

Generation of an Integrated Transcription Map of the *BRCA2* Region on Chromosome 13q12–q13

FERGUS J. COUCH,^{*} JOHANNA M. ROMMENS,^{†‡} SUSAN L. NEUHAUSEN,[§] CAROLE BELANGER,[¶] MARTINE DUMONT,[¶] KENNETH ABEL,^{||} RUSSELL BELL,^{**} SIMON BERRY,^{**} ROBERT BOGDEN,^{**} LISA CANNON-ALBRIGHT,[§] LINDA FARID,^{*} CHERYL FRYE,^{**} THOMAS HATTIER,^{**} TERESA JANECKI,^{**} PING JIANG,^{**} ROBERT KEHRER,^{**} JEAN-FRANCOIS LEBLANC,[¶] JODI MCARTHUR-MORRISON,[†] DAVID MCSWEENEY,[†] YOSHIO MIKI,^{††} YI PENG,^{*} CAROLLE SAMSON,[¶] MARIANNE SCHROEDER,^{**} SARAH C. SNYDER,^{**} MICHAEL STRINGFELLOW,^{**} CARRIE STROUP,^{**} BRAD SWEDLUND,^{**} JEFF SWENSEN,[§] DAVID TENG,^{**} SANJAY THAKUR,^{*} THANH TRAN,^{**} MARTINE TRANCHANT,[¶] JANE WELVER-FELDHAUS,^{**} ALEXANDER K. C. WONG,^{**} HIROAKE SHIZUYA,^{‡‡} FERNAND LABRIE,[¶] MARK H. SKOLNICK,[§] DAVID E. GOLDGAR,^{§¹} ALEXANDER KAMB,^{**} BARBARA L. WEBER,^{*} SEAN V. TAVTIGIAN,^{**²} AND JACQUES SIMARD^{¶^{2,3}}

^{*}Department of Medicine, Department of Pathology, and Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [†]Department of Genetics, Research Institute, The Hospital for Sick Children, and [‡]Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, M5G 1X8, Canada; [§]Genetic Epidemiology Group, Department of Medical Informatics, University of Utah School of Medicine, Salt Lake City, Utah 84132; [¶]Laboratory of Molecular Endocrinology, CHUL Research Center and Laval University, Quebec City, Quebec, G1V 4G2, Canada; ^{||}Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109; ^{**}Myriad Genetics Inc., 390 Wakara Way, Salt Lake City, Utah 84108; ^{††}Division of Biochemistry, Cancer Institute, Tokyo, Japan; and ^{‡‡}Division of Biology, California Institute of Technology, Pasadena, California 91125

Received March 22, 1996; accepted May 29, 1996

An integrated approach involving physical mapping, identification of transcribed sequences, and computational analysis of genomic sequence was used to generate a detailed transcription map of the 1.0-Mb region containing the breast cancer susceptibility locus *BRCA2* on chromosome 13q12–q13. This region is included in the genetic interval bounded by *D13S1444* and *D13S310*. Retrieved sequences from exon amplification or hybrid selection procedures were grouped into physical intervals and subsequently grouped into transcription units by clone overlap. Overlap was established by direct hybridization, cDNA library screening, PCR cDNA linking (island hopping), and/or sequence alignment. Extensive genomic sequencing was performed in an effort to understand transcription unit organization. In total, approximately 500 kb of genomic sequence was completed. The transcription units were further characterized by hybridization to RNA from a series of human tissues. Evidence for seven genes, two putative pseudogenes, and nine additional putative transcription units was obtained. One of the transcription units was recently identified as *BRCA2* but all others are novel genes of unknown func-

tion as only limited alignment to sequences in public databases was observed. One large gene with a transcript size of 10.7 kb showed significant similarity to a gene predicted by the *Caenorhabditis elegans* genome and the *Saccharomyces cerevisiae* genome sequencing efforts, while another contained a motif sequence similar to the human 2', 3' cyclic nucleotide 3' phosphodiesterase gene. Several retrieved transcribed sequences were not aligned into transcription units because no corresponding cDNAs were obtained when screening libraries or because of a lack of definitive evidence for splicing signals or putative coding sequence based on computational analysis. However, the presence of additional genes in the *BRCA2* interval is suggested as groups of putative exons and hybrid selected clones that were transcribed in consistent orientations could be localized to common physical intervals. © 1996

Academic Press, Inc.

INTRODUCTION

Transcript maps of chromosomal regions are important milestones to understanding the human genome (Collins and Galas, 1993). Physical and genetic maps provide the framework for detailed study of genomic regions while transcript maps are essential for the characterization of all genes present and an understanding of their genomic organization. Transcript

¹ Current address: Unit of Genetic Epidemiology, International Agency for Research on Cancer, Lyon Cedex 08, France.

² These authors should be considered equal contributors.

³ To whom correspondence should be addressed. Telephone: (418) 654-2264. Fax: (418) 654-2735. E-mail: Jacques.Simard@crchul.ula.val.ca.

maps also play a role in the identification of candidate genes for mapped disease genes.

Genes or gene fragments can be identified from genomic DNA by a variety of methods. The most common approaches are (1) exon amplification (Buckler *et al.*, 1991), (2) hybrid selection (Parimoo *et al.*, 1991; Lovett *et al.*, 1991), (3) analysis of genome sequence (Uner-Bacher and Mural, 1991), and (4) direct cDNA library screening (Brody *et al.*, 1995). The advantages and disadvantages of these methods have been described (Hochgeschwender, 1992; Harshman *et al.*, 1995). As it is unlikely that a single method will identify all the transcripts within a given genomic interval (Harshman *et al.*, 1995), we sought to develop a detailed transcription map of the *BRCA2* region by employing the first three of these techniques.

In late 1994, the breast cancer susceptibility locus *BRCA2* was mapped to the 6-cM region between the genetic loci *D13S289* and *D13S267* (Wooster *et al.*, 1994). This interval was subsequently reduced to 4 cM by identification of recombination events in linked families with early onset breast cancer using the markers *D13S289* and *D13S260* (Thorlacius *et al.*, 1995). Subsequent physical mapping revealed this region to be approximately 2 Mb in size, a region small enough to attempt a positional cloning strategy. A physical map of the 1.5-Mb region between *D13S1444* and *D13S310* was constructed (see below). Coincident with ongoing genetic and physical mapping studies, analysis of sporadic pancreatic cancer tumors resulted in identification of a single homozygous deletion of 250–300 kb located within the *BRCA2* interval between *D13S260* and *D13S171* (Schutte *et al.*, 1995a,b). As deletions at this interval are extremely rare (Teng *et al.*, 1996), the presence of the homozygous deletion within the delineated region defined by genetics suggested the existence of a tumor suppressor gene, possibly *BRCA2*. Also during the course of our study 900 kb of genomic sequence from the *BRCA2* region was made publicly available by the Sanger and Washington University sequencing centers. Most recently, a portion of the *BRCA2* cDNA sequence (7.3 kb of an estimated 10–12 kb) was reported (Wooster *et al.*, 1995). In parallel, the complete coding sequence and the exonic structure of the *BRCA2* gene, which is composed of 27 exons distributed over approximately 70 kb of genomic DNA, have been determined and deposited in GenBank (Accession No. U43746) (Tavtigian *et al.*, 1996).

We present a transcription map of the *BRCA2* region integrated with the physical and genetic maps of the region. The physical map spans the 1.5-Mb interval between *D13S1444* and *D13S310*. The most thoroughly analyzed segment of the transcription map spans the 750-kb interval bounded by *D13S1699* and *D13S171*. In addition to the *BRCA2* gene, we have identified 17 transcription units. Furthermore, several expressed segments including unique exons and cDNA fragments not contained within the assigned transcription units

have been isolated and accurately mapped to the region.

MATERIALS AND METHODS

Physical map assembly. Yeast artificial chromosome (YAC) clones from the Centre d'Etude du Polymorphisme Humain (CEPH) containing short tandem repeat (STR) markers *D13S289*, *D13S290*, *D13S171*, *D13S260*, and *D13S267* were identified from compiled maps at GenBank. The YACs were obtained from Research Genetics Inc. (Huntsville, AL), propagated, and verified using the STRs initially determined to be linked to *BRCA2* region. Additional STRs were obtained from CEPH, and sequence-tagged sites (STSs) were generated from YAC ends either by YAC vector-*Alu* PCR using primers as described (Neuhausen *et al.*, 1994) or by YAC vector-random primer PCR (Swensen 1996). These STSs were used to identify overlaps among the YACs. YAC end STSs were checked on the somatic cell hybrids NA11689 and NA11575 (Coriell Institute Camden, NJ) to ensure that all of the YAC ends were on 13q.

An initial, incomplete set of P1s and BACs in the region was identified using STRs and YAC-derived STSs to screen the Genome Systems P1 library and Simon Laboratory BAC library "A" (Shizuya *et al.*, 1992). P1 and BAC clone DNAs were prepared and ends were sequenced as described (Neuhausen *et al.*, 1994). Clone overlaps and relative orientations were determined by STS content mapping. Non-repetitive markers from the edges of the developing clone contigs were used to screen the P1 and BAC libraries again. Once the PAC library became available from Genome Systems, PAC screening was substituted for P1 screening. In this way, the developing P1/BAC/PAC clone contigs were expanded until a single contig spanning the entire genetically defined interval was completed. Di-, tri-, and tetranucleotide repeat markers were cloned from the P1s, BACs, and PACs by hybrid capture using biotinylated repeat-containing oligos (Swensen 1996). STR primer sequences and sizes of the products are given in Table 2. They are also available from the Genome Data Base (GDB).

