Genomics **60,** 295–308 (1999)
Article ID geno.1999.5927, available online at http://www.idealibrary.com on IDEAL®

Genome Duplications and Other Features in 12 Mb of DNA Sequence from Human Chromosome 16p and 16q

Brendan J. Loftus,* Ung-Jin Kim,† Victoria P. Sneddon,‡ Francis Kalush,*,[1] Rhonda Brandon,*,[1] Joyce Fuhrmann,* Tanya Mason,* Marie L. Crosby,* Mary Barnstead,* Lisa Cronin,* Anne Deslattes Mays,*,[1] Yicheng Cao,† Robert X. Xu,† Hyung-Lyun Kang,† Steve Mitchell,† Evan E. Eichler,§ Peter C. Harris,‡ J. Craig Venter,*,[1] and Mark D. Adams*,[2]

*The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850; †Division of Biology, The California Institute of Technology, Pasadena, California 91125; ‡MRC Molecular Haematology Unit, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, United Kingdom; and §Department of Genetics, Case Western Reserve University, Cleveland, Ohio 44106

Received March 15, 1999; accepted July 13, 1999

Several publicly funded large-scale sequencing efforts have been initiated with the goal of completing the first reference human genome sequence by the year 2005. Here we present the results of analysis of 11.8 Mb of genomic sequence from chromosome 16. The apparent gene density varies throughout the region, but the number of genes predicted (84) suggests that this is a gene-poor region. This result may also suggest that the total number of human genes is likely to be at the lower end of published estimates. One of the most interesting aspects of this region of the genome is the presence of highly homologous, recently duplicated tracts of sequence distributed throughout the p-arm. Such duplications have implications for mapping and gene analysis as well as the predisposition to recurrent chromosomal structural rearrangements associated with genetic disease.

INTRODUCTION

Virtually all of the research in human genomics to date has been focused on genes, and most progress has been made from sequencing of cDNAs. Genes are the core functionality of the genome but there are clearly other intriguing facets in the structure and organization of DNA on human chromosomes. The subject of this study, which includes sequence from both the p-arm of chromosome 16 and portions of the q-arm, has been mapped in some depth. A nearly complete yeast artificial chromosome (YAC) map as well as partial cosmid coverage and cosmid contigs anchored to the YAC map by sequence-tagged sites (STSs) spaced at an average of 150 kb apart has been described (Doggett *et al.*, 1995). We took advantage of this STS data as the starting point for building a sequence-ready map of the 20-Mb region adjacent to the centromere on 16p. Chromosome 16 is of interest as the location of several disease genes including those for polycystic kidney disease (*PKD1*) (European Polycystic Kidney Disease Consortium, 1994) and Batten disease (*CLN3*) (Mitchison *et al.*, 1993), in addition to the fragile loci FRA16B and FRA16A (Nancarrow *et al.*, 1994; Yu *et al.*, 1997). A number of leukemia subtypes have been associated with chromosome 16 abnormalities in this region (Liu *et al.*, 1996; Dissing *et al.*, 1998), and a series of low-abundance chromosome 16-specific repeats (CH16LARS) has been identified and postulated to be responsible for disease-causing genome rearrangements (Stallings *et al.*, 1992).

While the 20-Mb region of the p-arm targeted in this study has not been completely sequenced, an initial analysis of the results from almost half of this region has shed light on some intriguing features of this segment of the genome. Here we present a first view of 11.8 Mb of sequence data from both 16p and 16q that provides a glimpse of the complexities of the genome and allows an evaluation of a number of the current gene and exon prediction methods and the usefulness of expressed sequence tag (EST; Adams *et al.*, 1991, 1995; Hillier *et al.*, 1996) databases in gene structure determination. Approximately 6500 human genes have

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession Nos. AC002038–AC002042, AC002044, AC002045, AC002286–AC002289, AC002299–AC002303, AC002307, AC002310, AC002331, AC002394, AC002400, AC002425, AC002492, AC002506, AC002519, AC002544, AC002550, AC002551, AC002565, AC002990, AC003003, AC003007, AC003009, AC003010, AC003026, AC003034, AC003108, AC003119, AC003661, AC003964, AC003977, AC004020, AC004097, AC004123, AC004125, AC004131, AC004158, AC004381, AC004382, AC004513, AC004514, AC004525, AC004529, AC004531, AC004533, AC004605, AC004626, AC004682, AC004685, AC004787, AC005136, AC005638, AF001548–AF001552, U62317, U91318–U91327, U95737–U95743, U96629.

[1] Present address: Celera Genomics, 45 W. Gude Drive, Rockville, MD 20850.

[2] To whom correspondence should be addressed at present address: Celera Genomics, 45 W. Gude Drive, Rockville, MD 20850. Telephone: (240) 453-3700. Fax: (240) 453-4000. E-mail: AdamsMD@celera.com.

0888-7543/99 $30.00

been completely sequenced as cDNA clones (Adams *et al.*, 1995), representing only about 10% of the expected number of human genes (Fields *et al.*, 1994). Up to 75% of the human genes may be tagged by ESTs (Adams *et al.*, 1995; Hillier *et al.*, 1996), more than 1,000,000 of which now exist in public databases. However, the EST databases are biased for the 3′ untranslated portions of genes as well as for shorter and more abundant transcripts, and thus the coverage of transcribed sequence is far from complete. The accurate prediction of gene structure from raw genomic sequence data has been the focus of considerable development effort in computational biology for many years. An increasingly diverse set of very large EST databases and complete genome sequences of both eukaryotes (Goffeau *et al.*, 1996) and prokaryotes (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995; Bult *et al.*, 1996; Blattner *et al.*, 1997) has resulted in a substantial reevaluation of predictive methods for determining gene structure, with greater emphasis placed on incorporation of database matches than on *de novo* predictions (Huang *et al.*, 1997; Xu *et al.*, 1997). Gene indices (http://www.tigr.org/tdb/tdb.html) have been built for several species and contain assemblies of ESTs that are longer, more informative, and generally of higher quality than individual EST sequences. These assembled sequences facilitate the determination of gene structure as well as the identification of sequence similarity, alternative splice forms, polymorphic variations, and tissue expression. For these methodologies to be effective, the gene to be annotated must have matches in the databases being searched; it is therefore important to assess the completeness of the databases.

In this report we present an analysis of 11.8 Mb of sequence data from either side of the centromere of one of the best characterized autosomes (chromosome 16). One of the most intriguing observations is the presence of several copies of large (greater than 20 kb) duplicated regions (duplicons). The term “duplicon” is used to refer to a duplicated genomic segment of considerable size to distinguish these from common retroposed repetitive elements that are dispersed throughout the entire human genome. Some of these duplications include portions of the polycystic kidney disease gene (*PKD1*) (European Polycystic Kidney Disease Consortium, 1994) and the entire coding region of the translation initiation factor *eIF3-p110* gene (Asano *et al.*, 1997). The number and size of these duplicons and the presence of apparently transcribed sequences within them raise questions about the mechanisms by which the duplications arose, how they are maintained, and the role that they may play in chromosome and genome evolution.

MATERIALS AND METHODS

Map Construction

Two different fingerprinting methods were used to create contig maps and confirm the sizes of the bacterial artificial chromosomes (BACs). At Caltech, BAC DNA samples were end-radiolabeled, cleaved with two restriction endonucleases, and then separated on polyacrylamide gels. The data were analyzed using the Sanger Center's IMAGE 2.1 software (Marra *et al.*, 1997). The band size output of this program was used to create contig maps with FPC (Marra *et al.*, 1997). At The Institute for Genomic Research (TIGR), a different fingerprinting method was used to size BACs chosen for sequencing and provide data for assembly verification. In this method, BAC DNA samples were cleaved with *Hin*dIII or other restriction endonucleases, and the fragments were separated on agarose gels. After staining with fluorescent dye was performed, a fluorescence scanner was used to produce an image of the gel. The band sizes were read by IMAGE 3.6 (an updated version of IMAGE 2.1). BAC library screening was performed at Caltech and TIGR. Probes designed from STSs and BAC end sequences were used to select positive clones from BAC libraries created by Pieter De Jong's laboratory at the Roswell Park Cancer Institute and at The California Institute of Technology that were reported to have 20X coverage of the human genome (P. de Jong and H. Shizuya, pers. comm.). Minimally overlapping clones were sequenced to extend the length of the contigs and close the gaps between contigs. The resulting BAC map is shown in Fig. 1.

Sequencing

BAC clones selected for sequencing from the map were confirmed by re-end-sequencing and fingerprinting. Large-scale cultures were then subcloned into pUC18, essentially as described (Fleischmann *et al.*, 1995). Approximately 1000 pUC18 subclones were purified per 100 kb of BAC DNA, and each subclone was sequenced from each end using ABI dichlororhodamine or BigDye terminator kits. Base-calling was performed using phred (Ewing and Green, 1998). Sequence assembly was performed using TIGR Assembler (Sutton *et al.*, 1995). A suite of software tools was used to assess the quality and completeness of the sequence assemblies at the end of the random phase. Directed sequencing was performed as suggested by these software tools and as determined by manual inspection of contigs to close any remaining gaps (generally fewer than a dozen per 100 kb of BAC DNA) and improve the quality of weak regions. Each base was covered either on both strands or by two independent sequencing chemistries (e.g., dye-labeled terminators and dye-labeled primers). Independent assessment of sequence accuracy demonstrated that fewer than one error in 50,000 bases remained in finished BAC sequences.

Sequence Analysis

DNA and protein searches. Searches of the genomic sequence were performed on a number of DNA databases including GenBank (release 107.0) (Benson *et al.*, 1998) from which ESTs had been removed and databases of transcribed sequences from both the human (HGI) and the mouse (MGI) gene indices collated at TIGR (http://www.tigr.org). A nonredundant database of protein sequences (nr) from NCBI (http://www.ncbi.nlm.nih.gov/) was used for the DNA vs protein searches. BAC DNA sequences were masked for repeats by screening against a library of repetitive elements using Repeatmasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker). The analysis and annotation tool (AAT) (Huang *et al.*, 1997) was used to perform database searches and generate alignments. This tool consists of two sets of programs, one for comparison of the query sequence with a protein database and the other for comparing the sequence with a cDNA database. The database search program quickly identifies regions of similarity between the query and the database sequences, and the alignment program constructs an optimal alignment for each region of the query sequence and the database sequence.

