# Numerous Novel Annotations of the Human Genome Sequence Supported by a 5′-End–Enriched cDNA Collection

Betina M. Porcel, Olivier Delfour, Vanina Castelli, et al.

*Genome Res.* 2004 14: 463-471
Access the most recent version at doi:10.1101/gr.1481104



Resource

# Numerous Novel Annotations of the Human Genome Sequence Supported by a 5′-End–Enriched cDNA Collection

Betina M. Porcel,[1,5] Olivier Delfour,[1] Vanina Castelli,[1] Veronique De Berardinis,[1] Lucie Friedlander,[1,2] Corinne Cruaud,[1] Abel Ureta-Vidal,[1,3] Claude Scarpelli,[1] Patrick Wincker,[1] Vincent Schächter,[1] William Saurin,[1,4] Gabor Gyapay,[1] Marcel Salanoubat,[1] and Jean Weissenbach[1]

[1]*Genoscope-Centre National de Séquençage and CNRS UMR-8030, 91000 Evry, France*

A collection of 90,000 human cDNA clones generated to increase the fraction of "full-length" cDNAs available was analyzed by sequence alignment on the human genome assembly. Five hundred fifty-two gene models not found in LocusLink, with coding regions of at least 300 bp, were defined by using this collection. Exon composition proposed for novel genes showed an average of 4.7 exons per gene. In 20% of the cases, at least half of the exons predicted for new genes coincided with evolutionary conserved regions defined by sequence comparisons with the pufferfish *Tetraodon nigroviridis*. Among this subset, CpG islands were observed at the 5′ end of 75%. In-frame stop codons upstream of the initiator ATG were present in 49% of the new genes, and 16% contained a coding region comprising at least 50% of the cDNA sequence. This cDNA resource also provided candidate small protein-coding genes, usually not included in genome annotations. In addition, analysis of a sample from this cDNA collection indicates that ~380 gene models described in LocusLink could be extended at their 5′ end by at least one new exon. Finally, this cDNA resource provided an experimental support for annotations based exclusively on predictions, thus representing a resource substantially improving the human genome annotation.

[The sequence data from this study have been submitted to EMBL under accession nos. BX323813, BX323814, BX324295–BX465182, AL513551–AL583711.]

The draft sequences of the human genome (Lander et al. 2001; Venter et al. 2001), together with the completed sequence of several human chromosomes (Dunham et al. 1999; Hattori et al. 2000; Deloukas et al. 2001; Heilig et al. 2003; Skaletsky et al. 2003) have provided a tremendous amount of information to be exploited over the years to come. However, the efforts of the International Human Genome Sequencing Consortium to achieve quasi-completeness of the human genome sequence must be accompanied by an exhaustive inventory and description of the human genes. Because both ab initio gene predictions and genome sequence comparisons show limitations in detecting untranslated regions (UTRs), an effort in sequencing full-length transcripts is still necessary to identify 5′-UTRs and disclose the complete gene structures and splice variants. The full description and molecular characterization of gene structures is also a prerequisite for the definition of the proximal promoter region, a preliminary step in unraveling the regulation of gene expression.

**Present addresses: [2]LGI-BioInformatic, Aventis Pharma S.A., 94400, Vitry-Sur-Seine, France; [3]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB101SD, UK; [4]Genomining, 92120, Montrouge, France.**
**[5]Corresponding author.**
**E-MAIL betina@genoscope.cns.fr; FAX 33-1-60-87-25-14.**
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1481104. Article published online before print in February 2004.

High-throughput systematic sequencing of EST libraries has provided a wealth of information on many human gene transcripts. However, these EST sequences are often partial and, hence, insufficient to define the structure of the entire genes and encoded proteins. A number of extensive cDNA programs have been initiated since then to supply sequence, mapping, and expression data on the corresponding genes (Strausberg et al. 1999, 2002; Wiemann et al. 2001; Kikuno et al. 2002). These programs generated resources that were crucial for both the annotation of the human genome (Wiemann et al. 2001; Reymond et al. 2002) and the experimental analysis of gene function. Despite these important and productive efforts, a small but undefined number of genes have not yet been identified, and many gene models remain incomplete.

Sequencing of full-length transcripts followed by sequence alignment on a genomic reference sequence have recently been successfully used for the identification of exon structures of both human and other eukaryotic genes (Haas et al. 2002; Collins et al. 2003). We generated a collection of human cDNA sequences enriched for full-length inserts that we aligned on the human genome sequence assembly. This approach led to the identification of 552 novel human genes not documented in the LocusLink resource (Pruitt et al. 2000; Wheeler et al. 2003). Moreover, an extension of the 5′-ends of a substantial number of already annotated genes was observed in these sequence alignments. This new collection of cDNAs is therefore a valuable instrument for improving the quality of the existent human genome annotation.

# RESULTS

## The cDNA Collection and Its Analysis

mRNAs from nine human tissues—namely, neuroblastoma, placenta, fetal and adult brain, fetal liver, thymus, T- and B-cell lines, and the HeLa-cell line–were used to prepare cDNA libraries enriched for full-length inserts (see Methods). This set of cDNA libraries is hereafter referred to as the CNSLT cDNA resource. A total of 91,813 CNSLT cDNA clones were essentially submitted to single path pairwise end sequencing, producing >200,000 sequence reads (Table 1). The 5′ and 3′ sequence reads from each CNSLT cDNA clone were initially aligned on the repeat-masked human genome assembly (NCBI build 30) by using BLAST (see Methods). Ninety-six percent of the CNSLT cDNA clones (Table 2) yielded alignments that were used to define preliminary transcript models as described in Methods. A total of 1397 cDNA clones were considered to be putative chimeras due to the discrepant alignments of their 5′ and 3′ sequences on the genome and were excluded from the analysis. No alignment was observed between 3293 cDNA clones and the reference genomic sequence. This alignment was repeated with the 3293 nonmatching clones on NCBI build 33 that became available during revision of this article. A significant fraction of the clones (1369/3293, 42%) could be matched to the nearly finished human reference genomic sequence (for further details, see Methods). Furthermore, 228 of the 58% yet-nonmatching clones yielded an alignment with the mouse genome assembly.

