ORIGINAL PAPER

Generation of a transcription map distal to HLA-F

Stefano Goldwurm¹, Benjamin FH Van der Griend¹, Joanne L Banyer¹, Lara M Cullen¹, Anna Zournazi¹, Moira L Menzies¹, Frances Busfield¹, Peter FR Little² and Elizabeth C Jazwinska¹

¹*Clinical Sciences Unit, The Queensland Institute of Medical Research and Department of Medicine, University of Queensland, Brisbane, Australia*

²*Department of Biochemistry, Imperial College of Science, Technology and Medicine, London, UK*

We have constructed a transcription map covering a 2 Mb region beginning approximately 1 Mb distal to HLA-F. Cosmids isolated from a chromosome 6 library were positioned by YAC hybridisation, STS and fingerprint analysis. Using direct cDNA selection, exon trapping, and direct genomic sequence analysis, we identified 42 potential exonic fragments in this region. Six fragments corresponded to previously characterised genes, four previously broadly mapped to this region. Five fragments were similar to known genes, eight fragments matched ESTs and 10 of the remaining 23 novel fragments, gave a positive signal on northern analysis. All cDNA fragments were mapped to the YAC and cosmid contig covering the region and with respect to other known genes and STS in this area. The distribution of the cDNA fragments indicated their organisation in three clusters around CpG islands.

Keywords: chromosome 6; transcript map; cDNA selection; exon trapping

Introduction

The construction of a transcription map aids the characterisation of genes present within a genomic region and allows identification of their intron-exon organisation. Together with linkage analysis data, detailed transcription maps can greatly increase the speed of discovery of disease genes; indeed in some cases it is possible to move straight to the analysis of candidate genes already identified in a candidate region defined by linkage analysis. Now that large scale

sequencing of the human genome is well underway this 'positional candidate' approach is likely to become the predominant strategy for the identification of disease genes.¹

Several diseases are known to be linked to the major histocompatibility complex (MHC) region at 6p 21.3, one of the most gene-dense areas of the human genome. The *MHC* genes are divided into three classes and are clustered on the chromosome accordingly. Linkage of diseases to the *MHC* is sometimes due to the key role these genes have in regulation of immunity, but in other cases, such as dyslexia,² schizophrenia,³ and autism,⁴ the connection with the immune system is not so obvious, and predisposition may be caused by non-MHC related genes encoded within the region. One such disease is haemochromatosis (HC), a defect of iron metabolism, linked to the HLA-A locus in 1975⁵

Correspondence: Dr EC Jazwinska, SmithKline Beecham, New Frontier's Science Park North, Harlow, Essex CM19 5AD, UK

Received 1 December 1997; revised 30 January 1998; accepted 2 February 1998

but not thought to be due to a gene involved in immunity control.

The region distal to HLA-F, has not been studied as extensively as the MHC Class I, Class II and III regions. We previously constructed a YAC genomic contig covering this distal region,⁶ and subsequently we identified a new Krüppel-type zinc finger gene, *ZNF184*, in this region⁷ and characterised a genomic segment which contains several genes with homology to genes in the spinal muscular atrophy (*SMA*) gene region of chromosome 5q13.1 (Banyer *et al.*, unpublished data). Recently the gene for HC, initially termed HLA-H but later renamed *HFE* according to the accepted HUGO/GDB nomenclature,^{8,9} was identified.¹⁰⁻¹² *HFE*, located approximately 5 Mb distal of *HLA-A*, has high homology with Class I MHC molecules and is thus the most distal MHC Class I gene identified to date. The identification of *HFE* now further extends the distal border of the MHC previously thought to end at *HLA-F* just 200 kb distal of *HLA-A*.

Further characterisation of the region between *HLA-F* and *HFE* is now needed to determine the gene content of this region and to facilitate the identification of disease genes linked to this area. To achieve this goal, we constructed a cosmid contig covering approximately 2 Mb, beginning approximately 1 Mb distal of *HLA-F* and ending approximately 2 Mb proximal of *HFE*. We combined three rapid gene identification techniques – direct cDNA selection, exon trapping, and direct genomic sequencing – to isolate gene fragments. These were positioned in the cosmid contig with respect to other genes we have previously mapped to this area, thus generating a transcription map of this distal region of the MHC.

Materials and Methods

Cosmid Isolation

YACs 24ce5, 17ah2, 306e5, 259f7, 753h12, and 950h11 were obtained from CEPH and ICI libraries and organised in a contig as previously described.⁶

Localisation and integrity of YACs was verified by fluorescence *in situ* hybridisation (FISH). YAC DNA, prepared as previously described,⁶ was used to select Lawrist cosmids from a chromosome 6 specific gridded library (cosgrid) using recommended hybridisation conditions.¹³ To validate the cosmids selected from the cosgrid by hybridisation analysis, DNA from individual cosmids was prepared using Qiagen plasmid mini kits (Cat No. 12125) according to manufacturer's instructions, and *Bam*HI/*Sca*I digests were Southern blotted and probed with YAC DNA. Only those cosmids

clearly mapping to the original YAC used in their selection were analysed further.

PCR Analysis

Cosmid DNA was analysed for the presence of STS comprising YAC end-clones and microsatellites as previously described.^{6,14} Primer sequences, annealing temperatures and PCR product sizes of STS used are given in Table 2. All PCRs comprised 35 cycles and were performed in 25 µl reaction volumes containing 100 ng of each primer, 400 µM dNTPs, 1.5 mM MgCl₂, 1.25 U of Taq polymerase (Cetus USA), and 100 ng of template DNA.