Restriction analysis of P1s, PACs, and BACs. Five hundred nanograms of DNA from each clone was digested with *EcoRI* or *HindIII* and fractionated in 0.5% agarose gels. The restriction fragments were visualized with ethidium bromide staining under ultraviolet light, and clone lengths were determined by adding the sizes of individual restriction fragments, excluding those corresponding to vector fragments. The extent of overlap among neighboring P1s, BACs, and PACs was calculated by adding the sizes of shared restriction fragments.

Identification of STRs. STRs were identified by three methods: (1) probing immobilized cosmid DNA with a (CA)_n oligomer to identify repeats; (2) direct sequencing of P1 clones with a (CA)_n oligomer; and (3) hybrid capture of P1, BAC, and PAC fragments containing di-, tri-, and tetranucleotide repeats. Followed cloning and sequencing, oligonucleotides were selected and used in PCR as described (Neuhausen *et al.*, 1994). Details of each identified STR are indicated in Table 2.

Hybrid selection. Randomly primed cDNA was prepared from poly(A)⁺ RNA of mammary gland, ovary, testis, fetal brain, and placenta tissues and from total RNA of the Caco-2 cell line (ATCC HTB 37). Hybrid selection with pooled cDNA was carried out for two consecutive rounds of hybridization to immobilized P1 or BAC DNA as described previously (Tavtigian *et al.*, 1996; Rommens *et al.*, 1995). Two to four overlapping P1 and/or BAC clones were used in individual selection experiments. Approximately 200 to 300 individual colonies from each ligation (from each 250 kb of genomic DNA) were picked and gridded into microtiter plates for ordering and storage. Cultures were replica transferred onto Hybond-N membranes (Amersham). Initial analysis of the cDNA clones involved a prescreen for ribosomal sequences and subsequent cross screenings for detection of overlap and redundancy.

Plasmids from 25 to 50 clones from each selection experiment that did not hybridize in the prescreening phase were isolated for further

analysis. The retrieved cDNA fragments were verified to originate from individual starting genomic clones by hybridization to restriction digests of DNAs of the starting clones. The clones were tentatively assigned to groups based on the overlapping or nonoverlapping intervals of the genomic clones. The fragments were also hybridized to genomic DNAs from human leukocytes and from the rodent hybrid cell GM10898A (NIGMS) that contains human chromosome 13.

A secondary round of hybrid selection was also performed. Biotinylated DNA from 10 genomic clones, positioned between P1-759-D10 and BAC B213E7 on the physical map, was hybridized to pooled cDNA from human mammary gland, brain, lymphocyte, and stomach in individual experiments as previously described (Tavtigian *et al.*, 1996). Two X 96-clones were isolated from each experiment. Each of the 1920 clones was sequenced, checked for repetitive sequence, and assessed for redundancy by comparison to other DNA databases and cDNA sequence from the BRCA2 region. Contigs of clones were assembled where overlapping sequence was identified. Many of the hybrid selected clones identified in Fig. 1 represent hybrid clone contigs rather than individual clones.

Exon amplification. Exon amplification was performed according to the methodology described (Church *et al.*, 1994). A minimal overlapping set of BACs, PACs, and P1s from the BRCA2 region was utilized. Three different experiments were undertaken: (1) a single pool of 10 genomic clones; (2) six pools of 3 overlapping genomic clones each; and (3) 10 individual genomic clones. Genomic clones were digested with *Pst*I or *Bam*HI and *Bgl*III and ligated into *Pst*I or *Bam*HI sites of the pSPL3 splicing vector. End products of exon amplification were cloned into the pAMP1 plasmid with the Uracil DNA Glycosylase cloning system (Life Technologies, Inc.). Six thousand clones were picked, propagated in 96-well plates, stamped onto filters, and analyzed for the presence of vector and repeat sequences by hybridization. Each clone insert was amplified by PCR and tested for redundancy and human clone specificity by hybridization to membranes with exon DNA and dot blots to the parent genomic DNA clone. Candidate exons were sequenced, compared to DNA sequence databases, and used as probes for screening cDNA libraries.

cDNA library screening. To provide a preliminary screen of cDNA libraries for hybrid selected clones, PCR was performed using specific oligonucleotides derived from the sequences of each of the retrieved cDNA clones using DNA from five λ gt11 cDNA libraries (human mammary gland (HL 1037b), human breast (HL 1061b), ZR-75-1 human breast cancer cells (HL 1059b), human ovary (HL 1098b), and human testis (HL1010b) (Clontech)) as template. Hybrid-selected clones present in cDNA libraries (as detected by this approach) were used as probes to screen the corresponding positive libraries. Hybrid clones for which PCR was not attempted were used as probes to screen a pool of the libraries described above (2×10^5 clones for each library) as previously described (Rommens *et al.*, 1995). Prehybridization and hybridization were performed at 42°C in 50% formamide, 5× SSPE, 0.1% SDS, 5× Denhardt's mixture, 0.2 mg/ml denatured salmon testis DNA, and 2 μ g/ml poly(A). Dextran sulfate (4%, v/v) was included in the hybridization solution only. The filters were rinsed in 2× SSC for 10 min at room temperature and then washed in 2× SSC/0.1% SDS for 30 min at 60°C followed by two washes in 1× SSC/0.1% SDS for 20 min each at 60°C. The positive phages were retested for second and third screenings, as required, to obtain purified plaques for subcloning or sequencing.

Unique exons were amplified by PCR, double labeled with [α -³²P]-dCTP and [α -³²P]dGTP, and screened against a pool of libraries (HepG2 liver, breast, mammary, placental, and testis (Clontech)) containing 200,000 clones from each. The HepG2 and testis libraries, an ovarian oligo(dT) primed library (Stratagene), and a random-primed breast library (Clontech) were also screened individually if the pooled screen was negative. Prehybridization and hybridization were performed in 10 mM NaCl, 5% SDS, 10% dextran sulfate, 100 μ g/ml of salmon sperm DNA at 65°C. Membranes were washed in 3× SSC, 0.5% SDS at 65°C for a total of 60 min and exposed to film at -70°C overnight. The positive phages were retested for second and third screenings to obtain purified plaques for PCR-based sequencing.

DNA sequencing. Expressed sequences retrieved during the first round of hybrid selection and many of the clones retrieved from cDNA libraries were sequenced by with the dideoxy chain termination method (Sanger *et al.*, 1977) using the T7 Sequencing Kit (Pharmacia Biotech Inc.). Alternatively, expressed sequences retrieved during the second round of hybrid selection, many of the clones retrieved from cDNA libraries, and all genomic DNA subclones obtained from the P1s, BACs, and PACs were sequenced on ABI 377 sequencers using ABI Prism dye terminator cycle sequencing kits (Perkin-Elmer). Finally, sequencing directly from P1, BAC, or PAC templates utilized either the Cyclist DNA sequence kit (Stratagene) or the Amplicycle DNA sequencing kit (Perkin-Elmer) interchangeably.

Sequence analysis. All of the sequence data generated over the course of this project were assembled into a Genetic Data Environment (GDE) (Smith *et al.*, 1994) database or Wisconsin Sequence Analysis Package GCG program (Version 8, September 1994, Genetics Computer Group, 575 Science Drive Madison, WI 53711). Sequence alignment, assembly, and the parsing of exons across genomic sequences were performed within GDE. BLAST (Altschul *et al.*, 1990) and FastA (Pearson and Lipman, 1988) searches against both local and remote databases were also initiated from within GDE. Sequences from cDNAs described in this work have been submitted to the NCBI databases under the GenBank Accession numbers listed in Table 3.

Northern analysis. RNA from HepG2 (ATCC HB8065), T-47D (ATCC HTB 133), MCF-7 (ATCC HTB 22), JEG-3 (ATCC HTB 36), Caco-2 (ATCC HTB 37), A293 (ATCC CRL 1573), and HeLa (ATCC CCL2) cell lines was isolated using the TRI reagent (Molecular Research Center, Inc.) according to the manufacturer's protocol. RNA was electrophoresed on a 1.2% agarose-formaldehyde gel and transferred to a GeneScreen Plus membrane (NEN, Dupont). Following UV cross-linking, the membranes were hybridized with probes labeled by random priming with the same hybridizing solutions described for the cDNA library screening, except for the presence of 1% SDS instead of 0.1% SDS. The membranes were washed twice in 2× SSC/0.1% SDS at 20°C for 30 min followed by a stringency wash in 0.1× SSC/0.1% SDS at 50°C for 30 min. RNA hybridization was subsequently performed as described above with MTN filters from Clontech. RNA probes were removed between hybridizations by heating membranes to 95°C in 0.02× SSC and 0.01% SDS.