Gene structure annotation. A combination of three publicly available gene and exon prediction programs was used: Grail 2 version 1.3 (Xu *et al.*, 1994), Genefinder (Phil Green, University of Washington), and Genscan (Burge and Karlin, 1997). In the determination of gene structure, a gene model that best represented the results from the database matches and gene structure predictions was constructed. Each gene model was constrained by the requirement of a spliced database match spanning at least one end of each exon annotated

within a gene model. This resulted in the construction of gene structures contingent on the presence of database matches and did not always represent an entire gene structure due to the lack of EST coverage across some genes. This method of annotation underpredicts the number and size of genes in the absence of a full-length database match, but was deemed necessary due to the deficiencies in *de novo* gene structure prediction from genomic sequence.

RESULTS

A map of the region illustrating the location of sequenced BAC clones and contigs is depicted in Fig. 1. The overall GC content of the sequence is 43.97%. The total percentage of repetitive elements calculated using Repeatmasker was estimated at 42%, which is higher than the 34% predicted for DNA in the GC content range of 43–52% (Smit, 1996). A previous study has also indicated a higher frequency of repetitive DNA sequences in chromosome 16 (Okumura *et al.*, 1994). Short interspersed nuclear elements (SINEs) accounted for 21.5% of the genomic sequence, long interspersed nuclear elements (LINEs) account for 12.3%, and LTR elements account for 5.5%. This correlates to an increased percentage of all repeated elements compared to predicted frequencies for sequences in this GC range with a small increase in the percentages of SINEs and larger increases in expected values of LINEs and LTRs (Smit, 1996). A graph plotting the distribution of SINEs and LINEs across a number of BAC clones mapped along the p-arm illustrates the near-complementarity that appears to exist between the two classes of repeat throughout the human genome (Fig. 2). The graph also indicates a detectable inverse relationship between the frequency of LINE elements and the percentage of bases annotated as genes and the expected positive correlation between bases coding and GC content (Fig. 2).

Gene Structure Annotation (ESTs)

It was possible to incorporate 73% of the 6785 EST database matches into annotated gene models using a search strategy designed to maximize the number of informative database matches while eliminating matches due to the presence of repetitive elements. The remaining ESTs were not incorporated into gene models because they were determined to be uninformative matches on manual inspection or because they did not contain sequence overlap or clone links with an annotated gene. These additional sequences may represent unprocessed message or genomic sequence contaminants of the EST libraries or they may indicate the presence of another gene, the presence of an extended 3′ or 5′ untranslated region (UTR) of an annotated gene, or the presence of alternative splice forms. The percentage of matching ESTs (73%) that were used in the prediction of gene structure is in line with figures previously published for data from high-throughput genomic sequence (Bailey *et al.*, 1998). All but five of the annotated genes (94%) had at least one database match to a human expressed sequence; 75% of the annotated genes had database matches to expressed mouse sequences, and this figure increased to 81% when genes arising from duplications were omitted. Of the previously identified genes, 88% had a significant database match to an expressed mouse sequence. Sixty-two percent of unknown genes had mouse EST matches, a figure that increased to 74% omitting genes in duplicons.

De Novo Gene Structure Prediction

Genscan, Grail2 (Version 1.3), and Genefinder were used for *ab initio* gene-finding. Disparities were observed in the results of gene prediction programs, including different use of splicing sites or exons between different gene models and assignment of gene models to both DNA strands at the same position. The use of more than one gene prediction program is likely to result in increased accuracy as exons or splicing sites that are missed by one program could be predicted correctly by another.

Genes Identified during Annotation

Of the 84 gene assignments made, 44 genes were assigned function on the basis of similarity to previously known human genes or genes from another species, and 40 were designated as “unknown gene products”; 20 more DNA segments were annotated as pseudogenes (Table 1). Alternative splice forms were identified for 13 of the annotated genes on the basis of matches to EST sequences. UTRs were also annotated on the basis of EST database matches. Untranslated regions downstream of the coding region were annotated for 52 of the 84 annotated genes, upstream 5′ UTRs were found for 40 of the genes with 25 genes annotated with both 5′ and 3′ UTRs. All previously identified genes except *KIAA0220*, *SCNN1B*, a T-cell directed chemokine, the interleukin 4 receptor, and the antisense transcript of the B-cell maturation factor gene were found to have an existing mouse homologue or a significant database match to mouse expressed sequences.

A number of the annotated, previously identified genes are thought to contribute to known disease phenotypes, but in many cases knowledge of their immediate genomic environment was previously unavailable as they have been sequenced from cDNA clones. There is a cluster of three genes that are part of the superfamily of ABC transporters located in BAC clone A-962B4. The gene for multiple drug resistance (*MRP1*) lies in opposite transcriptional orientation to two putative multidrug resistance genes; the anthracycline resistance-associated (*ara*) gene and a previously undescribed *MRP1*-like gene. The *MRP1* and *ara* genes have previously been found to be co-overexpressed in some leukemia cell lines as part of the multifactorial drug resistance response (Longhurst *et al.*, 1996). The 3′ end of the *PM5* gene, which shares homology with conserved regions of the collagenase gene family (Templeton *et al.*, 1992), is also found in

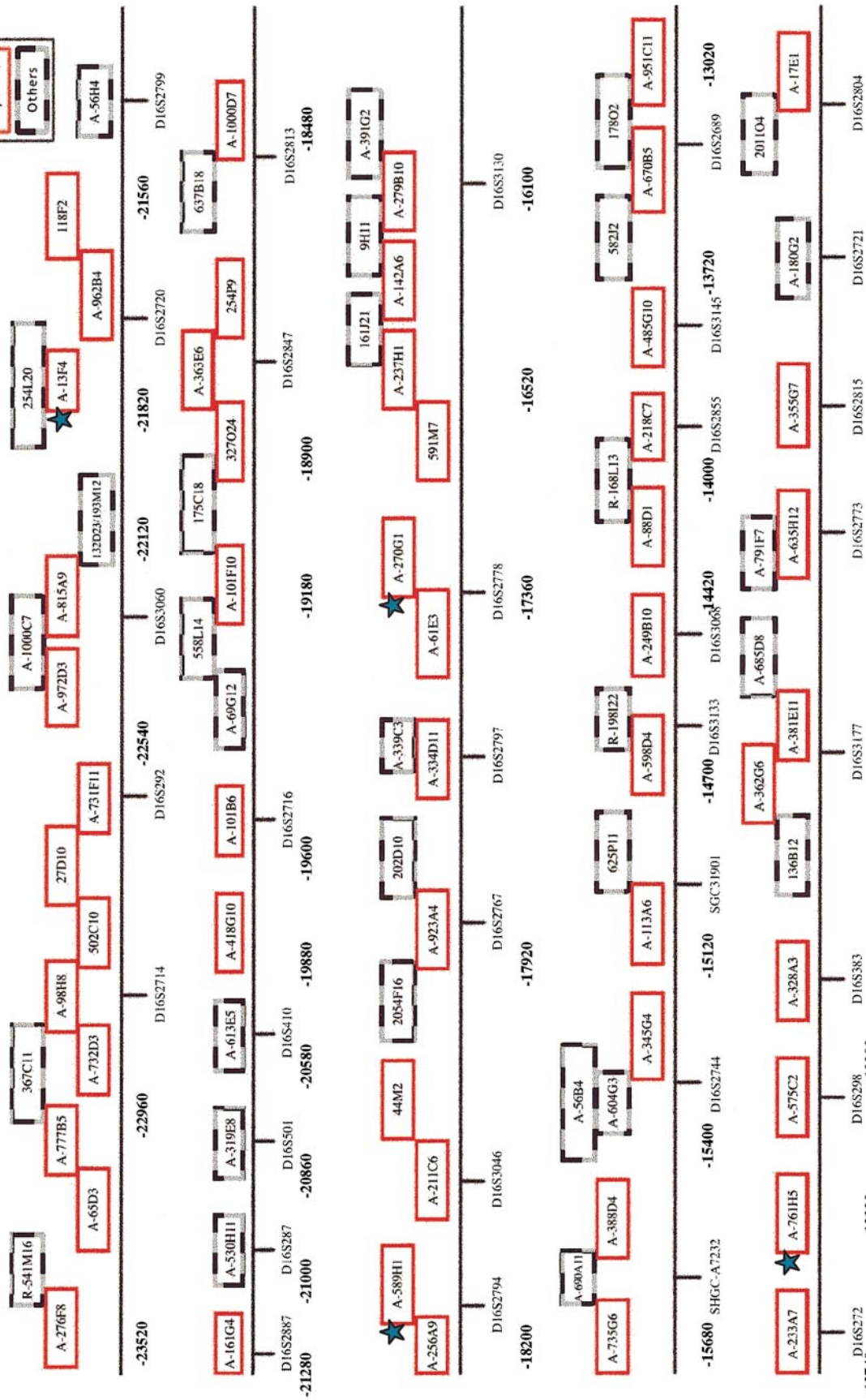

FIG. 1. Map of chromosome 16p13.1–p11.2 illustrating the location of sequenced BAC clones and contigs. The assembled physical map was based on STS content, BAC end sequences, and restriction fingerprints. Coordinates correspond to approximate distances in kilobase-pairs, measured from the centromere, as described in Cao *et al.* (1999). Clones marked with a star contain copies of Duplicon A.

