Human Molecular Genetics, 2004, Vol. 13, No. 22 **2737–2751**
doi:10.1093/hmg/ddh301
Advance Access published on September 22, 2004

Evolution of the tumor suppressor *BRCA1* locus in primates: implications for cancer predisposition

Adam Pavlicek[1,†], Vladimir N. Noskov[2,†], Natalay Kouprina[2], J. Carl Barrett[2], Jerzy Jurka[1] and Vladimir Larionov[2,*]

[1]Genetic Information Research Institute, Mountain View, CA, USA and [2]Laboratory of Biosystems and Cancer, National Cancer Institute, Bethesda, MD, USA

Received June 7, 2004; Revised September 1, 2004; Accepted September 14, 2004 DDBJ/EMBL/GenBank accession nos[‡]

Germ-line mutations in the *BRCA1* gene predispose affected individuals to breast and ovarian cancer syndromes. In an attempt to systematically analyze a broader spectrum of genetic changes ranging from frequent exon deletions and duplications to amino acid replacements and protein truncations, we isolated and characterized full size *BRCA1* homologues from a representative group of non-human primates. Our analysis represents the first comprehensive sequence comparison of primate *BRCA1* loci and corresponding proteins. The comparison revealed an unusually high proportion of indels in non-coding DNA. The major force driving evolutionary changes in non-coding *BRCA1* sequences was *Alu*-mediated rearrangements, including *Alu* transpositions and *Alu*-associated deletions, indicating that structural instability of this locus may be intrinsic in anthropoids. Analysis of the non-synonymous/synonymous ratio in coding portions of the gene revealed the presence of both conserved and rapidly evolving regions in the BRCA1 protein. Previously, a rapidly evolving region with evidence of positive evolutionary selection in human and chimpanzee had been identified only in exon 11. Here, we show that most of the internal *BRCA1* sequence is variable between primates and evolved under positive selection. In contrast, the terminal regions of *BRCA1*, which encode the RING finger and BRCT domains, experienced negative selection, which left them almost identical between the compared primates. Distribution of the reported missense mutations, but not frameshift and nonsense mutations, is positively correlated with BRCA1 protein conservation. Finally, on the basis of protein sequence conservation, we identified missense changes that are likely to compromise BRCA1 function.

INTRODUCTION

The *BRCA1* gene (MIM 113705) on chromosome 17q21.31 was identified on the basis of its linkage to early onset breast and breast–ovarian syndromes in women (1,2). The lifetime risk of breast and ovarian cancer among female *BRCA1* mutation carriers is 82% and 54%, respectively (3). The *BRCA1* gene contains 24 exons (22 coding and 2 non-coding) and covers a span of ~90 kb (4). Its coding region comprises ~5.6 kb and encodes a protein of 1863 amino acids. Exon 11, with 3427 bp, accounts for 61% of the CDS. The remaining exons are small, ranging from 37 to 311 bp.

BRCA1 has been implicated in a diverse array of biological processes, including the cellular response to DNA damaging agents, transcriptional regulation, ubiquitination and chromosome remodeling (reviewed in 5). Despite extensive studies, the function of the BRCA1 protein remains unclear. At present, over 30 proteins have been identified that bind to it directly, indirectly or as part of a larger multiprotein complex. BRCA1 contains three distinct protein-interacting regions: the RING finger domain, the RAD51 interaction domain and the BRCT domain. The N-terminal RING finger domain encompassing exons 2–6 has been implicated in interactions with at least five different proteins, including formation

*To whom correspondence should be addressed. Tel: +1 3014967941; Fax: +1 3014802772; Email: larionov@mail.nih.gov
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
‡AY365046 and AY589040–AY589042.

of stable heterodimers with BARD1 (reviewed in 6). The RING domain functions *in vitro* as an E3 ubiquitin ligase where it catalyzes the synthesis of monoubiquitin- and polyubiquitin-targeted proteins. The activity is greatly increased when BRCA1 is in a complex with its N-terminal binding partner BARD1. *In vivo*, BRCA1 and BARD1 co-localize in a cell-cycle-dependent manner and in response to DNA damage, suggesting a role for BRCA1 ubiquitin conjugation in DNA repair. The central RAD51 interaction domain is located at exon 11. It is involved in DNA repair and contains multiple protein-binding sites, including those for RAD51, the RAD50 complex and MSH2 (5). The C-terminal region was identified *in vitro* as a transcriptional co-regulator with some specificity for p53 and STAT1 co-activation. The region contains two ~90 amino acid sequence repeats called BRCT (BRCA1 C-terminal) (7) that have a weak similarity to other proteins involved in DNA repair, such as the yeast protein RAD9 and the mammalian protein XRCC1. This domain is particularly rich in protein–protein interaction sites (reviewed in 6), including binding domains for DNA helicase BACH1 (BRCA1-associated C-terminal helicase), which is involved in the repair of double-strand DNA breaks (8).

The Breast Cancer Information Core (BIC; http://research.nhgri.nih.gov/projects/bic/) is a database of more than 8500 mutations—including polymorphisms and rare variants—scattered along *BRCA1*, but only some of them are known to affect BRCA1 function. The vast majority of the disease-associated mutations result in truncated reading frames. The mutations include large genomic deletions and duplications involving one or more *BRCA1* exons (9,10) caused mostly by recombination between *Alu* repeats, which are particularly numerous in *BRCA1* (4). Nearly one-third of reported BRCA1 changes are missense mutations, and the functional consequence of most of them is uncertain. The reported disease-associated mutations are concentrated at the terminal RING and BRCT domains. Although those domains encompass only 13% of the entire protein, they house more than 90% of the missense changes known to be deleterious (11–13). It should be noted, however, that the sequence variants in the BIC database are based on voluntary submissions and do not represent an unbiased set of BRCA1 mutations.

Predictions regarding missense changes can be strengthened by comparative evolutionary analysis. Such analysis may be particularly helpful in the identification of low-penetrance missense changes in functionally important regions. Phylogenetic approaches can also determine whether certain residues have evolved more rapidly than predicted by a neutral theory, reflecting the action of positive (diversifying) selection. So far, complete CDS of *BRCA1* are available for only a few vertebrate species. A recent analysis of partial *BRCA1* exon 11 sequences in various mammalian species allowed the prediction of several missense mutations that would be more likely to affect BRCA1 function (14,15). Comparison of exon 11 sequences in non-human primates also revealed that the RAD51 interaction domain experienced strong positive selection during human evolution (16).

Here, we describe the isolation of genomic clones containing the entire *BRCA1* gene from chimpanzee, gorilla, orangutan and rhesus macaque. Comparison of the homologues allowed us to follow evolutionary changes in coding and non-coding regions of the *BRCA1* gene in primates and to extend the number of predicted amino acid changes that would affect gene function.

RESULTS

Genomic organization of the *BRCA1* genes in primates

We isolated *BRCA1* gene homologues from chimpanzee, gorilla, orangutan and rhesus macaque by transformation-associated recombination (TAR) cloning (Supplementary Material, Fig. S1) (17). The targeting hooks were developed from a promoter sequence and the 3′-untranslated region of human *BRCA1*. The vector allowed us to isolate the entire *BRCA1* gene as ~95 kb genomic fragments. The yield of *BRCA1*-positive clones from the four primate DNAs was approximately the same as that from human DNA (~1%). We isolated at least three independent genomic clones for each species. We converted one randomly chosen TAR clone from each species into a BAC and sequenced it with high accuracy using BAC DNA as a template.

The overall *BRCA1* structure was conserved in all five primates (Fig. 1). A conserved promoter region containing CpG islands was followed by 24 exons with conserved exon–intron boundaries. Multiple alignment of the genes revealed an unusually high proportion of insertions and deletions (indels) (Table 1; Fig. 1B and G). Pairwise identity in the aligned segments (with indels excluded) ranged from 93 to 99% for human–chimpanzee. The identity dropped to 72–73% (between macaque and hominoids) when indels were included (Table 1). Many of the rearrangements were linked to the activity of *Alu* repeats (see later).

The promoter region and the 5′ end of the human *BRCA1* gene is duplicated ~40 kb upstream, and homologous recombination in this segment occasionally causes deletion of the promoter region (18,19). We analyzed polymorphisms in human genomic clones containing the duplicated segments that were available in sequence databases for evidence of gene conversion (i.e. correspondence between the 5′ *BRCA1* pseudogene and the functional copy) between the segments, but did not detect any (data not shown). Nor did analysis of BIC mutations detect any hallmark of gene conversion (data not shown).

Alu repeats shape *BRCA1* genes in primates

Most of the detected long indels appear to be associated with *Alu* sequences. *Alu* elements are ~280 bp long, and in order to detect possible *Alu*-mediated rearrangements we concentrated on 45 indels ≥250 bp. The majority of the long rearrangements took place in the lineage leading to hominoid primates and in the macaque branch (Fig. 1B; Supplementary Material, Table S1). The ancestral orangutan–gorilla–chimpanzee–human lineage (25–14 million years ago—MYA; 20) accumulated nine *Alu* insertions (six from the *AluY* and three from the *AluS* subfamilies), resulting in a gain of 2755 bp (*Alu* insertions and duplications of the target sites). The same lineage exhibited two deletions not found in the rhesus macaque: one deletion (655 bp) was caused by homologous