RESULTS AND DISCUSSION

Physical and Genetic Map Integration

Two complete genomic clone contigs spanning the BRCA2 region as defined by genetic loci were assembled. A yeast artificial chromosome contig was obtained using the "Infoclone" database of the GDB. The relative positions and overlap of these YACs were verified and refined by STS content mapping using STR markers and YAC end-derived STSs. Cosmid libraries constructed from YACs 964B2, 931F4, 979H8, 951A3, 746G10, and 941D4 were screened with a C_{A_n} probe to identify new STRs. Cosmid ends were also sequenced to provide additional STSs for P1, BAC, and PAC library screening. An incomplete set of P1 and BAC clones of the region was obtained using the STRs and YAC-derived STSs by screening the Genome Systems P1 library and the Simon Laboratory BAC library "A" (Shizuya *et al.*, 1992). Initial groups of P1 and BAC clones were subsequently connected by using STSs developed from their respective ends. Once the PAC library became available from Genome Systems, PAC screening was substituted for P1 screening. Table 1 contains primer sequences and product sizes for the STSs used

TABLE 1

Primer Sequences and Sizes for the Sequence-Tagged Sites Derived from P1, BAC, and PAC Clones in the *BRCA2* Region

STS name	Forward primer	Reverse primer	STS size (bp)
P1-1101-F9 SP6	GGG TAA CTG GAC GTA AAG AC	GTG AGT TGA AAT GCT GTC TG	186
P1-1101-F9 T7	GAG TTC ACG ATT AAT TCT TAG G	GTG GTT CTA CCA AAC AGA CA	163
P1-981-C5 SP6 ^a	Repetitive	Repetitive	
P1-981-C5 T7	CTA TTC CCA TGA AAT AAA GG	GTT GCT TCA GAG GTT AAG TC	178
B52F10 SP6	GCC TAA AAG TTG TAG TGC TG	AAT GAG GGG AAT CGA GTC TA	144
B52F10 T7	ATT AGG AGG GAA AAT AAG AAG G	CAA GCC TGG ACA TCA ATG AG	126
P1-759-D10 SP6	CTT TAC CGT GGA AAC GCT TA	TGA GAG GTG AGG ATG TCT GC	181
P1-759-D10 T7	GCT ACA GCC ACA AAC TTA TGA	GAG TAC TGG GCA CAG AAA GA	249
P1-931-D1 SP6	AGA CAG AGA ATC TCA ACT GG	TTT GAT TTT CAC AGC AGA TG	155
P1-931-D1 T7 ^a	Repetitive	Repetitive	
P1-106-A2 SP6	GTG ACA GTA AGC TTC CTT G	AAA GAA CAT CCT TAG TTT GAC	153
P1-106-A2 T7	CAG GAA GAA CTG GAG GTT AG	ACG CTG CTT TGT TAT TTA GG	200
P1-1282-C12 SP6	CCA GGC TGA GCA ACA GAG AA	TGA CAG AGC AAG ACC CCA TC	150
P1-1282-C12 T7	ACT GCT GAT GAG TTT TGG TG	ACA GTG GTA GAC CAT CCA TA	200
B489G4 SP6	GAA TGA GGG CAA GGA ACA C	TTG ACA GCA AGC CAG TGA TA	227
B489G4 T7	GCT TTG AAT GTG GCC CAA CA	ATA TGT GTA AGA CGC GGG TG	177
PAC-25-7K SP6	TAT TAT CTC TTT GAA GTG G	ACT ACT AAA TTC CTG CTA C	172
PAC-25-7K T7	TTG TTG TTT TGT CTG AGG	TAC CCA GCT ACT TGA AGA	141
P1-294-F6 SP6 ^a	Repetitive	Repetitive	
P1-294-F6 T7	GCT TAC AAT ACG CAA CTT AC	AAT ATC TTA AAT GGT CAC AGG	196
P1-919-D6 SP6	AGA TGC TTA CTG GCA CTT AC	CAA TAT AGG AGG CCT AGT GTC	151
P1-919-D6 T7	TGA GAC CTT TTA TCA TCT GC	CAA GCT AAT CTG ACC AAG TG	247
B86E4 SP6	GGTATTACACCTGTTTTGCTC	GATGGATAATTCATCCCATAAC	130
B86E4 T7	CCC TCC TCC TAA GGT CAC TA	CCG TCT CAC TGG AGA GAT AA	118
B213E7 SP6	ACC AAC AAT CTC TTC AAG G	ACC AAC ACT GCT GAC ACC	240
B213E7 T7	GCT CCC TCT TAT TCC TTG	CTG TAT TTC AGG CAC TGG	220
PAC-38-14A SP6	AAT GGA GAG GAG GCT GTT A	GAA AAT TAA GAG CCC AGA A	75
PAC-38-14A T7	ATA CAT ATT TTC TCA AGG GAT A	CAT TGG CTT TTT GTC CTCT	98
P1-1149-C3 SP6	GGG TAG AAC AGG CAT TCG TC	CCT CAA CTT AGA TGG TGC CAG	100
YAC-979H8R ^b	CAG GAG GCT CAC AGC TCA GG	CAA GCA GAG CCA AAG GTG AG	120
P1-546-D4 SP6	CAA TAA AGG AAT ATG TGT AGA TAC	TCC ATC AAC TTA CTA TAC AAC TCC	150
P1-546-D4 T7	CCA CCC CTG CAT GGA CTC TG	CAT GGG TGT CAC GTG GAC TC	130
B2767G6 SP6	TAT TTC ATC CAA CCA TGT GC	GGA AAC CCA TTC TAT TAC AG	174
B276G6 T7	CAA GAC AGT GCA AGT GGT AG	TCC TCA TCG GAG TCG TCA	201
B484C6 SP6	AAC TGA GGT CTG CTA TTC A	ATT CTC TTC TGA CTT GGT A	179
B484C6 T7	TGA CGG AGC AGT GAG AAG	GTT TCC CGC TAC CAA GTC	188
P1-339-C7 SP6	ATG AGG AGG TTT ATC CAG TC	ATC AGC ACC ATT TGA AAT CC	117
P1-339-C7 T7	GTA AGA ACT TAC CAG CCA AGA	TCA CAG CAG GAT GGT TGA AG	153
B722G3 SP6	GTC CTT CCC TAG ACT GCA C	TAT GAA TGA CGC TTC TGG AG	154
B722G3 T7	TTG ATG TGT TGC TGG ATT C	GAT TCA CAG CCG AAC TCT AC	244

^a The P1-981-C5 SP6, P1-931-D1 T7, and P1-294-F6 SP6 end sequences were repetitive and did not provide good landmarks. Sequencing with P1-1101-F9 T7F, P1-106-A2 T7F, and PAC-25-7K T7F, respectively, confirmed map continuity.

^b The YAC-derived marker 979H8R was used to identify both P1-546-D4 and P1-1149-C3, confirming continuity.

to establish continuity across the region. Digested P1, BAC, and PAC clones were also screened by hybrid capture using biotinylated repetitive sequence primers to identify new STRs.

Following completion of the P1/PAC/BAC contig, the

BRCA2 candidate region was further refined by genetic recombinants in two *BRCA2* families. Thirteen polymorphic STRs were examined, 9 of which were new. The primer sequences, products sizes, number of observed alleles, and heterozygosity of these are pre-

TABLE 2
New STR Markers Located in the Region between D13S1444 and D13S310

Locus	Primer name	Primer sequence	Average size of STS (bp)	Number of alleles	Het ^a	<i>n</i>
D13S1444	tdj3820-R tdj3820-LB	AATGACTTTATCTACATGAAT CCCTTGCATGGAAAATTGTAAAG	186	9	0.80	70
D13S1700	M4247.4A.2F1 M4247.4A.2R2	ACCATCAAACACATCATCC AGAAAGTAACTTGGAGGGAG	292	18	0.89	76
D13S1699	M4659-SFB M4659-SRB	AGACAGAGAATCTCAACTGG TTTGATTTTTCACAGCAGATG	155	6	0.67	78
D13S1698	YS-GA9-CR1 YS-GA9-CL1	GTCCATACCACTAAGTCTGAC AACCTCAGGCTAATAGTCTCA	177	10	0.63	82
D13S1697	B489G-3C11FB B489G-3C11RB	CTGAAGGTTGGGGTGATTG CGTAATCCCARCGACTTGA	221	4	0.47	76
D13S1701	MB561A#FA2 MB561A#RB	GAATGTGCGAAGAGCTTGTC AAACATACGCTTAGCCAGAC	298	8	0.88	74
D13S1695	M53702C6FA M53702C6RA	AGAATCATTGCCTACTTA GATAACTTACCAGCATGTGA	246	11	0.79	78
D13S1696	YS-GB10TR1 YS-GB10TL1	CTTCAGAACTTATTAAGACCTTAG CTGGTTGTTTTAGAACTCATACT	214	6	0.56	82
D13S1694	YS-AC6-AAR1 YS-AC6-AAL1	CAATAAGCCCACTTGGGATAC CCATGTGAGGCACCTGTAAG	230	5	0.78	80

^a Observed heterozygosity from *n* independent chromosomes.

sented in Table 2. K107 had an affected individual who carried the predisposing haplotype telomeric to *D13S1444*, and K2043 had an affected individual who only carried the predisposing haplotype centromeric to *D13S310*. As shown in Fig. 1, the minimum tiling path across this reduced interval consists of 11 P1s, 6 BACs, and 2 PACs. The contig presented spans approximately 1.8 Mb as two additional BAC clones that extend across the region are also included. The map was consistent with STS content mapping of 51 additional overlapping P1/BAC/PAC clones from this region (data not shown).

