onto the corresponding positions in the 1.5-Mb sequence of the contig NT_002588 using the PipTools programs. It is anticipated that the annotations of the contig NT_002588 will be updated to include gene identity in the future. Nevertheless, the annotation process described for this example can be used for any genomic sequence.

To extract annotations from a GenBank BAC file, the program *genbank2exons* reads the coordinates of genes, mRNAs, and/or CDSs from the GenBank text-file format and converts them to the format of an *exons* file for *PipMaker*. The *exons* file directs *PipMaker* to label the name of each gene, specify the direction of transcription with an arrow, and draw and number the exons as boxes above the pip, as illustrated in Fig. 2.

In the example using the mouse MHC contig, annotations in the BAC-specific coordinates need to be transferred to the proper positions in the genomic contig NT_002588. There are several options for transferring annotations from a GenBank file. For example, an adjustment of the coordinates from one GenBank file to those of an overlapping sequence can be accomplished by a preliminary sequence alignment using the *PipMaker* server. Then, the program *transform-pos* (described in more detail below) uses the alignment information to convert the coordinates of a subsequence of DNA to a larger contig (in a simple case) or a different assembly of the sequence (where gaps or rearrangements may be present). For the example, each BAC sequence would be aligned to the genomic contig prior to conversion of the annotations to their coordinates in the larger sequence.

