Hum Genet (2002) 110 : 111–121
DOI 10.1007/s00439-001-0646-6

ORIGINAL INVESTIGATION

Ester Rozenblum · Pia Vahteristo · Therese Sandberg · Jon Thor Bergthorsson · Kirsi Syrjakoski Don Weaver · Karin Haraldsson · Hrefna Kristin Johannsdottir · Paula Vehmanen · Savita Nigam Natalie Golberger · Christiane Robbins · Evgenia Pak · Amalia Dutra · Elizabeth Gillander Dietrich A. Stephan · Joan Bailey-Wilson · Suh-Hang Hank Juo · Tommi Kainu · Adalgeir Arason Rosa Bjork Barkardottir · Heli Nevanlinna · Ake Borg · Olli-P. Kallioniemi

A genomic map of a 6-Mb region at 13q21-q22 implicated in cancer development: identification and characterization of candidate genes

Received: 3 August 2001 / Accepted: 9 October 2001 / Published online: 14 December 2001

Abstract Chromosomal region 13q21-q22 harbors a putative breast cancer susceptibility gene and has been implicated as a common site for somatic deletions in a variety of malignant tumors. We have built a complete physical clone contig for a region between D13S1308 and AFM220YE9 based on 18 yeast artificial chromosome and 81 bacterial artificial chromosome (BAC) clones linked together by 22 genetic markers and 61 other sequence tagged sites. Combining data from 47 sequenced BACs (as of June 2001), we have assembled *in silico* an integrated 5.7-Mb genomic map with 90% sequence coverage. This area contains eight known genes, two hypothetical proteins, 24 additional Unigene clusters, and approximately 100 predicted genes and exons. We have determined the cDNA and genomic sequence, and tissue expression profiles for the KIAA1008 protein (homologous to the yeast mitotic control protein dis3+), KLF12 (AP-2 repressor), progesterone induced blocking factor 1, zinc finger transcription factor KLF5, and LIM domain only-7, and for the hypothetical proteins FLJ22624 and FLJ21869. Mutation screening of the five known genes in 19 breast cancer families has revealed numerous polymorphisms, but no deleterious mutations. These data provide a basis and resources for further analyses of this chromosomal region in the development of cancer.

E. Rozenblum, P. Vahteristo, T. Sandberg, J. T. Bergthorsson, and K. Syrjakoski contributed equally to this work

E. Rozenblum · D. Weaver · P. Vehmanen · S. Nigam
N. Golberger · C. Robbins · E. Gillander · D.A. Stephan
T. Kainu · O.-P. Kallioniemi (✉)
Cancer Genetics Branch,
National Human Genome Research Institute,
National Institutes of Health, Bethesda, MD 20892, USA
e-mail: okalli@nhgri.nih.gov,
Tel.: +1-301-4352896, Fax: +1-301-4027957

J. Bailey-Wilson · S.-H.H. Juo
Inherited Disease Research Branch,
National Human Genome Research Institute,
National Institutes of Health, Bethesda, MD 20892, USA

E. Pak · A. Dutra
Genetic Diseases Research Branch,
National Human Genome Research Institute,
National Institutes of Health, Bethesda, MD 20892, USA

K. Syrjakoski
Laboratory of Cancer Genetics,
Tampere University and University Hospital,
33520 Tampere, Finland

P. Vahteristo · H. Nevanlinna
Obstetrics and Gynecology, Helsinki University Central Hospital,
00029 Helsinki, Finland

T. Sandberg · K. Haraldsson · A. Borg
Department of Oncology, University Hospital,
221 85 Lund, Sweden

J.T. Bergthorsson · H.K. Johannsdottir · A. Arason
R.B. Barkardottir
Laboratory of Cell Biology, House 14, Department of Pathology,
University Hospital of Iceland, 101 Reykjavik, Iceland

Introduction

We recently reported on a putative breast cancer candidate predisposition locus at 13q21-q22 based on linkage analysis of 77 Finnish, Icelandic, and Swedish families that were negative for mutations in the BRCA1 and BRCA2 genes (Kainu et al. 2000). A multipoint heterogeneity LOD score of 3.46 was obtained for the interval between the genetic markers D13S1296 and D13S1308. Comparative genomic hybridization and loss of heterozygosity (LOH) studies have suggested that 13q21 is also a common region of somatic loss in sporadic breast cancer (Larramendy et al. 2000; Wistuba et al. 1998), malignant fibrous histiocytoma (Larramendy et al. 1997; Mairal et al. 1999), and prostate cancer (Dong et al. 2000). In prostate cancer, the minimal deletion area spans up to 3 Mb between markers D13S275 and D13S162 partly overlapping with the site implicated in breast cancer predisposition.

Because of several clues suggesting involvement of the 13q21-q22 area in cancer predisposition and progression, we undertook physical and transcript mapping of this locus, by using a combination of traditional positional cloning methods and new approaches based on the human genome draft sequence. The goal was to generate resour-

ces for cloning potential cancer-associated genes, focusing particularly at a region between markers D13S1302 and D13S162. The genes in the region were identified by analysis of genomic sequence and cDNA cloning. The exon-intron structure and expression pattern was determined for seven of the genes in the region: genes for KIAA1008 (homologous to the yeast mitotic control protein DIS3), PIBF1 (progesterone induced blocking factor 1), KFL5 (Kruppel-like factor 5, also known as IKLF), KLF12 (Kruppel-like factor 12, also known as AP-2 repressor), LMO7 (LIM domain only-7), hypothetical protein FLJ21869, and hypothetical protein FLJ22624. Mutation screening of the five known genes was performed in 19 breast cancer families showing linkage to 13q21-q22 in order to investigate whether the genes are involved in breast cancer predisposition.

Materials and methods

Yeast artificial chromosome and bacterial artificial chromosome map

Eighteen mega yeast artificial chromosomes (YACs) were selected from the Centre d'Étude du Polymorphisme Humain (CEPH) human YAC library by using genetic markers present in the interval between markers D13S1308 and D13S162 (Whitehead Institute for Biomedical Research/MIT Center for Genome Research, Human Physical Mapping Project: http://www-genome.wi.mit.edu/). Genome Systems (St. Louis, Mo.) human bacterial artificial chromosome (BAC) library releases I and II were screened by polymerase chain reaction (PCR), first with the existing genetic markers and then with newly defined sequence tagged sites (STSs). BAC DNA was isolated following the Qiagen (Santa Clara, Calif.) Plasmid Midi Kit (catalog no. 12143 or 12145). BAC insert ends were sequenced as described previously (Carpten et al. 2000) to generate additional STSs and to identify flanking clones by further PCR screening.

PCR for STS mapping

PCR was carried out in 25-µl volumes in 96-well plates, in 1× PCR buffer II (Perkin Elmer), 2.25 mM $MgCl_2$ (Perkin Elmer), 0.2 mM dNTPs (Gibco-BRL), 0.8 µM of each of the primers, and 0.114 U/µl AmpliTaq Gold Polymerase (Perkin Elmer) with 60 ng template in 2 µl. The following PCR program was used: one initial denaturation step at 94°C for 10 min, 34 cycles of 94°C for 30 s, 50°C or 55°C for 30 s, and 72°C for 1 min, followed by a final extension at 72°C for 10 min. PCRs were performed in an MJ Research thermocycler, PTC-225, and the products were checked in a 2% agarose gel.