Transcribed Sequence Identification

Exon amplification. Three experiments were performed. A total of 6000 clones were picked and gridded for analysis. Prescreening with repetitive DNA and vector DNA eliminated 60–70% of clones. PCR amplification of the remaining 30% revealed that two-thirds did not contain inserts. The final 10% were then analyzed for redundancy and location by hybridization to gridded sets of all candidate exon clones and to blots of genomic DNA from the original cloned genomic DNAs. A total of 75 unique exons were obtained and were sequenced. Twelve exons mapped outside the *D13S1444* to *D13S310* region contained repetitive sequences or were shown to result from cryptic splicing of the vector. Each of the remaining candidate exons was used for cDNA library screening. One or more cDNA clones were isolated for 14 of the 63 candidate exons tested.

Hybrid selection. Hybrid selection was also used to retrieve transcribed sequences from groups of two to four genomic clones, each spanning 250–300 kb as described under Materials and Methods. As the physical

map developed, selection was also applied to single genomic clones to ensure that all regions were examined or that regions appearing to be poorly represented in the cDNAs from the grouped clones could be reassessed. Two schemes were employed to analyze the selected cDNA fragments. In the first scheme, hybridization with individual clones was used to determine map position and to provide an assessment of redundancy. The fragments were also verified to originate from chromosome 13 by hybridization to a hamster hybrid containing chromosome 13 as its only human material. Two selected clones were found that hybridized to discrete *EcoRI* fragments in B213E7 DNA and to fragments of corresponding sizes in chromosome 13 DNA but that also hybridized strongly to additional fragments in total human DNA; these were not analyzed further. The remaining 95 of a total of 97 clones were then analyzed by sequencing and used for cDNA library screening. The results for those clones yielding cDNAs are presented in Table 3.

The second scheme of hybrid selection experiments identified a total of 1920 retrieved clones. These clones were screened for the presence of repetitive elements. The remaining nonrepetitive clones were sequenced, condensed according to overlap, and aligned with overlapping sequences from all other putative transcripts in our database. Potential cDNA sequence contigs were also aligned with genomic sequence and scanned for the presence of splice junctions. These sequence contigs revealed unique transcription units or extended the sequences available for identified transcription units. Specific clones from this series were also used to screen cDNA libraries (Table 3).

Genomic DNA sequencing. Genomic DNA sequences obtained by sampling from across the entire

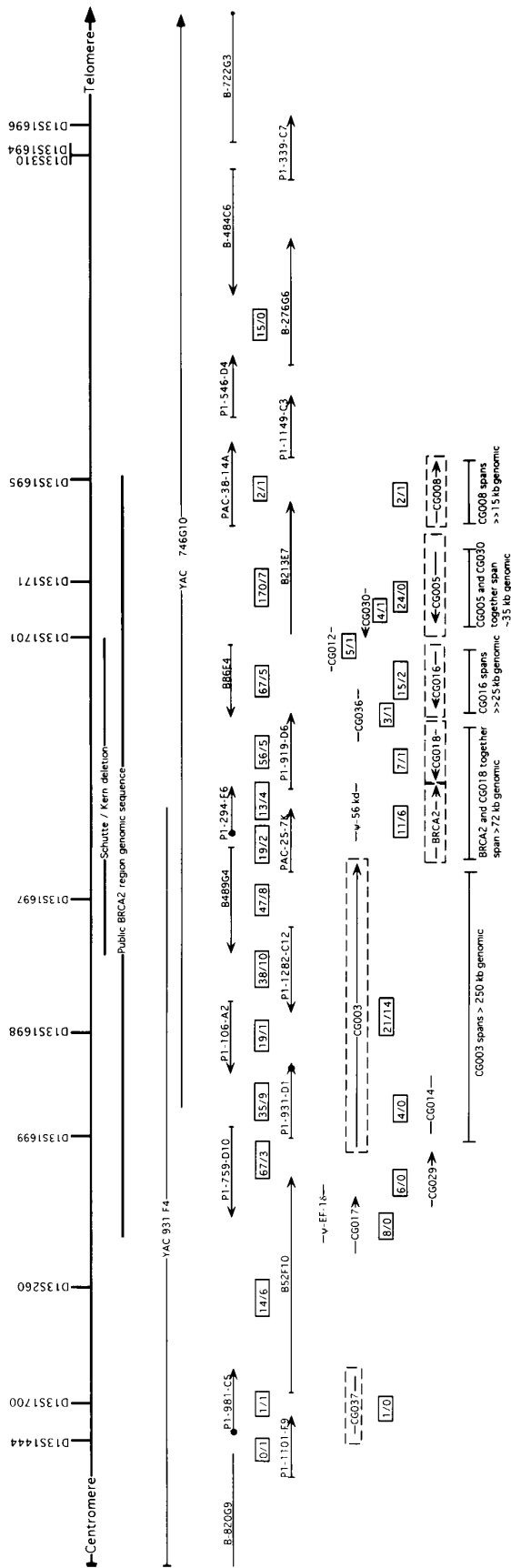


FIG. 1. Integrated genomic/transcript map of the BRCA2 region. The polymorphic loci are shown above the physical map that was assembled from the YACs 931F4 and 746G10 at the BRCA2 locus. The two polymorphisms at D13S310 are close together and could not be ordered. The minimal tiling contig of P1, BAC, and PAC clones indicated as horizontal arrows are shown below the YACs. The T7 and SP6 ends are indicated as \rightarrow and \leftarrow , respectively. Filled circles indicate clone ends too repetitive to provide useful landmarks. The number of hybrid-selected clone contigs and nonrepeated exon-trapped clones (respectively) that mapped to each P1, BAC, or PAC is indicated by the two numbers in the solid box associated with each genomic clone. Transcribed contigs mapping to overlap regions were counted only once. Transcription units are placed below the P1/BAC/PAC contig. Candidate transcription units are represented by horizontal lines with their name superimposed. Where the orientation of transcription is known, it is indicated by an arrowhead. The Class I transcription units are surrounded by hatched boxes. The subset of hybrid-selected clone contigs and nonrepeated exon-trapped clones, respectively, that mapped to each candidate transcription unit is indicated by the two numbers in the solid box associated with each unit. Alignment of the exons of the transcription units across the genomic sequence allowed a minimum estimate of the genomic intervals spanned by some of the transcription units. These sizes, where known, are indicated below the transcription units.

TABLE 3
Summary of Transcription Units

Transcription unit classes	Probes			Isolated cDNA clones			Transcription units																							
	Name	Size (kb)	Genomic localization	Number of cDNA clones	Size of longest contig (kb)	GenBank Accession No.	Name	Transcript size (kb)	Remarks																					
I	GT563	0.65	B489G4	2	10.7	U50534	CG003	2.1, 3.3, 3.5, 6.5, 8.0, and 11.0	Similarity with <i>C. elegans</i> F21H11.2 protein (GenBank U11279)																					
	GT571	0.6	B489G4	2																										
	GT599	0.75	P1-106-A2	2																										
	GT601	0.99	B489G4	6																										
	GT637	0.4	P1-1282-C12 and B489G4	1																										
	GT653	0.8	P1-1282-C12	3																										
	wXBE1A3	0.18	P1-1282-C12 and B489G4	1																										
	wXBE1C10	0.27	P1-1282-C12 and B489G4	2																										
	wXBE1E11	0.24	P1-1282-C12 and B489G4	1																										
	wXPA1B11	0.14	P1-106-A2	1																										
	wXPH1H10	0.13	P1-1282-C12 and B489G4	1																										
	wXPE1A4	0.12	P1-1282-C12 and B489G4	1																										
	GT566	0.9	B231E7	7						2.1	U50532	CG005	2.3, 3.2, 4.4, and 9.5	Similarity with <i>C. elegans</i> F26A1.14 protein (GenBank U27312)																
	GT575	0.35	B231E7	1																										
	GT578	0.6	B213E7	3																										
	GT607	0.7	B231E7	3																										
	GT610	0.55	B231E7	8																										
	GT616	0.41	B213E7	10																										
	GT641	0.45	B213E7	1																										
GT658	0.75	B213E7	1																											
GT659	0.75	B213E7	8																											
GT568	0.9	B231E7	3	2.3	U50535	CG006	First 82 nucleotides identical to the 3' end of exon II, while the remaining 2.2 kb correspond to 5' end of intron II of CG005 gene																							
GT603	0.75	B231E7	2																											
GT617	0.75	B231E7	1																											
GT577	0.5	B231E7	2					1.37	U50536						CG011	Intron III of CG005 gene														
GT576	0.51	B231E7	1														0.68	U50537	CG033	Intron II of CG005 gene, EST H05940										
wXBG1D11	0.19	PAC-38-14A	2																		1.85	U50533	CG008	1.4, 5.5, and 8.0	No similarity					
GT642	0.75	P1-294-F6	1																							10.9	U43746	CG013 (BRCA2)	0.4, 1.0, and 11.0	Tavtigian <i>et al.</i> (1996)
GT713	0.41	PAC-25-7K	1																											
wXBF1B6	0.3	P1-294-F	1																											
wXPF1B8	0.3	P1-294-F6	2																											
WXPFI A5	0.5	P1-294-F6	3																											
GT597	0.6	B86E4	2							2.8	U50529	CG016	4.4	No similarity																
GT660	0.55	B86E4	2																											
GT615	0.6	B86E4	1																											
GT677	0.6	P1-919-D6 and B86E4	1																											
wXPB1F12	0.25	B86E4	1																											