Fingerprint Analysis

Manual and automated fingerprint analyses of cosmids were performed as follows. For manual fingerprinting, *Bam*HI/*Sca*I digested Lawrist cosmid DNA was electrophoresed and visualised on 0.8% agarose gels. For automated fingerprinting *Hind*III/*Sau*3AI digests were performed on 2 µl of microprep DNA in 96 well microtitre plates, as previously described.¹⁵ Samples were run on denaturing polyacrylamide gels (4% acrylamide/bisacrylamide (19:1), 7 M urea) in TBE, against size markers consisting of lambda DNA cut with *Sau*3AI and end labelled with [³²S]dATP (Amersham UK). Gels were run at 75 W for 1.5 hours, fixed and autoradiographs were digitised using an Amersham Filmreader Scanner and analysed using Image2.1 gel analysis software. (<http://www.sanger.ac.uk/fingerprinting/imageprocessing/overview.html>). Contigs were assembled using Contig C software for assembling physical maps from fingerprints, derived from VAX CONTIG9 programs.¹⁶ The Contig C package is available by anonymous ftp from <ftp.sanger.ac.uk/pub/contigc/>. A band tolerance of 0.7 mm and random match probabilities of less than 10⁻⁴ were used in the analysis.

cDNA Selection

Four direct cDNA selection experiments were performed as follows:

- (1) using YACs 259f7 and 306e5 subcloned into lambda Fix (Stratagene);
- (2) using YACs 753h12 and 950h11 subcloned into Supercos (Stratagene);
- (3) using Lawrist cosmids from chromosome 6 libraries selected with YACs 306e5, 259f7, 753h12, and 950h11;
- (4) using Lawrist cosmids selected with YAC 17ah2.

Pools of cDNA fragments were amplified from the following λgt10 or λgt11 human libraries: placenta (American Type Culture Collection, Rockville, MD, USA ATCC 77399), thymus (ATCC 77081), hepatocellular carcinoma Hep G2 cell line (ATCC 77400), normal liver (ATCC 77402), neonatal liver (ATCC 77427), and small intestine (Clontech HL1133b). cDNA selection was performed as previously described¹⁷ with the following modifications: amplified cDNA (3 µg) was blocked with sonicated human placental DNA (2 µg), sonicated yeast AB1380 DNA (1 µg) and pBluescript (Stratagene) DNA digested with *Eco*RV (2 µg). After the second round of selection the cDNAs were size-selected (0.3–1 kb) by agarose gel electrophoresis and gel purification, PCR amplified, digested with *Eco*RI and ligated in pBluescript. Resulting clones were screened by hybridisation with the corresponding

YAC probe. Only those clearly mapping to YACs were sequenced.

Exon Trapping

Exon trapping was performed as originally described,¹⁸ using Lawrist cosmids isolated with YAC 259f7, and using Supercos (Stratagene) USA subcloned YACs 753h12 and 950h11. The cosmids were pooled, digested with *Bam*HI and *Bgl*II and then ligated into *Bam*HI digested exon trapping vector pSPL3B (generously supplied by Dr T Burn of Integrated Genetics Inc., USA) and used with the exon trapping system (GIBCO BRL, Cat. No. 18449-017).

DNA Sequencing

This was performed with the appropriate vector primers using the fluorescent-based dideoxynucleotide chain terminating method¹⁹ and Taq DyeDeoxyl™ Terminator premix (Applied Biosystems USA) according to the manufacturer's instructions. All samples were processed by an Applied Biosystem model 373A automated DNA sequencer.

Mapping of Clones to the YAC and Cosmid Contig

Purified inserts from clones were ³²P-labelled with a Random Primed DNA Labelling kit (Boehringer Mannheim Germany) and used as probes against Southern blots of *Eco*RI digested YAC DNA,⁶ and *Bam*HI/*Sca*I digested Lawrist cosmid DNA. Blots were washed to a final stringency of 0.1 × SSC, 0.1% SDS at 65°C for 20 min and exposed to X-ray films O/N at -70°C.

Northern Analysis

Total human RNA (25 µg) from brain (Clontech USA, Cat. No. 64020-1), liver (Clontech, Cat. No. 64022-1), placenta (Clontech, Cat. No. 64024-1), and thymus (Clontech, Cat. No. 64028-1) was electrophoresed and blotted according to standard procedures²⁰ and probed with ³²P-labelled inserts of clones as described above. Blots were washed to a final stringency of 0.5 × SSC, 0.1% SDS at 65°C for 20 min, exposed for 4-8 days and visualised using a phosphorimager. Band size was estimated by comparison with a molecular weight marker (GIBCOBRL USA, Cat. No. 15620-016).

Direct Genomic Sequencing

DNA was prepared from cosmids using Qiagen plasmid mini kits and sequenced using Lawrist primers T7 and Sp6. Automated sequencing was performed as described above.

Sequence Analysis

All sequences were compared against the GenBank (National Centre for Biotechnology Information USA) sequence database, searching for identities with known genes, ESTs or structural domains using the programs BLASTN²¹ and BLASTX.²⁶ Clones corresponding to cryptic spliced products of the exon trapping vector pSPL3 B, or due to contaminations such as *E. coli*, yeast, mitochondrial, and ribosomal sequences, or containing mainly repeat elements, were not considered further. Clones identical to ESTs in the database, were extended where possible using the GCG Fragment Assembly System.

