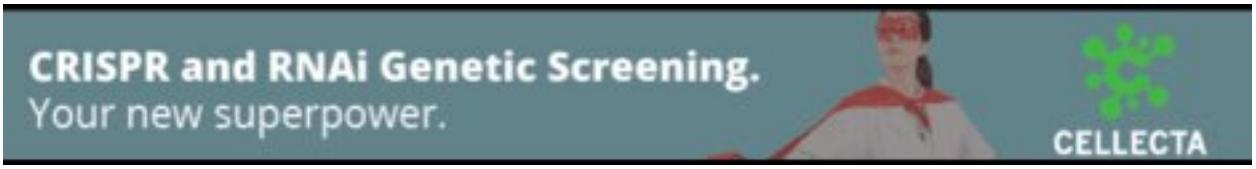
Comparing Vertebrate Whole-Genome Shotgun Reads to the Human Genome

Rui Chen, John B. Bouck, George M. Weinstock, et al.

Genome Res. 2001 11: 1807-1816
Access the most recent version at doi:10.1101/gr.203601



Letter

Comparing Vertebrate Whole-Genome Shotgun Reads to the Human Genome

Rui Chen,[2] John B. Bouck,[1] George M. Weinstock, and Richard A. Gibbs

Department of Molecular and Human Genetics, Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA

Multi-species sequence comparisons are a very efficient way to reveal conserved genes. Because sequence finishing is expensive and time consuming, many genome sequences are likely to stay incomplete. A challenge is to use these fragmented data for understanding the human genome. Methods for using cross-species whole-genome shotgun sequence (WGS) for genome annotation are described in this paper. About one-half million high-quality rat WGS reads (covering 7.5% of the rat genome) generated at the Baylor College of Medicine Human Genome Sequencing Center were compared with the human genome. Using computer-generated random reads as a negative control, a set of parameters was determined for reliable interpretation of BLAST search results. About 10% of the rat reads contain regions that are conserved in the human genomic sequence and about one-third of these include known gene-coding regions. Mapping the conserved regions to human chromosomes showed a 23-fold enrichment for coding regions compared with noncoding regions. This approach can also be applied to other mammalian genomes for gene finding. These data predicted ~42,500 genes in the human, slightly more than reported previously.

The draft sequence of the human genome provides a huge challenge of how to interpret its biological function (I.H.G.S. Consortium 2001; Venter et al. 2001). One of the most important and powerful methods for annotation is through comparative genomics. Pioneering studies in mouse–human comparisons show that both coding and regulatory gene regions can be identified through sequence conservation (Lichtarge et al. 1996; Ansari-Lari et al. 1997; O'Brien et al. 1999; Bouck et al. 2000b; Gelfand et al. 2000; Roest Crollius et al. 2000; Wasserman et al. 2000). The puffer fish, *Tetraodon nigroviridis*, also provides an excellent data set for comparison with the human (Roest Crollius et al. 2000). No single cross-species comparison can identify all elements of interest, and additional data from yet more related genomes are required. In addition, new tools and careful tuning of data search parameters are needed to speed annotation efforts.

A diverse collection of information from the human genome, including genomic sequence, transcripts, protein sequence, and gene function annotation has been stored in various databases (Box 1). One of the best sources for integrated human genomic sequence data is the Goldenpath database, in which human genome draft sequences have been ordered and mapped to individual chromosomes. Several databases have also been established for integrating data from human transcripts, such as RefSeq, the human transcript database (HTDB), and UniGene database (Bouck et al. 2000a; Pruitt and Maglott 2001). These genomic and transcript-based databases provide the primary resources for in silico exploration of the human genome.

In combination with sequence searching and gene modeling programs, these databases predict 30,000–40,000 human genes (Ewing and Green 2000; Roest Crollius et al. 2000; I.H.G.S. Consortium 2001; Venter et al. 2001). Less than half of these have been confirmed using rigorous methods, such as large-scale cDNA sequencing or RT-PCR of expressed sequences, therefore there is considerable interest in using newly available cross-species data for validation. Currently one of the fastest primary tools available for performing such large-scale genome comparison is BLAST (Altschul et al. 1997). The relationship between the statistics calculated by the BLAST program and biologically meaningful matches, however, is affected by many factors, such as the size of the database and the evolutionary distance between the two sequences. Furthermore, no theoretical proof has been found for calculating the statistical significance of gapped nucleotide sequence alignment. Therefore, empirical testing and careful tuning of BLAST searches, to establish parameters to distinguish real sequence matches from spurious alignments, is necessary.

In this paper, using rat whole-genome shotgun (WGS) reads, we determined a set of parameters suitable for a mammalian cross-species homology comparison. We also describe a method to identify low-frequency repetitive elements that can otherwise complicate cross-species searching. We conclude that, similar to the mouse, rat WGS reads can be used to analyze most of the genes that are conserved between human and rodents (Bouck et al. 2000b). The approaches can be applied to other mammalian genomes for gene finding. We estimate that there are ~42,500 genes in the human, slightly more than reported previously.

RESULTS

Determination of Parameters for Analyzing Similarity Search Results

The WGS reads used in the study were an initial rat WGS data set generated by the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC, URL http://

[1]Present address: Celltech Research and Development, Bothell, WA 98021, USA.
[2]Corresponding author.
E-MAIL ruichen@bcm.tmc.edu;FAX (713) 798-5741.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.203601.