TABLE 3—Continued

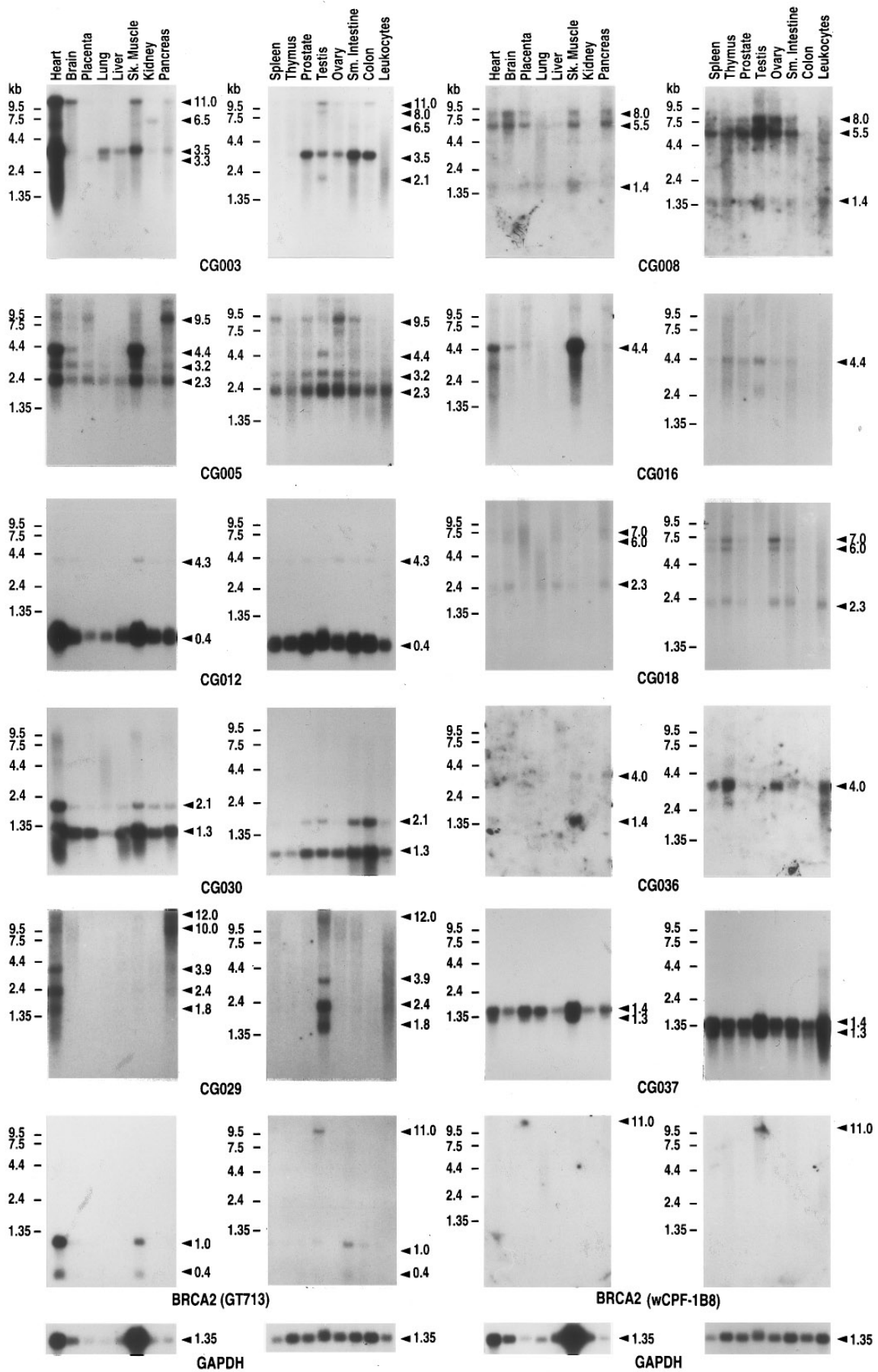
Transcription unit classes	Probes			Isolated cDNA clones			Transcription units		
	Name	Size (kb)	Genomic localization	Number of cDNA clones	Size of longest contig (kb)	GenBank Accession No.	Name	Transcript size (kb)	Remarks
	GT711	0.75	P1-919-D6	1	2.05	U50527	CG018	2.3, 6.0, and 7.0	EST R61510, EST R41970, EST F03825, EST Z38606 EST F02671
	wXBF1A6	0.25	P1-919-D6	1					
	GT627	0.4	B52F10	21	1.43	U50523	CG037	1.3 and 1.4	EST R69912, EST T33994, EST H74118, EST R63793 EST R81762, EST R64207
II	GT565	0.45	B86E4	6	1.42	U50530	CG012	0.4 and 4.3	Colinear with genomic sequence EST H18362
	GT661	0.38	B86E4	9					
	Hexp17.1E05	0.6	P1-759-D10	5	1.62	U50525	CG029	1.8, 2.4, 3.9, 9.5, and 12.0	Colinear with genomic sequence, no similarity
	Hexp10.2F08 wXBG1C7	0.32 0.15	B213E7 B213E7	1	4.9	U50531	CG030	1.3 and 2.1	Colinear with genomic sequence EST L44472
	GT574	0.41	P1-919-D6 and B86E4	1	0.95	U50528	CG036	1.4 and 4.0	Colinear with genomic sequence No similarity
III	GT572	0.6	P1-931-D1	1	1.2	U50526	CG014	ND	Colinear with genomic sequence, no similarity
	Hexp12.1G01	0.7	P1-759-D10	1	2.3	U50524	CG017	ND	Colinear with genomic sequence, no similarity
	GT573	0.75	P1-931-D1	3	0.75	U50538	—	ND	Colinear with genomic sequence, no similarity
	GT605	0.5	B213E7	1	0.5	U50539	—	ND	Colinear with genomic sequence, no similarity
	GT650	0.75	B276C6	2	0.45	U50540	—	ND	No similarity
IV	—	—	P1-759-D10 and B52F10	—	—	—	ψ -EF-1 δ	ND	Putative pseudogene: altered form of transcription elongation factor-1 δ (TEF-1D)
	—	—	P1-294-F6	—	—	—	ψ -56 kDa	ND	Putative pseudogene: non-coding copy of interferon inducible 56-kDa protein (GenBank Accession No. M24594)

Note. A summary of the retrieved clones and the corresponding isolated cDNA is listed. The clones are grouped according to the transcription units shown in Fig. 1. The longest cDNA "contig" or "contigs" of each transcription unit as given were obtained by compiling overlapping sequences of selected or trapped clones and their corresponding cDNAs obtained by library screening. The contig sequence was assigned a GenBank Accession number as indicated.

interval facilitated the merging of groups of transcribed contigs and revealed transcription unit organization and genomic structure. Sequences of P1, BAC, or PAC clones were obtained with oligonucleotides corresponding to clone vector ends or from internal oligonucleotides designed from selected or trapped transcribed sequences. Additional sequences were obtained from shotgun cloning of small fragments of the larger genomic clones. In total, 30% of the sequence of the region from BAC B52F10 to PAC-38-14A was obtained. cDNA contigs generated by exon trapping, hybrid selection, and cDNA library screening were aligned with the genomic sequences to help identify splice junctions and resolve issues of transcription unit organization. However, with this limited amount of genomic sequence, spatial relationships between the retrieved transcribed contigs proved difficult to establish even when the genomic clones of origin were known. During the course of this work 900 kb of genomic sequences from the in-

terval outlined in Fig. 1 became publicly available from the Sanger and Washington University sequencing centers. These genomic sequences were aligned with our genomic sequences and could be merged into a set of 160 sequence contigs.

Assembly of transcription units. Contigs of transcribed sequences were grouped into tentative transcription units based on (1) map position, (2) overlap of sequence following isolation of longer cDNA clones from cDNA libraries that provided for extension of contigs or for merging of small contigs, (3) recognition of poly(A) tails and corresponding polyadenylation signals, and/or (4) RNA hybridization analysis. The results are summarized in Fig. 1 and Table 3. Identities to public ESTs are also indicated in Table 3. The RNA hybridization results are shown in Fig. 2. In total, four classes of transcription units were assigned. The first class included seven transcription units or genes for which complete (three) or partial transcription contigs



(four) were isolated, each with putative extended coding potential. Furthermore, these contigs aligned with genomic sequence such that gene structure with exons and introns could be recognized following alignment. Each of these genes, including CG003, CG005, CG008, CG013 (*BRCA2*), CG016, CG018, and CG037, detected discrete transcripts in human tissues as shown in Fig. 2 and will be described individually below.

The second class included the four transcription units CG012, CG029, CG030, and CG036, for which extended cDNA sequence was obtained from isolated cDNAs. Their respective expression profiles in a variety of human tissues were distinct and thus suggest that they correspond to discrete transcription units (Fig. 2). Further, their physical position indicated dispersal across the interval and thus also supported the likelihood of each being a discrete unit. However, all of these transcript contigs appeared to be colinear with genomic DNA and did not appear to possess extended coding capacity. Whether these units encoded polypeptides was therefore uncertain.

