

## Understanding missense mutations in the BRCA1 gene: An evolutionary approach

Melissa A. Fleming, John D. Potter, Christina J. Ramirez, Gary K. Ostrander, and Elaine A. Ostrander

*PNAS* 2003;100;1151-1156; originally published online Jan 16, 2003;  
doi:10.1073/pnas.0237285100

**This information is current as of December 2006.**

|  |   |
|--|---|
| <b>Online Information &amp; Services</b> | High-resolution figures, a citation map, links to PubMed and Google Scholar, etc., can be found at:<br><a href="http://www.pnas.org/cgi/content/full/100/3/1151">www.pnas.org/cgi/content/full/100/3/1151</a>   |
| <b>Supplementary Material</b>            | Supplementary material can be found at:<br><a href="http://www.pnas.org/cgi/content/full/0237285100/DC1">www.pnas.org/cgi/content/full/0237285100/DC1</a>   |
| <b>References</b>                        | This article cites 44 articles, 22 of which you can access for free at:<br><a href="http://www.pnas.org/cgi/content/full/100/3/1151#BIBL">www.pnas.org/cgi/content/full/100/3/1151#BIBL</a><br><br>This article has been cited by other articles:<br><a href="http://www.pnas.org/cgi/content/full/100/3/1151#otherarticles">www.pnas.org/cgi/content/full/100/3/1151#otherarticles</a> |
| <b>E-mail Alerts</b>                     | Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or <a href="#">click here</a> .   |
| <b>Rights &amp; Permissions</b>          | To reproduce this article in part (figures, tables) or in entirety, see:<br><a href="http://www.pnas.org/misc/rightperm.shtml">www.pnas.org/misc/rightperm.shtml</a>  |
| <b>Reprints</b>                          | To order reprints, see:<br><a href="http://www.pnas.org/misc/reprints.shtml">www.pnas.org/misc/reprints.shtml</a>   |

Notes:

# Understanding missense mutations in the *BRCA1* gene: An evolutionary approach

Melissa A. Fleming\*, John D. Potter<sup>†</sup>, Christina J. Ramirez\*, Gary K. Ostrander<sup>‡</sup>, and Elaine A. Ostrander\*<sup>§</sup>

\*Divisions of Clinical Research and Human Biology, and <sup>†</sup>Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024; and <sup>‡</sup>Departments of Biology and Comparative Medicine, Johns Hopkins University, Baltimore, MD 21218

Communicated by Leland H. Hartwell, Fred Hutchinson Cancer Research Center, Seattle, WA, November 27, 2002 (received for review July 25, 2002)

**The role of missense changes in *BRCA1* in breast cancer susceptibility has been difficult to establish. We used comparative evolutionary methods to identify potential functionally important amino acid sites in exon 11 and missense changes likely to disrupt gene function, aligning sequences from 57 eutherian mammals and categorizing amino acid sites by degree of conservation. We used Bayesian phylogenetic analyses to determine relationships among orthologs and identify codons evolving under positive selection. Most conserved residues occur in a region with the highest concentration of protein-interacting domains. Rapidly evolving residues are concentrated in the RAD51-interacting domain, suggesting that selection is acting most strongly on the role of *BRCA1* in DNA repair. Investigation of the functional role of missense changes in breast-cancer susceptibility should focus on 38 missense changes in conserved and 3 in rapidly evolving regions of exon 11.**

breast cancer | exon 11 | gene evolution | missense change | *BRCA1*

Protein-truncating mutations distributed across *BRCA1* are associated with an increased cumulative lifetime risk of breast (60–80%) and ovarian (20–40%) cancer (reviewed in ref. 1). Known mutations in breast-cancer susceptibility genes have been collated in the Breast Cancer Information Core (BIC) database (2, 3). Nearly half the reported changes in *BRCA1* are frameshift mutations and thus expected to be disease associated (2). Most of the rest are missense changes; 323 of these have been reported in 1,735 individual entries. Disease-association status is known for only a fraction of these: in the RING finger domain (4) and the C-terminal region of the protein (5, 6). Case-control and family studies are underpowered to draw conclusions regarding the remainder; these are not highly penetrant alleles (7–9).

