Computational and Biological Analysis of 680 kb of DNA Sequence from the Human 5q31 Cytokine Gene Cluster Region

Kelly A. Frazer, Yukihiko Ueda, Yiwen Zhu, et al.

Genome Res. 1997 7: 495-512
Access the most recent version at doi:10.1101/gr.7.5.495

RESEARCH

Computational and Biological Analysis of 680 kb of DNA Sequence from the Human 5q31 Cytokine Gene Cluster Region

Kelly A. Frazer, Yukihiko Ueda, Yiwen Zhu, Vincent R. Gifford, Maria R. Garofalo, Narla Mohandas, Christopher H. Martin, Michael J. Palazzolo, Jan-Fang Cheng, and Edward M. Rubin[1]

Human Genome Center, Lawrence Berkeley National Laboratory (LBNL), Berkeley, California 94720

With the human genome project advancing into what will be a 7- to 10-year DNA sequencing phase, we are presented with the challenge of developing strategies to convert genomic sequence data, as they become available, into biologically meaningful information. We have analyzed 680 kb of noncontiguous DNA sequence from a 1-Mb region of human chromosome 5q31, coupling computational analysis with gene expression studies of tissues isolated from humans as well as from mice containing human YAC transgenes. This genomic interval has been noted previously for containing the cytokine gene cluster and a quantitative trait locus associated with inflammatory diseases. Our analysis identified and verified expression of 16 new genes, as well as 7 previously known genes. Of the total of 23 genes in this region, 78% had similarity matches to sequences in protein databases and 83% had exact expressed sequence tag (EST) database matches. Comparative mapping studies of eight of the new human genes discovered in the 5q31 region revealed that all are located in the syntenic region of mouse chromosome 11q. Our analysis demonstrates an approach for examining human sequence as it is made available from large sequencing programs and has resulted in the discovery of several biomedically important genes, including a cyclin, a transcription factor that is homologous to an oncogene, a protein involved in DNA repair, and several new members of a family of transporter proteins.

[The sequence data described in this paper are available via the internet at http://www-hgc.lbl.gov/sequence-archive.html.]

The Genome Project has shifted only recently to the sequencing phase for humans (Marshall and Pennisi 1996) while significant progress has already been made on the sequencing of selected model organisms. The genomic sequence of several organisms, including two eubacteria (Fleischmann et al. 1995; Fraser et al. 1995), an archaeon (Bult et al. 1996), and the extensively studied eukaryote, *Saccharomyces cerevisiae* (Walsh and Barrell 1996), have already been completed. The strategy employed to computationally identify and analyze putative genes in these model organisms has consisted of identifying protein-coding open reading frames (ORFs) followed by a search of the databases to determine whether these ORFs are homologs of previously characterized genes. Surprisingly, almost one-half of the protein-coding ORFs revealed during the analyses of these model organism genomes have shown no homology to previously characterized genes (Dujon 1996). In contrast to the genomes of model organisms, in which contiguous and annotated sequence data were released at defined intervals, human genomic sequence is being released to the public domain in noncontiguous minimally annotated fragments. Because the human genome is significantly larger and more complex, it is clear that the approaches employed to analyze it will have to vary from the approaches used previously for the genomes of model organisms.

The annotation of human genomic sequence is facilitated greatly by the availability of the large public expressed sequence tag database (dbEST) so that transcribed regions of the genome can be identified, whether or not homology to a previously characterized gene is present. Even in the absence of protein and EST similarity matches, gene prediction programs can provide important clues about the lo-

[1]Corresponding author.
E-MAIL emrubin@lbl.gov; FAX (510) 486-6816.

cation of new genes in unannotated human genomic sequence. Although many putative new genes can thus be computationally predicted, ascertaining whether the putative new gene is transcribed and the tissues in which it is expressed can only be determined by experiment.

The 5q31 region is of particular interest for large-scale sequencing because of the presence of the cytokine gene family and the fact that a quantitative trait locus associated with inflammatory diseases has been mapped in this region (Marsh et al. 1994; Bleecker et al. 1995; Postma et al. 1995). The cytokines *interleukin-3* (*IL-3*), *IL-4, IL-5, IL-13,* and *granulocyte–macrophage colony-stimulating factor* (*GM-CSF*) are clustered within a 1 Mb region of each other on chromosome 5q31 (Saltman et al. 1993; Nimer and Uchida 1995; Smirnov et al. 1995). Although the loci encoding these proteins are not homologous at the nucleotide or amino acid level, they are considered a gene family because of their localization, overlap in biological activities, and secondary and tertiary structural similarities. It is remarkable that because of the uniqueness of each cytokine, new ones have been discovered not based on homology to old ones but solely using classical methods of cloning genes based on biological activity. Because of the clustering of the cytokine gene family on chromosome 5q31, it has been hypothesized that other, as yet unidentified, interleukins may be located in this region.

In this report we have computationally and biologically analyzed the genomic organization of the cytokine gene cluster region on human chromosome 5q31 and compared it to the genomic organization of the syntenic region on mouse chromosome 11q. As a result of our efforts we have identified a large number of new genes, determined their expression patterns and, in many cases, predicted their possible functions.

RESULTS

Physical Map, Sequencing Strategy, and Contig Assembly

Overlapping P1 and P1-derived artificial chromosome (PAC) clones from the 1-Mb region on chromosome 5 containing the cytokine gene family were isolated by screening genomic libraries (J. Cheng 1996, Chromosome 5 Physical Mapping, http://www-hgc.lbl.gov/clone-info.html) by hybridization and PCR. The minimum tiling contig of the region, which is represented schematically in Figure 1, consists mainly of P1 clones, although one gap in the contig is filled with a PAC clone. The directed sequencing strategy used to generate the data has been described previously (Martin et al. 1995). The sequence of the entire 1-Mb region on chromosome 5 containing the cytokine gene family is being generated and released as assembled fragments ~3 kb in size (LBNL 1996, LBNL/BDGP Sequence Archive, http://www-hgc.lbl.gov/sequence-archive.html). We used the BSPASS program (S. Pitluck 1996, Towards Automated Assembly for the Directed Sequencing Strategy, http://www-hgc.lbl.gov/inf/spass.html) to build the currently available overlapping sets of assembled 3-kb DNA sequence into 34 blocks of sequence, ranging in size from 1.6 to 95.8 kb, which together compose a total of 680 kb of sequence.

Computational Analysis

Seventeen new genes were discovered computationally in the 5q31 region and their expressions verified using the strategy diagrammed in Figure 2. The distribution of these 17 new genes as well as the 6 genes previously known to lie in 5q31 are illustrated in Figure 1.

Comparison of protein translations of the 5q31

Figure 1 Physical map of the megabase region containing the cytokine gene cluster on human chromosome 5q31. The scale on the *left* is in kilobases. In the middle, YAC clones are represented by the medium-length bars, P1 and PAC clones are represented by the shorter bars. The long bar on the right represents a composite map of the overlapping P1 and PAC (H37) clones in the region. The regions depicted in red have been sequenced and analyzed in this study while the regions depicted in black have not been sequenced and/or analyzed. Known and putative genes are color coded according to the computational method by which they were identified as described in Fig. 2. (*Dark blue*) Genes that were sequenced previously. Except for Ril, all of the genes that were sequenced previously had also been localized to this region of 5q31. (*Light blue*) Putative genes identified by homology to known proteins. (*Green*) Putative genes identified by EST matches. (*Purple*) Putative genes identified by analysis of GRAIL-predicted exons and verified by expression studies. The genomic regions containing known and putative genes that have protein database matches are indicated by an increase in the width of the composite map bar. The direction of transcription for each gene is indicated by a vertical arrow. Locations of identical EST matches are indicated by the horizontal arrows on the *left* side of the composite map. The number next to the arrow indicates how many EST matches were found at each location. An arrow without a number represents a single EST match.

sequence data with the proteins in GenPept identified 18 matches: 6 (33%) were exact matches to known genes mapped previously in this region, 1 (6%) was an exact match to a previously identified human gene whose location had not yet been determined, 4 (22%) were orthologs of characterized genes in the rat, mouse, and *Drosophila* organisms, and 7 (39%) were partial matches that ranged from being recognizable motifs to being moderately homologous with known genes (Table 1; Figs. 1 and 3).

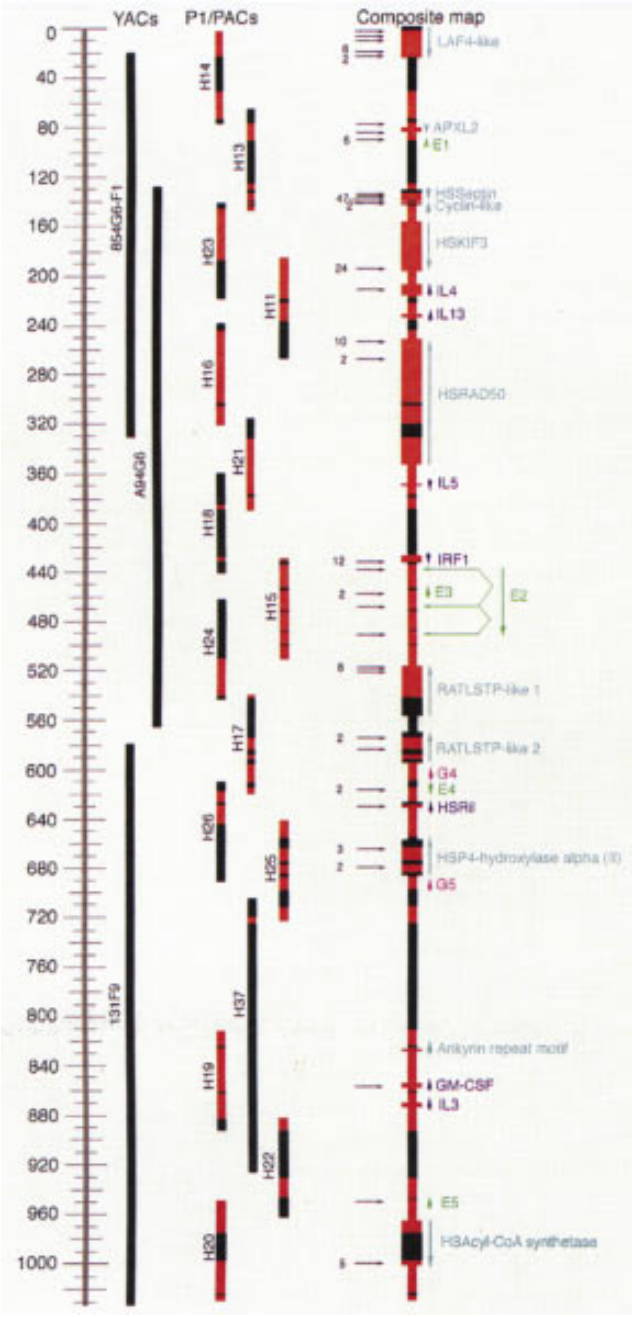

Figure1 (*See facing page for legend.*)

Analysis of these GenPept database matches localized, determined the direction of transcription, and predicted the functions of the new 5q31 putative genes based on their similarities to known genes. In addition, comparison of human and mouse or rat orthologous genes frequently identified the splice sites in the human gene.

