Original Contributions

Mammalian Genome
Incorporating Mouse Genome

Mammalian Genome 12, 673–677 (2001).
DOI: 10.1007/s00335-001-2048-4

© Springer-Verlag New York Inc. 2001

Computational analysis of full-length mouse cDNAs compared with human genome sequences

Shinji Kondo,[1,*] Akira Shinagawa,[1,*] Tetsuya Saito,[1,*] Hidenori Kiyosawa,[1] Itaru Yamanaka,[1] Katsunori Aizawa,[1,2] Shiro Fukuda,[1,2] Ayako Hara,[1] Masayoshi Itoh,[1] Jun Kawai,[1] Kazuhiro Shibata,[1,2] Yoshihide Hayashizaki[1,2,3]

[1]Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan
[2]CREST, Japan Science and Technology Corporation (JST), 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan
[3]University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan

Received: 18 January 2001 / Accepted: 17 May 2001

Abstract. Although the sequencing of the human genome is complete, identification of encoded genes and determination of their structures remain a major challenge. In this report, we introduce a method that effectively uses full-length mouse cDNAs to complement efforts in carrying out these difficult tasks. A total of 61,227 RIKEN mouse cDNAs (21,076 full-length and 40,151 EST sequences containing certain redundancies) were aligned with the draft human sequences. We found 35,141 non-redundant genomic regions that showed a significant alignment with the mouse cDNAs. We analyzed the structures and compositional properties of the regions detected by the full-length cDNAs, including cross-species comparisons, and noted a systematic bias of GENSCAN against exons of small size and/or low GC-content. Of the cDNAs locating the 35,141 genomic regions, 3,217 did not match any sequences of the known human genes or ESTs. Among those 3,217 cDNAs, 1,141 did not show any significant similarity to any protein sequence in the GenBank non-redundant protein database and thus are candidates for novel genes.

Introduction

Identification and annotation of genes in the human genome are under way, exhaustively using the resources available. While ESTs are the most abundant (more than 3 million sequences), their use is severely limited, because a large proportion are contaminating sequences, which include unspliced introns, genomic DNA, and spurious transcriptions. Only those genomic regions aligned with ESTs that show clear signals, such as splice junctions, are annotated (Hattori et al. 2000; Venter et al. 2001). Although *ab initio* methods are useful for extracting plausible genes and approximating the gene structures (Burset and Guigo 1996), their predictions require support from ESTs or protein sequences before they can be annotated. An unknown proportion of their predictions are false positives, while some true genes escape their scrutiny. A study of the calibration of a routinely used gene predictor, GENSCAN (http://genes.mit.edu/GENSCAN.html) reported that it correctly predicted all exons for only 20% of the human genes tested (Dunham et al. 1999). Under such circumstances, alignment of full-length cDNAs with genomic sequences should provide the most reliable way of identifying genes and determining their whole structures.

The number of coding exons and their sizes are fairly well conserved between orthologous mouse and human gene pairs. Batzoglou et al. (2000) studied 117 orthologous gene pairs and reported that the number of exons is identical for 95% of the pairs and that the lengths of corresponding exons are identical for 73% of the pairs. The coding regions tend to show 85 ± 7% identity at the nucleotide and protein levels (Makalowski and Boguski 1995). In contrast, corresponding introns show only a weak identity—approximately 35%, which is nearly the background sequence identity rate for random sequences (Batzoglou et al. 2000). Thus, it is quite feasible that nearly all the exons of human genes in the human genome can be detected by using full-length cDNAs of their mouse counterparts as probes.

Materials and methods

The mouse cDNAs. To detect genes in the human genome, we used 21,076 full-length mouse cDNAs and 40,151 EST sequences. Details of the 21,076 full-length cDNAs (corresponding to between approximately 13,000 and 16,000 non-redundant genes) can be found in the report of Kawai et al. (2001). Although the redundancy of the ESTs is undetermined, they were not expected to significantly overlap the full-length cDNAs, because they were sampled after an effective subtraction was applied to the tissue libraries by using previously sequenced cDNAs as drivers (Carninci et al. 2000). The ESTs are single-read sequences to be used for full-length assembly. The read lengths given by the three sequencers we used are 600 ± 80 bases [RISA System (Shibata et al. 2000)], 840 ± 120 bases (ABI377 and ABI3700, Applied Biosystems Inc.) and 1,020 ± 150 bases (Licor DNA 4200, Licor).

