Construction of a High-Resolution 2.5-Mb Transcript Map of the Human 6p21.2–6p21.3 Region Immediately Centromeric of the Major Histocompatibility Complex

Nicos Tripodis,[1,3] Sophie Palmer,[2] Sam Phillips,[2] Sarah Milne,[2] Stephan Beck,[2] and Jiannis Ragoussis[1,4]

[1]*Genomics Laboratory, Division of Medical and Molecular Genetics, Guy's Campus, GKT School of Medicine, King's College London SE1 9RT, UK;* [2]*The Sanger Centre, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK*

We have constructed a 2.5-Mb physical and transcription map that spans the human 6p21.2–6p21.3 region and includes the centromeric end of the MHC, using a combination of techniques. In total 88 transcription units including exons, cDNAs, and cDNA contigs were characterized and 60 were confidently positioned on the physical map. These include a number of genes encoding nuclear and splicing factors (Ndr kinase, HSU09564, HSRP20); cell cycle, DNA packaging, and apoptosis related [p21, HMGI(Y), BAK]; immune response (*CSBP*, *SAPK4*); transcription activators and zinc finger-containing genes (*TEF–5, ZNF76*); embryogenesis related (*Csa–19*); cell signaling (*DIPP*); structural (*HSET*), and other genes (*TULP1*, *HSPRARD*, *DEF–6*, *EO6811*, cyclophilin), as well as a number of RP genes and pseudogenes (*RPS10*, *RPS12–like*, *RPL12–like*, *RPL35–like*). Furthermore, several novel genes (a *Br140–like*, a *G2S–like*, a *FBN2–like*, a *ZNF–like*, and *B1/KIAA0229*) have been identified, as well as cDNAs and cDNA contigs. The detailed map of the gene content of this chromosomal segment provides a number of candidate genes, which may be involved in several biological processes that have been associated with this region, such as spermatogenesis, development, embryogenesis, and neoplasia. The data provide useful tools for synteny studies between mice and humans, for genome structure analysis, gene density comparisons, and studies of nucleotide composition, of different isochores and Giemsa light and Giemsa dark bands.

The 6p21.3 band of chromosome 6 is one of the most intensely studied regions of the human genome, because of the presence of the major histocompatibility complex (MHC) that occupies 4 Mb within that chromosomal segment.

The mouse MHC is situated on chromosome 17 within the fourth distal inversion, in(17)4 of the *t* haplotype. The colinearity of genes between the mouse and human MHC has been shown to extend out of the human MHC toward both the centromere and the telomere (Tripodis et al. 1998). Sporadic cases have been reported associating the HLA region in humans with spontaneous abortion, transmission distortion, and infertility, all characteristics of the mouse *t* complex (for review, see Kostyu 1994) and it is possible that the association extends toward the regions flanking the MHC.

The 6p21.3 region has been also associated with neoplasia, as a number of chromosomal rearrangements with breakpoints have been observed in various human cancers (Johansson et al. 1993; Williams et al. 1997; Xiao et al. 1997). Several genes believed to be involved directly in cancer have been physically mapped within the 6p21.2–21.3 region. These are the high mobility group *HMGI(Y)* gene (Friedmann et al. 1993), the cyclin-dependent kinase inhibitor 1A (also known as *p21*; el-Deiry et al. 1993; Harper et al. 1993), and the *PIM1* oncogene (Cuypers et al. 1986).

To characterize this region in detail, we have constructed a physical map (Tripodis et al. 1998) that extends >2.5 Mb from the centromeric end of the MHC and marked by the *HKE-3* gene, to include the marker D6S291. As the main tool for transcript map construction we used the cDNA selection technique (Lovett et al. 1991; Korn et al. 1992) and complete genomic sequencing followed by sequence analysis. These were enhanced by applying exon trapping (Church et al. 1994), isolation and sequencing of GC-rich fragments (Shiraishi et al. 1995), and direct screening of cDNA libraries in selected areas.

[3]*Present address: Division of Molecular Genetics (H5), The Netherlands Cancer Institute, Plesmanlaan 121, Amsterdam 1066CX, The Netherlands.*
[4]Corresponding author.
E-MAIL ioannis.ragoussis@kcl.ac.uk; FAX 0171-955-4444.

10:454–472 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

This is the first report of a detailed transcription map of the region immediately centromeric of the MHC in humans. The transcription map and genomic sequence encompass the boundary of a Giemsa-positive and a Giemsa-negative band. The application of several gene identification techniques combined with complete genomic sequencing and analysis has allowed direct comparison between gene prediction programs, expressed sequence taq (EST) mapping information, and isolated exons and cDNAs. The resulting detailed transcription map fully integrates the existing and generated gene information and includes the identification and elucidation of the organization for a number of genes. The data provide valuable resources, not only to medical genetics but also to studies of genome organization.

RESULTS

Direct Screening of a cDNA Library and exon trapping in a 410-kb region encompassing *TCP–11* and *ZNF76*

The close mapping of *ZNF76* and *TCP-11* (Ragoussis et al. 1992), as well as a relatively high number of *Bss*HII restriction sites in the same area (Tripodis et al. 1998) indicated the presence of CpG islands and consequently, additional coding sequences in this 410-kb region. Two different approaches were used to generate additional coding sequences: direct screening of a cDNA library and exon trapping.

Direct screening of the U937 cDNA library resulted in the isolation of two overlapping cDNA clones of ~2 kb length, B1 and Δ5. The two clones map on cosmids 15a and 4N (Fig. 1). They are part of the 3′-untranslated region (UTR) of the 6.3-kb mRNA *KIAA0229*.

Twelve cosmids from the *TCP11* region were pooled in a single exon trapping experiment. The individual exons had the following characteristics: Clones x2A6 and x13H9 (Table 1) have a 100% homology with *KIAA0229* and therefore, must be exons of this mRNA. Clone x9C3 was mapped in the T3 end fragment of cosmid 15a, which contains the gene encoding TCP-11. Clones x1C7 and x8F2 (Table 1) have continuous open reading frames (ORFs) but no similarity was identified in any database search. Clone x8F2 is partially overlapping with a selected cDNA (clone 4.2/G5; Table 1), which increases the possibility of it being a bona fide exon. Clone x2E11 (Table 1) maps in cosmids 44N22 and 33F1 (Fig. 1) and has a continuous ORF. No identity with known sequences was found in electronic searches. Two selected cDNA clones contain the entire sequence of 2E11. These putative exons constitute 75% of the insert containing gridded clones.

Generation of GC-Rich Fragments in the Area Centromeric to p21

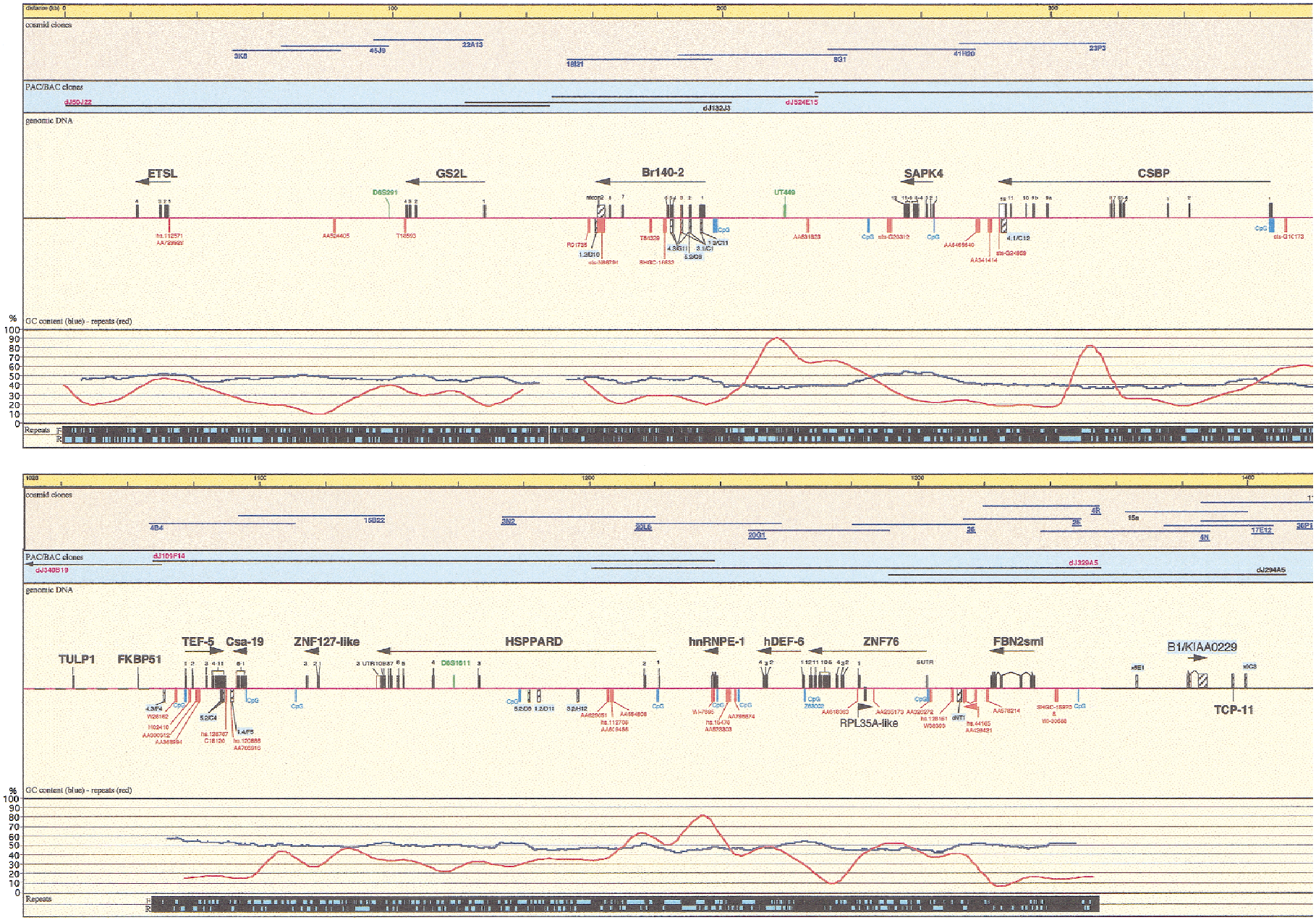
The isolation of GC-rich fragments as an alternative to sample sequencing was tried on three cosmids that map centromeric to the *p21* gene: 54P5, 27D8, and 23H4 (Fig. 1). The particular region was selected because of the presence of *Bss*HII sites indicative of CpG islands and relatively high GC content.

Six fragments were isolated from the three cosmids and multiple duplicate clones analyzed. The fragments fall in two groups, the 23H4-A1 group, which is 725 bp long and has a T_m = 79.9°C and the 23H4-B4 group, which is 961 bp long and has a T_m = 79.1°C (Table 1). Searching the DNA databases revealed an almost 100% identity of both sequences with part of the mitochondrion genome but also 100% homology with the 5′-UTR of the *Wnt-13* gene, which is located at chromosome 1p13 (Katoh et al. 1996).

cDNA Selection

We used three different cDNA libraries (human fetal brain, fetal liver, and adult muscle) to cover a wide spectrum of expressed sequences, and 44 cosmids covering 1.7 Mb of DNA. We analyzed 1920 selected cDNA clones (1.13 clones per screened kilobase) in total. The mean redundancy of cDNA clones screened back to the cDNA filters is 32.76, whereas only 4 of 37 cDNA clones did not map back to a cosmid. Hybridization with ribosomal protein-like sequences revealed 165 positive clones (8.59% of all cDNAs) and 195 clones (10.15%) had a high number of *Alu* repeats. Taking these out, 740 true positive clones (or 38.54% of the

Figure 1 (*See pages 456–458.*) Schematic representation of the transcript map of the 6p21.2–6p21.3 boundary region, immediately centromeric of the MHC. All physical distances are indicated in the top bar as kilobases and thin lines represent 20-kb intervals. Only the relevant minimal number of clones is indicated. Blue cosmid names were used in cDNA selection, underlined in exon trapping, yellow highlighted in isolation of GC rich fragments. PAC/BAC clones are indicated with black lines (prefix dJ for PACs). Clones are drawn to scale. PAC dJ40L2 is extending on either side, as the arrows indicate. Sequenced clones are indicated in red. PAC dJ340B19 is only partially sequenced; therefore, the location of the identified transcripts is tentative. The genomic DNA shows sequenced parts (red lines). (Black boxes) Mapped and predicted exons; (red boxes) mapped ESTs and Unigene contigs; (light blue boxes) predicted CpG islands; (green boxes) genetic markers; (hatched boxes) selected cDNA clones, cDNA contigs, and exons (names with blue background); (open boxes) 3′ and 5′ untranslated regions (all drawn to scale). Linking lines were used to indicate multiple matches of EST and cDNA contigs on the genomic sequence and to denote linked exons, where this is not obvious. Numbers above black boxes indicate the relevant exon number. The names of genes are indicated above the exons; the arrows show the orientation of transcription. The percentage of GC (blue graph lines) and percentage of repeats (red graph lines) are indicated under the sequenced genomic DNA and in the last black stripe the exact position of repeats is indicated as turquoise boxes for the Forward and Reverse strand. Position 0 is the true centromeric end of PAC clone dJ50J22.


cosmid clones
PAC/BAC clones
genomic DNA
ETSL
GS2L
Br140-2
SAPK4
CSBP
GC content (blue) - repeats (red)
Repeats
TULP1
FKBP51
TEF-5
Csa-19
ZNF127-like
HSPPARD
hnRNPE-1
hDEF-6
ZNF76
FBN2sml
B1/KIAA0229
RPL35A-like
TCP-11