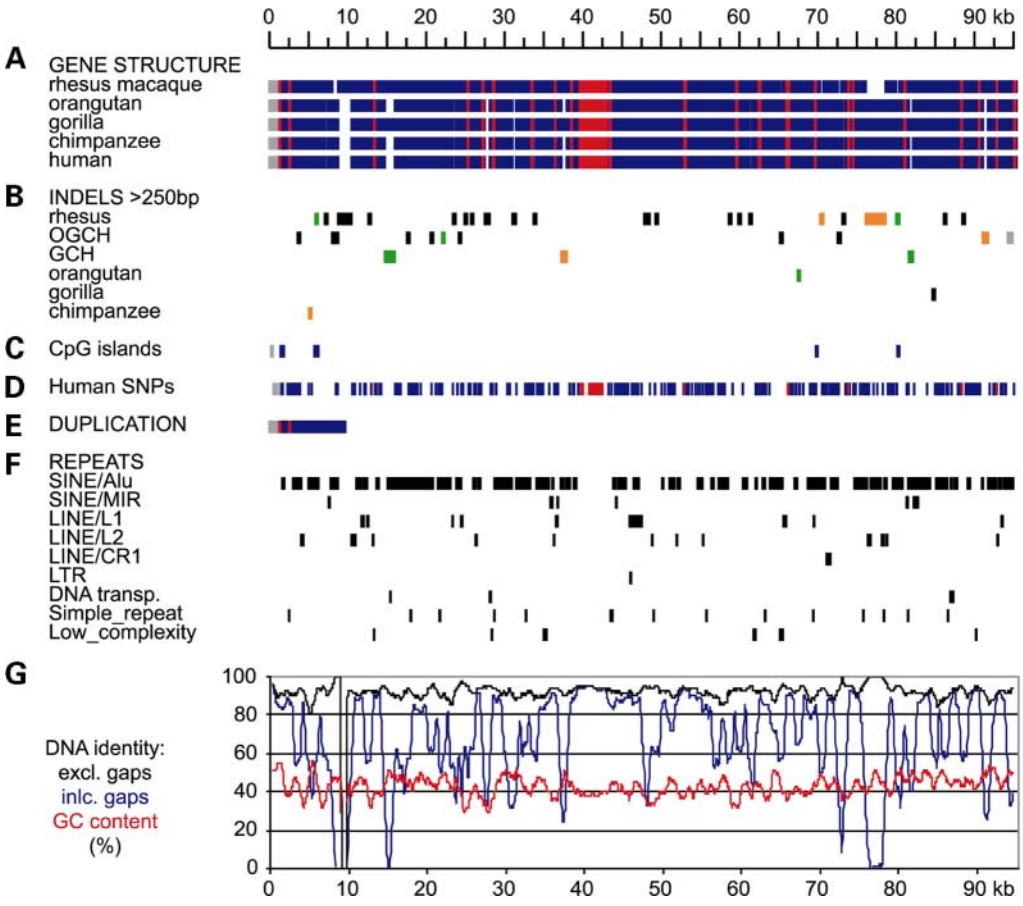

Figure 1. Structure and evolution of the *BRCA1* gene in primates. The scale of all plots corresponds to the consensus sequence on the basis of multiple alignment of five genomic *BRCA1* clones. (**A**) Schematic representation of the alignment. Promoter regions, exons and introns are gray, red and blue, respectively. Plain white segments correspond to gaps. (**B**) Positions of long (≥250 bp) insertions/deletions. All insertions are *Alu* retropositions and are black. Deletions by *Alu–Alu* homologous recombination are orange and deletions by non-homologous recombination are green. The gray segment in the last intron corresponds to a variable *Alu* array (see text). Individual rows represent different lineages. 'OGCH' corresponds to the orangutan–gorilla–chimpanzee–human ancestral lineage and 'GCH' to the gorilla–chimpanzee–human lineage. (**C**) Positions of predicted CpG islands. (**D**) Positions of human SNPs. Intron SNPs are blue and exon SNPs are red. (**E**) Location of a ∼14 kb-long (including 4.2 kb upstream) segmental duplication. (**F**) Distribution of repetitive elements. (**G**) DNA identity and GC content. DNA identity was calculated with (blue) and without (black) indel positions. We used a 1 kb sliding window with 100 bp overlaps for all plots. The GC profile corresponds to the consensus sequence; individual sequences have nearly identical profiles.

Table 1. Pairwise identity (%) of aligned primate *BRCA1* genes

	Macaque	Orangutan	Gorilla	Chimpanzee	Human
Macaque		93.4	93.6	93.72	93.68
Orangutan	73.13		97.07	97.13	97.16
Gorilla	71.81	90.44		98.74	98.78
Chimpanzee	72.06	90.44	96.73		98.96
Human	72.31	91.06	96.83	97.25	

We calculated pairwise identities for five complete *BRCA1* genes and included all promoter regions, introns and exons. The values above the diagonal show DNA identities calculated without indels, those under the diagonal show DNA identities calculated with indels (gaps).

Alu–Alu recombination, the other (317 bp) probably by non-homologous recombination. The lineage leading to rhesus macaque was particularly rich in indel variety. We detected two deletions caused by homologous *Alu–Alu* recombination and two caused by non-homologous recombination, leading to a total loss of 3647 bp in the primate consensus sequence. There were 23 macaque-specific *Alu* insertions (19 from *AluY*, three from *AluS* subfamilies and one short Alu fragment) that together with target-site duplication contributed 7547 bp. Indel variation, and especially *Alu* retrotranspositions, ceased in recent hominoid lineages (Fig. 1B). After separation of the orangutan, there were three deletions in the lineage leading to human, chimpanzee and gorilla (14–7 MYA; 20); one (671 bp) was caused by homologous *Alu–Alu* recombination and two (519 and 1279 bp) were probably caused by non-homologous recombination. There was one 5353 bp deletion in chimpanzee caused by *Alu–Alu* recombination and one gorilla-specific *AluYc1* insertion. There was also an orangutan-specific 280 bp deletion in the non-repetitive DNA. In addition, the last intron contained a cluster of *Alu* sequences that was unstable, because no sequence pairs share the same indel pattern.

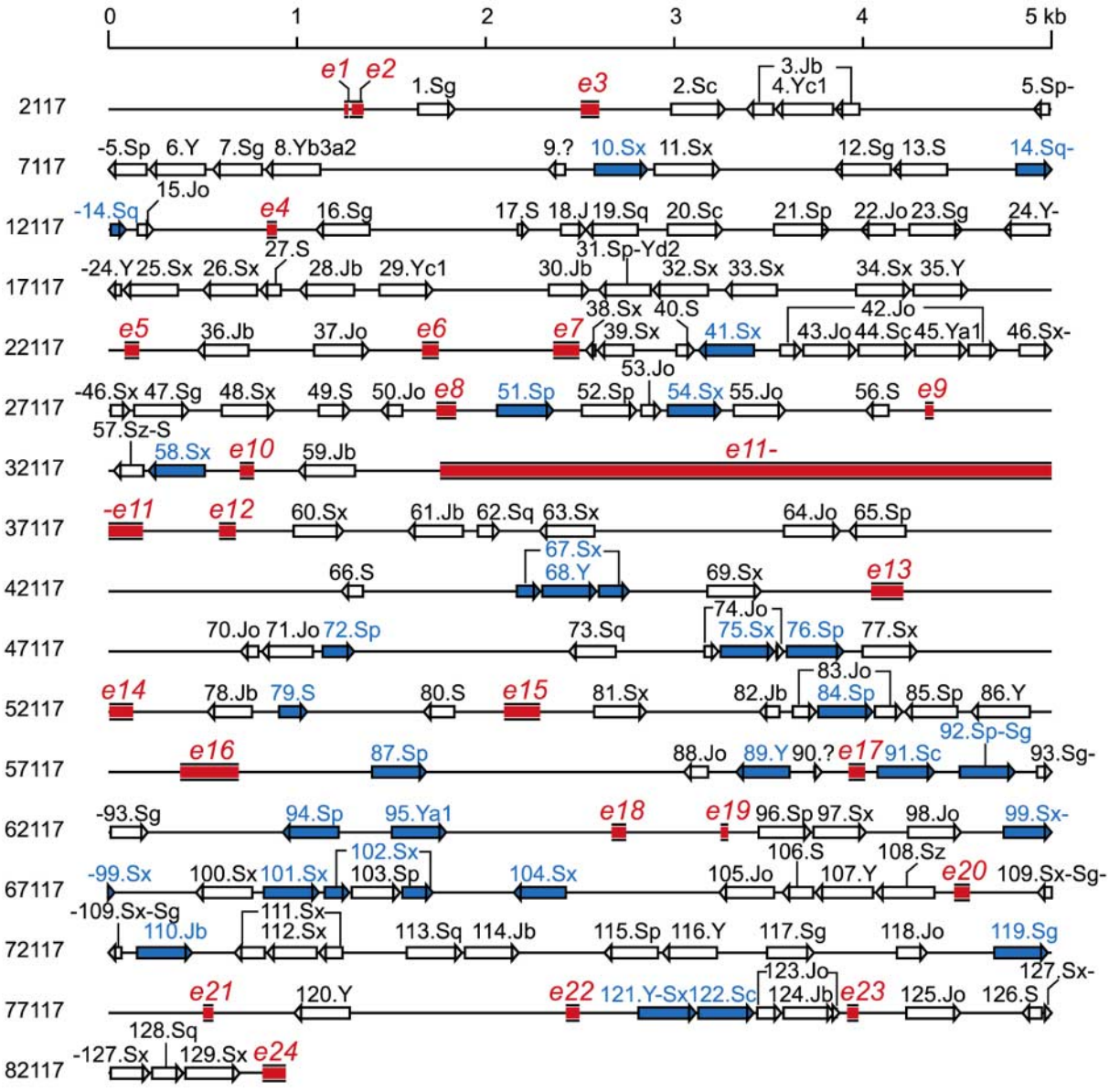

Figure 2. *Alu* elements and human *BRCA1* rearrangements. The figure shows *Alu* repeats in human *BRCA1*. Positions correspond to a 117 kb *BRCA1* genomic sequence (GenBank accession number L78833). Exons are depicted as red rectangles and *Alu*s as arrows. *Alu*s known to be involved with human exon deletions/duplications are blue (Supplementary Material, Table S2). *Alu5* and *Alu6* fused by homologous recombination in chimpanzee, *Alu92* and *Alu93* in macaque. *Alu7* and *Alu8* as well as *Alu99* and *Alu104* either were fused in rhesus macaque (probably by non-homologous recombination) or the original breakpoints were lost. *AluSp* and *AluYd2* fused to form *Alu31* in the common ancestor of hominoid primates after separation from macaque. *AluSz* and *AluS* elements fused to *Alu57* by homologous recombination in the common ancestor of African hominoid after separation of orangutan. *AluSx–AluSg* fused by homologous recombination to *Alu*109, which is found in African hominoids. *Alu121* is a product of *AluY,* and *AluSx* formed by homologous recombination in the ancestor of hominoids. *Alu128* and *Alu129* contain several independent indels in different species. Seven elements represent new insertions in the hominoid lineage (*Alu4*, *Alu10*, *Alu11*, *Alu21*, *Alu29*, *Alu35* and *Alu86*).

To summarize, all 33 insertions >250 bp were caused by *Alu* recombinations. *Alu* sequences also contributed significantly to long (>250 bp) deletions. Five deletions were caused by homologous *Alu–Alu* recombination and six by non-homologous recombination. Distribution of the long deletions along the *BRCA1* genes was not random (Fig. 1B); large deletions were, as expected, found in the largest introns. *Alu* insertions were more dispersed. However, there appears to be a hotspot for retroposition, with eight independent *Alu* insertions at positions around 8–10.5 kb (intron 3). In conclusion, our analysis shows that *Alu*s were the main force shaping primate *BRCA1* genes.

Most *Alu* repeats involved in disease-associated genomic rearrangements are retained in non-human primates

The human *BRCA1* gene contains 129 *Alu* elements, which is equivalent to ~42% of the sequence or ~1 per 0.7 kb (4; Figs 1F and 2). This high density of *Alu*s appears to be the

main source of large genomic rearrangements identified in patients with a hereditary predisposition to breast and ovarian cancers (9,10,21–27). So far, 19 different types of *Alu*-associated germ-line *BRCA1* rearrangements ranging in size from 0.5 to 23.8 kb have been described in the literature (Supplementary Material, Table S2; Fig. 2).