Determination of loci. The RIKEN mouse cDNAs (RepeatMasked by RepeatMasker) (http://ftp.genome.washington.edu/RM/RepeatMasker.html) were BLASTN-searched (-p blastn -e 1.0) (http://www.ncbi.nlm.nih.gov/BLAST) against the human draft sequences [GoldenPath sequences (http://genome.cse.ucsc.edu/goldenPath/) assembled by the Center for Biomolecular Science and Engineering, University of California, Santa Cruz]. We observed that typically aligned regions of a cDNA are scattered throughout the human genome, flanked by introns. For each aligned region, we inspected its strand and positions, on both the cDNA and genome coordinates, and collected those that we considered to belong to a single locus. (See Supplementary information: Table 1a–b, Fig. 1 and Locus determination method.)

Computer-based annotations were given to each locus, such as the total

*These authors contributed equally to this work.

Correspondence to: S. Kondo; e-mail: rgscerg@gsc.riken.go.jp

The EST sequence data described in this paper have been submitted to the DDBJ data library under accession nos: BB609426–BB667065

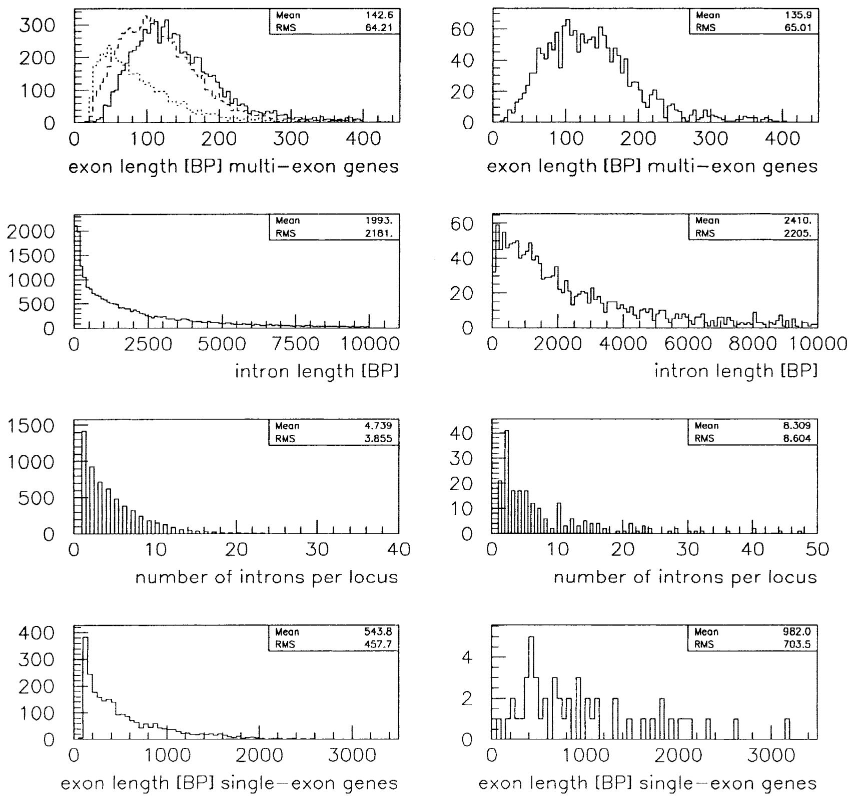

Fig. 1. Comparison of gene structures of the loci detected by the RIKEN full-length mouse cDNAs with those reported for genes annotated on human Chr 21 (http://hgp.gsc.riken.go.jp/chr21/Genetable.html). The figures on the left illustrate the features of the loci detected by the RIKEN mouse cDNAs, and those on the right show the annotated features of the genes on Chr 21. **Top:** Lengths of exons (aligned regions) in multiple-exon loci. The top-left figure shows distributions for aligned regions not overlapping GENSCAN exons (dotted line), aligned regions overlapping GENSCAN exons (dashed line), and GENSCAN exons overlapping the aligned regions (solid line). **Second:** Lengths of introns per locus. **Third:** Number of introns per locus. **Bottom:** Lengths of exons (aligned regions) in single-exon loci.