Similarity searches against dbEST localized 155 exact matching human and 19 highly similar (85% or greater identity at nucleotide level) mouse ESTs in this 1-Mb region of 5q31. To avoid using an arbitrary percent nucleotide identity cutoff, high scoring human ESTs were assessed to be an exact match or not based on individual inspection. Analysis of the human EST data revealed 19 genes that contain at least one exact match, of which 4 were known genes, 10 had been identified in the GenPept database searches based on similarities to known genes (Table 1), and 5 were new putative novel genes (Table 2). We named these new putative novel genes E1–E5 to indicate that they had been identified by EST matches. Further analysis of the human and mouse EST matches localized and determined the direction of transcription of the E1–E5 genes, as well as yielded information about the spatial expression patterns of the putative genes in 5q31 and the relative abundance and splice sites of their transcripts (Figs. 1 and 3). For example, the *Homo sapiens* (*HS*) *septin* putative gene had 51 exact human EST matches, obtained from six different tissues, thus suggesting that it is a highly expressed ubiquitous gene. On the other hand, *E1, E2, E3, E5, HSRil, cyclin-like, IL-4,* and *GM-CSF* had only one or two exact EST matches and *IL-13, IL-5,* and *IL-3* had no exact EST matches, suggesting that these genes are either expressed at low levels or in a tissue-specific manner, as is known to be the case for the interleukin genes.

The gene prediction program GRAIL was employed to locate exons in the 5q31 sequence. GRAIL predicted a total of 484 exons, of which 259 (54%) were associated with known genes or the putative genes identified through the GenPept and dbEST database searches, 67 (14%) were associated with repetitive elements, such as the LINE family called L1, which contain ORFs, and 158 (32%) were completely novel. These 158 novel GRAIL predicted exons were analyzed to identify new genes without either protein or EST database matches. The criteria used to group the GRAIL predicted exons into putative genes maximized the likelihood of identifying new interleukins. All the members of the 5q31 cytokine family, although lacking sequence homology, possess four exons spanning between 2 and 9

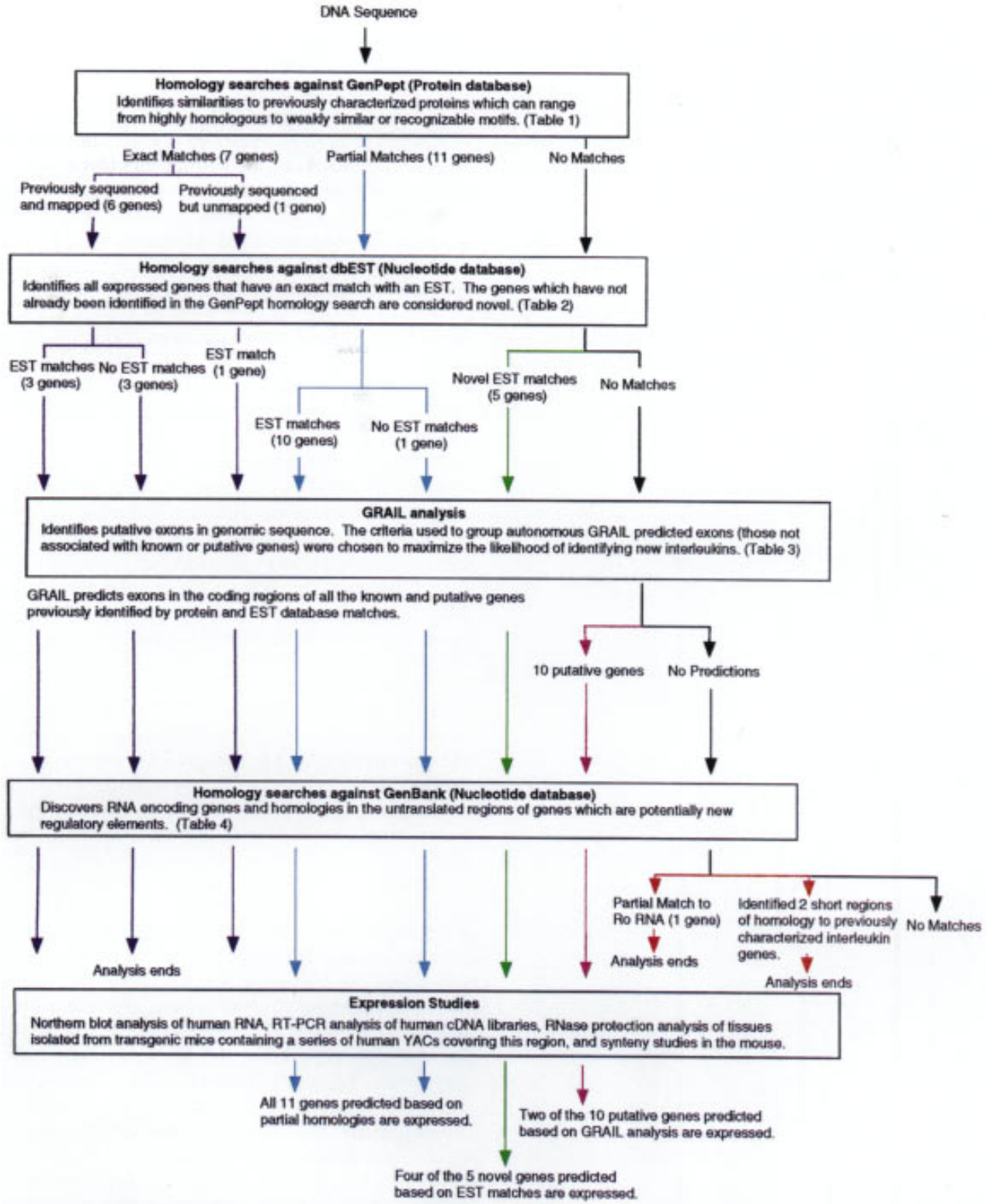

Figure 2 Approach used to computationally and biologically analyze the 5q31 sequence data. The homologies and expression patterns of the putative genes identified during this analysis are described in Tables 1–4.

kb of genomic sequence. Thus, the novel GRAIL-predicted exons were analyzed to identify genomic regions that fit the following criteria: (1) contained four or more excellent or good GRAIL-predicted exons on the same strand of DNA sequence; (2) of the GRAIL predicted exons, at least four had to be located within 9 kb of one another; and (3) the opposite strand could not obviously code for a gene. Ten genomic regions were identified that fit the above criteria and were named *G1–G10* to indicate that they had been identified by GRAIL analysis (Table 3; Fig. 3).

To assess the performance of GRAIL we examined the GRAIL predictions of five genes, *IRF1*, *IL-4*, *IL-5*, *IL-13*, and *GM-CSF*, in which the genomic structure, including the intron/exon boundaries, has been reported previously in detail. Of the 25 coding exons comprising these five genes, GRAIL Ia identified 18 (72%) with a false-positive rate of 10% (percent of predicted exons that are not real), whereas GRAIL II identified 21 (84%) of the exons and had a false-positive rate of 19%. If the GRAIL Ia and II exon predictions are combined, 23 (92%) of the known exons were identified with a false-positive rate of 15%. For the set of 21 exons recognized by GRAIL II, 11 (52%) had both splice junctions predicted correctly and 20 (95%) had at least one splice junction predicted correctly, whereas GRAIL Ia predicted correctly one of the two splice junctions for only 2 (11%) of the exons it recognized, and none of the exons had both splice junctions correctly predicted. It should be noted that GRAIL is a ''learning'' program, and the specific results will vary over time as the genes used in the training set are changed.

Comparison of the 5q31 DNA sequence data with the DNA sequences in GenBank identified a putative new RNA encoding gene and two regions, which are potential DNA regulatory sequences, homologous to sequences in the untranslated regions of known interleukin genes (Table 4; Fig. 3). Both of these regions of homology were short but statistically significant; one was to a region 5′ of *IL-13* (86 bp) (91% identity), and the other was to the third intron of *IL-4* (52 bp) (78% identity); these regions were named IL-13SH and IL-4SH, respectively.

