Complete genomic sequence and analysis of 117 kb of human DNA containing the gene BRCA1.

T M Smith, M K Lee, C I Szabo, et al.

Genome Res. 1996 6: 1029-1049
Access the most recent version at doi:10.1101/gr.6.11.1029

RESEARCH

Complete Genomic Sequence and Analysis of 117 kb of Human DNA Containing the Gene *BRCA1*

Todd M. Smith,[1] Ming K. Lee,[2] Csilla I. Szabo,[2] Nicole Jerome,[1] Mark McEuen,[1] Matthew Taylor,[1] Leroy Hood,[1] and Mary-Claire King[2,3]

[1]Department of Molecular Biotechnology and [2]Departments of Medicine and Genetics, University of Washington Medical School, Seattle, Washington 98195

Over 100 distinct disease-associated mutations have been identified in the breast–ovarian cancer susceptibility gene *BRCA1*. Loss of the wild-type allele in >90% of tumors from patients with inherited *BRCA1* mutations indicates tumor suppressive function. The low incidence of somatic mutations suggests that *BRCA1* inactivation in sporadic tumors occurs by alternative mechanisms, such as interstitial chromosomal deletion or reduced transcription. To identify possible features of the *BRCA1* genomic region that may contribute to chromosomal instability as well as potential transcriptional regulatory elements, a 117,143-bp DNA sequence encompassing *BRCA1* was obtained by random sequencing of four cosmids identified from a human chromosome 17 specific library. The 24 exons of *BRCA1* span an 81-kb region that has an unusually high density of *Alu* repetitive DNA (41.5%), but relatively low density (4.8%) of other repetitive sequences. *BRCA1* intron lengths range in size from 403 bp to 9.2 kb and contain the intragenic microsatellite markers D17S1323, D17S1322, and D17S855, which localize to introns 12, 19, and 20, respectively. In addition to *BRCA1*, the contig contains two complete genes: *Rho7*, a member of the *rho* family of GTP binding proteins, and *VAT1*, an abundant membrane protein of cholinergic synaptic vesicles. Partial sequences of the *IA1–3B* B-box protein pseudogene and *IFP 35*, an interferon induced leucine zipper protein, reside within the contig. An *L21* ribosomal protein pseudogene is embedded in *BRCA1* intron 13. The order of genes on the chromosome is: centromere–*IFP 35–VAT1–Rho7–BRCA1–IA1–3B*– telomere.

[The sequence data described in this paper have been submitted to GenBank under accession no. L78833.]

Inherited mutations in the breast–ovarian cancer susceptibility gene, *BRCA1* (Hall et al. 1990; Miki et al. 1994), confer a lifetime risk of breast cancer greater than 80% and an increased risk of ovarian cancer (Newman et al. 1988; Easton et al. 1993; Ford et al. 1994). Evidence for tumor suppressive function of *BRCA1* derives from the high proportion (87%) of truncating germ-line mutations, which likely represent loss-of-function alterations (Castilla et al. 1994; Friedman et al. 1994b; Futreal et al. 1994; Miki et al. 1994; Simard et al. 1994; Friedman et al. 1995b; Gayther et al. 1995, 1996; Hogervorst et al. 1995; Hosking et al. 1995; Merajver et al. 1995; Shattuck-Eidens et al. 1995; Struewing et al. 1995; Szabo and King 1995; Takahashi et al. 1995; Couch et al. 1996; Durocher et al. 1996; FitzGerald et al. 1996; Johansson et al. 1996; Langston et al. 1996; Serova et al. 1996); loss of the wild-type allele in >90% of breast and ovarian patients with inherited *BRCA1* mutations (Smith et al. 1992; Friedman et al. 1994a; Neuhausen and Marshall 1994); decreased levels of *BRCA1* expression in breast (Thompson et al. 1995) and ovarian (R. Hernandez, M. Skelly, C. Laird, M.-C. King, and A. Gown, in prep.) tumors from patients not selected for family history; accelerated growth of normal and malignant mammary epithelial cells upon experimental inhibition of *BRCA1* expression with antisense oligonucleotides (Thompson et al. 1995); and inhibition of malignant breast and ovarian cancer cell growth in culture as well as suppression of MCF7 breast cancer cell tumorigenesis in mice by

[3]Corresponding author.
E-MAIL mcking@u.washington.edu; FAX (206) 616-4295.

overexpression of wild-type *BRCA1* (Holt et al. 1996).

Few somatic point mutations or frame-shift alterations have been identified thus far (Futreal et al. 1994; Hosking et al. 1995; Merajver et al. 1995; Takahashi et al. 1995), suggesting that somatic inactivation of *BRCA1* may occur through different mechanisms, such as interstitial chromosomal deletion or epigenetic silencing of *BRCA1* expression. Gross somatic mutations at *BRCA1* are frequently found in breast and ovarian tumors from patients not selected for family history. Loss of heterozygosity (LOH) in the *BRCA1* region ranges from 40% to 80% among sporadic breast carcinomas (Cropp et al. 1993; Saito et al. 1993; Ford et al. 1994) and from 30% to 70% among sporadic ovarian carcinomas (Russell et al. 1990; Cliby et al. 1993; Yang-Feng et al. 1993; Takahashi et al. 1995). *BRCA1* transcript expression in breast carcinomas of patients not selected for family history ranges from none detectable to <50% of normal levels (Thompson et al. 1995) and protein expression in ovarian carcinomas is lost (R. Hernandez, M. Skelly, C. Laird, M.-C. King, and A. Gown, in prep.).

Characterization of the genomic sequence of *BRCA1* may lead to identification of putative transcriptional regulatory sequences and structural elements that potentially contribute to chromosomal instability. We present the complete sequence and analysis of a 117,143-bp region of human chromosome 17 that encompasses the 81-kb *BRCA1* gene.

RESULTS

Mapping BRCA1 Cosmid Clones

Clones containing portions of the *BRCA1* gene were isolated from a chromosome 17 specific cosmid library by screening arrayed filters with probes prepared from PCR products of *BRCA1* exons amplified from human genomic DNA (Friedman 1994b). Overlapping cosmids were identified by PCR using the same primer pairs as for the preparation of the probes. From these screens, 10 overlapping cosmids were identified that spanned the *BRCA1* gene (Table 1), four of which (*BRCA1*-5, 1-7, 1-8, and 1-9) were selected for sequencing.

Sequencing

The four overlapping cosmids containing portions of the *BRCA1* gene were sequenced from randomly selected M13 subclones (Deininger 1983; Wilson et al. 1994) using fluorescence based automated detection (Hunkapiller et al. 1991). The 117,143-bp sequence (Fig. 1; GenBank accession no. L78833) was assembled from three separate overlapping contigs of 42 kb, 36 kb, and 40 kb. For each contig, between 1300 to 1500 ABI trace files were analyzed with the phred base-calling program (P. Green and B. Ewing, unpubl.; see Methods). The resulting sequence strings were assembled using phrap (P. Green, unpubl.) to give an average redundancy of eight- to tenfold for the entire region after vector-containing and low-

Table 1. BRCA1 Cosmid Contig

	P-8	P-1	exon1	exon2	exon3	exon5	exon6	exon7	exon8	exon9	exon11C	exon11P	exon12	D17S1323	exon13	exon14	exon15	exon16	exon17	exon18	exon19	D17S1322	exon20	D17S855	exon21	exon22	exon23	exon24	
43H2	+	+	+	+																									BRCA1-2
48C3	+	+	+	+	+																								BRCA1-3
154B12	+	+	+	+	+	+																							BRCA1-4
145E5	+	+	+	+	+	+	+	+																					BRCA1-10
73F5	+	+	+	+	+	+	+	+	+	+	+	+	+																BRCA1-5
141H4						+	+	+	+	+	+	+	+	+	+	+	+												BRCA1-6
14F8														+	+	+	+	+	+	+	+	+	+						BRCA1-7
38H2																		+	+	+	+	+	+	−	+	+	+		BRCA1-1
87E7																				+	+	+	+	+	+	+	+	+	BRCA1-8
90F2																							+	+	+	+	+	+	BRCA1-9

Chromosome 17 specific cosmid library designations and contig identifiers used in the text are indicated. Primers used to map the extent of the cosmid inserts are described for BRCA1 (Friedman et al. 1994b, 1995b) and microsatellites (Anderson et al. 1993; Neuhausen et al. 1994).

quality sequences were removed. Complete assembly of the 36-kb contig (*BRCA1*-7/8) was possible with the combined data from two cosmids (*BRCA1*-7 and *BRCA1*-8), whereas the 42-kb (*BRCA1*-5) and 40-kb (*BRCA1*-9) contigs required additional data to close gaps that remained after an initial phase of random sequencing. To obtain sequence spanning the gaps, primers were designed from existing data, and single-stranded M13 templates known to contain the region of interest were sequenced using dye terminator chemistry (Wilson et al. 1994). Assembly accuracy was verified by comparison of restriction enzyme digests of the cosmid DNA (data not shown) to computer-generated restriction maps, and by alignment with known cDNA sequences.

Accuracy of the sequence data was estimated using three methods: observing discrepancies in overlapping regions between cosmid clones, observing discrepancies between genomic sequence data and published cDNA sequences from this region, and estimating quality values using phred and phrap for each nucleotide of the sequence. The 946-bp overlap between *BRCA1*-5 and *BRCA1*-7/8 did not show any discrepancies. Between *BRCA1*-7/8, and *BRCA1*-9 there are two overlapping regions (resulting from a 16-kb deletion in *BRCA1*-8) of 2500 bp and 7467 bp. The 2500-bp overlap did not show any discrepancies, and the 7467 bp overlap showed four discrepancies: Two were a result of sequencing errors that were resolved by replacing the consensus but low-quality nucleotides with nucleotides assigned high-quality values; the other two likely represent haplotype variants. These data indicate an error estimate of less than one error in 3000 bases. When the 5711-bp *BRCA1* cDNA sequence (Miki et al. 1994); (GenBank accession no. U14680) was aligned with the germ-line sequence, six discrepancies [five mismatches and one insertion/deletion (indel)] were detected. Of these, only one mismatch in the coding sequence was definitely a result of an incorrect base in the germ-line consensus sequence. The other four mismatches and indel are all in the 5′ untranslated region (UTR) and may represent either sequencing discrepancies or polymorphisms. Differences between the genomic sequence and published cDNA sequences of two other genes within the contig, *Rho7* (GenBank accession no. X95456) and *VAT1* (GenBank accession no. U18009), were confirmed to be true discrepancies between the data sets.

Based on automated quality values from phrap, which include quality values obtained from analysis of the peaks in a sequence trace by phred and confirmation of bases from overlaps between individual sequence strings within the project, ~95% of the data is estimated to have one error or less in 10,000 bases, and 98% of the data is estimated to be >99.9% accurate. Regions that are sequenced on both strands (~99% of the total 117,143-bp contig) result in the highest quality values. The remaining 1% of the sequence that was determined from a single strand contained at least three overlapping high-quality sequences

Figure 1 Global analysis of the *BRCA1* contig based on the data bases available, July 1996. (*A*) Graph of the frequency of CpG dinucleotides divided by the frequency of GpC dinucleotides. Dinucleotide frequencies were calculated in a 1000-nucleotide window moved at 100-nucleotide intervals using the program CpG (T. Smith, unpubl.). (*B*) Structure of the genes and gene fragments found in the 117,143-bp contig. (*C*) Distribution of *Alu* interspersed repeat elements in the contig. The human specific (*AluY*) elements are highlighted in red. (*D*) Distribution of non-*Alu* interspersed repeats, including L1, Mer, Mir, and Mlt families, and a complete LTR element. (*E*) All simple-sequence repeats of ⩾10 nucleotides found using the program sputnik (C. Abajian, unpubl.); STS repeats known to be polymorphic are highlighted in red. (*F*) Locations of exons predicted by GRAIL (Xu et al. 1994). (*G*) Locations of exons predicted by genefinder (P. Green, unpubl.). (*H*) Blastn search of the 117,143-bp DNA sequence against dbEST (NCBI). For this search interspersed repeats were masked (by replacing repeat region sequence with strings of x's or n's) using cross_match (P. Green, unpubl.) and a library of interspersed repeat sequences (A. Smit, unpubl.) (masked repeats displayed in *C* and *D*). The resulting sequence data was used to search dbEST with blastn (Altschul et al. 1990) and the accession numbers of the sequences (subject) predicted to align with the contig sequence (query) by blastn were obtained over the network and used to make a library of subject sequences. Cross_match was then used to refine the alignments by searching the query sequence against the library of subject sequences. The three steps of this process were automated with FindMatches (T. Smith, unpubl.; see Methods). The results from the analysis are plotted as a three-dimensional histogram where the *x*-axis represents the contig, the *y*-axis gives the number of sequences aligning in a particular region, and the *z*-axis lists bins of decreasing identity between the query and subject sequences. (*I*) Blastn search against the nonredundant version (nr) of GenBank (NCBI). This analysis was carried out as in *H*.