Analysis of the junction regions revealed that at least 26 different *Alu* elements are involved in the rearrangements (Supplementary Material, Table S2; Fig. 2). We found these high-risk elements to be remarkably stable in hominoid primates, having been conserved in chimpanzee, gorilla, orangutan and rhesus macaque. Whereas in rhesus macaque a high-risk *Alu92* fused with *Alu93*, and a high-risk *Alu99* fused with *Alu104* deleting another high-risk *Alu101* and an *Alu102* element, there was no loss of 'dangerous' *Alu*s in the hominoid primates. Only one of seven *Alu* insertions in the hominoid lineage, *Alu10*, was linked with the disease-associated rearrangements. *Alu10* inserted after macaque separated from the other lineages and was involved in a deletion in the *BRCA1* promoter region that resulted from its recombination with an upstream *Alu* repeat (28).

Structure and evolution of the *BRCA1* CDS and protein

We analyzed the *BRCA1* CDS from five primate species (Fig. 3). All of them encode proteins that comprise 1863 amino acids. The rhesus macaque protein has an in-frame 3 bp deletion in exon 11 that resulted in the loss of serine 287 and one 3 bp insertion in exon 11 that inserted isoleucine 1020. Many of the base substitutions in primate homologues (70/300 or 23% of all variable positions) involved CpG.

Table 2 shows pairwise identities in the *BRCA1* CDS. As expected, coding regions tend to be conserved more than full-length genes with promoters and introns. For proteins (Table 3), human–chimpanzee identity was 98.22% and human–gorilla identity was 98.01%. The greatest identity (98.65%) was found for gorilla and chimpanzee. At the same time, primate BRCA1s are relatively distinct from the mammalian examples shown in Table 3, and the chicken protein shows so little similarity to the mammalian protein that in many places we could not align the two.

DNA and protein conservation profiles fluctuate significantly along the sequence lengths (Fig. 3F). The terminal RING and BRCT domains are the most conserved, and the central parts are variable. We detected a similar pattern of conservation when we compared the human BRCA1 protein with the canine, rat, murine and chicken orthologues (29–31; Supplementary Material, Figs S2 and S3).

Analysis of substitutions in individual primate branches revealed an increased non-synonymous/synonymous ratio ($\omega = K_a/K_s$) in the human and chimpanzee lineages (Fig. 4), indicating positive selection (adaptive evolution). Positive selection was particularly strong in exon 11, as has been previously shown (16,32), but it was also detectable in exons 12–16 (Fig. 3D). This segment contains three human-specific non-synonymous substitutions (one non-conservative) and no synonymous ones. The chimpanzee branch exhibits conservative replacement of two residues in the same region. In both terminal lineages, non-conservative amino acid changes appear primarily in the first half of the BRCA1 protein (Fig. 3D, red bars), whereas conservative changes appear primarily in the second half. Terminal segments, on the other hand, have been under conservative negative selection (Fig. 3D–F). Indeed, both DNA and protein sequences conservation are significantly negatively correlated with the ω ratio, the correlation is -0.399 in the case of DNA identity ($P = 0.040$; Spearman's rank coefficient) and -0.663 ($P < 0.001$) for protein identity (Supplementary Material, Table S3). The non-synonymous rate significantly varies along the CDS, however, detailed analysis of the codon adaptation index (CAI) rules out the possibility that this variation is driven by selection on optimal codon usage (Fig. 3E; Supplementary Material, Table S3).

Conservation of specific structures in the RING and BRCT domains and sites of phosphorylation

The majority of the known cancer-causing BRCA1 missense mutations are localized in the RING finger and BRCT domains (33 and references therein). Using the available crystal structure of the domains (12,34), we compared primate BRCA1 with known BRCA1 proteins to investigate in detail the interspecies conservation within the RING and BRCT domains.

The RING domain found in various proteins is characterized by a conserved pattern of eight cysteine and histidine residues forming a pair of Zn^{2+}-binding sites (I and II). The BRCA1 RING domain is, as expected, strongly conserved within those sites (Fig. 5). In addition, the regions close to the active sites—the central α-helix, a β-strand and adjacent segments—are strictly conserved, not only among the analyzed primates but also in xenopus, chicken, dog, mouse and rat. In primates, the few replacements observed were limited mostly to the long N- and C-terminal α-helices. Surprisingly, *in vitro* mutations in the site II domain do not disrupt the conformation needed for its proper dimerization with its heterodimeric partner BARD1, suggesting that the main function of the conserved residues near sites I and II is interaction with other proteins, such as the ubiquitin conjugation enzymes (35).

The C-terminus of BRCA1 has a more complex structure, consisting of two BRCT repeats connected by a 23 amino acid linker (7,34; Fig. 6). The peripheral regions harbor the majority of variable sites in the N-terminal repeat. The inner region consists of three highly conserved structural motifs—sheets β3–4 and helix α2. The linker region is variable except for the alpha helix αL. The C-terminal repeats are relatively more flexible, but sheet β2′ and some neighboring residues are highly conserved. Interestingly, the above-mentioned conserved region overlaps the BACH1 helicase interaction regions (36). The majority of replacements occurred between mammal and other vertebrates; changes were nearly absent in primates. Gaps were not allowed except at the C-end of the linker and the most C-terminal part of the BRCA1 proteins (Fig. 6).

The two BRCT repeats in BRCA1 interact through three α-helices—α2 from the N-terminus and α1′ and α3′ from the C-terminal repeat. Similar tandem BRCT repeats are common in other proteins, such as 53PB1, RAD9, RAD4 and DNA ligase IV (34). The sequence alignment between the proteins shows conservation of α1′ and α3′ helices

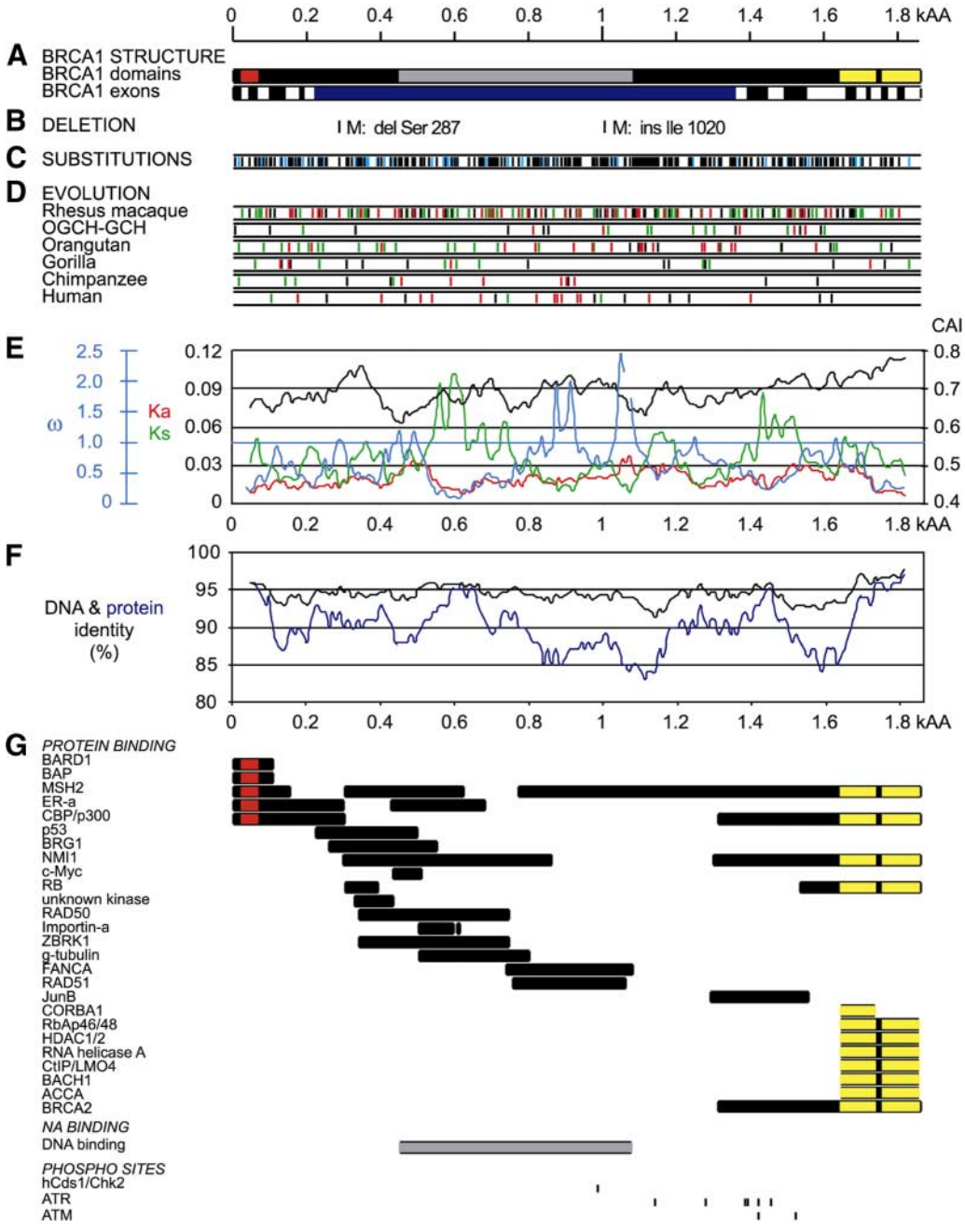

Figure 3. Structure of the *BRCA1* CDS and proteins, and evolution in primates. The scale of all plots corresponds to the protein alignment of 1863 amino acids, and positions in the CDS were scaled accordingly. Note that the CDS starts from the third exon; the non-coding first two exons and the non-coding part of the third and last exons are excluded. (**A**) Structure of human *BRCA1* CDS and protein. The first scheme shows the positions of the major domains in the BRCA1 protein. The RING finger domain is marked in red, the BRCT terminal repeats in yellow and the DNA-binding domain in gray. The second scheme shows the positions of the coding exons 3–24 in the CDS (the odd exons are black and the even ones are white). The largest exon, exon 11, is blue. **(B)** Positions of rhesus macaque-specific indels. (**C**) Variable positions in the CDS alignment. Blue positions correspond to CpG dinucleotides. (**D**) CDS substitutions during evolution. We derived the expected ancestor CDS using maximum-likelihood codon reconstruction implemented in PAML. Non-synonymous/synonymous ($\omega = K_a/K_s$) ratios were free to vary in all branches. Positions marked in green correspond to synonymous changes in a given lineage. Bars representing non-synonymous change are black if conservative and red if non-conservative. We defined conservative amino acid replacements as those with $P > 0.5$ on the basis of the Gonnod PAM250 substitution matrix (63). (**E**) Comparison of ω, K_a and K_s rates with the codon adaptation index (CAI). K_a (red) and K_s (green) values and the ω ratio (blue) are for all branches (fixed ω ratio), CAI is the average for all five primates (note that CAI differences between the species are small). The blue horizontal line corresponds to $\omega = 1$. We set the window to 300 bp (100 amino acids) with a 30 bp (10 amino acids) step. (**F**) Conservation at the nucleotide and protein levels in primates. The *Y*-axis corresponds to the proportion of conserved (identical) positions in the CDS and the protein alignment. The former uses a 300 bp overlapping window and 30 bp steps, the latter 100 amino acids windows and 10 amino acids steps. (**G**) Known binding sites on the *BRCA1* protein (after 6,67,68). The filled boxes show positions of *BRCA1* interaction with the names of the interacting agents to the left. Protein interactions are labeled in black; overlaps with the RING structural domain and/or BRCT are red and yellow, respectively. The DNA-binding domain is gray; black bars at the bottom indicate positions of phosphorylated serine/tyrosine residues.