During the characterization of CG012, it appeared that some isolated cDNAs did not align precisely to the genomic DNA sequence. It was not possible to establish that these cDNAs were actually transcribed from the chromosome 13q12–q13 region. That a gene family was involved appeared reasonable but also limited the conclusions that could be drawn from the observed expression in the human tissues even though it was known that the cDNA that was used for hybridization to RNA was verified to originate from chromosome 13.

The third class of transcription units included sets of contigs built largely by exon or hybrid selected clones with longer cDNA clones. They differed from class II units in that no RNA hybridization analysis was performed. They included the units designated CG014, CG017, GT573, GT605, and GT650 in Table 3. From the limited sequence that was obtained, four appeared to be colinear with genomic sequence. GT650 could not be analyzed as no genomic sequence from B276G6 was available. This class of transcription unit may correspond to genes that code for polypeptides, but insufficient information was obtained to establish this.

The fourth class included two transcription units for which evidence was obtained to suggest that they were pseudogenes. The genomic sequence revealed frameshift alterations in a sequence nearly identical to the gene for a transcription elongation factor (*TEF-1d*) in the overlap interval of B52F10 and P1-759-D10. Two hybrid selected clones corresponding to this sequence were obtained, one with identity to the reported gene sequence and one with

perfect alignment to the genomic sequence with internal frameshifts. The second pseudogene aligned with the gene for an interferon-inducible 56-kDa protein. This sequence also contained frameshifts when compared to the GenBank entry of the original gene (Accession No. M24594) and was positioned in intron 24 of the *BRCA2* gene.

In summary, a total of 27 of the 63 trapped exons were accounted for by sequence alignment to transcription units. The numbers of exon clones isolated from specific genomic clones and those present in the individual transcription units are indicated in Fig. 1. The other 36 putative exons are spread across the entire region with a noted number of 6 from the B52F10 BAC genomic clone alone. A total of 563 hybrid-selected clone contigs are shown on the map in Fig. 1, 111 of which are contained within the 7 candidate genes and the transcription units as indicated. Similarly, as with exon-trapped clones, many of the hybrid selected clones are not contained within the transcription units from Table 3. The organization of transcription units provided some insight into the distribution of these nonassembled transcribed sequences.

Organization of transcription units. Together with the sequence generated during the course of this work and with the genomic sequence made publicly available, each candidate gene or transcription unit could be aligned. Additional information including direction of transcription, an estimate of the minimum size of the corresponding genomic region, and relative proximity (Fig. 1). This exercise also enabled the ordering many of the independent sequence contigs across the region, though individual contigs completely within a gap or an interval of nonoriented sequence would have been missed.

The interval between CG037 and CG003 spans more than 100 kb. It contains one Class II transcription unit (CG029) and one Class III transcription unit (CG017) as well as singlet exon-trapped and hybrid-selected clones. The genomic clone B52F10 was not used in a second round of hybrid selection, and only a limited amount of gene assembly was feasible. Given the number of nonassembled exons that were found to map to B52F10 (see previous section), there is a strong possibility that the interval contains unidentified genes.

CG003 spans at least 250 kb of genomic sequence. The first intron of CG003 spans more than 29 kb, and the apparent inclusion of the Class III transcription unit CG014 within the intron is noteworthy. It was not possible to establish whether CG014 did contain equivocal coding capacity as the sequence that was

FIG. 2. Hybridization of cDNAs to poly(A)⁺ RNA. Representative clones of 11 transcription units are shown hybridized to 2 μg of poly(A)⁺ RNAs of the tissues indicated. The sizes of the markers listed on the left and the sizes of mRNAs, as indicated by the arrows on the right, are given in kilobases (kb). As control for the amount of RNA present in each lane, the blots were hybridized with a glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) cDNA fragment. Increased levels of RNA in heart and skeletal muscle relative to other tissues are apparent. Blots were exposed to Kodak X-Omat AR film with one intensifying screen at -70°C. 0.5 to 10 days were needed to optimize signal.

available was limited. This transcription unit may correspond to an intron sequence as retrieval of transcribed but unspliced sequence by either hybrid selection or cDNA library screening is plausible. However, of the 28 unique trapped exons that mapped to the region encompassing CG003, only 14 corresponded to CG003 itself and 9 distinct exons mapped to within the P1-931-D1 genomic clone. While cryptic splicing artifacts can be generated, it is unlikely that so many would be isolated from a single physical interval. Further, based on RNA hybridization, these exons did not correspond to the large or alternative transcripts seen for CG003 (data not shown). These results together suggest the possibility of at least one additional gene in this physical interval.

The 3' end of CG003 and the 5' end of BRCA2 both map to the overlapping interval of B489G4 and PAC-25-7K. The alignment of end exons of CG003 and BRCA2 reveals that the intervening region contained approximately 24 kb of genomic sequence. The positioning or order of 9 hybrid selected clones (2 of which overlap to form a contig) is restricted to the region of the 3' introns of CG003, the 5' introns of BRCA2, and the region between the 2 genes. In contrast, 24 kb of genomic sequence, including the 5' end of intron 11 of BRCA2 extending into intron 24, contained 48 hybrid selected clones that were assembled into 9 contigs and 13 singlets (Note that only the contigs are indicated in Fig. 1.) In addition, the putative introns did not give rise to trapped exons. Taken together, it appears unlikely that any additional genes occur between CG003 and BRCA2. This evaluation was feasible as the genomic sequence was nearly complete in this region. Further, given that BRCA2 is not an abundantly expressed gene (Tavtigian *et al.*, 1996; Fig. 2), the depth of detection of transcribed sequences by the exon trapping and hybrid selection procedures was adequate. The majority of exons and transcribed contigs from the region were explained as *bona fide* BRCA2 sequences.

Although the exact polyadenylation site of BRCA2 is unknown, the available and converging 3' ends of BRCA2 and CG018 are spaced by only 2 kb. CG018 and CG016 are transcribed in the same direction and map to the overlapping genomic clones P1-919-D5 and B86E4, respectively. Comparison of the size of their longest cDNA contig with their transcript sizes estimated from Northern blots reveals that neither of these genes has been entirely cloned. In addition, the Class II transcription unit CG036 plus a large number of hybrid selected clones that have not been assigned to any named transcription unit map between them. It is likely that additional exons of CG016 or CG018 are also present.

CG016 and CG005, which are also transcribed in the same direction, map to the overlapping genomic clones B86E4 and B213E7, respectively. The polyadenylation signal of the CG005 splice isoform encoding the longest putative open reading frame currently identified lies 11 kb telomeric of the SP6 end of B86E4, in B213E7.

However, genomic sequences across CG016 are somewhat sparse, so the distance from the 5' end of CG016 to the SP6 end of B86E4 is unknown. A very large number of hybrid-selected clones, almost all transcribed in the same direction as CG005, were captured from the 11 kb of genomic sequence immediately downstream of the CG005 polyadenylation signal. The sequence of the Class II transcription unit CG030 was assembled from only some of these clones. A second Class II transcription unit, CG012, also maps between CG005 and CG016. CG030 and CG012 may be representatives of *bona fide* genes lying between CG005 and CG016. Alternatively, they may be included in introns or be a part of a larger transcript that includes both CG005 and CG016 in a single transcription unit. This latter possibility has not yet been ruled out and is supported by the common 4.4-kb transcript that was seen in RNA of several tissues with representatives of each of these two assigned units (Fig. 2).

A large number of hybrid-selected clones plus 7 exon-trapped clones map to the interval between the currently known 5' ends of CG005 and CG008. Genomic sequences across CG008 were very sparse, so only a limited amount of alignment was feasible, thus restricting additional analysis.

Characterization and expression patterns of transcription units. The analysis of the sequence of each of the transcription units and the RNA survey analysis of a range of human tissues provided for the identification and characterization of 6 new genes. CG013 corresponds to BRCA2 and is positioned essentially in the center of the region analyzed. Its partial (Wooster *et al.*, 1995) and complete (Tavtigian *et al.*, 1996) putative amino acid sequence has been reported. The structural organization of the BRCA2 gene, which consists of 27 exons spanning 60–70 kb, has also been described (Tavtigian *et al.*, 1996). This gene was initially merged from 3 candidate transcription contigs with the identification of the large open reading frame of a 4.9-kb exon from the genomic DNA sequence. cDNAs corresponding to this gene were identified from a subset of 6 trapped exons and 11 hybrid-selected clones, and the final sequence was obtained by overlap consensus. The expression of this gene as determined by RNA hybridization and RT-PCR has been described (Tavtigian *et al.*, 1996). Hybridization of three different cDNA probes to human multiple tissue Northern filters revealed an 11- to 12-kb transcript that was detectable in testis, thymus, and placenta (Tavtigian *et al.*, 1996, Fig. 2) thus suggesting that little, if any, of the BRCA2 mRNA sequence is missing from our composite cDNA. As illustrated in Fig. 2, smaller transcripts were also detected in several tissues by using the cDNA probe GT713, which corresponds to BRCA2 exons 3–7, but were not detected with the probe wCPF1B8.1, corresponding to the 3' end of exon 11 through to exon 20. The significance of these smaller transcripts is unclear.