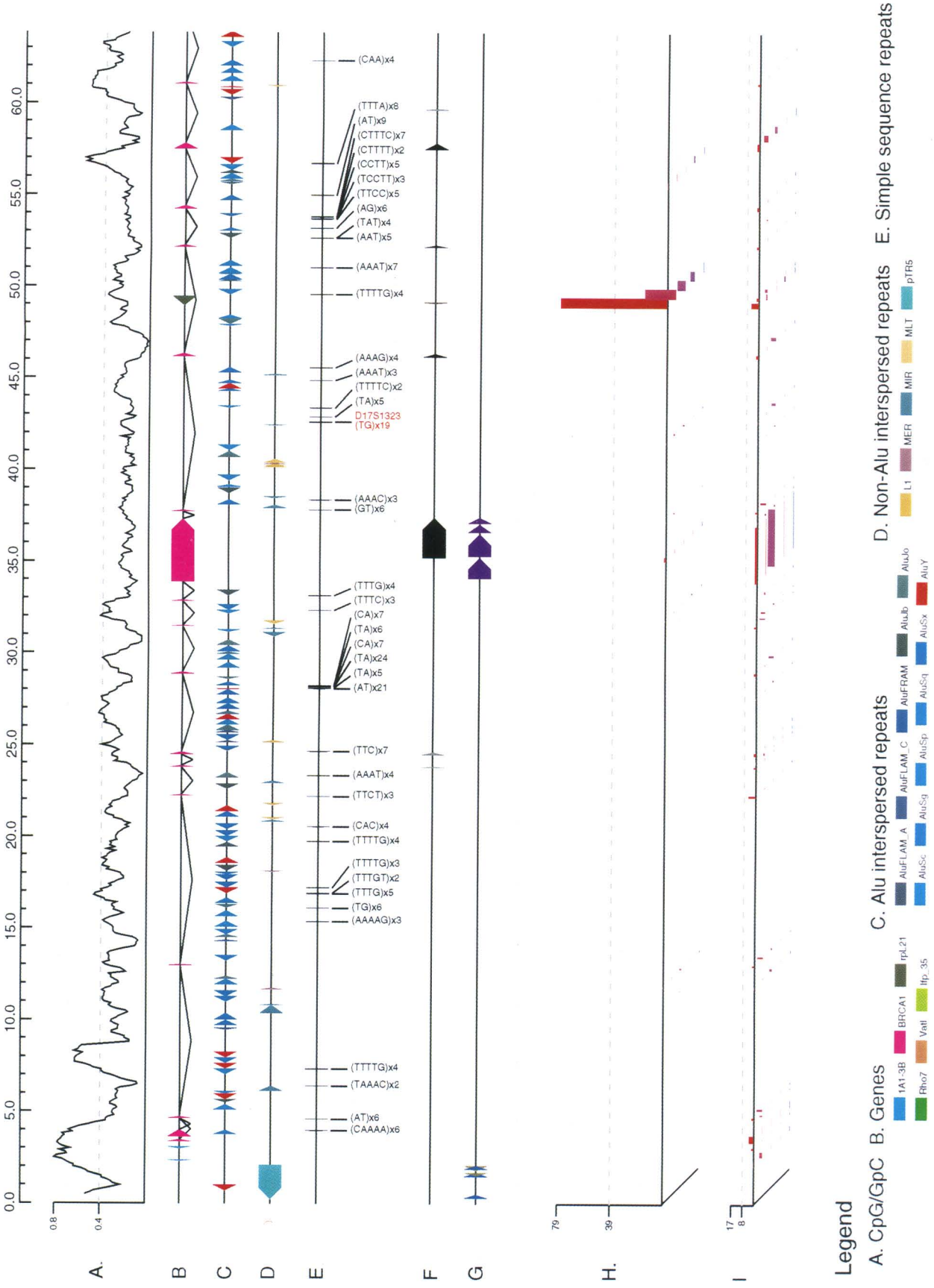

Figure 1 *(See previous page for legend; figure continued on facing page)*

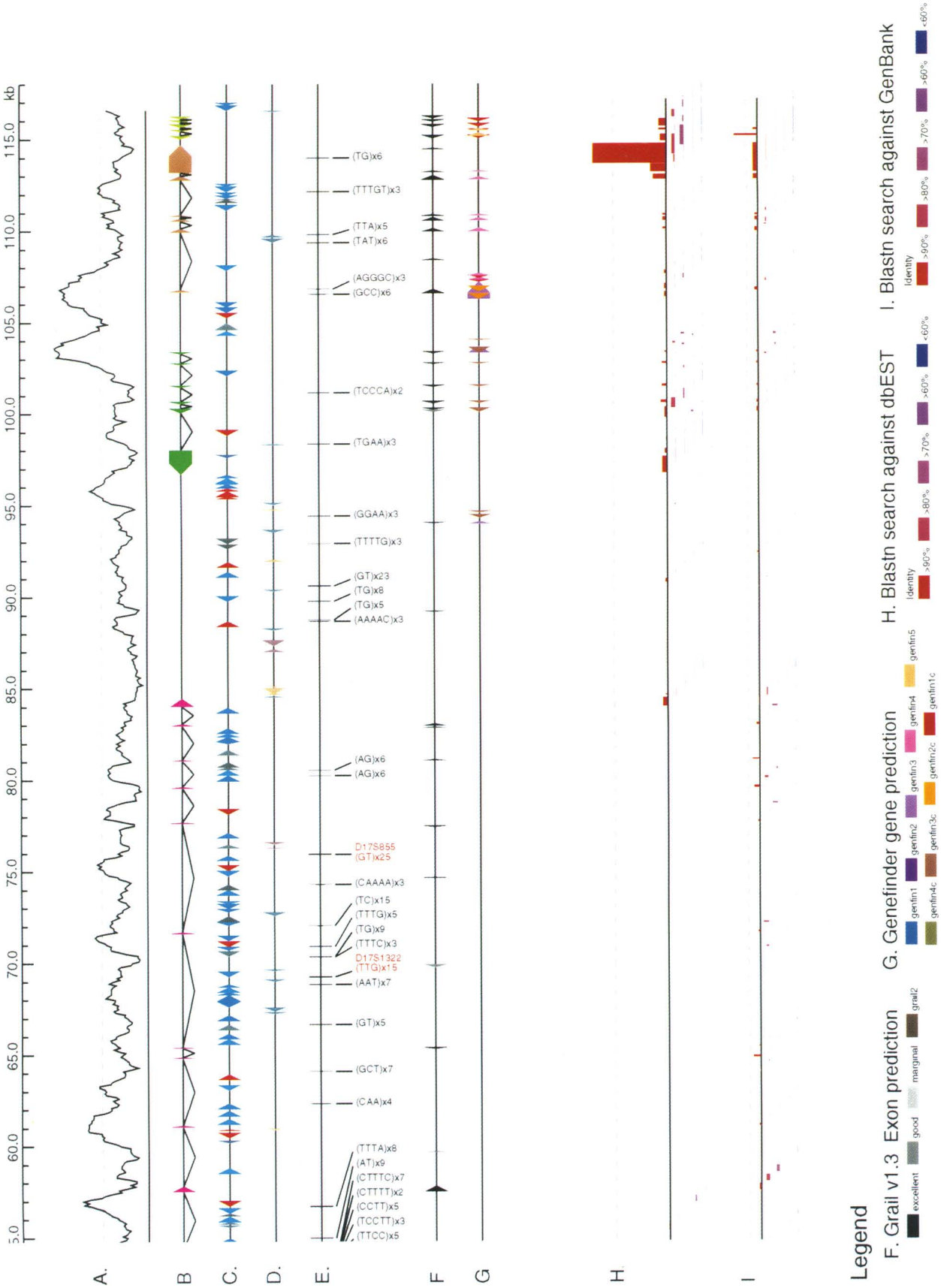

Figure 1 (*Continued*)

from independent clones. From these estimates the sequence should contain approximately one error in 3000–6000 bases.

Analysis of the 117,143-bp Contig

Repetitive Elements

Low-complexity DNA sequences containing interspersed repetitive elements and simple repeating units or homopolynucleotide stretches were identified using cross_match (P. Green, unpubl.) and a library of human repeat sequences (A. Smit, unpubl.; see Methods) and masked prior to carrying out data-base searches. This analysis identified 138 individual *Alu* elements within the *BRCA1* gene, comprising 41.5% of the 81-kb sequence (Fig. 1, Table 2). Ninety-four of these *Alu* elements are complete (containing >90% of the consensus length in the alignment), whereas the remaining 44 represent partial elements ranging in length from 69–231 nucleotides of 70–100% sequence homology to the consensus sequences of *Alu* subfamilies (Deininger 1989; Jurka 1995; Batzer et al. 1996). The density of *Alu*s is significantly lower in the regions flanking *BRCA1*, with 32 elements comprising 21.8% of the flanking sequence. In addition to the *Alu* sequences, 46 other interspersed repetitive elements comprise 6.6% of the contig, including fragments belonging to the L1, Mer, Mir, and Mlt families, as well as a complete long terminal repeat (LTR) element, pTR5 (La Mantia et al. 1989), at the 5′ end of the *BRCA1* contig. L1 fragments occur more frequently within the *BRCA1* gene, whereas Mer and Mlt elements are more frequent in the flanking sequences (Table 2). A total of 41.9% of the *BRCA1* contig is made up of interspersed repetitive elements. With the exception of a single *Alu* half element (FLAM) in the *Rho7* gene (described below) none the repetitive elements appear to be in naturally transcribed regions within the contig. The data masking procedure with cross_match is very efficient: Only a few regions of similarity were identified in the data-base searches by virtue of repeat elements. Without masking, well over 300 sequence similarities to repetitive elements with blastn P values $<10^{-70}$ were identified throughout the contig.

Table 2. Interspersed Repeats in the *BRCA1* Contig

	BRCA1 gene (nucleotides 3,344–84,436)			Rest of locus (nucleotides 1–3,343, 84,435–117,143)		
Subfamily	**number**	**bases**	**% region**	**number**	**bases**	**% region**
AluFLAM_A	1	106	0.1	0	0	0.0
AluFLAM_C	3	358	0.4	0	0	0.0
AluFRAM	1	69	0.1	1	170	0.5
AluJb	12	2925	3.6	4	773	2.1
AluJo	20	4066	5.0	1	369	1.0
AluSc	6	1655	2.0	2	298	0.8
AluSg	13	3202	4.0	2	354	1.0
AluSp	15	4305	5.3	3	900	2.5
AluSq	14	3155	3.9	4	1,046	2.9
AluSx	37	9545	11.8	7	1,675	4.7
AluY	16	4255	5.3	8	2,086	5.8
Total *Alu*	138	33,641	**41.5**	32	7,671	**21.3**
L1	7	969	1.2	1	107	0.3
MER	5	374	0.5	2	478	1.3
MIR	17	2,352	2.9	8	948	2.6
MLT	1	95	0.1	2	412	1.1
SVA	1	108	0.1	1	50	0.1
pRT5 (LTR)	0	0	0	1	1849	5.1
Total non-*Alu*	31	3,898	**4.8**	15	3,844	**10.5**

In addition to the interspersed repetitive elements, 68 simple sequence repeats (SSRs) of at least 10 nucleotides in length were detected by the sputnik program (C. Abajian, unpubl.; see Methods) (Fig. 1). Of these, 15 form six overlapping blocks with two or more different or alternating repeat unit types and 43 are individual repeats extending 15 bp or more. Thirty-two of the SSRs are not components of *Alu* elements (Table 3).

Comparison of Repeat Elements in BRCA1 *and Other Genes*

The distribution of repeat elements within *BRCA1* was compared with that of other genes selected from version 95.0 of the GenBank data base (Benson et al. 1996). First, all entries in GenBank with the words "human" or "*Homo sapiens*" along with the word "complete" on the definition line were identified. Entries identified by this method but which contained no coding sequence were discarded. The remaining sequences were screened to eliminate redundant entries. If more than one entry contained DNA sequences of the same gene, the most recent entry was retained and all others were discarded. Finally, any entry lacking an indication of introns was also discarded. After this screening process, 326 entries were retained for analysis. Sequences between the first base of the first exon and the last base of the last exon in each entry were analyzed and compared with the sequence within the same boundaries of *BRCA1*.