aligned region, number of introns, and matching of each aligned region with exons predicted by GENSCAN. (Supplementary information: Table 1b)

While stringent thresholds for the annotated features effectively pick up only human orthologs or close paralogs, non-stringent thresholds can be used to thoroughly detect even loci of distant paralogs. Since the draft sequences are still incomplete and contain errors, even the known human genes could be aligned only partially, as described in the next section; approximately 10% of them cannot yet be aligned at all (Lander et al. 2001). After testing certain sets of thresholds, we found that a simple threshold of the total aligned region is sufficiently effective to guarantee a high sensitivity without loss of much specificity. As a default, we considered loci that aligned with cDNAs over 100 or more bases. The identity was not a relevant parameter, since we consistently observed 75% or better identity (expected from the 85 ± 7% reported above) for the aligned regions of 100 or more bases, as described later. We evaluate our method with respect to the sensitivity and specificity in the next section.

A "TAKERU 2000" system (http://www.nabe-intl.co.jp), a Beowulf PC cluster equipped with a 16 CPUs (Pentium III, 800 MHz), was employed for the task of locus determination and annotation; searching the 61,227 mouse cDNAs against the whole human genome could be completed within 20 h.

Sensitivity, specificity, and fragmentation. To assess the sensitivity of the above procedure, we aligned 120 orthologous mouse sequences for human genes annotated on Chromosomes (Chrs) 21 and 22 with the genomic sequences of those chromosomes to compare the detected locus positions with their annotated positions. We could align 117 (98%) of them at the annotated positions (seven of them were aligned slightly off the annotated positions, and two of them were aligned on the opposite strand); 92 of them could be aligned over 50% or more of their lengths, and 109 could be aligned over 30% or more of their lengths. (Accession numbers of these sequences and the results of their alignment are provided in Supplementary information: Tables 2a–i.) In addition, we aligned 10,239 known human gene sequences in the RefSeq database (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html). At 95% or better identity, we could align 8,882 of them over 50% or more of their lengths and could align 5,936 of them over 90% or more of their lengths.

For two genes (Supplementary information: Tables 2b–c), our method split each single gene into two separate loci. The alignment generated 21 additional loci—20 corresponded to annotated loci of their paralogs and 1 did not correspond to any annotated locus (Supplementary information: Tables 2d–f).

We tested the specificity by using 100 sequences chosen randomly from the UniGene mouse clusters (http://www.ncbi.nlm.nih.gov/UniGene/index.html). When we aligned those sequences with the two chromosomes, our method located 11 loci, 7 of which did not have annotations of their orthologous and homologous human genes on the chromosomes (Supplementary information: Table 2i).

In summary, the sensitivity of our method is 98%, and the rate of false positives is at the most 6%. We observed fragmentation of a single gene into two separate loci in fewer than 2% of the sequences used.

Results

Number of loci identified. We found 35,141 non-redundant loci that aligned with the mouse cDNAs over 100 or more bases. (See Supplementary information: Figs. 2 and 3 for their chromosomal distribution and correlation with the gene richness parameter.) Owing to the redundancy among the mouse cDNAs, a number of the loci are hit by more than 1 cDNA. In such cases, we assigned to the locus the single representative giving the largest total aligned region (see Supplementary information: Fig. 4). We also observed that a single cDNA detected more than one locus in many instances, because it can locate its paralogs and pseudogenes (see Supplementary information: Tables 1a–b and Fig. 1). If a member of a large family exists among the mouse cDNAs, it could hybridize with nearly all its members. Thus, the single cDNA could also detect loci of other family member genes. After assigning a single representative to each locus, we obtained 20,177 non-redundant representative cDNAs. Thus, the 20,177 representative cDNAs form 35,141 loci. This implies that a large proportion of human genes belong to families. This tendency remained invariant even when we selected only loci that aligned at a much higher strin-