The exon boundaries and the structure of each of the transcript models were further defined by using the sim4 algorithm (see Methods). Such transcript models, supported by the cDNA clones, were obtained for 97% (84,865 out of 87,123 clones) of the CNSLT cDNA resource.

The LocusLink resource serves as a central source to integrate sequence, standard nomenclature, gene, and protein descriptions (Pruitt et al. 2000; Wheeler et al. 2003). Therefore, this resource was the one chosen as a reference to pursue our analysis. We restricted this study to the subset of LocusLink entries identified by a LocusID tag through the use of their curated RefSeq mRNA or GenBank transcripts (for details, see Methods). Transcript models were obtained for 98% of such mRNA sequences.

Transcript models derived from the CNSLT cDNA clones and RefSeq or GenBank transcripts were merged by clustering, resulting in a total of 11,124 forward and 10,817 reverse clusters considered as gene models (see Methods). Of these gene models, 4241 genomic regions supported by the CNSLT cDNA resource were not documented by a LocusID (see Methods). A total of 2041 out of these 4241 regions, which corresponded to the CNSLT clones with overlapping 5′ and 3′ sequence reads (46%) were submitted for further analysis.

**Table 1. Sequencing Statistics by Human Tissue**

| Human tissue | No. of clones[a] | No. of reads |
|---|---|---|
| Adult brain | 1,689 | 4,378 |
| Fetal brain | 12,395 | 29,904 |
| Placenta | 31,400 | 71,881 |
| Thymus | 2,767 | 4,949 |
| Neuroblastoma | 21,394 | 48,365 |
| Fetal liver | 4,617 | 10,324 |
| HeLa cells | 4,050 | 10,115 |
| B cells (Ramos cell line) | 6,079 | 15,134 |
| T cells (Jurkat cell line) | 7,421 | 16,899 |

[a]Libraries were generated by 5′-end enrichment, as described in Methods.

## Manual Curation of Gene Models

It is well known that automated gene-finding procedures, especially in eukaryotes, are not yet fully accurate in terms of specificity or sensitivity or in the detection of exon and gene boundaries (Hogenesch et al. 2001). The procedure described above is not an exception. Therefore, the 2041 potential new gene models were subjected to manual inspection. The strategy followed for expert annotation of the gene models is outlined in Figure 1. Gene models were grouped into two different subcategories: spliced and unspliced models. To purge our analysis of possible genomic DNA contamination, 598 unspliced models devoid of a CDS of at least 300 bp (29% of the clusters) were automatically excluded from the manual validation. However, it is important to point out that such products could correspond to transcripts encoding small proteins or untranslated portions of real coding transcripts. Hence, a total of 1444 out of the 2041 original clusters were subjected to human curation, to validate the novel genes supported by the CNSLT cDNA resource. About 21% of these gene models (300/1444), which corresponded to dubious data (see Methods), were omitted from the analysis. We later searched the curated set of unspliced gene models against the human genome assembly looking for processed pseudogenes. The remaining 1072 clusters were later analyzed for their coding properties. Finally, a MegaBLAST comparison (Zhang et al. 2000) was performed against a more recent version of the genome sequence (NCBI build 31, freeze November 2002) in order to check for novelty.

The search for coding regions was performed for the reconstructed sequence of the genes mapping to nonannotated regions, according to the LocusLink resource (see Methods). In-frame stop codons upstream of the initiator ATG were present in 49% (273/552) of the new genes supported by the CNSLT cDNA resource. This number was consistent with that observed by using cDNA resources containing the 5′ ends of the transcripts (Suzuki et al. 2000; Wiemann et al. 2001). In 16% of the cases (44/273), the coding region spans at least 50% of the sequence rebuilt on the genome assembly.

After this curation procedure, the remaining 1072 proposed gene models (see Methods) were classified into three different categories: (1) 226 known genes, corresponding to genes included in LocusLink in the course of the analysis; (2) 552 novel genes, which have a CDS of at least 300 bp without a LocusID identifier (single-exon genes accounted for 31% of the total number of novel genes); and (3) 294 putative genes, corresponding to gene models with no LocusID identifier and no significant CDS. A summary of the number of novel, putative, and known genes supported by the CNSLT resource in each chromosome is given in Table 3. Novel genes were found even for chromosomes with annotation that has recently been updated (Reymond et al. 2002; Collins et al. 2003). In addition, among the 226 recently annotated models from category 1, 6% were based only on ab initio or in silico predictions (LocusLink Release 21.02.03): the CNSLT cDNA resource now provides experimental support for these models. Finally, the use of the CNSLT resource made it possible to fuse contiguous, yet distinct, gene models into a single model (data not shown).

The novel and putative genes, which were supported only by the CNSLT cDNA resource, were finally realigned to the NCBI build 33 in order to ascertain their position on the nearly "finished" version of the human reference genomic sequence (Table 4; for details, see Methods). A total of 548 out of the 552 proposed new gene models yielded alignments on the new release of the reference human genome sequence.

Only four of the novel genes, previously located on chromosomes 1, 8, and 17, did not align on the current human ref-

**Table 2. Statistics on Preliminary Transcript Models**

| | No. of clones/transcripts | No. of clones/transcripts matching the NCBI build 30 assembly |
|---|---|---|
| CNSLT cDNA clones | 90,416[a] | 87,123 (96%) |
| RefSeq seq with LocusID[b] | 17,146 | 16,841 (98%) |
| GenBank seq with LocusID[b] | 49,262 | 48,217 (98%) |

[a]Number of clones after removal of putative chimeras (see Results).
[b]Subset of RefSeq or GenBank transcripts identified by a LocusID in the LocusLink resource (see Methods).

erence sequence. Likewise, of the 294 putative genes defined on the build 30, only one, previously located on chromosome 3, did not match the current release of the human genome sequence. In some instances, novel and putative genes were fused in a single genomic region.

To evaluate the impact on the annotation of the "essentially complete" human genome sequence, we later re-examined the novelty of the gene models defined by the CNSLT resource using the Ensembl genes as the set of reference annotation (Hubbard et al. 2002). Novel and putative genes were associated by clustering with the Ensembl human genes (ENSG), generated by the Ensembl gene builder for NCBI build 33, which resulted in 23,575 clusters (see Methods). Of these, 335 genomic regions covered by our novel genes were still not documented by the Ensembl annotation. Furthermore, 268 of the putative genes map to 266 genomic regions devoid of Ensembl annotation (Table 5; for details, see Methods). In particular, seven out of the eight novel genes supported by the CNSLT cDNA clones identified on chromosome 21, as well as nine out of the 16 located on chromosome 22, were not identified by the Ensembl annotation pipeline.