Table 3. Non-*Alu* SSRs in the *BRCA1* Contig

Repeat type	Begin (nucleotide)	End (nucleotide)	Gene	Intron	Marker
(AT)×6	4587	4599	*BRCA1*	1	
(TAAAC)×2	6395	6409	*BRCA1*	2	
(TTCT)×3	22175	22189	*BRCA1*	3	
(AAAT)×4	23317	23333	*BRCA1*	5	
(AT)×21	28032	28075	*BRCA1*	7	
(TA)×5	28088	28099	*BRCA1*	7	
(TA)×24	28125	28174	*BRCA1*	7	
(CA)×7	28174	28188	*BRCA1*	7	
(TA)×6	28196	28208	*BRCA1*	7	
(CA)×7	28208	28222	*BRCA1*	7	
(GT)×6	37797	37810	*BRCA1*	12	
(TG)×19	42582	42621	*BRCA1*	12	D17S1323
(TA)×5	42876	42887	*BRCA1*	12	
(AAAT)×3	44853	44868	*BRCA1*	12	
(TTCC)×5	53645	53665	*BRCA1*	14	
(TCCTT)×3	53666	53684	*BRCA1*	14	
(CCTT)×5	53687	53708	*BRCA1*	14	
(CTTTT)×2	53754	53768	*BRCA1*	14	
(GCT)×7	64072	64093	*BRCA1*	17	
(TTG)×15	69217	69263	*BRCA1*	19	D17S1322
(GT)×25	75906	75956	*BRCA1*	20	D17S855
(TG)×5	88724	88735	intergenic		
(TG)×8	89743	89759	intergenic		
(GT)×23	90565	90611	intergenic		
(GGAA)×3	94405	94418	intergenic		
(TGAA)×3	98327	98342	intergenic		
(TCCCA)×2	101127	101141	*Rho7*	3	
(GCC)×6	106513	106531	intergenic		
(AGGGC)×3	106804	106819	intergenic		
(TAT)×6	109336	109355	*VAT1*	1	
(TTA)×5	109784	109799	*VAT1*	1	
(TG)×6	113978	113990	*VAT1*	3′UTR	

The length of analyzed sequences for these entries ranged from 436 to 175,019 bases. Overall, *Alu*s and miscellaneous types of repeats occur with about the same frequencies. Given the different distribution of intron lengths in the genes, and that repeat elements most frequently occur in intronic sequences, the percentage of repeats were compared with the percentage of intronic sequence for each entry. The percentage of repeats is modestly correlated with the percentage of intronic sequences [correlation (r^2) = 0.52]. *BRCA1* is a relatively large gene with introns comprising 90.9% of its sequence; 46.3% of its sequence consists of repeat elements. Among the 14 large genes (>30,000 bases) in the analyzed set, the average percentage of intronic sequence is 89.8% (range: 70.3% to 98.41%) while the average repeat element content is 30.39% (range: 3.39% to 50.76%). Only three genes had higher overall *Alu* densities than *BRCA1*: apolipoprotein c-I (VLDL; GenBank accession no. M20903) with 60.8% *Alu*s; Blym transforming gene (GenBank accession no. K01884) containing 53.7% *Alu*s, and apolipoprotein c-IV (APOC4; GenBank accession no. HSU32576) with 41.3% *Alu* composition.

The BRCA1 *Gene*

Comparative analysis of the genomic sequence for *BRCA1* and the cDNA sequence (Miki et al. 1994) revealed the positions of all 24 exons (Fig. 2A, Table 4). Characterization of an aberrant *BRCA1* cDNA clone in the original report (Miki et

Table 4. Positions of Exons for Genes in the *BRCA1* Contig

	Start	End	Exon		Start	End	Exon
BRCA1				*Rho7*			
	3344	3464	1a		98031	96693	6[c]
	3621	3998	1b		100302	100054	5[c]
	4620	4718	2		100673	100539	4
	12955	13008	3		101551	101442	3
	22201	22278	5		102774	102687	2
	23778	23866	6		103386	103285	1[c]
	24473	24612	7	*VATI*			
	28853	28958	8[a]		106659	106789	1
	31443	31488	9[a]		109927	110134	2
	32810	32886	10[a]		110520	110690	3
	33872	37297	11[a]		110796	110885	4
	37700	37788	12[a]		112774	113015	5
	46156	46327	13[a]		113178	114716	6
	52118	52244	14	*IFP35*			
	54211	54401	15		115216	115033	5
	57494	57804	16		115546	115440	4
	61038	61125	17		115846	115660	3
	64782	64859	18[a]		116055	115949	2
	65360	65400	19[a]		116275	116128	1
	71598	71681	20[a]				
	77620	77674	21				
	79543	79616	22[a]				
	81034	81094	23[a]				
	82936	83872	24[b]				
	84012	84436	24[b]				

[a]Exons are shifted relative to those reported in the *BRCA1* cDNA sequence (U14680) so that the introns begin with a GT donor sequence and end with an AG acceptor sequence.
[b]The 3′-UTR region is predicted by alignment with available human 3′-UTR sequence data (U68041) and a mouse cDNA sequence (U36475). A poly(A) signal sequence is located at nucleotide 84413.
[c]In the *Rho7* gene the final exon (6) is predicted by alignment with EST sequences and contains a polyA signal sequence. Exon 5 is possibly truncated, and the first exon positions are based on alignment with the *Rho7* cDNA sequence (X95456).

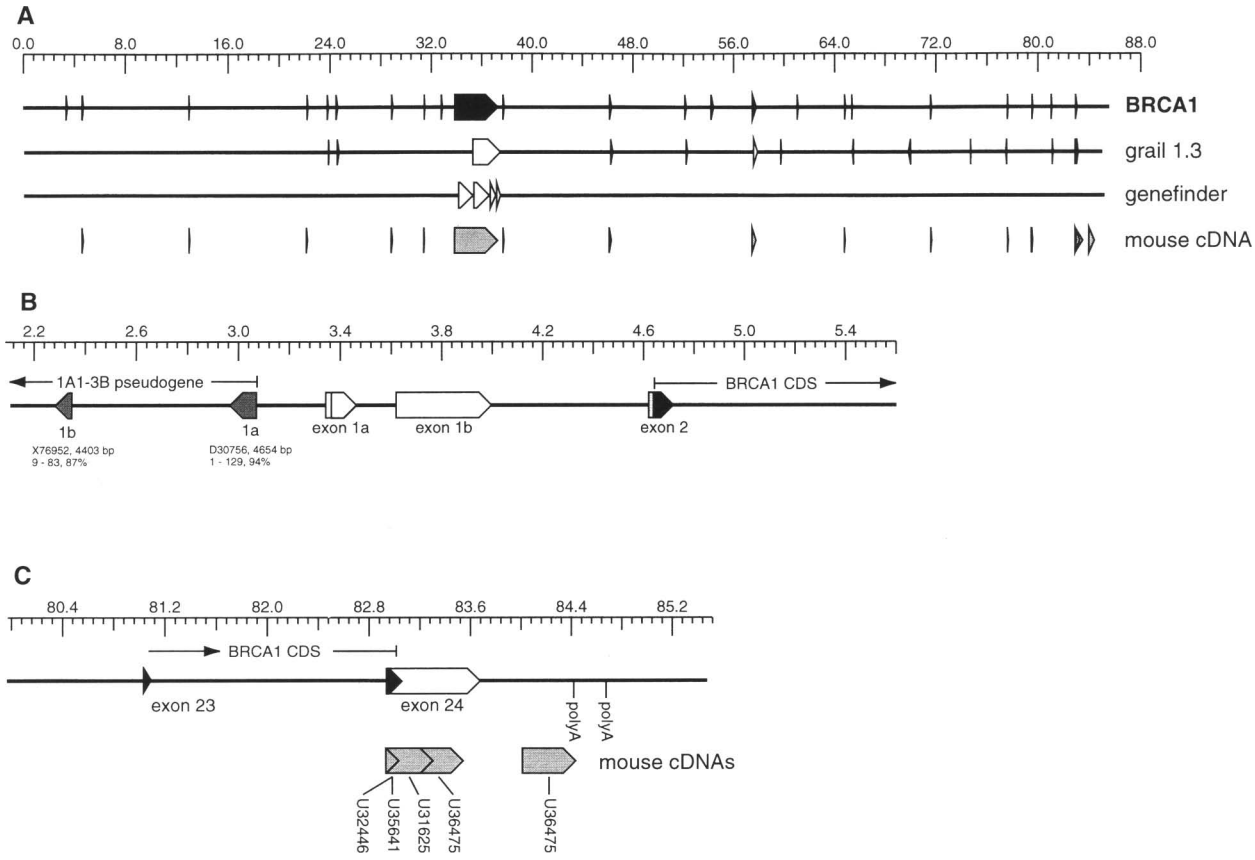

Figure 2 Map of the *BRCA1* gene. (*A*) Structure of the gene predicted by alignment with the human *BRCA1* cDNA (U14680), dark boxes. Below this alignment are exons predicted by grail and genefinder from the complete genomic sequence. The shaded boxes at the bottom represent alignments with five mouse *BRCA1* cDNAs (U32446, U35641, U31625, U36475, and U68174). (*B*) Expansion of the 5′ end of *BRCA1* showing the alternate starting exons of the *BRCA1* gene and exons 1a and 1b of the *1A1-3B* pseudogene. (*C*) Expansion of the 3′ end of the *BRCA1* gene showing the 3′ UTRs, corresponding mouse cDNA sequences, and poly(A) signal sequence sites.

al. 1994) led to the misidentification of an inserted *Alu* element as exon 4. Not normally found in *BRCA1* transcripts, insertion of this *Alu* would lead to introduction of a STOP codon. Hence, *BRCA1* exons and introns are numbered 1a, 1b, 2, 3, 5, 6, and so on. Most of the exon/intron boundaries were accurately predicted in the published cDNA sequence (HSU14680); however, 11 boundaries required adjustment with the available genomic data to preserve the nearly invariant consensus 5′ GT and 3′ AG dinucleotides of intron boundaries (Table 4; GenBank accession no. L78833). The coding regions between the two sequences were in perfect agreement at both the nucleotide and amino acid levels.

The *BRCA1* gene is conserved in mammals (Miki et al. 1994), and five complete cDNA sequences from the mouse have been determined (GenBank accession nos. U31625, U32446, U35641, U36475, and U68174). All these sequences show a high degree of similarity when aligned against the genomic human sequence using the program cross_match (P. Green, unpubl.). Overall the human and mouse sequences are 76% identical at the nucleotide level with exons 2 (87%), 3 (90%), 5 (90%), 12 (85%), 19 (91%), and 21 (87%) containing the highest identities and the identities of the other exons ranging from 68% to 83%. The U36475 cDNA sequence identified two additional 3′ regions of high similarity between human and mouse (Fig. 2C), one at nucleotides 83,326–83,545 (two windows of 95% identity at nucleotides 83,326–83,349, and 74% identity at nucleotides 83,432–83,545), and the other at nucleotides 84,012–84,436 (80% identity). These homologies overlap the 833-bp

segment of 3′-UTR sequence (96% identity to nucleotides 83,061–83,893) derived from a human placental *BRCA1* cDNA clone (Friedman et al. 1995b; GenBank accession no. U68041). The presence of a putative poly(A) signal sequence at nucleotide 84,416 and the high similarity of the human and mouse sequences suggests that nucleotides 83,061–84,416 correspond to the 3′ UTR.

A 3798-nucleotide fragment containing putative promoter sequences for *BRCA1* has been cloned and sequenced (Xu et al. 1995; GenBank accession no. U37574). Our data are virtually identical to this sequence: 13 mismatches and one indel were observed; nine of the differences occur at Ns in the U37574 sequence. Analysis of this region originally suggested that it represents a bidirectional promoter controlling expression of *BRCA1* and *1A1-3B,* a B-box containing protein with homology to *CA125* (Brown et al. 1994). However, subsequent characterization of a 300-kb region revealed a duplication of the 5′ ends of both *BRCA1* and *1A1-3B* (Brown et al. 1996). Although the full-length genes are separated by ~50 kb, 5′ pseudogene sequences of each exist within 550 bp of the transcription initiation site of the other but in the opposite orientation. The genomic *BRCA1* and U37574 sequences are identical to the *1A1–3B* pseudogene exon 1a and 1b sequences (Brown et al. 1996; Fig. 2B) but differ from the alternative first exon sequences in *1A1-3B* transcripts characterized from an ovarian tumor cell line (exon 1b; Campbell et al. 1994) and the myeloblast cell-line KG1 (exon 1a; N. Nomura, unpubl.; GenBank accession no. D30756).