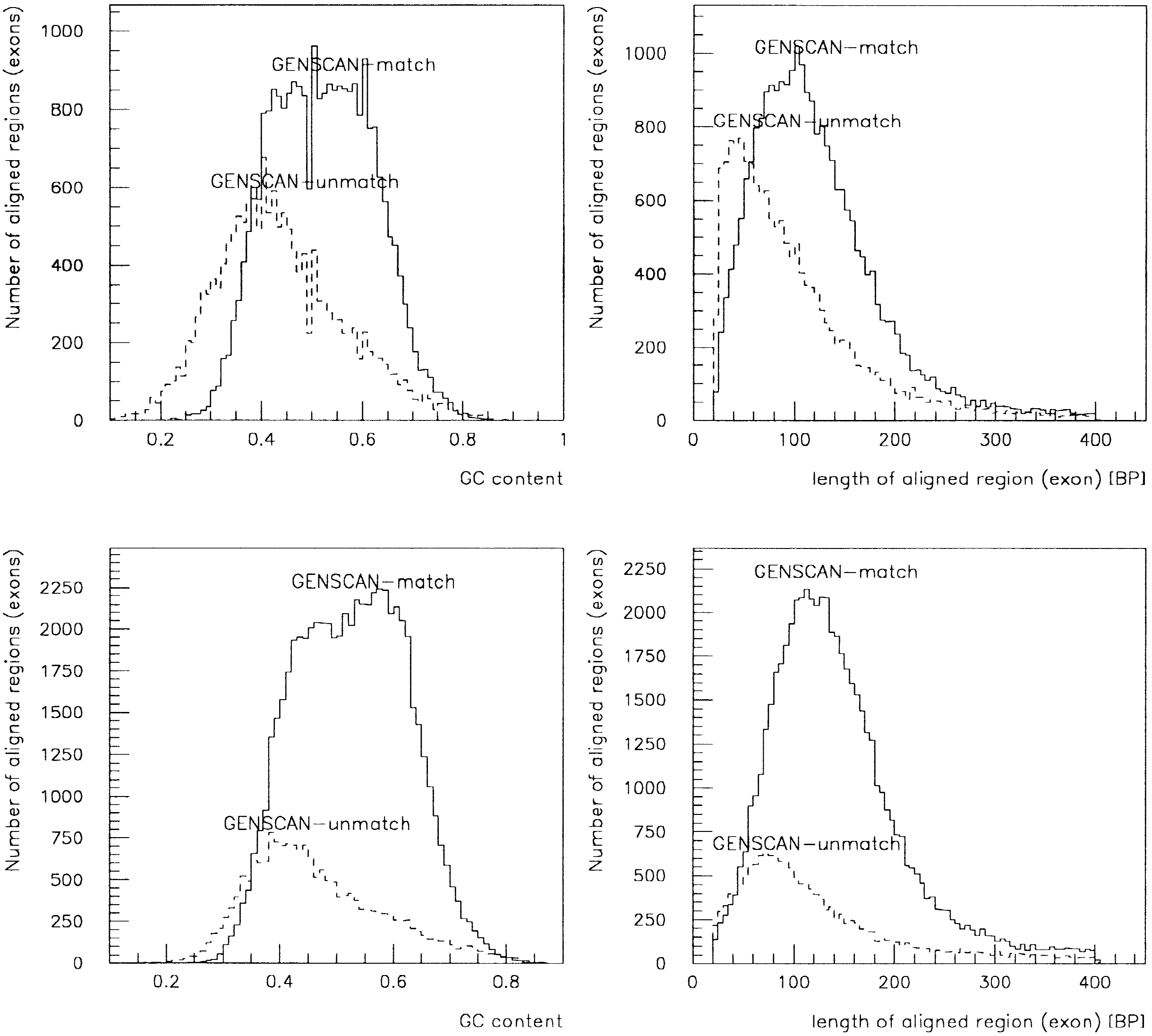

Fig. 2. GC-content and lengths of aligned regions matching or not matching GENSCAN exons. Figures on the left show the distribution of aligned regions with given GC-contents, and figures on the right show the distribution of their lengths. Solid lines indicate the regions matching GENSCAN exons, and dashed lines indicate the regions not matching them. The two top figures show the results from the loci detected by RIKEN full-length mouse cDNAs. Only the locus best matching each cDNA was selected. The results observed for the loci detected by human RefSeq sequences are shown in the two bottom figures. Loci that aligned at 95% or better identity over 50% or more of their lengths were used.

gency (over 300 or more bases) with the cDNAs; i.e., 20,172 loci are formed by 13,148 representative cDNAs.