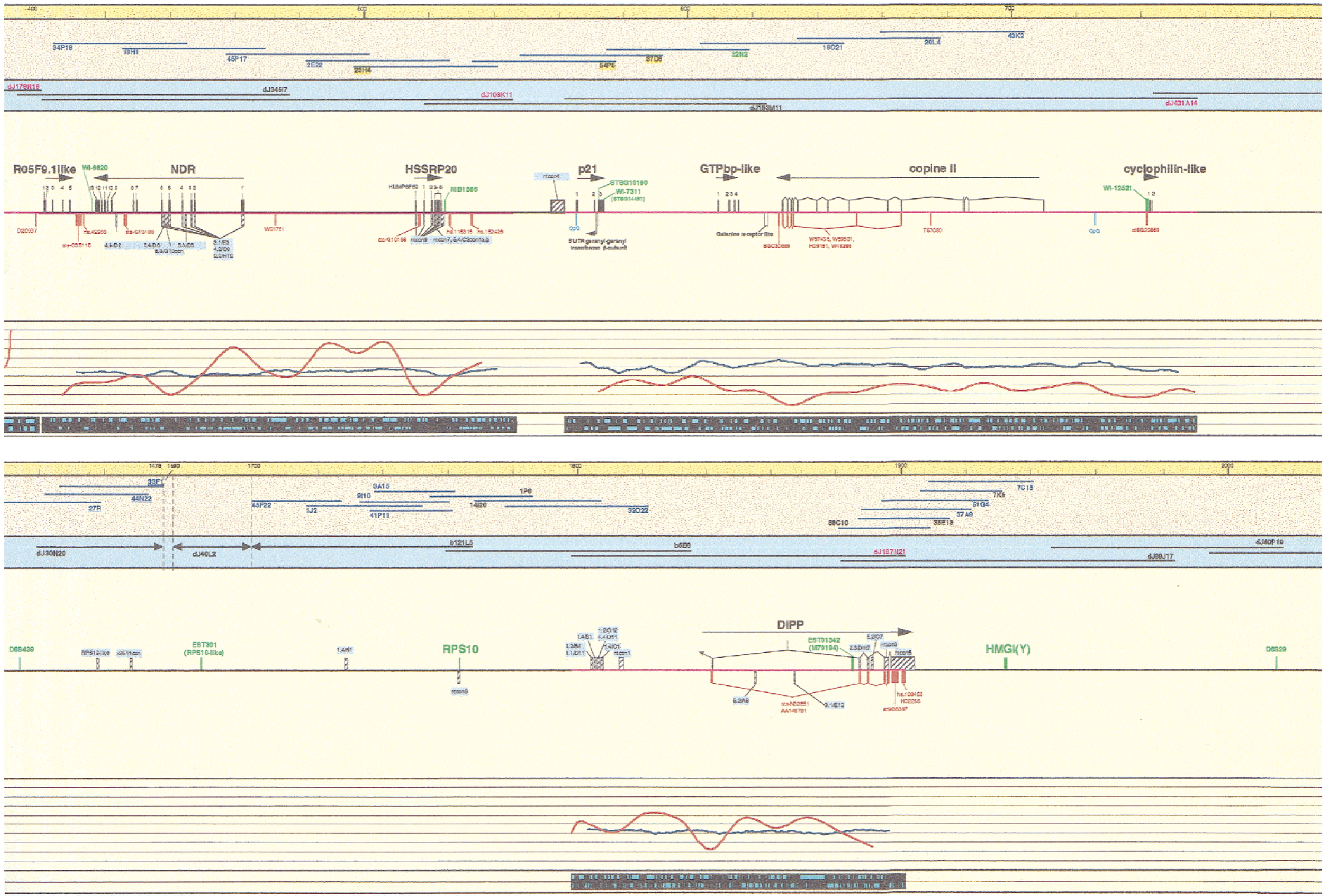


Figure 1 (*See p. 455 for legend.*)

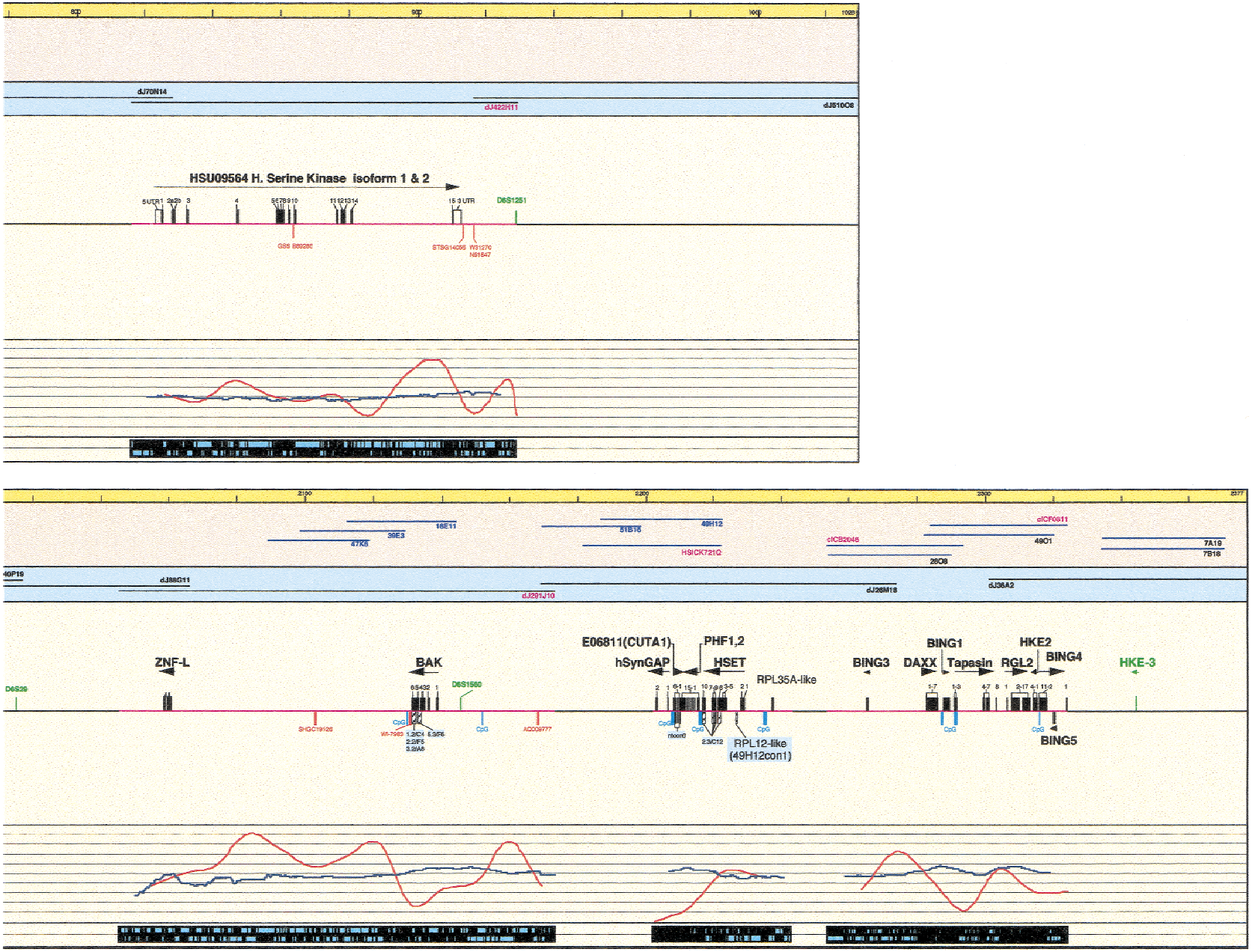


Figure 1 (*See p. 455 for legend.*)

total picked clones) were identified from a 1.7-Mb region (0.43 clones per screened kilobase).

From the identified 740 cDNA clones a subset of 177 was sequenced, resulting in a total of 100 kb of sequence generated from both strands. All the sequenced clones were screened for overlaps and contigs were created. In specific cases the extension toward the 5′ and 3′ end was done using the rapid amplification of cDNA ends (RACE) PCR method as well as electronic walking.

Therefore, 77 unique cDNAs and cDNA contigs were formed (Table 1), of which 55 were confidently positioned on the existing physical map (Fig. 1; Table 1). The expression pattern of each contig was assessed, based on the cDNA library of origin. More than half of all cDNAs and cDNA contigs originate from the fetal brain library (Table 1). In some cases additional information regarding the expression pattern of a particular contig was obtained by screening Northern blots. A comprehensive list showing the clone names, GenBank identification, size, percentage of repeats, BlastN and BlastX results, and homologies to the Unigene, dbest, and htgs databases is shown in Table 1 and their map position is shown in Figure 1.

Genomic Sequencing

Eleven PAC clones and 3 cosmids are currently sequenced and have resulted in a total of ~1.7 Mb of sequence, representing ~70% of the region. Analysis and annotation of this sequence was performed using the comprehensive programs available at the UK HGMP Resource Centre (NIX) and the HPREP suite and viewed in ACEDB. Furthermore, each PAC sequence was screened with the sequences of all the generated clones (cDNAs, exons, GC-rich fragments) and overlaps were established between coding and genomic sequences. The PAC sequences for dJ50J22, dJ524E15, dJ179N16, dJ108K11, dJ431A14, dJ422H11, dJ109F14, and dJ187N21 are complete and fully annotated and are available from the European Molecular Biology Laboratory (EMBL) and GenBank databases (see Table 2 for accession no.), whereas the rest can be seen at the Sanger Centre's web page (http://www.sanger.ac.uk).

Using these analytical tools it was possible to identify the exact position of all selected cDNAs that correspond to the sequenced PAC clones and correlate them to known genes, to position new genes identified by sequence similarity and predicted exons, and locate other transcription units (sequenced ESTs and Unigene contigs) and polymorphic markers. The structure of a large number of genes was elucidated and the fine physical mapping of the region was verified and enhanced (Fig. 1).

The cosmid clones 12K9 (HSICK0912) and 21K7 (HSICK0721) (Janitz et al. 1999) that fully overlap with the cosmid clone 49H12 were sequenced to completion at the Sanger Centre, resulting in 40 kb of genomic fully annotated sequence (presented in Fig. 1). Data from two additional cosmid clones, namely 46B20 (HSICB2046) and 11F8 (HSF0811) (Herberg et al. 1998a,b), were also included for completion purposes.

Construction of a Transcript Map

In total we have assigned 88 transcribed sequences to this region (Table 1). Of these 23 are known genes and 11 show significant similarity to previously described transcripts. This group of 34 transcripts is presented in Table 3 ordered from centromere to telomere and includes structural, expression, and interspecies data.

Identified and Mapped Genes

Analysis of the genomic sequence of clone dJ50J22 revealed two putative coding sequences similar to the ETS transcription factor (four putative exons) and the *GS2*-expressed sequence (four putative exons), which has been assigned to chromosome X (direct submission, HSU03886). The two putative genes are tentatively called *ETSL* (for ETS Like) and *GS2L* (Table 3).

Evidence for the *Br140-2* gene is given by the cDNAs 1.3/C11, 3.1/C1, 4.3/G11, 5.2/C6, 1.2/D10, and ntcon2, which are located on clone dJ524E15. The first four cDNAs show strong homology with the *Br140* gene, (Thompson et al. 1994; Gregorini et al. 1996). The ORFs of 1.3/C11 and 4.3/G11 indicate that the cysteine and histidine residues responsible for the zinc finger formation (Fig. 2A) and the bromodomain region (Fig. 2B) of peregrin are highly conserved in this new gene. Sequence analysis of the genomic sequence confirmed the presence of a coding sequence homologous to *Br140*, tentatively named *Br140-2* (see Fig. 1). The expression profile of the gene, by Northern blot analysis and by listing the tissues of origin of overlapping ESTs, suggests a widely expressed gene (Table 3). Homology of *Br140* with the *AF10* and *AF17* human genes (Prasad et al. 1994) and the *CEZF* gene from *Caenorhabditis elegans* (Chaplin et al. 1995) suggests that they may form a new family of regulatory proteins (Gregorini et al. 1996) and *Br140-2* could be the latest member of this family. The stress-activated protein kinase 2 (*SAPK2*, also mentioned as p38 and p38β; Lee et al. 1994; Crawley et al. 1997) is a mitogen-activated kinase (MAP kinase) that encodes for the cytokine suppressive anti-inflammatory drug (CSAID) binding protein 1 (CSBP1) and its isoform (CSBP2). It was mapped on clone dJ179N16 and verified by cDNA clone 4.1/C12. The alternative exons 9a and 9b are both 79 nucleotides long and are within 4 kb from one another. Further analysis of this PAC revealed the presence of another stress-activated protein kinase centromeric of CSBP, namely *SAPK4* (or p38Δ; Kumar et al. 1997). The *SAPK4* gene is also encoded by 12 exons and transcribed proximally. The *SAPK4* and *SAPK3* genes