Promoter and Enhancer Elements

BRCA1 transcription initiates from either of two sites separated by 277 bp that encode alternative first exons 1a and 1b. Both transcription initiation sites are utilized in most tissues, although there is preferentially higher expression of the exon 1a transcript in mammary gland and of the exon 1b transcript in placenta (Xu et al. 1995). TATA boxes are not evident in the sequences 5′ of either exon 1a or 1b (Brown et al. 1994; Xu et al. 1995); however, both have features similar to initiator elements (Inr) and reside in GC-rich regions characteristic of TATA-less promoters (Azizkhan et al. 1993). Furthermore, GC boxes (GGGCGG), which bind the Sp1 transcription factor and have been shown to be required for interaction of transcription factor IID (TFIID) with TATA-less promoters, are present 5′ of exon 1a [163 and 233 nucleotides upstream of the exon 1a transcription initiation site (nucleotide 3344), i.e., at positions −163 and −233], exon 1b [−7, −46, −130 from exon 1b initiation (nucleotide 3621)] and overlapping exon 1a (−200 and −248 from exon 1b).

Other potential regulatory elements in the sequences preceding exons 1a and 1b were identified using SIGNALSCAN (Prestridge 1991) and the TRANSFAC data base (Wingender 1994) to identify transcription factor binding sites. Among those identified in the region preceding exon 1a are cyclic AMP regulatory element binding protein (CREB) at position −176, CCAAT binding factor (−149, −340), serum response factor (SRF: −148), polyomavirus enhancer A binding protein 3 (PEA3: −183), and pituitary transcription factor-1 (Pit-1: −6); sites preceding exon 1b are CREB (−59) and activator protein 2 (AP2: −10). Sequence alignment was also used to identify potential progesterone (PRE) and estrogen (ERE) response elements. With the exception of two imperfect ERE elements in introns 2 (nucleotide 7238) and 7 (nucleotide 25,455), only ERE half sites were detected, which although sufficient for estrogen receptor binding do not confer hormone- inducible transcriptional activation. A single putative PRE element was identified at nucleotide 1222, ~2 kb upstream of the start of transcription of exon 1a and exon 1b, and two matches were embedded in the coding regions of exon 2 (nucleotide 4668) and exon 11 (nucleotide 37,063).

Other Genes

In addition to *BRCA1*, five genes were identified within the 117,143-bp sequenced region (Fig. 1; Table 4). Most of these were known to map close to *BRCA1* and expressed tags had been identified in the search for *BRCA1*. Two are complete genes (*Rho7*, *VAT1*), one is incomplete (a 3′ portion of *IFP 35*), and two are pseudogenes (*rpL21* and two 5′ exons of 1A1-3B were identified). These genes were identified by a combination of homology searching (blastx and blastn; Altschul et al. 1990); searches against the nonredundant (nr) data base at the National Center for Biotechnology Information (NCBI) and blastn searches against dbEST (the data base of expressed sequence tags, using the e-mail or blastn client servers at NCBI), and exon prediction using grail (version 1.3; Xu 1994)

and genefinder (C. Wilson and P. Green, unpubl.).

Rho7: The *Rho7* gene (nucleotides 96,693–103,386) was identified initially as a homolog to the *rho* family of GTP-binding proteins (Chardin 1991) by a similarity search with blastp against the SWISS–PROT data base using deduced amino acid sequences from grail predicted exons. Recent searches using blatsn against the nr data base identified the putative cDNA corresponding to this gene (X95456; P. Chardin, unpubl.); only two bases out of 684 were in conflict. The reported cDNA sequence contains the coding sequence only; neither 5′ nor 3′ UTRs are in GenBank. A portion of the potential 5′ UTR was identified by genefinder (Fig. 3). Other exon/intron boundaries were identical when genefinder and similarity analysis (alignment against X95456) were compared. Grail 1.3 identified these exons as well but did not predict any 5′ UTR sequence, and split the 3′ terminal coding exon into two smaller exons. This gene is also highly similar to a mouse EST homolog (R74747; D. Beier and K. Brady, unpubl.); >90% identity is observed between the two nucleotide sequences. The only significant difference is that the mouse EST lacks the first exon relative to the human germ-line sequence. None of the *rho* homologs identified any additional exons 5′ or 3′ to the coding sequence; both grail and genefinder predicted the end of the last exon to precede the terminal A of the TGA stop codon by four bases to give a GT dinucleotide acceptor sequence in the intron. Five EST sequences (R42098, N66093, R15355, H48939, with 97–99% identity over the entire lengths of the clones, and R74748, a mouse cDNA that shows 74% identity) map close to the 3′ end of the *Rho7* gene. R42098, N66093, and R15355 form one region of similarity from nucleotides 96,693–98,031, which contains two potential poly(A) signal sites at nucleotides 96,710 and 97,592. R42098 (343-bp) and R15355 (457-bp) are derived from end sequences of clone yf90b08 (Washington University–Merck EST Project, unpubl.), which likely spans the 1338-bp region from nucleotides 96,693 to 98,031. Another region of similarity, aligning with H48939 and R74748, would extend the exon containing the TGA stop codon an additional ~240–400 bases. This is plausible because neither a GT

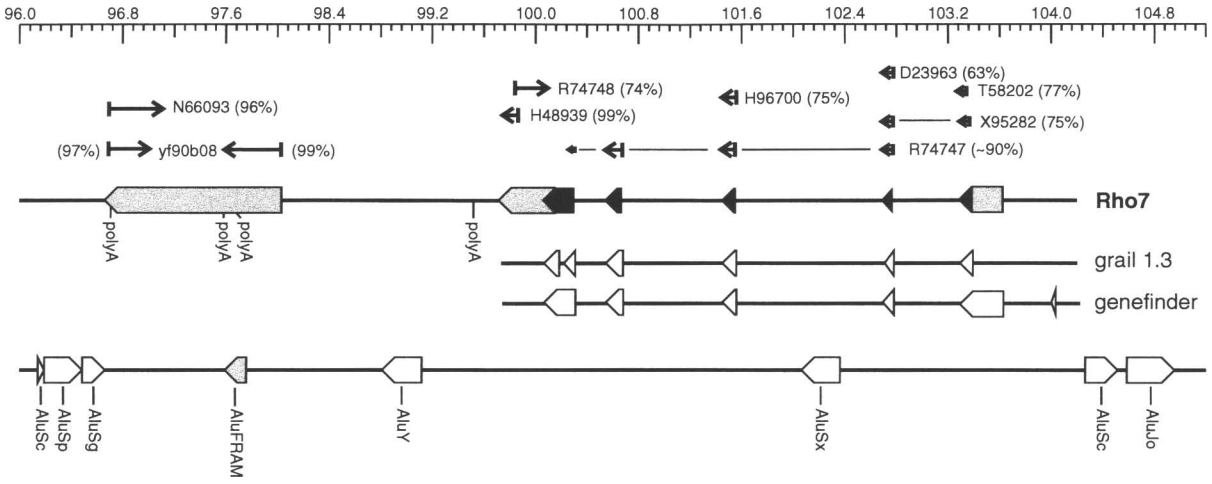

Figure 3 Map of the region from nucleotides 96,000 to 105,000 containing *Rho7*. The arrows at the *top* represent similar sequences identified by searches against the expressed sequence data base (dbEST). The identity of the pairwise alignment with the 117,143-bp contig is shown in parentheses. ESTs T58202, X95282, H96700, N66093, and yf90b08 are of human origin; R74747 and R74748 are of mouse origin; and D23963 is from rice. The next line shows the deduced structure of the *Rho7* gene. Dark boxes represent the coding region of the gene as determined by alignment with the *Rho7* cDNA sequence (X95456). Identity is >99%. The shaded regions correspond to exons predicted by alignments with human sequences from dbEST, or predicted by genefinder. *Below* the *Rho7* gene map are maps of exons predicted by grail 1.3 and genefinder. The last map shows the position of the *Alu* repeats in this region with the *Alu* half element that is predicted to be in the 3′ terminal exon of *Rho7* shaded. The CpG island at the beginning of this gene (103,020–104,040) overlaps with the putative first exon.

splice donor sequence nor a poly(A) signal sequence is present immediately after the TGA codon.

VAT1: Two cDNA clones (Friedman et al. 1995a; U18009, U25779) homologous to *VAT1*, an abundant membrane protein found in *Torpedo* cholinergic synaptic vesicles (Linial et al. 1989), were identified by a blastn search against the nr data base. These cDNAs identified previously in the search for *BRCA1* are highly similar (99–100% identities in both cases) to exon regions of the germ-line sequence from nucleotides 106,659 to 114,716. The putative poly(A) signal sequence starts at 114,696. The alignments were used to determine complete exon/intron structure for this gene (Fig. 4). Most of the exons were predicted by grail and genefinder as well; however, both programs failed to predict the 1538-bp 3′ terminal exon. In addition to the cDNA clones, 84 EST sequences that match the exon sequences with high similarity (97–100% identity over the length of the EST) were identified by blastn against dbEST. The majority of these sequences (64) map to the last exon, and 40 are confined to the last 500 bp of the gene. The high number of EST sequences found to map to the *VAT1* gene suggest that it is an abundant message. Two of the ESTs map to the first intron of this gene (Harshman et al. 1995).

IFP 35: At the terminal end of the *BRCA1* contig (nucleotides 115,000–117,134), a homology to a 282 amino acid interferon-induced leucine zipper protein *IFP 35* (Bange et al. 1994; SWISS-PROT accession no. P80217) was identified using grail and blastx. The complete cDNA sequence for this protein has not been deposited in GenBank. Conceptual translation of the putative exons for this gene resulted in an amino acid sequence that perfectly matched the reported *IFP 35* protein sequence from amino acid 41 to amino acid 263 (Fig. 5). The extreme 5′ exons for *IFP 35* are not included in the 117,143-bp contig, and the 19 carboxy-terminal amino acids did not align with the peptide deduced from genomic DNA sequence because of a frameshift in one of the data sets. The genomic *IFP 35* sequence (L78833) would require a 2-bp deletion at nucleotide 115,093 to generate a conceptual translation matching the published *IFP 35* peptide sequence. In the germ-line sequence, the stop codon for this gene is at nucleotide 115,023, ~10 bases further downstream than predicted by either grail or genefinder, and a protein five amino acids larger than the reported *IFP 35* would be encoded. Three EST sequences possibly extend the boundary of the 3′ terminal exon by ~180 bases. This would place the end of the *IFP 35* gene within 150 bases of the end of the *VAT1* gene. No potential poly(A) signal sequences could be detected within this region.

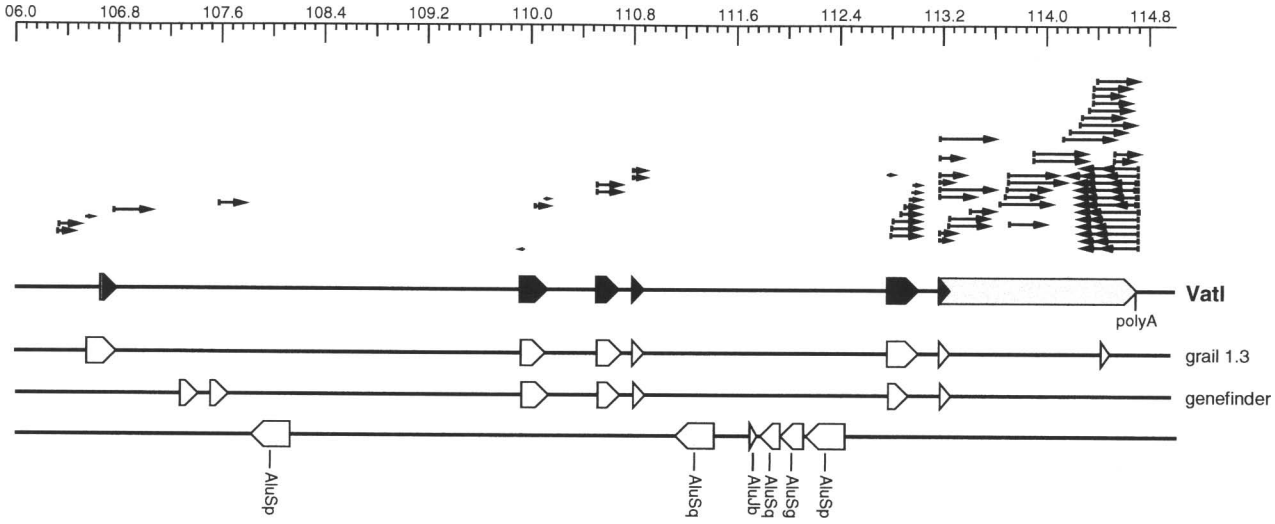

Figure 4 Map of the region from nucleotides 106,000 to 115,000 containing *VAT1*. Sequences from dbEST with high similarity to this region are shown above a map of the proposed *VAT1* gene. Sequences mapping to introns are unexpressed STSs isolated in the hunt for *BRCA1* and submitted to dbEST. The coding region for *VAT1* is indicated by dark boxes. 5′ and 3′ UTRs, determined from an alignment with U18009, are shown as shaded boxes. Exons predicted by grail and genefinder are shown as open boxes below the map of the gene.