A total of 210 Ensembl genes were covered by a novel gene defined by the CNSLT resource, which corresponded to 209 genomic regions. Fifty-two of these Ensembl annotations were tagged as novel Ensembl genes; these are now also confirmed by the CNSLT cDNA resource. To complete the analysis, we compared the CNSLT cDNAs now matching the NCBI build 33 and the Ensembl genes. In 74 cases, these cDNAs were located in genomic regions devoid of Ensembl annotation, indicating the existence of other new genes not identified by using the previous release of the human genome sequence.

A survey of the exon composition of the spliced models proposed for the new genes showed an average of 4.7 exons per gene, which was lower than reported mean values for extensively annotated genes (Heilig et al. 2003). However, the population of novel genes defined by the CNSLT resource was biased to small genes, because these were restricted to models based on CNSLT cDNA clones with 5′ and 3′ overlapping reads.

More than one transcript variant was observed for ~9% of the new genes.

To evaluate the proportion of complete 5′ ends in models

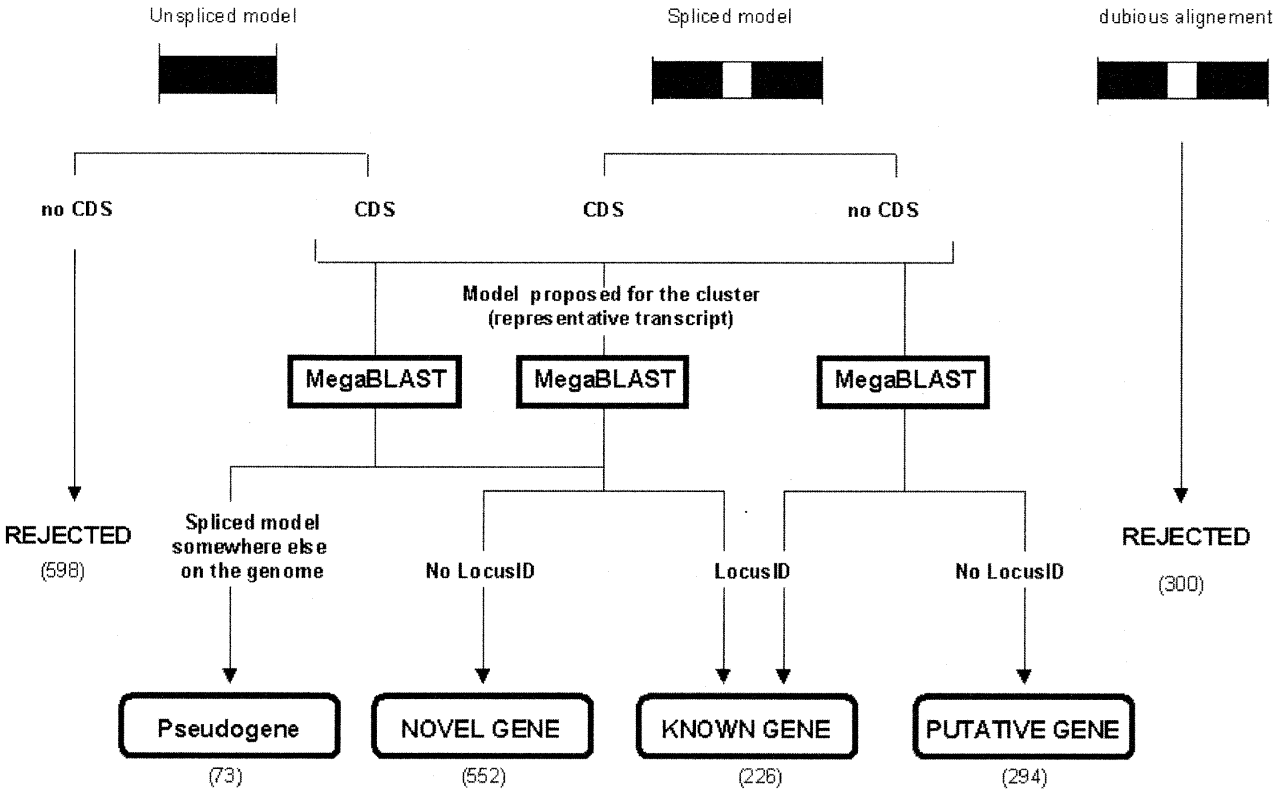


**Figure 1** Schematic outline of the strategy used to confirm the candidate genes. Numbers between brackets correspond to gene models and/or proposed gene models in each category. Description of the strategy followed for manual curation is shown in Results.

**Table 3. Chromosomal Distribution of Genes Supported by the Genoscope cDNA Resource**

| Chromosome | Novel genes | Putative genes | Known genes |
|---|---|---|---|
| 1 | 42 | 27 | 24 |
| 2 | 36 | 25 | 14 |
| 3 | 26 | 20 | 20 |
| 4 | 18 | 21 | 8 |
| 5 | 27 | 13 | 8 |
| 6 | 27 | 22 | 8 |
| 7 | 28 | 15 | 7 |
| 8 | 25 | 21 | 10 |
| 9 | 23 | 19 | 14 |
| 10 | 36 | 9 | 5 |
| 11 | 23 | 11 | 15 |
| 12 | 25 | 15 | 15 |
| 13 | 20 | 7 | 4 |
| 14 | 27 | 8 | 10 |
| 15 | 14 | 9 | 10 |
| 16 | 28 | 4 | 14 |
| 17 | 40 | 4 | 17 |
| 18 | 5 | 1 | 5 |
| 19 | 30 | 17 | 10 |
| 20 | 8 | 8 | 0 |
| 21 | 8 | 4 | 0 |
| 22 | 16 | 3 | 4 |
| X | 20 | 10 | 3 |
| Y | 0 | 1 | 0 |
| Total[a] | 552 (38%) | 294 (20%) | 226 (16%) |

[a]Total number of gene models is 1444 (for detailed explanation, see text).

proposed for the new genes after human curation, the presence of CpG islands in the 5′ regions was investigated (see Methods). A total of 344 CpG islands (60%) appeared in the vicinity of the 5′ end of the proposed novel models. Previous estimates have shown that between 60% and 67% of genes are associated with a CpG island at their 5′ end (Antequera and Bird 1999; Deloukas et al. 2001; Heilig et al. 2003).