Table 1. cDNA Selection Clones, Configs, and Exon Trapping Products

Name	GenBankID	Clones	Origin	Size(bp)	Repeats(%)	BLASTN-nr	dbest	Unigene	htgs	BLASTX	exon prediction
1.2/D10	AF034367	3	FB	700	Ø	Ø	Ø	Ø	dJ524E15	Ø	FEX
ntcon2	AF034173	3	FL	2275	Ø	Bcd	H77459	sts-N66794	dJ524E15	peregrin[3]	FEX/GRAIL/Genemark
4.3/G11	AF034196	1	FL	646	4.8	Br140	Ø	Ø	dJ524E15	peregrin	FEX/GRAIL/Genemark
3.1/C1	AF038248	1	FB	624	Ø	Ø	AA368577	n/a	dJ524E15	peregrin	FEX/GRAIL/Genemark
1.3/C11	AF038247	1	FB	405	Ø	Br140	R67512[1]	Ø	dJ524E15	peregrin	FEX/GRAIL/Genemark
5.2/C6	AF038424	1	FB	559	Ø	Ø	AA368577	Ø	dJ524E15	peregrin	FEX
2.3/B8	AF034182	1	cDNA	402	Ø	KIAA1286	N20665	hs.42318	dJ524E15	Ø	Ø
4.1/C12	AF034185	1	FB	285	Ø	CSBP	W88610	SHGC-9739	dJ179N16	CSBP	FEX/GRAIL/Genemark
4.4/D2	AF037642	1	FB	381	Ø	Ø	Ø	Ø	dJ108K11	Ø	Ø
5.4/D6	AF038242	1	cDNA	780	Ø	Ø	Ø	Ø	dJ108K11	Ø	FEX
5.3/G10con	AF038237	2	FB	654	42.97	Ø	Ø	Ø	dJ108K11	Ø	FEX
5.3/D5	AF038235	1	FL	368	Ø	Ø	Ø	Ø	dJ108K11	Ø	Ø
2.2/H12	AF034187	1	FB	553	Ø	Ndr kinase	Ø	Ø	dJ108K11	Ndr kinase	FEX/GRAIL
3.1/E3	AF034188	1	AM	573	Ø	Ndr kinase	Ø	Ø	dJ108K11	Ndr kinase	FEX/GRAIL
4.2/C8	AF034189	2	FB	850		Ndr kinase	Ø	Ø	dJ108K11	Ndr kinase	FEX/GRAIL
ntcon9	AF038250	7	FL-FB	1287	Ø	Srp20	AA314628	WI-11748	dJ108K11	HSRP20	FEX/GRAIL
ntcon7	AF038240	3	FL-FB	823	Ø	Srp20	AI221131	WI-11748	dJ108K11	HSRP20	FEX/GRAIL
5.4/C2con1a	AF038238	4	FL-FB-AM	976	Ø	Srp20	AA961285	WI-11748	dJ108K11	Ø	FEX/GRAIL
5.4/C2con1b	AF038239	7	FL-FB-AM	1015	Ø	Srp20	AA961285	WI-11748	dJ108K11	Ø	FEX/GRAIL
ntcon4	AF034175	8	FL-FB-AM	1438	Ø	LAP+p21prom	AA026038	Ø	dJ431A14	APBQL	Ø
23H4-A1	AF034178	1	GC-frag.	725	Ø	Mit.gen.	Ø	Ø	Ø	Mit.gen. rRNA 16S	n/a
23H4-B4	AF034179	9	AM	961	Ø	Mit.gen.	Ø	Ø	Ø	Mit.gen. rRNA 16S	n/a
4.3/F4	AF037640	1	FL	490	17.55	Ø	Ø	Ø	dJ109F14	Ø	FEX/GRAIL/Genemark
5.2/C4	AF034191	1	AM	612	Ø	hTEF-5	AA156876	hs.122124	dJ109F14	TEF-5	FEX/GRAIL/Genemark
1.4/F5	AF037629	3	FB	641	5.93	Ø	AA706916	hs.120886	dJ109F14	Ø	Genemark
5.2/D9	AF038234	1	FL	803	8.09	Ø	Ø	Ø	dJ109F14	Ø	FEX/Genemark
1.2/D11	AF034368	1	FB	333	Ø	Ø	Ø	Ø	dJ109F14	Ø	Ø
3.2/H12	AF037634	1	cDNA	708	52.68	Ø	Ø	Ø	dJ109F14	Ø	FEX/Genemark
cNT1	AF038658	1	FB	1008	2.58	Ø	AA983855	hs.128311	dJ329A5	Ø	FEX/GRAIL/Genemark
x5E1	AF034361	1	exon	86	Ø	Ø	Ø	Ø	Ø	Ø	n/a
B1	AF034186	1	U937	1753	Ø	KIAA0229	AA150872	KIAA0229	Ø	ankyrin	n/a
ntB1con	AF038251	11	FB-AM	2374	Ø	KIAA0229	W27616	KIAA0229	Ø	ankyrin	n/a
1.4/B6	AF038657	3	FL-FB	427	Ø	KIAA0229	Ø	KIAA0229	Ø	ankyrin	n/a
x2A6	AF034363	1	exon	107	Ø	KIAA0229	AA117840	KIAA0229	Ø	ankyrin	n/a
x13H9	AF034359	1	exon	75	Ø	KIAA0229	Ø	KIAA0229	Ø	ankyrin	n/a
x8F2	AF034171	1	exon	126	Ø	Ø	Ø	Ø	Ø	Ø	n/a
x1C7	AF034360	1	exon	68	Ø	Ø	Ø	Ø	Ø	Ø	n/a
8wcon1	AF038244	10	FL-FB-AM	855	42.69	RPS12	AA878066	stSG27460	dJ526L6	40S RPS12	n/a
x2E11	AF034362	1	exon	303	Ø	Ø	Ø	Ø	Ø	Ø	n/a
x2E11con	AF037633	2	FL-FB	1149	Ø	Ø	Ø	Ø	Ø	Ø	n/a
1.4/B1	AF034371	1	cDNA	389	58.61	Ø	Ø	Ø	dJ187N21	Ø	Genemark
ntcon8	AF038249	3	AM	685	Ø	RPS10	AI066801	WI-8076	AC005204	40S RPS10	n/a
1.1/D11	AF034364	1	cDNA	703	67.57	Ø	Ø	Ø	dJ187N21	Ø	FEX/GRAIL
1.2/B4	AF034366	1	cDNA	694	25.9	Ø	Ø	Ø	dJ187N21	Ø	FEX/GRAIL
1.2/G12	AF034370	2	FB	461	Ø	Ø	Ø	Ø	dJ187N21	Ø	FEX
1.4/C1	AF034372	1	FB	440	Ø	Ø	Ø	Ø	dJ187N21	Ø	Genemark
4.4/D11	AF034370	1	FB	191	Ø	Ø	Ø	Ø	dJ187N21	Ø	FEX
ntcon1	AF034172	3	FB	1155	10.74	Ø	Ø	Ø	dJ187N21	Ø	Ø
5.2/A8	AF038232	1	FB	438	37.21	Ø	Ø	Ø	dJ187N21	Ø	Ø
3.1/E12	AF034192	1	FB	572	13.29	Ø	AA094545	Ø	dJ187N21	Ø	FEX
5.2/C7	AF038233	1	FB	408	Ø	Ø	Ø	Ø	dJ187N21	Ø	FEX
ntcon3	AF034174	8	FB-AM	1411	4.18	Ø	N75093	sts-N33851	dJ187N21	Ø	FEX
2.3/D117	AF038252	1	AM	132	Ø	Ø	AA029904	hs.44950	dJ187N21	Ø	Genemark
4.4/D6sp6	AF038253	1	FB	393	Ø	Ø	W42653	Ø	dJ187N21	Ø	Genemark
ntcon5	AF034176	35	FL-FB-AM	7232	8.63	Ø	AA187630	stSG6397[4]	dJ187N21	Ø	FEX/GRAIL/Genemark
1.2/C4	AF038245	1	AM	299	Ø	BAK	AA421826	WI-7983	dJ291J10	BAK	FEX/GRAIL/Genemark
2.2/F5	AF038246	1	AM	462	Ø	BAK	AA421826	WI-7983	dJ291J10	BAK	FEX/GRAIL/Genemark
3.2/A8	AF034190	1	AM	817	Ø	BAK	H52672	WI-7983	dJ291J10	BAK	FEX/GRAIL/Genemark
5.3/F5	AF038236	1	FB	689	20.32	Ø	Ø	Ø	dJ291J10	Ø	FEX
49H12con1	AF037643	7	AM-FL	617	Ø	RPL12	AA653789	stSG8782	cICK0721Q[5]	60S RPL12	FEX/GRAIL/Genemark
2.3/C12	AF034183	1	FL	635	5.04	HUMMHCB	AA102137	SHGC-9619	cICK0721Q[5]	HSET	FEX/GRAIL/Genemark
ntcon6	AF034177	9	FL-FB-AM	660	Ø	EQ6811	AA846105	stSG1372	cICK0721Q[5]	CutA	FEX/GRAIL/Genemark
1.1/E12	AF034365	1	FL	737	31.87	Ø	T35793	hs.130860	Ø	Ø	n/a
1.2/D12	AF034369	1	FB	708	Ø	Ø	Ø	Ø	Ø	Ø	n/a
1-D06	AF037644	1	HGMP	76	Ø	KIAA0070	AI096598[1]	hs.3100	Ø	lysyl-tRNA synthetase	n/a
1-D09	AF037645	2	HGMP	278	Ø	Ø	Ø	Ø	Ø	His/Ala rich protein II[2]	n/a
2.1/A11	AF037630	1	FL	489	41.31	Ø	Ø	Ø	Ø	Ø	n/a
2.1/B5	AF034180	1	FB	690	45.22	Ø	W42653[1]	Ø	Ø	Ø	n/a
2.1/D5	AF034181	1	FL	441	10.34	Ø	Ø	Ø	Ø	Ø	n/a
2.3/G2	AF034184	2	FL	421	59.62	Ø	Ø	Ø	Ø	Ø	n/a
2.4/H6	AF037631	1	cDNA	685	62.34	Ø	Ø	Ø	Ø	Ø	n/a
2.4/H7	AF037632	1	cDNA	691	60.7	Ø	Ø	Ø	Ø	Ø	n/a
3.1/F7	AF034193	1	cDNA	676	32.4	Ø	R13537	WI-20496	AC005234	Ø	n/a
3.2/C10	AF034194	1	FL	488	41.39	Ø	Ø	Ø	Ø	Ø	n/a
4.1/H6	AF034195	1	cDNA	167	Ø	Ø	Ø	Ø	Ø	Ø	n/a
4.1/H10	AF037635	1	FB	486	Ø	Ø	Ø	Ø	Ø	Ø	n/a
4.2/F6	AF037636	1	cDNA	361	Ø	HD3	AA456572[1]	WI-12869	Ø	Ø	n/a
4.2/G5	AF037637	1	FL	517	19.54	Ø	Ø	Ø	Ø	Ø	n/a
4.2/G7	AF037638	1	cDNA	374	Ø	Ø	Ø	Ø	Ø	Ø	n/a
4.3/E8	AF037639	1	FB	419	Ø	Ø	H66797[1]	stSG6397	Ø	Ø	n/a
4.4/C3	AF037641	1	FB	437	Ø	Ø	AA321292	Ø	Ø	Ø	n/a
5.1/G6	AF038230	1	cDNA	694	56.59	Ø	Ø	Ø	Ø	Ø	n/a
5.1/G11	AF038229	1	cDNA	694	30.38	Ø	Ø	Ø	Ø	Ø	n/a
5.1/H6	AF038231	1	FB	783	33.95	Ø	Ø	Ø	Ø	Ø	n/a
5.4/D1	AF038241	1	cDNA	170	48.82	Ø	Ø	Ø	Ø	Ø	n/a
5.4/E11	AF038656	1	FB	191	61.78	Ø	Ø	Ø	Ø	Ø	n/a
5.4/G8	AF038243	1	FL	414	Ø	Ø	Ø	Ø	Ø	Ø	n/a
cNT3	AF038659	1	FB	750	Ø	SKI2W	AA074489[1]	WI-16314	HSMHCT8S23	Ski2W helicase	n/a

Table 2. Sequenced Clones

Clone name	Accession no.	Type	Status[a]	Contigs[b]	Sequenced[c]
dJ50J22	Z84484	PAC	F	1	146746
dJ524E15	Z84485	PAC	F	1	80908
dJ179N16	Z95152	PAC	F	1	172048
dJ108K11	Z85986	PAC	F	1	145616
dJ431A14	Z85996	PAC	F	1	195364
dJ422H11	Z99128	PAC	F	1	112984
dJ340B19	AL033519	PAC	U	53	128168
dJ109F14	AL022721	PAC	F	1	170245
dJ329A5	Z97832	PAC	U	6	179338
dJ187N21	Z98036	PAC	F	1	103146
dJ291J10	Z93017	PAC	U	8	134771
cICK721	AL021366	cosmid	F	1	40775
HSICB2046	Z97183	cosmid	F	1	39872
HSF0811	Z97184	cosmid	F	1	40127

Clone name	G+C (%)[d]	Isochore	Repeat (%)[e]	Genes[f]	Gene density
dJ50J22	47.7	H1/H2	29.5	2	0.14
dJ524E15	44.0	H1	44.5	1	0.12
dJ179N16	42.5	L2/H1	37.8	2	0.12
dJ108K11	43.6	L2/H1	46.6	3	0.21
dJ431A14	50.0	H2	25.9	4	0.20
dJ422H11	41.0	L1/L2	45.5	1	0.09
dJ340B19	44.1	H1	50.6	2	0.16
dJ109F14	50.2	H2	35.9	4	0.23
dJ329A5	48.1	H1/H2	32.7	4	0.22
dJ187N21	41.9	L1/L2	45.7	1	0.10
dJ291J10	49.7	H2	51.2	2	0.15
cICK721	50.1	H2	28.3	4	0.98
HSICB2046	49.9	H2	39.5	4	1.00
HSF0811	52.4	H3	30.6	6	1.50

Report of sequenced clones at the Sanger Centre that are part of the physical map of the 6p21.2–6p21.3 boundary region.
[a](F) Finished; (U) unfinished.
[b]Number of formed contigs.
[c]Produced sequence (bp).
[d]Percentage of GC and the corresponding isochore (according to Bernardi 1989).
[e]The percentage of repeats was estimated using RepeatMasker.
[f]The number of identified genes is used to calculate the gene density, expressed in number of genes per 10 kb.

form a distinct set of p38 MAP kinases, which seem to mediate the role of CSBP (Goedert et al. 1997). The sizes of all the exons of the two genes are identical (Fig. 1; Table 3). The two genes have a 61% amino acid identity. However, *SAPK4* occupies a genomic region that is ~one-tenth the genomic region of CSBP, due to its smaller introns. This suggests a duplication event of a single MAP kinase gene.