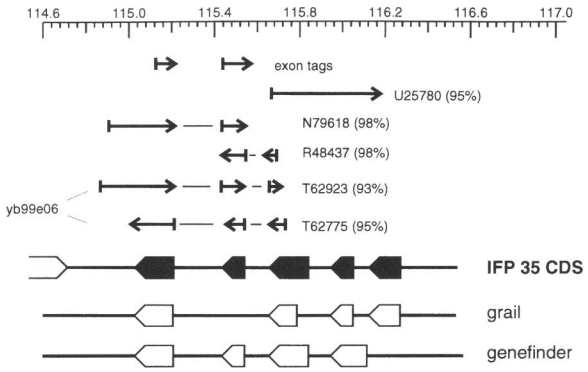

Figure 5 Map of the region from nucleotides 114,600 to 117,143 containing *IFP 35*. Coding exons of the interferon induced gene *IFP 35*, determined by alignments of the contig with the *IFP 35* protein sequence (P80217), are shown as filled boxes, and a portion of the 3′ terminal exon of *VAT1* is shown as a shaded box. Alignments of *IFP 35* yielded 100% identity over the first 227 amino acids. Above the line representing *IFP 35* are cDNA sequences identified from dbEST. T62923 and T62775 are opposite ends of clone yb99e06. U25780 is deposited in dbEST, but the sequence was determined in the hunt for the *BRCA1* gene. Numbers in parentheses denote the percent of nucleotides identical by alignment with the contig. In addition to the dbEST sequences, 19 exon tags (labeled arrows) were identified in the nr database at NCBI (U21518, U21520, U21522, U21523, U21525–U21527, U21531–U21533, U21536–U21539, U21541–U21545). Exons predicted by grail and genefinder are shown as open boxes. The 5′ end of this gene was not in the sequenced region and no poly(A) signal sequence could be detected except on the complementary strand at position 105,990, ~10 kb downstream from the *IFP 35* gene.

rpL21: A pseudogene of ribosomal protein L21 was identified in intron 13 of the *BRCA1* gene (Figs. 1 and 6). It was detected by similarity searching (11 similar sequences identified in the nr sequence data base and 111 similar sequences identified in dbEST with blastn as of July 1996) and grail 1.2; grail version 1.3 failed to identify any exons in this region. Two of the cDNA sequences (GenBank accession nos. U25789 and L38826) were identified in the search for *BRCA1* and predicted to map to chromosome 17q12-21 (Albertsen et al. 1994; Harshman et al. 1995). U25789 was predicted to map between markers D17S1321 and D17S1325, a ~600-kb region containing the *BRCA1* gene (Neuhausen et al. 1994). Alignment of the cDNA sequence and genomic sequence shows 93% identity over the entire 562 bases of the U25789 sequence and 95% to the first 497 bases (out of 763 total) of the L38826 sequence, suggesting that these cDNA sequences do not map to this site. Furthermore, conceptual translation of the U25789 and L38826 sequences produces full-length L21 peptides, whereas a conceptual translation of the L21 sequence identified herein results in a truncated product due to frame shifts caused by single base deletions. Thus it is likely that this germ-line sequence is a pseudogene, and the other cDNA clones map elsewhere in the genome.

Other Features

In addition to finding genes and repeat sequences, genomic sequencing allows the identification and precise placement of other landmarks on genetic maps. These include STS (sequence tagged site) markers and clones obtained from specific experiments on genomic DNA. Within the 117,143 bp of sequence data, five previously characterized STS markers were identified—D17S1323 (nucleotides 42,555–42,694), D17S1322 (nucleotides 69,171–69,293), D17S855 (nucleotides 75,856–76,010), UT6423 (L18209) at nucleotides 3138–3744 (97% identity), and CHLC.GCT17C04.P18510 (G10044) at nucleotides 63,905–64,148 (97% identity)—in addition to 63 novel SSRs. Several exon tags identified in the hunt for *BRCA1* were identified and placed on the contig map (Fig. 6).

Analysis of CpG and GpC content revealed that CpG dinucleotides are under-represented in the contig as a whole, with an average of 1.6 CpG pairs per 100 nucleotides (S.D. 1.97), compared with an average of 5.3 GpC pairs per 100 nucleotides (S.D. 2.89). The distributions of CpG and GpC pairs in the sequence are generally uniform and have about the same regions of high concentration. However, three regions in the sequence have unusually high CpG content (Fig. 1). The first spans nucleotides 1500–4150 and maps to the promoter region of the *BRCA1* gene (nucleotides 3343–3735). This region is also identified by two sequence homologies (Z57797 and Z57798, 99% identity) that derive from the ends of clone cpg197c5, which was isolated by a method to enrich for CpG containing DNA (Cross et al. 1994). The other two CpG-rich regions extend from nucleotides 102,680 to 104,650 and 105,240 to 108,100, overlapping the 5′ UTRs of *Rho7* and *VAT1*, respectively.

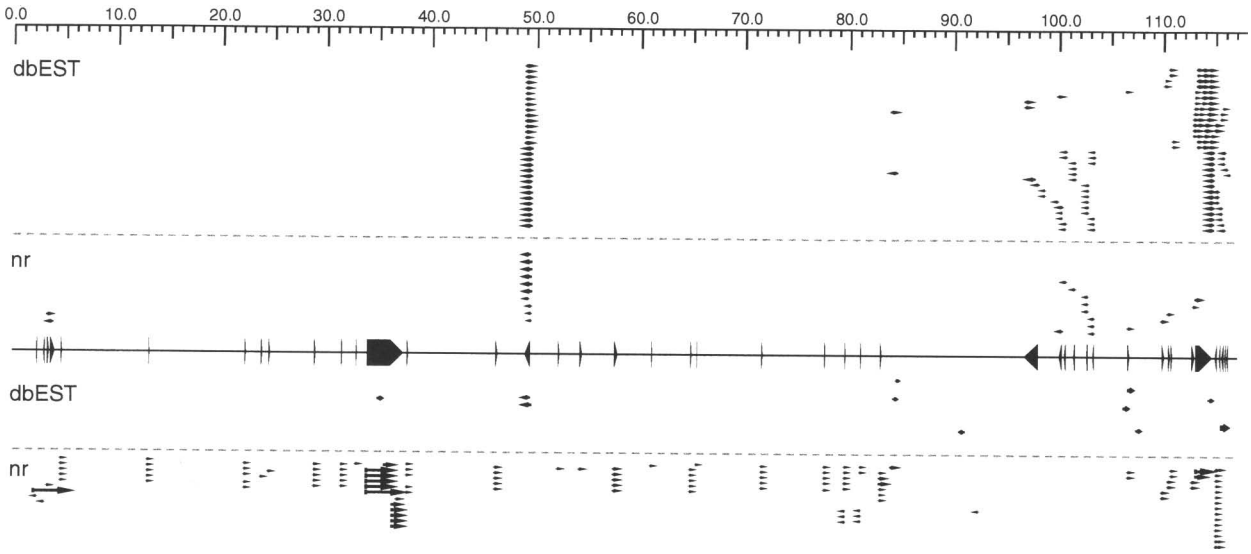

Figure 6 Summary of blastn searches against dbEST and GenBank (nr). The masked *BRCA1* contig was searched against the dbEST and nonredundant (nr) GenBank databases as described in Fig. 1. The resulting subject sequences were separated into two groups. Above the map of exons are ESTs from randomly selected cDNA clones. Below the map of exons are ESTs identified as the direct result of *BRCA1* research. No ESTs to the *BRCA1* coding sequence have been found from randomly selected cDNA clones.

DISCUSSION

Gene Identification and Description

A clear challenge in genomic sequencing is the ability to predict genes in DNA sequences from nucleotide composition. In the 117,143-bp *BRCA1* contig, three complete genes, two partial genes, and one pseudogene were identified by a combination of similarity searching and analysis with gene predicting computer algorithms: The *BRCA1* gene includes the coding sequence for a 1863 amino-acid protein and spans 81 kb of DNA with 24 exons ranging in length from 40 bp to 3425 bp. The introns range in length from 403 bp to 9193 bp; 10 introns are longer than 3 kb. In contrast, the other genes in the *BRCA1* contig are much smaller. *Rho7* includes a coding sequence that would yield a 227 amino-acid protein and spans 6.9 kb of DNA. The exons range in size from 87 bp to 1338 bp (putative 3′ UTR) and the introns range from 235 bp to 2023 bp. The *VAT1* coding sequence predicts a 301 amino-acid peptide in six exons spanning 8.1 kb. The exons range from 89 bp to 1538 bp and the introns from 104 bp to 3136 bp. The largest exons for the *Rho7* and *VAT1* genes appear to contain the 3′ UTRs.

Similarity analysis was the most successful method for accurately predicting coding regions and identifying exons found in UTRs, but relies on data from full-length cDNAs and fails when these sequences are not available or incomplete. Of the 227 ESTs retrieved from the dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_genes/) in July 1996 with high similarity to the *BRCA1* contig (Figs. 1 and 6), 111 mapped to the *rpL21* pseudogene and 84 mapped to *VAT1*, of which 64 were completely contained within the *VAT1* 3′ UTR. Among five ESTs that mapped to *BRCA1*, three were obtained in a highly intensive search for this gene: One derived from exon 11 and two from the 3′-UTR region. Neither of the two ESTs derived from randomly isolated cDNA sequences (human: H90415; mouse: W91622) identifies *BRCA1* coding sequences; both map to the 3′ UTR. Two gene-finding programs, grail and genefinder, were tested on the 117,143-bp contig. Both programs successfully identified almost all exons of genes in the contig other than *BRCA1* exons (Figs. 1–5). Genefinder predicted *BRCA1* exon 11; grail predicted ten *BRCA1* exons (Fig. 2). Both programs failed to identify 5′ or 3′ UTRs with high success, probably because they lack either a splice acceptor site (5′ UTR) or a splice donor site (3′ UTR) and do not have a nucleotide composition bias observed in coding regions.

The *rpL21* pseudogene and the 3′ UTR of

VAT1 contained the greatest number of matches in dbEST. Among the other genes within the sequenced *BRCA1* contig, the *Rho7* gene matched four ESTs which localize to the 3′ UTR and are derived from three independent clones. Because *Rho7* is a member of a large family of GTP-binding proteins homologous to *ras*, six other sequences in dbEST similar to this gene could be used to identify some of the coding regions. Of these six, three were from human (dBEST accession nos. H48939, T58202, and X95282), two were from mouse (dBEST accession nos. R74747 and R74748), and one (D23963) was isolated from rice (Fig. 3). Similarly, the *IFP 35* gene had four matches from three clones in dbEST when sequences determined specifically as a result of the hunt for the *BRCA1* gene were excluded. Unlike *Rho7*, however, the ESTs matching the *IFP 35* gene contained sequences covering nearly three exons at the 3′ end of the gene. This gene appears to have a short (200-bp) 3′ UTR.

Complete genomic sequences can be used to measure the level of gene sampling and coding information obtained from the sequence data generated from partial sequencing of random cDNA clones in expressed sequence tag (EST) projects. As of June 17, 1996, 51 of the 82 (62%) human disease genes that have been cloned positionally, including the *BRCA1* gene, had at least one match in dbEST. However, the only EST sequence in the data base matching coding sequence from the *BRCA1* gene was isolated in the hunt for *BRCA1* and not obtained as a result of random sequencing of cDNA libraries. Similar results were observed with the *BRCA2* gene sequence (data not shown). Four of the five *BRCA1* ESTs mapped to the 3′ UTR, as did the majority of ESTs from *Rho7* and *VAT1* (68/94, or 72%). Thus, analysis of the 117,143-bp contig leads to two important observations about EST data. First, primarily abundant genes are sampled by the random nature of EST sequencing. Normalizing the cDNA libraries improves this sampling, but low-abundance transcripts of genes with important functions will continue to be missed. Second, the EST sequences are biased toward the 3′ ends of genes. There is currently very little information on the structure of mammalian genes. However, if the genes in this locus are any indication, the 3′ UTRs of many genes may be 1–2 kb in length. This appears to be the predominant length of the clones used in EST sequencing projects, based on alignments of the sequences from both ends of individual clones. Thus, the bias for sampling 3′ sequences will identify largely 3′ UTRs, depending on the preparation of the clone library. If this is the case, only a small fraction of the ESTs will contain coding information, and the number of novel genes (i.e., lack of similarity in data bases) predicted by evaluation of EST data is likely to be an overestimate.