Visual mapping by fluorescence in situ hybridization

Metaphase fluorescence in situ hybridization (FISH) was performed for low-resolution physical mapping and to identify chimeric clones. Metaphases were prepared from stimulated peripheral blood lymphocytes by standard techniques (Lundsteen and Lind 1985). For high-resolution mapping, DNA fibers were obtained by incubation of the Epstein-Barr virus-immortalized lymphoblastoid cells in lysis buffer (2 M $MgCl_2$, 25 mM TRIS-HCl, 1% Triton X) for 30 min at room temperature. Microscope slides were dipped into the lysis buffer, then taken out vertically, air dried for 5 min, and fixed in methanol:acetic acid (3:1) for 20 min. FISH was performed with DNA labeled with spectrum orange, spectrum green (Vysis, Downers Grove, Ill.) and Cy5 (Life Science Boston, Mass.) by nick translation, essentially as described by Pinkel et al. (1986) and Lichter et al. (1988).

Transcript identification

All the expressed sequence tags (ESTs) and genes of the area between the genetic markers D13S1260 and D13S160 were screened by PCR against all possible YACs and BACs (GeneMap '98, http://www.ncbi.nlm.nih.gov/genemap98; GeneMap'99, http://www.ncbi.nlm.nih.gov/genemap99). Additional genes in the area were identified from the analysis of the genomic sequence data that became available later (see below).

Transcript extension

cDNA clones representing the genes of interest were obtained from Research Genetics (Huntsville, Ala.) and ATCC (Rockville, Md.) and further extended by the Rapid Amplification of cDNA Ends (Marathon RACE; Clontech, Palo Alto, Calif., catalog no. K1802–1) and by various methods of cDNA library screening (Rapid Screen cDNA libraries, Origene, Rockville, Md.; HUCL, Stratagene, La Jolla, Calif., catalog no. 937811; GeneTrapper cDNA Positive Selection System, Life Technologies, Rockville, Md., catalog no. 10356–020). Probes were designed from these sequences for Northern hybridizations in order to determine gene expression and the size of the transcripts.

Northern hybridization

Northern hybridizations were performed on multitissue blots supplied by Clontech (catalog nos. 7760–1, 7759–1, and 7767–1) with Hybrisol I (Intergen, Purchase, N.Y., catalog no. 54040) at 42°C. Washes were performed according to Sambrook et al. (1989) with a final stringency of 0.1×SSC (1×SSC=150 mM NaCl, 15 mM sodium citrate, pH 7.0) at 63°C.

Analysis of genomic sequence

The availability of the genomic sequence increased continuously as the project advanced, necessitating frequent comparisons between publicly available genomic sequences in the GenBank (http://www.ncbi.nlm.nih.gov:80/entrez/viewer.cgi) and at the Sanger Centre (http://www.sanger.ac.uk/) with cDNA sequences arising from our own cDNA cloning efforts and those from public EST efforts. We used the WebBLAST program for this purpose (http://www.ncbi.nlm.nih.gov/blast/blast.cgi; Ferlanti et al. 1999). Repetitive and vector sequences were first removed, and the edited sequences were used to search the nr, dbEST, and swissprot databases of NCBI with WebBLAST. Matches to dbEST were used to search the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene/).

In order to access all the available genomic sequence, all EST- and STS-based markers were periodically searched against the genomic sequence databases (http://www.ncbi.nlm.nih.gov/genome/sts/). Any new sequences were then BLASTed against the sequence databases in order to retrieve more overlapping BACs. Sequencher program was utilized to assemble the genomic contigs that were then analyzed with the GeneMachine program (http://genome.nhgri.nih.gov/genemachine/; Makalowska et al. 1999) to find putative expressed sequences. GeneMachine has a graphical output that places BLAST results from the nr and dbEST databases in their base-pair position, along the sequence that is submitted. In addition, it provides the base-pair positions of predicted exons and genes from the programs MZEF, Grail, Fgenes, and GenScan.

Mutation analysis

Conformation-sensitive gel electrophoresis (CSGE; Ganguly et al. 1993), as modified by Couch et al. (1996), and direct sequencing

(Carpten et al. 2000) were performed from both genomic DNA and cDNA extracted from lymphoblastoid cell lines established from patients or from whole blood sampled from patients. The sensitivity of the CSGE method is almost 90% (Ganguly et al. 1993). DNA isolation was performed by using phenol-chloroform-isoamyl alcohol (25:24:1; Sambrook et al. 1989), and RNA isolation was carried out with a Qiagen Rneasy Max Kit (catalog no. 75162). cDNA and genomic DNA derived from lymphoblastoid cell lines and from whole blood from 19 breast cancer families from Finland, Sweden, and Iceland were analyzed for mutations at each institution. The number of affected individuals per family range between 3 and 11, and the number of affected individuals tested per family was between one and three.

Results

Physical map of the 13q21-q22 critical region

A complete physical clone map was assembled for the 13q21-q22 region between markers D13S1308 and AFM220YE9 by using 18 CEPH human megaYACs from the WC13.3 contig (http://www-genome.wi.mit.edu/) and 35 BACs (Genome Systems). The map was confirmed by STS-PCR with 22 genetic markers, 12 EST-derived STSs, and 61 novel STS derived from BAC end sequencing. In addition, all clones were checked for location and chimerism by metaphase-FISH analysis and 25 BAC clones also by dual color fiber-FISH to ascertain clone order over short distances and to ensure closure of the contigs.

As genomic sequences became available, we identified 47 sequenced BACs from the GenBank and assembled them *in silico* into six genomic contigs and then integrated them into our existing physical clone map (Fig. 1, Table 1). The map spans approximately 6 cM in the Genethon genetic map (http://www.cephb.fr/bio/ceph-genethon-map.html) corresponding to about 5.7 Mb genomic DNA. Each of the six genomic contigs is anchored to the physical map by a genetic marker or EST and positioned relative to each other such that the size of the gaps correspond to the estimated size of the BACs that bridge them. Because of the changing nature of the incomplete public genomic maps (NCBI MapViewer and Human Genome Project Working Draft at UCSC), the overall base-pair position of our map should be considered as provisional. Although we do not have 100% sequence coverage, the BAC contig is complete. The gaps between contigs 2 and 3, and between contigs 3 and 4 were bridged by our BACs. The other gaps in the genomic sequence were bridged by unfinished sequenced BACs from the Human Genome Project. Altogether, only an estimated 400 kb of the 5.7 Mb of sequence between markers D13S1308 and AFM220YE9 remains to be sequenced.

Transcript identification

Based on recombinants in some breast cancer families, we considered the region from D13S1302 to D13S162 (3.3 Mb) to be the most important. We identified four known genes within this region: the genes for KIAA1008 protein (KIAA1008), KLF12, PIBF1, and KLF5 (Fig. 1, Table 2). In initial mapping data, LMO7 was located proximal to marker D13S162 but later was verified to be located distal to this marker. This 3.3-Mb region also contained 18 Unigene clusters (Unigene Build 135): Hs.28465, Hs.166425. Hs.55418, Hs.253076, Hs.212235, Hs.161159, Hs.242720, Hs.145953, Hs.332545, Hs.269658, Hs.271776. Hs.312503, Hs.133319, Hs.164467, Hs.171085, Hs.148515, Hs.97408, and Hs.278118.