Sequence comparisons with the pufferfish *Tetraodon nigroviridis* using Exofish (Roest Crollius et al. 2000) showed the presence of evolutionary conserved regions (ecores) for 34% of the proposed new gene models (188/552). Even though comparisons to the genomic fish sequence are not sufficient to identify all the exons, the sensitivity of the method attained 24% for exon identification on the new models supported by the CNSLT cDNA resources, which is similar to values previously reported (Heilig et al. 2003). However, in 20% of the novel gene models (110/552), at least half of the exons contain ecores defined by Exofish. In this subset, the occurrence of CpG islands close to the 5′ end was observed in 75% (82/110) of the cases, indicating that such gene models were complete or nearly complete.

An example of a novel gene supported by the CNSLT cDNA resource is shown in Figure 2A. This gene corresponds to a cluster of nine cDNA clones that match the reverse strand of the human genome assembly for chromosome 22. Five of these cDNA clones could be assembled on the genome sequence. Five different alternative transcripts were found for this gene, with the most abundant transcript variant chosen as representative for the gene. The model proposed corresponded to a structure of 18 exons confirmed by identifiable splice junctions. Thirteen out of the 18 exons were supported by linked *Tetraodon* ecores. A coding region of 272 animo acids was identified in the model proposed for this gene, comprising <50% of the sequence reconstructed on the genome. Moreover, a CpG island was found in the vicinity of the 5′ end of the gene.

Comparison of the virtual cDNA (see Methods) with the human genome assembly build 31 by means of MegaBLAST allowed us to map the proposed gene model on chromosome 22 (NT_011520 contig); it overlaps a hypothetical gene model (locus *LOC220686*), which had no experimental transcript evidence supporting the hypothetical model to date.

## Small Open Reading Frames

Small open reading frames (smORFs) were usually not included in the initial genome annotations. However, genes coding for proteins <100 amino acids are present and functionally active in higher eukaryotes (Kessler et al. 2003). These smORFs, including the putative genes defined above as category 3, could therefore correspond to either noncoding RNA transcripts or genes encoding short proteins, or could be part of larger gene models.

For the category 3 models, we observed an average of 3.4 exons per gene and the presence of a CpG island in the vicinity of the 5′ end for 47% (139/294) of the cases in this category. Because in most of the cases, human genes have a mouse counterpart with highly conserved exonic structure (Waterston et al. 2002; Guigo et al. 2003), we searched for mouse homologs of these putative human gene models (see Methods): 57% of category 3 models (167/294) match the mouse genome in at least one coding frame. Moreover, at least one of the exons predicted for 22% of such models (37/167) was further supported by an ecore shared with *Tetraodon*. As an example, a putative gene (accession nos. BX409763, BX435024, and BX460561) located on the forward strand of human chromosome 3 (Fig. 2B), matches its mouse counterpart on chromosome 6, overlapping the Riken cDNA C630007B19 gene (LocusID *320265*).

TBLASTX searches were also performed against the mouse genome sequence for a sample of 126 monoexonic smORFs, and matches were found in 75 of the cases (60%), 19 of which had a CpG island in the vicinity of the 5′ end. Although it is likely that a fraction of the monoexonic smORFs may correspond to pseudogenes, this strongly indicates that unspliced smORFs may encode a true protein or be part of larger gene models.

**Table 4. Features of Novel and Putative Genes Supported by the CNSLT Resource**

| | Novel genes | Putative genes |
|---|---|---|
| No. of genes (FANTOM2 cDNAs) | 286 (374) | 70 (9) |
| Alignment with the human genome | | |
| No. of genes/no. of clusters | 548/548 | 293/290 |
| No. of genes (no. of FANTOM2 cDNAs)/N. of clusters[a] | 160 (198)/157 | 13 (19)/13 |
| Alignment with the mouse genome | | |
| No. of genes/no. of clusters | 302/298 | 70/68 |
| No. of genes (no. of FANTOM2 cDNAs)/no. of clusters[a] | 172 (222)/168 | 23 (29)/21 |

[a]Clusters shared by CNSLT novel or putative genes and mouse FANTOM2 cDNA clones.

**Table 5. Comparison Between Novel and Putative Genes Supported by the CNSLT Resource and the Ensembl Annotation (ENSG and ENSMUG Genes)**

| | Novel genes | Putative genes |
|---|---|---|
| Alignment with the human genome | | |
| No. of genes/no. of clusters | 548/544 | 293/290 |
| No. of genes (no. of ENSG genes)/no. of clusters[a] | 215 (210)/209 | 26 (24)/24 |
| No. of genes alone/no. of clusters | 335/335 | 268/266 |
| Alignment with the mouse genome | | |
| No. of genes/no. of clusters | 302/298 | 70/68 |
| No. of genes (no. of ENSMUG genes)/no. of clusters[a] | 193 (193)/189 | 12 (11)/11 |
| No. of genes alone in a cluster/no. of clusters | 109/109 | 58/57 |

[a]Clusters shared by CNSLT novel or putative genes and Ensembl genes.

### Extension of the 5′ End of Annotated Genes

We observed a total of 4306 gene models that shared both types of transcripts (RefSeq/GenBank and CNSLT) in which the CNSLT cDNA mapped more 5′ than the alternative resource. An automated alignment procedure based on sim4 indicated that 36% of these 4306 gene models were extended by at least one exon. A sample of this subset of 36% (218 gene models) was subjected to the manual validation procedure used for the potential novel genes (see Methods). This manual curation identified one or more additional exons in 24% of the sample, indicating that altogether ~380 LocusLink models could be substantially extended. Furthermore, a CpG island could be identified in the vicinity of the extended 5′ end in 61% of the 218 models. In addition, an alternative exon could be detected for 12% of this sample in the course of the manual curation.