The *BRCA1* gene encodes a 1,863-aa protein with a single large region, exon 11, encoding some 60%. The gene is highly polymorphic, with many common single-base exon changes. Regions interacting with other proteins have been identified, but structural and biochemical properties of the protein remain largely unknown, making it difficult to predict the consequences of any single missense change (10). Available functional assays are time-consuming, expensive, and applicable only to C-terminal mutations (6, 11, 12). Predictions regarding missense changes can be strengthened by comparative evolutionary analysis to establish whether mutations cluster in conserved regions (13–15). Such analyses may be particularly helpful in identifying low-penetrance missense changes in functionally important regions. Phylogenetic approaches can also determine whether certain residues have evolved more rapidly than predicted by neutral theory (the ratio of the rate of nonsynonymous to synonymous substitution,  $\omega$ , >1), reflecting the action of positive (diversifying) selection (16, 17).

The ability to detect conserved (18–20) and rapidly evolving (21, 22) regions in *BRCA1* has been limited by the small number of cloned sequences available. Recently, portions of exon 11 have been used in phylogenetic studies to clarify the relationships among higher mammals (23–25). DNA sequences for exon 11, but not other regions of the gene, are available for representa-

tives of all 18 eutherian orders and a marsupial (metatherian). These diverse sequence data allowed the use of phylogenetic approaches to identify (i) regions of exon 11 conserved in species with mammary tissue (susceptible to breast cancer) and (ii) regions evolving under positive selection. We here rank known missense changes with a view to establishing priorities for functional and population studies.

## Methods

**Sequences.** *BRCA1* amino acid and nucleotide sequences were extracted from GenBank for 57 eutherian mammals, 1 marsupial, the chicken, and the frog (species and accession nos. are presented in Table 2, which is published as supporting information on the PNAS web site, www.pnas.org). Amino acid sequences were aligned using CLUSTALX 1.81 (26). The 57 mammalian nucleotide sequences were aligned to amino acid sequences by using CODONALIGN 1.0 (27) and were checked visually (see Figs. 3 and 4, which are published as supporting information on the PNAS web site).

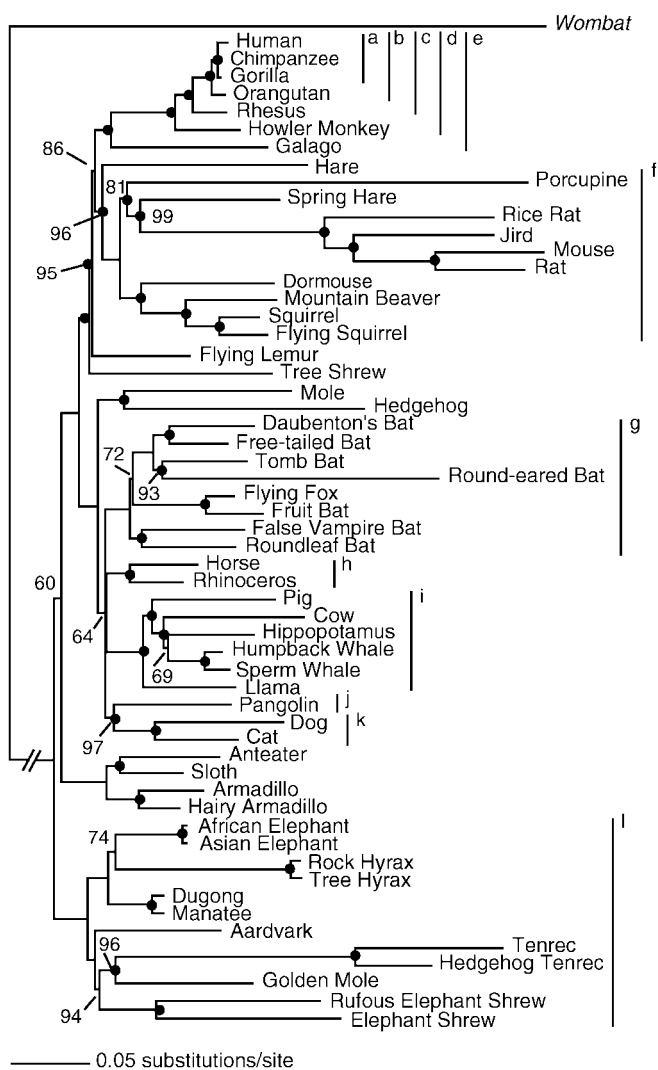
Analyses were based on sequences between human codons 282 and 1152 (76.4% of exon 11). Sequence data spanning this region are available for 53 eutherians and the marsupial, *Vombatus*. Codon numbers are based on the human sequence. Sequences start at codon 300 for *Lepus*, codon 316 for *Hystrix*, and codon 285 for *Erinaceus*, and end at codon 1072 for *Aplodontia*.