Telomeric of CSBP lie 5 exons, coding for a putative 310 amino acid protein with significant homology

Details of cDNA selection clones, contigs, and exon trapping products ordered from centromere to telomere. Clones below the black bar possibly fall outside the presented contig. (Clones) Number of clones that form a contig. (Origin) FB, fetal brain; FL, fetal liver; AM, adult muscle; cDNA, from cDNA selection, but origin could not be determined; HGMP, from the gridded HGMP library; exon, from the exon trapping; GC, frag., from the GC-rich isolation experiment. (Repeats) Percentage of repeats, identified by RepeatMasker. (BLASTN-nr) BlastN hits in the nonredundant sequence databases. (dbest) BlastN hits in the EST sequence databases. (htgs) BlastN hits in the high throughput sequence databases. (Unigene) The identified ESTs belong to a Unigene contig. (BLASTX) Best match is shown in each case. (Exon prediction) Predictions of the FEX, GRAIL, and Genemark algorithms from NIX.
[1]For these matches $P >$ le-100. For R67512, $P =$ 7e-20; AI096598, $P =$ 2e-10; W42653, $P =$ 2e-43; AA456; $P >$ 2e-16; H66797, $P =$ 4e-15; AA074489, $P =$ 3e-54.
[2]$P =$ 4.5e-7.
[3]$P =$ 2.1e-5.
[4]Also WI-15639.
[5]Also cICK0921Q.

Table 3. List of Identified and Mapped Genes in the 6p21.2–6p21.3 Boundary Region and Expression Profile

					expression profile																									homology				
Gene	Coding (bp)	Genomic (Kb)	Nr. of exons	Orientation	Brain	Breast	CNS	Colon	Ear	Eye	Foreskin	Heart	Kidney	Liver	Lung	Muscle	Neural	Ovary	Pancreas	Placenta	Prostate	Skin	Small Intestine	Spleen	Testis	Tonsil	Thyroid	Thymus	Uterus	Whole embryo	*M.musculus*	X.*laevis*	D.*melanogaster*	C. *elegans*	S.*cerevisiæ*
ETSL	519	9.8	4	P				+																	+	+					+				
GS2L	660	23.8	4	P	+	+						+	+		+		+			+										+	+		+	+	
Br140-2	2412	28.7	8	P	+		+	+	+			-		+	-		+			+	+				+	+	+				+		+	+	+
*CSBP	1080	83	12	P	+							+		+	+			+		+				+						+	+				
**SAPK4	1095	8.8	12	P	+	+		+			+	+	+	+	+			+		+				+	+				+	+	+				
R05F9.1L	933	8.2	5	D				+		+					+					+									+	+	+		+	+	+
*NDR	1397	48.5	13	P	+	+						+	+			+								+	+	+	+		+	+	+		+	+	
HSRP20	495	8.5	6	D	+	+		+	+	+	+	+	+	+	+	+		+	+	+	+			+	+	+	+	-	+	+	+		+	+	+
p21	494	9.4	3	D	+	+		+	+	+	+	+	+	+				+	+	+	+		+		+				+	+	+				
GTP-BPL	453	6.2	4	D													+														+		+		+
copine II	nd	48+	16+	P	+			+		+		+		+	+	+		+			+			+					+	+	+		+	+	
cyclophilinL	324	0.8	2	D	+	+		+			+	+	+	+	+		+			+	+	+							+	+	+		+	+	+
HSU09564	1967	90.2	15	D	+	+	+	+		+	+			+	+	+				+		+	+		+	+	+			+				+	+
**TULP1	1628	nd	nd	nd	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+
**FKBP51	1371	nd	nd	nd	+			+								+	+			+	+				+						+	+	+	+	+
*hTEF-5	1304	12	11	D	-								+	-	-	+			-	+											+					
**Csa-19	653	3	6	P	+		+	+		+		+		+	+	+	+	+		+	+	+			+		+		+	+	+		+	+	+
ZNF127L	375	3.7	3	P																											+		+		
PPAR-A	1325	85	10	P														+													+	+	+	+	
**hnRNPE-1	1070	2.1	3+	P	+		+	+		+		+	+	+	+	+	+	+	+	+	+	+		+	+		+		+	+	+		+	+	+
hDEF-6	947?	13.3	4	P								+			+		+												+		+				
ZNF76	1547	35.8	12	P	+	+						+	+	-	-	+				+					+	+			+		+	+	+	+	+
FBNS2sml	nd	13.6	7+	P	+																													+	
B1/KIAA0229	3542	14+	4+	D	+	+		+		+		-	+	-	+	+		+	-	-	+		+	+	+	+		+	+		+				
TCP-11	nd	nd	nd	nd																											+				
RPS10	497	0.5	1	nd	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+		+	+	+				+	+			+	+
*DIPP	516	120+	5+	P	+	+			+	+	+	+	+	+	+	+			+	+	+	+	+		+		+		+	+	+				
HMGI(Y)	625	10.1	8	nd	+	+		+				+		+	+	+		+	+	+	+				+	+				+					
ZNF-L	nd	2.3+	3+	P																															
*BAK	635	8.3	6	P		-		+					+		+	+		+			-					+					+				
HSET	1562	12	12	P								+		+																	+	+	+	+	+
PHF1,2	1373/1703	5	15	P	+		+					+					+			+	+				+	+				+					
CUTA1	660	1.5	6	D	+	+				+		+		+		+									+							+			
hSynGAP	nd	3.3+	2+	P			+										+																		

(Coding) The number of coding nucleotides; (Genomic) the extent of each gene over the genomic region. The number of exons is indicated. Where not all exons are known, a "+" sign is used. Alternatively spliced exons are included. Orientation is either proximal (P) or distal (D). The expression profile is indicated with + for EST established tissue expression and a gray box for experimental determination (Northern blot). (−) Negative Northern results. If there is no information, a blank box is left for that particular tissue. The last five columns indicate homology with other species. (*) Experimental expression data presented in this work and from relevant literature; (**) experimental expression data from relevant literature only.

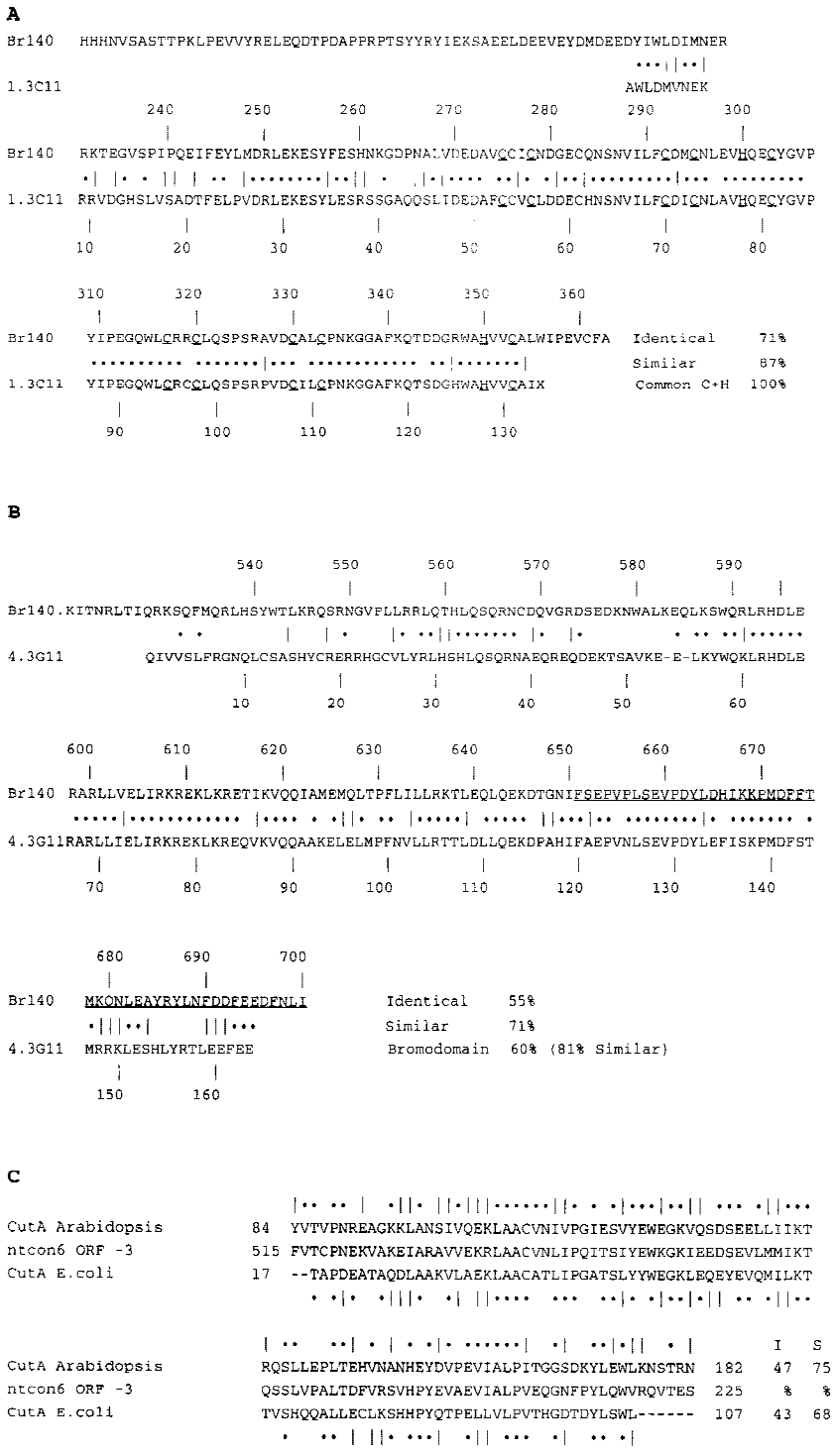


Figure 2 Identification of the *BR140-2* and the human *CUTA1* genes by cDNA selection. (*A*)The overlap between the translated (ORF1) cDNA clone 1.3/C11 and peregrin shows that 1.3/C11 shares the same cysteine (C) and histidine (H) (both underlined) residues with peregrin, which are involved in the formation of the zinc fingers of peregrin. Bullets denote identity and vertical lines similarity (also for *B* and *C*). (*B*) The overlap between the translated (ORF2) cDNA clone 4.3/G11 and peregrin shows that clone 4.3/G11 contains a domain highly similar to the bromodomain of peregrin. (*C*) The overlap between the translated (ORF3) of ntcon6 and two CutA protein sequences for *Escherichia coli* and *Arabidopsis thaliana* are shown. The percentage of identity and similarity between each sequence is shown under I and S, respectively, at the *bottom right*.

to a sequenced coding segment of a *C. elegans* cosmid (clone R05F9, U41533) and weak similarity to the SRP40 gene. This putative gene was designated *R05F9.1-like* (Fig. 1; Table 3).

The Ndr protein kinase, a highly conserved protein mainly localized in the nucleus (Millward et al. 1995), and the *HSRP20* gene were positioned accurately with the help of cDNA clones and contigs (Table 1).

The detailed genomic organization of the murine *SRp20* was described recently (Jumaa et al. 1997) and it was shown that alternative splicing occurs on exon 4 (Jumaa and Nielsen 1997). The alternative splicing is regulated by the antagonistic action of SRp20 in favor of use of exon 4 and of ASF/SF2 in skipping exon 4. The necessary sequence elements responsible for this regulatory process are within and immediately flanking exon 4 (Jumaa and Nielsen 1997). The sequence of the human HSRP20, as deposited in the databases, does not include a segment similar to the murine exon 4. However, the cDNA data (ntcon7 and ntcon9) and the genomic sequence of dJ108K11 show that not only is the same exon present in humans, but the flanking intronic sequences have also been conserved between the two species. This may well indicate that the human HSRP20 is regulated in a similar fashion to its murine counterpart.

The *p21* gene is a cyclin-dependent kinase inhibitor (el-Deiry et al. 1993; Harper et al. 1993). Two markers are part of the 3′-UTR of *p21* (WI-7311 and stSG10190). Interestingly, part of the complementary sequence of intron 2 is identical to the 5′-UTR of the geranyl-geranyl-transferase type-I (GGTase-I) β subunit (F.L. Zhang et al. 1994). Further telomeric, four putative exons were identified that encode for a protein that is similar to the GTP-binding domain of the Ras-related RAB proteins (Fig. 1; Table 3). This putative gene is designated *GTP-BPL* (for GTP-binding protein-like) and has a distal orientation

of transcription. Less than 10 kb from the 3′UTR of the *GTP-BPL* is a G protein-coupled receptor-like sequence, which has a weak similarity to the galanin-1 receptor gene (Habert-Ortoli et al. 1994).

A possible new member of the copines is present on PAC dJ431A14. The copines are a novel class of C2 domain-containing, calcium-dependent, phospholipid-binding proteins that are highly conserved and may function in membrane trafficking (Creutz et al. 1998). A number of ESTs that map in the same region increase the possibility that this is a transcribed and functional gene.

A cyclophilin-like gene has been identified by sequence analysis. The Unigene contig WI-12521 maps on the same area and increases the possibility that this is a functional cyclophilin-like gene. Genomic sequence from human *Xq22* suggests that the cyclophilin gene is located there and Haendler and Hofer (1990) note that there is a number of processed cyclophilin pseudogenes. However, all the sequenced pseudogenes lack introns, in contrast to the described transcript. A putative CpG island lies 15 kb upstream of exon 1, as expected from a housekeeping gene, supporting the evidence that this is a functional gene.