Three other genes were located immediately outside or at the boundary of the core region: dachshund (Dachshund) homolog (DACH), KIAA0603 gene product (KIAA00603) and ubiquitin carboxyl terminal esterase L3 (ubiquitin thiolesterase; UCHL3). In addition, the broad interval covered by the physical map contained eight additional Unigene clusters: Hs.27657, Hs.191614, Hs.59590, Hs.78961, Hs.332782, Hs.162240, Hs.190331 and Hs.101539.

Transcript characterization

Of the multiple known genes and various Unigene clusters mentioned above, the transcripts given below were further studied. Table 2 contains a summary of these transcripts.

Transcript of KIAA1008

KIAA1008 has a 40% identity with *Saccharomyces cerevisiae* mitotic control protein dis3+, a component of the yeast exosome (Shiomi et al. 1998). By Northern analysis, we found it to be expressed in spleen, thymus, prostate, testis, ovary, small intestine, colon and peripheral blood leukocytes (Fig. 2). We extended the mRNA sequence to 7.3 kb (GenBank AF330044) and determined its intron-exon structure (Fig. 3). The gene spans 26.5 kb of genomic sequence and is comprised of 21 exons. It encodes a protein of 958 amino acids. It is transcribed divergently from PIBF1 whose 5' end is separated from it by only 200 bp.

Transcript of PIBF1

By Northern analysis, we found PIBF1 to be expressed most prominently in the testis and in spleen, thymus, prostate, ovary, small intestine and colon (Fig. 2). We extended the mRNA sequence to 3.2 kb (GenBank AF33046), from which we determined the intron-exon structure (Fig. 3). The gene spans 234.4 kb of genomic sequence and is comprised of 18 exons. It encodes a protein of 758 amino acids.

Transcript of KLF5

KLF5 (IKLF, CKLF, BTEB2) is a transactivator that binds the GC-box and epidermal growth factor (EGF) response element. Our Northern analysis (Fig. 2) and mRNA exten-

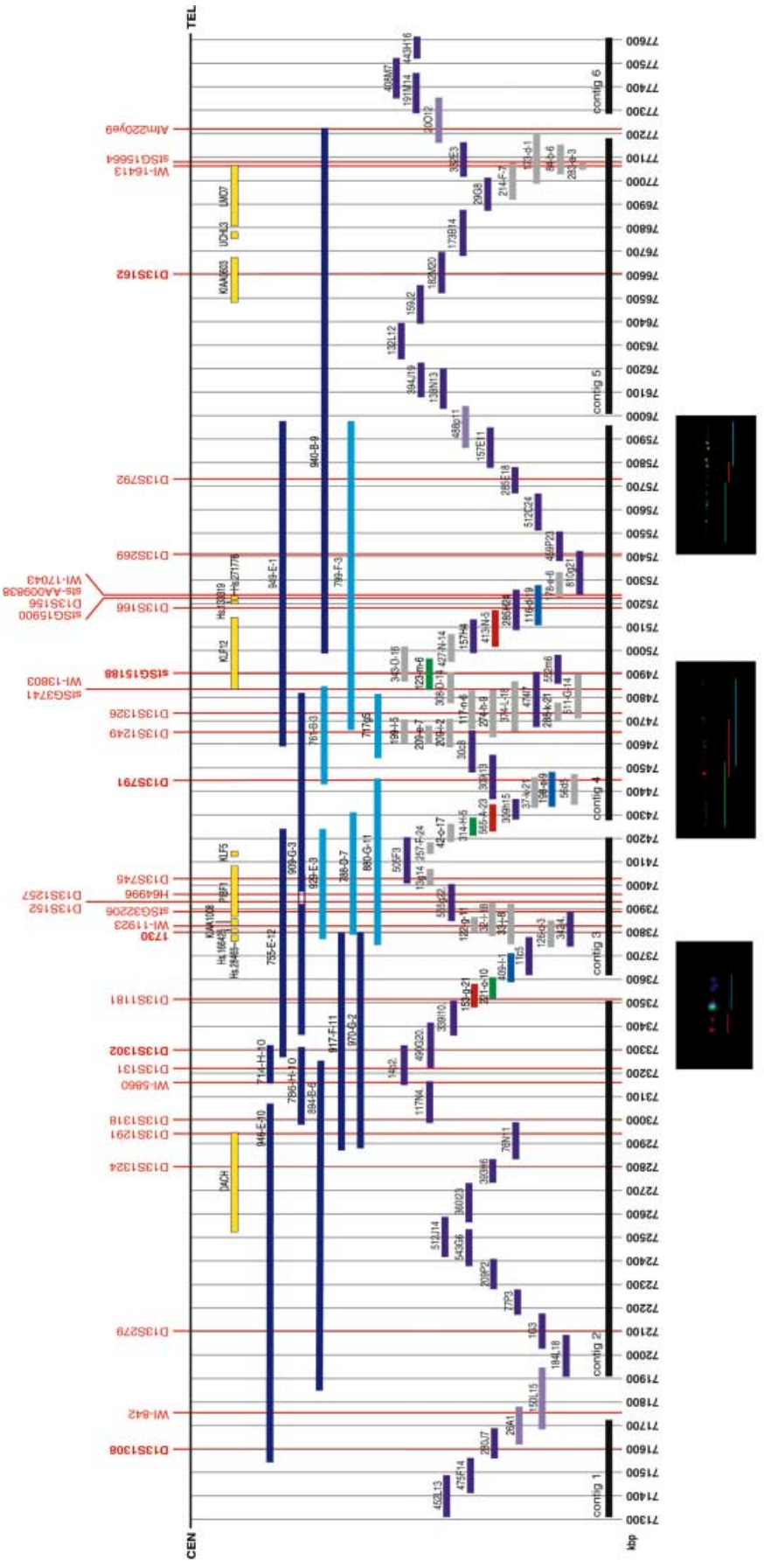

Fig. 1 Genomic contig of 13q21-q22 region. The corresponding BAC map indicates these BACs. The base pair positions in the map are provisional. Reference markers for each of the six contigs (*bold*) were calculated relative to the base pair position of marker D13S1302 in the October 7, 2000 freeze of the Human Genome Project Working Draft of the UCSCs map. Microscopic images represent fiber-FISH with three BAC clones labeled in three different colors