Results

Cosmid Contig

One hundred and seventy Lawrist cosmids were isolated and organised in bins according to YAC hybridisation patterns. Seventy-five cosmids were ordered from the proximal STS 814R to the distal STS D6S1558 by analysis of STS content combined with results of fingerprint analysis (Table 1). The bin arrangement provided an approximate location for cosmids which was then refined by STS and fingerprint analysis. Several cDNA fragments isolated by direct selection were positioned in the cosmid contig and used as STS to recognise overlapping cosmids (Table 2).

Generation of a Transcription Map

In all, 42 potentially exonic DNA fragments were isolated: 28 fragments by direct cDNA selection, eight by exon trapping, and six by direct genomic sequence analysis; all are subsequently referred to as cDNA fragments. cDNA fragments isolated by either direct cDNA selection or exon trapping were mapped to YAC and cosmid contigs by hybridisation. Where possible, results were confirmed by PCR analysis (see Table 2). cDNA fragments were positioned with respect to YACs, cosmids and previously characterised genes⁷ (Banyer *et al.*, unpublished) (see Figure 1). cDNA fragments mapped to 3 clusters, one in YAC 17ah2, one around the microsatellite D6S105, and the third on the distal end of YAC 950h11. The size, method of identification, and sequence analysis of each of the 42 cDNA fragments is given in Table 3.

Clones Corresponding to Known Genes

Six cDNA fragments were found to correspond to previously characterised genes, four were known to map in this area: B123 and Y247 were identical to histone genes, and F06 and F26 corresponded to the zinc finger gene *LD5*. The other cDNA fragments were mapped for the first time to this region: J13 was identical to the human heterogeneous nuclear ribonucleoprotein *H1* (hnRNP H1) gene, and D2 was identical to the tRNA-Val (CAC).

Clones with Homology to Known Genes

Three cDNA fragments (F44, F20 and F36, see Table 3) were homologous to domains of the Krüppel-related *ZNF* genes.²³ F44 was identical to the 3'-sequence of two ESTs, which enabled us to construct an extended sequence containing a Krüppel-type C2H2 zinc finger followed by a 3'UTR and a polyA tail. F20 contained Krüppel associated boxes (KRAB) A and B and some

Table 1 Cosmid contig

<i>Cosmid</i>	<i>STS</i>	<i>YAC</i>	<i>Clones</i>
P1452		24C	
F0445		24C	
K063		24C	F24
F0522		24C	F24
E1944		24C, 17A	F24
M1232		24C, 17A	F24
A1416		24C, 17A	F24
M2046		24C, 17A	F24
00241		24C, 17A	F24
B216	814R	24C, 17A	
D1027	814R, 849R	24C, 17A	F14, F54
H2341	814R, 849R	24C, 17A	F14, F54, F20
D0845	814R, 849R	24C, 17A	F14, F54, F20
D1419	849R	24C, 17A	F54, F20
L0614		24C, 17A	F54, F20
K1643	753L	24C, 17A	F54, F20, F44, F21, F08
M1412	753L	24C, 17A	F54, F20, F44, F21, F08
H0131	753L	24C, 17A	F54, F20, F44, F21, F08
N1813	753L	753, 17A	F36, F16, F45, F23, F50
C0926		17A	F36, F16, F45, F23, F47
K1837	19AR	753, 17A	F36, F16, F45, F23, F47, F26, F06
K0613		753, 17A	F45, F23, F47
J1513		753, 17A	F26, F06
I0636		753	
D0825	950L	753	
K114†	950L	753	
C1219*	306L, D6S306	753, 950	
P1343	306L, D6S306	306	
A0928†		753	B123, Y247
B1815		753, 950	
N0617		950	
B1046		753, 950	
L173		753, 950	
G072		753, 259	
O0103	D6S105	950	
J1334*	D6S105	950	
N117	D6S464	753, 950	
M0748*		753	
F087†		306, 259	G81, G60, G47
F107*		306, 259	G81, G60, G47
H152*		753, 950, 259	Y245, R1, R23
L0628†		753	Y245, R1, R23
G0426†		753, 259	R23, R21, R10, J13, J35, B296
L0424		753, 950	J13, J35
E153*		753, 950, 259	J13, J35
O0332		753, 950, 259	J13, J35
G072*		259	B296, ZNF184
H1648*		753, 259	ZNF184
J1615*		753	ZNF184, R19
B0947*		753, 950, 259	
A186*		950	D1
A206		950	
I2343*		950	
B1549*		950	
N078*		950	

Table 1 *Continued*

<i>Cosmid</i>	<i>STS</i>	<i>YAC</i>	<i>Clones</i>
F028*		950	
L2138*		753, 950	
B0128		753, 950	
A1436*		950	
L1244†		950	
O0435		950	
C1311*	D6S1260	753, 950	
B1253*	D6S1260	950	B256, G32, D2
D1438*		950	
P1251		950	
N0635*		950	
I0926		950	
I0926		950	
K0419*		950	
O0714*		950	G52, B14, G79, J5, B200, D6
B148*	D6S1558	950	G52, B14, G79, J5, B200, D3
M2452*		950	J23, G52, B14, G79, J5, B200, D8
M0119		950	G59, G97, J23, G52, B14, G79, J5, B200, G96
L0454*		950	J23, G52, B14, G79, J5, D7
N223		950	G79, J5, B200
P0416		950	J21, G59, G97, J23

List of 75 cosmids ordered from centromere to telomere. The presence of STS, YAC hybridisation, and clones are indicated.

*Cosmids sequenced at both ends with primers SP6 and T7; †Cosmids sequenced only at one end.