MRBAYES 2.01 (28) was used to construct a phylogenetic tree of *BRCA1* evolution by using nucleotide sequences of 57 eutherian mammals and *Vombatus* as an outgroup (Fig. 1). The Bayesian approach is similar to maximum likelihood (ML) in allowing the specification of complex models but is computationally much faster for large data sets (29). Rather than finding a single most likely tree, Bayesian analysis results in a set of equally likely trees. Support for particular branches can then be calculated by summing the posterior probabilities for a branch across these trees rather than through computation-intensive nonparametric bootstrapping.

We specified a general time-reversible model of sequence evolution, allowing for six rates of nucleotide substitution and differing base frequencies (30). These parameters were estimated from the data. Among-site rate variation was calculated by an approximation to the continuous gamma model using four rate categories (31). Rates for each codon position were estimated separately; adjacent sites were assumed to have correlated rates. The analysis was started from random trees for four simultaneous, independent chains, uniform prior probability distributions for tree topologies and the rate matrix, and a Dirichlet prior of 4.0 for base frequencies (28). The analysis was run for 250,000 and 500,000 generations; every hundredth tree was saved. Stable likelihood estimates were achieved after  $\approx$ 50,000 generations, so the first 20% of saved trees were discarded as “burn-in” when generating a consensus of equally

Abbreviations: AS, ancestral sequences; BIC, Breast Cancer Information Core.

<sup>§</sup>To whom correspondence should be addressed at: Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, D4-100, Seattle, WA 98109-1024. E-mail: eostrand@fhrc.org.



**Fig. 1.** A Bayesian phylogenetic tree of 2,622 bp of *BRCA1* exon 11. The numbers adjacent to internal nodes represent the posterior probability that a clade is correct based on a consensus of 4,000 trees with roughly equivalent likelihoods; unlabeled nodes represent posterior probabilities of 1.0. Nodes supported with posterior probabilities of <0.5 are considered not resolved and have been collapsed (resulting in polytomies; e.g., human, gorilla, and chimpanzee). Filled circles indicate nodes where ancestral sequences were calculated. Taxonomic categories greater than species level for primates and other taxa referred to in the text are identified by lines and letters: a, African hominids; b, great apes; c, catarrhine primates (Old World monkeys and apes); d, haplorhine primates (monkeys and apes); e, Primates; f, Rodentia; g, Chiroptera; h, Perissodactyla; i, Cetartiodactyla; j, Pholidota; k, Carnivora; and l, Afrotheria.

likely trees with estimates of posterior probabilities for each clade.

**Conserved Regions.** Homologous amino acid sites were categorized as (i) having a single residue in all sequences compared (fixed), (ii) including conservative substitutions only (conservative sites), or (iii) including nonconservative substitutions or gaps (nonconservative sites). We identified “conserved regions” of the gene by using a sliding window of 5 aa. These were identified as portions of the alignment that began and ended with fixed or conservative sites and that were comprised of  $\geq 80\%$  of such sites. One-sample run tests [two-tailed (32)] were used to determine whether fixed or conservative residues were associated.

To reduce the impact of sequencing errors and species-specific polymorphisms, we compared levels of amino acid conservation both among the species themselves and among sequences derived for their immediate ancestors (nodes just interior to terminal nodes; Fig. 1). These ancestral sequences were calculated by Bayesian phylogenetic analysis (28, 33) in which clades of sister taxa were constrained on the 250,000-generation consensus tree, and the most probable sequence was generated at each node. Additional nodes were constrained to retain the topology of the original tree (Fig. 1).

**Regions Evolving Under Positive Selection.** Positively selected sites were identified using codons as sites and allowing  $\omega$  to vary among sites (16). To identify residues and regions that may have been selected for new or modified functions specific to humans or primates, we identified amino acid substitutions that arose independently in the primate lineage leading to humans and classified sites according to their degree of conservation in non-primates. Substitutions at sites that were fixed or conservative in non-primates were deemed most likely to affect function.

**Missense Changes.** To determine which missense changes were most likely to affect function in humans, we compared the distribution of conserved sites and sites evolving under positive selection in the 57 eutherian species and in their immediate ancestors with that of missense changes reported in the BIC database. Nonconservative substitutions at fixed or conservative sites and conservative substitutions at fixed sites were of most interest. Predictions were derived using the Gonnet PAM 250 matrix (34) to identify missense changes involving nonconservative substitutions and the extent to which sites in the *BRCA1* sequence were conserved across ancestral sequences (AS method). We compared these predictions with those derived by the program SIFT (35), which estimates the degree of conservation as a continuous variable derived from information theory by using the number of amino acids observed at a homologous position. The associations of missense changes (conservative or nonconservative) with fixed or conservative amino acid sites and with particular portions of *BRCA1* were tested using  $\chi^2$  with correction for continuity (32).