Table 1 Novel STSs

Marker name	GenBank accession no.	Forward primer sequence I tou	Reward primer sequence	Product size (bp)
32i18F	G67302	CAGGGAAAGCGAACTACAGAAATC	AAGCATAAGGGGGAGCTTGATG	158
32i18R	G67303	CTTGAAGGATAACACATGGGC	CCAGTCAACCATCTCAACACAAAC	138
33j8F	G67304	GGAAGAGAGGGCTAGAAAAACACAG	GTGAGAAAGAAGGTGGGGAAATG	158
33j8R	G67305	GCCGAAGATAGTGAGGGAATG	GCTGTTGTAAAGATGCCAACTG	191
122g11F	G67306	TTCATCACCTGGCTTCCAAGAC	CAAGGAGTGCGTACAGACAGAAGAG	132
122g11R	G67363	CGGTAATTTTCCTTTCTTGCCC	GAACCAGCTTTTGCTTTCCCC	370
116d19F	G67307	GGGAAAGTAAGTGGCCTGTTG	CTAGATATGCCCAGTTCTACAAATC	204
116d19R	G67308	TTTCACTTGCTTTGACATGGTTTC	TTTAGCCCACTGACTCTTTTTTGC	115
178e6R	G67364	CCTCACAGAAACCTTCTGGG	AACTAACCTTGCCATTTC	104
283a3F	G67309	GGGCATTTTTAAGACCTGGAGAAG	CAGAGAAAGACAGGCATTTCCTTG	149
283a3R	G67310	GGGAAGTGGCTGATAAATGC	CAGACCATGCTCCTGTGAAG	103
84b6F	G67311	CTATGTTGGTCAGGCTGGTCTTG	CCCAGTCCATTAGTTCTCTGCTTTG	160
84b6R	G67312	GAAACAGAGCTGAAACACTGGTCTC	GCTCATAAGCCCCAAACCAATAG	249
173d1F	G67313	CGCAGGCACTTAAACATGGTATC	TGAAGGGGAAGCTCACAGTATCAC	177
173d1R	G67314	AAACAGTTTGGCAGTTCTCAACAAG	TCCCACTCATCAGTTGCTGGAC	176
123m6F	G67315	GCCACCCTTGAGATGTTGGAAC	ACTCAGGAGAGCCCCAGAATTG	231
123m6R	G67316	GCCACCATCTCAAAAATTGTCCAC	ACACCTACGAGGTAGCAAGCAGTC	216
42o17F	G67317	ACCTTTCCTGGGGTCTGAAC	CAGTGGTTTTTTGAGTCTCTATTTG	215
42o17R	G67318	TGGAGTTTGAACTGATGCAGTCTG	TGAAAGTGGGCAACCAGTATTCTC	375
153g21F	G67321	CTATCTTACATTCACCCGCTTC	CTTGTTAAACTCCAGGCATACG	106
153g21R	G67322	TTCAGGTCAGGTGCAGCAGTTC	CGAGGCAACAAGCATTAGGAAAG	112
221o10F	G67323	GGAAGGAGAGCACTGAGCTAAGTG	TTCAGGGGCTTTGCACAATG	233
221o10R	G67324	CTTAGTTTCTGCAACTCGGAGGAC	TAAAGACAAGGCTGTGGGTTCAG	195
209i2F	G67325	AAACGTGCCCATTTAAGCTGG	GCTAGGACAGGTCCCAACTTTTG	137
209i2R	G67326	TTCCTGACACCCTGGCTTTTC	GAAGATTGACATCATTTGGCAATTC	118
209e7F	G67327	GGAAAAGGGATTTCAGTCCAGAAG	GCTTTGATACTGCCAGTGTGTCATC	309
209e7R	G67328	AGGGGGTAAGGATCTCCTGACTTG	CACTGGGATGACCTTTGATTCTTG	169
199l5F	G67365	CCCATTTAAGCTGGTGAAATATC	CCAGGTCCCAACTTTTGAAC	124
199l5R	G67329	TTAGGGGTAAGGATCTCCTGACTTG	TCCACTGGGATGACCTTTGATTC	172
37k21R	G67330	CTATGAAATGGAGGAGATGAAAAC	TCTCAAATGCTTAAAGGCTCTATC	140
198e9F	G67331	GTACATGCCTGTGGTCTCAG	GGAGACAGGGTCTCACTTTG	135
198e9R	G67332	GGAACACAGCTACATCTTGCTGC	CCAGAACCTGGGAAGTGACTAATG	174
117n6F	G67333	GCACCACCAGAACTGGCTAAAAC	CAGAGGGCATGTTATTAGGGATCTG	142
117n6R	G67334	TCCTAATGTTCTCTGTCACAAGAAG	TTCTACTGCTTTGGGATCTATATTG	107
274h9F	G67335	TGACAGGGAAAAATCTCCAAG	AAACATCCATCTCAAAATG	188
274h9R	G67336	CACAAGCCTCGATGAGCTATG	AAGGTGAAGTAAAAGGTTAAGGAAG	264
288k21F	G67337	GCACACTATGCCTCCGTGTTAATG	GCTTTGGGAAGCCACGTATTTG	128
288k21R	G67338	TGGTGTTCGCTGGAACCTTTG	ACTAATGGCAAATCTGGCATCG	273
13g14F	G67339	GGAAGTGGAAAACATGAGTAAGCAG	TCAGTTGACTAGCCTGGACACAAG	113
56d5F	G67340	CAAAGAATAAGCTGCTTCTGGAC	TCTTAGTTACTGGGTATGTCCTAGC	141
56d5R	G67341	TTGGTTTTGCTGGGTCATCTTC	CATTAGGGAGAGTGGGATAAGGTTG	126
126o3F	G67362	GCCGAGGTTCTTTTCCTTG	CCTCCAGGCACTTACAATCTAAG	115
126o3R	G67342	GTGACTTGCTGATCGTGCATTG	CATTTGCCAGCCCATCTGTTC	258
308d14F	G67343	ATGAGACGGGCTTCACTGCATC	TTGACAAACTTTCTGTGCTGGTTG	158
308d14R	G67344	GTAATGCCTTACGTTTTAGAAAGTG	TGGTGTCTATTTTGACTGAACATC	127
374l18F	G67345	GCTCTTTTCATGCAGCTCCTAAATC	CCCATGCTCCATGTTTGTTTCTC	221
374l18R	G67346	CATTTAGGTCGGTGCCTGACAC	GAGATGTCCTTCACTGAAACTGGAG	249
511g14F	G67347	TCCTAATCTGTAAAATGGGGCTTG	TGCGTTTGGTAAATACTGTCCTTTG	109
511g14R	G67348	CCTGCCTCTATTCTTGCATACCAC	GCCATGAAGCAATGAGAAGTCAC	130
257f24F	G67349	TCTCCAAAATGACATGCTGTTC	CAAGAAAATTACCATCCCAGTG	112
257f24R	G67350	GAAATCATGGCTAATGCTCTAATC	GAAAAGGTTAAACCAAAACTACAGC	107
314h5F	G67351	CAGAGCCCATTCGCTATTAAAC	ATAAAGATTGGAGATGGTTACCTG	141
314h5R	G67352	ACTCGGGCTCAAATCTTTTC	TGGGAAAGAGGAATCCACAG	101
565a23R	G67353	TGCATTCTCCTCCAGACTTTG	GGTTTAGAGCAAAAGGGTATGTAG	159
413n5F	G67354	TTGTCCAAGTACAGGTATAAGGAAG	GAATTTAGGCAATCAAGTTGAC	163
413n5R	G67355	AAGACCTTGAGTGAGATGTGGTAG	AGAAAAAGCACGCAGTC	172
343d16F	G67356	ATTTAGCACTTCCTGAATGATGA	CACTGGGTAAAATGAGCTGAAAG	81
427n14F	G67357	CAGTTAAGGGGTTCAGAATCTAAG	CGCCATATTGTGTGAATCAATC	206
427n14R	G67358	CCTCCTTCAAAAAAGAGCAATC	AATCCCTCACTTCCAGAGACTAC	183
214f7F	G67359	GATTTGGCATTTAGAACTAGCTTC	GCTCAATATAATGTCTGGCTCAG	101
214f7R	G67360	AAAAACAAACATGCCTCCTTC	GCTCACAACCAAGTCTCCAATC	112

Table 2 Expressed sequences in the region between 13q markers D13S1302 and AFM220YE9