The HSU09564 serine kinase extends over at least 90 kb of genomic sequence. There are two alternative exons 2, designated 2a and 2b, which account for the isoforms 1 and 2 of the gene. It seems that HSU09564 plays a central role in the regulation of splicing in the nucleus by controlling the intranuclear distribution of SR splicing factors in interphase cells and therefore, in the reorganization of nuclear speckles during mitosis (Gui et al. 1994).

Although the PAC clone dJ340B19 is not fully sequenced, BlastN and BlastX analyses show that the *TULP1* gene is located there. The *TULP1* gene was cloned together with other members of the *tub* family of proteins, which have a highly conserved carboxyl terminus (North et al. 1997). *TULP1* mutations were found in nonsyndromic recessive retinitis pigmentosa patients (Hagstrom et al. 1998) and in all affected individuals from two arRP14-linked kindreds (Banerjee et al. 1998), indicating that the *TULP1* protein is essential for the normal physiology of the retina and responsible for the arRP14. Apart from the *TULP1* gene, the *FKBP51* gene and an *RPS15A-like* gene are also located in this genomic segment. The *FKBP51* gene is a highly conserved protein that has peptidylprolyl isomerase activity (Nair et al. 1997).

The *hTEF-5* gene (Jacquemin et al. 1996) is located at the centromeric end of clone dJ109F14 (Fig. 1). The 5.2/C4 clone is practically identical to the *hTEF-5* gene and maps on the cosmid 4B4. Comparison of the genomic sequence of the overlapping PAC clone dJ109F14 with the 5.2/C4 and the *hTEF-5* sequences revealed that the *hTEF-5* gene is encoded by 11 exons and the cDNA clone covers exons 8, 9, and 10. Analysis of the genomic sequence reveals that there are two alternative exons, designated 3a and 3b (Fig. 1), which account for two isoforms of the *hTEF-5* gene (Fig. 1; Table 3). The *hTEF-5* is a transcription factor that contains the TEA/ATTS DNA-binding domain and plays a role in myogenesis, cardiogenesis, CNS development, and organogenesis. The cDNA clone 5.2/C4 originates from the adult muscle library, in agreement with the reported expression pattern of the gene (Jacquemin et al. 1996).

In a tail-to-tail orientation with *hTEF-5* is the *Csa-19* gene, which (Fig. 1; Table 3) was cloned by selecting cDNA clones whose expression is down-regulated in the thymus by cyclosporin-A (Fisicaro et al. 1995). The *Csa-19* gene is highly conserved between mice and humans and expressed in many adult and fetal issues but predominantly in lymphoid organs (Table 3; Fisicaro et al. 1995). Further telomeric, a putative gene was identified. It has a weak similarity to ZNF127 (direct submission, U19107) and was tentatively designated *ZNF127L* (Fig. 1; Table 3).

The human peroxisome proliferator-activated receptor-Δ (*PPAR*–Δ, or *PPAR*-β; (Schmidt et al. 1992) is conserved in mouse and rat and was mapped previously by analysis of a somatic cell hybrid panel on chromosome 6p21.2–p21.1 (Yoshikawa et al. 1996). There are two CpG islands, one upstream of the 5′-UTR of the gene and another one in intron 6 (Fig. 1). Interestingly, a number of ESTs and 3 cDNA clones (5.2/D9, 1.2/D11, and 3.2/H12) have been mapped within intron 2 (Fig. 1). It is still unclear whether they are part of a yet unidentified gene within intron 2.

The heterogeneous ribonucleoparticle *hnRNP-1* (direct submission, X78137) was identified in the centromeric end of the clone and at least 3 putative exons were predicted (Fig. 1; Table 3). A putative CpG island (see below) is present immediately upstream exon 1 (Fig. 1), suggesting this is the true functional gene.

The human *DEF-6* gene was also identified (Hotfilder et al. 1999) and is closely related but not identical to the recently described B-cell-specific switch recombinase SWAP-70. The *ZNF76* gene is encoded by 12 exons and its 5′-UTR coincides with a mapped *Bss*HII site on the physical map (Tripodis et al. 1998; Fig. 1). Between the 5′-UTR and coding exon 1 there is a *RPL35-like* sequence. Further telomeric several putative exons were identified that also have a proximal orientation of transcription and encode for a protein similar to fibrillin-2 (Fig. 1; H. Zhang et al. 1994). The novel putative gene was tentatively called *FBN2sml* (for similar).

The *KIAA0229* transcript was identified as follows: the B1 cDNA, isolated by direct screening of the U937 cDNA library, was used to screen the cDNA clones from the cDNA selection experiment. The cDNA clones posi-

tive with B1 formed the contig B1con, which extends further upstream from B1, starting at residue 3918 of the *KIAA0229* mRNA. In addition to that, exon x13H9 identified 4 overlapping cDNA clones. They are all contained within the 419-bp long 1.4/B6 cDNA clone. Alignment of the sequence of clone 1.4/B6 with the *KIAA0229* sequence shows identity for the first 118 bp of 1.4/B6 between residues 2109 and 2218 just 10 bp upstream of exon x13H9.

The *KIAA0229* gene has its 3′-UTR at least 15 kb away from *TCP-11* and extends over an area of at least 27 kb toward *ZNF76*, which is 170 kb further centromeric (Fig. 1). The B1 cDNA was used to screen a series of Northern blots (Fig. 3), showing that it is predominantly expressed in testis as a 7.4-kb transcript, whereas transcripts with sizes of 9.5 and 4.5 kb were also detected. In the mouse a 4.5-kb signal was detected in spleen, testis, heart, lung, liver, and skeletal muscle (Fig. 3). Screening of total genomic DNA digests with B1 showed that it is present only once in the genome (not shown). Therefore, the B1 cDNA is part of a new gene with potential alternative splicing, resulting in mRNA species of sizes ranging from 9.5 to 4.5 kb and with predominant expression in human adult testis. The *KIAA0229* sequence deposited in the databases is 6335 bp long, which probably coincides with the 6-kb signals detected at the Northern blots. The predicted protein encoded by the *KIAA0229* mRNA shows similarity to the human ankyrin protein.

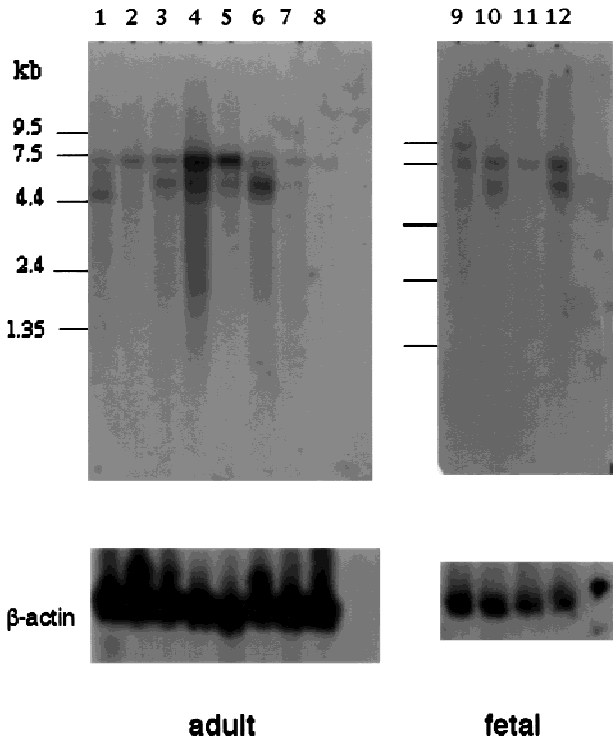


Figure 3 Expression pattern of B1 in a series of multiple tissue Northern blots from human adult (Clontech MTNII) and fetal tissues (Clontech). The tissues present in these blots are as follows: (Lane *1*) spleen; (land *2*) thymus; (land *3*) prostate; (lane *4*) testis; (lane *5*) ovary; (lane *6*) small intestine; (lane *7*) colon; (lane *8*) peripheral blood leukocytes; (lane *9*) fetal brain; (lane *10* fetal lung; (lane *11*) fetal liver; (lane *12*) fetal kidney. β-Actin control is shown underneath. Sizes are indicated on the side of each blot.

The *TCP-11* and *RPS10* (corresponding to ntcon8) genes and the *RPS10* pseudogene (*EST301*) were previously mapped (Tripodis et al. 1998). Further toward the telomere the genomic sequence of dJ187N21 revealed the diphosphoinositol polyphosphate phosphohydrolase (*DIPP*) gene (Safrany et al. 1998) and the cDNA contigs ntcont3 and ntcon5 are part of its 3′-UTR.

The *BAK* gene was identified and verified by the presence of cDNA clones. The detailed genomic structure of *BAK* is described elsewhere (Herberg et al. 1998c). The *HMGI(Y)* gene maps between *DIPP* and *BAK* (Tripodis et al. 1998). A zinc finger putative gene, identified by similarity searches and exon prediction, is situated centromeric to *BAK* (Fig. 1). Although it was not possible to determine the total number of exons, the transcript is proximally orientated. It is a C2H2 type of zinc finger protein and is marked as *ZNFL* (Table 3).

The *HSET* gene was previously mapped in the region as *HTCTEX-7* (Tripodis et al. 1998). It is encoded by 10 exons and was identified with cDNA 2.3/C12. A putative CpG island is present in the 5′-UTR, but this is close to an *RPL35a* pseudogene. Another *RPL12-like* sequence was identified in intron 2 (49H12con1).

The cDNA contig ntcon6 is encoded by 6 exons and is homologous to EO6811, which is a cDNA encoding a polypeptide secreted from human glioblastoma cells. The encoded peptide showed similarity to the *Arabidopsis thaliana* CutA protein (Fong et al. 1995; Fig. 2C), which is conserved across species from different taxa. In *Escherichia coli* the CutA protein is presumably involved in copper tolerance and also affects tolerance to zinc, cadmium, nickel, and cobalt salts (Fong et al. 1995). A putative CpG island was identified on the 5′ end of ntcon6. The *PHF1* and *PHF2* genes are located between *HSET* and ntcon6 with a CpG island at the 5′ end (see Fig. 1; Janitz et al. 1999). The *PHF1* gene is similar to the *Polycomb* of *Drosophila melanogaster* and has been previously mapped on chromosome 6p21.3 by in situ fluorescence hybridization (Coulson et al. 1998).

Further centromeric two exons were identified that have homology with the synaptic Ras–GTPase activating protein p135SynGA, identified in the rat (Chen et al. 1998). The identified exons belong to the human homolog, designated *hSynGAP*. It is centromerically transcribed and as it has a head-to-head orientation with ntcon6 (Fig. 1), it is possible that the two genes share 5′-UTRs. The available genomic sequence does not contain the entire coding sequence of the *hSynGAP*; therefore it was not possible to determine the

total number of exons and the genomic extent of this gene.

The two cosmid clones cICB2046 and cICF0811 overlap with cosmids and a PAC clone (dJ36A2) previously mapped in the centromeric end of the MHC (Tripodis et al. 1998). The two clones overlap for ~10 kb and together result in a 69.9-kb contiguous genomic segment. The detailed analysis and gene content of these two clones has been described elsewhere (Herberg et al. 1998a,b). In brief, the following gene order was established in this region: *BING3*, *DAXX* (*BING2*), *BING1*, *Tapasin*, *RGL2*, *HKE2*, *BING4*, and *BING5*. The *HKE3* (*RPS18*) gene is present on PAC clone dJ36A2 (Tripodis et al. 1998), telomeric of *BING5* (Fig. 1).

Putative Transcripts

Four ESTs/EST clusters were mapped on the dJ179N16 (Fig. 1). One of them is Unigene contig stSG23312, which includes ESTs derived from breast, heart, lung, muscle, parathyroid, placenta, and whole embryo. This contig is ~4 kb centromeric of the 3′-UTR of the *SAPK4* gene. Because no poly(A) signal was identified in between, this EST contig [containing a poly(A) signal] may be the true 3′-UTR of *SAPK4*.

At the 5′ end of the p21 gene outside the dJ431A14 sequence lies the cDNA contig ntcon4, which is highly similar to the bovine leucine aminopeptidase (LAP), a highly conserved protein (Taylor et al. 1984). ntcon4 is also highly similar to the promoter sequence of p21 (91% similarity with 1% gaps over the entire length of the nucleotide sequence of ntcon4) as shown from BLASTN searches. The ESTs identified belong to the Unigene contig hs.106750 (AOO6X39), which has been mapped on chromosome 4, between markers D4S1601 and D4S419. These results suggest that the identified sequence is represented at least twice in the human genome but no conclusion could be drawn from this work as to which is the functional transcript or whether both locations harbor functional genes.

Between *Csa-19* and *hTEF-5* is located the cDNA clone 1.4/F5. Sequence similarity searches revealed no identity of the clone with any known genes, proteins, or ESTs, even after obtaining additional sequence information using the RACE PCR method. Genomic DNA hybridizations revealed that it is single copy and the clone was used to screen a total RNA Northern blot. Three signals with approximate sizes of 0.6, 0.8, and 1.8 kb were seen in all tissues with varied intensity, the strongest being in kidney and skeletal muscle. Two strong signals from transcripts of ~9 kb and 4 kb were detected in spleen and one strong signal of ~4 kb in duodenum (data not shown).

The cDNA clone cNT1 was mapped back to cosmid 36, which overlaps with dJ329A5 and was found on the genomic sequence. Several ESTs and Unigene contigs flank the cNT1 sequence (Fig. 1), suggesting a yet undetermined gene.

The Unigene contig WI-30588 was mapped upstream of *FBN2sml* and it represents either a 5′-UTR of this gene or a separate transcript (Fig. 1).