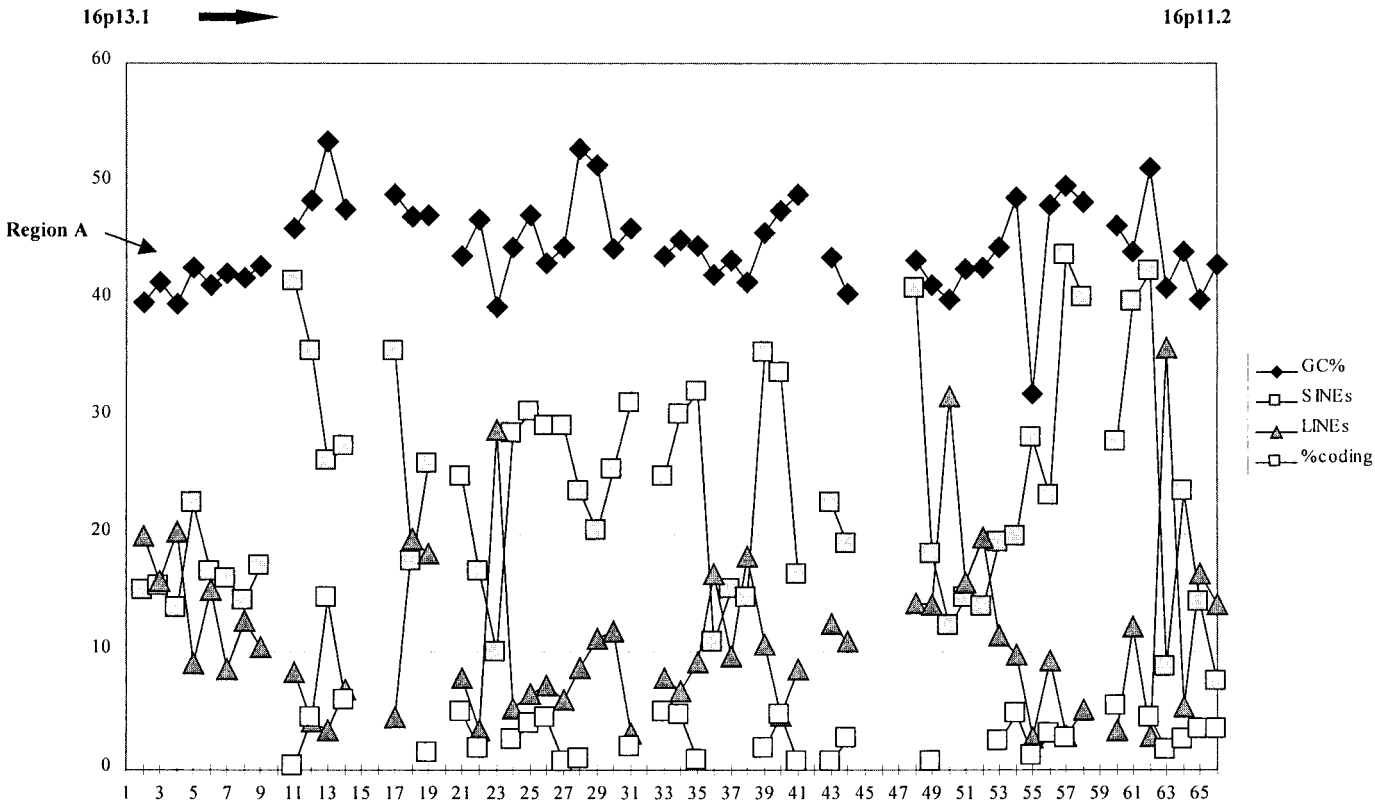

FIG. 2. General features of the completely sequenced BAC clones from the p-arm of chromosome 16 in order as they appear on the map (Fig. 1). Each point on the map represents an individually sequenced BAC clone. Points joined are for the purposes of clarity and do not indicate the presence of a sequence contig.

A-962B4. This region has been proposed to contain candidate genes for the pseudoxanthoma elasticum disorder (PXE) (Struk *et al.*, 1997). Both *MRP1* and *PM5* have been proposed as candidate genes for PXE (Van Soest *et al.*, 1997). The annotation of *ara* and an *MRP1*-like gene positioned between *MRP1* and *PM5* suggests that these might also be considered as potential candidate genes for PXE. It is possible to construct a single predicted transcript with an open reading frame spanning the length of apparent coding regions *ara* and the *MRP1*-like gene. Because the *ara* gene was isolated from a leukemia cell line, there exists the possibility that *ara* and the *MRP1*-like annotated gene actually form a single transcript.

Two kidney-expressed genes associated with hypertension were identified during annotation but were found to lie in different regions of chromosome 16: the beta subunit of the amiloride-sensitive sodium channel gene (*SCNN1B*) (Shimkets *et al.*, 1994), located on 16p12, and the *SA* gene (16p13.11) (Samani *et al.*, 1994). Two genes with similarity to the rat kidney-specific (*ks*) gene were found in the same region of the chromosome as the amiloride-sensitive sodium channel beta subunit gene. One was annotated as a homologue of the rat kidney-specific (*ks*) gene, sharing 92% similarity at the protein level with the other annotated as similar to rat *ks*, having a lower degree of similarity (82%). Both genes share a tissue distribution similar to that of the rat *ks* gene, being found only in kidney and liver ESTs, indicating a possible functional similarity. Both of these genes have a significant similarity to the proximal end of the human *SA* gene, sharing 81% (rat *ks* homologue) and 79% similarity with the human *SA* protein, respectively. This may indicate that these genes share some functionality with the human *SA* gene and may play a role in hypertension.

Five genes were assigned as homologues or having significant similarity to previously known rat genes, including the two *ks*-like genes described above. Four appeared to share a restricted tissue distribution similar to that proposed for the rat genes on the basis of EST database matches. A homologue of the rat *p8* gene with two alternative splice forms was annotated with greater than 83% similarity to the rat *p8* protein, which has been proposed as a putative transcription factor regulating pancreatic growth (Mallo *et al.*, 1997). The rat *p8* gene is activated in the pancreas; however, the human gene demonstrated a broader tissue expression profile. A homologue of the rat brain/kidney (*B/K*) gene with 99.2% similarity to the rat protein was found to display a limited distribution in terms of EST coverage being found only in Wilms tumor, uterus tumor, and brain tissue ESTs. The function of the rat *B/K* gene is unknown but it is thought to be involved in cyclic AMP-dependent processes due to the presence of sequences for phophorylation by cyclic AMP-dependent kinase (Kwon *et al.*, 1996). A homologue of the pancreas-specific rat zymogen granule membrane protein (*RHZG16P*), sharing 92% similarity with the rat gene at the protein level, was annotated with a restricted

TABLE 1

Genes Identified during the Annotation Process

Name assigned	Mouse hits	% Similarity	Splice	Human EST No.	Tissues
Myosin heavy chain (MYH11)	Yes	99.1		145	M
Multiple drug resistance (MRP1)	Yes	100	Splice	64	E, R, J
Anthracycline resistance-associated (ara)	Yes	98.8		5	FL, FS, FLU
PM5	Yes	99.4		188	M
Translation initiation factor (eIF3, p110)	Yes	99.8		250	M
Zona pellucida (ZP2)	Yes	100		0	
Mu-crystallin (Mu)	Yes	100		24	CE, T, P, IB
MRP1-like gene	Yes	66.5		25	M
B-cell maturation protein (BCMA)	Yes	99.6		11	BC, TC, BR
G1 to S phase transition (GSPT1)	Yes	99.8	Splice	95	M
KIAA0420	Yes	99.9		27	BR, O, T
Rat B/K homologue	Yes	99.2		14	B, WT, PG, UT, TC
Transcription factor (TFIIIC) (partial)	Yes	94.5		36	B, C, K, P, T, TC
Alpha fetoprotein binding protein	Yes	99.3		18	PI, TC, G
CD11a gene	Yes	99.8		60	AT, AO, E, Bl
Gene product with similarity to chp	Yes	80		5	SI, F
Protein kinase C beta subunit (PRKCB1)	Yes	100	Splice	72	M
BCMA antisense	No	100		0	
Acyl carrier protein (ACP) (partial)	Yes	100		85	M
Mito-glutamyl tRNA synthetase-like gene	Yes	62.3		4	SK
Rat P8 homologue	Yes	83	Splice	110	M
Batten disease gene (CLN3)	Yes	100	Splice	90	M
KIAA0220	No	100		30	M
Sodium channel (SCNN1B)	No	100		4	T
Kinesin-like gene	Yes	97.3		24	B, P, SY
MAZ-like gene (partial)	Yes	100	Splice	38	M
orip-like gene (partial)	Yes	99.4		76	M
SA gene	Yes	100		30	ER, CNS, O, PY, TC, U
Rat RHZG16P homologue	Yes	92		10	C
KIAA0370	Yes	100		175	M
CC chemokine (STCP-1)	Yes	100	Splice	117	M
CXC3 chemokine precursor	Yes	100		108	B, L, S, BR
T-cell directed chemokine	No	100		0	
RNA polymerase II subunit (hrpB33)	Yes	100		200	M
Rat kidney-specific (ks) homologue	Yes	92		8	K, L, S
Rat (ks)-like gene	Yes	82		2	T
Interleukin 4 receptor (IL4R)	No	100		64	O, PA, L, PG
Amyloid precursor protein (APP-B1)	Yes	100		76	M
Pre-mRNA splicing factor (PRP16)	Yes	99.9		97	M
Haptoglobin	Yes	99.9		296	M
Cell division control (CDC37)-like gene	Yes	80		0	
Transcription factor (NFAT4A)	Yes	100		15	T, J, LU
Haptoglobin-related protein	Yes	99.9		270	M
Gene with similarity to Dynein beta chain	No	48.8		0	0

tissue distribution matching large numbers of colon and irradiated colon derived human and mouse ESTs.

Genes with no significant match to previously identified genes in the databases accounted for 48% of the annotated genes identified during this study. A number of these unknown gene products showed a tissue specificity including A-334D11.3, which matched only colon derived ESTs; A-951C11.2, which matched only Jurkat cells; A-233A8.1, which matched only brain ESTs; and A-363E6.1, which matched only placental ESTs. In contrast, among the unknown gene products, those identified within BAC clones containing duplicon A (A-13F4, A-270G1, A-589H1, etc.) were found to have significant database matches to large numbers of human ESTs but none to corresponding ESTs from mouse or other species. This poses the question whether these annotated genes are actually transcribed or whether the EST matches are as a result of transcribed repeated elements from the 3′ UTRs of genes.