Name	GenBank accession no.	Unigene cluster[a]	Gene map 98/'99 marker	No. of ESTs in cluster	Base pair position
KIAA1008 protein	AF330044	Hs.323346	WI-11923	87	73851320–73824857
		Hs.179566	AA213647	45	
PIBF1	AF330046	Hs.43913	H64996	60	73851520–74085949
			stSG32206		
KLF5	AF132818	Hs.84728		167	74128290–74147500
KLF12	AF330041	Hs.23510	WI-13803	24	75144602–74835646
		Hs.291800		4	
		Hs.104492		9	
		Hs.151949	stSG3741	8	
		Hs.294018		68	
LMO7	AF330045	Hs.5978	WI-16413	186	76819212–77057403
		Hs.193380		19	
FLJ21869	AK025522	Hs.28465	stSG46189	10	73797125–73777818
FLJ22624	NM_024808	Hs.166425	1730	35	73797412–73816829
HBCL13.1	AF330043	Hs.271776	sts-AA009838	3	75242847–75240989
HBCL13.2	AF330042	Hs.133319		4	75277404–75275937

[a]Unigene Build 135

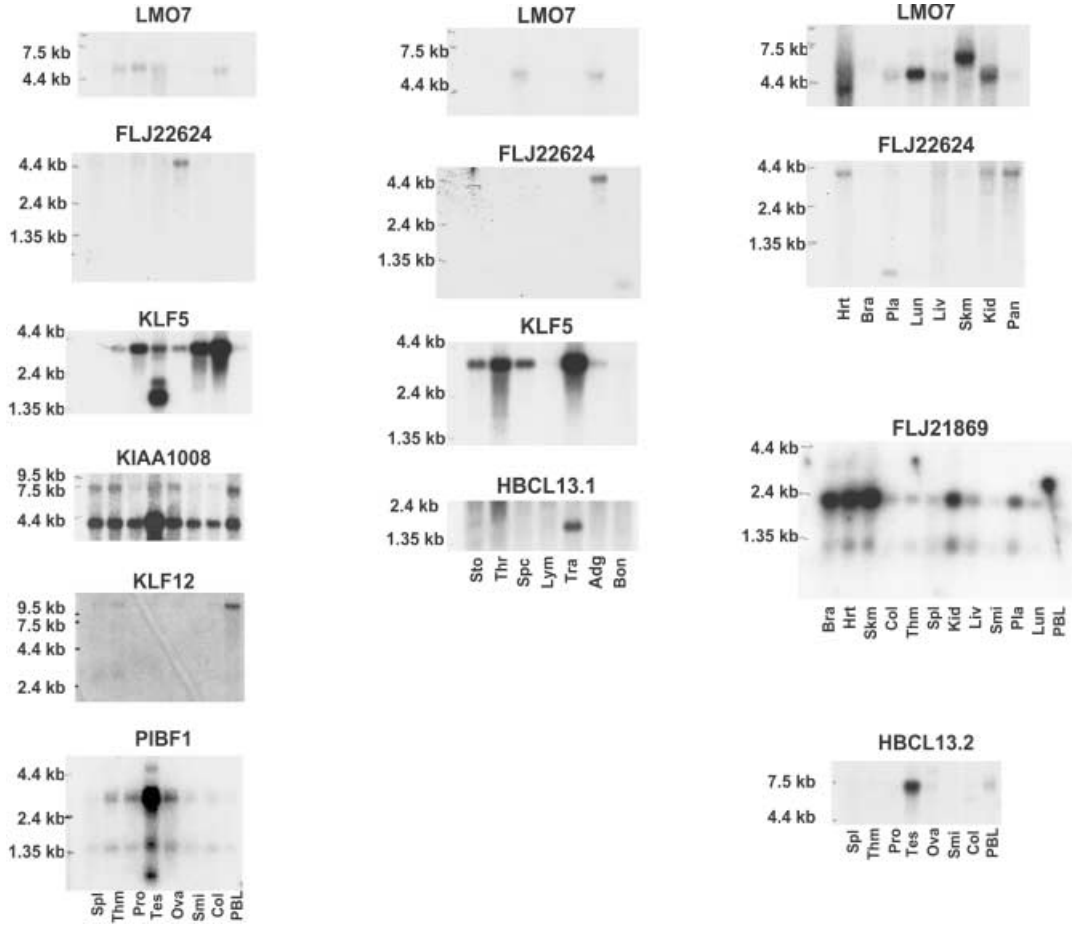

Fig. 2 Northern hybridizations to ascertain tissue distribution and size of the transcripts (*Adg* adrenal gland, *Bon* bone marrow, *Bra* brain, *Col* colon, *Hrt* heart, *Kid* kidney, *Liv* liver, *Lun* lung, *Lym* lymph node, *Ova* ovary, *Pan* pancreas, *PBL* peripheral blood leukocytes, *Pla* placenta, *Pro* prostate, *Skm* skeletal muscle, *Smi* small intestine, *Spc* spinal cord, *Spl* spleen, *Sto* stomach, *Tes* testis, *Thm* thymus, *Thr* thyroid, *Tra* Trachea)

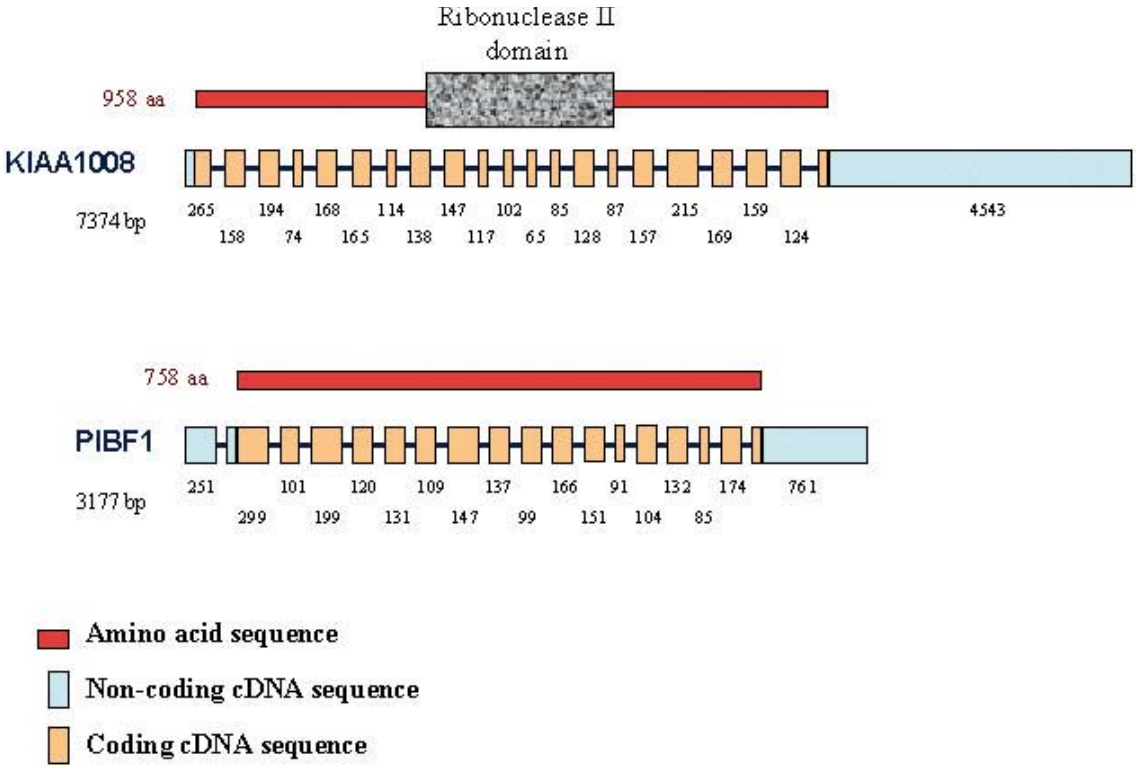

Fig. 3 Genomic structure and predicted protein structure of the KIAA1008 and PIBF1 genes. The exons and their sizes are indicated. The exons were determined by comparison of the extended full-length cDNA sequence with the HTGS database (NCBI). The sizes of the predicted amino acid sequence and protein domains that have a match in the BLOCKS database with an e-value <0.05 are marked and annotated

sion confirmed the previously published expression pattern and gene structure (Shi et al. 1999).