Box 1. Databases Used in This Study and Other Relevant Sources

http://genome.ucsc.edu/
GoldenPath: Integrated human genomic database. Contains the assembly of the human draft sequence and other annotation information. Genieknown gene set can be found in the database file.

http://www.ncbi.nlm.nih.gov/HTGS/
Human High throughput Phase3(HS3). Contains all finished human sequences.

http://www.ncbi.nlm.nih.gov/UniGene
UniGene database: Transcript database. Contains non redundant set of gene clusters which contain both mRNA and EST clones. It also contains many clusters with EST clones only. HS3 is now at PRI division at NCBI.

http://www.hgsc.bcm.tmc.edu/HTDB
Human Transcript Database(HTDB): A collection of cDNA clone sequences. EST clones are not included.

http://www.ncbi.nlm.nih.gov/blast
BLAST (Basic Local Alignment Search Tool): A set of local similarity search programs with high speed. BLASTN is for DNA search and TBLASTX is for six-frame translation of DNA search.

http://ftp.genome.washington.edu/cgi-bin/RepeatMasker
RepeatMasker: A program that screens DNA sequences for interspersed repeats and low-complexity DNA sequences.

www.hgsc.bcm.tmc.edu; National Center for Biotechnology Information [NCBI], URL http://www.ncbi.nlm.nih.gov). To develop effective search parameters, a pilot set of ~20,000 rat WGS reads were first filtered from low-quality data and contaminants, such as mitochondria, *Escherichia coli*, and phage DNA sequences. Because the rat and human are known to share many common and relatively frequent repetitive elements, we also masked the rat reads for known mammalian repetitive elements using the RepeatMasker program (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker; A.F.A. Smit and Green, P., unpubl.). The resulting 11,027 filtered reads, with an average read length of 522 bp (phred value ≥20), were used for sequence similarity searches against the UniGene, HTDB, HS3 (representing ~1 GB of human finished genomic sequence), and the Goldenpath databases. To achieve a relative high sensitivity, the BLASTN program was used with default searching parameters (word size = 11, gap open cost = 5, gap extension cost = 2, mismatch penalty = −3; see Discussion for details of choosing BLAST parameters) (Altschul et al. 1997).

A control data set was generated in parallel using random sequences. We took into account rat sequence-specific features, including read length, GC content, and the position of repetitive elements. Random reads with a similar nucleotide composition, length, and masked regions to real reads were generated. This set of random reads was then used in the same battery of sequence similarity searches as the filtered reads to determine the significance of the search results.

Searching results were initially compared on the basis of the BLAST bit score parameter. The distribution of the data generated from the filtered reads and the random reads was very different. Figure 1 shows that the score distribution from the filtered reads against the UniGene database forms a smooth curve that begins at a very high value and gradually decreases to 60 bits. In contrast, the results from the random reads never have a score >60 bits and are generally <50 bits. These results indicate that a hit with a score >60 bits is a significant match.

Similarly, the distribution of the BLAST *E*-values and alignment lengths were compared between the filtered reads and the randomly generated reads (Table 1). A dramatic difference was seen when the *E*-value is $<10^{-5}$ and the match length is >50 bp. In addition to the UniGene comparison, the same searches were conducted with three other databases: HTDB, HS3, and Goldenpath. Similar results were obtained and we found that the same set of parameters could be used for searching each genomic and transcript database (data not shown).

Based on these results, we classified matches that exceeded the threshold values for all three of these search parameters as strong hits, whereas matches that fulfilled at least one criterion were weak hits. According to this standard, the searching of the Human Phase 3 database with 11,027 reads yielded 9.3% (1030/11,027) strong hits and 6.8% (751/11,027) weak hits. In contrast, no strong hits and only five weak hits were found using the same number of random reads. Similar results were observed using the other three databases (Table 2). Therefore, there is a >100-fold difference in frequency even in the weak hit category, showing that our classification schema, based on these control sequences and BLAST parameters, is highly discriminating.

Filtering Out Potentially Unidentified Repetitive Elements

A critical issue is to distinguish bona fide gene hits from matches to low-frequency repetitive elements. Although only

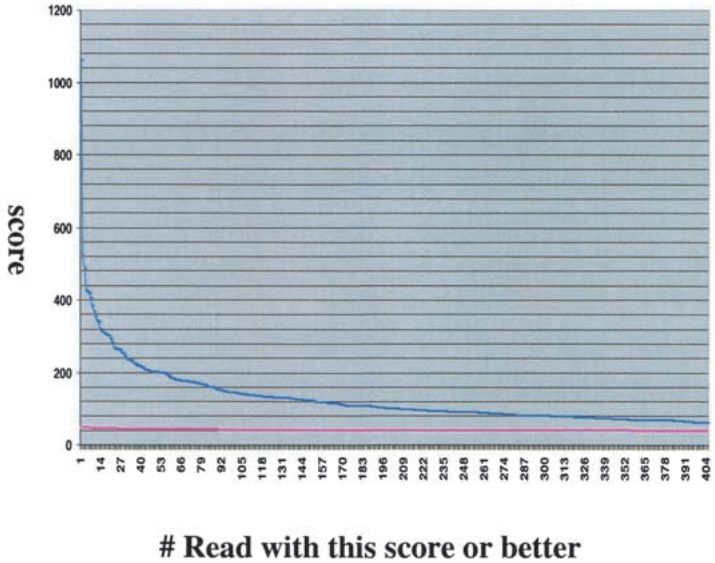


Figure 1 Comparison of the BLAST match score distribution of rat WGS reads versus random reads, and distribution of the top score of matches found when searching the UniGene database. WGS reads are in blue and random reads in red.

Table 1. Distribution of Search Results against the UniGene Database

Alignment length	WGS reads(%)	Random reads(%)	Score(bits)	WGS reads(%)	Random reads(%)	*E* value	WGS reads(%)	Random reads(%)
10–30	97.6	99.96	30–40	84.42	92.32	>1	84.37	92.28
30–50	0.96	0.04	40–50	13.85	7.68	0.1–1.0	10.96	6.87
>50	1.44	0	50–60	0.47	5.50E-03	0.01–0.1	2.2	0.71
			>60	1.26	0	0.001–0.01	0.74	0.13
						0.0001–0.001	0.27	5.50E-03
						<0.00001	1.06	0

Both the WGS-reads and random-reads results are shown and put side by side. A >100× difference is achieved with alignment length >50 bp, match score better than 60, or *E*-value <10-5.

a few percent of the WGS reads have frequent matches when searching against the genomic databases (Fig. 2), the large absolute number of hits generated by these reads will still produce a high background noise. Nevertheless, these reads should not be simply excluded from the analysis based on their hit frequency, as multiple hits from single sequence reads may represent interesting common domains shared by large groups of genes.