Table 2. Pairwise identity (%) of *BRCA1* CDS

	Macaque	Orangutan	Gorilla	Chimpanzee	Human
Macaque		96.13	96.13	96.26	96.13
Orangutan	96.08		98.33	98.51	98.3
Gorilla	96.08	98.33		99.33	99.14
Chimpanzee	96.2	98.51	99.33		99.28
Human	96.08	98.3	99.14	99.28	

We calculated pairwise identities for five CDS. The values above the diagonal show DNA identities calculated without indels, those under the diagonal show DNA identities calculated with indels.

and, to lesser extent, in α2′. On the other hand, comparative analysis of several BRCA1 proteins reveals that the α2′ helix is the most conserved of all helices. Both α1′ and α3′ contain several changes, including non-conservative replacements.

BRCA1 is phosphorylated at multiple sites, mainly on serine and to a lesser extent on tyrosine residues (37). Several phosphorylation sites, modified by at least three different kinases have been identified so far in human BRCA1 (Fig. 3G, bottom). All these positions are invariant in primates. Serine 988, 1280, 1387, 1457 and 1524 and tyrosine 1394 are invariant in canine, rat and mouse. Serine 1143 is replaced by phenylalanine in rat and mouse; serine 1423 is replaced by asparagine in mouse. On the other hand, 57% (128/224) of serine residues as well as 74% (23/31) of tyrosine residues were variable in mammals.

Human mutations and sequence conservation

We used the April 2004 version of the BIC database to collect *BRCA1* mutations. The set included 8588 mutations scattered over 1090 mostly exonic nucleotide positions (885 protein positions); 32% were missense, 54% were frameshift and 12% were nonsense mutations; other categories, such as synonymous mutations, splice variants or large indels, were less frequent. Out of 1246 distinct (non-redundant) protein mutations, 38% (473) were missense, 40% (502) frameshift and 14% (176) nonsense. To reduce the bias caused by a high proportion of founder mutations in the BIC, in the following parts we concentrated mostly on non-redundant sites with known mutations rather than on total number of mutations.

We investigated whether CpG-induced changes were over-represented in the BIC mutation database. The human *BRCA1* CDS contains 43 CpG dinucleotides (86 bp). We found 54 non-redundant DNA mutations within the CpGs (62.8% of CpG positions; Fig. 3C). The *BRCA1* CDS has 5503 non-CpG positions (5598−86 CpG sites) that have mutated in 1016 different places (18.5% of positions). Thus, the mutation frequency at CpG positions is greatly increased ($P < 0.0001$; chi-square test). Most of the mutations were missense mutations (46 mutations in 86 CpG sites, 411 mutations in 5503 non-CpG sites; $P < 0.0001$; chi-square test). In addition, comparison of the total (redundant) number of mutations revealed a significant bias towards CpGs sites (data not shown). Although the observed frequencies of CpG mutations could be influenced by biased submission into the BIC, our results indicate that the CpG dinucleotides present an increased risk of mutation, confirming the preliminary results obtained by Rodenhiser *et al.* (38), who analyzed the relatively small number of mutations available in the BIC at that time.

Figure 7C shows that the density of sites harboring at least one mutation as well as the distribution number of mutations are non-random, and there are two peaks in the N- and C-terminal domains. A separate analysis of missense, nonsense and frameshift mutation sites (Fig. 7D) revealed that the pattern is the strongest for missense mutations, and the density of positions with missense mutations was quite similar to the conservation profiles in Figure 3F; Supplementary Material, Figures S2 and S3. A correlation analysis (Supplementary Material, Table S4) showed that the density of missense sites correlates positively with DNA identity in primates and protein conservation in both primates and mammals. The obvious interpretation is that most mutations are not tolerated in high conservation areas and thus their abundance in the database reflects a detection bias for deleterious replacements. Indeed, BIC missense mutations tend to be non-conserved in the conserved regions compared with the flexible regions (Fig. 7E; Supplementary Material, Table S4), strongly supporting detection bias.

The distribution of frameshift and nonsense mutations, on the other hand, is independent of sequence conservation (Fig. 7D; Supplementary Material, Table S4). Therefore, it seems that frameshift and nonsense mutations have a similar phenotype. Presumably, the nonsense-mediated mRNA decay pathway (39,40) prevents the production of such truncated proteins. A disproportionate representation of founder mutations (such as those found in Ashkenazi Jewish individuals) in the BIC was partially eliminated by our focus on mutation sites and not on the actual number of mutations. Although some BIC bias may affect our results on frameshift and nonsense mutations, it is unlikely that it would strongly affect our conclusions on the distribution of missense mutations, because a typical (random-like) noise should decrease, not increase, the statistical significance of the results.

Prediction of deleterious missense mutations

We applied two related computer-based methods for prediction of deleterious missense mutations: (i) predictions using the SIFT program (41), and (ii) the ancestral sequence (AS) method (14). Both methods use evolutionary conservation of the BRCA1 protein to predict deleterious changes. Only highly variable positions are expected to tolerate non-conservative mutations (as estimated by a protein substitution matrix), whereas conserved positions are expected to tolerate only replacements that have similar physicochemical properties (conservative changes). The primate alignment contains a small number of replacements and thus cannot be used efficiently to distinguish polymorphic changes. Including chicken BRCA1 in the comparison, on the other hand, lowers alignment quality in many places and is likely to produce a misleading conservation profile and thus bias the statistics. The mammalian alignment seems to provide the best balance, and indeed, its predictive value was clearly superior (Supplementary Material, Table S4). We therefore used five

Table 3. Pairwise identity (%) of BRCA1 proteins

	Chicken	Dog	Rat	Mouse	Macaque	Orangutan	Gorilla	Chimpanzee	Human
Chicken		31.49	28.93	30.08	32.58	32.73	32.73	32.67	32.54
Dog	25.31		54.07	52.79	74.29	74.17	74.17	74.44	74.12
Rat	23.1	51.56		80.17	56.41	57.5	57.16	57.44	56.94
Mouse	24.05	50.55	79.47		55.0	56.49	56.26	56.32	56.04
Macaque	26.32	73.15	53.68	53.43		93.5	93.39	93.55	93.12
Orangutan	26.41	72.95	54.66	53.93	93.4		97.1	97.31	96.77
Gorilla	26.41	72.95	54.34	53.72	93.29	97.1		98.65	98.01
Chimpanzee	26.36	73.22	54.61	53.77	93.45	97.31	98.65		98.22
Human	26.26	72.9	54.13	53.51	93.02	96.77	98.01	98.22	

We calculated pairwise identities for nine protein sequences. The values above the diagonal show DNA identities calculated without indels, those under the diagonal show protein identities calculated with indels.

primate proteins and a canine, rat and mouse protein to predict the effect of mutations on human BRCA1. For all 473 different missense mutations reported in the BIC database, we predicted the effect on protein function and thus the predisposition to an increased risk of cancers (Supplementary Material, Table S5). In addition, using SIFT, we obtained a complete matrix of all possible replacements (even unreported ones) in the BRCA1 protein and an estimate of their deleterious effect (Supplementary Material, Table S6).

Tables 4 and 5 summarize the results. The proportion of BIC missense mutations predicted to be tolerated was about 38% by the SIFT method and 45% by the AS method. The other mutations would be expected to affect protein function and therefore predispose to breast and ovarian cancers. Both methods correctly predicted that a very small fraction of replacements would be tolerated in the area coding the conserved terminal RING and BRCT domains. It has been estimated that $<10\%$ of missense mutations are tolerated in these segments (11–13); our analysis predicted 2–15%. The predicted deleterious mutations listed in Supplementary Material, Table S5 would be a good set for correlations of phenotype with missense mutations in breast–ovarian cancer families and for studying the effects of the mutations on BRCA1 structure and function.

DISCUSSION

This work represents the first systematic study of evolutionary changes in the entire *BRCA1* locus in non-human primates. Our comparative analysis of *BRCA1* genes was simplified by the use of TAR cloning, a technique that allows the direct isolation of gene homologues (17,42). Using this cloning strategy, we isolated the genomic *BRCA1* clones containing 5′, 3′ and all intron sequences from chimpanzee, gorilla, orangutan and rhesus macaque. This allowed us to develop sequence data that are either not present in the sequence databases or present as poor quality draft sequences.

Interspecies comparisons revealed the existence of an unexpectedly high number of indels. The proportion of *BRCA1* indels was approximately three times higher than previously observed for full-length *ASPM* genes from the same primates (42). Most long indels were associated with *Alu* sequences. The majority of *Alu* insertions took place in the ancestral lineage leading to hominoid primates after the split of *Hominidae* (25–14 MYA) and the rhesus macaque branch; more recent hominoid lineages acquired mostly deletions. As no significant rearrangements involving other repetitive sequences were observed, we concluded that *Alu* repeats were the main contributors to evolution of the *BRCA1* non-CDS, and that the *BRCA1* gene represents a genomic hotspot for both retroposition and recombination of *Alu* repeats.

Our analysis of *BRCA1* gene homologues revealed that most *Alu* elements involved in genomic rearrangements in humans are retained in non-human primates. The fact that the high-risk repeats were not eliminated by selection during primate evolution suggests a role in gene expression. Alternatively, ineffective selection against late-onset diseases may explain the tolerance of many dangerous *Alu* repeats in ancestral lineages.

Recombination between *Alu* repeats is an important contributor to genetic disorders (reviewed in 43). Genomic rearrangements may account for up to 30% of all *BRCA1* mutations identified in breast cancer families (9,10,21–27). Given the high frequency of germ-line *Alu*-mediated *BRCA1* rearrangements, it would not be surprising if *Alu*s also contribute to at least some cases of sporadic breast and ovarian cancers by stimulating somatic recombinations, as has been recently suggested (44).

Using partial *BRCA1* CDS derived from exon 11, Huttley *et al.* (16) showed that the RAD51-interacting domain evolved under positive selection in human and chimpanzee. Comparison of primate BRCA1 proteins has shown that the positive selection was not restricted to the RAD51-interacting domain but extended to most of the protein sequence, including the part encoded in exons 12–16. The terminal parts of BRCA1 encoding the RING and BRCT domains experienced strong conservation both in human and non-human primate lineages as was previously reported for other vertebrates (29–31). Such a mosaic of positive and negative selection has been previously described for other proteins (42,45,46).