Transcript of KLF12

KLF12 (AP-2 repressor) binds to AP-2 transcription factor, which in turn regulates the expression of *c-erb-2* in human mammary carcinoma (Bosher et al. 1995). By Northern analysis, we found it to be expressed most prominently in peripheral blood leukocytes and more weakly in spleen, thymus, and colon (Fig. 2). We extended the mRNA sequence to 10.6 kb (GenBank AF330041). The gene spans 308.9 kb of genomic sequence and is composed of seven exons. It encodes a protein of 402 amino acids.

Transcript of LMO7

LMO7 contains a single cysteine-rich zinc-binding LIM domain involved in protein-protein interactions (Putilina et al. 1998). By Northern analysis, we found it to be expressed most prominently in lung, skeletal muscle and kidney, and in thymus, prostate, testis, colon, spinal cord, adrenal gland, placenta, and liver (Fig. 2). We extended the mRNA sequence to 7.2 kb (GenBank AF330045). The gene spans 238.2 kb of genomic sequence and is comprised of 27 exons. It encodes a protein of 1349 amino acids.

Transcript of hypothetical protein FLJ21869

Hs.28465 encodes a hypothetical protein, FLJ21869, with no homology to other known proteins. By Northern analysis, we found it to be expressed most prominently in brain, heart, and skeletal muscle, in kidney and placenta, and faintly in all tissues, as a 2.4-kb transcript (Fig. 2). Our mRNA extension confirmed the sequence in the Unigene database. The gene spans 19.1 kb of genomic sequence and is comprised of three exons. It encodes a short predicted protein of 82 amino acids. It is transcribed divergently from Hs.166425 whose 5' end is separated from it by only 287 bp.

Transcript of hypothetical protein FLJ22624

Hs.166425 encodes a hypothetical protein, FLJ22624, with no homology to known proteins. By Northern analysis, we found it to be expressed in adrenal gland, heart, ovary, pancreas, and kidney as a 4.4-kb transcript and in bone and placenta as a ~0.8-kb transcript (Fig. 2). Our mRNA extension confirmed the sequence in the Unigene database. The gene spans 19.2 kb of genomic sequence and is comprised of 11 exons. It encodes a predicted protein of 519 amino acids.

Transcript of HBCL13.1

By Northern analysis, we found HBCL13.1 (Table 2) to be expressed in trachea as a 1.8-kb transcript (Fig. 2) We extended its mRNA sequence to 1.8 kb (GenBank AF330043) but failed to uncover a probable coding sequence.

Transcript of HBCL13.2

By Northern analysis, we found HBCL13.2 (Table 2) to be expressed most prominently in testis and weakly in peripheral blood leukocytes as a 4.4-kb transcript (Fig. 2). We were only able to extend its mRNA sequence to 1.5 kb (GenBank AF330042) and failed to uncover a probable coding sequence.

The extended sequences for each mRNA were validated by RT-PCR. 3' ESTs matching these sequences revealed multiple possible polyadenylation sites. A list of these putative polyA sites and other clusters of ESTs are given for each of the known genes that were examined by Northern hybridization (Table 3).

Mutation analysis

Five of the genes, KIAA1008, KLF12, PIBF1, LMO7, and KLF5, and the available sequence from hypothetical protein FLJ22624 were screened for germline mutations in breast cancer families by using CSGE and direct sequencing. No frame shift or nonsense mutations were found. One of the variants is located in a splice acceptor site i.e. C34 (-3)T in the KLF12 gene (Table 4). This nucleotide change did not segregate with the disease in the families in which it was found. It was also present in healthy controls. Five missense mutations were found that were not shared by the other affected members of the respective families and were therefore considered as polymorphisms. Altogether 21 different polymorphisms were found in these five genes, most of which are novel and represent a resource for developing SNP markers for these genes (Table 4).

Discussion

We describe here the construction of a genomic map and the characterization of candidate genes in a genomic interval implicated in breast cancer predisposition and cancer progression. First, we constructed a 5.7-Mb genomic

Table 3 Multiple polyadenylation sites

mRNA	Location in bp	Nearest poly A signal	No.of 3' ESTs with nearby end	PolyA signal distance from polyA	Northern data
KIAA1008	3636	aauaaa	18	–29	
	3729	auuaaa	20	–20	Prominent band in all tissues
	4097	aauaga	5	–59	
	5441		51		Alu element
	7289	aauaaa	14	–18	Band varies intensity in various tissues
KLF12	2394	auuaaa	10	–21	Faint band in some tissues
	4657	cauaaa	3	–15	
	5978	aauaaa	7	–20	
	8594	aaugaa	2	–33	
	9127		10		False-primed from internal polyA stretch
	9540		5		False-primed from internal polyA stretch
	9773	aauaaa	2	–31	
	10590	aauaaa	23	–17	Prominent band
LM07	6021	aauaaa	50	–20	Only band in most tissues, smaller alternatively spliced isoform in brain
	6939	uauaaa	3	–27	
	7209	aauaaa	2	–18	Prominent band in skeletal muscle
KLF5	1613	aauaga	1	–18	
	1980		35		False-primed from internal polyA stretch
	3190	uuuaaa	2	–45	
	3539	aauaaa	27	–14	Prominent band
PIBF1	2469	aaaaca	1	–18	
	2757	auuaaa	3	–13	
	2824	aauaaa	39	–25	Prominent band in most tissues
	2895	uauaaa	1	–15	
	3122	aauaaa	7	–37	Faint band in most tissues

Table 4 Novel SNPs

Gene	Exon/Intron	Position[a] (bp)	Nucleotide change	Aminoacid change
AP-2 repressor	Intron 1	34(–3)[b]	C to T	Noncoding
	Exon 5	840	C to A	Gly to Gly
KIAA1 008 protein	Exon 5	806	G to A	Ser to Asn
	Exon 6	977	C to G	Thr to Arg
	Intron 7	1101(+57)	A to C	Noncoding
	Exon 14	1767	A to G	Thr to Thr
	Exon 15	1902	T to C	Ser to Ser
LM07	Intron 3	504(+136)	A to C	Noncoding
	Exon 6	852	C to T	Phe to Phe
	Intron 7	1100(+34)	C to T	Noncoding
	Intron 7	1100(+34)	C to A	Noncoding
	Exon 11	1825	A to G	Ile to Val
	Intron 13	2381(–3)	T to C	Noncoding
	Exon 14	2448	A to G	Leu to Leu
	Intron 19	2829(+42)	C to G	Noncoding
	Intron 19	2829(+42)	C to T	Noncoding
	Exon 23	3307	A to C	Asn to His
	Intron 22	3304(+7)	G to A	Noncoding
	Intron 26	4028(–44)	G to A	Noncoding
PIBF1	Intron 3	353(+19)	A to T	Noncoding
	Exon 4	499	A to G	Ile to Val
	Intron 5	672(+38)	G to A	Noncoding
	Intron 10	1323(–74)	del 4	Noncoding