To investigate these events, we first grouped reads based on the number of hits found when searching the HS3 genomic database. A plot of the distribution of the number of matches clearly showed a graph that becomes flat after six hits per read, which is therefore used as our cut-off hit number (Fig. 2A). We identified a similar number of 12 hits for Goldenpath, five for HTDB, and three for UniGene. Therefore, we consider reads that have more than six matches in the HS3 database abundant reads. Based on these results, we divided all reads into three categories–reads containing unique elements, medium represented elements, and abundant elements. Reads that show only one hit are unique elements. Reads that have more than one match but less than the cut-off number are considered medium represented elements, and reads that have more than the cut-off number of hits are classified as abundant elements and very likely contain repetitive elements. Based on this classification, ~7.6% and 4.2% of total reads belong to the abundant category when searching against the Goldenpath and HS3 databases, respectively.

To verify the identity of repetitive elements in the abundant class, we tested if these elements also have matches in the transcription databases. We reasoned that reads containing protein family domains will also match many entries when searching transcriptional databases, whereas reads that are nontranscribed, genomic repetitive elements will probably only have a few or even no matched entries in the transcriptional databases, as such repetitive elements are more frequently outside the coding regions. The results of manual checking of matches to the transcriptional databases are shown in Table 3. Reads that have many hits in both genomic and transcriptional databases most frequently contain domains shared by large gene families. In contrast, reads that have many hits in the genomic database, but only a few in the transcript database, predominantly contain repetitive elements. Examples include known repetitive elements HSAG-1 middle repetitive genetic elements and Human L1 putative reverse transcriptase gene insertion in hamster, which are not in the database of frequent repetitive elements we first used in the `RepeatMasker`. As shown in Figure 2B, after elimination of reads that probably contain repetitive elements, there are 2423 and 985 unique reads when searching either the Goldenpath or the HS3 databases. Together there are 21% of the 11,027 reads with at least one match in either one of these two genomic databases. Similarly, 3.2% of the reads (355/11,027) have matches in either of the two transcriptional databases. Overall, the comparison between the genomic and transcript databases provides an excellent method to eliminate potential repetitive elements.

Comparison between Bulk Rat WGS Reads and Human Genome Databases

Using the parameters and methods defined above, a total of 450,928 filtered rat WGS reads were analyzed (Fig. 3). As shown in Table 4, ~10% of all the reads (45,343/450,928) matched with the Goldenpath database. These reads generated a total of 76,447 matches with an average of 1.7 matches per read. Only 4.2% of all the reads (18,875/450,928) have matches in the HS3 database (with an average of 1.4 matches per read), consistent with the fact that the HS3 database is 2.35× smaller than Goldenpath (1.14 GB vs. 2.68 GB). As expected, most reads that have matches in the HS3 database were also identified in Goldenpath. Only 1210 reads are positive in HS3 but not in Goldenpath, whereas 27,678 reads are unique to searching the Goldenpath database (Table 4). Because this result was obtained from a large and random data set, we conclude that the Goldenpath database is a more comprehensive source for human genomic sequences and searching of it alone is sufficient. Together we found 46,553 rat WGS reads that matched to the human genome, ~10.3% of all reads.

The results of searching the Human Transcript Databases have also been examined. As shown in Table 4, 1.8% (8146/450,928) and 2.8% (12,684/450,928) of all reads have matches in the HTDB and the UniGene databases, respectively. These matches account for 37.3% (5714/15,305) and 10% (8744/

Table 2. Number of Strong and Weak Matches Obtained Using Either the Rat WGS Reads or Random Reads

	Stringent condition		Weak condition	
	WGS reads	random	WGS reads	random
GoldenPath	3015	0	1671	2
HS3	1030	0	751	5
HTDB	414	0	292	0
UniGene	405	0	225	0

Using the stringent condition, no matches have been obtained from random reads. Using the weak condition, only a few matches have been obtained from the random reads.

86,918) of all the records in HTDB and UniGene, respectively. When we compare these two results, only 626 reads are unique in HTDB, whereas 5164 reads are unique in UniGene database, indicating HTDB is less comprehensive than UniGene. When combined, a total of 13,309 reads have at least one match in the transcript databases, ~2.95% of the total reads. Most of these reads also have matches in the genomic databases. Only 12% of the reads (1638 /13,309) have matches in the transcript database but not in the genomic databases. This result is consistent with the fact that ~10% of the human genome sequences have not been included in either of the genomic databases we used. We further examined records that match to many reads, to verify the removal of repetitive elements. As shown in the Table 5, these records often are members from large gene families, such as ribosome RNA genes, zinc finger proteins, olfactory receptor genes, and homeo-domain-containing genes.

In summary, although only a small percentage of the rat WGS reads contain sequences that are conserved in the human genome (10.3% of all reads), 30% (2.95%/10.3%) of these matches can be mapped to exons, despite the fact that only about half of the genes are believed to be represented in the transcript databases. Therefore, comparison between the rat WGS reads and the human genome is a very fast and effective method to enrich for exon segments and provides a useful route for gene discovery. To analyze the correlation between conserved regions and genes further, we examined the distribution of these matches on each human chromosome using information from the Genome Browser Database. As shown in Figure 4, the density of the matches on individual chromosomes are plotted. We found that the number of hits is proportional to the size and gene density of individual chromosomes.

Comparison Between Reads from Transcribed and Nontranscribed Regions

To examine the relationship between sequence conservation and gene structure further, we analyzed matches that fall within known genes using the Genie-known gene set in the Goldenpath database. A total of 8290 Genie-known genes were anchored on individual chromosomes and spanned