Novel gene candidates. To identify candidates for novel genes, the 20,177 representative cDNAs were pairwise searched (-p blastn -e 1.0) against the four human gene repositories: RefSeq (10,239 human sequences), UniGene (2,231,796 human sequences), GenBank dbEST (3,049,782 human sequences) (http://www.ncbi.nlm.nih.gov/dbEST), and Ensembl transcription data (36,971 human sequences) (ftp://ftp.sanger.ac.uk/pub/ensembl/data/cdna/ensembl.cdna). In this search we did not include the mouse cDNAs that did not have significant alignment (100 bases or more aligned bases) with the human genome. After we discarded those that showed 70% or better identity over 100 or more aligned bases with any sequence in the above four gene databases, 3,202 cDNAs remained (817 full-length and 2,385 EST sequences). These 3,202 cDNAs were further searched (BLASTX with default parameters) against the GenBank non-redundant protein database (582,290 protein sequences from human and other species) (ftp://ncbi.nlm.nih.gov/blast/db/nr). Among the 3,202 cDNAs, 2,061 showed a significant match (25% or better identity over 50 or more aligned amino acid residues) with sequences in the protein database. Of the remaining 1,141 cDNAs, 231 were full-length cDNAs and 910 were ESTs. Thus, we present these 1,141 cDNAs as candidates for completely novel genes.

Exon–intron structures. To analyze the gene structures, we used only loci that best matched each full-length cDNA, giving the largest alignment (see supplementary information: Locus determination method). The features of exons and introns of the identified loci were compared with those annotated on Chr 21 (Fig. 1). The lengths of the exons matching the GENSCAN predictions and introns of those loci are comparable with those of Chr 21 (Figure 1). The fewer introns and shorter lengths of singleton (loci containing a single exon) exons of the loci identified by the RIKEN cDNAs are probably attributable to the fact that the RIKEN cDNAs represent relatively short genes—at 1,270 ± 660 bases, they are approximately 1,000 bases shorter than the known mouse