Our analysis revealed that the most conserved sequences form specific tertiary structures. In the RING domain, the most conserved residues were closely packed around the Zn^{2+}-binding sites. We speculate that these Zn^{2+}-binding sites interact with ubiquitin conjugation enzymes (35). Surprisingly, the most conserved part of the BRCT domain

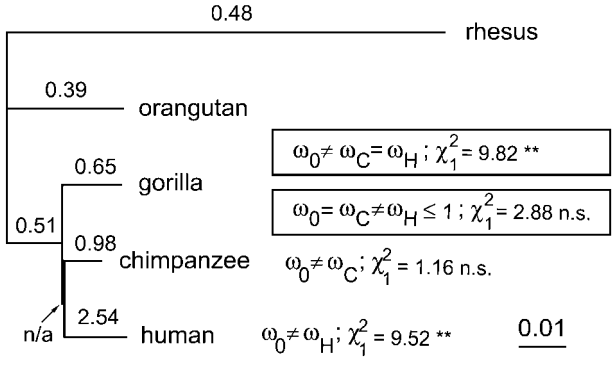

Figure 4. Phylogenetic tree and K_a/K_a ratio for BRCA1 proteins. For codons, we computed the phylogenetic tree and ω (K_a/K_a) ratios using the maximum-likelihood method implemented in PAML. The ω ratio was set free to vary in all branches. Branch labels mark the ω ratios for corresponding branches. We tested several hypotheses using likelihood ratio tests (69). The ω rate is designated as ω_H for the human lineage, ω_c for the chimpanzee lineage, and ω_o for all other lineages. **, indicates $P < 1\%$, $\chi^2_1 = 6.63$. In addition to testing the different ω in the human and chimpanzee lineages, we tested the hypotheses that (i) both the chimpanzee and human lineages evolved at a constant rate that differed from that of the rest of the tree (significant at $P < 1\%$; boxed) and (ii) the human branch evolved at a ω rate significantly >1 (not significant; boxed).

is not at an interface where the two BRCT repeats interact, but is found mostly around the inner part of the N-terminal repeat. It co-localizes with the critical residues for BRCA1–BACH1 interaction (36). BACH1 functions with BRCA1 as a mediator of double-strand break repair, and deleterious BACH1 mutations seem to predispose affected individuals to early-onset breast cancer (8,47). On the basis of interspecies BRCA1 comparisons, the majority, if not all, of the mutations in the RING Zn^{2+}-binding region and in the most conserved regions of the BRCT domain represent a group of mutations that strongly predispose to breast and ovarian cancers. Similarly, the strong conservation of certain phosphorylation sites indicates a critical role for them in protein function and therefore suggests that altered residues may result in cancer predisposition.

Amino acid substitutions resulting from nonsense, missense and frameshift mutations in the BIC database for BRCA1 were unevenly distributed along the protein. Although the frameshift and nonsense mutants did not exhibit any specific clustering, the frequency of missense mutants correlated positively with BRCA1 conservation. This strongly suggests that the clustering of missense mutants within conservative regions is driven by different phenotypic manifestations in the conserved and variable regions, and therefore by a detection bias for deleterious mutations.

Most of the 473 independent missense mutations reported in the BIC prior to April 2004 play an unknown role in breast cancer susceptibility. The effect of the mutations has been difficult to characterize because the function of some regions of the BRCA1 protein is poorly understood. Recently, interspecies comparisons have been made in an attempt to predict the role of missense changes in breast cancer susceptibility (14,15). By aligning exon 11 sequences from 57 eutherian and 8 marsupial mammals and categorizing amino acid sites by the degree of conservation, investigators identified 21 missense mutations that are likely to influence gene function and thereby contribute to cancer susceptibility. In our work, we applied a similar approach to analyze complete sequences of eight BRCA1 homologues. Our prediction yielded 55–62% deleterious missense mutations in the BIC, including those identified previously (14,15).

Our analysis is likely to overpredict deleterious mutations owing to the small number of sequences we used for comparison. At the same time, the use of BRCA1 sequences from more distant species might disguise some deleterious mutations as a result of fixation of mutations that are deleterious in humans (48). These cases can be explained either by compensatory effects of other mutations or by relaxed selection of late-onset phenotypes in the distantly related species (49,50). Therefore, adding the BRCA1 protein sequences from other primate species to the analysis may produce better estimates of mutation effects. While we will sequence other primate genes in future work, the predicted deleterious missense mutations in Supplementary Material (Tables S5 and S6) may be helpful for further detailed analyses of phenotypic correlation of missense mutations in breast–ovarian cancer families.

In conclusion, comparison of primate *BRCA1* gene homologues allowed us to reconstruct an evolutionary history of the entire *BRCA1* locus. The impact of *Alu* repeats, CpG dinucleotides and a mixture of positive selection and conservation of the CDS were the main factors that shaped *BRCA1* evolution in primates. Interspecies sequence comparisons also provided a basis for the identification of conservative amino acid residues in BRCA1 and for the prediction of missense changes that compromise BRCA1 function. Missense mutations that confer the highest predisposition to breast and ovarian cancers are located in the evolutionarily conserved regions, phosphorylated residues and especially in specific protein-binding domains. Genomic clones of *BRCA1* homologues with regulatory elements may also be used for comparative gene expression studies to identify the role of intron regions in gene regulation.

MATERIALS AND METHODS

TAR cloning of *BRCA1* gene homologues by *in vivo* recombination in yeast

To isolate the full *BRCA1* gene from the chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*) and rhesus macaque (*Macaca mulatta*) genomes, we used TAR vector pVC-BC1 containing 5′ and 3′ targeting sequences (hooks) of the human *BRCA1* gene (51). The 5′ targeting sequence of 769 bp corresponds to positions 2147–2915 in the genomic sequence L78833 (GI: 1698398) and the 3′ targeting sequence of 325 bp corresponds to positions 82 936–83 260 in the genomic sequence L78833 (GI: 1698398). We PCR-amplified the 5′ *BRCA1* sequence from genomic DNA using the primer pair BC1 (5′-CTCGAGG TCACTAAAACGAT-3′) and BC2 (5′-GAATTCCAGCATG CGTTGCGG-3′). We PCR-amplified the 3′ *BRCA1* sequence from genomic DNA using the primer pair BC3 (5′-GAATT CCAATTGGGCAGATGTGT-3′) and BC4 (5′-GGATCCAA GGGAGACTTCAAG-3′). We cloned the PCR products into a polylinker of a basic TAR vector as *Xho*I–*Eco*RI and

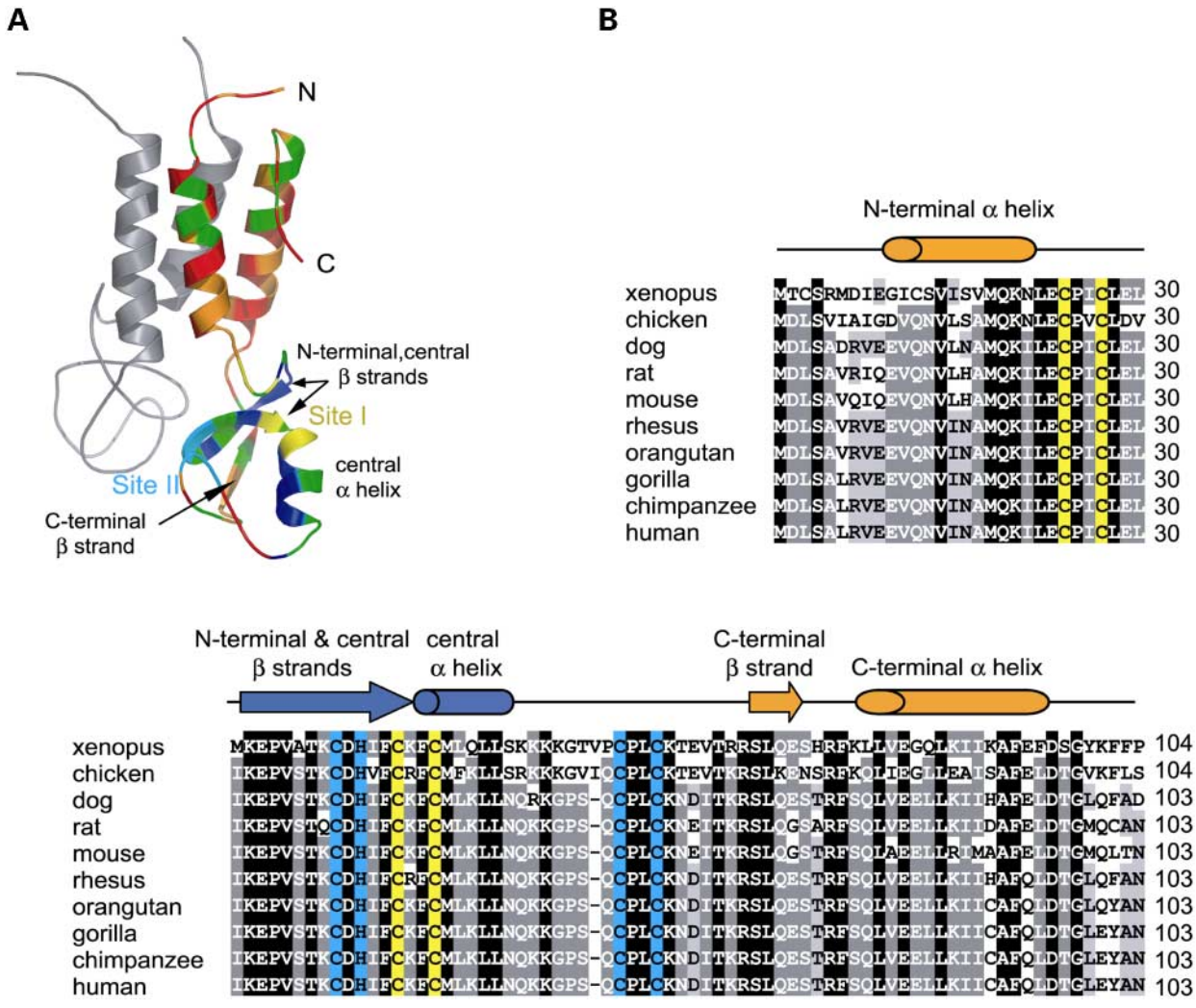

Figure 5. Structural localization of substitutions in the RING domain. (**A**) Structure of the RING domain of human BRCA1 (12,35). We show both the BRCA1 RING domain (various colors) and the interacting BARD1 domain (gray). Background is orange for the BRCA1 strand and blue for the RING domain. Sites I and II mark the pairs of Zn^{2+} ligand domains defining the RING domain. The RING domain is composed of one (central) α-helix, one β-sheet (previously described as two independent N-terminal and central sheets) (12,35) and variable regions without a clearly defined secondary structure. Coordinates were extracted from the pdb file 1MJ7 (12,35). We compared variable positions in vertebrate (primates plus mouse, rat, dog, chicken and xenopus) protein alignments. Positions with conservative changes ($P > 0.5$, Gonnet PAM250 matrix) are shown in green and non-conservative positions in red. (**B**) Protein alignment underlying the 3D structures. The Zn^{2+} ligand domains are highlighted in different colors.