Biological Analysis of Computationally Predicted Genes

Expression Studies Using Materials Developed from Human RNA

To biologically verify and determine the expression patterns of the computationally predicted genes, a variety of expression studies were performed (Tables 1–3). The putative genes were first examined for expression by RT–PCR analysis using four human cDNA libraries (infant brain, HeLa, placenta, and T-cell), and oligonucleotide primers were chosen based on homology matches and GRAIL predictions. RT–PCR analyses demonstrated that 10 of the 11 putative genes identified by similarities to known proteins, 4 of the 5 novel genes identified by EST matches, and 2 of the 10 putative genes predicted based on GRAIL analysis were expressed. Northern blot analyses were used to examine the transcript sizes and tissue distributions of the putative new genes identified by protein and EST database matches (Tables 1 and 2). Northern blot analyses were also used to decipher questions arising from the computational studies of the 5q31 sequence data. In several cases, it was unclear whether an EST corresponded to a particular gene. One reason was that the GenPept and dbEST database matches did not overlap, for example, the *APXL2* putative gene appears to be located in the intron of EST (N48057). An alternative reason was that the gene sequence was incomplete, as in the case of *HSacyl-CoA* (coenzyme A) synthetase and the ESTs at its 3′ end (Fig. 3). Examination of transcript sizes and expression patterns by Northern blot analyses determined that *APXL2* and EST (N48057) derive from same gene and similarly for *HSacyl-CoA* synthetase and the ESTs at its 3′ end.

Expression Studies Using Human YAC Transgenic Mice

To develop substrates for examining expression patterns and possible functions of the identified putative genes, we created a panel of transgenic mice containing three human yeast artificial chromosomes (YACs), 854G6-F1 (350 kb), A94G6 (450 kb), and 131F9 (500 kb), spanning 900 kb of the 5q31 region (Fig. 1). Previous studies of transgenic mice containing human genes on large insert vectors, such as YACs, bacterial artificial chromosomes (BACs), and P1s, have shown that the human transgenes are usually expressed in an appropriate spatial- and temporal-specific manner (Frazer et al. 1995; Smith et al. 1995). RNase protection assays, which are highly specific and detect a single-base-pair difference between the probe and the transcript being analyzed, were performed on numerous tissues isolated from the 5q31 YAC transgenic mice using probes for the *HSseptin*, *HSKIF3*, and *HSRAD50* putative genes (Table 1). Expression of the *HSseptin*

Table 1. Expression Analysis of 12 New Putative Genes on 5q31 Discovered Based on

Putative gene	Homology, function, and related comments[a]	BLASTX (*P* value)	RT–PCR analysis of cDNA libraries[b]
LAF-4-like	Homo sapien AF-4 (20-80%) (L13773) and LAF4 (23-80%) (U34360) proteins. AF-4 and LAF4 are putative transcription factors which may function in lymphoid development (Ma and Staudt 1996). AF-4 is an oncogene that is translocated in t(4:11) acute lymphoblastic leukemias (Domer et al. 1993; Morrissey et al. 1993; Nakamura et al. 1993).	LAF4: 3.3e-44 AF4: 1.8e-42	(+) Brain (+) HeLa (+) Placenta
APXL2	Xenopus laevis APX (23-61%) (Z14997) and Homo sapien APXL (22-56%) (X83543) proteins. APX may play a role in the regulation of amiloride-sensitive sodium channels (Staub et al. 1992). The function of APXL has not been experimentally determined (Schiaffino et al. 1995).	APX: 2.0e-14 APXL: 2.4e-4	(+-) Brain (+) HeLa (+) Placenta
HSSeptin2	Homologous to numerous members of a family of cytoskeletal proteins required for cytokinesis (Fares et al. 1995). Possibly is the ortholog of the Drosophila melanogaster Sep2 (49-76%) (U28966) and Saccharomyces cerevisiae CDC12 (33-64%) (X82498) proteins.	Sep2: 1.5e-109 CDC12: 1.2e-40	(+) Brain (+) HeLa (+) Placenta
Cyclin-like	Homologous to 12 cyclins in the conserved domain known as the "cyclin box" (Hadwiger et al. 1989). Cahnorhabditis elegans Ro2f2.1 (28-50%) (U00055) and Mus musculus cyclin G (28-38%) (Z37110) proteins. Cyclin-like most likely comprises a new family of cyclins.	Ro2f2.1: 1.0e-3 cyclin G: 7.5e-3	(+) Brain (-) Placenta
HSKIF3	Homologous to more than 50 different kinesins which are microtubule-associated motor proteins involved in intracellular transport. Most likely is the ortholog of the Mus musculus KIF3 (93-100%) (D12645) protein. KIF3 has been reported as being expressed predominantly in the murine brain (Aizawa et al. 1992).	KIF3: 0.0	(+) Brain (-) HeLa (-) Placenta
HSRAD50	Most likely is the ortholog of the Saccharomyces cerevisiae RAD50 (41-79%) (X14814) protein. RAD50 is required for DNA repair and meiosis-specific double-strand break formation (Johzuka and Ogawa 1995).	RAD50: 5.7e-44	(+) Brain (+) HeLa (+) Placenta
RATLSTP-like1	Homologous to 9 members of the transporter family of proteins. Most similar to the Rattus norvegicus RATLSTP (29-61%) (L27651) protein. RATLSTP is a putative liver specific transporter protein (Simonson et al. 1994).	RATLSTP: 3.1e-27	(+) Brain (+) HeLa (+) Placenta (+) T-Cell
RATLSTP-like2	Homologous to 2 members of the transporter family of proteins. Most similar to the Rattus norvegicus RATLSTP (42%) (L27651) protein. RATLSTP is a putative liver specific transporter protein (Simonson et al. 1994).	RATLSTP: 3.4e-3	
HSRil	Most likely is the ortholog of the Rattus norvegicus ril (93%) (X76454) protein. The ril protein contains LIM/double zinc finger domains and is likely to be involved in the regulation of normal growth control (Kiess et al. 1995). The mRNA encoding for ril in Homo Sapiens (X93510) has previously been isolated but not mapped or characterized.	Ril: 1.6e-12	
HSProlyl4-hydroxylase alpha(II)	Most likely is the ortholog of the Mus musculus alpha(II) subunit (83-100%) (U16163) of the prolyl4-hydroxylase protein. Prolyl4-hydroxylase protein catalyzes the post-translational formation of 4-hydroxyproline in collagens (Helaakoski et al. 1995).	alpha(II): 6.6e-67	(+) Brain (+) HeLa, (+) Placenta (-) T-Cell
Ankyrin repeat motif	Homologous to twelve proteins in the conserved 33 amino acid "ankyrin repeat motif". Homo Sapiens ANKY (40-60%) (X16609) and Mus musculus Ank-1 (40-60%) (X69063) proteins. The ankyrin repeat is a common motif found in gene products with diverse functions; it is believed to be involved in highly specific protein-protein interactions (Peters and Lux 1993).	ANKY: 3.9e-08 Ank-1: 1.1e-07	(+) Brain (+) HeLa (+) Placenta (+) T-Cell
HSAcyl-CoA synthetase	Most likely is the ortholog of the Rattus norvegicus brain long-chain acyl-CoA synthetase (LACS) (81-92%) (D10041). LACS play important roles in fatty acid metabolism. This LACS has been reported as being predominantly expressed in the brain (Fujino and Yamamoto 1992).	LACS: 1.7e-130	(+) Brain (-) HeLa (-) Placenta (+) T-Cell

[a]The percent amino acid identity and NCBI accession numbers of the homologous proteins are indicated in brackets.
[b]Signal strengths are represented as follows: Strong (++); moderate (+); weak (+−); negative (−).
[c]Two ESTs can be developed from the 5′ and 3′ ends of the same cDNA clone. The number of cDNA clones isolated from each tissue

Homology to Previously Characterized Genes

Number of tissue derivations of EST matches[c]	Size and expression patterns of transcripts by Northern blot analysis[b]	RNase protection assay using YAC transgenic mice[b]
Total number of ESTs: 11 Number of cDNA clones: 8 Aorta (1) (D63072) Retina (1) (W26686) Fetal Liver-spleen (5) (N78208) Hippocampus (1) (M79022)	10 kb Heart (+) Placenta (+) Liver (+-) Kidney (-) Brain (+-) Lung (+-) Skeletal Muscle (+) Pancreas (+)	
Total number of ESTs: 2 Number of cDNA clones: 1 Multiple sclerosis (1) (N48057) The region of APXL2 homologous to APX and AXPL lies within an intron of this EST.	3.5kb 7.2 kb 9.0 kb Heart (+-) (+-) (+-) Brain (+-) (+-) (+-) Placenta (+-) (+-) (+-) Lung (+-) (-) (-) Liver (++) (++) (++) Skeletal Muscle (-) (-) (+-) Kidney (-) (-) (-) Pancreas (+-) (+-) (+-)	
Total number of ESTs: 51 Number of cDNA clones: 37 Fetal Liver-spleen (16) (AA002107) Brain (11) (H46577) Placenta (7) (R66031) Multiple sclerosis (1) (N59157) Retina (1) (W22101) Colon (1) (D25728)	2 kb 2.9kb 4 kb 7 kb Heart (+) (+) (+) (+-) Brain (++) (++) (+-) (+-) Placenta (+-) (-) (+) (+-) Lung (+-) (-) (+-) (-) Liver (-) (-) (-) (+-) Skeletal Muscle (+-) (+-) (+-) (-) Kidney (+-) (-) (+-) (-) Pancreas (+-) (+-) (+) (+-)	Brain (++) Thymus (++) Liver (++) Kidney (++) Heart (++) Skeletal muscle (++) Skin (++) Small intestine (++) Lungs (++) Spleen (++) Stomach (++) Colon (++)
Total number of ESTs: 2 Number of cDNA clones: 2 Multiple sclerosis (1) (N50853) Brain (1) (R53169)	3 kb Heart (+) Placenta (-) Liver (-) Kidney (-) Brain (++) Lung (+-) Skeletal Muscle (+-) Pancreas (+)	
Total number of ESTs: 25 Number of cDNA clones: 18 Melanocyte (6) (N44937) Brain (5) (R51130) Placenta (3) (H12783) Multiple sclerosis (1) (N48093) Retina (1) (AA001064) Breast (1) (H44415) Heart (1) (W75947)	7.1 kb Heart (+-) Brain (+) Placenta (+-) Lung (-) Liver (-) Skeletal Muscle (+-) Kidney (-) Pancreas (+-)	Brain (++) Thymus (+) Liver (+-) Kidney (+-) Heart (+-) Skeletal muscle (+-) Skin (+-) Small Intestine (+-) Lungs (+-) Spleen (+-) Stomach (+-) Colon (+-)
Total number of ESTs: 12 Number of cDNA clones: 8 Fetal liver-spleen (4) (R98868) Melanocyte (1) (N24479) Placenta (2) (H93370) Retina (1) (H92341)	4kb 6.5kb 7kb 8kb Heart (+-) (+-) (+-) (+-) Brain (-) (-) (-) (-) Placenta (+-) (+-) (-) (-) Lung (-) (-) (-) (-) Liver (+-) (+-) ( ) (-) Skeletal Muscle (+-) (+-) (+-) (-) Kidney (-) (-) (-) ( ) Pancreas ( ) ( ) (+-) (+-)	Brain (+) Thymus (+) Liver (++) Kidney (+) Heart (+) Skeletal muscle (+) Skin (+) Small Intestine (+) Lungs (+) Spleen (+) Stomach (+) Colon (+)
Total number of ESTs: 9 Number of cDNA clones: 7 Fetal liver-spleen (2) (T84999) Breast (1) (R72710) Colon (1) (D25712) Retina (1) (H85369) Brain (1) (Z42456) Unknown (1) (Z38659).		
Total number of ESTs: 3 Number of cDNA clones: 3 Breast (2) (R55616) Pancreas (1) (W60592).	3.8 kb Heart (+-) Placenta (+-) Liver (-) Kidney (-) Brain (-) Lung (-) Skeletal Muscle (+-) Pancreas (-)	
Total number of ESTs: 1 Number of cDNA clones: 1 Breast (1) (R48104).		
Total number of ESTs: 6 Number of cDNA clones: 6 Fibroblast (1) (W48584) Melanocyte (1) (N42517) Parathyroid tumor (1) (W56255) Heart (1) (W81160) Brain (1) (R21366) Unknown (1) (F13103)	2 kb 5 kb 6.5 kb Heart (++) (+-) (+-) Brain (-) (-) (-) Placenta (+) (+-) (+-) Lung (+-) (-) (-) Liver (+-) (+-) (-) Skeletal Muscle (+) (+-) (+-) Kidney (-) (-) (-) Pancreas (++) (+-) (+-)	
None	Not Detected	
Total number of ESTs: 5 number of cDNA clones: 7 Fetal liver spleen (2) (T24040) Multiple sclerosis (1) (N47598) Retina (1) (H84656) THC[d] (L49746) comprised of cDNA clones isolated from the liver (2) and the eye (1)	2.7 kb 7 kb Heart (-) (-) Placenta (-) (-) Liver (-) (-) Kidney (-) (-) Brain (+) (+) Lung (-) (-) Skeletal Muscle (-) (-) Pancreas (-) (-)	