Potential transcripts just telomeric of TCP-11 are represented by exon x9C3, exon x2E11, and the cDNAs 1.4/B1 and 4.4/D11 (see Fig. 1 and for clone details, Table 1). These cDNAs and exons did not show homology to any known expressed sequences in the databases.

The proximal end of the dJ187N21 clone overlaps with the cosmids 32O22 and 14I20 as revealed by the presence of a group of cDNAs (1.1/D11,1.2/B4, 1.4/B1, 1.2G12, 4.4/D11, 1.4/C1) and ntcon1.

Within the DIPP introns are the cDNAs 5.2/A8, 3.1/E12, 2.3/D1, and 5.2/C7 (Fig. 1), but whether they form part of another expressed sequence is not known. The 5′ end of DIPP is outside the dJ187N21 sequence.

Comparison Between Experimental and Electronic Gene Identification

All known genes that have been identified by cDNA selection or exon trapping and are present within the sequenced segments were identified by the sequence analysis tools through a combination of BlastN hits to the nonredundant nucleic acid database and BlastX hits to protein databases (see Table 1). The exon prediction programs also correctly identified exons at these positions (Table 1). The 28 cDNAs that did match the available sequence but showed no similarities to sequences in the nonredundant database can be split into two groups: one group of 23, where at least one prediction program indicated the presence of an exon and thus supporting the evidence for the presence of a genuine transcribed sequence, and five cDNAs without any overlapping exon prediction (Table 1). Most of these five cDNAs localized within introns of known genes (for example, 4.4/D2 and 5.3/D5 within the Ndr locus; see Table 1; Fig. 1) and are most likely cDNA library artifacts.

Analysis of Nucleic Acid Content of the Sequenced Segments

The percentage of GC for all sequenced contigs was calculated for every 10 kb and plotted in overlapping windows of 100 bp (Fig. 1). The size of most contigs coincides with that of the sequenced clones and, in turn, this is approximately equal to a standard isochore unit (Bernardi 1989). The isochore content of each sequenced clone is shown in Table 2. Identification of CpG islands was visualized based on outputs provided by the NIX program and the putative CpG islands were indicated in Figure 1. Putative CpG islands were identified in the 5′-UTRs of most genes (*Br140-2*, *SAPK4*, *CSBP*, *p21*, *TEF-5*, *Csa-19 HSPPARD*, *hnRNPE-1*, *hDEF-6*,

ZNF76, *PHF1*, and *EO68011*), or between two genes that have a head-to-head orientation but yet undefined true 5′-UTRs (between copine II and cyclophilin-like, between *FBN2sml* and *B1/KIAA0229*, and between *RPL35A-like* and *HSET*). It is possible that the true 5′-UTRs of these genes are located near the identified CpG islands.

Using the Repeat Masker (of NIX) as output files, the percentage of all repeats was calculated for every sequenced 10 kb and plotted alongside the sequenced data and the percentage of GC content (Fig. 1). The exact location of the identified repeats is also provided in Figure 1. As with the percentage of GC, the average repeats content per sequenced contig was calculated (Table 2).

For every sequenced contig, the total number of identified genes was calculated (Table 2). The calculated averages of percent GC, repeats, and number of genes were used to estimate positive and negative correlations between these features. There is a significant strong positive correlation between GC content and gene number (Spearman's $\rho = 0.838$ at the 0.0001 significance level) and density (Spearman's $\rho = 0.821$ at the 0.0003 significance level), as expected. A relatively weaker, still significant negative correlation, is also present between percentage of repeats and both GC content (Spearman's $\rho = -0.587$ at the 0.027 significance level) and gene number (Spearman's $\rho = -0.600$ at the 0.023 significance level).

These correlations can be visually followed in much greater detail in Figure 1. Local increase of percent GC is observed in almost all cases where there are exons, whereas overall percentage of GC content is higher in the most gene-rich regions. The presence of repeats is less predictive, in terms of gene identification, but marked increases of repeats are observed in a number of large introns (such as intron 8 of *CSBP*, intron 1 of *NDR*, and intron 14 of *HSU09564*) and in a number of intergenic regions (such as, between *Br140-2* and *SAPK4*; *CSBP* and *R05F9.1-like*; *NDR* and *HSRP20*; *HSPPARD* and *hnRNPE-1*; and between *BING3* and *DAXX*). In all these cases this is coupled with a drop in percentage of GC.

DISCUSSION

We present a transcription map of the 6p21.2–6p21.3 boundary region that spans ~2.5 Mb. The region starts from the G-dark band 6p21.2 near the genetic marker D6S291, and ends immediately centromeric to the MHC, providing a link with the up-to-date extended MHC class II region. In total, 88 putative coding sequences have been isolated and characterised (Table 1) and 1.7 Mb of genomic sequence generated and analyzed (Table 2). This resulted in the identification of at least 35 genes, of which 23 have been described previously (Table 3). For most of these genes, the detailed genomic organization, the orientation of transcription, and transcription profiles have been determined (Fig. 1; Table 3). A number of pseudogenes was also identified and annotated on the available sequence data. At least 63 STSs, ESTs, and EST contigs have been also positioned on the existing physical map (Fig. 1).

The identification of transcripts was mainly based on cDNA selection, and additional methods (exon trapping, direct screening of cDNA libraries, and isolation of GC-rich fragments) were implemented only in selected areas. Direct comparison of the various techniques is, therefore, only valid for these areas. Direct screening resulted in the identification of one gene (*B1/KIAA0229*), which was also identified by cDNA selection and exon trapping. There was some overlap between exon trapping and cDNA selection, which provided further evidence that the trapped exons are part of functional genes. However, because more than half of the exon trapped region is yet to be sequenced, it is difficult to assess the true success rate of this technique.

The isolation of GC-rich fragments identified the 5′-UTR of the Wnt-13 gene. The Wnt genes are implicated in murine mammary malignancies and also in human malignancies (Katoh et al. 1996). It is interesting that Wnt-13 is within a segment of chromosome 1 that seems to be involved in a duplication event within the same chromosome. The corresponding locus on chromosome 1q23–q25 has been implicated in a triplication event between chromosomes 1, 6p21.3, and 9 (Katsanis et al. 1996). Therefore, it is possible that another Wnt gene or pseudogene is located at chromosome 9q34.

It seems that the Tm of the retained fragments is the most important selective criterion, together with the size. However, the use of this technique as a tool for scanning unsequenced cloned DNA is probably superseded by the plethora of already sequenced and mapped ESTs that present undoubtedly the best starting point for construction of transcript maps in most regions of the human genome.

The cDNA selection proved a very powerful technique in isolating and characterizing many genes in the region and supporting the exon trapping results. In this work, the cDNA libraries used clearly had a large number of genomic segments derived from unspliced RNA (i.e., ntcon7 was exclusively part of the intronic sequence of HSRP20, and many other cDNAs had parts of coding and parts of intronic sequence). This makes the subsequent effort for mapping and identifying a complete gene rather difficult. The intronic segments can have a number of repeats that may confuse the mapping by hybridization. At the same time, for every such cDNA the yield of translated sequence is lower, which in turn increases the amount of additional work

needed for identifying a complete gene sequence. The quality of the cDNA libraries and the size selection (full mRNA transcripts) are critical factors for the successful cloning of genes. Increasing the enrichment steps and selecting from a variety of cDNA libraries significantly increases the success ratio of this technique.

Our findings are in agreement with previous studies, where no single method of transcript identification is sufficient to identify all the transcription units within a genomic segment (Yaspo et al. 1995).

The obvious complement to these techniques is the genomic sequencing and the application of similarity searches and exon prediction programs. We have used this approach with significant success and the results suggest that we reached 100% recovery of transcribed units in most of the completely sequenced clones. The available genomic sequence of certain PAC clones significantly facilitated, not only the mapping of cDNA clones, but also the identification of the entire transcribed sequence of putative genes, and with the help of exon prediction programs, the genomic organization of these genes. Approximately half (28/48) of the sequences that resulted from cDNA selection and exon trapping and matched the available genomic sequence did not show any similarity to genes or protein sequences deposited in public databases (Table 1). The presence of transcripts in these positions was strongly supported by EST matches (in nine cases) and exon prediction results (in 23 cases; Table 1). As a result the generated material provided valuable experimental evidence for the presence of transcribed sequences at 19 positions where the sequence analysis tools had predicted exons, thus indicating that the combination of the two approaches is still important for the construction of complete transcript maps.

The identified genes belong to different functional groups. These included a number of nuclear and splicing factors (Ndr kinase, *HSU09564*, *HSRP20*. *R05F9.1L*); cell cycle, DNA packaging, and apoptosis related [*p21*, HMGI(Y), BAK, GTP-BPL]; immune response (*CSBP*, *SAPK4*, *FKBP51*); transcription activators and zinc finger-containing genes (*hTEF-5*, *ZNF76*, *Br140-2*, *ETSL*, *ZNF127L*, *ZNFL*); development and embryogenesis related (*Csa-19*, *B1/KIAA0229*); cell signaling (*DIPP*), structural (*HSET*), and other genes (*TULP1*, *HSPPAR-D*, *hDEF-6*, *EO6811/ntcon6*, *copine II*, *cyclophilinL*, *hnRNP-1*, *PHD1* and *PHD2*, *hSynGAP*), as well as a number of RP genes and pseudogenes (*RPS10*, *RPS18*, *RPS12-like*, *RPL12-like*, *RPL35-like*, and *35A-like*).

A possible duplication event can be traced in the case of the *CSBP* and *SAPK4* genes. It is also interesting that functionally closely related genes reside in the same general area, such as *HSRP20* and *HSU09564*. Because it has been shown that addition of *HSU09564* to a splicing reaction in vitro results in dose-dependent inhibition of pre-mRNA splicing, it would be interesting to see how the activity of the *HSRP20* will be affected by the two alternatively spliced forms of *HSU09564*.

The gene order of this segment is of great interest, as this region is syntenic to mouse chromosome 17. The colinearity of genes between the two species has been demonstrated in the past for the region immediately centromeric of the MHC (Kikuti et al. 1997; Tripodis et al. 1998). The presented data should help in establishing the extent of the colinearity. The *BAK* gene, *TCP-11*, *ZNF76*, *HSRP20*, and *hTEF-5* have murine homologs that have already been mapped in the mouse syntenic segment. In our previous work (Tripodis et al. 1998), we presented a human–mouse comparative map based on a limited number of genes in the region. The information presented in this paper provides a much more detailed gene content for the human segment. In this way, the transcription map of the mouse syntenic region should be greatly facilitated and, upon completion, it will be possible to define the exact blocks of genes that have been conserved between the two species.

The localization of *BAK* between *HMGI(Y)* and *HSET* raises questions over the possible role that *BAK* may have in the phenotypic characteristics associated with inversion of this region in the mouse *t* complex. The mouse *BAK* gene was recently characterized and maps on mouse chromosome 17B (Ulrich et al. 1997). It has 62% similarity with the human *BAK* and is also encoded by 6 exons (Ulrich et al. 1997). The regional assignment of *BAK* in the borders of a mouse chromosomal segment that is inverted in the *t* complex presents two possible scenarios involving *BAK*. In the first scenario, the *BAK* gene is within the inverted segment and in the second, it lies exactly on the border of the noninverted segment, that also contains *tctex-7A* and the mouse *H-2* complex. In both scenarios, when the inversion takes place it is possible that the transcriptional control of *BAK* changes because of new neighboring elements (that may be another gene, a promoter sequence, or a transcription activator factor). Because *BAK* belongs to the Bcl family of genes and is involved in apoptosis, it is tempting to assume that changes in the expression regulation of *BAK* may be involved in some of the phenomena leading to the embryonic developmental mutations and transmission ratio distortion associated with the *t* complex.

Analysis of the transcription profile of the identified genes and cDNAs was greatly facilitated by screening the EST databases and listing the libraries of origin of overlapping ESTs (shown as crosses in Table 3). The overlap between the experimentally established expression profiles (by Northern blot analysis, shown as gray boxes in Table 3) and the EST-established profiles is significantly high. The EST-derived expression profiles were in good agreement with experimentally ob-

tained data and in many cases they were complementary. Although it has been suggested that EST profiling can be used as an alternative cheap way to determine the expression of incompletely characterized genes, the accuracy of the data is sometimes questionable. For this reason we suggest that when considering the EST data one should pay attention to the total number of clones sequenced from that particular library, as a rough indication of the frequency of this clone/gene in the library. If a very high number of ESTs has been sequenced and there are only a few identical (or overlapping) clones present, then it is possible that the gene is expressed at very low levels in the tissue of origin. If, however, it is a frequent transcript and the Northern results are negative, then two possibilities arise: one is that the EST profile is derived from an expressed homolog and the other that the gene in question has a very defined time-window of expression, which is represented in only one of the two tissues of origin used in the Northern blot and cDNA library construction, respectively.

Calculation of the GC content of the sequenced clones and the percentage of repeats (Table 2) and comparison with the number of genes identified in each clone led to some interesting observations. When plotting the percentage of GC of every clone along the genomic DNA data, the colinearity between GC richness and high gene density becomes obvious (Fig. 1). Close inspection of the percent GC graph and the location of exons reveals how the local percentage of GC increases in the coding areas, irrespective of the overall percent GC content of the region. Calculation of correlation coefficients between percentage of GC, percentage of repeats, and gene density support the visual observations, provided by the graphs in Figure 1. As expected, a strong positive correlation exists between gene density and number and percentage of GC, whereas a negative correlation exists between percent repeats and percent GC and gene density. The marked drop in GC content and gene density observed in clone dJ187N21 (Fig. 1), and the high number of repeats in dJ291J10 may indicate the physical boundary of the Giemsa-positive 6p21.2 chromosomal band and the Giemsa-negative part of the 6p21.3. If the overall average GC content of the sequenced clones centromeric and telomeric of dJ187N21 is calculated, there is a ~5% difference between the two regions (the centromeric average is 45.9% and the telomeric average is 50.2%). This indicates an H2 isochore telomeric of dJ187N21 and an H1 isochore centromeric of dJ187N21. In this H1 region several L1 and L2 regions are present, as well as a number of genes with very long introns and repeat-rich regions, supporting the suggestion of a transition to a Giemsa dark band.