Table 2 Primer sequences, annealing temperatures, and sizes of the new Sequence-Tagged-Sites derived from the cDNA fragments isolated

<i>Clone</i>	<i>Forward primer</i>	<i>Reverse primer</i>	<i>Ann. T. (°C)</i>	<i>Size (bp)</i>
F54	CGATAAGCATCTATTTGCCTGG	AAAACAGGGAAGAACATGGAGG	60	272
F20	TGAGATGCACAGCAAGGAGC	AATAGCACCACCACCTCCTCC	62	131
F44	AAAATGTAGTCTGGCTGGCTGAGATG	TCGTGAAAGATTTCATTAGAGGC	58	262
F36	GAGAAGTTTAATGATGACTGAAGCT	AGAATTCATACCGCAAATAAACTC	62	150
F26	TCCGTTCTCTAGCTGATGTTCC	TACTAGGAGACACTAGGACCCCG	62	173
F06	AATTCCGGCCATATCAGTGC	GTTGAATGAGACCAGTGTTC	60	402
G81	ACCTTGATGCCACTCTGTACG	AGGATGGGCGATTGCTCCAGG	60	122
B256	ACAGCTGGATTGCAGTTGTGG	CCTATTTTGCTATCCTGTCC	62	175
ZNF184	CCTGACATTTCTGAAGAAGAGC	GGTTTGCCTGCTGTCTCTCTAACTGCC	62	130

spacer sequence. F20 and F44 mapped close to each other on the same cosmid (Figure 1), but amplification of cosmid DNA with primers designed from F20 and F44 sequences failed to yield a product. It is possible that the two cDNA fragments may represent two separate zinc finger genes or the zinc finger domain and the 3'UTR are separated by an intron. PCR amplification of F44, F20 and F36 from the corresponding cosmids and YACs was performed, in order to confirm their location (Table 2).

R19 was similar to the end of the coding region of the *P5-1* gene. This gene is a member of a multicopy gene family in the Class I region,²⁴ while another cDNA fragment, G81, was similar to *G9a*, a gene localised in the Class III region of the MHC²⁵ (Table 3). The

presence of G81 in the YAC and cosmid contigs was confirmed by PCR amplification (Table 2), and by direct genomic sequencing of cosmids F087 and F107 using primers designed from the sequence of G81.

Clones Corresponding to Novel Sequences

Thirty-one cDNA fragments corresponded to novel sequences. Eight of these matched ESTs so Northern analysis was performed on the remaining 26 cDNA fragments, ten of which showed a visible transcript (Table 3); the remaining 16 gave no visible signal. Three novel clones, F16, F45, F23 mapped close to one another and may be part of the same transcription unit. No test was carried out here to see if they were indeed linked. Two clones, G60 and G40, mapped to the same

Figure 1 Transcription map distal to HLA-F.

The complete YAC contig is shown and minimal cosmid contig and location of cDNA fragments is indicated in a region beginning approximately 1 Mb distal to HLA-F. STS shown comprise: right end of YAC 814d10, right end of YAC 849f12, left end of YAC 753h12, D6S1260 and the cDNA clones F54, F20, F44, F36, F26, F06, G81, B256, and the gene *ZNF184* (see also Table 2). Approximate location of genomic sequences similar to known genes or ESTs are indicated. Numbered boxes correspond to the three gene clusters. *Clones or genomic sequences that match an EST; @ those identical to known genes. \$ the Krüppel-type zinc finger gene, *ZNF184*,⁶ # cDNA fragments previously characterised and similar to genes or cDNA fragments isolated on 5q13.1.²⁷

Table 3 Summary of cDNA fragments

<i>Clone⁽¹⁾</i> (<i>GenBank</i> <i>Accession</i> <i>No.</i>)	<i>Size</i> (<i>bp</i>)	<i>EST</i> <i>match⁽²⁾</i>	<i>Exten</i> <i>seq.⁽³⁾</i> (<i>bp</i>)	<i>Homologies and remarks</i>	<i>Northern</i> <i>analysis</i>
F24a ¹ (U77497)	504	No		Novel. Best ORF is 84 aa long	No detectable bands on northern analysis
F14a ² (U77498)	254	No		Novel. Alu repeat from 27 to 120. One ORF of 84 aa	No detectable bands on northern analysis
F54a ² (U77499)	539	No		Novel. No ORFs	No detectable bands on northern analysis
F20a ¹ (U77500)	492	No		ORF of 163 aa containing a KRAB A and B box and some sequence of the spacer of a novel Krüppel related ZNF. 96% identity at nucleotide level with CpG island (Z55026) ²⁸ 77% identity at amino acid level with ZNF165 (X84801) ³⁹ 72% identical at amino acid level with mouse Zfp-38 (g1083566) ^{40, 41}	ND
F44a ¹ (U77501)	502	290735 272024	809	ORF of 71 aa containing a C2H2 zincfinger of the Krüppel type followed by a 3'UTR and polyA tail	ND
F21a ¹ (U77502)	514	No		Novel. 100% identical to CpG island (Z60040) ²⁸ Best ORF is 66 aa long	No detectable bands on northern analysis
F08a ¹ (U77503)	490	No		Novel. Alu repeat from 380–460. No ORFs	No detectable bands on northern analysis
F50a ¹ (U77504)	626	No		Novel. Alu repeat from 434–456 and 534–568. No ORFs	1 transcript of 8.0 kb
F36a ¹ (U77505)	208	37049 31747 301073	452	ORF of 60 aa containing a C2H2 zincfinger of the Krüppel type 80% identity at the amino acid level with mouse Zfp-92 (586935) ⁴¹ 73% identity at amino acid level with mouse Zfp-35 (141622) ⁴² 100% identical to 3'UTR of three ESTs. The 5'sequences (387 bp) of the ESTs contain KRAB A and B boxes	ND
F16a ¹ (U77506)	283	No		Novel. No ORFs	No detectable bands on northern analysis
F45a ¹ (U77507)	106	No		Novel. Best ORF is 32 aa long	No detectable bands on northern analysis
F23a ¹ (U77508)	750	No		Novel. Repeat with 60% similarity to many human genomic sequences No ORFs	No detectable bands on northern analysis
F47a ¹ (U77509)	694	No		Novel. Alu repeat from 455 to 575. No ORFs	1 transcript of 4.0 kb
F26a ¹	747			100% identical to human ZNF LD5-1 (U57796) ³⁸	ND
F06a ¹	416			100% identical to human ZNF LD5-1 (U57796) ³⁸	ND
B123a ¹ (U77510)	196			100% identical to human histone H2B.1 (M60751) ^{29, 33} 92% identity with CpG island (HS187F7F) ²⁸	ND