*Eco*RI–*Bam*HI fragments. Before performing the transformation experiments, we linearized the TAR cloning vector with *Eco* RI to release targeting hooks. We prepared genomic DNA samples from chimpanzee, gorilla, orangutan and rhesus macaque fibroblast culture cell lines (Coriell Institute for Medical Research) on agarose plugs. For transformations, we used the highly transformable *Saccharomyces cerevisiae* strain VL6–48 (*MATα, his3–Δ200, trp1–Δ1, ura3–52, lys2, ade2–101, met14*), which has *HIS3* deleted (52). Spheroplast transformation experiments were carried out as previously described (52). The yield of transformants per μg vector, 2 μg genomic DNA, and 5×10^8 spheroplasts was 3–10 colonies. We obtained approximately 300 His^+ transformants for each species. To identify clones positive for *BRCA1*, we examined yeast transformants by PCR using diagnostic primers 11F (5′-CTCAGTTCAGAGGCAACGAA-3′) and 11R (5′-GGAGCCCACTTCATTAGTAC-3′) specific for *BRCA1* exon 11. These primers amplify a 302 bp sequence by PCR. We isolated yeast genomic DNA from individual transformants or pools and PCR-amplified them as previously described (52). The yield of *BRCA1*-positive clones from human, chimpanzee, gorilla, orangutan and rhesus macaque genomic DNAs was ∼1%. To confirm that the copies were complete, we PCR-analyzed three independent TAR YAC isolates for each species using a set of primers specific for each of the 24 exons (1). We obtained the same size PCR products for each species isolate with each primer pair. Finally, we examined *Alu* profiles of the YACs after *Taq*I digestion and found that they were indistinguishable (data not shown). From these studies we concluded that we had isolated non-arranged

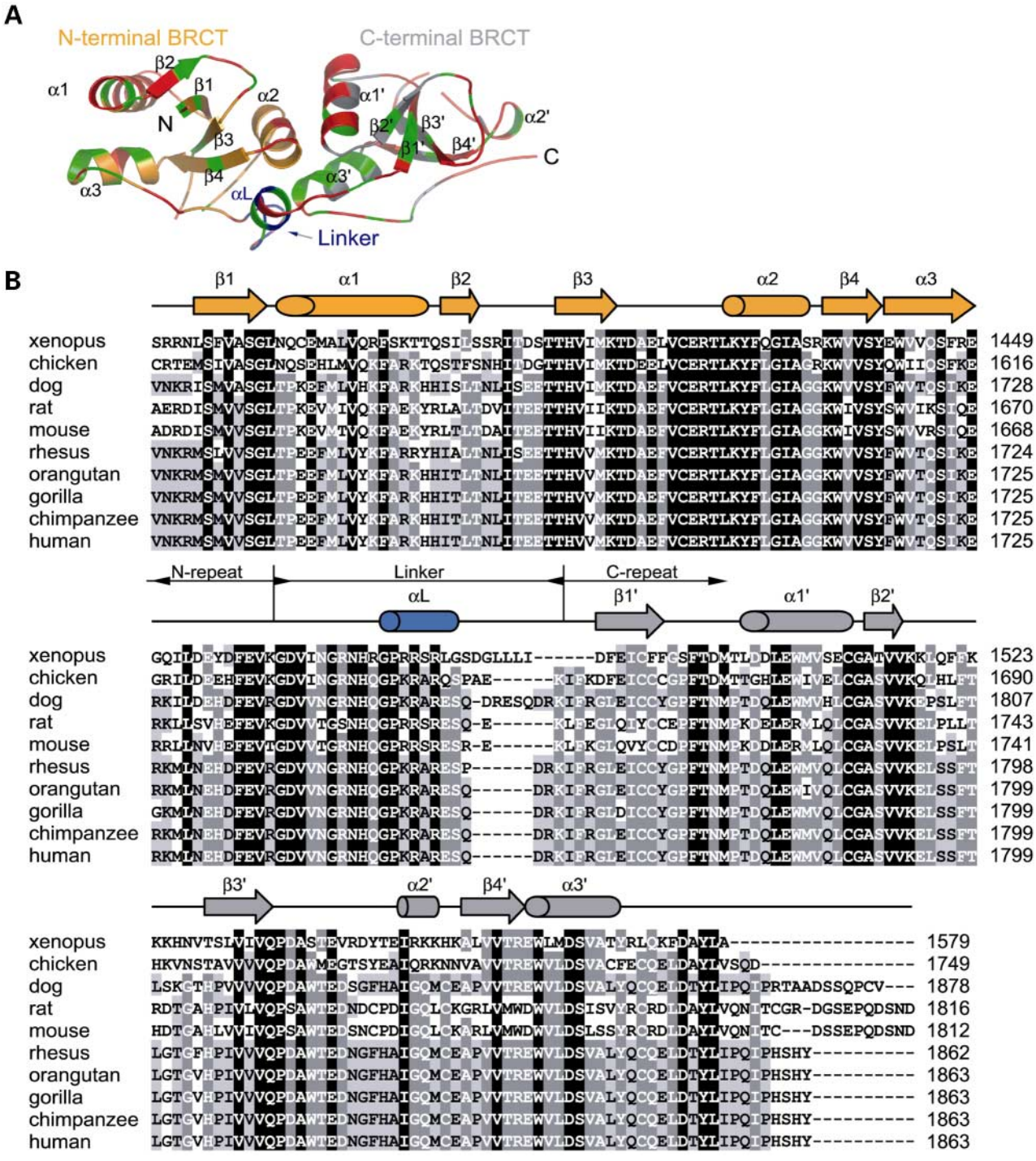

Figure 6. Structural localization of substitutions in the BRCT domain. (**A**) Structure of the BRCT region of human BRCA1 (34). The two BRCT repeats are separated by a 23 amino acids linker and all three domains are marked by different colors. We slightly modified the secondary structure coordinates from the pdb file 1JNX to match the data from the original paper by Williams *et al.* (34). We compared variable positions in vertebrate (primates plus mouse, rat, dog, chicken and xenopus) protein alignments. Positions with conservative changes ($P > 0.5$, Gonnet PAM250 matrix) are shown in green and non-conservative positions are in red. (**B**) Protein alignment underlying the 3D structures.

genomic copies of all the *BRCA1* gene homologues. We retrofitted the individual *BRCA1* YACs corresponding to each species into BACs by homologous recombination in yeast using BAC/Neo^R retrofitting vector BRV1 and used them to transform a *recA* DH10B *Escherichia coli* strain (52). Before sequencing, we confirmed the integrity of the inserts in BACs by *Not*I, *Hin*dIII, *Eco*RI and *Pst*I digestion. Chimpanzee, gorilla, orangutan and rhesus macaque TAR clones containing full-size *BRCA1* genes were directly sequenced from BAC DNAs (53). Identical *Alu*-profiles of

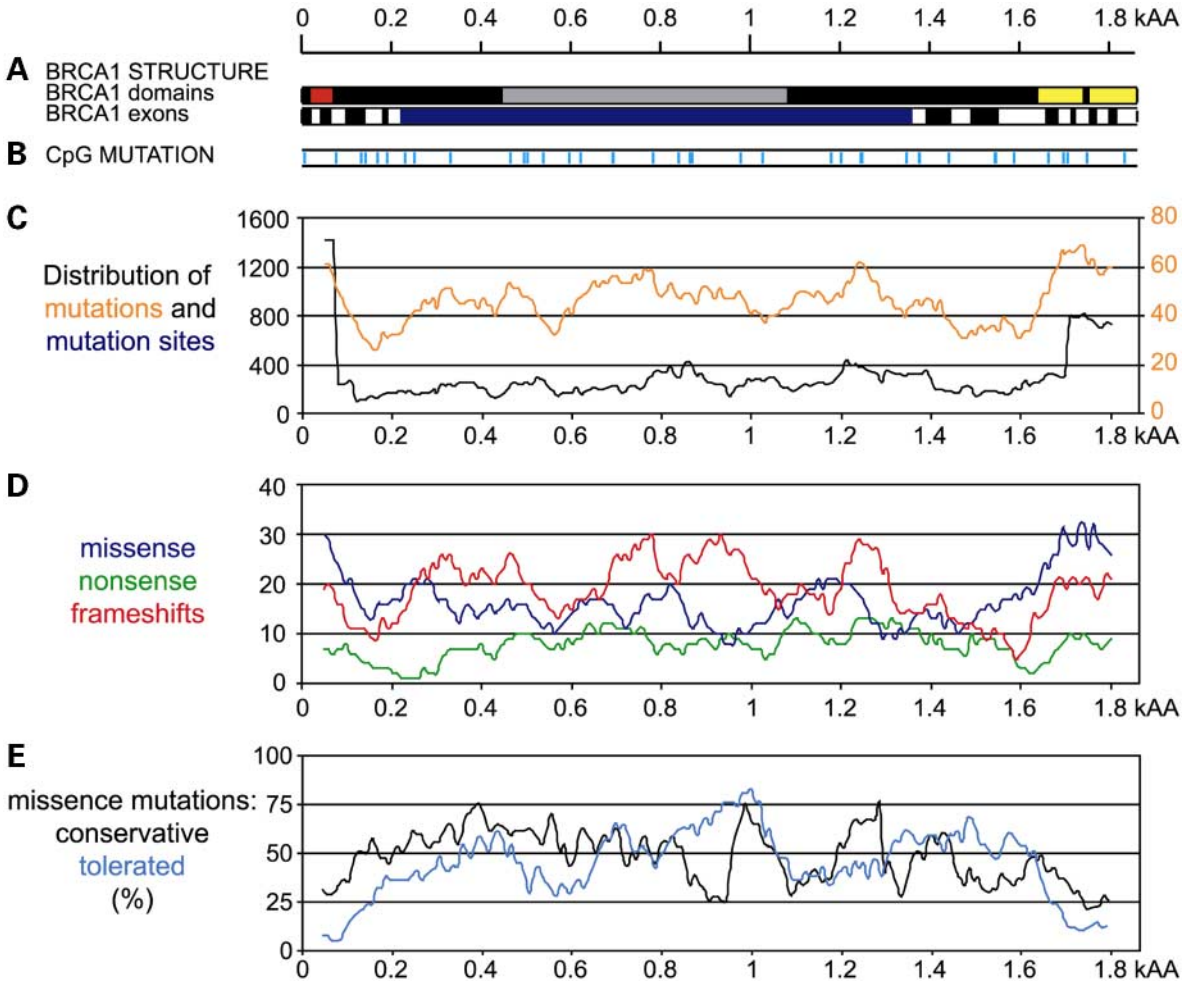

Figure 7. Positions of human mutations. Mutation positions were obtained from the BIC database. (**A**) Structure of the human *BRCA1* CDS and protein (Fig. 3A). (**B**) Positions of BIC mutations located within CpG dinucleotides. (**C**) Positions of known mutations in the human *BRCA1* protein (1090 positions, 885 sites on the protein sequence). We obtained the point mutations and small indel positions from the BIC database. The plot uses 100 amino acids overlapping windows and 10 amino acids steps. The black line (left *Y*-axis) measures the number of mutations per 100 amino acids; as many mutations are reported for the same sites, the number can exceed the window size. The orange line (right *Y*-axis) shows the proportion of non-redundant sites with reported mutations per 100 amino acids. The interval value is 0–100 and is equivalent to % mutation sites per window. (**D**) Distribution of three main mutation types along the human protein sequence. The plot uses 100 amino acids overlapping windows and 10 amino acids steps. Values in the three plots correspond to the percentage of sites with particular mutations per window. (**E**) Proportion of tolerated missense mutations from the BIC database compared with the proportion of conservative substitutions in the mammalian alignment. We estimated the effect of missense mutations using the SIFT program and an alignment of primate proteins with mouse, rat and canine orthologues (eight sequences). Any mutation with a SIFT score <0.05 was considered as changing protein function and all others as preserving protein function and therefore tolerated. We calculated the proportion of conservative substitutions in the mammalian alignment from all non-redundant missense changes in a given segment. Both profiles were obtained using 100 amino acids overlapping windows and 10 amino acids steps.

independent TAR isolates were considered as a conformation of indels. In addition, some indels were confirmed by PCR amplification from genomic DNAs (data not shown). All sequences were named and numbered according to the clone/accession identifier. Sequences were deposited into GenBank under accessions AY365046, AY589040, AY589041 and AY589042.