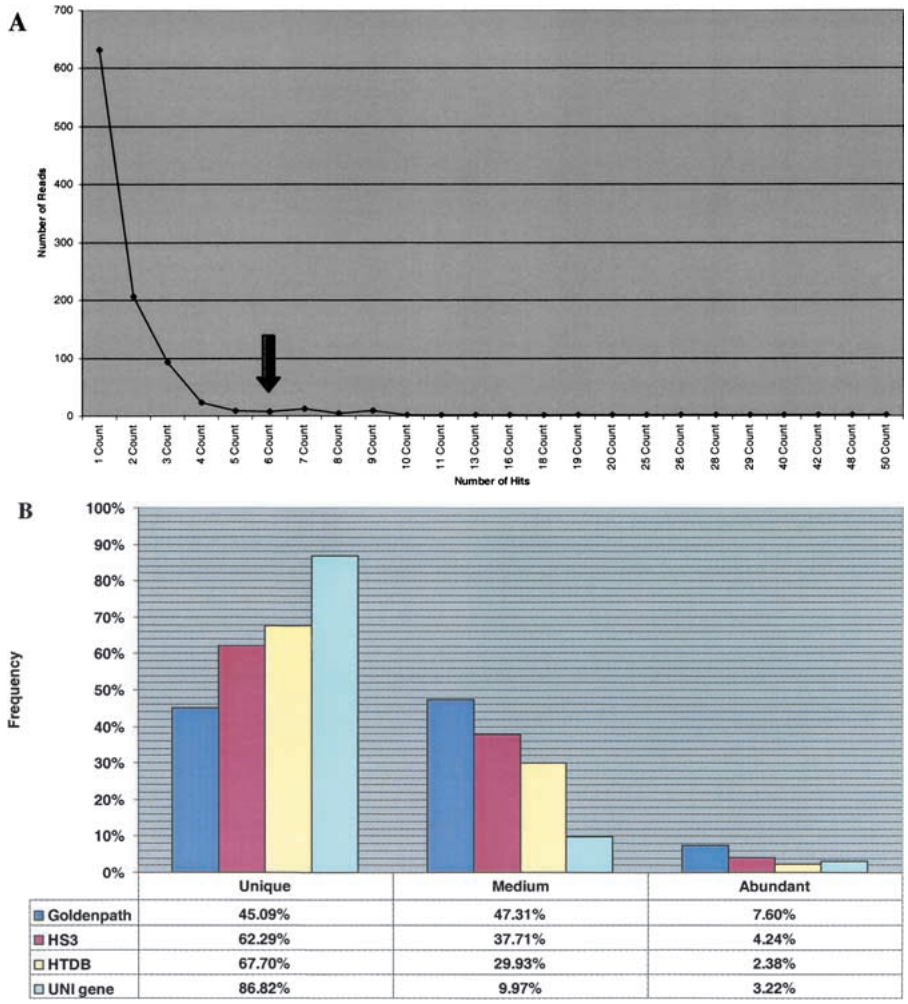


Figure 2 (*A*) Distribution of number of hits each read generated in searching of the HS3 database. The *abscissa* is the number of hits of one read to the HS3 database. The *ordinate* is the total number of reads that have corresponding number of hits. The curve drops fast initially and then becomes flat after six. (*B*) Distribution of number of matches each read obtained. Searches are against the Goldenpath, HS3, HTDB, and UniGene databases. Most of the reads have only one or a few matches.

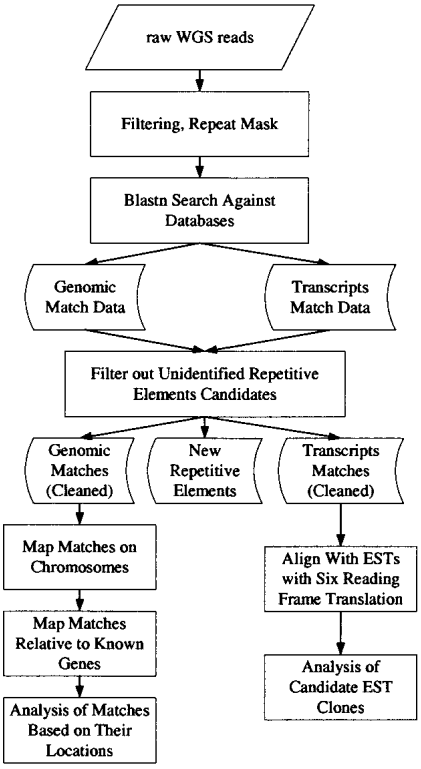


Figure 3 Flowchart of the overall procedure for analyzing conserved regions between rat and human. Raw WGS reads were filtered from low-quality data and contaminants, masked with known mammalian repetitive elements using the program RepeatMasker. The resulting sequences were used to search against several human genome databases. Potential repetitive elements were identified through comparison between genomic matches and transcripts matches. The final match results were analyzed further through integration with other information from the human genome.

~435 Mb. The end points and the intron/exon boundaries of each gene were determined by the alignment between the genomic sequence and the corresponding mRNA sequence. The total length of the 435-Mb region is divided into three categories: 11.9 Mb of coding exons (2.7%), 4.4 Mb of noncoding exons (1%), and 418 Mb of introns (96%), respectively. A total of 14,881 matches were mapped to these genes. Among them, 62% (9228/14,881) contain exons, 36.8% (5469/14,881) of the matches are exclusively in intron regions, and 1.2% (184/14,881) mapped to nontranslated exons. Considering that 62% of matches contain coding regions that only occupy 2.7% of the total sequence, this is a 23-fold enrichment of the translated region in sequences that are conserved between rat and human.

Once a conserved region is identified, it is important to determine whether it contains exons. Previous studies indicated that using appropriate combinations of searching parameters, it is possible to distinguish intron versus exon matches between the pufferfish and the human (Roest Crollius et al. 2000). We tried to test whether this is true for the rat and human comparison by comparing the match results obtained from reads that contain transcribed regions and those that only match to the intron regions of the known genes. Although this is not a perfect comparison, as the intronic regions of these genes may contain some unidentified exons, we were still able to detect some differences between these two sets of matches. Specifically, as show in Figure 5B, matches in the nontranscribed region tend to have a relatively low score with ~58% from 50–90 bits. In contrast, matches from the transcribed region tend to have higher score where only 37% are between 50–90 bits. The shift toward stronger matches is also very clear when we compare the length of these matches (Fig. 5A). Although 48% of the matches from the nontranscribed region is shorter than 80 bp, only 29% matches from the transcribed region are in the same category. These results indicate that conserved transcribed regions are more similar at the nucleic acid level to each other than the conserved nontranscribed regions. The difference, however, is probably not sufficient to distinguish these two types of sequences reliably.