is indicated in the first bracket. The NCBI accession number of only one EST derived from each tissue is given in the second bracket. [d]Tentative human consensus sequences (THCs) are composed of two or more cDNA clones that have been assembled based on overlap into a larger block of sequence than a typical EST (Adams et al. 1995).

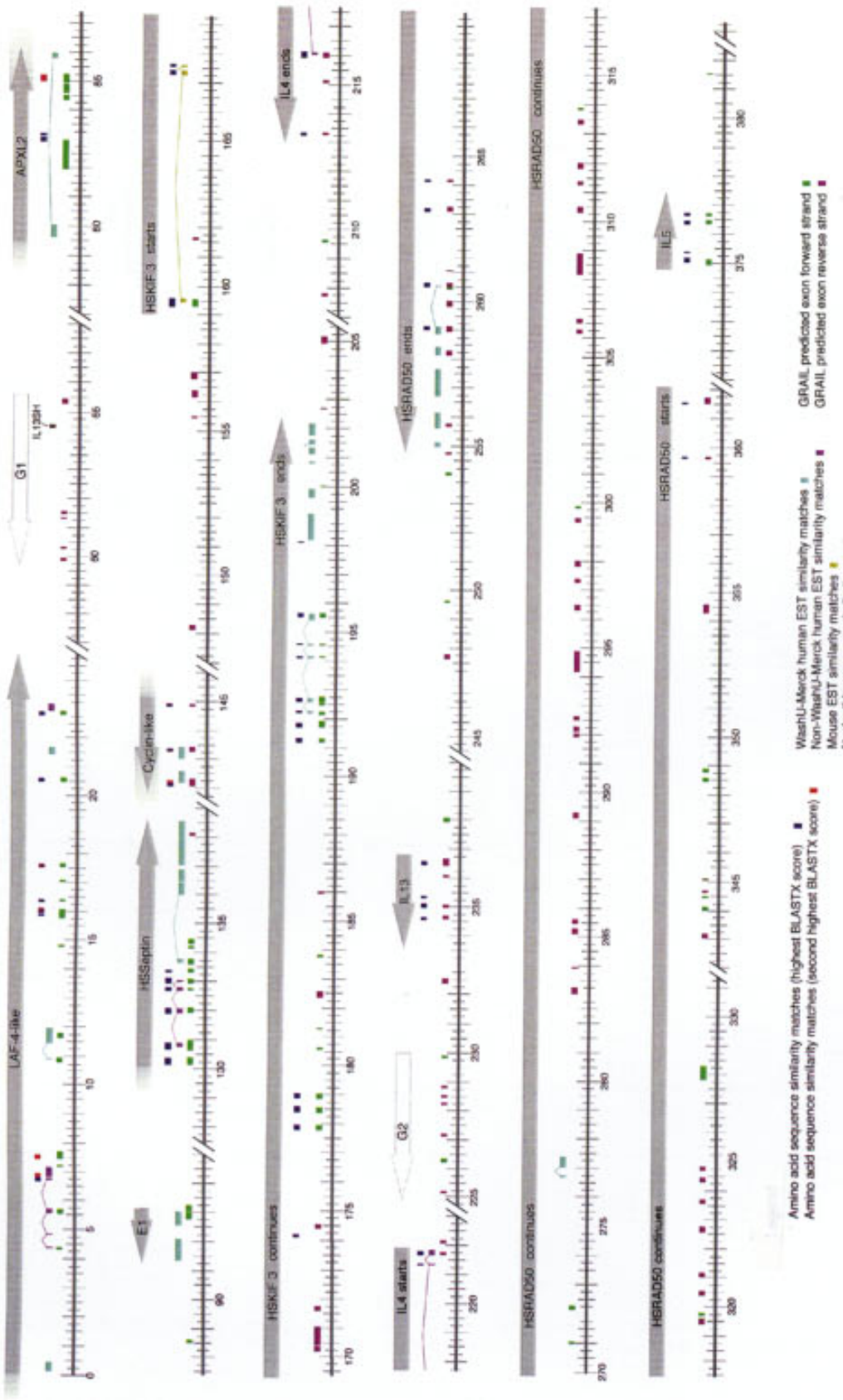

Figure 3 (*See p. 504 for legend.*)

Figure 3 (*Continued*)

gene was detected in all tissues examined by RNase protection assays and Northern blot analysis of human RNA. RNase protection assays are more sensitive than Northern blots; therefore, expression of the *HSKIF3* and *HSRAD50* genes was detected in some tissues by the former method of analysis but not by the latter. Nevertheless, tissues in which expression by Northern blot analysis was detected typically had stronger signals in the RNase protection assays than other tissues.

The 5q31 YAC transgenic mice that we have developed will serve as valuable reagents to study the details of expression, genomic organization, and biological properties of genes in this region. This is exemplified by RNase protection assays of several GRAIL-predicted exons in the *HSRAD50* gene, some of which are not homologous with the yeast *RAD50* gene. These GRAIL predicted exons all displayed identical patterns of expression and thereby supported the hypothesis suggested by the 5′ and 3′ similarity matches of the human and yeast *RAD50* genes, namely that *HSRAD50* is a single gene >100 kb in length.

Comparative Mapping of the 1-Mb Region Containing the Interleukin Gene Family in the Mouse and Human Genomes

The mouse interleukin gene family is located on the long arm of chromosome 11 in a region syntenic with human chromosome 5q31 (DeBry and Seldin 1996). The genes known to lie in this segment include (from proximal to distal) *IL-4*, *IL-13*, *IL-5*, *IRF1*, *GM-CSF*, and *IL-3*. To ascertain whether the new human genes that we identified in 5q31 are also located in this syntenic region of the mouse genome, we isolated a series of mouse YACs by PCR using primers directed to the mouse *IRF1* and *GM-CSF* genes. These mouse YACs were tested for the presence or absence of eight of the new genes (Fig. 4) by Southern analysis using hybridization probes generated from human genes. All eight of the human genes hybridized to the mouse YACs. Restriction analysis of the mouse YACs also indicated that the eight genes are in the same proximal-to-distal order in the mouse and human genomes (data not shown). These comparative mapping studies demonstrate that at least a significant fraction of the new human genes identified through our analysis of 5q31 sequence data is syntenic in the mouse and human genomes.

DISCUSSION

The approach that we employed to identify genes in a large segment of the human genome relied first on searching for sequence similarities in protein and EST databases, second on coding potential predictions, and finally on a panel of biological confirmation studies. This is different from the approach used to analyze the genome of *S. cerevisiae,* as well as other model organisms, which first identified ORFs and then compared the ORFs with sequence databases to determine protein and nucleic acid homologies. This latter approach was successful for the analysis of the yeast genome because of both its compactness (intergenic regions are short and introns are rare), and its simplicity—only 6%–7% of ORFs do not correspond to real genes (Dujon 1996). Because the overwhelming majority of yeast ORFs are real genes, biological confirmation of expression is not necessary for initial purposes. In comparison, the human genome is very complex. It contains large intergenic and intronic sequences, numerous repetitive elements some of which contain protein coding ORFs, and genes that are typically composed of multiple exons whose boundaries are difficult to predict. Although the strength of human gene rec-

Figure 3 Summary of the computational analysis of the DNA sequence data in the 5q31 region. The scale is in kilobases. The locations and directionality of the putative genes are indicated by arrows and were determined based on database matches and/or GRAIL predictions. Similarity matches to known proteins, ESTs, and to the noncoding regions of known genes are color-coded as described in the key. If the conceptual translation of a putative gene was homologous to more than one protein then the locations of the similarity matches to those proteins with the highest and second highest BLASTX scores are indicated. Matches with human ESTs generated by the WashU–Merck Human EST project are distinguished from matches with ESTs generated by other groups. Only exact human EST matches and mouse EST matches with 85% or greater nucleotide identity are shown. If an EST contained gaps and matched the genomic sequence in more than one place, this was presumed to be attributable to the splicing of an intron and the EST segments were joined together by a line. GRAIL predicted exons in the forward strands are distinguished from those in the reverse strand by their color, as described in the key. If GRAIL predicted exons at the same location in both the forward and reverse DNA sequence strands, those that correspond to the coding strand (i.e., those GRAIL predictions having the same direction as the putative gene) are shown.