As an alternative to annotating the full-length genomic

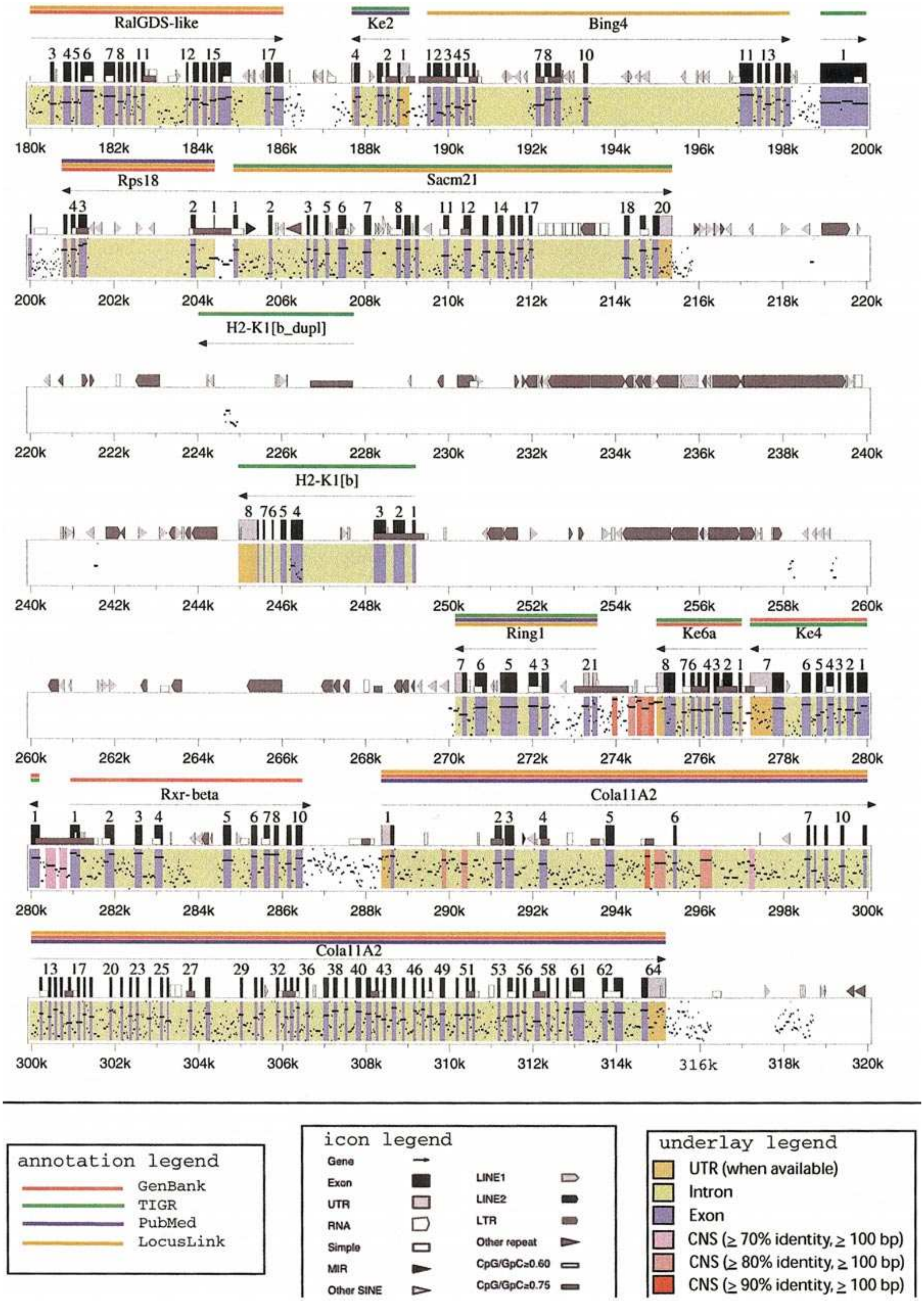


FIG. 2. Pairwise alignment and annotation of the mouse MHC region with orthologous human sequences. A 140-kb region (extracted from a larger alignment of 1.5 Mb) illustrates the use of annotation files to enhance an alignment. Icons on the top horizontal axis of the pip denote features in the mouse genomic DNA that were annotated using PipTools. Features in the pip are color-coded: blue (exons), orange (UTRs, when known), yellow (introns), red, and light red and pink (conserved noncoding sequences of ≥90, 80, or 70% identity, respectively, for a length of ≥100 bp). CpG islands are automatically calculated by *PipMaker* and displayed in the pip as short white boxes where the ratio CpG/GpC lies between 0.6 and 0.75 and as short dark gray boxes where the ratio CpG/GpC exceeds 0.75. Repeat families are represented by triangles and identified. Gene names and annotations for this mouse sequence are found in BAC file AF100956. The annotation legend shows colors of horizontal bars that serve as hyperlinks to other databases, as defined in the annotation file.

contig by transforming gene coordinates from each of the BAC sequences, aligning mRNA from each of the genes to the genomic sequence using a series of three PipTools programs can generate the annotations. As mentioned earlier, the first program, *genbank2exons,* extracts the gene coordinates in the form of an *exons* file. The program *exons2mrna* reads the *exons* file plus the corresponding FastA genomic sequence file, generates the putative mRNA sequence for each gene, and writes them to a single output file along with the GenBank file FastA header line. The mRNA sequences are aligned to the genomic sequence using a third program, *sim4,* which works in concert with the PipTools package, but is installed separately (available at http://bio.cse.psu.edu/). The program accepts a file containing multiple FastA sequences for comparison with a genomic sequence, such as are generated by the *exons2mrna* program. It also accepts and converts the coordinates of the coding sequences for inclusion in the *exons* file output. These features were added to the original *sim4* program, which is useful for aligning an individual mRNA or cDNA to a genomic sequence. The result of comparing the file of mRNA sequences with the genomic sequence produces an *exons* annotation file whose coordinates correspond to features in the genomic contig sequence.