An example of the extension of an annotated gene, provided by the CNSLT resource is shown in Figure 2C. This gene is supported by the DKFZP434K1772 cDNA (LocusID *54507*), a 12-exon structure located on the forward strand of chromosome 1. The human CNSLT cDNA clone (accession nos. BX329090, BX370116, BX399403, and BX399404; Fig. 2C), extends the annotated gene by seven exons, with a canonical dinucleotide GT–AG pattern for all donor and acceptor sites. Two of these exons are supported by human–*Tetraodon* ecores. Furthermore, this extension allowed the anchoring of the gene in the vicinity of a CpG island. The CDS from this annotated gene model is also extended by using this CNSLT cDNA. We estimate that the new model, initially based on the DKFZP434K1772 hypothetical protein, should now be complete.

### Additional Comparisons to Mouse Genome and Transcripts

To further characterize the set of novel and putative genes, we searched the FANTOM2 cDNA data set (Okazaki et al. 2002) for mouse cDNAs representing possible counterparts (see Methods). Fifty-two percent of the novel genes (286/552), as well as 24% of the putative genes (70/294), had a homolog in the FANTOM2 cDNA data set (Table 4).

In addition, 160 of the novel genes (29%) were covered by at least one FANTOM2 cDNA clone on the same genomic region of the human genome assembly (Table 4). The same procedure was applied by using the mouse genome sequence assembly, resulting in 172 novel genes clustered with at least one FANTOM2 cDNA clone (see Methods).

As it has already been shown, 40% of the human genome can be aligned with the mouse genome at the nucleotide level. Moreover, 99% of mouse genes have a homolog in the human genome, with highly conserved exonic structure (Waterston et al. 2002). The procedure used on the human Ensembl genes, applied to the mouse genome sequence assembly and the annotated Ensembl mouse genes (see Methods), resulted in 109 novel genes mapping to regions not documented by the Ensembl annotation pipeline in the mouse (Table 5).

## DISCUSSION

Integration of information on CpG islands and human–pufferfish ecores into the models proposed for each transcript followed by manual curation allowed us to group 1072 gene models into three main categories: novel, known, and putative genes. Novel genes were identified on all chromosomes except for chromosome Y. Although eight new genes were located on chromosome 21, 16 mapped to chromosome 22, possibly as a reflection of the higher gene content of chromosome 22 (Dunham et al. 1999). Both chromosomes 21 and 22 have been lately updated in their annotation by using new resources (Reymond et al. 2002; Collins et al. 2003). The use of the CNSLT resource resulted in the identification of novel genes even in regions of the genome that have been extensively characterized, even though the number of such new genes is expected to be small. About 20% of the regions supported by the CNSLT cDNA clones were classified as putative, with an exon/intron structure but no CDS according to our criteria (see Methods).

Recently small protein-coding genes (smORFs) were found in *Saccharomyces cerevisiae* by searching potential budding yeast ORFs against sequences from other related and nonfungal species (Kessler et al. 2003). The CNSLT cDNA resource provides a number of additional candidate smORFs that can be validated by further bioinformatic analyses and experimental approaches.

First exons, which are usually partially or completely noncoding, are overlooked by most gene-finding algorithms used for prediction of protein-coding genes. Although new methods are improving the accuracy of first exon predictions (Brent 2002), experimental verification is still a necessity for refining such gene models. The cDNA resource presented in this study allowed the extension of ~380 gene models described in LocusLink at their 5′ end, thereby improving the annotation of these already-known genes.

Sequence comparison to the mouse FANTOM2 cDNAs, together with later alignment with the mouse genome gave further support to the novel genes defined by the CNSLT cDNA resource. Furthermore, the CNSLT resource provides a number of additional candidate genes on the mouse genome, to be confirmed later by experimental approaches.

Even though several groups have focused on the generation of full-length cDNA libraries, there is still a lack of resources for