CG037, present on P-1981-C5, is the most centro-

meric of the genes identified (Fig. 1). This gene was abundantly expressed as a doublet mRNA of 1.3 and 1.4 kb in all 16 human tissues tested (Fig. 2). Twenty-one cDNA clones were obtained by screening with a single hybrid-selected clone. The compiled sequence of the overlapping cDNAs revealed the presence of a long open reading frame of 248 amino acids with 5' and 3' untranslated regions of 248 and 444 nucleotides, respectively. Database analysis indicated alignment with 45 different ESTs but not to any known genes at either the nucleotide or the amino acid sequence level. Only the six ESTs having greater than 90% identity over >300 nucleotides are indicated in Table 3. It is of interest that, in contrast to the 1.4-kb transcript that is expressed in all tissues surveyed, the second, and smaller, mRNA species was not detectable or was weakly expressed in several tissues. The nature of the difference between the transcripts was not explained from the sequencing analysis but likely results from tissue-specific alternative splicing or polyadenylation.

A large gene, CG003, spanned over 250 kb of genomic DNA. RNA hybridizations revealed two prominent transcript sizes of 11.0 and 3.5 kb. The larger message, which is composed of 61 exons, was the only one expressed in the brain but was broadly expressed together with the 3.5-kb mRNA in other tissues including heart, skeletal muscle, testis, placenta, pancreas, spleen, thymus, prostate, ovary, colon, and small intestine. Complete sequence of cDNAs corresponding to 10.7 kb was obtained. The amino acid sequence aligned with significant similarity to extended open reading frames that have been identified in both *S. cerevisiae* and *C. elegans* genomic sequencing efforts, but only 12 human ESTs have been reported that could be aligned. The function of the protein encoded by this gene is not known. This gene is clearly complex as hybridization with GT601 (nucleotides 9461 to 10460 of the composite cDNA) revealed additional transcripts of 8.0 and 2.1 kb in testis RNA and a prominent mRNA of 6.5 kb in the kidney. Further, only mRNAs of 3.5 and 3.3 kb were present in the lung. The mRNA of 3.5 kb that generally appeared most prominent was absent or expressed at low levels in placenta, kidney, spleen, and thymus. The doublet 3.5- and 3.3-kb transcript and the 9.5-kb transcript were also observed in the HepG2 cell lines by using GT601 as probe but only the 9.5-kb mRNA species was detectable when GT653 (nucleotides 4807 to 5657 of the composite cDNA) was used as probe (data not shown). The complete alignment and the hybridization of the probes GT601 and GT653 to genomic DNAs confirmed their origin as the *BRCA2* region. Additional experiments will be required to delineate the relation of these multiple transcripts and to obtain a better assessment of the large message. Its expression pattern was partially compromised by the varying abundance of mRNAs present on the RNA blots used; see control (GAPDH) samples in Fig. 2.

The gene CG018 involved up to three mRNAs of 7.0, 6.0, and 2.3 kb as determined by using wXBF1A6 (nu-

cleotides 231 to 448) as a probe. The larger pair of transcripts and the smaller transcript demonstrated distinct but restricted expression profiles as shown in Fig. 2. Only a portion of this gene was recovered, but an extended segment with good coding capacity and aligned consensus splicing signals were present. Alignment to the public databases was restricted to five ESTs.

CG016 was characterized by a restricted pattern of expression with a 4.4-kb mRNA being most abundant in heart, brain, skeletal muscle, thymus, prostate, testis, ovary, and small intestine. The cDNA probe used, wCPB1F12.T8, corresponds to nucleotides 451 to 902. Seven cDNAs were isolated with hybrid-selected clones and a single exon-trapped clone to yield a partial contig of 2.2 kb. This limited sequence was found to span more than 25 kb of genomic DNA and revealed the presence of a long putative open reading frame of 635 amino acids with 5' and 3' untranslated regions of 323 and 583 nucleotides, respectively. No similarities were identified in the databases.

A composite cDNA sequence of 2.1 kb was found for CG005. The sequence was assembled from 42 cDNA clones and contains a long open reading frame of 583 amino acids with 5' and 3' untranslated regions of 169 and 197 nucleotides, respectively. The predicted open reading frame is derived from six exons. Sequence analysis indicated similarity to 15 ESTs and a portion of the putative coding region revealed a good alignment with the *C. elegans* F26A1.14 protein and to human 2',3' cyclic nucleotide 3' phosphodiesterase over 177 amino acids. This gene is relatively complex as four different transcripts with lengths of 9.5, 4.4, 3.2, and 2.3 kb were detected by RNA blot analysis using GT616 (nucleotides 891 to 1301) as probe. All tissues displayed the smaller mRNA with several tissues expressing different combinations of the larger messages as shown in Fig. 2. Three additional transcribed contigs have been mapped to the genomic region containing CG005, as listed in Table 3. They include CG006, CG011, and CG033. Their role in the CG005 transcript remains to be precisely determined, although it was noted that the sequence of each of these contigs was a colinear genomic sequence and they are all transcribed in the same direction as CG005. Thus, they may correspond to products of hnRNA. CG006 included exon II and intron II boundaries of CG005 and has been reported as an EST.

The most telomeric Class I transcription unit, CG008, was derived from one exon and two hybrid-selected clones. Two cDNAs were isolated using the exon-trapped clone to yield a final transcribed composite contig of 1.8 kb. Three transcripts of 8.0, 5.5, and 1.4 kb were detected in most human tissues analyzed but relative levels varied considerably as shown in Fig. 2. Transcript analysis was carried out with wCBG.1D11#2, which corresponds to nucleotides 1 to 1066 of the composite cDNA. No similarity to known genes or proteins was observed despite the presence of

an extended segment with good coding capacity of 384 amino acids.

The other transcription units derived from hybrid-selected clones and cDNA library clones belonging to Classes II and III are presented in Table 3. Their composite sequences (with the possible exception of GT650) are colinear with genomic sequence. No extended open reading frames or consensus splice sites were identified within these sequences. No similarity with other sequences has been observed with the exception of single ESTs showing sequence identity with CG012 and CG030. Northern blot analysis has been performed for CG012, CG029, CG030, and CG036 and revealed distinct tissue-specific expression profiles (Fig. 2). CG030 is composed of a 4.9-kb contig derived from a single cDNA clone, 4 hybrid-selected clones, and 1 exon-trapped clone. Expression analysis identified two transcripts of 1.3 and 2.1 kb. The disparity between the expression data and the cDNA sequence data together with the absence of splice signals suggested that the 4.9-kb contig may contain chimeric cDNA cloning artifacts. It was noted that CG018, CG036, and CG016 are all located in a region of less than 100 kb; however, RNA hybridization analysis clearly reveals independent transcripts. GT573 is a transcription unit located in P1-931-D1 within an intron of the BRCA2 gene. GT573 consists of a 0.75-kb contig composed of three cDNA clones from cDNA libraries and the original hybrid-selected clone. GT573 has no significant homology to any previously described cDNA sequence.

CONCLUSIONS

Genome sequencing of prokaryotic genomes or whole chromosomes of *S. cerevisiae* or *C. elegans* has produced, as a matter of course, extended transcript maps. In contrast, extended human transcript maps are less detailed but have appeared most advanced where positional cloning projects of diseases or conditions have demanded the refined integration of physical data and candidate gene characterization. As these projects are driven by the endpoint set by the disease being studied, less emphasis is placed on the completion of the local transcription map. There are also the limitations caused by the lack of complete and oriented genomic sequence information or of the ability to analyze adequately the long stretches of DNA. The largest gene that was identified in this study, also corresponding to the largest mRNA and extending over 250 kb of genomic DNA, displayed strong similarity to putative open reading frames identified by the *C. elegans* and *S. cerevisiae* genomic sequencing projects. The extensive open reading frame of *C. elegans* spanned much less genomic distance and was used as a guide to complete the cloning of the entire long transcript of the human gene. When the genomic sequence of this human gene becomes polished, it will be interesting to compare the effectiveness of gene detection software in the human genomic sequence versus model organism genomic se-

quence. Further, this test case would provide for an evaluation of the extent of the complexity of genes that could be revealed based solely on the analysis of genomic sequence (i.e., can the relation of the additional and alternate sized transcripts be revealed?).

Although the retrieval of transcribed segments involves common procedures, the more challenging aspect of building transcription maps includes the assembly of retrieved sequences into transcription units. While the assembly in this study was supported by the generous amount of genomic sequence, extensive effort was still needed to verify contiguity of transcripts. This was, at least partially, a result of having observed several large and/or alternatively spliced transcripts. The genomic sequence did reveal that two transcription units identified were pseudogenes. A very serious limitation was noted with the restricted genomic sequencing initially performed. The long open reading frame of the large exon of the BRCA2 gene was missed by our first analysis and not recognized until more genomic sequence became available.