[a]Position of nucleotide change considering the cDNA sequence and nucleotide 1 as corresponding to the first Met. Numbers in parenthesis correspond to the 5' part of the intron if positive or to the 3' part of the intron if negative
[b]Splice acceptor site

map based on a combination of a physical assembly of 18 YACs and 35 BACs and an *in silico* assembly of draft DNA sequence for 47 BACs from the human genome sequencing program. The two maps validated one another and made it possible to achieve greater than 90% coverage of contiguous genomic sequence, the majority of which is finished. Second, analysis of the genomic sequence uncovered eight genes in the interval and two predicted proteins and identified an additional set of 24 Unigene clusters, some of which were localized to this region for the first time. Third, for the most interesting genes in the region, we determined their exon-intron structure and their expression profile in multiple tissues. Finally, we performed a mutation analysis of five genes of potential relevance to cancer (PIBF1, KIAA1008, KLF12, KLF5, and LMO7) in breast cancer families and identified novel SNPs for further genetic analyses of these genes and this genomic region.

During the time that this work was performed, the availability of genome sequence for this region increased from virtually zero to the present 90%. Assembly of a clone contig map that had complete coverage enabled us to ascertain that only about 400 kb of the total of 5.7 Mb remained to be sequenced in five small gaps that were closed with the novel BAC clones identified here. Taken together, our results provide a substantially more coherent and complete view of the region of interest compared with the fragmented unfinished sequence data available in the genomic databases for this region as of April 2001. For example, in the April 1, 2001 freeze, the NCBI map (http://www.ncbi.nlm.nih.gov/genome/guide/human/) placed the genomic sequence contig 1 (Fig. 1), which is anchored by marker D13S1308, approximately 30 Mb distal to our region at 13q31. However, two unfinished sequenced BACs, RP11–26a1 and RP11–150l15 (Human Genome Project), linked the contig where D13S1308 is located, to the rest of our map. The order of the markers in the UCSC map (http//:genome.ucsc.edu/), as of the December 12, 2000 freeze is consistent with the order presented here. Another example is the location of the gene for Acid Phosphatase 5, tartrate resistant (ACP5), which is included in the UCSC map and in the published human genome maps for this region (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). In our BLAST analyses of the transcripts in this area, this gene showed 99% homology to a sequence in chromosome 19p13.3-p13.2 and no homology to any other chromosome. Therefore, this gene was most likely incorrectly assigned to chromosome 13.

The global analysis of the sequence data indicates that this region in chromosome 13 is a gene-poor area, as suggested by the lower GC content (36%) reported by the International Human Genome Sequencing Consortium (2001). Besides the eight known genes, two predicted proteins, and 24 Unigene/EST clusters, we have also identified more than 100 other putative transcripts in the area based on gene and exon prediction programs from GeneMachine (http://genome.nhgri.nih.gov/genemachine/; Makalowska et al. 1999). It is likely that several of these candidate transcripts will be false positives. Moreover, even some of the correct predictions may correspond to different exons of the same gene. This is the case for several of the Unigenes mapped to this area. In cloning full-length

transcripts for the genes in this region, we have encountered several examples that illustrate the difficulty of defining the identity of individual genes from EST data, even in the presence of genomic sequence information. Multiple Unigenes often correspond to the same gene. For example, 5 distinct Unigenes (Unigene Built 135; Table 2) correspond to the KLF12 gene. Many of the various Unigenes for KLF12 are attributable to different polyadenylation sites (Beaudoing et al. 2000), such as Hs.104492, Hs.151949, and Hs.294018 (Table 2). Another way in which extra clusters of ESTs may be generated is demonstrated by the cluster of 3'ESTs that terminate at base pair 1979 of KLF5 (AF132818). The finding that a corresponding 2.0-kb band has not been seen in Northern analysis can be explained by false priming from a polyA stretch internal to the 3' untranslated region during the generation of the cDNA libraries.

Of the multiple real and putative transcripts in the 13q21-q22 region, five genes have biological functions that are potentially relevant in cancer. PIBF1 is released by activated lymphocytes during pregnancy and inhibits natural killer activity and prostaglandin synthesis, thereby protecting against abortion in vivo (Szekeres-Bartho et al. 1985, 1989, 1999). KIAA1008 (Nagase et al. 1999; Shiomi et al. 1998) is a gene homologous to the dis3+ gene, a component of the exosome in fission yeast. It encodes a mitotic-control protein with a PIN (PilT amino terminus) domain, which is a novel signaling domain superfamily conserved from yeast to mammals (Makarova et al. 1999). KIAA1008 has been found to be overexpressed in colorectal cancers and in lung metastases of murine colon-adenocarcinoma (Lim et al. 1997). KLF12 and KLF5 are members of the mammalian Kruppel-like transcription factor family (Turner and Crossley 1999). KLF12 (also called Ap-2 rep) is a zinc finger protein that binds to the AP-2α gene promoter and represses its expression (Roth et al. 2000). AP-2 is an inducer of Her-2 oncogene expression in a nontumorigenic immortalized mammary epithelial cell line (Bosher et al. 1995). KLF5 may be an activator of EGF (Shi et al. 1999). Finally, LMO7 is a member of a group of proteins that are believed to have a transcription regulatory function through protein-protein interactions and other roles (Putilina et al. 1998). The LIM motif in LMO7 is a cysteine-rich double zinc-finger domain present in human, mouse, rat and *Caenorhabditis elegans* proteins (Putilina et al. 1998). No deleterious mutations in any of these genes have been found in the 13q21-q22-linked breast cancer families so far analyzed suggesting that these genes may not have major significance as predisposition genes. However, further work to evaluate the potential significance of the missense mutations/polymorphisms is warranted, as is the analysis of somatic mutations, particularly in prostate cancer that shares this region as a deletion site (Dong et al. 2000).

In conclusion, we describe the transcript, physical, and genomic maps for a 5.7-Mb region between D13S1308 and AFM220YE9 implicated in cancer development and progression. This region harbors eight known genes, 24 additional Unigene clusters, and more than 100 predicted genes or exons. Five known genes were studied in detail at the transcriptional level and were screened for mutations in 19 breast cancer families. Although deleterious mutations were not found in these families, these five genes and all the other newly defined candidate transcripts in the interval provide a rich resource and starting material for the identification of potential cancer-associated genes from this chromosomal site.

Acknowledgements We thank Jeff Trent, John Carpten, and Raman Sood for helpful discussions, Iza Makalowska and Mike Galdzicki for their support with the GeneMachine program, and Darryl Leja and Jennifer Reed for their help with the graphical design of the figures.