Sequence analysis

We aligned genomic sequences using MAVID (54; http://baboon.math.berkeley.edu/mavid/) and proteins and protein-coding DNA sequences using Dialign2.1 (55; http://bibiserv.techfak.uni-bielefeld.de/dialign/). We edited alignments manually with the Seaview editor (56; http://pbil.univ-lyon1.fr/software/seaview.html). For prediction of CpG islands we used cpgplot (EMBOSS (57) http://www.hgmp.mrc.ac.uk/Software/EMBOSS/) with the default parameters (length $\geq$200; CpG/GpC $\geq$0.6; GC $\geq$0.5). We determined the CAI by cai (EMBOSS) with a human codon use library. Censor (58; http://www.girinst.org/Censor_ServerData_Entry_Forms.html), RepeatMasker (A.F.A. Smit and P. Green unpublished data; http://www.repeatmasker.org/), Repbase Update libraries (59; http://www.girinst.org/Repbase_Update.html) and TandemRepeatFinder (60; http://tandem.bu.edu/trf/trf.html) were used to identify repetitive elements. The segmental duplication of the 5′ *BRCA1* region was localized by local BLAT searches (61). Human single nucleotide polymorphism (SNP) data were extracted

Table 4. SIFT-based analysis of deleterious mutations

Selection (amino acids)	Non-redundant missense mutations			Redundant missense mutations		
	Total	Tolerated	%	Total	Tolerated	%
1–1863	473	180	38.1	2725	1200	44.0
RING (1–103)	46	2	4.3	295	4	1.4
BRCT (1646–1863)	93	14	15.1	359	79	22.0
Fleming (282–1103)	170	88	51.8	1080	580	53.7
All-Fleming (1–281; 1104–1863)	303	92	30.4	1645	620	37.7

We performed predictions of deleterious missense mutations by SIFT (41) using mammalian protein alignment (primates plus canine, rat and mouse). We considered SIFT scores ≤ 0.05 deleterious (for individual mutations see Supplementary Material, Table S5). We show global statistics for the complete human protein, terminal domains containing RING and BRCT motifs and comparisons with the segment analyzed by Fleming *et al.* (14) as well as the complete protein without the segment. Note that exon 11 encodes amino acids 223–1365, whereas Fleming *et al.* used amino acids 282–1103.

Table 5. Ancestral sequence (AS) method-based analysis of deleterious mutations

Selection (amino acids)	Non-redundant missense mutations			Redundant missense mutations		
	Total	Tolerated	%	Total	Tolerated	%
1–1863	473	214	45.2	2725	1458	53.5
RING (1–103)	46	3	6.5	295	5	1.7
BRCT (1646–1863)	93	14	15.1	359	45	12.5
Fleming (282–1103)	170	105	61.8	1080	786	72.7
All-Fleming (1–281; 1104–1863)	303	109	36.0	1645	672	40.9

We performed predictions of deleterious missense mutations by the ancestral sequence (AS) method (14) using mammalian protein alignment (primates plus canine, rat and mouse). We considered as deleterious all replacements in invariant positions and non-conservative replacements in conservative positions (for individual mutations, see Supplementary Material, Table S5). We show global statistics for the complete human protein, terminal domains containing RING and BRCT motifs, and comparisons with the segment analyzed by Fleming *et al.* (14) and the complete protein without this segment. Note that exon 11 encodes amino acids 223–1365, whereas Fleming *et al.* used amino acids 282–1103.

from dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). To avoid possible paralogous sequence variants from the 5′ segmental duplication, we extracted all chromosome 17 SNPs and localized them on chromosome 17 using local BLAT (61) searches; only SNPs with best hits within the *BRCA1* gene were considered.

We used SNAP (http://www.hiv.lanl.gov/content/hiv-db/SNAP/WEBSNAP/SNAP.html) to detect synonymous and non-synonymous substitutions (62). Gonnet PAM250 matrix (63) was applied to classify substitutions and human mutations as conservative or non-conservative. We considered changes as conservative if the score was >0.5 (14). Human *BRCA1* mutations were downloaded from the BIC (64). We used SIFT (41; http://blocks.fhcrc.org/sift/SIFT.html) to predict deleterious missense mutations. Protein structures were visualized in PyMOL (DeLano Scientific, San Carlos, CA; http://www.pymol.org).

We applied the codon maximum-likelihood method in codeml in PAML v. 3.13 (65; http://abacus.gene.ucl.ac.uk/software/paml.html) for reconstruction of phylogenetic trees, ancestral sequences and detection of positive selection. Branch lengths and ancestral sequences were reconstructed using a free ratio model for individual branches. Phylogenetic trees were drawn in TREEVIEW (66).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

We thank Marco Montagna, Tom Scholl and Sylvie Mazoyer for providing data on *Alu*-associated genomic rearrangements in *BRCA1*, Andrew Gentles for corrections and Miriam Bloom (SciWrite Biomedical Writing and Editing Services) for professional editing. This work was supported in part by National Institutes of Health Grant 2 P41 LM 06252-04A1 (J.J.).

REFERENCES

1. Futreal, P.A., Liu, Q., Shattuck-Eidens, D., Cochran, C., Harshman, K., Tavtigian, S., Bennett, L.M., Haugen-Strano, A., Swensen, J., Miki, Y. *et al.* (1994) *BRCA1* mutations in primary breast and ovarian carcinomas. *Science*, **266**, 120–122.
2. Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., Ding, W. *et al.* (1994) A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science*, **266**, 66–71.
3. King, M.C., Marks, J.H. and Mandell, J.B. for The New York Breast Cancer Study Group. (2003) Breast and ovarian cancer risks due to inherited mutations in *BRCA1* and *BRCA2*. *Science*, **302**, 643–646.
4. Smith, T.M., Lee, M.K., Szabo, C.I., Jerome, N., McEuen, M., Taylor, M., Hood, L. and King, M.C. (1996) Complete genomic sequence and analysis of 117 kb of human DNA containing the gene *BRCA1*. *Genome Res.*, **6**, 1029–1049.
5. Rosen, E.M., Fan, S., Pestell, R.G. and Goldberg, I.D. (2003) *BRCA1* gene in breast cancer. *J. Cell Physiol.*, **196**, 19–41.
6. Welcsh, P.L. and King, M.C. (2001) *BRCA1* and *BRCA2* and the genetics of breast and ovarian cancer. *Hum. Mol. Genet.*, **10**, 705–713.