An integrated analysis of the transcribed sequences of the interval between *D13S1444* and *D13S1695* revealed a series of at least seven transcription units with extended and substantial coding potential, two pseudogenes, and at least nine additional transcription units as revealed by identification of short transcribed segments and RNA hybridization. Precise location for each transcription unit could be established from the refined physical map achieved with the overlapping P1, PAC, and BAC genomic clones. Evidence was obtained that the analysis was incomplete and that additional transcription units were likely to be present, as a large number of putatively transcribed contigs and a number of trapped exons did not fit into the assigned transcription units. The refined analysis of the region at and near the BRCA2 gene indicated that the majority of the transcribed contigs could be assigned if genomic sequence was available. Some of the transcription units simply may not have been characterized sufficiently to find their open reading frames. Efforts to attempt to link the transcribed contigs or exons, especially for those that appeared to group within short physical distances of each other, would aid in clarifying their existence. Anticipated completion of the genomic sequence by the Sanger and Washington University sequencing centers in combination with the transcription maps that were generated to find BRCA2 candidates will provide a more complete picture of the transcribed sequences at the *BRCA2* locus.

At least four genes, CG016, CG018, BRCA2, and a portion of CG003, are deleted in a pancreatic carcinoma that has been reported (Schutte *et al.*, 1995a,b). While involvement of BRCA2 in the development of pancreatic cancer is reasonable, additional roles for these other genes cannot be excluded.

ACKNOWLEDGMENTS

B.L.W. is supported by NIH Grants CA57601 and CA61231. F.J.C. is supported by NIH grant CA67403. J.M.R. and J.S. are Scholars

of the Medical Research Council (MRC) of Canada while F.L. is a MRC Career Scientist. J.S. and F.L. are supported by MRC and Endorecherche Inc. S.N., M.H.S., and D.E.G. are supported by U.S. Army Grant DAMD17-94-J-4260 and NIH Grants CA55914, CA48711, CA65673, RR00064, CN05222, and CA42014. J.M.R. is a member of the Canadian Genetic Diseases Network.

REFERENCES

- Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410.
- Brody, L. C., Abel, K. J., Castilla, L. H., Couch, F. J., McKinley, D. R., Yin, G., Ho, P. P., Merajver, S., Chandrasekharappa, S. C., Xu, J., Cole, J. L., Struewing, J. P., Valdes, J. M., Collins, F. S., and Weber, B. L. (1995). Construction of a transcription map surrounding the *BRCA1* locus of human chromosome 17. *Genomics* 25: 238–247.
- Buckler, A. J., Chang, D. D., Graw, S. L., Brook, J. D., Haber, D. A., Sharp, P. A., and Housman, D. E. (1991). Exon amplification: A strategy to isolate mammalian genes based on RNA splicing. *Proc. Natl. Acad. Sci. USA* 88: 4005–4009.
- Church, D. M., Stotler, C. J., Rutter, J. L., Murrell, J. R., Trofatter, J. A., and Buckler, A. J. (1994). Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nature Genet.* 6: 98–105.
- Collins, F., and Galas, D. (1993). A new five-year plan for the U.S. Human Genome Project. *Science* 262: 43–46.
- Harshman, K., Bell, R., Rosenthal, J., Katcher, H., Miki, Y., Swensen, J., Gholami, Z., Frye, C., Ding, W., and Dayananth, P. (1995). Comparison of the positional cloning methods used to isolate the *BRCA1* gene. *Hum. Mol. Genet.* 4: 1259–1266.
- Hochgeschwender, U. (1992). Toward a transcriptional map of the human genome. *Trends Genet.* 8: 41–44.
- Lovett, M., Kere, J., and Hinton, L. M. (1991). Direct selection: A method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci. USA* 88: 9628–9632.
- Neuhausen, S. L., Swensen, J., Miki, Y., Liu, Q., Tavtigian, S., Shattuck-Eidens, D., Kamb, A., Hobbs, M. R., Gingrich, J., Shizuya, H., Kim, U. J., Cochran, C., Futreal, P. A., Wiseman, R. W., Lynch, H. T., Tonin, P., Narod, S., Cannon-Albright, L., Skolnick, M. H., and Goldgar, D. (1994). A P1-based physical map of the region from D17S776 to D17S78 containing the breast cancer susceptibility gene *BRCA1*. *Hum. Mol. Genet.* 3: 1919–1926.
- Parimoo, S., Patanjali, S. R., Shukla, H., Chaplin, D. D., and Weissman, S. M. (1991). cDNA selection: Efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc. Natl. Acad. Sci. USA* 88: 9623–9627.
- Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85: 2444–2448.
- Rommens, J., Durocher, F., McArthur, J., Tonin, P., Leblanc, J-F., Allen, T., Samson, C., Ferri, L., Narod, S., Morgan, K., and Simard, J. (1995). Generation of a transcription map at the HSD17B locus centromeric to *BRCA1* at 17q21. *Genomics* 28: 530–542.
- Rommens, J. M., Mar, L., McArthur, J., Tsui, L.-C., and Scherer, S. (1994). Towards a transcriptional map of the q21–q22 region of chromosome 7. In "Identification of Transcribed Sequences" (U. Hochgeschwender and K. Gardiner, Eds.), pp. 65–79, Plenum, New York.
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74: 5463–5467.
- Schutte, M., da Costa, L. T., Hahn, S. A., Moskaluk, C., Hoque, A. T. M. S., Rozenblum, E., Weinstein, C. L., Bittner, M., Meltzer, P. S., Trent, J. M., Yeo, C. J., Hruban, R. H., and Kern, S. (1995a). Identification by representational difference analysis of a homozygous deletion in pancreatic carcinoma that lies within the *BRCA2* region. *Proc. Natl. Acad. Sci. USA* 92: 5950–5954.
- Schutte, M., Rozenblum, E., Moskaluk, C. A., Guan, X., Hoque, A. T., Hahn, S. A., da Costa, L. T., de Jong, P. J., and Kern, S. E. (1995b). An integrated high-resolution physical map of the DPC/BRCA2 region at chromosome 13q12. *Cancer Res.* 55: 4570–4574.
- Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci. USA* 89: 8794–8797.
- Smith, S. W., Overbeek, R., Woese, C. R., Gilbert, W., and Gillevet, P. M. (1994). The genetic data environment an expandable GUI for multiple sequence analysis. *Comput. Appl. Biosci.* 10: 671–675.
- Swensen, J. (1996). PCR with random primers to obtain sequence from yeast artificial chromosome insert ends or plasmids. *Biotechniques* 20: 486–491.
- Tavtigian, S. V., Simard, J., Rommens, J., Shattuck-Eidens, D., Couch, F., Neuhausen, S., Merajver, S., Thorlacius, S., Offit, K., Stoppa-Lyonnet, D., Belanger, C., Bell, R., Berry, S., Bogden, R., Chen, Q., Davis, T., Dumont, M., Frye, C., Hattier, T., Jammulapati, S., Janecki, T., Jiang, P., Kehrer, R., Leblanc, J. F., Mitchell, J. T., Peng, Y., Samson, C., Schroeder, M., Snyder, S., Stringfellow, M., Stroup, C., Swedlund, B., Swensen, J., Teng, D., Thomas, A., Tran, T., Tranchant, M., Weaver-Feldhaus, J., Wong, A. K. C., Shizuya, H., Eyfjord, J. E., Cannon-Albright, L., Labrie, F., Skolnick, M., Weber, B., Kamb, A., and Goldgar, D. E. (1996). The complete *BRCA2* gene and mutations in chromosome 13q-linked kindreds. *Nature Genet.* 12: 333–337.
- Teng, D. H-F, Bogden, R., Mitchell, J., Baumgard, M., Bell, R., Berry, S., Davis, T., Ha, P. C., Kehrer, R., Jammulapati, S., Chen, Q., Offit, K., Skolnick, M. H., Tavtigian, S. V., Jhanwar, S., Swedlund, B., Wong, A. K. C., and Kamb, A. (1996). Low incidence of *BRCA2* mutations in breast carcinoma and other cancers. *Genomics* 13: 241–244.
- Thorlacius, S., Tryggvadottir, L., Olafsdottir, G. H., Jonasson, J. G., Ogmundsdottir, H. M., Tulinius, H., and Eyfjord, J. E. (1995). Linkage to *BRCA2* region in hereditary male breast cancer. *Lancet* 346: 544–545.
- Unerbacher, E. C., and Mural, R. (1991). Locating protein-coding regions in human DNA sequences by multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* 88: 11261–11265.
- Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., Micklem, G., Barfoot, R., Hamoudi, R., Patel, S., Rice, C., Biggs, P., Hashim, Y., Smith, A., Connor, F., Arason, A., Gudmundsson, J., Ficenec, D., Kelsell, D., Ford, D., Tonin, P., Bishop, D. T., Spurr, N. K., Ponder, B., Eeles, R., Peto, J., Devilee, P., Cornelisse, C., Lynch, H., Narod, S., Lenoir, G., Eglisson, V., Barkadottir, R. B., Easton, D. F., Bentley, D. R., Futreal, P. A., Ashworth, A., and Stratton, M. R. (1995). Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* 378: 789–792.
- Wooster, R., Neuhausen, S. L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T., Averill, D., Fields, P., Marshall, G., Narod, S., Lenoir, G. M., Lynch, H., Feunteun, J., Devilee, P., Cornelisse, C. J., Menko, F. H., Daly, P. A., Orminston, W., McManus, R., Pye, C., Lewis, C. M., Cannon-Albright, L. A., Peto, B., Ponder, Skolnick, M. H., Easton, D. F., Goldgar, D., and Stratton, M. (1994). Localization of a breast cancer susceptibility gene, *BRCA2*, to chromosome 13q12–13. *Science* 265: 2088–2090.