In the event that a GenBank file does not contain annotations, candidate exons can be found by aligning the genomic sequence to ESTs or through *ab initio* prediction. The program *blest* (also installed separately from http://globin.cse.psu.edu/ftp/dist/blest) is a modified version of *sim4* that is used to find near-identity matches between genomic sequences and a database of ESTs (such as TIGR; http://www.tigr.org/tdb/mgi/searching/reports.html). The *blest* distribution includes a related program, *summarize* (http://globin.cse.psu.edu/ftp/dist/blest/blest_readme.html), which will sort the putative exons and introns and point out inconsistencies in the data, such as overlapping exon and intron coordinates. One of the PipTools programs, *blest2exons,* creates an *exons* annotation file from the aligned ESTs in the *blest* output. Since EST clusters do not always contain a full gene, the results from *blest* serve as a first step in the annotation process to identify the position of genes within a genomic sequence. This process is especially useful for large, unannotated genomic sequences. In subsequent steps a user can make the assignments more precise by aligning curated cDNAs obtained from the Entrez database (http://www.ncbi.nlm.nih.gov/Entrez/) using the program *sim4.*

The *ab initio* gene prediction programs offer an alternative route to gene identification. One such program that can handle large genomic sequences is *Genscan.* The PipTools program called *genscan2exons* converts output from a *Genscan* analysis into an *exons* file.

Finding and Masking Repetitive DNA

Repeats are another feature commonly annotated within a genomic sequence. Information about the positions of repeats within a sequence can be generated using the *RepeatMasker* server (A. F. A. Smit and P. Green, unpublished resource) or extracted directly from a GenBank entry to make a *repeats* file for *PipMaker* using the program *genbank2repeats.* Both types of files direct *PipMaker* to mask repeats during the alignment process and to draw distinctive icons for each class of repeats on the top horizontal axis of the pip (Fig. 2). A user also can convert the text-based *RepeatMasker* output to a simplified *repeats* file using the program *rmask2repeats.* This is required for users of the *Laj* viewer, a Java applet for viewing and manipulating pairwise alignments ([13]; available at http://bio.cse.psu.edu/), which does not accept *RepeatMasker* format. This simplified format also allows manipulation of the *repeats* file by other programs in the PipTools package that work on lists of coordinates (e.g., *shift-pos*).

Utility of Other Prealignment Tools

Regulatory elements often contain CpG islands that are found at the 5′ ends of genes. The program *find-cpg* can be used to generate a list of all CpG islands in a genomic sequence (based on calculations of %G+C and the ratio of CpG to GpC [14]) and returns the information as a list of coordinates. *PipMaker* automatically identifies CpG islands within the first sequence and designates their position by short rectangles along the horizontal axis of the pip (Fig. 2). The *find-cpg* program provides the coordinates of CpG islands for use with the *Laj* viewer.

Color underlays may be used to add clarity to the pips and dot plots generated by *PipMaker* (Fig. 2). The program *exons2underlays* automatically converts an *exons* file to an *underlay* file. The underlay file specifies coordinates and colors to be used on the pip to indicate features from the exons file. It also has an option to color only the top half of the pip for features in the forward strand and the bottom half for features in the reverse strand. Analogous to the *exons2underlays* program, *genscan2underlays* generates an underlay file directly from a *Genscan* output file.

Once the desired annotation files for exons, repeats, and underlays are completed, the genomic sequence is aligned with another DNA sequence using *PipMaker.* As an example, a 140-kb subregion of the mouse MHC comparison is shown as a pairwise alignment (Fig. 2) with annotations generated from many of the resources described above.

Postalignment Programs for Data Manipulation and Evaluation

One common need is to convert coordinates of features such as exons in one sequence to the corresponding coordinates in the second sequence. Two PipTools, *where-hit* and *transform-pos,* make use of *PipMaker* output files in either the *concise* or the *lav* format (i.e., raw *Blastz* output [13]). When given a position in the first sequence, the program *where-hit* returns the corresponding position in the second sequence, which may be in a multiple-contig,