Other Features

CpG Islands

The *BRCA1* contig contains three CpG islands, one in front of each complete gene, *BRCA1*, *Rho7*, and *VAT1*. CpG islands consist of short stretches (1–2 kb) of unmethylated GC-rich DNA that do not show any suppression of CpG dinucleotides. CpG island DNA accounts for ~2% of the human genome and has been found in front of all studied housekeeping genes and many genes with a tissue restricted patterns of expression (Cross et al. 1994). In vertebrate genomes, CpG dinucleotides are suppressed to one-fourth of their expected frequency, believed to be caused by transamination of 5-methyl-cytosine. Like other genes with CpG islands the *BRCA1*, *Rho7*, and *VAT1* genes have their first exon contained within CpG islands. The *BRCA2* gene (GenBank accession no. U43746; Tavtigian et al. 1996) is also preceded by a CpG island. These data indicate that identifying potential genes within a chromosomal region by virtue of CpG islands would be a highly efficient strategy in physical cloning projects.

Simple Sequence Repeats

Additional sequence features of importance are SSRs or "microsatellites," which are used extensively as genetic markers because of their high level of polymorphism (Tautz 1989). Complete genomic sequencing gives one access to all such markers, many of which will be useful for high-resolution mapping of recombination events in linkage and physical mapping studies and for LOH analyses. Sixty-eight simple sequence repeats were identified in the *BRCA1* contig, of which 54 lie within the *BRCA1* gene. Among the intragenic SSRs, D17S855, D17S1322, and D17S1323 are previously characterized polymorphic markers. Three other sequences include at least 15 consecutive repeats of the dinucleotides.

High Density of Alu *Repeats in* BRCA1

The *BRCA1* contig has one of the highest densi-

ties of *Alu* repeats reported to date. Of 326 loci in GenBank analyzed for repeat sequences, relatively few (41 genes, or 12.6%) have *Alu* densities >20% and only three genes have *Alu* densities greater than *BRCA1*. A significant technical challenge posed by the high repeat content of genomic DNA from this region was cosmid instability, possibly exacerbated by the high copy vector (SuperCos; Wahl et al. 1987) used for preparation of the chromosome 17 specific library. Reduced deletion rates achieved under modified antibiotic selection enabled recovery of full-length clones in low yield. Sequence analysis of cosmid deletion breakpoints revealed that they occurred at *Alu*s, indicating that these repetitive elements can serve as templates for recombination in *E. coli*. High *Alu* density also led to sequence errors, probably as a result of "dropping" of bases within *Alu*-specific poly(A) sequences during DNA synthesis, analogous to slippage of eukaryotic DNA polymerases along monotonic base runs (Kunkel 1986; Greenblatt et al. 1996). The problem was more significant when M13 template clones were sequenced with AmpliTaq than with TaqFS, and was overcome by requiring a higher (i.e., >8-fold) redundancy of shotgun sequence reads. This redundant sequencing strategy also resolved discrepancies between the predicted and observed restriction maps of the cosmids caused by misassembly of the contig at *Alu* sequences.

In the human genome there appears to be a positive correlation between gene density and *Alu* density (for review, see Korenberg and Rykowski 1988; Holmquist 1992). In situ hybridization studies have shown that *Alu* sequences are localized predominantly to the gene-rich R (reverse) bands of metaphase chromosomes (Korenberg and Rykowski 1988), regions which are preferentially involved in homologous and nonhomologous chromosomal exchange processes (for review, see Morgan and Crossen 1977). A number of disease-associated genetic rearrangements and deletions involve *Alu* sequences: several Philadelphia chromosome BCR–ABL translocation breakpoints (Chen et al. 1989a,b); an inversion-deletion in β-globin (Glanzmann thrombasthemia; Li and Bray 1993); and intragenic deletions in lysyl hydroxylase (Ehlers-Danlos syndrome Type VI: Heikkinen et al. 1994; Pousi et al. 1994), low-density lipoprotein receptor (familial hypercholesteremia: Lehrman et al. 1987), apolipoprotein B (hypobetalipoproteinemia: Huang et al.1989), adenosine deaminase (ADA-SCID: Berkvens et al. 1990), and complement component C1 (hereditary angioedema: Stoppa-Lyonnet et al. 1990).

The mechanism by which these *Alu–Alu* rearrangements are mediated is not yet clear. In a survey of recombination sites involving *Alu* elements (Rudiger et al. 1995), a 26-bp core was identified that contains a pentanucleotide motif (CCAGC) that is part of prokaryotic *chi*, an 8-bp sequence known to stimulate *recBC* recombination in *E. coli*. Within the *BRCA1* gene there are 19 of these *Alu* core sequences, and 28 in the entire sequence contig. Although deletion breakpoints of the *BRCA1* cosmids were at *Alu* sequences, only one cosmid (*BRCA1*-8) appeared to contain the 26-bp core sequence at its deletion site, indicating that mechanisms other than chi-mediated recombination were involved in instability of the other cosmids. Sequence analysis of the Philadelphia chromosome translocation region involving chromosomes 22 (*ABL*) and 9 (*BCR*) identified breakpoints within *Alu* elements, near *Alu* repeats, and in inter-*Alu* sequences (T'oth and Jurka 1994; Chissoe et al. 1995). These results indicate that *Alu* elements may mediate chromosomal rearrangements indirectly, either by promoting mispairing and illegitimate recombination of nonhomologous *Alu* dense regions, or by the formation of inter- or intrachromosomal hairpin loop structures that could trigger chromosomal translocations and deletions (Deininger and Schmid 1976; Lehrman et al. 1986; Chen et al. 1989a,b). These observations in combination with the dearth of inactivating somatic point mutations of *BRCA1* in sporadic breast and ovarian carcinomas suggest that somatic inactivation of *BRCA1* may occur by interstitial deletion or rearrangement promoted by the high density of *Alu* elements.

Some *Alu* subfamilies include sequence motifs found in hormone response elements [HRE consensus sequence (A/G)G(G/T)T(C/G)(A/G); Vansant and Reynolds 1995], suggesting a direct role in regulation of gene expression. One such naturally occurring *Alu* element that precedes the keratin K18 gene and contains four HREs confers a 35-fold increase in retinoic acid-inducible transcription of a reporter gene in transfected cells relative to control constructs lacking one or more HREs (Vansant and Reynolds 1995). Within the *BRCA1* gene one strong match to the K18 *Alu*, including two HRE motifs, was found at position 5137 in intron 2. In addition, a recently identified subclass of *Alu* repeats containing the ERE

consensus sequence GGTCAnnnTGGTC$(n)_9$-TGACC can function as estrogen receptor-dependent transcriptional enhancers in yeast (Norris et al. 1995). The *BRCA1* contig shows two matches to this consensus sequence in *BRCA1* introns 2 and 7 (see also: Norris 1995). The possibility that *Alu* repeats containing EREs are functionally significant in regulating *BRCA1* expression is particularly appealing, given that *BRCA1* is expressed in hormonally responsive tissues (Gudas et al. 1995, 1996; Lane et al. 1995; Marquis et al. 1995; Vaughn et al. 1996) and given the epidemiologic association of estrogen and breast cancer risk (Kelsey 1993).

Analysis of the genomic sequence from a 117,143-bp region of human chromosome 17 encompassing the 81-kb *BRCA1* gene revealed the precise organization of *BRCA1* exons and introns; the locations of the *IFP 35*, *VAT1*, *Rho7* genes, and *1A1-3B* and *L21* pseudogenes; CpG islands preceding *BRCA1*, *Rho7*, and *VAT1*; and the positions of the intragenic microsatellite markers D17S1323, D17S1322, and D17S855. *BRCA1* contains an unusually high density of *Alu* elements (41.5%), suggesting that interstitial deletion or rearrangement promoted by these repetitive sequences may contribute to somatic inactivation of *BRCA1*. Complete understanding of the regulation of BRCA1 expression throughout development, in various tissues, and in response to external cues, as well as misregulation during tumorigenesis awaits experimental delineation of relevant promoter and enhancer elements and identification of mechanisms leading to inactivation.

METHODS

Identification and Mapping of Cosmid Clones

Overlapping cosmid clones containing fragments of the *BRCA1* gene were isolated from a chromosome specific 17 cosmid library constructed at the Los Alamos National Laboratory. Nylon filters arrayed with the DNA from individual clones were hybridized with ^{32}P-labeled *BRCA1* PCR products amplified from human genomic DNA using exon specific primers (Friedman et al. 1994b). Ten cosmids were identified as containing fragments of *BRCA1* and four of these were required to cover the gene completely.

Sequencing

Each cosmid was sequenced by a shotgun strategy (Deininger 1983; Wilson et al. 1994; Rowen et al. 1996). Cosmid DNA was isolated by anion exchange column purification (Qiagen, Inc.) following manufacturer's recommendations, and 15 µg was sheared by sonication (Deininger 1983). The resulting fragments were treated with mung bean nuclease to generate blunt ends and fragments of 1.6–3.0 kb in length were purified from agarose gels using a Qiaex gel extraction kit (Qiagen, Inc.). These fragments (~0.2–0.4 µg) were ligated in a 20-µl reaction volume (10 units T4 DNA ligase, U.S. Biochemical) to 0.1 µg of *Sma*I digested and calf intestinal alkaline phosphatase (Boehringer Mannheim Biochemicals)-treated M13mp18 DNA. The ligation mixture was used to transform *E. coli* strain XL-1 Blue MRF′ (Stratagene, Inc.). Single-stranded DNA was purified from phage grown from randomly selected clones by NaI extraction (Wilson 1993) and cycle sequenced with Prism DNA sequencing kits (Perkin-Elmer), using either MJ or Perkin-Elmer 9600 thermal cyclers. Sequencing reactions were analyzed by automated fluorescence detection on Perkin-Elmer 373 DNA sequencers.

Assembly and Analysis

Sequence Chromatogram files were analyzed with the phred base-calling program (P. Green and B. Ewing, unpubl.; http://www.bozeman.mbt.washington.edu/phrap.docs/phred.html) to generate files containing sequence data and quality assignments. Overlapping sequences, ~900 out of 1400 obtained per cosmid, were assembled into contiguous sequences by the phrap assembly program (P. Green, unpubl.; http://www.bozeman.mbt.washington.edu/phrap.docs/phrap.html), and the data were viewed in Consed (C. Abajian, D. Gordon, and P. Green, unpubl.).

Computer-aided DNA sequence analysis was carried out with a variety of tools. Repetitive DNA sequences were identified using RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) and a library of repetitive elements (A. Smit, unpubl.; http://bozeman.mbt.washington.edu/RM/RepeatMaster.html). Microsatellite repeats were identified with sputnik (C. Abajian, unpubl.; http://www.genome.washington.edu/sputnik.html). Grail (Xu et al. 1994) (Xgrail server) and genefinder (C. Wilson and P. Green, unpubl.) were used to identify putative exons. Similarity searches were performed with the blast programs (Altschul et al. 1990) using the servers at NCBI. Other similarity alignments were carried out with FASTA (Pearson 1990) and cross_match (P. Green, unpubl.). DNA composition was analyzed with DiNucleotides and PercentGC (T. Smith, unpubl.; http://weber.u.washington.edu/~soundbat) in addition to the GCG package (Genetics Computer Group 1994). Postscript figures of the data were produced using DrawMap (T. Smith, unpubl.).

ACKNOWLEDGMENTS

We thank Colin Wilson for analyzing sequences with an early version of genefinder. This work was supported by the National Institutes of Health Grant RO1-CA27632; Human Genome Distinguished Postdoctoral Fellowship (T.M.S.) sponsored by the U.S. Department of Energy, Office of Health and Environmental Research, and administered by the Oak Ridge Institute for Science and Education; and National Institutes of Health Postdoctoral Fellowship F32-CA66293 (C.I.S.). M.C.K. is an American Cancer Society Research Professor.

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

REFERENCES

Albertsen, H., S. Smith, S. Mazoyer, E. Fujimoto, J. Stevens, B. Williams, P. Rodriguez, C. Cropp, P. Slijepcevic, M. Carlson, M. Robertson, P. Bradley, E. Lawrence, T. Harrington, Z. Sheng, R. Hoopes, N. Sternberg, A. Brothman, R. Callahan, B. Ponder, and R. White. 1994. A physical map and candidate genes in the BRCA1 region on chromosome 17q12-21. *Nature Genet.* **7:** 472–479.

Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Anderson, L.A., L. Friedman, S. Osborne-Lawrence, E. Lynch, J. Weissenbach, A. Bowcock, and M.C. King. 1993. High density genetic map of the BRCA1 region of chromosome 17q12-q21. *Genomics* **17:** 618–623.