It is likely that the conservation pattern will be different between the coding and noncoding regions in that the coding regions can still be translated into conserved proteins despite nucleic acid substitution. To test this hypothesis, using the TBLASTX program, we performed six frame translations of the matched reads and realigned them with the human genomic sequences. As shown in Figure 5C, although coding region and intronic alignments could not be separated completely, more stringent alignment was detected in coding regions compared with the nucleotide alignment discussed above. We found a clearly different distribution between the coding regions and the intronic regions when increasing the match stringency. For example, 60.3% of the exon matches have >30 identical amino acids, whereas only 35% of the intron matches have the same feature. Similarly, with the match stringency cutoff set at 100 bits, 34% of the exonic alignments

Table 3. List of Reads Having the Most Matches in Searches of Both the Genomic and the Transcript Databases

WGS read name	Genomic database match	Transcripts database match	Gene description
TUWAG1U1875.scf	116	93	Human zinc finger protein
TUWAG1U2516.scf	83	67	Kruppel related zinc finger protein
TUWAA1U36846.scf	95	33	Human zinc finger protein
TUWAA1U12396.scf	46	19	Human olfactory receptor protein
TUWAA1U16676.scf	22	10	Human olfactory receptor (OR17-4)
TUWAA1U30976.scf	155	9	Human ribosomal protein
TUWAA1U37056.scf	19	6	Human 18S rRNA gene
TUWAG1U0855.scf	71	5	Human ribosomal protein
TUWAA1U24176.scf	26	5	Human H3.3 histone class C

These reads all contain domains that are shared by large gene families.

and only 15% of the intronic matches meet this standard. Therefore, a further enrichment of exon matches could be achieved by using the TBLASTX search with the threshold value of 30 amino acids and a score >100 bits.

Overall, both nucleotide and amino acid comparison allowed some discrimination of introns from coding regions through the stringency of the search. Unlike the pufferfish comparison, however, it is difficult to completely distinguish exon versus intron matches solely based on the alignment between the rat and the human, becasue intron matches with very long alignment and high scores have been found in our data set. Nevertheless, by combining the nucleic acid and amino acid alignment and carefully choosing match parameters, some enrichment of exon matches could be obtained and up to 80% of strong matches containing exon regions was achieved.

Analysis of Matches Between the Rat WGS Reads and Human ESTs

The results presented above indicate that in addition to the sequence alignment, other information is required to distinguish coding from intron regions among the conserved segments. One valuable resource is human expressed sequence tag (EST) data that, despite contamination with genomic sequences, is a rich representation of expressed genes. Interestingly, we found that rat WGS/human EST matches in the UniGene database have a strong bias toward known genes compared to isolated ESTs. As shown in Table 6A, the known gene clusters average 13-fold more rat WGS read hits compared with the EST-only clusters (0.75 vs. 0.058 match per cluster). These data are consistent with the notion that EST clusters in the Unigene database are relatively gene-poor (Roest Crollius et al. 2000). In fact, of the total 8744 UniGene entries matched by the rat WGS reads, 54% (4715/8,744) of them are known genes.

Based on results shown above, matches with good alignment at both nucleic acid and amino acid levels are likely to represent exons. Therefore, we reasoned that ESTs likely to contain gene-coding regions could be identified using the rat and human sequence comparison. To test this hypothesis, those 4377 rat WGS reads that contained nucleotide matches with EST clusters in the UniGene database were searched against the Unigene database again using TBLASTX. As shown in Table 6B, we found that 57% of these reads still only match to EST clones after translation and we reasoned that these EST clones were likely to contain exons from unidentified genes. When we examined 10 randomly selected cases that have a bit score >90 (which accounts for 24% of this category), two cases mapped to genes that are not part of the collection in the UniGene database (data not shown). The other eight cases are very good candidates for new genes (Fig. 6). In fact, when we used these EST clones to search the nonredundant database, several types of evidence were found. First, homologs of some of these EST clones could be found in a third species, such as in mouse (Fig. 6A,B). Second, in some cases, putative protein homologies were identified in other species, such as *Drosophila* and *Caenorhabditis elegans* (Fig. 6B). Third, some of the EST clones could be mapped to a putative coding region predicted by the gene finding programs, such as Genie (data not shown). Therefore, comparison between the rat WGS reads with the human sequence data is potentially a very powerful way to identify EST clones that contain gene-coding regions.

Table 4. Search Results of all Four Databases

Database	Total	Unique	Shared
Goldenpath	45343	27678	17665
HS3	18875	1210	
UniGene	12783	5264	7519
HTDB	8145	626	
Genomic	46553	34882	11671
Transcripts	13309	1638	

The total number of reads matching individual databases is shown. Genomic database is the combination of the Goldenpath and HS3 database. The transcripts database is the combination of the UniGene and HTDB database. See text for details.

DISCUSSION

Statistical Significance of WGS Sequencing BLAST Search Results

To identify parameters that can be used to search human ge-

Table 5. List of Reads Having the Most Matches in Searches of the Transcripts Database

Gene ID	Gene description	No. of hits
gi|285928|dbj|D14718|HUMHMG1	Human chromosomal protein HMG1 related gene	32
gi|531475|emb|X80910|HSPPP1CB	PPP1CB gene; protein phosphatase 1.	34
gi|32326|emb|X12597|HSHMG1	Human mRNA for high mobility group-1 protein (HMG-1)	35
gi|968887|dbj|D63874|HUMFM1	*Homo sapiens* high-mobility group (nonhistone chromosomal) protein 1(HMG1).	36
gi|4519269|dbj|AB011414|AB011414	Human zinc finger protein ZNF136	36
gi|430993|gb|U02478|HSU02478	trithorax (*Drosophila*) homolog	36
gi|337494|gb|M36072|HUMRPL7A	Human ribosomal protein L7a (surf 3) large subunit mRNA	37
gi|2792017|emb|Y10530|HSHTPCR2	*H. sapiens* gene encoding putative olfactory receptor	39
gi|498735|emb|X78932|HSHZF9	Human repressor transcriptional factor (ZNF85) mRNA	39
gi|2565195|gb|AF000381|HSAF000381	Human 18S rRNA gene	39
gi|1017721|gb|U35376|HSU35376	Human repressor transcriptional factor (ZNF85) mRNA	42
gi|1262328|dbj|D42073|HUMRCN	Human mRNA for reticulocalbin, calcium binding domain	43
gi|487784|gb|U09367|HSU09367	Human zinc finger protein ZNF136	55
gi|337377|gb|K03432|HUMRGEA	Human 18S rRNA gene	71
gi|1497857|gb|U34995|HSU34995	glyceraldehyde-3-phosphate dehydrogenase (GAPD)	82
gi|35052|emb|X53778|HSNGMRNA	*H. sapiens* hng mRNA for uracil DNA glycosylase, GAPD	84

These reads all contain domains that are shared by large gene families. No known repetitive elements have been observed.