TABLE 2: Tools for manipulating and analyzing pairwise alignment data

Program	Comments
where-hit	Translates a specified position in one sequence to the corresponding position in the other sequence
transform-pos	Translates a file of address ranges in one sequence to the corresponding ranges in the other sequence
strong-hits	Extracts matches from an alignment according to user specified criteria for percentage identity, length of sequence match, gap size, and identity differences across gaps

working-draft form. The program returns each range of matching coordinates with a header line to indicate the origin of the match. An automated method of converting all sequence coordinates in an annotation file to their corresponding positions in an aligning sequence is provided by the tool *transform-pos.* The aligning sequence may be from another species or an updated version of the first sequence in which the gene positions have changed. This program reads a file containing position intervals, which are specified with respect to one sequence, and converts them to the corresponding positions in a second sequence based on the alignments of the two sequences. If a position aligns with several regions in the other sequence, then only the first one encountered is reported; thus it is best if the alignment is produced using *PipMaker's* "single coverage" or "chaining" options. Both *where-hit* and *transform-pos* have an *inverse* option that allows a user to specify coordinates in the second sequence and obtain the corresponding positions in the first sequence.

A major goal of many comparative genomic analyses is the identification of highly conserved noncoding sequences (CNSs). This requires the prior annotation of exons, since coding exons are also highly conserved. The *strong-hits* tool (Table 2) identifies conserved noncoding regions, i.e., conserved sites that do not overlap coding regions defined in a given *exons* file. This program uses a *concise* or *lav* alignment file from *PipMaker* in conjunction with a user-specified criterion for stringency, e.g., ≥70% identity and ≥100 bp in length. Additionally, the user can specify the allowed length of each gap in a conserved region and the difference in percentage identity across the gap, e.g., maximum gap size of 2 bp, maximum step size of 5%.

Experimental Analysis of Conserved Noncoding Sequences in Pairwise Alignments

Blastz computed an alignment (Fig. 2) between the mouse genomic contig NT_002588 and the orthologous human sequence from the complete MHC assembled at the Sanger Institute (October 1999 version; http://www.sanger.ac.uk/HGP/Chr6/MHC.shtml). The program *strong-hits* identified CNSs in the alignment file that met the criteria of ≥70% identity and ≥100 nt in length with no gaps. The pip shows these elements with a color underlay that corresponds to their percentage identity (90, 80, or 70%) colored red, light red, or pink, respectively. Over 200 candidate regions were found within the 1.5-Mb mouse sequence, located both within and between transcription units. Furthermore, several of these sites are proximal to the start site of transcription for characterized genes, suggesting a role as putative regulatory elements.

It is possible that newly characterized genes are absent from the GenBank annotations for the mouse MHC sequence and could account for some of these apparent CNSs. We tested the possibility that the CNSs align to transcripts appearing in the EST database at NCBI (http://www.ncbi.nlm.nih.gov/dbEST/) or the TIGR EST cluster database. However, *Blast* searches of these databases failed to find any expressed sequences that mapped to the conserved regions of either the mouse or the human genomic DNA. Furthermore, the programs *Genscan* and *MZEF* [15] showed that the conserved regions do not contain open reading frames and are unlikely to be coding exons. We verified that the programs correctly identified exons in the annotated genes (data not shown). Thus a computational analysis of the conserved elements supports the conclusion that they are not protein-coding exons.

DNase I hypersensitive sites (HSs) frequently correspond to the sites of regulatory elements [16], and we chose an 18-kb region (located in the mouse sequence at positions 272–290 kb) containing several CNSs to test experimentally for HSs. One cluster of CNSs is found at position 280–281 kb, between the divergently transcribed genes *H2-Ke4* and *Rxrb,* and in a CpG island. Another cluster of CNSs falls between positions 274 and 275 kb of the alignment, 5′ of *Ring1* and 3′ of *H2-Ke6a* [17]. These two genes are transcribed in the same direction and, along with *Rxrb,* are expressed in multiple tissues, including blood (Unigene cDNA library report, http://www.ncbi.nlm.nih.gov/UniGene/). Thus the position of the CNSs between *H2-Ke6a* and *Ring1* suggests that they may play a role in the regulated expression of these genes.