Azizkhan, J., D. Jensen, A. Pierce, and M. Wade. 1993. Transcription from TATA-less promoters: Dihydrofolate reductase as a model. *Crit. Rev. Eukaryotic Gene Exp.* **3:** 229–254.

Bange, F.-C., U. Vogel, T. Flohr, M. Kiekenbeck, B. Denecke, and E.C. Bottger. 1994. IFP 35 is an interferon-induced leucine zipper protein that undergoes interferon-regulated cellular redistribution. *J. Biol. Chem.* **269:** 1091–1098.

Batzer, M.A., P.L. Deininger, B.-U. Hellmann, J. Jurka, D. Labuda, C.M. Rubin, C.W. Schmid, E. Zietkiewicz, and E. Zuckerkandl. 1996. Standardized nomenclature for Alu repeats. *J. Mol. Evol.* **42:** 3–6.

Benson, D., M. Boguski, D. Lipman, and J. Ostell. 1996. GenBank. *Nucleic Acids Res.* **24:** 1–5.

Berkvens, T., H. van Ormondt, E. Gerritsen, P. Khan, and A. van der Eb. 1990. Identical 3250 bp deletion between two AluI repeats in the ADA genes of unrelated ADA-SCID patients. *Genomics* **7:** 486–490.

Brown, M., C.-F. Xu, H. Nicolai, B. Griffiths, J. Chambers, D. Black, and E. Solomon. 1996. The 5′ end of the BRCA1 gene lies within a duplicated region of human chromosome 17q21. *Oncogene* **12:** 2507–2513.

Brown, M.A., H. Nicolai, C.F. Xu, B.L. Griffiths, K.A. Jones, E. Solomon, L. Hosking, J. Trowsdale, D.M. Black, and R. McFarlane. 1994. Regulation of BRCA1. *Nature* **372:** 733.

Campbell, I., H. Nicolai, W. Foulkes, G. Senger, G. Stamp, G. Allan, C. Boyer, K. Jones, R.J. Bast, E. Solomon, et al. 1994. A novel gene encoding a B-box protein within the BRCA1 region at 17q21.1. *Hum. Mol. Genet.* **3:** 589–594.

Castilla, L.H., F.J. Couch, M.R. Erdos, K.F. Hoskins, K. Calzone, J. Garaber, J. Boyd, M.B. Lubin, M.L. Deshano, L.C. Brody, F.S. Collins, and B.L. Weber. 1994. Mutations in the BRCA1 gene in families with early-onset breast and ovarian cancer. *Nature Genet.* **8:** 387–391.

Chardin, P. 1991. Small GTP-binding proteins of the ras family: A conserved functional mechanism? *Cancer Cells* **3:** 117–126.

Chen, S., Z. Chen, L. D'Auriol, M. Le Coniat, D. Grausz, and R. Berger. 1989a. Phl+bcr_ acute leukemias: Implications of Alu sequences in a chromosomal translocation occurring in the new cluster region within the BCR gene. *Oncogene* **4:** 195–202.

Chen, S., Z. Chen, M.-P. Font, L. D'Auriol, C.-J. Larsen, and R. Berger. 1989b. Structural alterations of the BCR and ABL genes in Phl positive leukemias with rearrangements in the BCR first intron: Further evidence implicating Alu sequences in the chromosomal translocation. *Nucleic Acids Res.* **17:** 7631–7642.

Chissoe, S.L., A. Bodenteich, Y.F. Wang, Y.P. Wang, D. Burian, S.W. Clifton, J. Crabtree, A. Freeman, K. Iyer, L. Jian, et al. 1995. Sequence and analysis of the human ABL gene, the BCR gene, and regions involved in the Philadelphia chromosomal translocation. *Genomics* **27:** 67–82.

Cliby, W., S. Ritland, L. Hartmann, M. Dodson, K.C. Halling, G. Keeney, K.C. Podratz, and R.B. Jenkins. 1993. Human epithelial ovarian cancer allelotype. *Cancer Res.* **53:** 2393–2398.

Couch, F., B. Weber, and B.C.I. Core. 1996. Mutations and polymorphisms in the familial early-onset breast cancer (BRCA1) gene. *Human Mutat.* **8:** 8–18.

Cropp, C.S., M.H. Champeme, R. Lidereau, and R. Callahan. 1993. Identification of three regions on chromosome 17q in primary human breast carcinomas which are frequently deleted. *Cancer Res.* **53:** 5617–5619.

Cross, S.H., J.A. Charlton, X. Nan, and A.P. Bird. 1994. Purification of CpG islands using a methylated DNA binding column. *Nature Genet.* **6:** 236–244.

Deininger, P.L. 1983. Random subcloning of sonicated DNA: Application to shotgun DNA sequence analysis. *Anal. Biochem.* **129:** 216–223.

———. 1989. SINEs: Short interspersed repeated DNA in higher eukaryotes. In *Mobile DNA* (ed. M. Howe and D. Berg), p. 972. American Society for Microbiology, Washington, D.C.

Deininger, P. and C. Schmid. 1976. An electron microscope study of the DNA sequence organization of the human genome. *J. Mol. Biol.* **106:** 773–790.

Durocher, F., D. Shattuck-Eidens, M. McClure, F. Labrie, M.H. Skolnick, D.E. Goldgar, and J. Simard. 1996. Comparison of BRCA1 polymorphisms, rare sequence variants and/or missense mutations in unaffected and

breast/ovarian cancer populations. *Hum. Mol. Genet.* **5:** 835–842.

Easton, D.F., D.T. Bishop, D. Ford, G.P. Crockford, and T.B.C.L. Consortium. 1993. Genetic linkage analysis in familial breast and ovarian cancer: Results from 214 families. *Am. J. Hum. Genet.* **52:** 678–701.

FitzGerald, M., D. MacDonald, M. Krainer, I. Hoover, E. O'Neil, H. Unsal, B. Silva-Arrieto, D. Finkelstein, P. Beer-Romero, C. Englert, D. Soroi, B. Smith, J. Younger, J. Garber, R. Duda, K. Mayzel, K. Isselbacher, S. Friend, and D. Haber. 1996. Germ-line *BRCA1* mutations in Jewish and non-Jewish women with early-onset breast cancer. *N. Engl. J. Med.* **334:** 143–149.

Ford, D., D.F. Easton, D.T. Bishop, S.A. Narod, D.E. Goldgar, and B.C.L. Consortium. 1994. Risks of cancer in BRCA1-mutation carriers. *Lancet* **343:** 692–695.

Friedman, L.S., E.A. Ostermeyer, E.D. Lynch, C.I. Szabo, L.A. Anderson, P. Dowd, M.K. Lee, S.E. Rowell, J. Boyd, and M.-C. King. 1994a. The search for BRCA1. *Cancer Res.* **54:** 6374–6382.

Friedman, L.S., E.A. Ostermeyer, C.I. Szabo, P. Dowd, E.D. Lynch, S.E. Rowell, and M.-C. King. 1994b. Confirmation of BRCA1 by analysis of germline mutations linked to breast and ovarian cancer in ten families. *Nature Genet.* **8:** 399–404.

Friedman, L.S., E.A. Ostermeyer, E.D. Lynch, P. Welcsh, C.I. Szabo, J.E. Meza, L.A. Anderson, P. Dowd, M.K. Lee, S.E. Rowell, et al. 1995a. 22 genes from chromosome 17q21: Cloning, sequencing, and characterization of mutations in breast cancer families and tumors. *Genomics* **25:** 256–263.

Friedman, L.S., C.I. Szabo, E.A. Ostermeyer, P. Dowd, L. Butler, T. Park, M.K. Lee, E.L. Goode, S.E. Rowell, and M.-C. King. 1995b. Novel inherited mutations and variable expressivity of BRCA1 alleles, including the founder mutation 185delAG in Ashkenazi Jewish families. *Am. J. Hum. Genet.* **57:** 1284–1297.

Futreal, P.A., Q. Liu, D. Shattuck-Eidens, C. Cochran, K. Harshman, S. Tavtigian, L.M. Bennett, A. Haugen-Strano, J. Swensen, Y. Miki, K. Eddington, M. McClure, C. Frye, J. Weaver-Feldhaus, W. Ding, Z. Gholami, P. Soderkvist, L. Terry, S. Jhanwar, A. Berchuck, J.D. Iglehart, J. Marks, D.G. Ballinger, J.C. Barrett, M.H. Skolnick, A. Kamb, and R. Wiseman. 1994. BRCA1 mutations in primary breast and ovarian carinomas. *Science* **266:** 120–122.

Gayther, S., W. Warren, S. Mazoyer, P. Russell, P. Harrington, M. Chiano, S. Seal, R. Hamoudi, E. van Rensburg, A. Dunning, R. Love, G. Evans, D. Easton, D. Clayton, M. Stratton, and B. Ponder. 1995. Germline mutations of the BRCA1 gene in breast and ovarian cancer provide evidence for a genotype-phenotype correlation. *Nature Genet.* **11:** 428–433.

Gayther, S.A., P. Harrington, P. Russell, G. Kharkevich, R.F. Garkavtseva, B.A.J. Ponder, and U.F.O.C.S. Group. 1996. Rapid detection of regionally clustered germ-line BRCA1 mutations by multiplex heteroduplex analysis. *Am. J. Hum. Genet.* **58:** 451–456.

Genetics Computer Group, Inc. 1994. The Wisconsin Sequence Analysis Package, Version 8.0. Madison, WI.

Greenblatt, M., A. Grollman, and C. Harris. 1996. Deletions and insertions in the p53 tumor suppressor gene in human cancers: Confirmation of the DNA polymerase slippage/misalignment model. *Cancer Res.* **56:** 2130–2136.

Gudas, J., T. Li, H. Nguyen, D. Jensen, F. Rauscher, and K. Cowan. 1996. Cell cycle regulation of *BRCA1* messenger RNA in human breast epithelial cells. *Cell Growth Diff.* **7:** 717–723.

Gudas, J.M., H. Nguyen, T. Li, and K.H. Cowan. 1995. Hormone-dependent regulation of BRCA1 in human breast cancer cells. *Cancer Res.* **55:** 4561–4565.

Hall, J.M., M.K. Lee, B. Newman, J.E. Morrow, L.A. Anderson, B. Huey, and M.-C. King. 1990. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250:** 1684–1689.

Harshman, K., R. Bell, J. Rosenthal, H. Katcher, Y. Miki, J. Swenson, Z. Gholami, C. Frye, W. Ding, P. Dayananth, K. Eddington, F. Norris, P. Bristow, R. Phelps, T. Hattier, S. Stone, D. Shaffer, S. Bayer, C. Hussey, T. Tran, M. Lai, P.J. Rosteck, M. Skolnick, D. Shattuck-Eidens, and A. Kamb. 1995. Comparison of the positional cloning methods used to isolate the BRCA1 gene. *Hum. Mol. Genet.* **4:** 1259–1266.

Heikkinen, J., T. Hautala, K.I. Kivirikko, and R. Myllyla. 1994. Structure and expression of the human lysyl hydroxylase gene (PLOD): Introns 9 and 16 contain Alu sequences at the sites of recombination in Ehlers–Danlos syndrome type VI patients. *Genomics* **24:** 464–471.

Hogervorst, F.B.L., R.S. Cornelis, M. Bout, M. van Vliet, J.C. Oosterwijk, R. Olmer, B. Bakker, J.G.M. Klijn, H.F.A. Vasen, H. Meijers-Heijboer, F.H. Menko, C.J. Cornelisse, J.T. den Dunnen, P. Devilee, and G.-J.B. van Ommen. 1995. Rapid detection of BRCA1 mutations by the protein truncation test. *Nature Genet.* **10:** 208–212.

Holmquist, G.P. 1992. Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet.* **51:** 17–37.

Holt, J., M. Thompson, C. Szabo, C. Robinson-Benion, C. Arteage, M.-C. King, and R. Jensen. 1996. Growth retardation and tumor inhibition by BRCA1. *Nature Genet.* **12:** 298–302.

Hosking, L., J. Trowsdale, H. Nicolai, E. Solomon, W. Foulkes, G. Stamp, E. Signer, and A. Jeffreys. 1995. A somatic BRCA1 mutation in an ovarian tumor. *Nature Genet.* **9:** 343–344.

Huang, L.-S., M. Ripps, S. Korman, R. Deckelbaum, and J. Breslow. 1989. Hypobetalipoproteinemia due to an apolipoprotein B gene exon 21 deletion derived by Alu–Alu recombination. *J. Biol. Chem.* **264:** 11394–11400.

Hunkapiller, T., R.J. Kaiser, B.F. Koop, and L. Hood.