## Results

**Phylogeny of *BRCA1*.** Variation in sequence length among mammals resulted in an alignment of 940 codons for the 57 eutherian mammals and *Vombatus*. Insertions in a variety of taxa were removed from the analysis because they are phylogenetically uninformative (Table 2). Three sites with amino acid deletions in the primate lineage leading to humans were retained, resulting in an 874-codon alignment between human codons 282 and 1152.

The topology of the phylogenetic tree of *BRCA1* (Fig. 1) is well supported; all but 7 of 54 (13%) clades resolved with posterior probabilities  $>0.90$ . The relationships of the African hominids (human, chimpanzee, and gorilla) and Chiroptera, Perissodactyla/Cetartiodactyla, and Pholidota/Carnivora were not resolved.

**Conserved Regions.** Only 24 (2.8%) of the 871 human amino acid residues are fixed among eutherian mammals, and another 75 (8.6%) are conservative (Fig. 2a). Fixed residues are not randomly distributed across exon 11; most are adjacent to other fixed residues ( $z = 5.56$ ,  $P < 0.001$ ) or conservative sites ( $z = 8.53$ ,  $P < 0.001$ ). The majority of fixed residues (70.8%) were identified across codons 282–554 ( $\chi^2 = 16.81$ ,  $df = 1$ ,  $P < 0.001$ ), an interval that includes putative interacting domains for at least 15 different proteins or complexes (reviewed in ref. 36; Fig. 2); 5 conserved regions (amino acid identity  $\geq 80\%$ ), 5–12 residues in length, were identified here. Another three fixed residues



occur in the two conserved regions encompassed by the RAD51-interacting domain (codons 758-1064). An eighth conserved region includes a single fixed 3' residue in a region of unknown function. Twenty-one (87.5%) of the fixed residues are also fixed in *Vombatus*, 20 (83.3%) in *Gallus*, and 14 (58.3%) in *Xenopus*. Amino acid substitutions occur at fixed residues in two conserved regions in *Vombatus* (regions 2 and 5), three in *Gallus* (regions 1, 3, and 8), and all but region 7 in *Xenopus* (Fig. 2a). All conserved regions consisted of >70% ( $79.8 \pm 1.7\%$ ) fixed and conservative sites when *Gallus* and *Xenopus* sequences were included in the comparison.

Pairwise comparisons of fixed amino acid conservation between humans and other eutherian species reveal high levels of amino acid identity (Table 2). Between humans and other primates, pairwise conservation is  $90.6 \pm 4.0\%$  on average, and conservation remains high between humans and other eutherian orders ( $69.8 \pm 1.6\%$  on average). Conservation is lower between humans and *Vombatus* (41.5%) and between humans and *Gallus* or *Xenopus* (25.4% and 19.5%, respectively). Only small portions of the latter two sequences could be readily aligned to the mammal sequences; the *Xenopus* sequence was considerably shorter than others.

The discrepancy between simultaneous and pairwise estimates of amino acid conservation in eutherians ( $2.8$  vs.  $69.8 \pm 1.6\%$  on average) is due in part to the large number of sequences in the comparison, each potentially contributing one or more non-disease-associated polymorphisms and sequencing errors. We maximized our ability to identify functionally important regions by deriving ancestral sequences for each pair of sister taxa (Fig. 1) and comparing these “ancestors” rather than terminal taxa. This approach reveals 167 (19.1%) fixed (Fig. 2a) and 274 (31.5%) conservative residues.

**Regions Evolving Under Positive Selection.** Three sites had posterior probabilities of 0.5 or greater of being under positive selection in eutherian mammals: codons 801, 886, and 890 (Fig. 2b). All are included in the RAD51-interacting domain and show evolutionary change in the primate lineage leading to humans: codon 890 diverged in humans; the other two diverged in the haplorhine primates (see Table 3, which is published as supporting information on the PNAS web site). These sites also showed evidence of evolution in other eutherian lineages.

Since the divergence of primates from non-primates, 132 amino acid sites in human *BRCA1* have evolved. There is a trend toward more substitutions in the primate lineage leading to humans in the RAD51-interaction domain than elsewhere ( $\chi^2 = 3.59$ , 1 df,  $P = 0.08$ ). SIFT (30) identified three residues in the RAD51-interacting domain that evolved in the primate lineage as likely to affect function: G813 (also identified by the AS method) evolved in African hominids from glutamic acid; N822 and P871 evolved in humans from threonine and leucine, respectively.