Large Intrachromosomal Duplications

A striking feature of the DNA sequenced in this study is the presence of a number of intrachromosomal duplicons of greater than 20 kb found interspersed throughout 16p (Fig. 3). These duplications contain both partial and entire previously identified genes and

TABLE 1—*Continued*

Name assigned	Mouse hits	Splice form	Human EST No.	Tissues
Unknown A334D11.1	Yes		19	MU, H, B
Unknown A-334D11.3	Yes		4	C
Unknown A-270G1.1	Yes		18	HE, BC, MY
Unknown A-270G1.2	Yes		148	UT, BC
Unknown A-13F4.3	No	Splice	140	M
Unknown A-13F4.5	No	Splice	138	M
Unknown A-211C6.1	No		21	M
Unknown 735G6.2	Yes		10	H, K
Unknown 735G6.5	Yes		110	M
Unknown A-362G6.1	Yes		32	C, FH, TC, IF, MY, L
Unknown A-362G6.2	Yes		91	FB, IB, PU, FS, FL
Unknown A-589H1.1	No	Splice	140	M
Unknown A-551G9.5	No		112	M
Unknown 327O24.1	Yes		22	C, MU. FH, MS, IB, AD
Unknown 327O24.3	No		8	L, S, T, TC, BC
Unknown A-61E3.4	Yes		46	M
Unknown 101F10.1	Yes		6	R, PL, GB, BR
Unknown 101F10.1	Yes		18	ET, O, HE, IB
Unknown 101F10.1	No		51	M
Unknown A-363E6.1	No			PL
Unknown A-363E6.2	Yes		35	M
Unknown A-635H12.2	Yes		40	M
Unknown A-951C11.2	No		2	J
Unknown A-3H8.1	No	Splice	11	T, MS, BT
Unknown A-575C2.4	No		31	M
Unknown 44M2.1	Yes		40	M
Unknown 44M2.2	Yes		35	PY, T, O, BC, L, S
Unknown 44M2.4	Yes		2	TC, G
Unknown 625P11.1	No		10	B, IB, FB, EY
Unknown A-152E5.4	Yes	Splice	104	M
Unknown A-152E5.6	Yes		60	M
Unknown A-152E5.8	Yes		88	FL, FS, FC
Unknown A-152E5.5	Yes		120	M
Unknown A-233A8.1	Yes		3	B
Unknown A-355G7.1	No		13	C, LU, PA, P, O, L, S
Unknown A-319E8.1	Yes		17	M
Unknown A-319E8.2	No		2	FB, FK
Unknown A-761H5.5	No	Splice	108	M
Unknown A-69G12.1	No		74	M
Unknown A-388D4.1	Yes		3	T, L, S

Note. Genes were divided into those for which a tentative function could be assigned based on sequence similarity and genes of unknown identity (termed unknown gene products). In the case of unknown gene products, the name is composed of the locus name consisting of the BAC clone name and a unique identifier. The presence of a significant match to a mouse expressed sequence is noted. For genes with assigned names the percentage similarity to the protein matched is indicated. The number of human EST database matches is noted, as is a summary of the tissues from which these ESTs were derived as well as the presence of putative alternative splice forms if any. The tissues are coded as follows: M, multiple; BR, breast; E, embryo; BC, B-cells; R, retina; O, ovary; J, Jurkat cells; U, uterus; L, liver; C, colon; S, spleen; K, kidney; LU, lung; PI, pineal gland; FL, fetal liver; G, germ cell tumor; FS, fetal spleen; AT, activated T-cells; FLU, fetal lung; AO, aorta; B, brain; E, endothelial cells; FB, fetal brain; Bl, blood; IB, infant brain; SI, small intestine; T, testis; F, fetus; P, prostate; SK, skeletal; TC, tonsillar cells; SY, synovial; CE, cerebellum; ER, ear; P, prostate; PY, parathyroid; TY, thymus; PA, pancreas; GB, gall bladder; PG, pituitary gland; HE, HeLa cells; H, heart; MY, melanocytes; FH, fetal heart; PU, pregnant uterus; EY, eye; FC, fetal colon; PL, placenta; MS, multiple sclerosis; FK, fetal kidney; BT, breast tumor; AD, adrenal gland; ET, endometrial tumor; NE, neuroepithelium; WT, Wilms tumor; UT, uterine tumor; HS, hippocampus; CB, cerebellum; MU, muscle.

putative transcripts. The exact size of some of these duplicons is unclear as they are complex in structure with components being present or absent among different representative copies, and therefore approximate sizes are given (Fig. 3). When located adjacent to one another, the duplicated elements are tandemly arranged, but it is not yet known whether all have the same orientation on the chromosome.

BAC clone A-13F4. A generalized description of the basic duplicated elements sequenced to date is outlined (Fig. 3). An example of duplications containing two copies of a *PKD1*-like gene (also termed the homologous gene loci, *HG1* and *HG2;* European Polycystic Kidney Disease Consortium, 1994) is found on BAC clone A-13F4. BAC clone A-13F4 contains two copies of a *PKD1*-like gene, one of the best studied duplicons on chromosome 16. Several copies of the *HG* loci were predicted to map to 16p13.1 when the *PKD1* gene was first identified in 16p13.3 (European Polycystic Kidney Disease Consortium, 1994). Analysis of A-13F4 reveals that the two *HG* loci (*HG1* and *HG2*) lie immediately adjacent to each other in the same orientation as

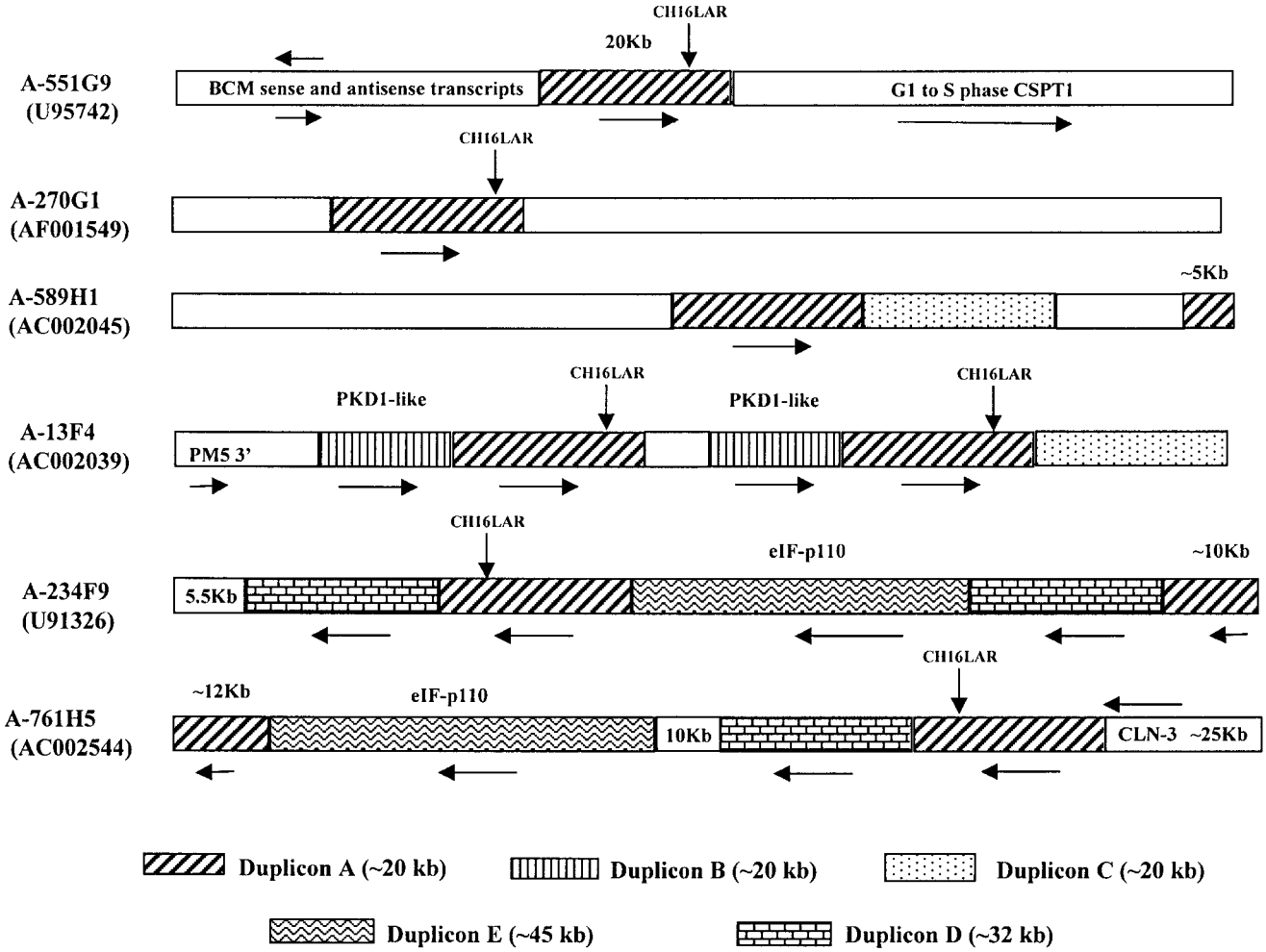

FIG. 3. BAC clones containing various intrachromosomally duplicated regions. The sizes for the duplications given are approximate.

shown (Fig. 4). The 5′ regions of both loci are similar to the 5′ end of *PKD1*, although there are significant differences (see below), and both *HG* loci have very similar 3′ regions that are not related to *PKD1*.