The identification of putative CpG islands is in accordance with the location of most 5′ UTRs of genes and may even be used as a starting search point in those cases where the 5′ UTRs of genes are not yet known.

The detailed map of the gene content of this chromosomal segment provides a number of candidate genes for playing a role in biological phenomena that have been associated with the MHC and the mouse *t* complex and include spermatogenesis, development, embryogenesis, and neoplasia. The data also provide useful tools for synteny studies between mice and humans, for genome structure analysis, gene density comparisons, and nucleotide composition of different isochores and Giemsa light and Giemsa dark bands. Furthermore, these data suggest that the centromeric end of the MHC may extend further than previously thought. Equally interesting, perhaps, will be the ability to screen the genomic sequence for yet unknown elements that may play a role in the initiation of replication of DNA, and folding of DNA and possibly elements that are involved in long-range interactions on the transcription level.

METHODS

Direct cDNA Library Screening

The U937 cDNA library was plated in high density onto four duplicate filters, to a total final number of approximately 1 million clones. One set of filters was prepared for hybridization, and the other served as the master plate. The library was screened with *Bss*HII containing *Eco*RI fragments from cosmids 36, 25, 4R, 4N, 15a, and 27R (Fig. 1). Single clones were identified by hybridization on *Xho*I Southern blots of the individual clones with the original probes.

Exon Trapping

Exon trapping was performed using the pSPL3b vector (Church et al. 1994) and as described elsewhere (Yaspo et al. 1995). Twelve cosmids forming a minimum tile path covering the region were pooled in a single exon trapping experiment. The pooled cosmids were 3N2, 20L6, 20G1, 36, 25, 4R, 4N, 36P16, 17E12, 27R, 44N22, and 33F1 (Fig. 1). From a total of 1344 isolated exon trap clones (3.3 clones per exon trapped genomic kilobase), 591 (43.9% of the picked clones) had an insert, of which 131 (22.1% of the insert containing clones) were positive with the HIV intronic sequence of the pSPL3b vector and 55 (9.3% of the insert containing clones) had an Alu repeat insert. Thus, the number of true positive clones was 405 (or 30.1% of the total picked clones).

Preferential Isolation of GC-Rich DNA Restriction Fragments

The isolation of GC-rich DNA restriction fragments (Shiraishi et al. 1995) involved the sequential digestion of 600 ng of cosmid DNA with four enzymes, *Mse*I (TTAA), *Tsp*509I (AATT), *Nla*III (CATG), and *Bfa*I (CTAG). The digests were subsequently run in denaturing gradient gel electrophoresis (Sambrook et al. 1989) maintained in a water bath at 59.5°C, over a period of 12 hr at 10 V/cm. The gel was visualized by staining with ethidium bromide and the DNA-containing gel

slices were cut with sterile scalpels. The DNA was extracted using electrophoresis in DEAE–cellulose membrane (Sambrook et al. 1989), subcloned into appropriately prepared pGEM-5Zf, pGEM7-Zf (Promega), and pUC18 (TaKaRa) vectors, and multiple clones isolated and stored at −80°C. The clones were analyzed by hybridization and sequencing.

cDNA Selection

Forty-four cosmids, representing ~1.7 Mb of the region, were pooled in a single experiment. The cosmids used for this experiment are—from centromere to telomere—3K8, 45J9, 22A13, 18I21, 8G1, 41H20, 23P3, 34P18, 18H1, 45P17, 2E22, 52N2, 18D21, 20L4, 43K2, 4B4, 15B22, 3N2, 20L6, 20G1, 36, 25, 4R, 4N, 17E12, 36P16, 27R, 44N22, 33F1, 45F22, 1J2, 9I10, 41P11, 3A15, 32O22, 37A9, 51G4, 7C15, 47K16, 39E3, 16E11, 51B16, 49H12, and 1J20 from a cosmid contig that maps in the MHC class III region (Fig. 1). Direct cDNA selection was performed as described by Korn et al. (1992). Three random primed cDNA libraries, with tissue-specific adapters, were used simultaneously. The cDNA libraries used were a human fetal brain (mixture of eight brains ranging from 21 to 26 weeks), a human fetal liver (mixture of three livers of 22, 23, and 26 weeks), and a human adult muscle (normal male, 30 years old). A total of 1920 clones was picked and stored in 96-well plates and high-density filters were generated. Part of the resulting selected cDNA eluate was used as a probe to screen a Southern blot with *Eco*RI digests of the cosmids used for the cDNA selection experiment. The same blot was then hybridized with total human DNA to identify the bands with a high content of repeats. In this way specific bands were isolated for screening the generated material including whole cosmid DNA from every cosmid that was used in the cDNA selection experiment, and selected cDNAs were screened back to the cosmids.

5′ and 3′ Extension of Coding Sequences With RACE PCR

The Marathon-Ready cDNA kit (Clontech) was used for extending the coding sequences of identified cDNA clones toward the 5′ and 3′ directions, applying the RACE PCR technology. were 5.2H45′ (CTCCACACAAA CCACCCTCGACATC), 5.2H45′N (GTTCATAGGTTGTGGCTGGCTTCTG), 5.2H43′ (TGTCGTTGATTTAAAGGAGCCAGTGC), 5.2H43′N (CTATGGTGTCTCTGGTCGCGTCTG), and 5.2H4NEW (CTTCCAGCATCTTGAGCTTGTC) for ntcon2, 1.4F55′ (CCCACTAGGAGCCCCAAACT GTCAC), 1.4F55′N (GAAATTCCTGGAGGCTGCACTGG), 1.4F53′ (GTTCTCAAGACAGGTCCAGCCAGCTAC), 1.4F53′N (TATCTGAAGGCTCGTGTGCTAATCGC), and 1.4F5NEW (TGTGTTCTTCAGTTCAGGGAC) for cDNA clone 1.4/F5, X7L2 (AGCACCTCACTGTCTTCCTCG), X7L3 (CCAAACAAAAGACAGGAGCAGAGAGG), X7R1 (ATGATGAGCCCTGTTCCTGC), and X7R2 (TGCGTCAACCTCATCCCTCAGATTACATCC) for ntcon6, and 34H1R (GCTCTTTATGAACTGCCCAACTTCCTGTCC) and 34H1REV (TATCGGTTTGACGGTTACACTTGG) for the B1/*KIAA0229* transcript. The reactions and cycling conditions were the same for the primary and nested PCR amplifications and were according to the recommendations of the manufacturer. The generated single bands were isolated from the gel using GeneClean II (Bio101, Inc.) and subcloned using the TA cloning kit (Invitrogen). The generated clones were then sequenced.

Sequencing

For genomic sequencing whole clones were randomly subcloned into M13mp18 and pUC18. Recombinant clones (80% M13s, 20% pUCs) were picked, amplified, and purified in 96-well microtitre plates (Beck and Alderton 1993; Mardis 1994; A. Smith and L. Baron, unpubl.). The DNA sequence was determined using enzymatic dideoxy chain termination sequencing chemistry and automated ABI 373/377 DNA sequencers (Applied Biosystems, Foster City, CA). The generated reads were quality clipped, screened for cloning and sequencing vectors, and assembled into contigs as described previously (Avis et al. 1997). For all other sequencing (of cDNA, exon trapped, and GC-rich clones) the automated ABI 377 DNA sequencer was used and the finished sequences were stored in GeneJockeyII (Biosoft) formatted files.

Electronic Analysis of the Sequencing Data

The finished genomic sequence was analyzed and annotated using the HPREP analysis suite (G. Micklen and R. Durbin, unpubl.) and the Sanger Centre analysis strategy (http://www.sanger.ac.uk/Teams/Informatics/Humana/). General sequence data manipulations were performed using the GeneJockeyII program (Biosoft). PCR-related analysis was done using the Amplify v1.2 program (University of Wisconsin). Searches of the public databases were done via the World Wide Web, primarily at the Baylor College developed "BCM Search Launcher" (Worley et al. 1995; Smith et al. 1996) (http://gc.bcm.edu:8088/search-launcher/launcher.html) and at the National Centre for Biotechnology Information supported BLAST and Gapped-BLAST web pages (http://www.ncbi.nlm.nih.gov/; Altschul et al. 1997). The NIX platform (http://menu. hgmp.mrc.ac.uk/menu-bin/Nix/Nix.pl) was also used for both the genomic and potential mRNA sequences. In this platform the sequence is masked for repeat regions using Washington University's RepeatMasker program and Blast searches are started using the masked sequence against the various databases. The following exon-finding programs are run using the masked sequence: Grail, Genemark, Fex, Hexon, Fgene. The trnascan program is run on the sequence to identify t-RNA genes. The algorithm to determine CpG islands is based on the definition of CpG Islands by Gardiner-Garden and Frommer (1987). Sequence exploration and gene discovery version 1.3 was used (Informatics Group, Oak Ridge National Laboratory, Oak Ridge, TN, USA). The SPSS statistical package (Windows version, 7.5) was used for calculating Spearman's ρ values and levels of significance.

Northern Blots

The following Northern blots were commercially obtained: Human Multiple Tissue Northern (MTN) Blot (lot 47886 and 21/165), Human Multiple Tissue Northern (MTN) Blot II (lot 46969), Human Fetal Multiple Tissue Northern (MTN) Blot II (lot 49927), and Mouse Multiple Tissue Northern (MTN) Blot (lot 2x155) (all from Clontech) and Adult Human Tissue Total RNA Blot (serial no. H347) (BIOS Laboratories). Hybridization of Northern blots was performed following the recommendations of the manufacturer. Hybridization took place over a period of at least 18 hr and usually over 24 hr at 42°C. Washing of the blots was also according to the manufacturer's recommendations, typically including the high stringency wash. The exposure times ranged from 14 days to 1 month.

Southern Blotting, Hybridization, and High-Density Gridding

Southern blotting, probe labeling, competition, hybridization, and high-density gridding onto nylon filters (Hybond-N+, Amersham) were performed as described in Tripodis et al. (1998).

ACKNOWLEDGMENTS

This work was supported by grants from the United Kingdom's MRC (G9215293) and EC (GENE-CT93-0075). Human sequencing at the Sanger Centre is funded by the Wellcome Trust. We would like to thank M-L. Yaspo, B. Korn, and A.M. Poustka for help with exon trapping and cDNA selection and all members of the chromosome 6 project group (http://www.sanger.ac.uk/HGP/Chr6/).

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

REFERENCES

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

Avis, T., E.K. Clark, T.L. Flack, M. Mohammadi, S. Milne, D. Niblett, S. Palmer, S. Phillips, C. Smalley, M. Tagney, K.L. Thorpe, B. Tubby, J. Westrop, and S. Beck. 1997. The chromosome 6 sequencing project at the Sanger Centre. *DNA Sequence* **8:** 131–136.

Banerjee, P., P.W. Kleyn, J.A. Knowles, C.A. Lewis, B.M. Ross, E. Parano, S.G. Kovats, J.J. Lee, G.K. Penchaszadeh, J. Ott, S.G. Jacobson, and T.C. Gilliam. 1998. TULP1 mutation in two extended Dominican kindreds with autosomal recessive retinitis pigmentosa. *Nature Genet.* **18:** 177–179.

Beck, S. and R.P. Alderton. 1993. A strategy for the amplification, purification, and selection of M13 templates for large-scale DNA sequencing. *Anal. Biochem.* **212:** 498–505.

Bernardi, G. 1989. The isochore organization of the human genome. *Annu. Rev. Genet.* **23:** 637–661.

Chaplin, T., P. Ayton, O.A. Bernard, V. Saha, V. Della Valle, J. Hillion, A. Gregorini, D. Lillington, R. Berger, and B.D. Young. 1995. A novel class of zinc finger/leucine zipper genes identified from the molecular cloning of the t(10;11) translocation in acute leukemia. *Blood* **85:** 1435–1441.

Chen, H.J., M. Rojas-Soto, A. Oguni, and M.B. Kennedy. 1998. A synaptic Ras-GTPase activating protein (p135 SynGAP) inhibited by CaM kinase II. *Neuron* **20:** 895–904.

Church, D.M., C.J. Stotler, J.L. Rutter, J.R. Murrell, J.A. Trofatter, and A.J. Buckler. 1994. Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nature Genet.* **6:** 98–105.

Coulson, M., S. Robert, H.J. Eyre, and R. Saint. 1998.The identification and localization of a human gene with sequence similarity to Polycomblike of Drosophila melanogaster. *Genomics* **15:** 381–383.

Crawley, J.B., L. Rawlinson, F.V. Lali, T.H. Page, J. Saklatvala, and B.M. Foxwell. 1997. T cell proliferation in response to interleukins 2 and 7 requires p38MAP kinase activation. *J. Biol. Chem.* **272:** 15023–15027.

Creutz, C.E., J.L. Tomsig, S.L Snyder, M.C. Gautier, F. Skouri, J. Beisson, and J. Cohen. 1998. The copines, a novel class of C2 domain-containing, calcium-dependent, phospholipid-binding proteins conserved from Paramecium to humans. *J. Biol. Chem.* **273:** 1393–1402.