Table 3 Continued

Clone ⁽¹⁾ (GenBank Accession No.)	Size (bp)	EST match ⁽²⁾	Exten seq. ⁽³⁾ (bp)	Homologies and remarks	Northern analysis
Y247a ¹ (U77511)	219			100% identical to histone H2A.1b (L19778). 93% identity with CpG island, clone 21g2 (Z65161) ²⁸	ND
G81a ¹ (U77513)	341	No		The sequence between 120 and 206 is duplicated between 255 to 341. Similar to G9a protein (X69838) ²⁵ but with a different exon organisation. It contains one ORF of 113 aa in which a stretch of 80 aa has 93% identity to G9a protein. Another ORF of 113 aa is orientated in the opposite direction and has 90% identity over 20 aa to G9a protein	ND
G60a ¹ (U77514)	96	c-1hd06 ⁽⁴⁾	1904	Both clones are novel. The sequences in the database of the 25 ESTs positive with G60 and G47 overlap and enabled the construction of a consensus sequence of 1904 bp. The best ORF is 311 aa long which is 38% identical and 48% homologous, when conservative substitution are included, to <i>Vibrio cholerae</i> neuraminidase (M83562) over 155 aa	No detectable bands on northern analysis
G47a ¹ (U77515)	183				
Y245a ² (U77516)	156	No		Novel. Similarity at a nucleotide level (60%) with zincfinger domains. One ORF of 52 aa	1 transcript of 3.1 kb
R1b ^b (U81872)	184	No		Novel at the nucleotide level. 87% identical to 5 μ mus musculus EST (clone 699468). At the amino acid level, it contains two possible ORFs of which one is 43% identical and 68% homologous, when conservative substitution are included, to <i>Caenorhabditis</i> <i>elegans</i> protein (1072187)	1 transcript of 2.5 kb
R23b ^b (U77517)	79	No		Novel. Three possible ORFs of 26 aa	No detectable bands on northern analysis
R21b ^b (U77518)	131	No		Novel. Two possible ORFs of 43 aa	No detectable bands on northern analysis
R10b ^b (U77519)	139	No		Novel. One ORF of 46 aa	No detectable bands on northern analysis
J13b ^b (U77520)	391			100% identical to human heterogeneous nuclear ribonucleoprotein hnRND H1 (L22009)	ND
J35b ^b (U77521)	149	No		Novel. One ORF of 49 aa	1 transcript of 2.4 kb
B296a ¹ (U77522)	447	No		Novel. Contains a LINE 1 repeat. The best ORF is 53 aa long	No detectable bands on northern analysis
R19b ^b (U77523)	88	No		80% identical at nucleotide level with human P5-1g (L06175) ²⁴ It contains 3 possible ORFs of 29 aa of which one is 50% identical to human P5-1g (L06175) while the other two do not have significant homologies	ND
D1c ^c (U81873)	600	244350 hfbdz16		Novel. Its best ORF is 64 aa long. Sequence of cosmid A186 with SP6	ND
B256a ¹ (U77524)	632	No		Novel. Alu repeat from 600 to 640. No ORFs	No detectable bands on northern analysis

Table 3 *Continued*

<i>Clone</i> ⁽¹⁾ (<i>GenBank</i> <i>Accession</i> <i>No.</i>)	<i>Size</i> (<i>bp</i>)	<i>EST</i> <i>match</i> ⁽²⁾	<i>Exten</i> <i>seq.</i> ⁽³⁾ (<i>bp</i>)	<i>Homologies and remarks</i>	<i>Northern</i> <i>analysis</i>
G32 ^{a1} (U77525)	296	No		Novel. One ORF of 98 aa	3 transcripts of 2.7 kb, 1.7 kb, and 1 kb
D2 ^c (U81874)	700	No		100% identical to human tRNA-Val, variant CAC (X17152) Genomic sequence of cosmid B1253 with T7	ND
G96 ^{a1} (U77526)	285	No		Novel. The best ORF is 71 aa long	1 transcript of 2.3 kb
G97 ^{a1} (U77529)	188	No		Novel at nucleotide level. One ORF of 62 aa 45% identical and 58% homologous, when conservative substitution are included, to putative E5 gene product of <i>human papilloma virus</i> (X77858)	No detectable bands on northern analysis
J23 ^b (U77531)	147	No		Novel. Two ORFs of 48 aa	3 transcripts of 6.8 kb, 3.7 kb, and 2.0 kb
B14 ^{a1} (U77527)	592	No		Novel. 100% identical to fetal brain cDNA d12_1g (Z70706) on 5q13 ²⁶ No ORFs	1 transcript of 6.2 kb
D3 ^c (U81875)	700	286057 c12283		Novel. 100% identical to cDNA g6_1 on 5q13 ²⁶ No ORFs Sequence of cosmid B148 with T7	ND
D6 ^c (U81876)	600	139751 csg3334		Novel. No ORFs. Sequence of cosmid O0714 with T7	ND
D7 ^c (U81877)	643	649509 EST5463 1		Novel. 100% identical to cDNA d8_2g on 5q13 ²⁶ The best ORF is 68 aa long. Sequence of cosmid L0454 with SP6	ND
D8 ^c (U81878)	550	277747		Novel. The best ORF is 101 aa long. Sequence of cosmid M2452 with SP6	ND
G59 ^{a1} (U77530)	68	c-3db10 37911 203790 26920 121676 22716 b417	458	Novel. One ORF of 22 aa. Matches the 3' end of 7 ESTs. The 5' sequence of these 7 ESTs overlap building a contig of 865 bp, that is novel sequences	1 transcript of 2 kb