7. Koonin, E.V., Altschul, S.F. and Bork, P. (1996) BRCA1 protein products... Functional motifs... *Nat. Genet.*, **13**, 266–268.
8. Cantor, S.B., Bell, D.W., Ganesan, S., Kass, E.M., Drapkin, R., Grossman, S., Wahrer, D.C., Sgroi, D.C., Lane, W.S., Haber, D.A. and Livingston, D.M. (2001) BACH1, a novel helicase-like protein, interacts directly with *BRCA1* and contributes to its DNA repair function. *Cell*, **105**, 149–160.
9. Puget, N., Torchard, D., Serova-Sinilnikova, O.M., Lynch, H.T., Feunteun, J., Lenoir, G.M. and Mazoyer, S. (1997) A 1-kb *Alu*-mediated germ-line deletion removing *BRCA1* exon 17. *Cancer Res.*, **57**, 828–831.
10. Petrij-Bosch, A., Peelen, T., van Vliet, M., van Eijk, R., Olmer, R., Drusedau, M., Hogervorst, F.B., Hageman, S., Arts, P.J., Ligtenberg, M.J. *et al.* (1997) *BRCA1* genomic deletions are major founder mutations in Dutch breast cancer patients. *Nat. Genet.*, **17**, 341–345.
11. Monteiro, A.N., August, A. and Hanafusa, H. (1996) Evidence for a transcriptional activation function of BRCA1 C-terminal region. *Proc. Natl Acad. Sci. USA*, **93**, 13595–13599.
12. Brzovic, P.S., Rajagopal, P., Hoyt, D.W., King, M.C. and Klevit, R.E. (2001) Structure of a BRCA1–BARD1 heterodimeric RING–RING complex. *Nat. Struct. Biol.*, **8**, 833–837.
13. Vallon–Christersson, J., Cayanan, C., Haraldsson, K., Loman, N., Bergthorsson, J.T., Brondum-Nielsen, K., Gerdes, A.M., Moller, P., Kristoffersson, U. *et al.* (2001) Functional analysis of BRCA1 C-terminal missense mutations identified in breast and ovarian cancer families. *Hum. Mol. Genet.*, **10**, 353–360.
14. Fleming, M.A., Potter, J.D., Ramirez, C.J., Ostrander, G.K. and Ostrander, E.A. (2003) Understanding missense mutations in the *BRCA1* gene: an evolutionary approach. *Proc. Natl Acad. Sci. USA*, **100**, 1151-1156.
15. Ramirez, C.J., Fleming, M.A., Potter, J.D., Ostrander, G.K. and Ostrander, E.A. (2004) Marsupial BRCA1: conserved regions in mammals and the potential effect of missense changes. *Oncogene*, **23**, 1780–1788.
16. Huttley, G.A., Easteal, S., Southey, M.C., Tesoriero, A., Giles, G.G., McCredie, M.R., Hopper, J.L. and Venter, D.J. (2000) Adaptive evolution of the tumour suppressor *BRCA1* in humans and chimpanzees. Australian breast cancer family study. *Nat. Genet.*, **25**, 410–413.
17. Kouprina, N. and Larionov, V. (2003) Exploiting the yeast *Saccharomyces cerevisiae* for the study of the organization and evolution of complex genomes. *FEMS Microbiol. Rev.*, **27**, 629–649.
18. Brown, M.A., Lo, L.J., Catteau, A., Xu, C.F., Lindeman, G.J., Hodgson, S. and Solomon, E. (2002) Germline *BRCA1* promoter deletions in UK and Australian familial breast cancer patients: Identification of a novel deletion consistent with *BRCA1*:psi*BRCA1* recombination. *Hum. Mutat.*, **19**, 435–442.
19. Puget, N., Gad, S., Perrin-Vidoz, L., Sinilnikova, O.M., Stoppa-Lyonnet, D., Lenoir, G.M. and Mazoyer, S. (2002) Distinct *BRCA1* rearrangements involving the *BRCA1* pseudogene suggest the existence of a recombination hot spot. *Am. J. Hum. Genet.*, **70**, 858–865.
20. Goodman M. (1999) The genomic record of Humankind's evolutionary roots. *Am. J. Hum. Genet.*, **64**, 31–39.
21. Montagna, M., Santacatterina, M., Torri, A., Menin, C., Zullato, D., Chieco-Bianchi, L. and D'Andrea, E. (1999) Identification of a 3 kb *Alu*-mediated *BRCA1* gene rearrangement in two breast/ovarian cancer families. *Oncogene*, **18**, 4160–4165.
22. Puget, N., Sinilnikova, O.M., Stoppa-Lyonnet, D., Audoynaud, C., Pages, S., Lynch, H.T., Goldgar, D., Lenoir, G.M. and Mazoyer, S. (1999) An *Alu*-mediated 6-kb duplication in the *BRCA1* gene: a new founder mutation? *Am. J. Hum. Genet.*, **64**, 300–302.
23. Puget, N., Stoppa-Lyonnet, D., Sinilnikova, O.M., Page, S., Lynch, H.T., Lenoir, G.M. and Mazoyer, S. (1999) Screening for germ-line rearrangements and regulatory mutations in *BRCA1* led to the identification of four new deletions. *Cancer Res.*, **59**, 455–461.
24. Rohlfs, E.M., Chung, C.H., Yang, Q., Skrzynia, C., Grody, W.W., Graham, M.L. and Silverman, L.M. (2000) In-frame deletions of *BRCA1* may define critical functional domains. *Hum. Genet.*, **107**, 385–390.
25. Rohlfs, E.M., Puget, N., Graham, M.L., Weber, B.L., Garber, J.E., Skrzynia, C., Halperin, J.L., Lenoir, G.M., Silverman, L.M. and Mazoyer, S. (2000) An *Alu*-mediated 7.1 kb deletion of *BRCA1* exons 8 and 9 in breast and ovarian cancer families that results in alternative splicing of exon 10. *Genes Chromosomes Cancer*, **28**, 300–307.
26. Montagna, M., Dalla Palma, M., Menin, C., Agata, S., De Nicolo, A., Chieco-Bianchi, L. and D'Andrea, E. (2003) Genomic rearrangements account for more than one-third of the *BRCA1* mutations in northern Italian breast/ovarian cancer families. *Hum. Mol. Genet.*, **12**, 1055–1061.
27. Hogervorst, F.B., Nederlof, P.M., Gille, J.J., McElgunn, C.J., Grippeling, M., Pruntel, R., Regnerus, R., van Welsem, T., van Spaendonk, R., Menko, F.H. *et al.* (2003) Large genomic deletions and duplications in the *BRCA1* gene identified by a novel quantitative method. *Cancer Res.*, **63**, 1449–1453.
28. Swensen, J., Hoffman, M., Skolnick, M.H. and Neuhausen, S.L. (1997) Identification of a 14 kb deletion involving the promoter region of *BRCA1* in a breast cancer family. *Hum. Mol. Genet.*, **6**, 1513–1517.
29. Abel, K.J., Xu, J., Yin, G.Y., Lyons, R.H., Meisler, M.H. and Weber, B.L. (1995) Mouse *BRCA1*: localization sequence analysis and identification of evolutionarily conserved domains. *Hum. Mol. Genet.*, **4**, 2265–2273.
30. Szabo, C.I., Wagner, L.A., Francisco, L.V., Roach, J.C., Argonza, R., King, M.C. and Ostrander, E.A. (1996) Human, canine and murine *BRCA1* genes: sequence comparison among species. *Hum. Mol. Genet.*, **5**, 1289–1298.
31. Orelli, B.J., Logsdon, J.M., Jr and Bishop. D.K. (2001) Nine novel conserved motifs in BRCA1 identified by the chicken orthologue. *Oncogene*, **20**, 4433–4438.
32. Yang, Z. and Nielsen, R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*, **19**, 908–917.
33. Shen, D. and Vadgama, J.V. (1999) *BRCA1* and *BRCA2* gene mutation analysis: visit to the Breast Cancer Information Core (BIC). *Oncol. Res.* **11**, 63–69.
34. Williams, R.S., Green, R. and Glover, J.N. (2001) Crystal structure of the BRCT repeat region from the breast cancer–associated protein BRCA1. *Nat. Struct. Biol.*, **8**, 838–842.
35. Brzovic, P.S., Meza, J.E., King, M.C. and Klevit, R.E. (2001) BRCA1 RING domain cancer-predisposing mutations. Structural consequences and effects on protein–protein interactions. *J. Biol. Chem.*, **276**, 41399–41406.
36. Joo, W.S., Jeffrey, P.D., Cantor, S.B., Finnin, M.S., Livingston, D.M. and Pavletich, N.P. (2002) Structure of the 53BP1 BRCT region bound to p53 and its comparison to the BRCA1 BRCT structure. *Genes Dev.*, **16**, 583–593.
37. Okada, S. and Ouchi, T. (2003) Cell cycle differences in DNA damage-induced BRCA1 phosphorylation affect its subcellular localization. *J. Biol. Chem.*, **278**, 2015–2020.
38. Rodenhiser, D., Chakraborty, P., Andrews, J., Ainsworth, P., Mancini, D., Lopes, E. and Singh, S. (1996) Heterogenous point mutations in the *BRCA1* breast cancer susceptibility gene occur in high frequency at the site of homonucleotide tracts, short repeats and methylatable CpG/CpNpG motifs. *Oncogene*, **12**, 2623–2629.
39. Liu, H.X., Cartegni, L., Zhang, M.Q. and Krainer, A.R. (2001) A mechanism for exon skipping caused by nonsense or missense mutations in *BRCA1* and other genes. *Nat. Genet.*, **27**, 55–58.
40. Perrin-Vidoz, L., Sinilnikova, O.M., Stoppa-Lyonnet, D., Lenoir, G.M. and Mazoyer, S. (2002) The nonsense-mediated mRNA decay pathway triggers degradation of most *BRCA1* mRNAs bearing premature termination codons. *Hum. Mol. Genet.*, **11**, 2805–2814.
41. Ng, P.C. and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucl. Acids Res.*, **31**, 3812–3814.
42. Kouprina, N., Pavlicek, A., Mochida, G.H., Solomon, G., Gersch, W., Yoon, Y.H., Collura, R., Ruvolo, M., Barrett, J.C., Woods, C.G. *et al.* (2004) Accelerated evolution of the *ASPM* gene controlling brain size begins prior to human brain expansion. *PLoS Biol.*, **5**, 1–11.
43. Kapitonov, V.V., Pavlicek, A. and Jurka, J. (2004) Anthology of human repetitive DNA. In Meyers, R.A. (ed.) *Encyclopedia of Molecular Cell Biology and Molecular Medicine.* Wiley–VCH, Vol. 1, pp. 251–305.
44. Frolov, A., Prowse, A.H., Vanderveer, L., Bove, B., Wu, H. and Godwin, A.K. (2002) DNA array-based method for detection of large rearrangements in the *BRCA1* gene. *Genes Chromosomes Cancer*, **35**, 232–241.
45. Endo, T., Ikeo, K. and Gojobori T. (1996) Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.*, **13**, 685–690.
46. Kreitman, M. and Comeron, J.M. (1999) Coding sequence evolution. *Curr. Opin. Genet. Dev.*, **9**, 637–641.
47. Cantor, S., Drapkin, R., Zhang, F., Lin, Y., Han, J., Pamidi, S. and Livingston, D.M. (2004) The BRCA1-associated protein BACH1 is a DNA helicase targeted by clinically relevant inactivating mutations. *Proc. Natl Acad. Sci. USA*, **101**, 2357–2362.
48. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

49. Kondrashov, A.S., Sunyaev, S. and Kondrashov, F.A. (2002) Dobzhansky–Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA*, **99**, 14878–14883.
50. Gao, L. and Zhang, J. (2003) Why are some human disease-associated mutations fixed in mice? *Trends. Genet.*, **19**, 678–681.
51. Annab, L.A., Kouprina, N., Solomon, G., Cable, P.L., Hill, D.E., Barrett, J.C., Larionov, V. and Afshari, C.A. (2000) Isolation of a functional copy of the human *BRCA1* gene by transformation-associated recombination in yeast. *Gene*, **250**, 201–208.
52. Kouprina, N. and Larionov, V. (1999) Selective isolation of mammalian genes by TAR cloning. In Dracopoli, N.C., Haines, J.L., Korf, B.R., Moir, D.T., Morton, C.C., Seidman, C.E., Seidman, J.G. and Smith, D.R. (eds) *Current Protocols in Human Genetics.* John Wiley & Sons, Vol. 1, pp. 5.17.1–5.17.21.
53. Polushin, N., Malykh, A., Malykh, O., Zenkova, M., Chumakova, N., Vlassov, V. and Kozyavkin, S. (2001) 2′-modified oligonucleotides from methoxyoxalamido and succinimido precursors: synthesis, properties, and applications. *Nucleosides Nucleotides Nucl. Acids*, **4–7**, 507–511.
54. Bray, N. and Pachter, L. (2004) MAVID: Constrained Ancestral Alignment of Multiple Sequences. *Genome Res.*, **14**, 693–699.
55. Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.
56. Galtier, N., Gouy, M. and Gautier, C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.*, **12**, 543–548.
57. Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends. Genet.*, **16**, 276–277.
58. Jurka, J., Klonowski, P., Dagman, V. and Pelton, P. (1996) CENSOR-a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.*, **20**, 119–121.
59. Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.*, **16**, 418–420.
60. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucl. Acids Res.*, **27**, 573–580.
61. Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
62. Korber, B. (2000) HIV signature and sequence variation analysis. In Rodrigo, A.G. and Learn, G.H. (eds) *Computational Analysis of HIV Molecular Sequences.* Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 55–72.
63. Gonnet, G.H., Cohen, M.A. and Benner, S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1145.
64. Szabo, C., Masiello, A., Ryan, J.F. and Brody, L.C. (2000) The breast cancer information core: database design, structure, and scope. *Hum. Mutat.*, **16**, 123–131.
65. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
66. Page, R.D.M. (1996) TREEVIEW: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.*, **12**, 357–358.
67. Hohenstein, P. and Fodde, R. (2004) Of mice and (wo)men: genotype–phenotype correlations in *BRCA1*. *Hum. Mol. Genet.,* **12**, R271–R277.
68. Hohenstein, P. (2004) Tumor Suppressor Genes in Tumorigenesis. PhD thesis, Leiden University, The Netherlands.
69. Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.,* **15**, 568–573.