**Missense Changes.** In BIC, 139 missense changes are reported at 126 sites in *BRCA1* exon 11. These are randomly distributed across fixed or conservative sites in the 57 eutherian mammals ( $\chi^2 = 0.22$ , 1 df,  $P > 0.80$ ) and in their immediate ancestors ( $\chi^2 = 0.12$ , 1 df,  $P > 0.90$ ). There is no difference in the distribution of conservative and nonconservative missense changes relative to fixed and conservative versus nonconservative sites in either eutherian mammals ( $\chi^2 = 0.52$ , 1 df,  $P > 0.90$ ) or ancestral sequences ( $\chi^2 = 0.12$ , 1 df,  $P > 0.70$ ).

Based on the level of conservation observed, we identified 38 of the 139 missense changes as likely to affect protein function (Table 1, Fig. 2a). Three of these changes affect residues that are fixed in all 57 eutherian mammals (F461L, G462R, and P514R) and another three are nonconservative substitutions affecting conservative sites (E765G, R866C, and E1060A). When ances-

tral sequences are compared, 32 additional changes occur at fixed sites or are nonconservative changes at conservative sites (Table 1). SIFT predicted that 70 changes would affect protein function, including 36 of these 38. Two changes predicted to disrupt function by the AS method had SIFT scores of 0.05 and 0.08 (H476R and M1137T, respectively) and were predicted to be tolerated.

In addition to the 38 changes, another 3 missense changes are predicted to affect function because they involve residues that are rapidly evolving or have recently evolved in humans at sites that are fixed or conservative in other lineages (Fig. 2b). Two of these occur at a site under positive-selection, G890V and G890R. The third, S708Y, affects a conservative site in non-primate ancestral sequences that evolved a nonconservative substitution in haplorhine primates. Two other changes, K739I and R841W, affect conservative sites that had been fixed in non-primate ancestral sequences (Q in both) but experienced conservative substitutions during human evolution and were already included in the 38 changes above.

To address the issue of potential rates of false positives and negatives when using the AS method, we applied it to a well studied protein with missense changes known to affect function: beta-globin. We used 15 diverse eutherian sequences from GenBank and 10 deleterious mutations from ref. 37 (see Tables 4 and 5 and Figs. 5 and 6, which are published as supporting information on the PNAS web site). The AS method correctly predicted that all 10 mutations would affect function: 6 nonconservative and 2 conservative changes at fixed sites, and 2 nonconservative changes at conservative sites. In contrast, simply aligning the 15 eutherian sequences would have resulted in one false negative because of a nonconservative change in the hamster that was not retained in the ancestral sequence of the hamster, rat, and mouse. SIFT also correctly assigned the 10 changes but with “low confidence” because the median sequence conservation was too high.

Using the available sequence from human, dog, rat, mouse, *Gallus*, and *Xenopus* (see Fig. 7 and Table 6, which are published as supporting information on the PNAS web site), we performed a similar comparison for 22 missense changes of known effect in *BRCA1* (RING domain and the two 3' BRCT domains according to the BIC and ref. 6). We did not apply the AS method because there were so few sequences from such divergent species. The 16 changes known to be detrimental were at 10 fixed sites in the six taxa and were correctly predicted to affect function. However, of six changes with no functional effect, two were correctly predicted to be tolerated (both conservative changes at a conservative site due to a single difference in either *Gallus* or *Xenopus*) but four occurred at fixed sites and were incorrectly predicted to be detrimental. Thus, the false-positive rate for the *BRCA1* RING and BRCT domains was 20% (4/20), and there were no false negatives.

## Discussion

**Phylogeny of *BRCA1*.** The phylogenetic tree for *BRCA1* (Fig. 1) is largely consistent with recent eutherian phylogenies derived from multiple genes and by both maximum likelihood (ML) and Bayesian methods of inference (25, 38–40). All phylogenies maintain the same sister taxa relationships between species, support four major clades of eutherian mammals, and refute a basal position for rodents in the Eutheria. Clades observed to be poorly resolved (posterior probabilities <0.70) or unresolved in this study are also poorly supported in other studies, including relationships among the African hominids (25), among the orders Chiroptera, Cetartiodactyla, Perrisodactyla, and Carnivora + Pholidota (25, 38, 40), and among the four major clades themselves (25, 38, 40).