A summary of the 42 cDNA fragments is listed. They are ordered from centromere to telomere as in Figure 6.2. ND=northern analysis was not done; aa=amino acids; ORF=open reading frames.

(1)Methods used to isolate the clone: ^a=cDNA selection (^{a1}=clone isolated from placental, thymus and HepG2 cDNA libraries; ^{a2}=clone isolated from normal liver, neonatal liver and small intestine cDNA libraries); ^b=exon trapping; ^c=direct genomic sequencing from cosmids.

(2)ESTs are reported only when 100% identical.

(3)ESTs sequences were used to build an extended sequence. The homologies and remarks are based on these extended sequences when possible.

(4)The longest EST is indicated. Another 24 ESTs were used to build the extended sequence, these were: 429617, 247072, 268118, 250895, 267563, 258539, 72850, 342122, 249213, 297863, 187644, 26525, 143207, 31259, 248870, 214184, 140279, 136248, 124712, 249627, 134569, 429381, 487925, 70401.

cosmid and were identical to ESTs that shared the same 3' end sequence. These ESTs were used to build an extended sequence of 1940 bp. The longest open reading frame of this gene was 311 amino acids long and showed homology to a neuraminidase from *Vibrio cholerae* (Table 3). Northern analysis was performed for G60 and G40 but did not detect a transcript in the tissues tested.

Nine novel cDNA fragments mapped at the distal end of YAC 950h11. Three of these were identical to cDNAs isolated from 5q13.1.²⁶ Analysis of this region has identified fragments with strong similarity to genes and STS from the SMA locus at 5q13.1, indicating that this distal region of 6p21.3 may be paralogous with 5q13.1²⁷ [now published].

All cDNA fragments detected only one major fragment when used as probes on Southern blots of both EcoRI digested YACs and BamHI/ScaI digested cosmids.

Discussion

We report here the construction of a cosmid contig and the isolation of 42 cDNA fragments mapping to a 2 Mb region distal to HLA-F. The cDNA fragments were found to be distributed in three clusters located around CpG islands as indicated by the restriction map of this region.⁶ Furthermore, four fragments, two in YAC 17ah2 (F20 and F21) and two around D6S105 (B123 and Y247), were very similar to CpG islands previously isolated using a methylated DNA binding column.²⁸

Combining the transcript map described here with three previously described transcript maps focused around HLA-A²⁹⁻³¹ and HFE provides a framework map for an extensive genomic region of approximately 6 Mb from HLA-A to HFE. Only one transcript identified here (B123) was found to be identical to a transcript in a previous report.²⁹ Analysis of all these transcription maps shows that the region distal of HLA-A contains many genes apparently not involved in immunity. In addition, we have been able to confirm previously reported features of this region. For example, the presence of two cDNA fragments homologous to histone genes supports the presence of a cluster of such genes mapped to 6p21-6p22 by FISH analysis³³ and subsequently confirmed to be located in this region.^{29,32} The isolation of the cDNA fragment identical to the tRNA-Val, variant CAC,³¹ is consistent with the description of a tRNA gene in YAC 753h12,³³ and

with the report of a cluster of tRNA genes localised by FISH analysis to 6p21.³⁴

Fragments Similar to MHC Genes

We identified 2 cDNA fragments similar to genes in the MHC region, R19 similar to the P5 multicopy gene family and G81 similar to a Class III gene. In the MHC region, other multicopy gene families besides the P5 group have been found,^{35,36} and some form repeated clusters of genes, for example a cluster comprising genes related to HLA, PERB11, and P5 families has been reported leading to the suggestion that clusters of genes within the MHC, and not only single genes, have been duplicated.³⁷ Thus if R19 is found to be a new gene of the P5 family, it is possible that the HLA and PERB11 genes will also be located here.

ZNF Gene Cluster

In total seven Krüppel-related zinc finger genes were localised in our transcript map, five of these were mapped in the present study (*ZNF184*, *LD5*, *F20*, *F44*, and *F36*) and two (*ZNF165* and *SRE-ZBP*) were previously reported in this region.^{29,38} Three of these, *ZNF184*, *ZNF165*, and *LD5* have been shown to be predominantly expressed in the testis^{7,38,39} and two, *F20* and *F36*, have strong homology with the murine *ZNF* genes, *Zfp-38*, *Zfp-92*, and *Zfp-35*, known to have a role in spermatogenesis.⁴⁰⁻⁴² The location and expression of the *ZNF* genes identified here correlate well with those from analyses of the syntenic region of the mouse which have shown the presence of a cluster of *ZNF* genes involved in embryogenesis and spermatogenesis.^{43,44}

SMA Homology Register

Nine novel cDNA fragments mapped to the most distal portion of our contig. Three of these were identical to cDNAs identified in the SMA region.²⁶ We previously identified cDNA fragments with more than 90% similarity to genes from the SMA region at 5q13.1 in this cluster indicating the presence of an interchromosomally duplicated region.²⁷ The SMA region contains many deletions, duplications, pseudogenes, and repetitive elements and is thus thought to be an unstable area of the human genome.⁴⁵ The presence of this putatively unstable gene-rich region at 6p21.3 may be of importance to the identification of disease genes mapped by linkage disequilibrium to this area.