Table 2. Expression and Structural Analysis of Five Putative Genes Discovered by EST Matches

Putative gene	Number and tissue derivations of EST matches[a]	BLASTN (*P*-value)	GRAIL–predicted exons and related comments[b]	RT–PCR analysis of cDNA libraries	Size and expression patterns of transcripts by Northern[c]
E1	Total number of ESTs: 9 Number of cDNA clones: 6 Brain (3) (R51144) Multiple sclerosis (3) (N49602)	R51144: 1.0e-142	The ESTs which comprise E1 are most likely in the 3' untranslated region of the gene. Three excellent GRAIL predicted exons exist in the non-coding strand.	(+) Brain (+) HeLa (-) Placenta (-) T-Cell	3.6 kb Heart (-) Brain (++) Placenta (-) Lung (-) Liver (-) Skeletal Muscle (-) Kidney (+-) Pancreas (+)
E2	Total number of ESTs: 2 Number of cDNA clones: 1 Lung (1) (W07522).	W07522: 2.7e-133	Eight exellent, six good and four moderate predicted exons spanning approximately 54 kb. The cDNA clone from which EST W07522 is derived also spans approximately 54 kb.	(+) Brain (+) HeLa, (+) Placenta (+) T-Cell	Not Detected
E3	Total number of ESTs: 2 Number of cDNA clones: 1 Fetal liver-spleen (1) (R92251)	R92251: 1.7e-118	One excellent and one good GRAIL predicted exons spanning approximately 1 kb. Gene E3 is in an intron of Gene E2.	(+) Brain (+) HeLa (+) Placenta (+) T-Cell	9.5kb 11.5 kb 14 kb Heart (+-) (+-) (+-) Brain (-) (-) (-) Placenta (+-) (-) (-) Lung (-) (-) (-) Liver (+-) (-) (-) Skeletal Muscle (+-) (+-) (+-) Kidney (-) (-) (-) Pancreas (+-) (+-) (+)
E4	Total number of ESTs: 2 Number of cDNA clones: 1 Fetal liver-spleen (1) (W89188) Numerous ESTs which are 80-90% identical to the genomic sequence are located in the same region as W89188.	W89188: 8.9e-144	Three excellent, seven good and two moderate GRAIL predicted exons spanning approximately 7.3 kb of genomic sequence.	(-) Brain (+) HeLa (-) Placenta (+) T-Cell	Not Detected
E5	Total number of ESTs: 1 Number of cDNA clones: 1 Retina (1) (H85711)	H85711: 1.3e-16	Two good GRAIL predicted exons spanning approximately 3.5 kb.	(-) Brain (-) HeLa (-) Placenta (-) T-Cell	Not Detected

[a]Two ESTs can be developed from the 5′ and 3′ ends of the same cDNA clone. The number of cDNA clones isolated from each tissue is indicated in the first set of parenthesis. The NCBI acession number of only one EST developed from each tissue is given in the second set of parenthesis.

[b]Predicted exons from GRAIL versions Ia and II are counted separately even if the two GRAIL versions predicted overlapping exons. Only GRAIL–predicted exons in the coding strand are included unless otherwise stated.

[c]Signal strengths are represented as follows: Strong (++); moderate (+); weak (+−); negative (−).

Table 3. Expression and Structural Analysis of 10 Putative Genes Identified by Analysis of GRAIL-Predicted Exons

Potential coding region	GRAIL–predicted exons and related comments[a]	RT–PCR analysis of cDNA libraries
G1	Two excellent and four good GRAIL predicted exons spanning approximately 5.6 kb of genomic sequence.	(–) Brain (–) HeLa (–) Placenta (–) T-Cell
G2	One excellent, four good and one moderate GRAIL predicted exons spanning approximately 3.7 kb of genomic sequence.	(–) Brain (–) HeLa, (–) Placenta (–) T-Cell
G3	Two excellent and two good predicted exons spanning approximately 5.4 kb of genomic sequence	(–) Brain (–) HeLa (–) Placenta (–) T-Cell
G4	One good GRAIL predicted exon approximately 120 bp in size was expressed. In the opposite strand there are three excellent, two good and one moderate GRAIL predicted exons spanning approximately 10.9 kb of genomic sequence, however, several of these predicted exons were tested by RT-PCR analysis and were not expressed.	(+) Brain (+) HeLa (–) Placenta (+) T-Cell
G5	Three excellent and one good GRAIL predicted exons spanning approximately 7.2 kb of genomic sequence.	(+) Brain (–) HeLa (+) Placenta (+) T-Cell
G6	Two excellent and two good GRAIL predicted exons spanning approximately 1.3 kb of genomic sequence.	(–) Brain (–) HeLa (–) Placenta (–) T-Cell
G7	One excellent and four good GRAIL predicted exons spanning approximately 1.3 kb of genomic sequence.	(–) Brain (–) HeLa (–) Placenta (–) T-Cell
G8	Two excellent, two good and one moderate GRAIL predicted exons spanning approximately 3.3 kb of genomic sequence.	(–) Brain (–) HeLa (–) Placenta (–) T-Cell
G9	One excellent and seven good GRAIL predicted exons spanning approximately 4.0 kb of genomic sequence.	
G10	Three excellent and one good GRAIL predicted exons spanning approximately 3.9 kb of genomic sequence.	(–) Brain (–) HeLa, (–) Placenta (–) T-Cell

[a]Predicted exons from GRAIL versions Ia and II are counted separately even if the two GRAIL versions predicted overlapping exons. Only GRAIL-predicted exons in the coding strand are included unless otherwise stated.

be standard practice to analyze large segments of human genomic sequence that contains gaps and frequently will code for multiple genes. It is considerably easier therefore to annotate human sequence by first identifying the locations of putative genes based on protein and EST database matches and to then use gene recognition programs to identify those genes without database matches.

In this study we computationally and biologically annotated 680 kb of noncontiguous genomic sequence in a 1-Mb region of the 5q31 cytokine gene cluster region resulting in the identification and verification of 23 genes, 16 of which had not been reported previously. Approximately one-third of the computationally identified new 5q31 genes were not homologous to previously characterized genes. These results suggest that similar to *S. cerevisiae* and the other model organisms, a large proportion of the human genes remaining to be discovered will be novel and lack homology with any of the currently existing sequences in databases. Of the 11 new human 5q31 genes that were homologous to known genes, the quality of the sequence similarities ranged from recognizable motifs to highly conserved orthologs of mouse and rat genes. Naturally, the greater the level of sequence similarity between the new 5q31 gene and its homolog, the more confident we were in using that similarity as a form of sequence annotation and assigning a putative function.

ognition programs lies in accurate prediction of coding regions (exons), their weakness is in splicing these exons together to correctly predict gene structure. This is especially problematic if a long genomic sequence contains exons from multiple genes or if it contains some gaps. For the foreseeable future it will

Of the 23 genes identified in 5q31, 19 (83%) had at least one exact matching EST in dbEST. These data are congruent with the recent analysis of the ESTs generated by the WashU–Merck Human EST

Table 4. Analysis of 5q31 DNA Sequence Matches to the Noncoding Regions of Known Genes

Region name	Homologies and related comments[a]	BLASTN (*P* value)
Ro RNA-like	Homologous to the Homo sapien gene hY1 (V00584) encoding a cytoplasmic Ro RNA. Nucleotides: 76–114; identities = 39/39bp (100%) and 38–74; identities = 33/37bp (89%).	V00584: 8.9e-12
IL13SH	Homologous to a region 5' of the human IL13 (U10307) gene. Nucleotides: 1131–1216; identities = 79/86bp (91%). Also homologous to regions 5' of the human IL13 precursor (U31120) gene. Nucleotides: 1181–1306; identities = 107/126bp (84%) and 1165–1198; identities = 33/34bp (97%).	U10307: 4.9e-19 U31120: 4.8e-32
IL4SH	Homologous to a region in intron 3 of the human IL4 (M23442) gene. Nucleotides: 7471–7522; identities = 42/52bp (78%). Also homologous to two regions in intron 3 of the Bos taurus IL4 (U14159) gene. Nucleotides: 829–880; identities = 41/52bp (78%) and 631–659; identities = 23/29bp (79%).	M23442: 4.9e-2 U14159: 4.7e-3

[a]The NCBI accession numbers of the homologous genes are indicated in parentheses.