**Table 1. The 38 (of 139) missense changes reported in BIC that are predicted to affect function because they occur at sites that are fixed or conservative in eutherian ancestral sequences**

| Type of missense change and amino acid site   | Missense change*  | SIFT score† |
|---|-------------------|-------------|
| Nonconservative changes at fixed sites        | T374I             | <0.01       |
|   | I456T             | 0.03        |
|   | <b>G462R</b>      | <0.01       |
|   | R507I             | 0.01        |
|   | <b>P514R</b>      | <0.01       |
|   | S741F             | 0.04        |
|   | <b>E765G</b>      | <0.01       |
|   | S784L             | 0.02        |
|   | <b>R866C</b> (10) | <0.01       |
|   | <b>E1060A</b>     | <0.01       |
|   | R1074G            | 0.03        |
|   | R1076T            | <0.01       |
|   | P1136R            | 0.01        |
|   | R331S (3)         | <0.01       |
| Nonconservative changes at conservative sites | R466G             | <0.05       |
|   | R504C             | 0.03        |
|   | R612G             | 0.01        |
|   | D695Y             | 0.03        |
|   | K739I             | 0.01        |
|   | N810Y (4)         | <0.01       |
|   | D825Y             | <0.05       |
|   | R841W (60)        | <0.01       |
|   | M1137T            | TOL (0.08)  |
|   | Q284R             | 0.01        |
|   | E300D             | 0.01        |
| Conservative changes at fixed sites           | I379M (10)        | <0.01       |
|   | <b>F461L</b> (2)  | <0.01       |
|   | H476R (3)         | TOL (0.05)  |
|   | D522N             | 0.01        |
|   | E597K (4)         | <0.01       |
|   | N609S             | 0.02        |
|   | E624K             | 0.01        |
|   | S632N             | 0.04        |
|   | S741C             | <0.05       |
|   | A807S             | 0.01        |
|   | S1027N            | 0.03        |
|   | K1109N (2)        | 0.01        |
| Q1144H  | 0.01              |             |

\*Missense changes affecting sites that were fixed in all 57 eutherian sequences and nonconservative changes affecting sites that were conservative in all 57 are in bold. Numbers in parentheses indicate the number of times (>1) that the missense change has been reported in the BIC.

†SIFT scores of <0.05 indicate missense changes that are likely to affect function. TOL indicates substitutions predicted by SIFT to be “tolerated” and is followed by a SIFT score when contrary to predictions based on eutherian ancestral sequences.

**Conserved Regions.** We identified eight regions in BRCA1 where amino acid conservation across eutherian mammals was  $\geq 80\%$  (Fig. 2). By comparing the chicken *BRCA1* sequence with that of primates, dog, rat, and mouse, Orelli *et al.* (20) identified “conserved sequence motifs” by inspection that correspond to regions 3–8 described here (18). In these six regions, we determined that 80.5% of fixed and conservative sites are also maintained in *Gallus* and *Xenopus*, suggesting strong conservation of function across three classes. Regions 1 and 2 have not been previously described as conserved, but 75% of fixed and conservative sites here are also found in *Gallus* and *Xenopus*. Thus, none of these conserved regions is unique to mammals.

Five of the conserved regions are at the 5' end of exon 11 (codons 282–554), which includes putative interacting sites for several proteins thought to be involved in transcription (Fig. 2b; reviewed in ref. 36). Conserved regions 6 and 7 are in the

RAD51-interacting domain (codons 758–1064), which has a documented role in DNA double-stranded break (DSB) repair. The RAD50-interacting site (codons 341–748), also implicated in DSB repair, includes conserved regions 3–5. To our knowledge, conserved region 8 does not correspond to any functional region of BRCA1, although five of the six fixed or conservative residues in this region are also conserved in *Gallus* and *Xenopus*.

**Regions Evolving Under Positive Selection.** We detected positive selection at three sites in the RAD51-interaction domain either in humans or the primate lineage leading to humans. Change occurred after the divergence of humans from African great apes (codon 890) and of the monkeys and apes from other primates (codons 801 and 886). Evolution at these three sites has also occurred in other taxa (e.g., Rodentia and Afrotheria), suggesting that the selection pressures involved are not unique to primates. SIFT predicted functional consequences for another three substitutions in RAD51 (at codons 813, 822, and 871) that evolved in humans and other primates.

The RAD51-interacting domain has been previously identified as a putative site of positive selection. Hurst and Pal (22) used pairwise comparisons to detect differences between human and dog versus mouse and rat and found a high  $\omega$  at codons 944 and 949. They note that this is consistent with the finding by Huttley *et al.* (21) of a peak in the nonsynonymous substitution rate around codon 900.