Novel cDNA Fragments

Thirty-one novel putative exonic DNA fragments were present in the transcription map. They did not contain recognisable domains, and were not similar to the other novel sequences previously reported in this area.²⁹ Eight of these 31 novel fragments matched EST sequences in the database. The EST database is a powerful resource, facilitating the recognition and the isolation of larger cDNAs enabling the construction of extended sequences. By comparing clones G60 and G47 to the EST database, we isolated 25 ESTs and were able to construct an extended sequence of 1940 bp corresponding to a novel gene. Ten of the novel cDNA fragments isolated by cDNA selection and exon trapping identified a visible transcript in northern analysis, the remaining 16 did not produce a detectable signal. It is possible that cDNA fragments not giving a detectable transcript on northern analysis may represent a scarce transcript not present in the tissues studied. Indeed, G60 and G47, two clones matching 25 ESTs in the database, were negative on the northern analysis. However, the possibility that some novel fragments isolated are non-transcribed sequences cannot be excluded.

In conclusion, we constructed a cosmid contig and transcript map distal of HLA-F and identified 42 potential exonic fragments. This transcription map provides a useful resource for the identification of disease genes in linkage disequilibrium within this genomic region.

Acknowledgements

This work was supported by a Program Grant from the National Health and Medical Research Council of Australia. SG was supported by a QIMR Bancroft Centre postgraduate scholarship and the Associazione di Amici di Gastroenterologia del Granelli di Milano. The collaboration with Peter Little was supported by a Biomedical Research Collaboration Award from the Wellcome Trust. We are grateful to Dr Michael Lovett and Dr Richard del Mastro, University of Texas SW Medical Centre, Dallas, USA, for their advice and expertise in cDNA selection, and to Dr Tim Burn, Integrated Genetics Inc., USA, for help and advice in exon trapping. We thank Dr Marina Castellano and Dr Nicholas Hayward for their helpful suggestions and critical reading of the manuscript.

References

- Collins FS: Positional cloning moves from predilectional to traditional. *Nat Genet* 1995; **9**: 347–350.
- Cardon LR, Smith SD, Fulker DW, Kimberling WH, Pennington BF, DeFries JC: Quantitative trait locus for reading disability on chromosome 6. *Science* 1994; **266**: 277–279.
- Schwab SG, Albus M, Hallmayer J *et al*: Evaluation of a susceptibility gene for schizophrenia on chromosome 6p by multipoint affected sib-pair linkage analysis. *Nat Genet* 1995; **11**: 325–327.
- Rogers T, Hallmayer J, Hebert J *et al*: A linkage study of autism. Abstract for the 'Human Genetics Society of Australia' meeting, September 1996 Adelaide, Australia.
- Simon M, Pawlowsky Y, Bourel M, Fauchet R, Genetet B: Hemochromatosis idiopathique, Maladie associée à l'antigène tissulaire HL-A3? *Nouv Presse Med* 1975; **4**: 1432.
- Burt MJ, Smith DJ, Pyper WR, Powell LW, Jazwinska EC: A 4.5-megabase YAC contig and physical map over the hemochromatosis region. *Genomics* 1996; **33**: 153–158.
- Goldwurm S, Menzies ML, Banyer JL, Powell LW, Jazwinska EC: Identification of a novel Krüppel-related zinc finger gene mapping to 6p21.3. *Genomics* 1997; **40**: 486–489.
- Mercier B, Mura C, Feret C: Putting a hold on 'HLA-H'. *Nat Genet* 1997; **15**: 234.
- Bodmer JG, Parham P, Albert ED, Marsh SGE: Putting a hold on 'HLA-H'. *Nat Genet* 1997; **15**: 234–235.
- Feder JN, Gnirke A, Thomas W *et al*: A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet* 1996; **13**: 399–408.
- Jazwinska EC, Cullen LM, Busfield F *et al*: Haemochromatosis and HLA-H. *Nat Genet* 1996; **14**: 249–251.
- Jouanolle AM, Gandon G, Jezequel P *et al*: Haemochromatosis and HLA-H. *Nat Genet* 1996; **14**: 251–252.
- Zehentner G, Lehrach H: The reference library system – sharing biological material and experimental data. *Nature* 1994; **367**: 489–491.
- Dib C, Faure S, Fizames C *et al*: A comprehensive genetic map of the human genome based on 5264 microsatellites. *Nature* 1996; **380**: 152–154.
- Ivens AC, Little PRF: Cosmid clones and their application to genome studies. In: Glover DM, Hames BD (eds). *DNA cloning 3: Complex Genomes* IRL Press 1997, 1–47.
- Sulston JF, Mallet F, Staden R, Durbin R, Horsnell T, Coulson A: Image analysis of restriction enzyme fingerprint autoradiograms. *Comput Appl Biosci* 1998; **5**: 101–106.
- Lovett M: Direct selection of cDNAs using genomic contig. In: Seidman J, Seidman C (eds). *Current Protocols in Human Genetics* Wiley Interscience: pp 6.3.1–6.3.15.
- Buckler AJ, Chang DD, Graw SL *et al*: Exon amplification: A strategy to isolate mammalian genes based on RNA splicing. *Proc Natl Acad Sci USA* 1991; **88**: 4005–4009.
- Prober JM, Trainor GL, Dam RJ *et al*: A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 1987; **238**: 336–341.
- Sambrook J, Fritsch EF, Maniatis T: *Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 1989.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990; **215**: 403–410.