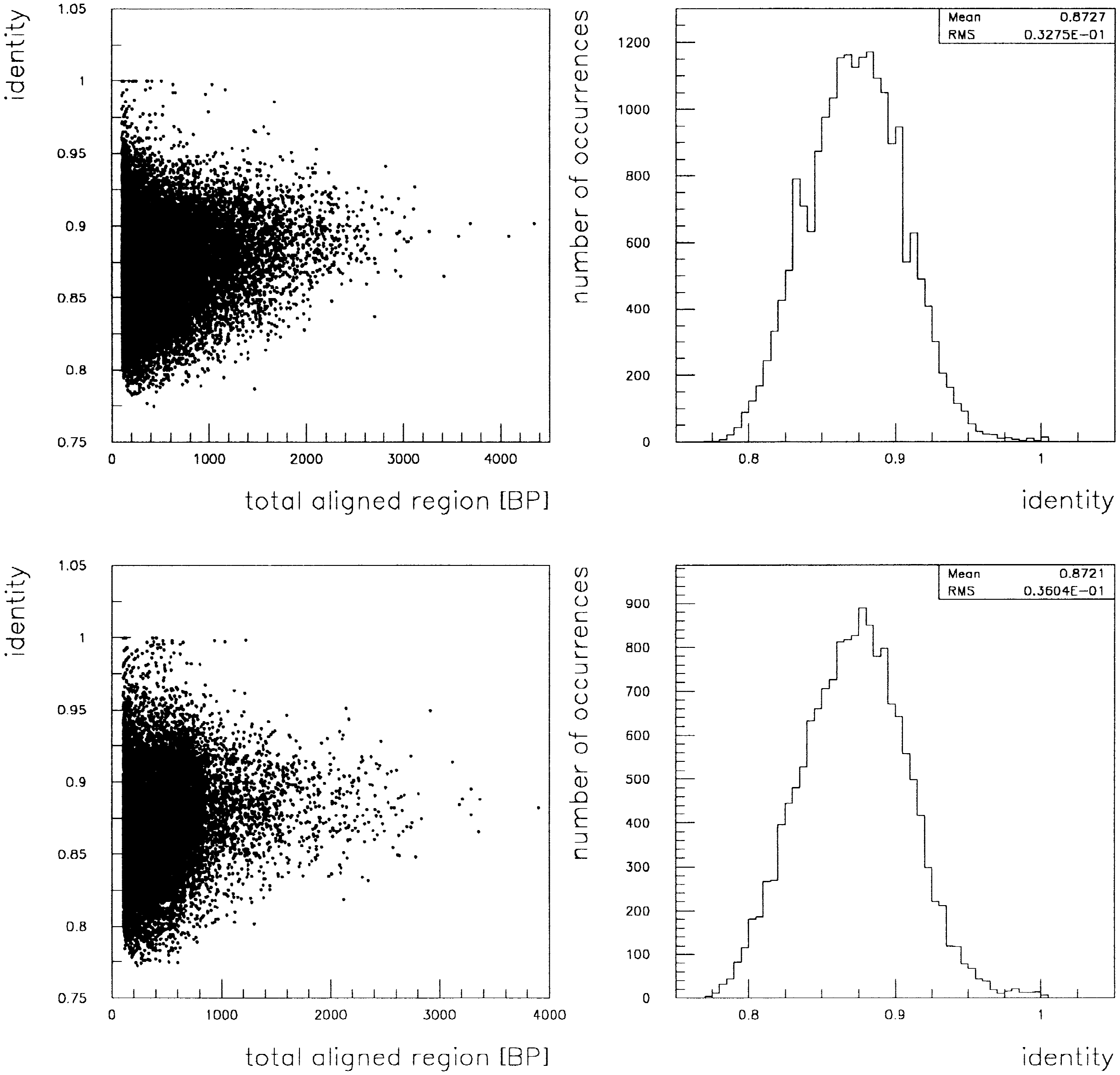

Fig. 3. The identities observed for the aligned regions between the RIKEN full-length mouse cDNAs and the human genome, and for those between RIKEN mouse cDNAs and UniGene human sequences. **Top left:** Dependence of identity variation on the size of the total aligned region (RIKEN mouse cDNAs vs. human genome sequences). **Top right:** Numbers of loci with given identifies (RIKEN mouse cDNAs vs. human genome sequences). Only the locus best matching each cDNA was selected. **Bottom left:** Identities vs. sizes of the total aligned regions (RIKEN mouse cDNAs vs. UniGene human sequences). **Bottom right:** Distribution of identities (RIKEN mouse cDNAs vs. UniGene human sequences). The RIKEN mouse cDNAs were aligned pairwise with the 2,231,796 UniGene sequences, and the best match with each RIKEN mouse cDNA was selected.

genes (2,200 ± 1500 bases). (We gave higher priority to short genes in the full-length assembly.) We observed that the exons not matching GENSCAN exons are approximately 50–60 bases shorter than exons typically predicted by GENSCAN. This implies a potential bias of GENSCAN against short exons. We return to this subject in the next section.

Exon and gene density vs. GC content. We computed the GC-content of each aligned region of the human genome and observed a large discrepancy (51 ± 10% vs. 43 ± 11%) in the GC-content between the regions matching and not matching GENSCAN exons (Fig. 2). To confirm this potential bias, we also examined the GC-content of the regions aligned with the human RefSeq sequences. We used only those regions given by the 8,882 sequences that aligned at 95% or better identity over 50% or more of their lengths. The tendencies remained unchanged for the regions (exons) of the RefSeq sequences that aligned under more stringent similarity conditions (Fig. 2). This strongly implies that GENSCAN tends to miss exons of small size and/or low GC-content.

We also computed the dependence of the loci density on the GC-content. The relative density of loci increased approximately 10-fold as the GC content shifted from 30% to 60% (Supplementary information: Fig. 5). This is consistent with the result reported by Lander et al. (2001).

Identity. Nucleotide-level identity in coding regions between orthologous mouse and human gene pairs is reported to be 85 ± 7% (Makalowski and Boguski 1995). The nucleotide-level identity of the aligned regions between the mouse cDNAs and the human genome turned out to be 88 ± 6% (Fig. 3), which agrees with the value reported above. Figure 3 also shows the dependence of the identity on the size of the aligned regions (exons). We observed that the identities converge toward approximately 88% as the sizes of the total aligned regions increase. An identical convergence is reported by Makalowski and Boguski (1995). The pairwise alignment of the mouse cDNAs with UniGene human sequences also showed the same tendency (Fig. 3). We selected only the best match (the largest aligned bases) of each cDNA with the UniGene sequences.