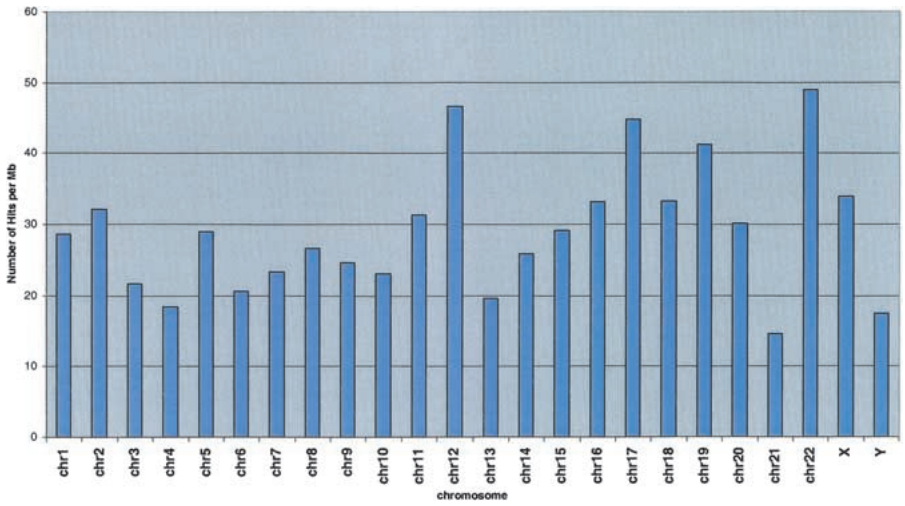


Figure 4 The density plot of matches on individual human chromosomes. The *ordinate* is the number of matches divided by the size of this chromosome in megabases. Therefore, the bigger the value, the more dense the match. The most dense chromosomes are 12 and 22 and the most sparse is chromosome 21.

nome sequences with WGS reads from another species, we first generated pseudo-random reads, which could serve as a negative search control. These random reads were generated so that they reflect both the base composition and the known repetitive element positions in real reads. By comparing the search results between these random reads and actual reads, we found that rat/human matches with a BLAST bit score >60, match length >50 bp, and an e-value $<10^{-5}$ are extremely likely to represent real cross-species matches. To find relatively weak matches, we also retain hits that satisfy at least one of these three conditions, which can still have a signal-to-noise ratio >100:1. Based on these criteria, we found that among 450,000 rat WGS reads, ~3% contain known transcribed regions that are conserved in the human genome. This result is consistent with the estimation that ~90% of the genes are conserved between rodent and human and ~3% of the mammalian genome are coding regions (I.H.G.S.Consortium 2001). Moreover, using this set of parameters, we found that the number of matches obtained are proportional to the size of the database we searched against. For example, the size of the Goldenpath database is 2.4× bigger than that of the HS3 database and we found 2.4× more matches searching against Goldenpath. This indicates that the occurrence of nonspecific matches in our study is low.

Identification of Repetitive Elements

Some WGS rat reads are similar to sequences that appear many times in the human genome. These sequences could be unknown genomic repetitive elements shared by rat and human, or else coding sequences of protein domains shared by gene families and other conserved functional and regulatory elements. We showed it is relatively easy to identify common transcribed gene domain sequences, as these also have matches in transcript databases. It was more difficult to distinguish the other two types of repeat elements—the repetitive elements and other abundant elements that may have roles in processes such as gene regulation, chromosomal structure, etc.

Choosing BLAST Search Parameters

The sensitivity and the speed of the BLAST search are affected by the set the parameters used. One of the most important parameters is the word size. Generally speaking, the larger the word size, the less sensitive the search results, and the faster the search speed. Because we are interested in recovering potential regulatory regions that are conserved between human and rat, we have chosen the default word size of 11 nucleotides. Choosing an even smaller word size will reduce the speed and make the search too expensive. It has been shown that choosing a set of stringent search parameters can exclude intronic matches between human and pufferfish (Roest Crollius et al. 2000). This is not the case, however, between human and rat, as considerably large amounts of intronic matches between human and rat persist at high stringency.

Table 6. Search Results of the UniGene Database

A.

	Entries in the UniGene Database	Matched reads
Known gene cluster	11,751	8841
EST-only cluster	75,167	4373

B.

	No match	Known gene	EST only
Number of reads	613	1269	2495
Percentage	14%	29%	57%

(A) In the UniGene database, 11,751 clusters/entries represent known genes (based on the keyword "mRNA" in the description line of the entry), while most of the rest of 75,167 clusters contain only EST clones. In the 12,684 rat WGS reads that matched to the UniGene database, 8841 of them matched to known genes while only 4373 reads matched to the EST clusters. (B) TBLASTX search results. Among the 4373 reads, 14% of the rat WGS reads do not have good matches in the database. Twenty-nine percent of these reads matched with known genes and 57% of the reads still match only to EST clones after translation.