- 22 Gish W, States DJ: Identification of protein coding regions by database similarity search. *Nat Genet* 1993; **3**: 266-272.
- 23 Bellefroid EJ, Poncelet DA, Lecocq PJ, Revelant O, Martial JA: The evolutionarily conserved Krüppel-associated box domain defines a subfamily of eukaryotic multifingered proteins. *Proc Natl Acad Sci USA* 1991; **88**: 3608-3612.
- 24 Vernet C, Ribouchon MT, Chimini G, Jouanolle AM, Sidibe I, Pontarotti P: A novel coding sequence belonging to a new multicopy gene family mapping within the human MHC class I region. *Immunogenetics* 1993; **38**: 47-53.
- 25 Milner CM, Campbell RD: The *G9a* gene in the human major histocompatibility complex encodes a novel protein containing ankyrin-like repeats. *Biochem J* 1993; **290**: 811-818.
- 26 Morrison KE, Qureshi SJ, Anderson S *et al*: Novel transcribed sequences represented in the complex genomic region 5q13. *Bioch Bioph Acta* 1996; **1308**: 97-102.
- 27 Banyer J, Goldwurm S, Cullen L, van der Griend B, Zournazi A, Smit D, Powell L, Jazwinska E: The Spinal Muscular Atrophy gene region at 5q13.1 has a paralogous chromosomal region on chromosome 6 at 6p21.3.
- 28 Cross SH, Charlton JA, Nan X, Bird AP: Purification of CpG islands using a methylated DNA binding column. *Nature* 1994; **6**: 236-244.
- 29 Gruen JR, Nalabolu SR, Chu TW *et al*: A transcription map of the major histocompatibility complex (MHC) class I region. *Genomics* 1996; **36**: 70-85.
- 30 Amadou C, Ribouchon MT, Mattei MG *et al*: Localization of new genes and markers to the distal part of the human major histocompatibility complex (MHC) region and comparison with the mouse: New insights into the evolution of mammalian genomes. *Genomics* 1995; **26**: 9-20.
- 31 Totaro A, Rommens JM, Grifa A *et al*: Hereditary hemochromatosis: Generation of a transcription map within a refined and extended map of the HLA Class I region. *Genomics* 1996; **31**: 319-326.
- 32 Ruddy DA, Kronmal GS, Lee VK *et al*: A 1.1-Mb transcript map of the hereditary hemochromatosis locus. *Genome Res* 1997; **7**: 441-456.
- 33 Albig W, Drabent B, Kunz J, Kalff-Suske M, Grzeschick KH, Doenecke D: All known human H1 histone genes except the H1(0) gene are clustered on chromosome 6. *Genomics* 1993; **16**: 649-654.
- 34 Buckland RA, Maule JC, Sealey PG: A cluster of transfer RNA genes (TRM1, TRR3, and TRAN) on the short arm of human chromosome 6. *Genomics* 1996; **35**: 164-171.
- 35 Pichon L, Hampe A, Giffon T, Carn G, Legall JY, David V: A new non-HLA multigene family associated with PERB11 family within the MHC class I region. *Immunogenetics* 1996; **44**: 259-267.
- 36 Venditti CP, Harris JM, Geraghty DE, Chorney MJ: Mapping and characterisation of non HLA multigene assemblages in the human MHC class I region. *Genomics* 1994; **22**: 257-266.
- 37 Leelayuwat C, Pinelli M, Dawkins RL: Clustering of diverse replicated sequences in the MHC. Evidence for en bloc duplication. *J Immunol* 1995; **155**: 692-698.
- 38 Beutler ET, Gelbart T, West C, Kuhl W, Lee P: A strategy for cloning the hereditary hemochromatosis gene. *Blood Cells Mol Dis* 1995; **21**: 207-216.
- 39 Tisrosvoutis KN, Divane A, Jones M, Affara NA: Characterisation of a novel zinc finger gene (ZNF165) mapping to 6p21 that is expressed specifically in testis. *Genomics* 1995; **28**: 485-490.
- 40 Chowdhury K, Goulding M, Walther C, Imai K, Fickenscher H: The ubiquitous transactivator Zfp-38 is up-regulated during spermatogenesis with differential transcription. *Mech Dev* 1992; **39**: 129-142.
- 41 Noce T, Fujiwara Y, Sezaki M, Fujimoto H, Higashinakagawa T: Expression of a mouse zinc finger protein gene in both spermatocytes and oocytes during meiosis. *Dev Biol* 1992; **153**: 356-367.
- 42 Cunliffe VP, Koopman P, McLaren A, Trowsdale J: A mouse zinc finger gene which is transiently expressed during spermatogenesis. *EMBO J* 1990; **9**: 197-205.
- 43 Kostyu DD: HLA: fertile territory for developmental genes? *Crit Rev Immunol* 1994; **14**: 29-59.
- 44 Crossley PH, Little PF: A cluster of related zinc finger protein genes is deleted in the mouse embryonic lethal mutation tw18. *Proc Natl Acad Sci USA* 1991; **88**: 7923-7927.
- 45 Lewin B: Genes for SMA: multum in parvo. *Cell* 1995; **80**: 1-5.