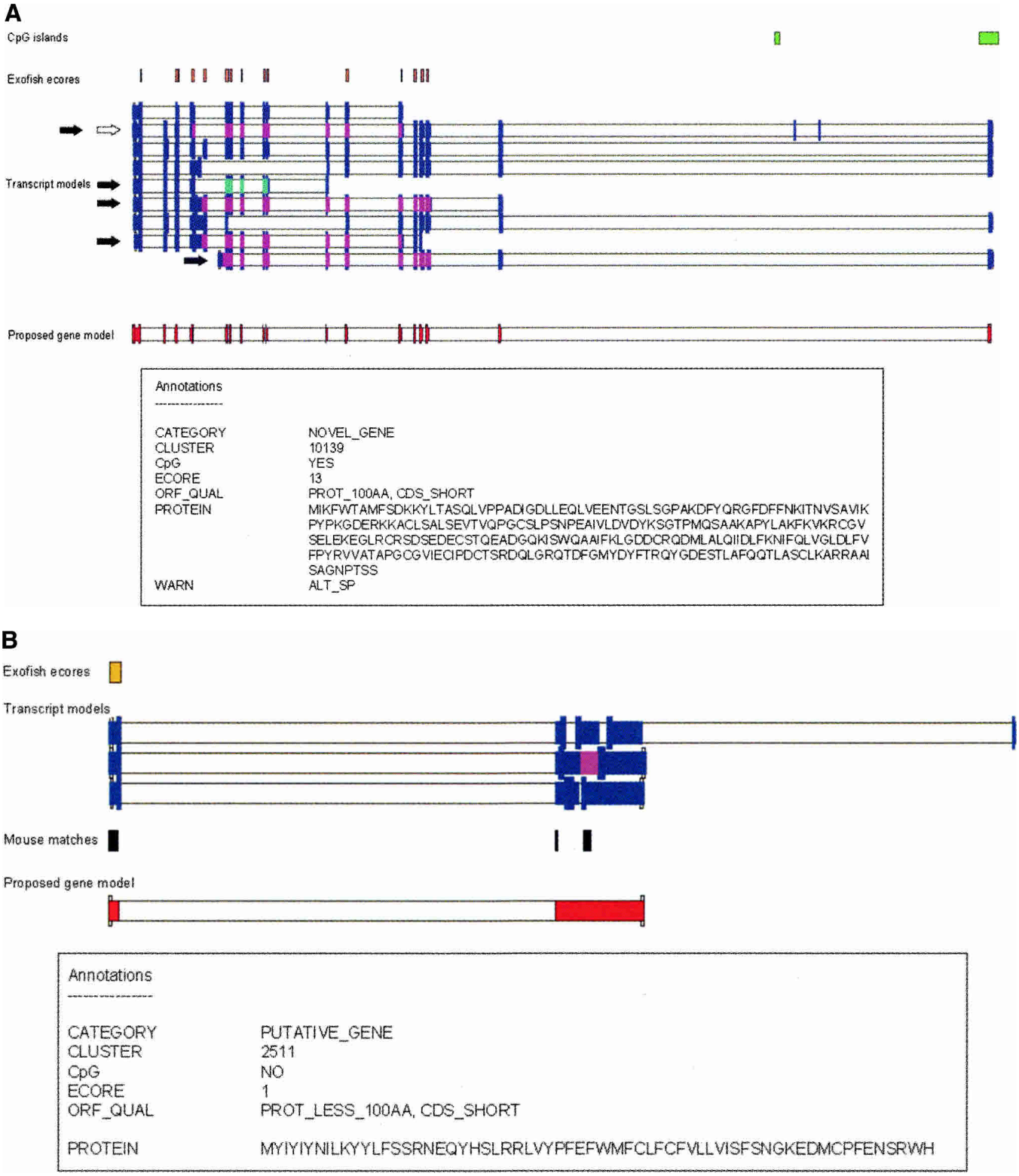


**Figure 2** (Continued on next page)

identifying a significant fraction of genes and transcripts in their entirety (Kristiansen and Pandey 2002). Of these, the CNSLT cDNA resource substantially increases the available human gene catalog. Examples of extension of models already based on full-length cDNAs (Wiemann et al. 2001) illustrate the high efficiency of the procedure used for cDNA construction. The establishment of a transcript catalog, which will eventually feature a nonredundant set of full-coding sequences, supported by clones