We tested for DNase I hypersensitive sites corresponding to the location of these CNSs in the genomic DNA of a mouse erythroleukemia cell line, MEL [18]. Following treatment with increasing amounts of DNase I, DNA was prepared from nuclei, digested with *Bgl*II, and electrophoresed. The resulting Southern blots were hybridized with a probe specific for the left end of the 18-kb *Bgl*II fragment, revealing a strong DNase I hypersensitive site between *Rxrb* and *H2-Ke4.* This corresponds to the cluster of CNSs between 280 and 281 kb (Fig. 3). The same probe showed weak cutting by DNase I 3′ of *H2-Ke6a,* in the

doi:10.1006/geno.2002.7018

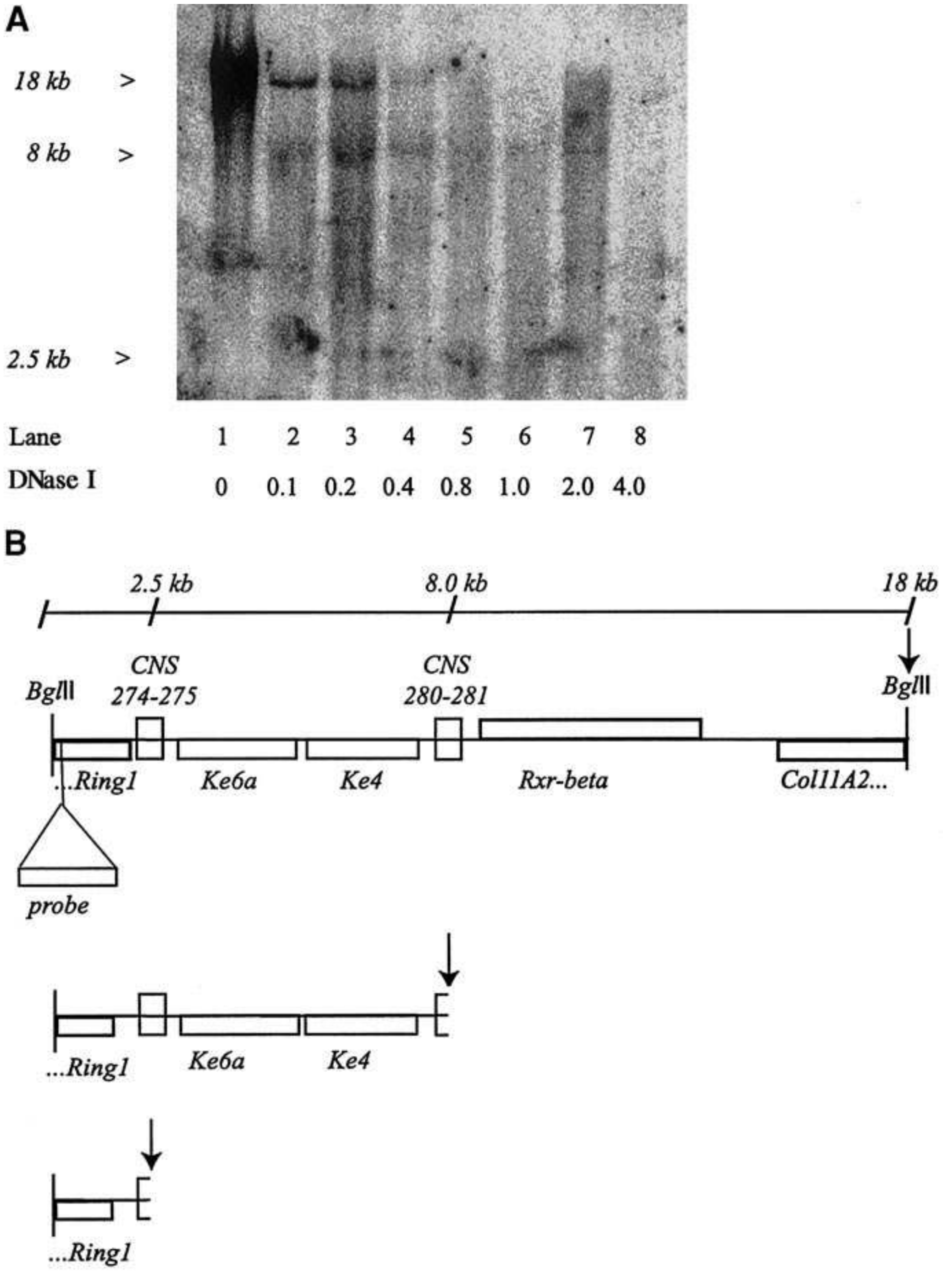


FIG. 3. (A) DNase I sensitivity of a conserved region in the MHC locus. Indirect end labeling of genomic DNA from a MEL cell line indicates hypersensitivity to DNase I at two positions that correspond to the locations of conserved noncoding regions in the genomic sequence alignment. The site of cleavage was calculated for each band based on the distance migrated by a ladder of DNA fragments of known molecular weight. Samples loaded onto each lane include MEL genomic DNA with no DNase I (lane 1) and 0.1, 0.2, 0.4, and 0.8 μg/ml DNase I (lanes 2–5, respectively). (B) The restriction fragments of the genomic DNA are shown with respect to the hybridization site of the probe. The parental fragment of 18 kb is reduced to 8 kb upon digestion of the HS at position 280 kb and further reduced to 2 kb upon cleavage at position 274 kb.

region corresponding to the cluster of CNSs between 274 and 275 kb.

Supplemental Information on the PipTools Package