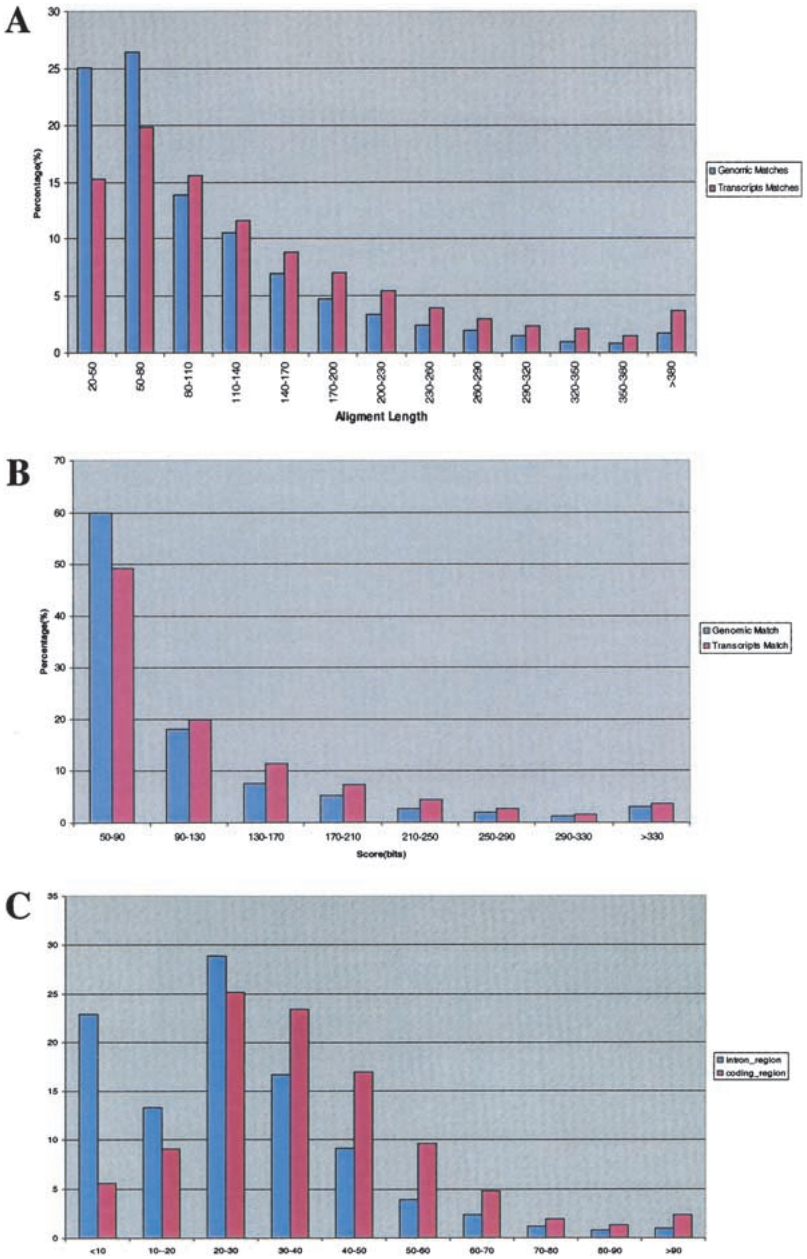


Figure 5 Comparison of the distribution of score and alignment length between matches from nontranscribed and transcribed regions. Columns in blue represent matches from the nontranscribed region, whereas red columns are matches from the transcribed region. (*A*) The distribution of the match length of nucleotide alignment. A shift to longer alignment is observed in the matches from the coding region. Similarly, in *B*, the average score from intron region is also lower than that from the coding region. (*C*) The distribution of match length of amino acid alignment. A more clear shift to better alignment in the coding region has been seen.

To distinguish the real matches and the random matches, we instead decided a cut-off value by comparing the distribution of matches between real reads and simulated reads.

WGS Reads Provide a Potential Resource to Discover Conserved Domains and New Genes

These data indicate that most conserved genes can be revealed through the rat WGS reads and the human genome comparison. About 3% of the WGS reads contain conserved coding regions, consistent with the anticipated percentage of the coding sequence in the human genome. Furthermore, with a coverage of 7.5% of the rat genome, 40% of the known gene entries of the UniGene database and 37% of the HTDB database were identified. This result is consistent with the notion that the recovery rate by random sequences is independent of the size of the database, but determined by the sequence coverage (Ewing and Green 2000). Based on the rat/human comparison, we can estimate the total number of genes in the human genome. Because 14881 out of 76447 genomic matches are localized to the `Genie`-known gene set (a total of