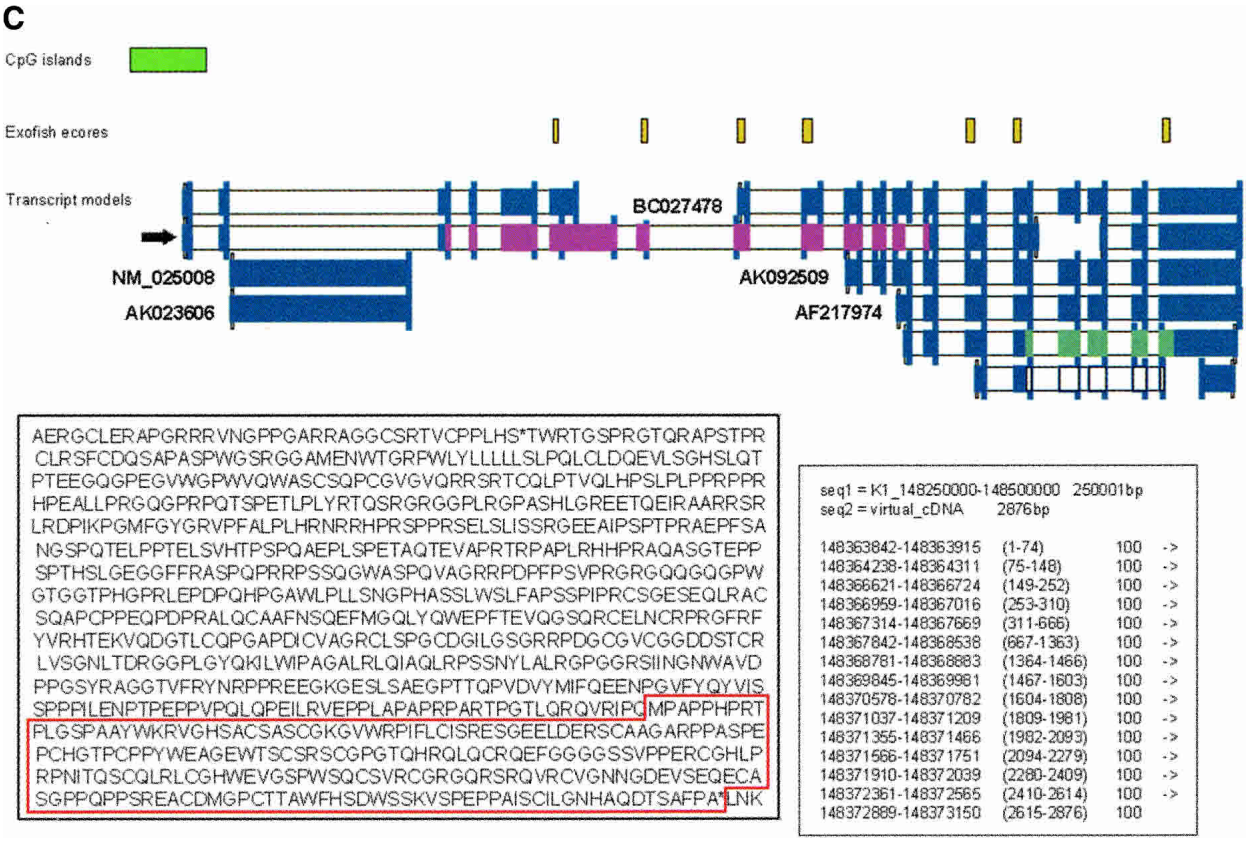


**Figure 2** (*A*) Example of a new gene supported by the CNSLT cDNA resource. Human curation of the identified gene models was performed by using a graphical interface. The transcript models of cDNA clones are represented by blue bars, whereas the model proposed for the gene is represented at the bottom of the figure by red bars (for a detailed explanation, see Results). An *empty arrow* indicates the CNSLT cDNA clone used for the construction of the proposed gene model. *Filled arrows* indicate the cDNA clones assembled on the genome. CpG islands are represented by green boxes; human-*Tetraodon* ecores, by orange boxes. Coding regions with an in-frame stop codon upstream of an initiator ATG are represented by magenta bars. When such stop codons could not be identified, the coding regions are represented by pale turquoise bars. Annotations found for this proposed gene model are listed in the boxed part of the figure. PROT_100AA indicates CDS of at least 300 bp; CDS_SHORT, coding region spanning <50% of the model sequence; and ALT_SP, alternative splicing. (*B*) Example of a putative gene. The transcript models of cDNA clones are represented by blue bars, whereas the model proposed for the gene is represented at the bottom of the figure by red bars (for a detailed explanation, see Results). A human–*Tetraodon* ecore, which supports the first exon of the proposed gene model, is represented by an orange box. A coding region of 64 amino acids with an in-frame stop codon upstream of an initiator ATG is represented by magenta bars. Matches to the mouse genome are indicated by black bars. Annotations found for this proposed gene model are listed in the boxed part of the figure. PROT_LESS_100AA indicates CDS <300 bp. (*C*) Example of an extension of an already annotated gene, using the CNSLT cDNA resource. The transcript models of cDNA clones or RefSeq and GenBank transcripts are represented by filled blue bars. CpG islands are represented by green boxes, and human–*Tetraodon* ecores by yellow boxes. The *filled arrow* points to the CNSLT cDNA clone extending the annotated gene. (*Inset, right*) The exons predicted by the alignment of the virtual cDNA sequence and the human genome assembly, using the sim4 algorithm (for detailed explanation, see text). Color code for coding regions: stop-ATG-stop, magenta bars; ATG-stop, pale turquoise bars; and stop-ATG, white bars. (*Inset, left*) The extension of the CDS from BC027478 (red box) using the CNSLT cDNA resource.

for each human gene transcript, will remain of paramount importance in the study of protein functions.

## METHODS

### Library Construction

The cDNA libraries were generated on the pCMVSPORT6 vector by Life Technologies, a division of Invitrogen Corporation. Briefly, first-strand cDNA was synthesized from polyA+mRNA by using Invitrogen Superscript II RT and an oligo-(dT) primer containing a NotI site. The 5′ end was enriched, and double-stranded cDNA was digested with NotI and cloned into the NotI and EcoRV sites of the pCMVSPORT6 vector.

### Alignment of cDNA Sequences to the Human Genome

Human transcripts identified by a LocusID (as of October 9, 2002) in the LocusLink resource: curated mRNA RefSeq sequences (NM_ sequences) and human GenBank sequences (release 131) were chosen as reference for the annotation. Of these, human transcripts corresponding to dubious or unproven genes, such as LOCs, were excluded from that reference data set. We used a two-step strategy to align the CNSLT cDNA clones or human transcripts on the genomic reference sequence. Preliminary transcript models were created based on the alignments of the 5′ and 3′ EST sequence reads derived from the cDNA clones and the repeat-masked genomic reference sequence (NCBI build 30). The repeats taken into account by the masking procedure were limited to Alu sequences and microsatellites. The HSPs obtained by the BLAST comparisons (W = 20, X = 8; Altschul et al. 1990) were combined in a coherent manner, consistent with their position on the reference sequence. In this way, one or several models were built for each transcript, composed of one or several tentative exons based on the alignment with the reference sequence. The model with the highest total score defined by the sum of the scores of each HSP (total score ≥1000) was selected as the preliminary transcript model that underwent further analysis. The unmasked regions of such preliminary transcript models were extended by 10 kb of genomic sequence on each end and re-

aligned with the cDNA clones by using the sim4 algorithm (A = 4, R = 2; Florea et al. 1998). This procedure defined transcript models with a high fraction of bona fide intron–exon boundaries. These transcript models, together with LocusID-tagged RefSeq and GenBank transcripts, were fused in gene models by a single linkage clustering approach, in which transcript models from the same genomic region and same strand sharing at least 100 bp are merged in a single model.

### Human Curation

Gene models defined automatically were subjected to manual curation in order to redefine, when needed, proposed gene structures, to resolve errors in splice site prediction, and to characterize alternative splicing events. When the use of sim4 yielded problematic alignments, leading to dubious introns, the EST_GENOME algorithm (Mott 1997) was used to realign the cDNA and genomic sequence and verify the location of introns. Human expertise was particularly required to detect dubious data, essentially due to clustering or experimental artifacts (e.g., genomic contamination of RNA used for cDNA library construction). Human expertise was also required to choose the representative transcript that defined the proposed gene model used for further analyses.

Comparative sequence analysis versus the *T. nigroviridis* genome using Exofish (Roest Crollius et al. 2000) and CpG island detection (http://compbio.ornl.gov/grailexp), in +2 kb or −0.5 kb of the first exon, were performed on the proposed models.

### CDS Determination

Exon sequences of a proposed gene model were concatenated in order to reconstruct a so-called "virtual cDNA" on which start and stop codons were identified. Each coding region, delimited by either the start codon or the 5′ extremity of the first exon, and a stop codon or the ending of the last exon were selected for further analysis if their length corresponded to at least 300 bp. The longest potential CDS was retained and chosen as the putative CDS for the proposed gene model.

### Search for Homologous Mouse Genes

Different approaches were used for the identification of candidate counterparts of our novel and putative genes in the mouse genome. In the first instance, putative gene models, as well as single-exon models with no CDS, were compared to the MGSCv3 mouse assembly by using TBLASTX (W = 5, X = 25, T = 75; Altschul et al. 1997) for the detection of potential smORFs. Potential homologs in the mouse genome draft were found by using a cutoff score of $E \leq 1 \times 10^{-4}$ and identity ≥40%.

A comparison against the mouse FANTOM2 cDNA clones was performed for the identification of possible mouse counterparts: The mouse FANTOM2 cDNAs were filtered by establishing the association to a novel and/or putative human gene; in a second step, the selected FANTOM2 cDNAs were mapped and clustered together with the novel and/or putative human genes on the human and mouse genome. Briefly, for each human novel and putative gene model (846 genes), we identified those FANTOM2 mouse cDNAs (data set of 60,770 mouse full-length cDNA clones; Okazaki et al. 2002) that shared a high sequence similarity from a BLAST alignment (Altschul et al. 1990) using W = 11 and X = 13 as settings. The score for each possible mouse cDNA is as follows: human gene pairwise alignment was then calculated as the sum of the score of each HSP (total score). For each novel and putative human gene, only the best match, corresponding to a FANTOM2 mouse cDNA with the highest "total score" (calculated as the sum of the score of each HSP), was retained if the total score ≥250. These FANTOM2 cDNAs (473), together with the subset of FANTOM2 mouse cDNAs, were later aligned on the human reference genomic sequence (NCBI build 33) and MGSCv3 mouse genome assembly (UCSC version mm3) by BLAST. Different settings were used for human–human (W = 20, X = 8), mouse–mouse (W = 20, X = 8), and mouse–human (W = 11, X = 13) comparisons. Single-linkage clustering was performed for each genome as already described above (for details, see Alignment of cDNA Sequences to the Human Genome).