Project, which indicated that between 50% and 80% of all the human genes have at least one exact EST match (Hillier et al. 1996). Greater than 85% of the human ESTs in dbEST were deposited by this project, which for the most part generated ESTs from oligo(dT)-primed normalized libraries constructed from 17 different tissues. Therefore, it has been predicted that genes that are primarily expressed in cell types and tissues, such as T cells, thymus, testes, and others, not used to generate the WashU–Merck ESTs will be underrepresented in dbEST. This prediction was confirmed in our study, where three of the five interleukins, which are predominantly expressed in T cells, had no EST matches. Despite the fact that the libraries used to generate these ESTs were normalized, one-third of the ESTs in the 5q31 region matched the *HSseptin* gene. This demonstrates that even though normalization brings the frequency of all cDNA clones to within a narrow range, highly expressed transcripts will still be present at a greater frequency. Because the libraries used to generate the WashU–Merck ESTs were oligo(dT)-primed and normalized, which favors truncated clones over their longer counterparts, the majority of the ESTs in dbEST are in the 3′-untranslated regions of genes. These ESTs are extremely useful for gene-based mapping strategies and sequence annotation of the 3′-untranslated regions of genes but are less useful for sequence annotation of coding regions. Because the average aligned nucleotide and amino acid identity of mouse and human orthologous genes is 85% in the coding region (Makalowski et al. 1996), the 400,000 mouse ESTs currently being generated (Washington University and Howard Hughes Medical Institute Mouse EST Project; http://genome.wustl.edu/est/mouse/≠est.mpg.html) are potentially a great resource to annotate coding regions of human genomic DNA, as the libraries used to generate these mouse ESTs should contain longer insert cDNA clones because of improved methods of normalization and subtraction (Bonaldo et al. 1996).

The performance of gene recognition programs in identifying genes lacking both protein and EST database matches depends on the criteria used for analysis. The criteria used in this study to group GRAIL predicted exons into putative genes were chosen in part to identify new interleukins in the 5q31 region. Of the 10 putative genes predicted based on GRAIL analysis, two, G4 and G5, were confirmed to be expressed by RT–PCR analysis. Because G4 and G5 may represent nonhomologous coding regions of the *RATLSTP-like2*, and *HSpropyl4-hydroxylase* genes, respectively, they are not counted as separate genes in this study. The GRAIL-predicted coding regions of several putative genes identified by this analysis, G1, G2, G3, and G8, share features with the GRAIL-predicted coding regions of the interleukin genes (Fig. 3); however, their expressions were not detected by RT–PCR analysis.

It is worthwhile to examine what types of genes would have been missed by our analysis. These omissions would include genes that both lack database matches and whose sequences would not be recognized by GRAIL, such as genes whose tran-

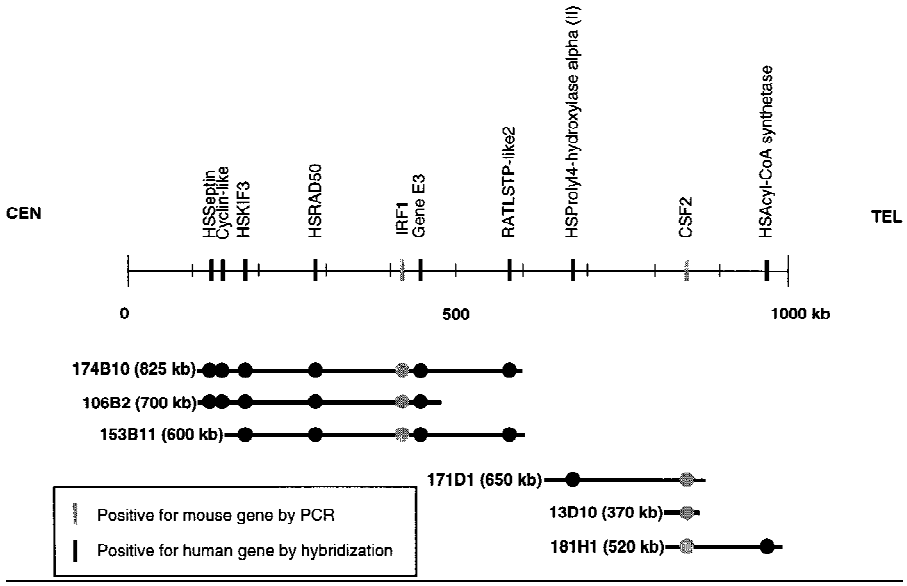

Figure 4 Content mapping of mouse YAC clones from the 11q region. The map at the *top* is based on the relative order and spacing of the genes in the syntenic 5q31 region in the human genome.

scripts are not translated into proteins. Because of the criteria we used to group the novel GRAIL-predicted exons into putative genes, our analysis also would not have identified genes that lack a database match and either code for short proteins or have long introns.

To biologically verify the computationally predicted genes, we used RT–PCR analyses of human cDNA libraries to rapidly determine whether a predicted gene was expressed. We then relied on Northern blot analysis of human RNA and RNase protection assays of 5q31 human YAC transgenic mice to determine transcript sizes and expression patterns, and to clarify ambiguous situations such as whether two expressed sequences are part of the same gene. In general, these two methods of analyses as well as EST tissue-specific frequencies, indicated similar expression patterns of the new 5q31 genes. In several cases, however, one method indicated that a gene was expressed in a particular tissue while the other methods indicated that it was not. Because the substrates used for these analyses were generated from RNA isolated at different developmental stages, the differences observed may be attributable to temporal regulation of gene expression. EST data errors may be another explanation for some of the discrepancies observed. For example, ~2% of the WashU–Merck cDNA clones are derived from intronic or intergenic sequences (Hillier et al. 1996); therefore, it is possible that the single EST match defining the E5 putative gene, which is not detected by the RT–PCR and Northern analyses, may be spurious.

In the future, whole genome expression studies are likely to be performed using automated high-throughput expression technologies (Lander 1996). These technologies, however, will not be suitable for exquisitely examining expression patterns spatially, temporally, and under variable stimuli, which for many genes may offer invaluable clues about function. YAC transgenic mice can provide much of this information and because of the large size of the human genomic insert it is possible to examine multiple human genes in a single line of animals. For example, in this study with three lines of transgenic mice each containing a single human YAC, we have the potential to examine in detail the in vivo properties of >900 kb of human DNA containing >20 genes.

In these studies we have employed a strategy that allows an in-depth computational and biological analysis of human genomic sequence data as they are produced. We have discovered a dense clustering of genes in the 5q31 cytokine gene cluster region and identified several genes that can be tested to determine whether they are the quantitative trait locus associated with inflammatory diseases that has been mapped to this region. Furthermore, we have demonstrated that the new genes identified in the 5q31 region are conserved in content and order in the syntenic region of the mouse genome.

METHODS

Computational Analysis

Alu and other human repetitive elements were identified and masked using a combination of BLASTN version 1.4 (Altschul et al. 1990) (default parameters) searches and XBLAST (Claverie and States 1993; Claverie 1996) filtering against the Alu.327.dna (National Center for Biotechnology Information; ftp://ncbi.nlm.nih.gov/pub/jmc/alu/) and humrep (National Center for Biotechnology Information; ftp://ncbi.nlm.nih.gov/repository/repbase/REF1/humrep.ref) databases, respectively. Protein translations of the masked 5q31 sequences

were compared with the sequences in GenPept release 94 using BLASTX with default parameters and the SEG (Wootton and Federhen 1996) plus XNU (Claverie and States 1993; Claverie 1996) programs to remove low entropy similarities. The HSPcrunch program (Sonnhammer and Durbin 1994) was used to simplify the BLASTX output. The masked 5q31 sequences were compared with the dbEST release 072996 and GenBank release 93 using BLASTN with default parameters. The unmasked 5q31 sequences were analyzed for potential coding regions using XGRAIL (v. 1.3b; Uberbacher et al. 1996).

RT–PCR Analysis

cDNA libraries were purchased from Clontech Laboratories (catalog nos. HL3003a, HL5014a, HL5013a, and HL5007a) and amplified according to the manufacturer's recommendations. PCR amplifications of known and putative genes were performed as follows: 50 ng of cDNA from each amplified library was mixed with 80 μM of each deoxyribonucleoside triphosphate, 20 ng of each oligonucleotide primer, 2 μl of 10× PCR buffer (Boehringer), and 1 unit of *Taq* polymerase (Boehringer) in a 20-μl volume. The samples were amplified using standard PCR reaction conditions in an automated (Perkin Elmer Cetus) thermal cycler for a total of 35 cycles.