PKD1-like areas. The similarity between *HG1* and *PKD1* starts 1365 bp 5′ to the putative *PKD1* transcriptional start site (Hughes *et al.*, 1995) and extends 87 bp into *PKD1* exon 33. Two major deletions distinguish these loci (Fig. 4). A deletion of greater than 10 kb is seen in the intervening sequence following exon 1 (IVS1) of *HG1* compared to *PKD1*, and a region of greater than 4 kb containing most of *PKD1* exon 15 and all of exon 16 is also absent from *HG1* (Fig. 4). The similarity between *HG2* and *PKD1* begins in IVS1 of *PKD1*, 1408 bp before exon 2. Therefore *HG2* has no exon 1 or 5′ promoter sequences. The sequence similarity continues through to an analogous position in exon 33, but exons corresponding to 15 and 16 of *PKD1* are present in *HG2*. Analysis of the coding sequence shows 118 sequence differences between *HG1* and *PKD1*, with 206 seen in the larger corresponding area between *HG2* and *PKD1*, representing sequence variation of 1.84 and 2.07%, respectively. Among these sequence differences, 73.03% are common to both *HG* loci. The majority of changes are base substitutions, some resulting in amino acid changes. However, three single base changes common to *HG1* and *2* (exons 18, 23, and 25), one unique to *HG1* (exon 8), and one specific to *HG2* (exon 15) are nonsense mutations. Two frameshifting mutations are seen in *HG1* (exons 23 and 29), and one is seen in HG2 (exon 15). Three mutations that disrupt consensus splice sites are found in *HG1* (the first in exon 11). Analysis of the intronic sequence from *HG1* and *HG2* showed a mutation level of 2.18% in *HG1* and 2.64% in *HG2* when compared with *PKD1*. The intronic mutation frequency is similar to that observed in exonic sequence. Twenty-six deletions/insertions (ranging from 4 to 28 bp) were observed in the intronic areas of the *HG* loci compared with *PKD1*.

HG-specific areas. A region of duplicated sequence is shared at the 3′ ends of the *HG1* and *HG2* loci, but is not found at the original *PKD1* locus. This sequence of 20,443 bp, termed duplicon A, has also been identified within a number of other BAC clones but not associated with *PKD1* sequence (Fig. 3). The identity of the two copies of duplicon A within BAC A-13F4 is greater than 99%, and the sequence variation seen between these copies and copies of the duplicon in the other clones (Fig. 3) is less than 5%. Analysis of *HG*

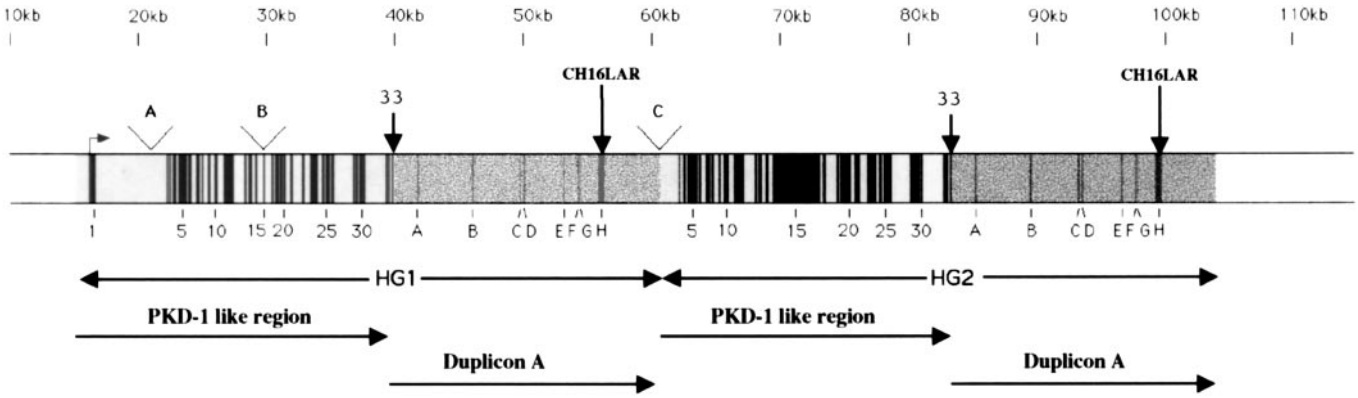

FIG. 4. Diagram of BAC clone A-13F4 showing the two copies of the PKD1-like genes HG1 and HG2. Exons common to PKD1 are numbered, while those present within the HG associated duplicated regions are lettered. The genomic region similar to PKD1 is shown in yellow, and the HG genomic areas (duplicon A) are shown in gray. HG1 exons are coded as PKD1-like (red) or HG specific (pink), with dark and light blue illustrating the corresponding areas of HG2. Every fifth PKD1-like exon is numbered (corresponding to PKD1 exon numbers), and each HG specific exon is lettered (A–H). Major deletions in the HG region compared to PKD1 are illustrated with arrows. A is a 10,436-bp deletion in IVS1 (7984–18,419 nt of PKD1 genomic sequence). B is a 4312-bp deletion of all but 92 bp of exon 15, IVS15, exon 16, and 415 bp of IVS16 (27,498–31,809 nt). C results in removal of 5′ promoter sequence, exon 1, and IVS1 up to 1596 bp before exon 2. Gaps in the genomic sequence corresponding to HG1 exon 1, IVS 21, IVS 22, HG2 IVS 20, exon IVS 21, and IVS 22 make analysis of these areas impossible. The location of the transition from PKD1-like to duplicon A sequence is shown at PKD1 exon 33, and the positions of the CH16LAR repeats are indicated.

cDNA sequence from clones HG4 and 11/21 (European Polycystic Kidney Disease Consortium, 1994) shows that up to eight exons are encoded by duplicon A, covering a maximum of 1156 bp and terminating with a polyadenylation site and a poly(A) tail. Sequence data from the two cDNAs provide some evidence of alternative splicing with 11/21 containing exons 31, IVS-31, and exon 32 as a single exon that is spliced to *HG* exon B (Fig. 4), while HG4 splices directly from *PKD1*-like exon 31 to *HG* exon A. Variations of this putative transcript are observed in each of the other copies of duplicon A. The last *HG*-specific exon contains five copies of an ~55-bp repeated unit also found in the 3′ UTR of another chromosome 16 gene, *KIAA0220* (Nagase *et al.*, 1996), identified on BAC clone A-61E3. The *HG*-specific regions also contain the previously described chromosome 16-specific low-copy-number repeat element (CH16LAR) that has been mapped to p13, p12, p11, and q22 on chromosome 16 (Stallings *et al.*, 1992) (Fig. 4). The overall structure of the *HG* genes suggests that they will not encode large proteins similar to *PKD1*. *HG2* does not have the *PKD1* first exon or promoter region and is therefore unlikely to be transcribed. The structure of *HG1* suggests that it may be transcribed; previously, *HG* transcripts have been characterized by Northern blotting (European Polycystic Kidney Disease Consortium, 1994). The calculated size of the *HG1* transcript (~8 kb) and its structure suggest that it may transcribe the ~8.5-kb mRNA, termed *HG-C* (European Polycystic Kidney Disease Consortium, 1994). The splicing, nonsense, and frameshifting mutations indicate that any protein product would be short, probably terminating by exon 8 and resulting in a product of ~55 kDa.

Other intrachromosomal duplicated regions. In addition to the *PKD1/HG* duplications, other genes including the translation initiation factor gene (*eIF3-p110*) and a number of putative genes of unknown function were found to be duplicated (Fig. 3). Unlike *PKD1*, the *eIF3-p110* gene was found to be duplicated over its entire coding region but it has not been determined which of the BAC clones contains the original locus of *eIF3-p110* or whether either copy is transcribed. The *eIF3-p110* duplicon (duplicon E) shared 99.6% identity between the two copies on BAC clones A-234F9 and A-761H5 over a genomic region of 23,593 bp containing the entire coding region of the gene. For the intronic sequence, the similarity between the duplicons is 99.5% with a 9-bp deletion in A-234F9 and deletions in A-234F9 within a dinucleotide (CA) repeat accounting for almost all of the sequence differences. At the level of coding sequence, the similarity between the predicted *eIF-p110* transcripts in both BAC clones is 99.9%. The *eIF-p110* transcript from BAC clone A-761H5 is identical to the previously published *eIF-p110* sequence (Asano *et al.*, 1997), and the *eIF-p110* transcript from BAC clone A-234F9 is 99.9% similar containing two sequence differences that result in the alteration of two adjacent amino acids within the predicted protein translation.

Another region of apparently recent duplication contains two copies of a 12,066-bp duplicon (Duplicon F) with 99.8% similarity separated by 25,405 bp found within BAC clone A-923A4. Unlike the duplicons previously described, the two copies are inverted with respect to one another, and neither contains CH16LAR sequences. Duplicon F is flanked by genes encoding proteins with a high degree of similarity (92 and 82%, respectively) to the rat kidney-specific (*ks*) gene (AF062389). The source of the duplication is unclear. This BAC contains the highest percentage of LTR elements of any sequenced clone in this analysis (17.41%), and these may have contributed to the instability of this locus. Duplicon F itself contains a large number of

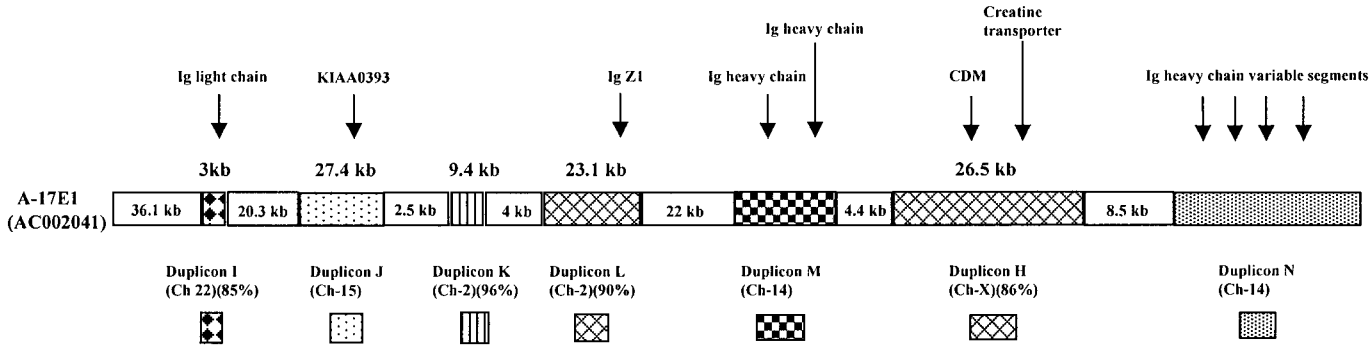

FIG. 5. BAC clone A-17E1 containing interchromosomally duplicated regions. Arrows indicate the position of genes (pseudogenes) within the sequence. Approximate sizes are given for both the duplicated regions and the distances between these regions.