Cuypers, H.T., G. Selten, A. Berns, and A.H.G. van Kessel. 1986. Assignment of the human homologue of Pim-1, a mouse gene implicated in leukemogenesis, to the pter-q12 region of chromosome 6. *Hum. Genet.* **72:** 262–265.

el-Deiry, W.S., T. Tokino, V.E. Velculescu, D.B. Levy, R. Parsons, J.M. Trent, D. Lin, W.E. Mercer, K.W. Kinzler, and B. Vogelstein. 1993. WAF1, a potential mediator of p53 tumor suppression. *Cell* **75:** 817–825.

Fisicaro, N., M. Katerelos, J. Williams, D. Power, A. D'Apice, and M. Pearse. 1995. Identification of genes downregulated in the thymus by cyclosporin-A: preliminary characterization of clone CSA-19. *Mol. Immunol.* **32:** 565–572.

Fong, S.T., J. Camakaris, and B.T. Lee. 1995. Molecular genetics of a chromosomal locus involved in copper tolerance in Escherichia coli K-12. *Mol. Microbiol.* **15:** 1127–1137.

Friedmann, M., L.T. Holth, H.Y. Zoghbi, and R. Reeves. 1993. Organization, inducible-expression and chromosome localization of the human HMG-I(Y) nonhistone protein gene. *Nucleic Acids Res.* **21:** 4259–4267.

Gardiner-Garden, M. and M. Frommer. 1987. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196:** 261-282.

Goedert, M., A. Cuenda, M. Craxton, R. Jakes, and P. Cohen. 1997. Activation of the novel stress-activated protein kinase SAPK4 by cytokines and cellular stresses is mediated by SKK3 (MKK6); comparison of its substrate specificity with that of other SAP kinases. *EMBO J.* **16:** 3563–3571.

Gregorini, A., F.I. Sahin, D.M. Lillington, J. Meerabux, V. Saha, P. McCullagh, M. Bocci, S. Menevse, S. Papa, and B.D. Young. 1996. Gene BR140, which is related to AF10 and AF17, maps to chromosome band 3p25. *Genes Chromosomes & Cancer* **17:** 269–272.

Gui, J.F., W.S. Lane, and X.D. Fu. 1994. A serine kinase regulates intracellular localization of splicing factors in the cell cycle. *Nature* **369:** 678–682.

Habert-Ortoli, E., B. Amiranoff, I. Loquet, M. Laburthe, and J.F. Mayaux. 1994. Molecular cloning of a functional human galanin receptor. *Proc. Natl. Acad. Sci.* **91:** 9780–9783.

Haendler, B. and E. Hofer. 1990. Characterization of the human cyclophilin gene and of related processed pseudogenes. *Eur. J. Biochem.* **190:** 477–482.

Hagstrom, S.A., M.A., North, P.M. Nishina, E.L. Berson, and T.P. Dryja. 1998. Recessive mutations in the gene encoding the tubby-like protein TULP1 in patients with retinitis pigmentosa. *Nature Genet.* **18:** 174–176.

Harper, J.W., G.R. Adami, N. Wei, K. Keyomarsi, and S.J. Elledge. 1993. The p21 Cdk-interacting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases. *Cell* **75:** 805–816.

Herberg, J.A., S. Beck, and J. Trowsdale. 1998a. TAPASIN, DAXX, RGL2, HKE2 and four new genes (BING 1, 3 to 5) form a dense cluster at the centromeric end of the MHC. *J. Mol. Biol.* **277:** 839–857.

Herberg, J.A., J. Sgouros, T. Jones, J. Copeman, S.J. Humphray, D. Sheer, P. Cresswell, S. Beck, and J. Trowsdale. 1998b. Genomic analysis of the Tapasin gene, located close to the TAP loci in the MHC. *Eur. J. Immunol.* **28:** 459–467.

Herberg, J.A., S. Phillips, S. Beck, T. Jones, D. Sheer, J.J. Wu, V. Prochazka, P.J. Barr, M.C. Kiefer, and J. Trowsdale. 1998c. Genomic structure and domain organisation of the human Bak gene. *Gene* **211:** 87–94.

Hotfilder, M., S. Baxendale, M.A. Cross, and F. Sablitzky. 1999. Def-2, -3, -6 and -8, novel mouse genes differentially expressed in the haemopoietic system. *Br. J. Haematol.* **106:** 335–344.

Jacquemin, P., J.A. Martial, and I. Davidson. 1996. Human TEF-5 is preferentially expressed in placenta and binds to multiple functional elements of the human chorionic somatomammotropin-B gene enhancer. *Biol. Chem.* **272:** 12928–12937.

Janitz, K., A. Wild, S. Beck, S. Savasta, G. Beluffi, A. Ziegler, and A. Volz. 1999. Genomic organisation of the HSET locus and the possible association of HLA-linked genes with Immotile Cilia syndrome (ICS). *Immunogenetics* **49:** 644–652.

Johansson, M., C. Dietrich, N. Mandahl, G. Hambraeus, L. Johansson, P.P. Clausen, F. Mitelman, and S. Heim. 1993. Recombinations of chromosomal bands 6p21 and 14q24

characterise pulmonary hamartomas. *Br. J. Cancer* **67:** 1236–1241.
Jumaa, H. and P.J. Nielsen. 1997. The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation. *EMBO J.* **16:** 5077–5085.
Jumaa, H., J.L. Guenet, and P.J. Nielsen. 1997. Regulated expression and RNA processing of transcripts from the Srp20 splicing factor gene during the cell cycle. *Mol. Cell. Biol.* **17:** 3116–3124.
Katoh, M., M. Hirai, T. Sugimura, and M. Terada. 1996. Cloning, expression and chromosomal localization of *Wnt-13*, a novel member of the Wnt gene family. *Oncogene* **13:** 873–876.
Katsanis, N., J. Fitzgibbon, and E.M.C. Fisher. 1996. Paralogy mapping: Identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. *Genomics* **35:** 101–108.
Kikuti, Y.Y., G. Tamiya, A. Ando, L. Chen, M. Kimura, E. Ferreira, K. Tsuji, J. Trowsdale, and H. Inoko. 1997. Physical mapping 220 kb centromeric of the human MHC and DNA sequence analysis of the 43-kb segment including the RING1, HKE6, and HKE4 genes. *Genomics* **42:** 422–435.
Korn, B., Z. Sedlacek, A. Manca, P. Kioschis., D. Konecki, H. Lehrach, and A. Poustka. 1992. A strategy for the selection of transcribed sequences in the Xq28 region. *Hum. Mol. Genet.* **1:** 235-242.
Kostyu, D.D. 1994. HLA: Fertile territory for developmental genes? *Crit. Rev. Immunol.* **14:** 29–59.
Kumar, S., P.C. McDonnell, R.J. Gum, A.T. Hand, J.C. Lee, and P.R. Young. 1997. Novel homologues of CSBP/p38 MAP kinase: Activation, substrate specificity and sensitivity to inhibition by pyridinyl imidazoles. *Biochem. Biophys. Res. Commun.* **235:** 533–538.
Lee, J.C., J.T. Laydon, P.C. McDonnell, T.F. Gallagher, S. Kumar, D. Green, D. McNulty, M.J. Blumenthal, J.R. Heys, S.W. Landvatter et al. 1994. A protein kinase involved in the regulation of inflammatory cytokine biosynthesis. *Nature* **372:** 739–746.
Lovett, M., J. Kere, and L.M. Hinton. 1991. Direct selection: A method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci.* 88:9628-9632.
Lux, S.E., W.T. Tse, J.C. Menninger, K.M. John, P. Harris, O. Shalev, R.R. Chilcote, S.L. Marchesi, P.C. Watkins, V. Bennett, S. McIntosh, F.S. Collins, U. Francke, D.C. Ward, and B.G. Forget. 1990. Hereditary spherocytosis associated with deletion of human erythrocyte ankyrin gene on chromosome 8. *Nature* **345:** 736–739.
Mardis, E.R. 1994. High-throughput detergent extraction of M13 subclones for fluorescent DNA sequencing. *Nucleic Acids Res.* **22:** 2173–2175.
Millward, T., P. Cron, and B.A. Hemmings. 1995. Molecular cloning and characterization of a conserved nuclear serine (threonine) protein kinase. *Proc. Natl. Acad. Sci.* **92:** 5022–5026.
Nair, S.C., R.A. Rimerman, E.J. Toran, S. Chen, V. Prapapanich, R.N. Butts, and D.F. Smith. 1997. Molecular cloning of human FKBP51 and comparisons of immunophilin interactions with Hsp90 and progesterone receptor. *Mol. Cell Biol.* **17:** 594–603.
North, M.A., J.K. Yan, Y. Naggert, K. Noben-Trauth, and P.M. Nishina. 1997. Molecular characterization of TUB, TULP1, and TULP2, members of the novel tubby gene family and their possible relation to ocular diseases. *Proc. Natl. Acad. Sci.* **94:** 3128–3133.
Prasad, R, D. Leshkowitz, Y. Gu, H. Alder, T. Nakamura, H. Saito, K. Huebner, R. Berger, C.M. Croce, and E. Canaani. 1994. Leucine-zipper dimerization motif encoded by the AF17 gene fused to ALL-1 (MLL) in acute leukemia. *Proc. Natl. Acad. Sci.* **91:** 8107–8111.
Ragoussis, J., G. Senger, I. Mockridge, P. Sanseau, S. Ruddy, K. Dudley, D. Sheer, and J. Trowsdale. 1992. A testis-expressed Zn finger gene (ZNF76) in human 6p21.3 centromeric to the MHC is closely linked to the human homolog of the t-complex gene tcp-11. *Genomics* **14:** 673–679.
Safrany, S.T., J.J. Caffrey, X. Yang, M.E. Bembenek, M.B. Moyer, W.A. Burkhart, and S.B. Shears. 1998. A novel context for the "MutT" module, a guardian of cell integrity, in a diphosphoinositol polyphosphate phosphohydrolase. *EMBO J.* **17:** 6599–6607.
Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. Molecular Cloning. *A laboratory manual*, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Schmidt, A., N. Endo, S. Rutledge, R.Vogel, D. Shinar, and G.A. Rodan. 1992. Identification of a new member of the steroid hormone receptor superfamily that is activated by a peroxisome proliferator and fatty acids. *Mol. Endocrinol.* **6:** 1634–1641.
Shiraishi, M., L.S. Lerman, and T. Sekiya. 1995. Preferential isolation of DNA fragments associated with CpG islands. *Proc. Natl. Acad. Sci.* **92:** 4229–4233.
Smith, R.F., B.A. Wiese, M.K. Wojzynski, D.B. Davison, and K.C. Worley. 1996. BCM Search Launcher—An integrated interface to molecular biology data base search and analysis services available on the World Wide Web. *Genome Res.* **6:** 454–462.
Taylor, A., T. Surgenor, D.K. Thomson, R.J. Graham, and H. Oettgen. 1984. Comparison of leucine aminopeptidase from human lens, beef lens and kidney, and hog lens and kidney. *Exp. Eye Res.* **38:** 217–229.
Thompson, K.A., B. Wang, W.S. Argraves, F.G. Giancotti, D.P. Schranck, and E. Ruoslahti. 1994. BR140, a novel zinc-finger protein with homology to the TAF250 subunit of TFIID. *Biochem. Biophys. Res. Commun.* **198:** 1143–1152.
Tripodis, N., R. Mason, S.J. Humphray, A.F. Davies, J.A. Herberg, J. Trowsdale, D. Nizetic, G. Senger, and J. Ragoussis. 1998. Physical map of human 6p21.2–6p21.3: Region flanking the centromeric end of the major histocompatibility complex. *Genome Res.* **8:** 631–643.
Ulrich, E., A. Kauffmann-Zeh, A.O. Hueber, J. Williamson, T. Chittenden, A. Ma, and G. Evan. 1997. Gene structure, cDNA sequence, and expression of murine Bak, a proapoptotic Bcl-2 family member. *Genomics* **44:** 195–200.
Williams, A.J., W.L. Powell, T. Collins, and C.C. Morton. 1997. HMGI(Y) expression in human uterine leiomyomata. Involvement of another high-mobility group architectural factor in a benign neoplasm. *Am. J. Pathol.* **150:** 911–918.
Worley, K.C., B.A. Wiese, and R.F. Smith. 1995. BEAUTY: An enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. *Genome Res.* **5:** 173–184.
Xiao, S, M.L. Lux, R. Reeves, T.J. Hudson, and J.A. Fletcher. 1997. HMGI(Y) activation by chromosome 6p21 rearrangements in multilineage mesenchymal cells from pulmonary hamartoma. *Am. J. Pathol.* **150:** 901–910.
Yaspo, M.L., L. Gellen, R. Mott, B. Korn, D. Nizetic, A.M. Poustka, and H. Lehrach. 1995. Model for a transcript map of human chromosome 21: Isolation of new coding sequences from exon and enriched cDNA libraries. *Hum. Mol. Genet.* **4:** 1291–1304.
Yoshikawa, T., Z. Brkanac, B.R. Dupont, G-Q. Xing, R.J. Leach, and S.D. Detera-Wadleigh. 1996. Assignment of the human nuclear hormone receptor, NUC1 (PPARD), to chromosome 6p21.1–p21.2. *Genomics* **35:** 637–638.
Zhang, F.L., R.E. Diehl, N.E. Kohl, J.B. Gibbs, B. Giros, P.J. Casey, and C.A. Omer. 1994a. cDNA cloning and expression of rat and human protein geranylgeranyltransferase type-I. *J. Biol. Chem.* **269:** 3175–3180.
Zhang, H., S.D. Apfelroth, W. Hu, E.C. Davis, C. Sanguineti, J. Bonadio, R.P. Mecham, and F.T. Ramirez. 1994b. Structure and expression of fibrillin-2, a novel microfibrillar component preferentially located in elastic matrices. *J. Cell. Biol.* **124:** 855–863.

Received April 29, 1999; accepted in revised form February 2, 2000.