The PipTools are simple, independent programs designed to run on a desktop computer. They are distributed as source code under the GNU Public License and can be modified (please see the "README" and "COPYING" files in the distribution package for details). Most of the tools are written in C, requiring an ANSI C compiler to install them. A few are written in Perl, requiring a Perl interpreter. Both C and Perl should be available for all common computing platforms. An all-Perl version of the PipTools is also provided as an alternative. These distribution packages are available for download as compressed zip archives and contain several types of documentation in HTML format. The page called "What's New" provides a summary of the changes in each released version of the PipTools and "Installation" has instructions on installing the programs. The "Program Reference" provides a detailed description of each program, what it does and how to use it, while "File Formats" describes the various formats for input and output files. Finally, a page of "Tips and Examples" explains how to use the tools from a biologist's perspective, including recipes for accomplishing common tasks.

Discussion

Gene annotations and predictions are now available for the human genome via the Human Genome Browser (http://genome.ucsc.edu/index.html), but central repositories for annotations of other species are currently unavailable. We describe a suite of tools that allows a user to annotate any genomic sequence that is aligned by the *PipMaker* server. These annotations can be generated from various sources including GenBank files and *Genscan* output. Additionally we describe tools that, when used in combination, provide an efficient means of annotating long genomic sequences for an alignment. Furthermore, they aid in the evaluation of aligned sequences by finding noncoding regions of high conservation or the corresponding position of a feature in the orthologous sequence.