LTR elements (32.72%) and a small number of SINEs (1.9%) but does not contain any LINE elements.

The haptoglobin (*Hp*) locus found on BAC clone A-259H10 represents another site of instability. It appears that this locus has undergone a homologous unequal crossover event resulting in a triplication of the locus. This triplication was followed by deletion of one locus in humans (Oliviero *et al.*, 1985). This previously documented duplication has been reported to have arisen as the result of a nonhomologous, probably random, crossing-over within different introns of two *Hp1* genes (Maeda *et al.*, 1984).

Interchromosomal Duplications

In addition to intrachromosomal duplications, analysis of the genomic sequence revealed two distinct interchromosomally duplicated regions both derived from Xq28. BAC clone 37914 carries a 9852-bp duplicated region (duplicon G) containing a partial adrenoleukodystrophy (*ALD*) pseudogene. The *ALD* partial pseudogene comprises the terminal 4 exons of the 10 exons of *ALD*. The duplicated region shares 94.1% identity with the published sequence (Brenner *et al.*, 1997). BAC clone A-17E1 also contains a number of duplicated regions from other chromosomes (Fig. 5). For some of these duplications, the ancestral locus has not been completely sequenced, so it is possible that the sizes of these duplications may be underestimated. Duplicon I contains a partial immunoglobulin (Ig) light chain pseudogene whose original locus is on chromosome 22 (Kawasaki *et al.*, 1997). Duplicon J contains the KIAA0393 pseudogene whose original locus is on chromosome 15. Downstream of duplicon J are two identifiable duplicated regions: duplicon K and duplicon L, both with significant sequence identity to BAC clone A-101B6, which has been assigned to chromosome 2. Duplicons L and K share a common region of 8540 bp with 96% sequence identity. Within the region unique to duplicon L lies a Z1 Ig V(K)1 pseudogene that is thought to have transposed from its original locus on chromosome 2 (Borden *et al.*, 1990). Duplicon M contains two Ig heavy chain variable region pseudogenes whose original locus is on chromosome 14 (Hobart *et al.*, 1981). These pseudogenes have also been found on 15q11.2 and 16p11.2 (Tomlinson *et al.*, 1994). Duplicon H, whose original locus is on Xq28, contains a creatine transporter pseudogene and a partial pseudogene of the tumor-associated antigen (*CDM*) protein containing the three terminal exons of the original four (Brenner *et al.*, 1997). Duplicon N contains four Ig heavy chain pseudogene segments of the Ig heavy chain gene. A region consisting of seven Ig heavy chain variable segments over 160 kb of DNA, which may represent the Ig heavy chain duplicons, has been previously described as resulting from a translocation from chromosome 14 (Nagaoka *et al.*, 1994).

Duplication of STS Markers

One impact of the various sequence duplications observed is that STS markers contained within these repeated regions do not in fact tag unique locations in the genome although it is likely that they appeared to be unique at the lower resolution of the map on which they were originally placed. A listing of STS markers known to be present more than once either within or between individual BAC clones that are not a result of legitimate overlaps is given in Table 2. This resulted in difficulty in accurately placing a number of the BACs on the physical map that contain duplicated regions including BAC clones A-234F9 and A-551G9.

DISCUSSION

A number of observations reported here may represent more generalized features of both chromosome 16 and the genome as a whole. The overall percentage of bases designated as coding (2.1%) as well as the average gene density (1 gene every 100 kb) is lower than the genome average. This may be explained in part because we did not annotate a gene based solely on results of computational gene finding programs, as *de novo* gene prediction is not sufficiently accurate. This is especially true of genes comprising multiple exons separated by large tracts of noncoding sequence for which prediction of intron/exon boundaries and assembly of the predicted exons into the correct model of gene structure are particularly error-prone (Xu *et al.*, 1997).

The current attempts to incorporate both DNA and protein database matches into such predictions and the

TABLE 2

Duplicated STS Markers

STS	BAC Clones
SHGC-36058	A-589H1, A-761H5, A-13F4 (X2)
SHGC-32146	A-234F9 (X2), A-761H5
D11S1053	A-589H1, A-88D1, A-17E1, A-100B4, 101B6, A-13F4
A002D07	A-100B4, A-101B6 (X3), A-29B12, A-270G1, A-88D1, A-17E1 (X4), 37914, 1000D7, A-355G7 (X2), A-27D10, A-161G4, A-10F4, 44M2(X2), A-542B9, A-926E7, A-731F11, 502C10, A-345G4
D16S3353	A-551G9, A-234F9, A-589H1, A-13F4 (X2)
D11S2442	A-2A8, A-142A6, A-101B6, A-88D1 (X2), A-17E1, A-345G4, A-355G7 (X2), 44M2 (X2)
SGC33003	A-13F4 (X2)
D11S2560	A-363E6, A-88D1, 254P9, A-911E12, 625P11 (X2), A-67A1, A-952F10, A-224D6, A-362G6

Note. STS sequences present in more than one copy and that do not appear to represent legitimate overlaps between BAC clones. Where sequences appear more than once in the same clone, an X followed by the number is used to denote the number of times the STS appears. Each of the BAC clones was derived from the Caltech Human BAC Library CIT987SK.

increasing use of expressed sequence databases from a variety of species should result in significant improvements in this area (Bailey *et al.*, 1998). Splice sites confirmed by EST matches can also help to constrain the possible combinations of predicted exons that contribute to the final gene structure definition. Another factor contributing to the relatively low apparent coding content is the fact that frequently ESTs do not cover 100% of the coding region of a gene particularly at the 5′ end. One particular region of 16p (region A (Fig. 2)) from 16p13.11, encompassing 718 kb of sequence and containing BAC clones A-65D3, A-777B5, A-732D3, A-98H8, 502C10, 27D10, and A-731F11, was found to be particularly gene poor. No gene structure could be annotated for this region as none of the EST database matches spanned exons. One indication that this region is indeed gene poor is that there is an obvious difference in overall GC content between this region and the rest of the sequence (Fig. 2).

Although the quantity of sequence presented here represents only a very small fraction of the entire human sequence, it illustrates in a number of ways the complexity of a large continually evolving genome. Sequence analysis of BAC clones 37914 and A-17E1 in the pericentromeric region of 16p11.2 suggests the presence of multiple duplications that originated from other locations in the genome. These observations support previous reports (Tomlinson *et al.*, 1994; Eichler *et al.*, 1996, 1997) that the region near the centromere of the short arm of chromosome 16 is composed of a complex patchwork of genomic segments that have been transposed from other locations in the human genome. The considerable degree of sequence similarity (>90%) indicates that these duplicons arose quite recently (<20 mya) and implicate the pericentromeric region of 16 as a site of considerable evolutionary instability. Recent investigations into the structure and sequence of other pericentromeric regions (Zimonjic *et al.*, 1997; Regnier *et al.*, 1997; Ritchie *et al.*, 1998; Eichler, 1998) suggest that extensive interchromosomal duplications may be a general property of many human pericentromeres.

In the case of intrachromosomal duplications the presence of duplicated sequences adjacent to both copies of the *PKD1*-like gene on BAC A-13F4 has likely contributed to the instability of this region. Two other examples of duplications in proximity to genes involved in disease-causing rearrangements have been found: the G1 to S phase transition (*GSPT1*) gene with breakpoints in acute nonlymphocytic leukemia (Ozawa *et al.*, 1992) and the deletion in *CLN3* seen in Batten disease patients (Mitchison *et al.*, 1997). The phenomenon of low-copy repeats as the cause of DNA rearrangement associated with human disease has been described previously for many conditions. These include color blindness (Nathans *et al.*, 1986), hypertension secondary to glucocorticoid-remediable aldosteronism (Lifton *et al.*, 1992), hemophilia (Lakich *et al.*, 1993) and Smith–Magenis syndrome (Chen *et al.*, 1997). In the case of *PKD1*, one mechanism by which the presence of highly homologous copies of a gene might contribute to disease phenotype is by acting as a reservoir for mutations via gene conversion events (Watnick *et al.*, 1998).

The CH16LAR sequences have a core short repeat similar to that of minisatellite sequences that are associated with high rates of mutation and recombination (Stallings *et al.*, 1992). The presence of chromosome-specific repetitive sequences may destabilize the surrounding genomic sequence, leading to frequent duplication events or unequal cross-over in meiosis that results in the multiplication of all or parts of genes within the region. As additional copies of these repeats are sequenced, we expect that the number of variable segments and the number of duplicated genes will increase, and in fact, it is not yet possible to construct a complete consensus sequence of the repeated region. It is possible that these CH16LARs are not themselves responsible for the duplications observed in this project even though copies of the repeated sequence are found interspersed throughout the most frequent of the duplicons, duplicon A (Fig. 3).

Of the duplicated loci, BAC clone A-13F4, which contains the *HG1* and *HG2* loci, is the best studied and provides clues to how this region may have evolved. Previous FISH, Southern blotting, and Northern blotting studies suggested the presence of at least three

HG loci and possibly several more (European Polycystic Kidney Disease Consortium, 1994). These data and recent preliminary evidence of BACs containing alternatively arranged *HG* loci (unpublished observations) indicate that the genes described here are not the full complement of *HG* loci. Lack of evidence of other *HG* loci close to this tandemly repeated pair, however, indicates that these genes will be separated over a significant area within proximal 16p.