Current methods for detecting positive selection at particular amino acid sites are best suited for detecting sites under continual selection pressure in multiple lineages (17, 41, 42). Studies in which amino acid sites have been identified as evolving under positive selection include the rapid evolution of virus-coat proteins (16) and proteins involved in fertilization in response to sexual selection (43). The physical interaction between amino acids encoded by *BRCA1* exon 11 and the RAD51 gene, although apparently critical for an appropriate response to DNA damage (44), is not well understood; most evidence points to the interaction being indirect (44–46). Why diversifying selection should be acting on the RAD51-interacting domain across a number of lineages, including humans, is not clear, but it suggests that this region is undergoing adaptive evolution and must be functionally important.

**Missense Changes.** We identified 38 of 139 BIC missense changes as affecting potentially functional residues in exon 11 by using the AS method. The SIFT program identified 36 of these 38 and an additional 34. SIFT identified more changes because it gives equal weight to substitutions in all taxa. In contrast, the AS method assumes (i) that substitutions not shared by sister taxa are likely to be sequencing errors or polymorphisms and (ii) that sites that have maintained nonconservative substitutions in even one pair of sister taxa are unlikely to affect function. The AS method correctly predicted the functional effects of >85% of the known detrimental missense changes in beta-globin and the RING and BRCT domains of *BRCA1*, confirming that this approach can identify high-priority mutations for further study.

Most of these mutations have been reported only once, so the contribution to disease susceptibility is not evaluable. The large number of changes predicted to disrupt function in exon 11 suggests that multiple pathways, perhaps associated with the multiple interacting proteins, influence cancer susceptibility.

Another three missense changes considered high priority for further study affect sites under positive selection or recent evolution. Two nonconservative missense changes occur at a positively selected site (G890V and G890R). Three more occur at sites that evolved substitutions during the evolution of humans but were fixed or conservative in non-primate ancestral sequences; two of these (K739I and R841W) affect sites included among the 38 identified by the AS method. R841W is the only

possibly functional missense change with sufficiently high prevalence (60 reports in BIC) to allow evaluation of its role in cancer susceptibility (47–50).

## Conclusions

We identified eight regions of very high amino acid conservation in eutherian mammals that are also well conserved in the single representatives of Aves and Amphibia in which *BRCAl* has been cloned. Five of these regions, at the 5' end of exon 11, overlap putative interaction domains for multiple proteins. Two other conserved regions are located at the 3' end, which is known to be involved in double-stranded break (DSB) repair. Possible

interacting proteins are known for seven of these regions, but the function of the eighth region is not known; its conservation even in Amphibia suggests that it is important. Evidence of positive selection acting on residues in the RAD51-interaction domain, both in humans and in other eutherian lineages, suggests that this region is undergoing adaptive evolution. Finally, consistent with our initial aim, we identified 41 missense changes likely to influence function and thereby contribute to cancer susceptibility. These sites are high priority for functional analyses.

We thank Steve Henikoff, Piri Welch, and Kathi Malone for helpful discussions. This work was supported by National Institutes of Health Grants K05 CA-90754-01 (to E.A.O.) and U24 CA-78164 (to J.D.P.).