The utility of the tools is illustrated in the annotation and analysis of a 1.5-Mb genomic sequence from the mouse MHC class II and III locus that is aligned by the *PipMaker* server. Annotation of the genes in this region reveals a number of unannotated, conserved, noncoding sequences that are putative regulatory elements. Two clusters of CNSs were examined for evidence of function. One cluster is located between divergently transcribed genes *H2-Ke4* and *Rxrb* and the other is at the 3′ end of the *H2-Ke6a* gene, preceding the *Ring1* transcription unit. Sequences between divergently transcribed genes are expected to contain promoters and possibly enhancers, and of course they can contain silencers or other negative regulators as well. Such positive and negative elements may be interspersed with each other, and they may be dependent upon integration in a chromosome at a particular site for full function. Hence a full analysis of these putative regulatory elements is a major project. We decided to test initially for a physical characteristic common to many regulatory elements (both positive and negative), i.e., the formation of DNase hypersensitive sites in nuclei and chromatin. The genes around the clusters of CNSs are expressed in several tissues, including blood cells. We showed that a strong hypersensitive site is located between the *H2-Ke4* and the *Rxrb* genes in nuclei from murine erythroleukemia cells. This coincides with one cluster of CNSs. It also lies within a large CpG island, which may also contribute to the sensitivity to DNase I. However, other CpG islands located in the same DNA fragment do not form strong hypersensitive sites, arguing that the CNSs are also needed for hypersensitive site formation.

A second cluster of CNSs located between *H2-Ke6a* and *Ring1,* which are transcribed in the same direction, also forms a hypersensitive site in nuclei from murine erythroleukemia cells. However, this site is not as accessible as that seen between *H2-Ke4* and *Rxrb,* despite the facts that the CNSs between *H2-Ke6a* and *Ring1* have a higher percentage identity and are also located within and close to CpG islands. One might expect that the sequence 3′ of *H2-Ke6a* (located to the left of it in Fig. 2) serves as a strong terminator of transcription, preventing transcriptional interference with the *Ring1* gene. It is possible that this could be accomplished by packing these sequences into a chromatin structure that is less accessible than the chromatin containing a promoter or enhancer. Information on the level of expression of these genes is not available, but this could have a strong effect on the degree of sensitivity to DNase I. It is difficult to design a single assay to test for a wide array of possible functions in gene regulation. A search for DNase I hypersensitive sites may find more types of regulatory elements than routine cell-transfection experiments, but it will not necessarily find all regulatory sequences.

Extensive annotation of the features within a genomic region increases the accuracy of identifying regulatory elements based on sequence conservation. Furthermore, easy manipulation of annotation files will become increasingly valuable as multispecies sequence alignments become more common. Thus far, the PipTools package is geared toward using annotations that are provided in GenBank or generated *de novo.* However, a similar series of tools is in development for use with genomic annotations from the Human Genome Browser. As the assembly of the human genome matures, this repository will become a primary source of genomic sequence and annotation.

We are currently able to identify putative regulatory elements that stand out as having a high level of sequence conservation. Tools such as *strong-hits* facilitate the analysis of alignments and selection of putative regulatory regions for further bioinformatic examination. The tools package was developed for use in individual laboratories that want to maintain local versions of annotated alignments for their chosen genomic loci. This work should complement the efforts of centralized, genome-wide repositories of ordered genomic fragments and their predicted coding regions. As more genomic sequences are assembled and made available, new genes or regulatory elements will need to be annotated for use in comparative analyses. Many of the tools described here will expedite this process.

Materials and Methods

Genomic annotations and alignment. Annotations for the GenBank files describing BAC sequences AF110520, AF100956, AF027865, AF050157, AF030001, AF049850, AF109906, AF109905, AF109719, and AC007080 were downloaded in their GenBank text format and extracted using the tool *genbank2exons.* An exons file was created for the genomic contig NT_002588

doi:10.1006/geno.2002.7018

by aligning it (in FastA format) to the sequence in each BAC file using the *PipMaker* server and then applying the tool *transform-pos* to transfer the coordinates of exons in sequence 2 (each BAC) to the aligning positions in sequence 1 (NT_002588). The contig sequence was submitted to RepeatMasker for an analysis of the repeats. An underlay file was created using the program *exons2underlays*. An alignment of the mouse MHC region with the orthologous region from human was generated by submitting the mouse sequence NT_002588 to the *PipMaker* server, along with the *exons*, *underlays*, and *repeats* files. The full human MHC consensus from the Sanger Center (October 1999 version) was submitted to *PipMaker* as sequence 2. Strongly conserved regions were identified by running the program *strong-hits* on the concise output from the alignment.