The previous evidence of *HG* mRNAs and the structure of *HG1* suggest that this is a transcribed gene. It is therefore likely that the ~1.4 kb of sequence 5′ to the transcript will form a functional promoter, similar to *PKD1*. It is not yet clear whether the small protein product predicted from this mutated locus is translated or is stable; studies with N-terminal antibodies will be required to test this. A stable *HG1* protein product, although very different from the full-length *PKD1* protein, would be interesting. Such a product would contain the signal sequence, the leucine-rich repeats, and possibly the lectin domain and hence could compete for ligands of these motifs in the extracellular environment. Analysis of sequence differences between the *HG* loci and *PKD1* gives some clues about its evolution. The degree of sequence divergence with *PKD1* (~2%) suggests that the separation from *PKD1* occurred relatively recently in evolutionary time. Preliminary data suggest that this may have occurred during primate evolution (unpublished observations). The finding of similar levels of sequence divergence in the *HG1* exons and introns suggests that this transcript may not generate a functional protein product. The greater sequence similarity between the *HG* loci than with *PKD1* suggests that duplication of *HG* occurred subsequent to the initial event transferring a copy of the 5′ part of *PKD1* from 16p13.3 to this more proximal area. It is also interesting that duplicon A including the promiscuous repeat, CH16LAR, is found in the 3′ region of the *HG* as well as adjacent to other recent duplications (Fig. 3) and suggests that this sequence may have played a role in the further reiteration of the *HG*-specific region (in some cases including the *PKD1*-related area) within proximal 16p.

In a more generalized sense, the presence of these duplicated regions has important implications for the utility of any methodology that relies on the use of sequence as a measure of uniqueness. These include the use of exact or nearly exact EST matches as evidence of the presence of a gene, the use of STS data as a measure of unique location in the genome, and restriction fingerprinting methods of determining clone overlap. The presence of duplicated regions has hampered the process of map building and particularly the construction of ordered contig maps for the p-arm as apparently unique locations to which STSs had been mapped were found to be duplicated. This has resulted in an inability to place a number of the BAC clones containing duplicated regions on the chromosome 16 map including BAC clones A-234F9 and A-551G9. Large, low-copy-number repeats are unlikely to be restricted to chromosome 16, and it is likely that more examples will emerge as the sequencing of the genome progresses. For the BACs to be placed uniquely on the chromosome 16 physical map, unique sequence adjacent to each repeat must be identified and mapped. As genome sequencing progresses, identification of these unique regions will be facilitated.

The exact role of duplicated regions in gene or genome function is one of the mysteries of the genome that will not be revealed by sequence analysis alone. It can be expected, though, that gene structure, and perhaps other elements such as regulatory regions, might eventually be predictable from sequence alone. That will be an accomplishment to equal that of completing the genome sequence itself. In the meantime, early genome sequence data will continue to contribute to the identification of new genes as well as provide a framework by which current mechanisms of gene and gene structure prediction can be evaluated.

ACKNOWLEDGMENTS

We are grateful for the excellent sequencing and finishing work provided by many members of the TIGR Sequencing Core Facility; we are grateful to Norman Doggett for providing the initial map information and to Lizin Zhou, Xiaoying Lin, and Steve Rounsley for discussions on issues of annotation. This work was supported by Grant R01-HG01464 from NHGRI to M.D.A.

REFERENCES

Aaronson, J. S., Eckman, B., Blevins, R. A., Borkowski, J. A., Myerson, J., Imran, S., and Elliston, K. O. (1996). Toward the development of a gene index to the human genome: An assessment of the nature of high-throughput EST sequence data. *Genome Res.* **6**(9): 829–845.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., *et al.* (1991). Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252**(5013): 1651–1656.

Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., Lee, N. H., Kirkness, E. F., Weinstock, K. G., Gocayne, J. D., White, O., *et al.* (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377**(6547 Suppl.): 3–174.

Asano, K., Kinzy, T. G., Merrick, W. C., and Hershey, J. W. (1997). Conservation and diversity of eukaryotic translation initiation factor eIF3. *J. Biol. Chem.* **272**(2): 1101–1109.

Bailey, L. C., Jr., Searls, D. B., and Overton, G. C. (1998). Analysis of EST-driven gene annotation in human genomic sequence. *Genome Res.* **8**(4): 362–376.

Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., and Ouellette, B. F. (1998). GenBank. *Nucleic Acids Res.* **26**(1): 1–7.

Blattner, F. R., Plunkett, G., III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**(5331): 1453–1474.

Borden, P., Jaenichen, R., and Zachau, H. G. (1990). Structural features of transposed human VK genes and implications for the mechanism of their transpositions. *Nucleic Acids Res.* **18**(8): 2101–2107.

Brenner, V., Nyakatura, G., Rosenthal, A., and Platzer, M. (1997). Genomic organization of two novel genes on human Xq28: Compact head to head arrangement of IDH gamma and TRAP delta is conserved in rat and mouse. *Genomics* **44**(1): 8–14.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., and Venter, J. C. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**(5278): 1058–1073.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**(1): 78–94.

Cao, Y, Kang, H. L., Xu, X., Wang, M., Dho, S. H., Huh, J. R., Lee, B.-J., Kalush, F., Bocskai, D., Ding, Y., Tesmer, J. G., Lee, J., Moon, E., Jurecic, V., Baldini, A., Weier, H.-U., Doggett, N. A., Simon, M. I., Adams, M. D., and Kim, U.-J. A 12 Mbp complete coverage BAC contig map in human chromosome 16p13.1–11.2. *Genome Res.* **9**(8): 763–774.

Chen, K. S., Manian, P., Koeuth, T., Potocki, L., Zhao, Q., Chinault, A. C., Lee, C. C., and Lupski, J. R. (1997). Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nat. Genet.* **17**(2): 154–163.

Dauwerse, J. G., Jumelet, E. A., Wessels, J. W., Saris, J. J., Hagemeijer, A., Beverstock, G. C., van Ommen, G. J., and Breuning, M. H. (1992). Extensive cross-homology between the long and the short arm of chromosome 16 may explain leukemic inversions and translocations. *Blood* **79**(5): 1299–1304.

Dissing, M., Le Beau, M. M., and Pedersen-Bjergaard, J. (1998). Inversion of chromosome 16 and uncommon rearrangements of the CBFB and MYH11 genes in therapy-related acute myeloid leukemia: Rare events related to DNA-topoisomerase II inhibitors? *J. Clin. Oncol.* **16**(5): 1890–1896.

Doggett, N. A., Goodwin, L. A., Tesmer, J. G., Meincke, L. J., Bruce, D. C., Clark, L. M., Altherr, M. R., Ford, A. A., Chi, H. C., Marrone, B. L., *et al.* (1995). An integrated physical map of human chromosome 16. *Nature* **377**(6547 Suppl.): 335–365.

Eichler, E. E., (1998). Masquerading repeats: Paralogous pitfalls of the human genome. *Genome Res.* **8**(8): 758–762.

Eichler, E. E., Budarf, M. L., Rocchi, M., Deaven, L. L., Doggett, N. A., Baldini, A., Nelson, D. L., and Mohrenweiser, H. W. (1997). Interchromosomal duplications of the adrenoleukodystrophy locus: A phenomenon of pericentromeric plasticity. *Hum. Mol. Genet.* **6**(7): 991–1002.

Eichler, E. E., Lu, F., Shen, Y., Antonacci, R., Jurecic, V., Doggett, N. A., Moyzis, R. K., Baldini, A., Gibbs, R. A., and Nelson, D. L. (1996). Duplication of a gene-rich cluster between 16p11.1 and Xq28: A novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum. Mol. Genet.* **5**(7): 899–912.

European Polycystic Kidney Disease Consortium (1994). The polycystic kidney disease 1 gene encodes a 14 kb transcript and lies within a duplicated region on chromosome 16. *Cell* **77**(6): 881–894.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**(3): 186–194.

Fields, C., Adams, M. D., White, O., and Venter, J. C. (1994). How many genes in the human genome? *Nat. Genet.* **7**(3): 345–346.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**(5223): 496–512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., *et al.* (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**(5235): 397–403.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science* **274**(5287): 546, 563–567.

Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfing, T., Schellenberg, K., Marra, M., *et al.* (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6**(9): 807–828.

Hobart, M. J., Rabbitts, T. H., Goodfellow, P. N., Solomon, E., Chambers, S., Spurr, N., and Povey, S. (1981). Immunoglobulin heavy chain genes in humans are located on chromosome 14. *Ann. Hum. Genet.* **45**(Pt. 4): 331–335.

Huang, X., Adams, M. D., Zhou, H., and Kerlavage, A. R. (1997). A tool for analyzing and annotating genomic sequences. *Genomics* **46**(1): 37–45.

Hughes, J., Ward, C. J., Peral, B., Aspinwall, R., Clark, K., San Millan, J. L., Gamble, V., and Harris, P. C. (1995). The polycystic kidney disease 1 (PKD1) gene encodes a novel protein with multiple cell recognition domains. *Nat. Genet.* **10**(2): 151–160.

Jiang, J., and Jacob, H. J. (1998). EbEST: An automated tool using expressed sequence tags to delineate gene structure. *Genome Res.* **8**(3): 268–275.

Kawasaki, K., Minoshima, S., Nakato, E., Shibuya, K., Shintani, A., Schmeits, J. L., Wang, J., and Shimizu, N. (1997). One-megabase sequence analysis of the human immunoglobulin lambda gene locus. *Genome Res.* **7**(3): 250–261.

Kwon, O. J., Gainer, H., Wray, S., and Chin, H. (1996). Identification of a novel protein containing two C2 domains selectively expressed in the rat brain and kidney. *FEBS Lett.* **378**(2): 135–139.

Laabi, Y., Gras, M. P., Carbonnel, F., Brouet, J. C., Berger, R., Larsen, C. J., and Tsapis, A. (1992). A new gene, BCM, on chromosome 16 is fused to the interleukin 2 gene by a t(4;16)(q26;p13) translocation in a malignant T cell lymphoma. *EMBO J.* **11**(11): 3897–3904.

Lakich, D., Kazazian, H. H., Jr., Antonarakis, S. E., and Gitschier, J. (1993). Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat. Genet.* **5**(3): 236–241.

Lifton, R. P., Dluhy, R. G., Powers, M., Rich, G. M., Cook, S., Ulick, S., and Lalouel, J. M. (1992). A chimaeric 11 beta-hydroxylase/aldosterone synthase gene causes glucocorticoid-remediable aldosteronism and human hypertension. *Nature* **355**(6357): 262–265.