1991. Large-scale and automated DNA sequence determination. *Science* **254:** 59–67.

Johannsson, O., E.A. Ostermeyer, S. Håkansson, L.S. Friedman, U. Jonahsson, G. Sellberg, K. Brøndum-Nielsen, V. Sele, H. Olsson, M.-C. King, and A. Borg. 1996. Founding BRCA1 mutations in hereditary breast and ovarian cancer in Southern Sweden. *Am. J. Hum. Genet.* **58:** 441–450.

Jurka, J. 1995. Origin and evolution of *Alu* repetitive elements. In *Impact of short interspersed elements (SINEs) on the host genome* (ed. R.J. Maraia), pp. 25–41. Landes Company, Austin, TX.

Kelsey, J.L., ed. 1993. Breast cancer. *Epidemiol. Rev.* **15:** 1–263.

Korenberg, J. and M. Rykowski. 1988. Human genome organization: *Alu*, lines and the molecular structure of metaphase chromosome bands. *Cell* **53:** 391–400.

Kunkel, T. 1986. Frameshift mutagenesis by eucaryotic DNA polymerases in vitro. *J. Biol. Chem.* **261:** 13581–13587.

La Mantia, G., G. Pengue, D. Maglione, A. Pannuti, A. Pascucci, and L. Lania. 1989. Identification of new human repetitive sequences: Characterization of the corresponding cDNAs and their expression in embryonal carcinoma cells. *Nucleic Acids Res.* **17:** 5913–5922.

Lane, T., C. Deng, A. Elson, M. Lyu, C. Kozak, and P. Leder. 1995. Expression of Brca1 is associated with terminal differentiation of ectodermally and mesodermally derived tissues in mice. *Genes & Dev.* **9:** 2712–2722.

Langston, A., K. Malone, J. Thompson, J. Daling, and E. Ostrander. 1996. *BRCA1* mutations in a population-based sample of young women with breast cancer. *N. Engl. J. Med.* **334:** 137–142.

Lehrman, M., D. Russell, J. Goldstein, and M. Brown. 1986. Exon-Alu recombination deletes 5 kilobases from the low density lipoprotein receptor gene, producing a null phenotype in familial hypercholesteremia. *Proc. Natl. Acad. Sci.* **83:** 3679–3683.

———. 1987. Alu–Alu recombination deletes splice acceptor sites and produces secreted low density lipoprotein receptor in a subject with familial hypercholesteremia. *J. Biol. Chem.* **262:** 3354–3361.

Li, L. and P.F. Bray. 1993. Homologous recombination among three intragene Alu sequences causes an inversion-deletion resulting in the hereditary bleeding disorder Glanzmann thrombasthenia. *Am. J. Hum. Genet.* **53:** 140–149.

Linial, M., K. Miller, and R.H. Scheller. 1989. VAT-1: An abundant membrane protein from *Torpedo* cholinergic synaptic vesicles. *Neuron* **2:** 1265–1273.

Marquis, S., J. Rajan, A. Wynshaw-Boris, J. Xu, G.-Y. Yin, K. Abel, B. Weber, and L. Chodosh. 1995. The developmental pattern of *Brca1* expression implies a role in differentiation of the breast and other tissues. *Nature Genet.* **11:** 17–26.

Merajver, S.D., T.M. Pham, R.F. Caduff, M. Chen, E.L. Poy, K.A. Cooney, B.L. Weber, F.S. Collins, C. Johnston, and T.S. Frank. 1995. Somatic mutations in the BRCA1 gene in sporadic ovarian tumours. *Nature Genet.* **9:** 439–443.

Miki, Y., J. Swensen, D. Shattuck-Eidens, P.A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L.M. Bennett, W. Ding, R. Bell, J. Rosenthal, C. Hussey, T. Tran, M. McClure, C. Frye, T. Hattier, R. Phelps, A. Haugen-Strano, H. Katcher, K. Yakumo, Z. Gholami, D. Shaffer, S. Stone, S. Bayer, C. Wray, R. Bogden, P. Dayananth, J. Ward, P. Tonin, et al. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266:** 66–71.

Morgan, W. and P. Crossen. 1977. The frequency and distribution of sister chromatid exchanges in human chromosomes. *Hum. Genet.* **38:** 271–278.

Neuhausen, S.L. and C.J. Marshall. 1994. Loss of heterozygosity in familial tumors from three BRCA1-linked kindreds. *Cancer Res.* **54:** 6069–6072.

Neuhausen, S.L., J. Swensen, Y. Miki, Q. Liu, S. Tavtigian, D. Shattuck-Eidens, A. Kamb, M.R. Hobbs, J. Gingrich, H. Shizuya, U.-J. Kim, C. Cochran, P.A. Futreal, R.W. Wiseman, H.T. Lynch, P. Tonin, S. Narod, L. Cannon-Albright, M.H. Skolnick, and D.E. Goldgar. 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.

Newman, B., M.A. Austin, M. Lee and M.-C. King. 1988. Inheritance of human breast cancer: Evidence for autosomal dominant transmission in high-risk families. *Proc. Natl. Acad. Sci.* **85:** 33044–33048.

Norris, J., D. Fan, C. Aleman, J.R. Marks, P.A. Futreal, R.W. Wiseman, J.D. Iglehart, P.L. Deininger, and D.P. McDonnell. 1995. Identification of a new subclass of Alu DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *J. Biol. Chem.* **270:** 22777–227782.

Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183:** 63–98.

Pousi, B., T. Hautala, J. Heikkinen, L. Pajunen, K.I. Kivirikko, and R. Myllyla. 1994. Alu–Alu recombination results in a duplication of seven exons in the lysyl hydroxylase gene in a patient with the type VI variant of Ehlers–Danlos syndrome. *Am. J. Hum. Genet.* **55:** 899–906.

Prestridge, D. 1991. SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements. *Comput. Appl. Biosci.* **7:** 203–206.

Rowen, L., B. Koop, and L. Hood. 1996. The complete 685-kilobase DNA sequence of the human β T cell receptor locus. *Science* **272:** 1755–1762.

Rudiger, N.S., N. Gregersen, and B.-M.C. Kielland. 1995.

One short well conserved region of Alu-sequences is involved in human gene rearrangements and has homology with prokaryotic chi. *Nucleic Acids Res.* **23:** 256–260.

Russell, S.E., G.I. Hickey, W.S. Lowry, P. White, and R.J. Atkinson. 1990. Allele loss from chromosome 17 in ovarian cancer. *Oncogene* **5:** 1581–1583.

Saito, H., J. Inazawa, S. Saito, F. Kasumi, S. Koi, S. Sagae, R. Kudo, J. Saito, K. Noda, and Y. Nakamura. 1993. Detailed deletion mapping of chromosome 17q in ovarian and breast cancers: 2-cM region on 17q21.3 often and commonly deleted in tumors. *Cancer Res.* **53:** 3382–3385.

Serova, O., M. Montagna, D. Torchard, S.A. Narod, P. Tonin, B. Sulla, H.T. Lynch, J. Feunteun, and G.M. Lenoir. 1996. A high incidence of BRCA1 mutations in 20 breast-ovarian cancer families. *Am. J. Hum. Genet.* **58:** 42–51.

Shattuck-Eidens, D., M. McClure, J. Simard, F. Labrie, S. Narod, F. Couch, B. Weber, L. Castilla, L. Brody, L. Friedman, E. Ostermeyer, C. Szabo, M.-C. King, S. Jhanwar, K. Offit, L. Norton, T. Gilewski, M. Lubin, M. Osborne, D. Black, M. Boyd, M. Steel, S. Ingles, R. Haile, A. Lindblom, A. Borg, D.T. Bishop, E. Solomon, P. Radice, G. Spatti, et al. 1995. A collaborative survey of 80 mutations in the BRCA1 breast and ovarian cancer susceptibility gene: Implications for presymptomatic testing and screening. *J. Am. Med. Assoc.* **273:** 535–541.

Simard, J., P. Tonin, F. Durocher, K. Morgan, J. Rommens, S. Gingras, C. Samson, J.-F. Leblanc, C. Belanger, F. Dion, Q. Liu, M. Skolnick, D. Goldar, D. Shattuck-Eidens, F. Labrie, and S.A. Narod. 1994. Common origins of BRCA1 mutations in Canadian breast and ovarian cancer families. *Nature Genet.* **8:** 392–398.

Smith, S.A., D.F. Easton, D.G.R. Evands, and B.A.J. Ponder. 1992. Allele losses in the region 17q12-q21 in familial breast and ovarian cancer non-randomly involve the wild-type chromosome. *Nature Genet.* **2:** 128–131.

Stoppa-Lyonnet, D., P. Carter, T. Meo, and M. Tosi. 1990. Clusters of intragenic Alu repeats predispose the human C1 inhibitor locus to deleterious rearrangements. *Proc. Natl. Acad. Sci.* **87:** 1551–1555.

Struewing, J.P., L.C. Brody, M.R. Erdos, R.G. Kase, T.R. Giambarresi, S.A. Smith, F.S. Collins, and M.A. Tucker. 1995. Detection of eight BRCA1 mutations in ten breast/ovarian cancer families, including one family with male breast cancer. *Am. J. Hum. Genet.* **57:** 1–7.

Szabo, C.I. and M.-C. King. 1995. Inherited breast and ovarian cancer. *Hum. Mol. Genet.* **4:** 1811–1817.

Takahashi, H., K. Behbakht, P.E. McGovern, H.-C. Chiu, F.J. Couch, B.L. Weber, L.S. Friedman, M.-C. King, M. Furusato, V.A. LiVolsi, A.W. Menzin, P.C. Liu, I. Benjamin, M.A. Morgan, S.A. King, B.A. Rebane, A. Cardonick, J.J. Mikuta, S.C. Rubin, and J. Boyd. 1995. Mutation analysis of the BRCA1 gene in ovarian cancers. *Cancer Res.* **55:** 2998–3002.

Tautz, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res.* **17:** 6463–6471.

Tavtigian, S.V., J. Simard, J. Rommens, F. Couch, E.-D. Shattuck, S. Neuhausen, S. Merajver, S. Thorlacius, K. Offit, L.-D. Stoppa, C. Belanger, R. Bell, S. Berry, R. Bogden, Q. Chen, T. Davis, M. Dumont, C. Frye, T. Hattier, S. Jammulapati, T. Janecki, P. Jiang, R. Kehrer, J.F. Leblanc, D.E. Goldgar, et al. 1996. The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. *Nature Genet.* **12:** 333–337.

Thompson, M.E., R.A. Jensen, P.S. Obermiller, D.S. Page, and J.T. Holt. 1995. Decreased expression of BRCA1 accelerates growth and is often present during sporadic breast cancer progression. *Nature Genet.* **9:** 444–450.

T'oth, G. and J. Jurka. 1994. Repetitive DNA in and around translocation breakpoints of the Philadelphia chromosome. *Gene* **140:** 285–288.

Vansant, G. and W.F. Reynolds. 1995. The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. *Proc. Natl. Acad. Sci.* **92:** 8229–8233.

Vaughn, J., P. Davis, M. Jarboe, F. Huper, A. Evans, R. Wiseman, A. Berchuck, J. Iglehart, P. Futreal, and J. Marks. 1996. *BRCA1* expression is induced before DNA synthesis in both normal and tumor-derived breast cells. *Cell Growth Diff.* **7:** 711–715.

Wahl, G., K. Lewis, J. Ruiz, B. Rothenberg, J. Zhao, and G. Evans. 1987. Cosmid vectors for rapid genomic walking, restriction mapping, and gene transfer. *Proc. Natl. Acad. Sci.* **84:** 2160–2164.

Wilson, R., R. Ainscough, K. Anderson, C. Baynes, M. Berks, J. Bonfield, J. Burton, M. Connell, T. Copsey, J. Cooper, et al. 1994. 2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans. *Nature* **368:** 32–38.

Wilson, R.K. 1993. High-throughput purification of M13 templates for DNA sequencing. *BioTechniques* **15:** 414–416.

Wingender, E. 1994. Recognition of regulatory regions in genomic sequences. *J. Biotechnol.* **35:** 273–280.

Xu, C.F., M.A. Brown, J.A. Chambers, B. Griffiths, H. Nicolai, and E. Solomon. 1995. Distinct transcriptional start sites generate two forms of BRCA1 mRNA. *Hum. Mol. Genet.* **4:** 2259–2264.

Xu, Y., R. Mural, M. Shah, and E. Uberbacher. 1994. Recognizing exons in genomic sequence using GRAIL II. *Genet. Eng. N.Y.* **16:** 241–253.

Yang-Feng, T.L., H. Han, K.C. Chen, S.B. Li, E.B. Claus, M.L. Carcangiu, S.K. Chambers, J.T. Chambers, and P.E. Schwartz. 1993. Allelic loss in ovarian cancer. *Int. J. Cancer* **54:** 546–551.

Received August 29, 1996; accepted in revised form October 2, 1996.