RT–PCR analyses were performed using the following primers chosen based on database matches and GRAIL predictions: *LAF-4-like* (120 bp)—forward, 5′-TGCTCTTTGGAAAGCTGTTCAGCTTGG-3′; reverse, 5′-TCTTCAGTGACCATTCCACAGAAGATC-3′; *APXL2* (114 bp)—forward, 5′-CAGGAAAGCTGGATCGTGTGG-3′; reverse, 5′-CCGGATGCTGGAACTCCAAGG-3′; *HSseptin2* (125 bp)—forward, 5′-GAACATTATTCCCATCATGCC-3′; reverse, 5′-CATCATCCGTGGCCAACTG-3′; *cyclin-like* (235 bp)—forward, 5′-AGCTTGGCCACAGGATGGGCATTAAGCCT-3′; reverse, 5′-AAGGATTCCTCTGAGGCAGC-3′; *HSKIF3* (220 bp)—forward, 5′-CCTCAAGTTCTCTGCGAAGTTC-3′; reverse, 5′-CTTGAAACAAAGCTCGACATG-3′; *HSRAD50* (91 bp)—forward, 5′-TTCACAATCTCTGAGCACTGATCG-3′; reverse, 5′-AAGATTTTGTGGAGCTTTTAGGACGT-3′; *RATLSTP-like1* (118 bp)—forward, 5′-CACCACCAGAGTGCCACG-3′; reverse, 5′-TATTTTATGCATTTGGCTACATGG-3′; *HSprolyl4-hydroxylase α(II)* (153 bp)—forward, 5′-GAGCTCCAGGGCACGGTGCAG-3′; reverse, 5′-GAAGGGGACTATTATCATACG-3′; ankyrin repeat motif (125 bp)—forward, 5′-GCTGCAAAATGCAGAGGTGATT-3′; reverse, 5′-GGGTGTACTCCTCTACATTATGCATG-3′; *HSacyl-CoA synthetase* (124 bp)—forward, 5′-GATCCTGAGGATACTGCGACTG-3′; reverse, 5′-GTGAGTGAACCAGTAGGCAAG-3′; *E1* (241 bp)—forward, 5′-AAGTCGAGCTCCTTCAGCAAG-3′; reverse, 5′-TGATTGGGCTGCAGTCTTG-3′; *E2* (185 bp)—forward, 5′-AGGAACGTGGTTTAATTGTGCA-3′; reverse, 5′-GTCTGTTGCTGTATATGGCAGAGCT-3′; *E3* (275 bp)—forward, 5′-AGCCACAGCACTTAGAGCTGAG-3′; reverse, 5′-AAGTGGATCAGTGAGCACAGCT-3′; *E4* (238 bp)—forward, 5′-GGAATAGGTGTCCCTGGGAAC-3′; reverse, 5′-GCAAAGAGACCAAGCATGTCTG-3′; *E5* (150 bp)—forward, 5′-AGATTCACATGAATTAGGAGCTACAC-3′; reverse, 5′-ATGAGATGCTGCTTTGAGCCCTTGG-3′; *G1* (99 bp)—forward, 5′-TTGCACAGGCAAGTAACCAGGCCATC-3′; reverse, 5′-TAGGCTGCGGAATTCTTCTCTGATCAGTA-3′; *G2* (90 bp)—forward, 5′-CATCACCACCACCATCATCATTAATAGCAC-3′; reverse, 5′-AGACCTGGGGCCACCTGCGACTTC-3′; *G3* (73 bp)—forward, 5′-ACCACTCCACACCTCGCAGCT-3′; reverse, 5′-CCCTGGCTTGTTCTTGGTGCC-3′; *G4* (120 bp)—forward, 5′-TGGCACGGCAGTATAAGGC-3′; reverse, 5′-CAGTGCCAGTGTACAGGGCC-3′; *G5* (76 bp)—forward, 5′-GAGAGCTGGGCACCTCCAAGCAG-3′; reverse, 5′-TCCAGAGTACCCCTGCCAGAGATG-3′; *G6* (164 bp)—forward, 5′-GTACCAGAAGCTCCCTGTCAACCC-3′; reverse, 5′-TCAGTCTCTCAGCCACAGAGTCCG-3′; *G7* (161 bp)—forward, 5′-CCATGTCCATGTAAACCTTCGTTATGTG-3′; reverse, 5′-GGATGAGGCTGGGTCATCTGGTGC-3′; *G8* (121 bp)—forward, 5′-AGCTTGGAACACCAGGACAGGGAAC-3′; reverse, 5′-CTGCTCTCCCAGGCTGATGGCCAG-3′; *G10* (118 bp)—forward, 5′-GTGAATCCTTCTGACTGTGCTATAG-3′; reverse, 5′-ACTCAATCACACTCAAGTCAATCTC-3′.

Southern and Northern Hybridization Analyses

Mouse YAC DNA was separated on Bio-Rad CHEF Mapper with 120 sec initial pulse to 20 sec final pulse time at 6 V/cm for 22 hr. DNA was then transferred to a Nytran Plus membrane (Schleicher & Schuell). An adult human multiple tissue total RNA blot was purchased from Clontech Laboratories. Probe DNA fragments of known or putative genes were amplified by PCR and purified from agarose gels on DEAE membranes (Schleicher & Schuell) using standard protocols. Approximately 25 ng of purified DNA was labeled with [α-^{32}P]dCTP using Megaprime DNA labeling systems (Amersham Life Science) according to the manufacturer's protocol. Both Southern and Northern blots were prehybridized in 0.5 M sodium phosphate (pH 7.2), 1 mM EDTA, and 7% SDS for 1 hr at 65°C. Hybridization was then carried out in the same solution with addition of 100 μg/ml of sheared salmon sperm DNA and radiolabeled probe, and incubated overnight at 65°C. The filters were washed once with 2× SSC–0.1% SDS, twice with 0.5× SSC–0.1% SDS, and if necessary once with 0.1× SSC–0.1% SDS at 68°C, and then exposed to X-ray films with an intensifying screen at −80°C.

Hybridization probes were generated using the following primers: *LAF-4-like* (323 bp)—forward, 5′-AGCTACACTGATACAAGTGGACCTAA-3′; reverse, 5′-GGGGAAGACTTAGACTCCTTCTTT-3′; *APXL2* (253 bp)—forward, 5′-TCCCTGCTGCAGCGACTCCGGCTCC-3′; reverse, 5′-AAGGTCGTCCCTGATGGCGTCCAGTTG-3′; *HSseptin2* (202 bp)—forward, 5′-AGACCGGCATTGGCAAAT -3′; reverse, 5′-GCCTCTTGCCTCTCATCCTT-3′; *cyclin-like* (235 bp)—forward, 5′-AGCTTGGCCACAGGATGGGCATTAAGCCT-3′; reverse, 5′-AAGGATTCCTCTGAGGCAGC-3′; *HSKIF3* (262 bp)—forward, 5′-CAGAGAAGCCAGAAAGCTGC-3′; reverse, 5′-TTGTAGCCTTCAAGTACAGAATCAAT-3′; *HSRAD50* (225 bp)—forward, 5′-CACTTTCTGAGGACCTACATTTCTATG-3′; reverse, 5′-AGTCGCTCACAGCAGCGTA-3′; *RATLSTP-like2* (172 bp)—forward, 5′-GAACAGAAATTCTTGGCAAGTCAGTT-3′; reverse, 5′-ACCACAGCGGGACACACAG-3′; *HSprolyl4-hydroxylase α(II)* (199 bp)—forward, 5′-TCTCCTTACCAAGGGAGAGCA-3′; reverse, 5′-CCGCTCGGCTACAATGAAG-3′; *HSacyl-CoA synthetase* (262 bp)—forward, 5′-CTTGTTTCAACAGAGTTTGTCCTC-3′; reverse, 5′-CCTGCTCACCTCTACTTCTTCTGA-3′; *E1* (241 bp)—forward, 5′-AAGTCGAGCTCCTTCAGCAAG-3′; reverse, 5′-TGATTGGGCTGCAGTCTTG-3′; *E3* (275 bp)—forward, 5′-AGCCACAGCACTTAGAGCTGAG-3′; reverse, 5′-AAGTGGATCAGTGAGCACAGCT-3′.

Production of Transgenic Mice

Because YAC (854G6) was too large to isolate intact, it was truncated using the acentric YAC deletion vector pBCL (Lewis

et al. 1992). The derivative clones were phenotyped as Lys^-, Trp^-, and Ura^+, and the sizes of the fragmented YACs were determined by pulse-field gel electrophoresis. In this study derivative clone 1 (350 kb) was used. YAC DNA (854G6 no. 1, A94G6, and 131F9) was isolated as described previously (Frazer et al. 1995), with the following modification: The DNA was dialyzed overnight on a 0.05-mm dialysis filter (Millipore) against injection buffer [10 mM Tris-HCl (pH 7.5), 0.1 mM EDTA, 100 mM NaCl]. The isolated DNA, at a final concentration of ~1 ng/ml, was microinjected into fertilized FVB hybrid mouse eggs using standard procedures.

RNase Protection Assays

Total RNA was extracted from various tissues of 3- to 4-week-old mice using RNA STAT-60 (TelTestB) according to the manufacturer's instructions. Radiolabeled antisense riboprobes were generated using the MAXIscript kit (Ambion, Inc., Austin, TX) from DNA templates generated by PCR using the following probes: *HSseptin2* (125 bp)—forward, 5′-GAACATTATTCCCATCATGCC-3′; reverse, 5′-CATCATCCGTGGCCAACTG-3′; *HSKIF3* (195 bp)—forward, 5′-TTTTTCTTCAATATCCAAGCGTT-3′; reverse, 5′-AGATCTATTTTGGTTATTTTATTTCCG-3′; *HSRAD50* probe (140 bp)—forward, 5′-TTTACCTAACAGTGAACCTGTGAC-GTT-3′; reverse, 5′-CCAGAGCATGTGCAAGAGATACTTAC-3′. RNase protection assays were performed using the RPA II kit (Ambion, Inc., Austin, TX) according to the manufacturer's recommendations. The protected fragments were separated on a 7.5% acrylamide denaturing gel, and the gel was exposed to X-ray film.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants to E.M.R. (PPG HL18574); Human Genome Distinguished Postdoctoral Fellowship (K.A.F.) sponsored by the U.S. Department of Energy, Office of Health and Environmental Research, and administered by the Oak Ridge Institute for Science and Education. E.M.R. is an American Heart Association Established Investigator. Research was conducted at the Lawrence Berkeley National Laboratory (Department of Energy contract DE-AC0376SF00098), University of California, Berkeley.

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

REFERENCES

Adams, M.E., A.R. Kerlavage, R.D. Fleischmann, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Weinstock, J.D. Gocayne, O. White, G. Sutton, J.A. Blake, R.C. Brandon, M.W. Chiu, R.A. Clayton, R.T. Cline, M.D. Cotton, J. Earle-Hughes, L.D. Fine, L.M. Fitzgerald, W.M. Fitzhugh, J.L. Fritchman, N.S.M. Geoghagen, A. Glodek, C.L. Gnehm, M.C. Hanna, E. Hedblom, P.S. Hinkle, Jr., J.M. Kelley, K.M. Klimek, J.C. Kelley, L.I. Liu, S.M. Marmaros, J.M. Merrick, R.F. Moreno-Palanques, L.A. McDonald, D.T. Nguyen, S.M. Pellegrino, C.A. Phillips, S.E. Ryder, J.L. Scott, D.M. Saudek, R. Shirley, K.V. Small, T.A. Spriggs, T.R. Utterback, J.F. Weidman, Y. Li, R. Barthlow, D.P. Bednarik, L. Cao, M.A. Cepeda, T.A. Coleman, E.J. Collins, D. Dimke, P. Feng, A. Ferrie, C. Fischer, G.A. Hastings, W.W. He, J.S. Hu, K.A. Huddleston, J.M. Greene, J. Gruber, P. Hudson, A. Kim, D.L. Kozak, C. Kunsch, H. Ji, H. Li, P.S. Meissner, H. Olsen, L. Raymond, Y.F. Wei, J. Wing, C. Xu, G.L. Yu, S.M. Ruben, P.J. Dillon, M.R. Fannon, C.A. Rosen, W.A. Haseltine, C. Fields, C.M. Fraser, and J.C. Venter. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 377: 3–174.