Liu, P. P., Wijmenga, C., Hajra, A., Blake, T. B., Kelley, C. A., Adelstein, R. S., Bagg, A., Rector, J., Cotelingam, J., Willman, C. L., and Collins, F. S. (1996). Identification of the chimeric protein product of the CBFB-MYH11 fusion gene in inv(16) leukemia cells. *Genes Chromosomes Cancer* **16**(2): 77–87. [Published erratum appears in *Genes Chromosomes Cancer* 1997, **18**(1):71]

Longhurst, T. J., O'Neill, G. M., Harvie, R. M., and Davey, R. A. (1996). The anthracycline resistance-associated (ara) gene, a novel gene associated with multidrug resistance in a human leukaemia cell line. *Br. J. Cancer* **74**(9): 1331–1335.

Maeda, N., Yang, F., Barnett, D. R., Bowman, B. H., and Smithies, O. (1984). Duplication within the haptoglobin Hp2 gene. *Nature* **309**(5964): 131–135.

Mallo, G. V., Fiedler, F., Calvo, E. L., Ortiz, E. M., Vasseur, S., Keim, V., Morisset, J., and Iovanna, J. L. (1997). Cloning and expression of the rat p8 cDNA, a new gene activated in pancreas during the acute phase of pancreatitis, pancreatic development, and regeneration, and which promotes cellular growth. *J. Biol. Chem.* **272**(51): 32360–32369.

Marra, M. A., Kucaba, T. A., Dietrich, N. L., Green, E. D., Brownstein, B., Wilson, R. K., McDonald, K. M., Hillier, L. W., McPherson, J. D., and Waterston, R. H. (1997). High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7**(11): 1072–1084.

Mitchison, H. M., Munroe, P. B., O'Rawe, A. M., Taschner, P. E., de Vos, N., Kremmidiotis, G., Lensink, I., Munk, A. C., D'Arigo, K. L., Anderson, J. W., Lerner, T. J., Moyzis, R. K., Callen, D. F., Breuning, M. H., Doggett, N. A., Gardiner, R. M., and Mole, S. E. (1997). Genomic structure and complete nucleotide sequence of the Batten disease gene, CLN3. *Genomics* **40**(2): 346–350.

Mitchison, H. M., Thompson, A. D., Mulley, J. C., Kozman, H. M., Richards, R. I., Callen, D. F., Stallings, R. L., Doggett, N. A., Attwood, J., McKay, T. R., *et al.* (1993). Fine genetic mapping of the Batten disease locus (CLN3) by haplotype analysis and demonstration of allelic association with chromosome 16p microsatellite loci. *Genomics* **16**(2): 455–460.

Nagaoka, H., Ozawa, K., Matsuda, F., Hayashida, H., Matsumura, R., Haino, M., Shin, E. K., Fukita, Y., Imai, T., Anand, R., *et al.* (1994). Recent translocation of variable and diversity segments of the human immunoglobulin heavy chain from chromosome 14 to chromosomes 15 and 16. *Genomics* **22**(1): 189–197.

Nagase, T., Seki, N., Ishikawa, K., Ohira, M., Kawarabayasi, Y., Ohara, O., Tanaka, A., Kotani, H., Miyajima, N., and Nomura, N. (1996). Prediction of the coding sequences of unidentified human genes. VI. The coding sequences of 80 new genes (KIAA0201–KIAA0280) deduced by analysis of cDNA clones from cell line KG-1 and brain. *DNA Res* **3**(5): 321–339, 341–354.

Nancarrow, J. K., Kremer, E., Holman, K., Eyre, H., Doggett, N. A., Le Paslier, D., Callen, D. F., Sutherland, G. R., and Richards, R. I. (1994). Implications of FRA16A structure for the mechanism of chromosomal fragile site genesis. *Science* **264**(5167): 1938–1941.

Nathans, J., Piantanida, T. P., Eddy, R. L., Shows, T. B., and Hogness, D. S. (1986). Molecular genetics of inherited variation in human color vision. *Science* **232**(4747): 203–210.

Okumura, K., Menninger, J., Stallings, R. L., Doggett, N. A., and Ward, D. C. (1994). In situ hybridization mapping of human chromosome 16: Evidence for a high frequency of repetitive DNA sequences. *Cytogenet. Cell Genet.* **67**(1): 61–67.

Oliviero, S., DeMarchi, M., Carbonara, A. O., Bernini, L. F., Bensi, G., and Raugei, G. (1985). Molecular evidence of triplication in the haptoglobin Johnson variant gene. *Hum. Genet.* **71**(1): 49–52.

Ozawa, K., Murakami, Y., Eki, T., Yokoyama, K., Soeda, E., Hoshino, S., Ui, M., and Hanaoka, F. (1992). Mapping of the human GSPT1 gene, a human homolog of the yeast GST1 gene, to chromosomal band 16p13.1. *Somat. Cell Mol. Genet.* **18**(2): 189–194.

Regnier, V., Meddeb, M., Lecointre, G., Richard, F., Duverger, A., Nguyen, V. C., Dutrillaux, B., Bernheim, A., and Danglot, G. (1997). Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. *Hum. Mol. Genet.* **6**(1): 9–16.

Ritchie, R. J., Mattei, M. G., and Lalande, M. (1998). A large polymorphic repeat in the pericentromeric region of human chromosome 15q contains three partial gene duplications. *Hum. Mol. Genet.* **7**(8): 1253–1260.

Samani, N. J., Whitmore, S. A., Kaiser, M. A., Harris, J., See, C. G., Callen, D. F., and Lodwick, D. (1994). Chromosomal assignment of the human SA gene to 16p13.11 and demonstration of its expression in the kidney. *Biochem. Biophys. Res. Commun.* **199**(2): 862–868.

Shimkets, R. A., Warnock, D. G., Bositis, C. M., Nelson-Williams, C., Hansson, J. H., Schambelan, M., Gill, J. R., Jr., Ulick, S., Milora, R. V., Findling, J. W., *et al.* (1994). Liddle's syndrome: Heritable human hypertension caused by mutations in the beta subunit of the epithelial sodium channel. *Cell* **79**(3): 407–414.

Smit, A. F. (1996). The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6**(6): 743–748.

Stallings, R. L., Doggett, N. A., Okumura, K., and Ward, D. C. (1992). Chromosome 16-specific repetitive DNA sequences that map to chromosomal regions known to undergo breakage/rearrangement in leukemia cells. *Genomics* **13**(2): 332–338.

Stallings, R. L., Whitmore, S. A., Doggett, N. A., and Callen, D. F. (1993). Refined physical mapping of chromosome 16-specific low-abundance repetitive DNA sequences. *Cytogenet. Cell Genet.* **63**(2): 97–101.

Struk, B., Neldner, K. H., Rao, V. S., St Jean, P., and Lindpaintner, K. (1997). Mapping of both autosomal recessive and dominant variants of pseudoxanthoma elasticum to chromosome 16p13.1. *Hum. Mol. Genet.* **6**(11): 1823–1828.

Sutton, G., White, O., Adams, M. D., and Kerlavage, A. R. (1995). TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.* **1**(1): 9–19.

Szpirer, C., Riviere, M., Szpirer, J., Levan, G., Guo, D. F., Iwai, N., and Inagami, T. (1993). Chromosomal assignment of human and rat hypertension candidate genes: Type 1 angiotensin II receptor genes and the SA gene. *J. Hypertens.* **11**(9): 919–925.

Templeton, N. S., Rodgers, L. A., Levy, A. T., Ting, K. L., Krutzsch, H. C., Liotta, L. A., and Stetler-Stevenson, W. G. (1992). Cloning and characterization of a novel human cDNA that has DNA similarity to the conserved region of the collagenase gene family. *Genomics* **12**(1): 175–176.

Tomlinson, I. M., Cook, G. P., Carter, N. P., Elaswarapu, R., Smith, S., Walter, G., Buluwela, L., Rabbitts, T. H., and Winter, G. (1994). Human immunoglobulin VH and D segments on chromosomes 15q11.2 and 16p11.2. *Hum. Mol. Genet.* **3**(6): 853–860.

van Soest, S., Swart, J., Tijmes, N., Sandkuijl, L. A., Rommers, J., and Bergen, A. A. (1997). A locus for autosomal recessive pseudoxanthoma elasticum, with penetrance of vascular symptoms in carriers, maps to chromosome 16p13.1. *Genome Res.* **7**(8): 830–834.

Watnick, T. J., Gandolph, M. A., Weber, H., Neumann, H. P. H., and Germino, G. G. (1998). Gene conversion is a likely cause of mutation in PKD1. *Hum. Mol. Genet.* **7**(8): 1239–1243.

Xu, Y., Einstein, J. R., Mural, R. J., Shah, M., and Uberbacher, E. C. (1994). An improved system for exon recognition and gene modeling in human DNA sequences. *Ismb* **2:** 376–384.

Xu, Y., Mural, R. J., and Uberbacher, E. C. (1997). Inferring gene structures in genomic sequences using pattern recognition and expressed sequence tags. *Ismb* **5:** 344–353.

Xu, Y., and Uberbacher, E. C. (1997). Automated gene identification in large-scale genomic sequences. *J. Comput. Biol.* **4**(3): 325–338.

Yu, S., Mangelsdorf, M., Hewett, D., Hobson, L., Baker, E., Eyre, H. J., Lapsys, N., Le Paslier, D., Doggett, N. A., Sutherland, G. R., and Richards, R. I. (1997). Human chromosomal fragile site FRA16B is an amplified AT-rich minisatellite repeat. *Cell* **88**(3): 367–374.

Zhang, J., and Madden, T. L. (1997). PowerBLAST: A new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res.* **7**(6): 649–656.

Zimonjic, D. B., Kelley, M. J., Rubin, J. S., Aaronson, S. A., and Popescu, N. C. (1997). Fluorescence in situ hybridization analysis of keratinocyte growth factor gene amplification and dispersion in evolution of great apes and humans. *Proc. Natl. Acad. Sci. USA* **94**(21): 11461–11465.