Discussion

Although computer predictions have been reported to perform well, the statistics were based on a relatively small number of genes that do not necessarily represent the whole gene population (http://genes.mit.edu/Limitations.html). The most routinely used predictor, GENSCAN, has proven to be extremely useful in gene identification; however, it still misses a certain proportion of exons and genes. Using other types of software might alleviate this problem. However, it is advisable to assess the potential biases of individual programs with a sufficient number of genomic sequences. Such calibration efforts should lead to the most effective usage of individual programs and stimulate the development of more accurate *ab initio* methods.

The number of novel gene candidates we found in this study was fairly small compared with the number detectable with available ESTs and protein sequences; this suggests that the number of unexplored human genes is also relatively small. Although we have attempted to sequence new genes by applying effective subtraction and normalization to the tissue libraries, we have so far sequenced fewer than 10,000 clones in each of most libraries, which typically contain around 10^7 to 10^8 clones. A further sequencing beyond the current number of clones may find more genes of low expression level.

We have presented an initial computational analysis of the human genomic loci located by mouse cDNAs. To verify the candidate loci detected, extensive manual curation is still required. As the mouse genome sequences become publicly available, comparison of the mouse cDNAs with the mouse genome will yield further information to refine this initial analysis. The results will be reported when available.

Acknowledgment. The authors thank the following people, listed in alphabetical order: for their discussion, encouragement, and technical assistance: J. Adachi, A. Arai, K. Arakawa, T. Akimura, P. Carninci, M. Furuno, T. Hanagaki, H. Hara, N. Hayatsu, K. Hiramoto, T. Hiraoka, T. Hirozane, Y. Hodoyama, F. Hori, K. Imotani, Y. Ishii, N. Kikuchi, Y. Kojima, H. Konno, M. Kouda, S. Koya, T. Matsuyama, M. Nakamura, K. Nomura, R. Numazaki, T. Okido, T. Ono, M. Ohno, N. Osato, R. Saito, C. Sakai, K. Sakai, H. Sano, K. Sano, D. Sasaki, K. Sato, Y. Shibata, T. Shiraki, Y. Sogabe, Haru. Suzuki, Hiro. Suzuki, A. Tagawa, F. Takahashi, S. Takaku-Akahira, T. Tanaka, T. Toya, A. Watahiki, T. Yamamura, A. Yasunishi, K. Toshida, K. Yoshida and S. Yoshida. They also thank N. Kazuta, L.K. Sato, Y. Shigemoto, and H. Torigoe for their secretarial assistance. This study has been mainly supported by a Research Grant for the RIKEN Genome Exploration Research Project, CREST (Core Research for Evolutional Science and Technology), and the ACT-JST (Research and Development for Applying Advanced Computational Science and Technology) of Japan Science and Technology Corporation (JST) to YH. All of these research funds are sponsored by the Science and Technology Agency of the Japanese Government. This work was also supported by a Grant-in-Aid for Scientific Research on Priority Areas and Human Genome Program, from the Ministry of Education, Science and Culture, and by a Grant-in-Aid for a Second Term Comprehensive 10-Year Strategy for Cancer Control from the Ministry of Health and Welfare to Y. Hayashizaki.

References

Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. Genome Res 10, 950–958

Burset M, Guigo R (1996) Evaluation of gene structure prediction programs. Genomics 34, 353–367

Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, et al. (2000) Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. Genome Res 10, 1617–1630

Dunham I, Hunt AR, Collins JE, Bruskiewich R, Beare DM, et al (1999) The DNA sequence of human chromosome 22. Nature 402, 489–495

Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, et al (2000) The DNA sequence of human chromosome 21. Nature 405, 311–319

Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, et al (2001) Functional annotation of a full-length mouse cDNA collection. Nature 409, 685–690

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al (2001) The human genome: initial sequencing and analysis. Nature 409, 860–921

Makalowski W, Boguski S (1995) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci USA 16, 9407–9412

Shibata K, Itoh M, Aizawa K, Nagaoka S, Sasaki N, et al (2000) RIKEN integrated sequence analysis system (RISA system)-384-format sequencing pipeline with 384 multi-capillary sequencer. Genome Res 10, 1757–1771

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al (2001) The sequence of the human genome. Science 291, 1304–1350