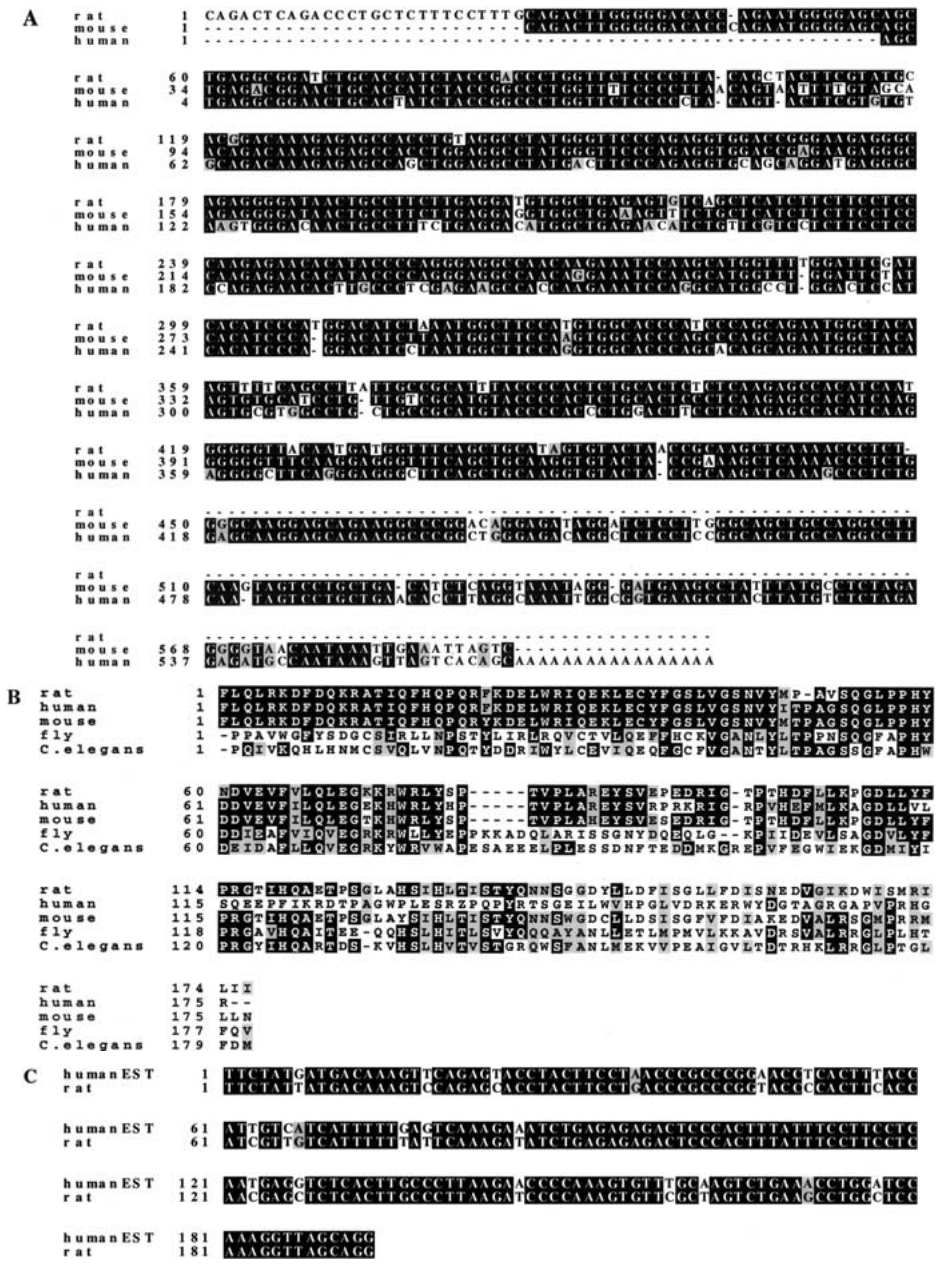


Figure 6 Sequence alignment of three different cases between EST and the rat WGS reads. (*A*) DNA alignment of region conserved in human, mouse and rat. The GenBank entry number for the human EST is AW139193 and the mouse cDNA is AK0055990. (*B*) Protein alignment of region conserved in human, mouse, and rat. In addition, the putative protein is likely to be conserved in *Drosophila* and in *Caenorhabditis elegans*. The letter z in the human sequence indicates a stop codon. The fact that the mouse and rat sequences continue to be very similar to each other before and after this stop codon indicates some sequencing or cloning error in the human sequence. The GenBank entry number for the human EST is BE784449, the mouse cDNA is AK013451, the fly is AAF45939.1, and the *C. elegans* is AAB53055.1. (*C*) DNA alignment of region that is only observed in human and rat; no mouse EST has been found. The human EST GenBank accession number is BF028817.

8290 genes), the number of genes in the human genome could be calculated as 42,587 (76447*8290/14881). This estimation disregards the difference in conservation and size between known genes and those unknown genes, and it also assumes that all human genes could be found by the rat/human comparison. Even with this assumption, this resulting number is slightly higher than the current estimation of number of genes in the human genome, indicating that most if not all of the human genes are conserved in the rat genome. Therefore, most of the entries in the current transcript database can be recovered through the rat WGS reads and human comparison.

METHODS

Generation of Rat WGS Sequence Reads

Genomic DNA was extracted from rat liver using the QIAGEN Genomic-tip System. After mechanical shearing, DNA fragments with a size of 1–3 kb were isolated by gel electrophoresis and cloned into a M13 vector using the double adaptor method (Andersson et al. 1996). Individual M13 plaques were picked and sequenced using the BODIPY dye primer chemistry (Metzker et al. 1996).

Sequence Data and Programs

The human genomic data used in the paper were downloaded from databases described in Box 1. Masking of known repetitive elements was performed using the `RepeatMasker` program. Sequence similarity searches were performed using BLAST (Altschul et al. 1997). The rest of the process was performed by using ad hoc scripts.

ACKNOWLEDGMENTS

We thank the rat production groups and the informatics group at HGSC for support. This work was supported by grant number HG02395 from the NHGRI and NHLBI at the National Institutes of Health. The rat genome WGS used in this study was generated at the BCM-HGSC during the year 2000. The rat genome sequencing project is now underway as a collaborative effort among the BCM-HGSC, Celera Genomics, Genome Therapeutics and other parties, funded by the NHLBI and NHGRI.

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

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Andersson, B., Wentland, M.A., Ricafrente, J.Y., Liu, W., and Gibbs, R.A. 1996. A 'double adaptor' method for improved shotgun library construction. *Anal. Biochem.* **236:** 107–113.

Ansari-Lari, M.A., Shen, Y., Muzny, D.M., Lee, W., and Gibbs, R.A. 1997. Large-scale sequencing in human chromosome 12p13: Experimental and computational gene structure determination. *Genome Res.* **7:** 268–280.

Bouck, J., McLeod, M.P., Worley, K., and Gibbs, R.A. 2000a. The human transcript database: A catalogue of full length cDNA inserts. *Bioinformatics* **16:** 176–177.

Bouck, J.B., Metzker, M.L., and Gibbs, R.A. 2000b. Shotgun sample sequence comparisons between mouse and human genomes. *Nature Genet.* **25:** 31–33.

Ewing, B. and Green, P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes [see Comments]. *Nature Genet.* **25:** 232–234.

Gelfand, M.S., Koonin, E.V., and Mironov, A.A. 2000. Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res.* **28:** 695–705.

International Human Genome Sequencing (I.H.G.S.) Consortium. 2001. Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. *Nature* **409:** 860–921.

Lichtarge, O., Bourne, H.R., and Cohen, F.E. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257:** 342–358.

Metzker, M.L., Lu, J., and Gibbs, R.A. 1996. Electrophoretically uniform fluorescent dyes for automated DNA sequencing. *Science* **271:** 1420–1422.

O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J., Nash, W.G., Wienberg, J., Stanyon, R., Copeland, N.G., Jenkins, N.A., Womack, J.E., and Marshall Graves, J.A. 1999. The promise of comparative genomics in mammals. *Science* **286:** 458–462.

Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137–140.

Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., Saurin, W., and Weissenbach, J. 2000. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nature Genet.* **25:** 235–238.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., and Holt, R.A. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Wyeth, W., Wasserman, M.P., Thompson, W., Fickett, J.W., and Lawrence, C.E. 2000. Human-mouse genome comparison to locate regulatory sites. *Nature Genet.* **26:** 225–228.

Received May 15, 2001; accepted in revised form August 16, 2001.