The subset of CNSLT cDNA clones with no match to the build 30 release of the human genome assembly was subjected to the same process.

### Comparison of Novel and Putative Genes Defined by the CNSLT Resource With the Ensembl Annotation

ENSG predicted for the NCBI build 33 of the human genome, corresponding to near full-length cDNA and/or protein sequences already available in the public sequence databases (20,218 known genes), plus novel evidence-based predictions (3081 novel genes), were retrieved by using the EnsMart data mining toolset at Ensembl (http://www.ensembl.org/EnsMart/). The Ensembl genes for which the location was tagged as "unknown chromosome" were excluded from the analysis. The human novel and putative genes defined by using the CNSLT cDNA resource were fused together with the Ensembl genes by a single linkage clustering approach, as already described. The same procedure was performed by using the Ensembl gene models (24,948 ENSMUG genes) predicted for the NCBI build 30 mouse assembly.

## ACKNOWLEDGMENTS

This work was supported by the French Ministry of Research (grant no. 9950275). We thank Carole Dossat and Olivier Jaillon for support on Exofish, Ralph Eckenberg on CDS identification, and Sumitta Samair and Eric Pelletier for support with the presentation of the data. We thank Chris Gruber, Wu Bo Li, and Joel Jessee for cDNA library construction. We wish to thank as well François Sigaux, Philippe Dessen, and Jacques Haiech for helpful discussions at various stages of the project; Susan Cure for her help in writing the manuscript; and the technical staff of Genoscope for its essential contribution to the experimental part of the work.

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., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

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.

Antequera, F. and Bird, A. 1999. CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr. Biol.* **9:** R661–667.

Brent, M.R. 2002. Predicting full-length transcripts. *Trends Biotechnol.* **20:** 273–275.

Collins, J.E., Goward, M.E., Cole, C.G., Smink, L.J., Huckle, E.J., Knowles, S., Bye, J.M., Beare, D.M., and Dunham, I. 2003. Reevaluating human gene annotation: A second-generation analysis of chromosome 22. *Genome Res.* **13:** 27–36.

Deloukas, P., Matthews, L.H., Ashurst, J., Burton, J., Gilbert, J.G., Jones, M., Stavrides, G., Almeida, J.P., Babbage, A.K., Bagguley, C.L., et al. 2001. The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414:** 865–871.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.

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.

Guigo, R., Dermitzakis, E.T., Agarwal, P., Ponting, C.P., Parra, G., Reymond, A., Abril, J.F., Keibler, E., Lyle, R., Ucla, C., et al. 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1019 additional genes. *Proc. Natl. Acad. Sci.* **100:** 1140–1145.

Haas, B.J., Volfovsky, N., Town, C.D., Troukhan, M., Alexandrov, N., Feldmann, K.A., Flavell, R.B., White, O., and Salzberg, S.L. 2002. Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.* **3:** research0029.0021–research0029.0012.

Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K., et al. 2000. The DNA sequence of human chromosome 21. *Nature* **405:** 311–319.
Heilig, R., Eckenberg, R., Petit, J.L., Fonknechten, N., Da Silva, C., Cattolico, L., Levy, M., Barbe, V., De Berardinis, V., Ureta-Vidal, A., et al. 2003. The DNA sequence and analysis of human chromosome 14. *Nature* **421:** 601–607.
Hogenesch, J.B., Ching, K.A., Batalov, S., Su, A.I., Walker, J.R., Zhou, Y., Kay, S.A., Schultz, P.G., and Cooke, M.P. 2001. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106:** 413–415.
Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30:** 38–41.
Kessler, M.M., Zeng, Q., Hogan, S., Cook, R., Morales, A.J., and Cottarel, G. 2003. Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome. *Genome Res.* **13:** 264–271.
Kikuno, R., Nagase, T., Waki, M., and Ohara, O. 2002. HUGE: A database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.* **30:** 166–168.
Kristiansen, T.Z. and Pandey, A. 2002. Resources for full-length cDNAs. *Trends Biochem. Sci.* **27:** 266–267.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Mott, R. 1997. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13:** 477–478.
Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.
Pruitt, K.D., Katz, K.S., Sicotte, H., and Maglott, D.R. 2000. Introducing RefSeq and LocusLink: Curated human genome resources at the NCBI. *Trends Genet.* **16:** 44–47.
Reymond, A., Camargo, A.A., Deutsch, S., Stevenson, B.J., Parmigiani, R.B., Ucla, C., Bettoni, F., Rossier, C., Lyle, R., Guipponi, M., et al. 2002. Nineteen additional unpredicted transcripts from human chromosome 21. *Genomics* **79:** 824–832.
Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., et al. 2000. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat. Genet.* **25:** 235–238.
Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423:** 825–837.
Strausberg, R.L., Feingold, E.A., Klausner, R.D., and Collins, F.S. 1999. The mammalian gene collection. *Science* **286:** 455–457.
Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., Altschul, S.F., et al. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci.* **99:** 16899–16903.
Suzuki, Y., Ishihara, D., Sasaki, M., Nakagawa, H., Hata, H., Tsunoda, T., Watanabe, M., Komatsu, T., Ota, T., Isogai, T., et al. 2000. Statistical analysis of the 5′ untranslated region of human mRNA using "Oligo-Capped" cDNA libraries. *Genomics* **64:** 286–297.
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., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A., et al. 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31:** 28–33.
Wiemann, S., Weil, B., Wellenreuther, R., Gassenhuber, J., Glassl, S., Ansorge, W., Bocher, M., Blocker, H., Bauersachs, S., Blum, H., et al. 2001. Toward a catalog of human genes and proteins: Sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.* **11:** 422–435.
Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. 2000. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7:** 203–214.

## WEB SITE REFERENCES

http://compbio.ornl.gov/grailexp; Grail Experimental Gene Discovery Suite Web site.
http://www.ensembl.org/EnsMart/; EnsMart data mining toolset retrieval of annotated genomes.

*Received April 30, 2003; accepted in revised form December 2, 2003.*