Aizawa, H., Y. Sekine, R. Takemura, Z. Zhang, M. Nangaku, and N. Hirokawa. 1992. Kinesin family in murine central nervous system. *J. Cell Biol.* 119: 1287–1296.

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

Bleecker, E.R., P.J. Amelung, R.C. Levitt, D.S. Postma, and D.A. Meyers. 1995. Evidence for linkage of total serum IgE and bronchial hyperresponsiveness to chromosome 5q: A major regulatory locus important in asthma. *Clin. Exp. Allergy* 25: (Suppl. 2) 84–88; (discussion) 95–96.

Bonaldo, M.D.F., G. Lennon, and M.B. Soares. 1996. Normalization and subtraction—Two approaches to facilitate gene discovery. *Genome Res.* 6: 791–806.

Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. Fitzgerald, R.A. Clayton, J.D. Gocayne, A.R. Kerlavage, B.A. Dougherty, J.F. Tomb, M.D. Adams, C.I. Reich, R. Overbeek, E.F. Kirkness, K.G. Weinstock, J.M. Merrick, A. Glodek, J.L. Scott, N.S.M. Geoghagen, J.F. Weidman, J.L. Fuhrmann, J.C. Venter, et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273: 1058–1073.

Claverie, J.M. 1996. Effective large-scale sequence similarity searches. *Methods Enzymol.* 266: 212–227.

Claverie, J.M. and D.J. States. 1993. Information enhancement methods for large scale sequence analysis (proteins). *Comput. Chem.* 17: 191–201.

DeBry, R.W. and M.F. Seldin. 1996. Human/mouse homology relationships. *Genomics* 33: 337–351.

Domer, P.H., S.S. Fakharzadeh, C.S. Chen, J. Jockel, L. Johansen, G.A. Silverman, J.H. Kersey, and S.J. Korsmeyer. 1993. Acute mixed-lineage leukemia t(4;11)(q21;q23) generates an *MLL-AF4* fusion product. *Proc. Natl. Acad. Sci.* 90: 7884–7888.

Dujon, B. 1996. The yeast genome project–What did we learn? *Trends Genet.* 12: 263–270.

Fares, H., M. Peifer, and J.R. Pringle. 1995. Localization and possible functions of Drosophila septins. *Mol. Biol. Cell* 6: 1843–1859.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton,

E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science* 269: 496–512.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley, et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397–403.

Frazer, K.A., G. Narla, J.L. Zhang, and E.M. Rubin. 1995. The apolipoprotein(*a*) gene is regulated by sex hormones and acute-phase inducers in YAC transgenic mice. *Nature Genet.* 9: 424–431.

Fujino, T., and T. Yamamoto. 1992. Cloning and functional expression of a novel long-chain acyl-CoA synthetase expressed in brain. *J. Biochem. (Tokyo)* 111: 197–203.

Hadwiger, J.A., C. Wittenberg, H.E. Richardson, M. de Barros Lopes, and S.I. Reed. 1989. A family of cyclin homologs that control the G1 phase in yeast. *Proc. Natl. Acad. Sci.* 86: 6255–6259.

Helaakoski, T., P. Annunen, K. Vuori, I.A. MacNeil, T. Pihlajaniemi, and K.I. Kivirikko. 1995. Cloning, baculovirus expression, and characterization of a second mouse prolyl 4-hydroxylase alpha-subunit isoform: Formation of an α_2 β_2 tetramer with the protein disulfide-isomerase/beta subunit. *Proc. Natl. Acad. Sci.* 92: 4427–4431.

Hillier, L., G. Lennon, M. Becker, M.F. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. Dubuque, A. Favello, W. Gish, M. Hawkins, M. Hultman, T. Kucaba, M. Lacy, M. Le, N. Le, E. Mardis, B. Moore, M. Morris, J. Parsons, C. Prange, L. Rifkin, T. Rohlfing, K. Schellenberg, M.B. Soares, F. Tan, J. Thierrymeg, E. Trevaskis, K. Underwood, P. Wohldman, R. Waterston, R. Wilson, and M. Marra. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* 6: 807–828.

Johzuka, K. and H. Ogawa. 1995. Interaction of Mre11 and Rad50: Two proteins required for DNA repair and meiosis-specific double-strand break formation in *Saccharomyces cerevisiae*. *Genetics* 139: 1521–1532.

Kiess, M., B. Scharm, A. Aguzzi, A. Hajnal, R. Klemenz, I. Schwarte-Waldhoff, and R. Schafer. 1995. Expression of ril, a novel LIM domain gene, is down-regulated in Hras-transformed cells and restored in phenotypic revertants. *Oncogene* 10: 61–68.

Lander, E.S. 1996. The new genomics—Global views of biology. *Science* 274: 536–539.

Lewis, B.C., N.P. Shah, B.S. Braun, and C.T. Denny. 1992. Creation of a yeast artificial chromosome fragmentation vector based on lysine-2. *Genet. Anal. Tech. Appl.* 9: 86–90.

Ma, C. and L.M. Staudt. 1996. *LAF-4* encodes a lymphoid nuclear protein with transactivation potential that is homologous to *AF-4*, the gene fused to *MLL* in t(4;11) leukemias. *Blood* 87: 734–745.

Makalowski, W., J.H. Zhang, and M.S. Boguski. 1996. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* 6: 846–857.

Marsh, D.G., J.D. Neely, D.R. Breazeale, B. Ghosh, L.R. Freidhoff, E. Ehrlich-Kautzky, C. Schou, G. Krishnaswamy, and T.H. Beaty. 1994. Linkage analysis of *IL4* and other chromosome 5q31.1 markers and total serum immunoglobulin E concentrations. *Science* 264: 1152–1156.

Marshall, E. and E. Pennisi. 1996. NIH launches the final push to sequence the genome. *Science* 272: 188–189.

Martin, C.H., C.A. Mayeda, C.A. Davis, C.L. Ericsson, J.D. Knafels, D.R. Mathog, S.E. Celniker, E.B. Lewis, and M.J. Palazzolo. 1995. Complete sequence of the bithorax complex of Drosophila. *Proc. Natl. Acad. Sci.* 92: 8398–8402.

Morrissey, J., D.C. Tkachuk, A. Milatovich, U. Francke, M. Link, and M.L. Cleary. 1993. A serine/proline-rich protein is fused to HRX in t(4;11) acute leukemias. *Blood* 81: 1124–1131.

Nakamura, T., H. Alder, Y. Gu, R. Prasad, O. Canaani, N. Kamada, R.P. Gale, B. Lange, W.M. Crist, P.C. Nowell, et al. 1993. Genes on chromosomes 4, 9, and 19 involved in 11q23 abnormalities in acute leukemia share sequence homology and/or common motifs. *Proc. Natl. Acad. Sci.* 90: 4631–4635.

Nimer, S.D. and H. Uchida. 1995. Regulation of *granulocyte-macrophage colony-stimulating factor* and *interleukin 3* expression. *Stem Cells* 13: 324–335.

Peters, L.L. and S.E. Lux. 1993. Ankyrins: Structure and function in normal cells and hereditary spherocytes. *Semin. Hematol.* 30: 85–118.

Postma, D.S., E.R. Bleecker, P.J. Amelung, K.J. Holroyd, J. Xu, C.I. Panhuysen, D.A. Meyers, and R.C. Levitt. 1995. Genetic susceptibility to asthma–bronchial hyperresponsiveness coinherited with a major gene for atopy. *N. Engl. J. Med.* 333: 894–900.

Saltman, D.L., G.M. Dolganov, J.A. Warrington, J.J. Wasmuth, and M. Lovett. 1993. A physical map of 15 loci on human chromosome 5q23-q33 by two-color fluorescence in situ hybridization. *Genomics* 16: 726–732.

Schiaffino, M.V., M.T. Bassi, E.I. Rugarli, A. Renieri, L. Galli, and A. Ballabio. 1995. Cloning of a human homologue of the *Xenopus laevis* APX gene from the ocular albinism type 1 critical region. *Hum. Mol. Genet.* 4: 373–382.

Simonson, G.D., A.C. Vincent, K.J. Roberg, Y. Huang, and V. Iwanij. 1994. Molecular cloning and characterization of a novel liver-specific transport protein. *J. Cell Sci.* 107: 1065–1072.

Smirnov, D.V., M.G. Smirnova, V.G. Korobko, and E.I. Frolova. 1995. Tandem arrangement of human genes for *interleukin-4* and *interleukin-13:* Resemblance in their organization. *Gene* 155: 277–281.

Smith, D.J., Y. Zhu, J. Zhang, J.F. Cheng, and E.M. Rubin. 1995. Construction of a panel of transgenic mice containing a contiguous 2-Mb set of YAC/P1 clones from human chromosome 21q22.2. *Genomics* 27: 425–434.

Sonnhammer, E.L. and R. Durbin. 1994. An expert system for processing sequence homology data. *Ismb* 2: 363–368.

Staub, O., F. Verrey, T.R. Kleyman, D.J. Benos, B.C. Rossier, and J.P. Kraehenbuhl. 1992. Primary structure of an apical protein from *Xenopus laevis* that participates in amiloride-sensitive sodium channel activity. *J. Cell Biol.* 119: 1497–1506.

Uberbacher, E.C., Y. Xu, and R.J. Mural. 1996. Discovering and understanding genes in human DNA sequence using GRAIL. *Methods Enzymol.* 266: 259–281.

Walsh, S. and B. Barrell. 1996. The *Saccharomyces cerevisiae* genome on the World Wide Web. *Trends Genet.* 12: 276–277.

Wootton, J.C. and S. Federhen. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* 266: 554–571.

Received December 27, 1996; accepted in revised form March 5, 1997.