References

Beaudoing E, Freier S, Wyatt JR, Claverie J-M, Gautheret D (2000) Patterns of variant polyadenylation signal usage in human genes. Genome Res 10:1001–1010

Bosher JM, Williams T, Hurst HC (1995) The developmentally regulated transcription factor AP-2 is involved in c-erbB-2 overexpression in human mammary carcinoma. Proc Natl Acad Sci USA 92:744–747

Carpten JD, Makalowska I, Robbins CM, Scott N, Sood R, Connors TD, Bonner TI, Smith JR, Faruque MU, Stephan DA, Pinkett H, Morgenbesser SD, Su K, Graham C, Gregory SG, Williams H, McDonald L, Baxevanis AD, Klingler KW, Landes GM, Trent JM (2000) A 6-Mb high-resolution physical and transcription map encompassing the hereditary prostate cancer 1 (HPC1) region. Genomics 64:1–14

Couch FJ, Farid LM, DeShano ML, Tavtigian SV, Calzone K, Campeau L, Peng Y, Bogden B, Chen Q, Neuhausen S, Shattuck-Eidens D, Godwin AK, Daly M, Radford DM, Sedlacek S, Rommens J, Simard J, Garber J, Merajver S, Weber BL (1996) BRCA2 germline mutations in male breast cancer cases and breast cancer families. Nat Genet 13:123–125

Dong JT, Chen C, Stultz BG, Isaacs JT, Frierson HF Jr (2000) Deletion at 13q21 is associated with aggressive prostate cancers. Cancer Res 60:3880–3883

Ferlanti ES, Ryan JF, Makalowska I, Baxevanis AD (1999) WebBLAST 2.0: an integrated solution for organizing and analyzing sequence data. Bioinformatics 15:422–423

Ganguly A, Rock MJ, Prockop DJ (1993) Conformation-sensitive gel electrophoresis for rapid detection of single-base differences in double-stranded PCR products and DNA fragments: evidence for solvent-induced bends in DNA heteroduplexes. Proc Natl Acad Sci USA 90:10325–10329

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921

Kainu T, Juo S-HH, Desper R, Schäffer AA, Gillanders E, Rozenblum E, Freas-Lutz D, Weaver D, Stephan D, Bailey-Wilson J, Kallioniemi O-P, Tirkkonen M, Syrjäkoski K, Kuukasjärvi T, Koivisto P, Karhu R, Holli K, Arason A, Johannesdottir G, Bergthorsson JT, Johannsdottir H, Egilsson V, Barkardottir RB, Johannsson O, Haraldsson K, Sandberg T, Holmberg E, Grönberg H, Olsson H, Borg A, Vehmanen P, Eerola H, Heikkilä P, Pyrhönen S, Nevanlinna H (2000) Somatic deletions in hereditary breast cancers implicate 13q21 as a putative novel breast cancer susceptibility locus. Proc Natl Acad Sci USA 97:9603–9608

Larramendy ML, Tarkkanen M, Blomqvist C, Virolainen M, Wiklund T, Asko-Seljavaara S, Elomaa I, Knuutila S (1997) Comparative genomic hybridization of malignant fibrous histiocytoma reveals a novel prognostic marker. Am J Pathol 151: 1153–1161

Larramendy ML, Lushnikova T, Bjorkqvist AM, Wistuba II, Virmani AK, Shivapurkar N, Gazdar AF, Knuutila S (2000) Comparative genomic hybridization reveals complex genetic changes in primary breast cancer tumors and their cell lines. Cancer Genet Cytogenet 119:132–138

Lichter P, Cremer T, Borden J, Manuelidis L, Ward DC (1988) Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. Human Genetics 80:224–234

Lim J, Kuroki T, Ozaki K, Kohsaki H, Yamori T, Tsuruo T, Nakamori S, Imaoka S, Endo M, Nakamura Y (1997) Isolation of murine and human homologues of the fission-yeast dis3+ gene encoding a mitotic-control protein and its overexpression in cancer cells with progressive phenotype. Cancer Res 57: 921–925

Lundsteen C, Lind AM (1985) A test of climate room for preparation of chromosome slides. Clin Genet 28:260–262

Mairal A, Terrier P, Chibon F, Sastre X, Lecesne A, Aurias A (1999) Loss of chromosome 13 is the most frequent genomic imbalance in malignant fibrous histiocytomas. A comparative genomic hybridization analysis of a series of 30 cases. Cancer Genet Cytogenet 111:134–138

Makalowska I, Ryan JF, Baxevanis AD (1999) GeneMachine: a unified solution for performing content-based, site-based, and comparative gene prediction methods (abstract). CSHL Genome Meeting. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV (1999) Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res 9: 608–28

Nagase T, Ishikawa K, Suyama M, Kikuno R, Hirosawa M, Miyajima N, Tanaka A, Kotani H, Nomura N, Ohara O (1999) Prediction of the coding sequences of unidentified human genes. XIII. The complete sequences of 100 new cDNA clones from brain, which code for large proteins in vitro. DNA Res 6:63–70

Pinkel D, Straume T, Gray JW (1986) Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. Proc Natl Acad Sci USA 83:2934–2938

Putilina T, Jaworski C, Gentleman S, McDonald B, Kadiri M, Wong P (1998) Analysis of a human cDNA containing a tissue-specific alternatively spliced LIM domain. Biochem Biophys Res Commun 252:433–439

Roth C, Schuierer M, Gunther K, Buettner R (2000) Genomic structure and DNA binding properties of the human zinc finger transcriptional repressor AP-2rep (KLF12). Genomics 63:384–390

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Shi H, Zhang Z, Wang X, Liu S, Teng CT (1999) Isolation and characterization of a gene encoding human Kruppel-like factor 5 (IKLF): binding to the CAAT/GT box of the mouse lactoferrin gene promoter. Nucleic Acids Res 27:4807–4815

Shiomi T, Fukushima K, Suzuki N, Nakashima N, Noguchi E, Nishimoto T (1998) Human dis3p, which binds to either GTP- or GDP-Ran, complements *Saccharomyces cerevisiae* dis3. J Biochem (Tokyo) 123:883–90

Szekeres-Bartho J, Kilar F, Falkay G, Csernus V, Torok A, Pacsa AS (1985) The mechanism of the inhibitory effect of progesterone on lymphocyte cytotoxicity. I. Progesterone-treated lymphocytes release a substance inhibiting cytotoxicity and prostaglandin synthesis. Am J Reprod Immunol 9:15–18

Szekeres-Bartho J, Autran B, Debre P, Andreu G, Denver L, Chaouat G (1989) Immunoregulatory effects of a suppressor factor from healthy pregnant women's lymphocytes after progesterone induction. Cell Immunol 122:281–294

Szekeres-Bartho J, Barakonyi A, Polgar B, Par G, Faust Z, Palkovics T, Szereday L (1999) The role of gamma/delta T cells in progesterone-mediated immunomodulation during pregnancy: a review. Am J Reprod Immunol 42:44–48

Turner J, Crossley M (1999) Mammalian Kruppel-like transcription factors: more than just a pretty finger. Trends Biochem Sci 24:236–240

Venter JC, Adams MD, et al (2001) The sequence of the human genome. Science 291:1304–1351

Wistuba II, Behrens C, Milchgrub S, Syed S, Ahmadian M, Virmani AK, Kurvari V, Cunningham TH, Ashfaq R, Minna JD, Gazdar AF (1998) Comparison of features of human breast cancer cell lines and their corresponding tumors. Clin Cancer Res 4:2931–2938