Determination of DNase I hypersensitive sites. Nuclei from MEL cells were isolated as described in Dhar *et al.* [19]. Nuclei were treated with increasing amounts of DNase I for 10 min at 37°C. DNA was isolated from nuclease-treated cells and Southern blots were conducted as described in Molete *et al.* [20]), except that the enzyme *Bgl*II was used in the restriction digestion. The ^{32}P-labeled probe that hybridized to the blot was synthesized as an oligonucleotide that corresponded to the nonrepetitive sequence at the left end of the 18-kb *Bgl*II fragment.

Acknowledgments

We thank Monique Rjinkels (Baylor College of Medicine) for careful testing of the programs and installation processes and Pallavi Eswara (The Pennsylvania State University) for critically reading the manuscript. This work was supported by NHGRI Grants HG02238 and HG02325 and NIDDK Grant DK27635.

References

1. Fortna, A., and Gardiner, K. (2001). Genomic sequence analysis tools: A user's guide. *Trends Genet.* **17:** 158–164.
2. Makalowski, W., Zhang, J., and Boguski, M. S. (1996). Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* **6:** 846–857.
3. Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B., and Lander, E. S. (2000). Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.* **10:** 950–958.
4. 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.
5. Mallon, A. M., *et al.* (2000). Comparative genome sequence analysis of the Bpa/Str region in mouse and man. *Genome Res.* **10:** 758–775.
6. Hardison, R. C. (2000). Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16:** 369–372.
7. Schwartz, S., *et al.* (2000). PipMaker—A Web server for aligning two genomic DNA sequences. *Genome Res.* **10:** 577–586.
8. Mayor, C., *et al.* (2000). VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046–1047.
9. Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.
10. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967–974.
11. Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J. Comp. Biol.* **7:** 203–214.
12. Elnitski, L., Riemer, C., Schwartz, S., Hardison, R., and Miller, W. (2002). PipMaker: A World Wide Web server for genomic sequence alignments. *Curr. Protocols* (in press).
13. Wilson, M. D., *et al.* (2001). Comparative analysis of the gene-dense ACHE/TFR2 region on human chromosome 7q22 with the orthologous region on mouse chromosome 5. *Nucleic Acids Res.* **29:** 1352–1365.
14. Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.* **196:** 261–282.
15. Zhang, M. Q. (1997). Identification of protein coding regions in the human genome based on quadratic discriminant analysis. *Proc. Natl. Acad. Sci. USA* **94:** 565–568.
16. Gross, D. S., and Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57:** 159–197.
17. Lovering, R., *et al.* (1993). Identification and preliminary characterization of a protein motif related to the zinc finger. *Proc. Natl. Acad. Sci. USA* **90:** 2112–2116.
18. Friend, C., Scher, W., Holland, J. G., and Sato, T. (1971). Hemoglobin synthesis in murine virus-induced leukemic cells in vitro: Stimulation of erythroid differentiation by dimethyl sulfoxide. *Proc. Natl. Acad. Sci. USA* **68:** 378–382.
19. Dhar, V., Nandi, A., Schildkraut, C. L., and Skoultchi, A. I. (1990). Erythroid-specific nuclease-hypersensitive sites flanking the human beta-globin domain. *Mol. Cell. Biol.* **10:** 4324–4333.
20. Molete, J. M., Petrykowska, H., Sigg, M., Miller, W., and Hardison, R. (2002). Functional and binding studies of HS3.2 of the beta-globin locus control region. *Gene* **283:** 185–197.

Sequence data from this article have been deposited with the GenBank Data Library under Nos. AF027865; 1997b, AF030001; 1998b, AF109906; 1998a, AF109905; 1998e, AF049850; 1998f, AF050157; 1998d, AF110520; 1998c, AF100956; 1999a, AC007080; 1999b, AF109719.