1. Lee, W. H. & Boyer, T. G. (2001) *Lancet* **358**, Suppl, S5.
2. Shen, D. & Vadgama, J. V. (1999) *Oncol. Res.* **11**, 63–69.
3. Szabo, C., Masiello, A., Ryan, J. F. & Brody, L. C. (2000) *Hum. Mutat.* **16**, 123–131.
4. Brzovic, P. S., Meza, J. E., King, M. C. & Klevit, R. E. (2001) *J. Biol. Chem.* **276**, 41399–41406.
5. Monteiro, A. N., August, A. & Hanafusa, H. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 13595–13599.
6. Vallon-Christersson, J., Cayanan, C., Haraldsson, K., Loman, N., Bergthorsson, J. T., Brondum-Nielsen, K., Gerdes, A. M., Moller, P., Kristofferson, U., Olsson, H., et al. (2001) *Hum. Mol. Genet.* **10**, 353–360.
7. Malone, K. E., Daling, J. R., Neal, C., Suter, N. M., O'Brien, C., Cushing-Haugen, K., Jonasdottir, T. J., Thompson, J. D. & Ostrander, E. A. (2000) *Cancer* **88**, 1393–1402.
8. Newman, B., Mu, H., Butler, L. M., Millikan, R. C., Moorman, P. G. & King, M. C. (1998) *J. Am. Med. Assoc.* **279**, 915–921.
9. Loman, N., Johannsson, O., Kristofferson, U., Olsson, H. & Borg, A. (2001) *J. Natl. Cancer. Inst.* **93**, 1215–1223.
10. Collins, F. S. (1996) *N. Engl. J. Med.* **334**, 186–188.
11. Hayes, F., Cayanan, C., Barilla, D. & Monteiro, A. N. (2000) *Cancer Res.* **60**, 2411–2418.
12. Humphrey, J. S., Salim, A., Erdos, M. R., Collins, F. S., Brody, L. C. & Klausner, R. D. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 5820–5825.
13. Ganesh, S., Agarwala, K. L., Amano, K., Suzuki, T., Delgado-Escueta, A. V. & Yamakawa, K. (2001) *Biochem. Biophys. Res. Commun.* **283**, 1046–1053.
14. Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. (2001) *Nat. Genet.* **28**, 281–285.
15. Koeberl, D. D., Bottema, C. D., Ketterling, R. P., Bridge, P. J., Lillicrap, D. P. & Sommer, S. S. (1990) *Am. J. Hum. Genet.* **47**, 202–217.
16. Nielsen, R. & Yang, Z. (1998) *Genetics* **148**, 929–936.
17. Yang, Z. & Bielawski, J. P. (2000) *Trends Ecol. Evol.* **15**, 496–503.
18. Abel, K. J., Xu, J., Yin, G. Y., Lyons, R. H., Meisler, M. H. & Weber, B. L. (1995) *Hum. Mol. Genet.* **4**, 2265–2273.
19. Szabo, C. I., Wagner, L. A., Francisco, L. V., Roach, J. C., Argonza, R., King, M. C. & Ostrander, E. A. (1996) *Hum. Mol. Genet.* **5**, 1289–1298.
20. Orelli, B. J., Logsdon, J. M., Jr., & Bishop, D. K. (2001) *Oncogene* **20**, 4433–4438.
21. Huttley, G. A., Eastal, S., Southey, M. C., Tesoriero, A., Giles, G. G., McCredie, M. R., Hopper, J. L. & Venter, D. J. (2000) *Nat. Genet.* **25**, 410–413.
22. Hurst, L. D. & Pal, C. (2001) *Trends Genet.* **17**, 62–65.
23. Teeling, E. C., Scally, M., Kao, D. J., Romagnoli, M. L., Springer, M. S. & Stanhope, M. J. (2000) *Nature* **403**, 188–192.
24. Adkins, R. M., Gelke, E. L., Rowe, D. & Honeycutt, R. L. (2001) *Mol. Biol. Evol.* **18**, 777–791.
25. Madsen, O., Scally, M., Douady, C. J., Kao, D. J., DeBry, R. W., Adkins, R., Amrine, H. M., Stanhope, M. J., de Jong, W. W. & Springer, M. S. (2001) *Nature* **409**, 610–614.
26. Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. (1998) *Trends Biochem. Sci.* **23**, 403–405.
27. Hall, B. G. (2000) CODONALIGN (University of Rochester, Rochester, NY).
28. Huelsenbeck, J. P. (2000) MRBAYES (University of Rochester, Rochester, NY).
29. Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. (2001) *Science* **294**, 2310–2314.
30. Swofford, D. L., Olsen, G. J., Wadell, P. J. & Hillis, D. M. (1996) in *Molecular Systematics*, eds Hillis, D. M., Moritz, C. & Mable, B. K. (Sinauer, Sunderland, MA), pp. 407–514.
31. Yang, Z. (1994) *J. Mol. Evol.* **39**, 306–314.
32. Siegel, S. (1956) *Nonparametric Statistics Behavioral Sciences* (McGraw-Hill, New York).
33. Huelsenbeck, J. P. & Bollback, J. P. (2001) *Syst. Biol.* **50**, 351–366.
34. Gonnert, G. H., Cohen, M. A. & Benner, S. A. (1992) *Science* **256**, 1443–1445.
35. Ng, P. C. & Henikoff, S. (2001) *Genome Res.* **11**, 863–874.
36. Welch, P. L. & King, M. C. (2001) *Hum. Mol. Genet.* **10**, 705–713.
37. Thein, S. L. (1999) *Br. J. Haematol.* **107**, 12–21.
38. Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W. & Springer, M. S. (2001) *Science* **294**, 2348–2351.
39. Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A. & O'Brien, S. J. (2001) *Nature* **409**, 614–618.
40. Eizirik, E., Murphy, W. J. & O'Brien, S. J. (2001) *J. Hered.* **92**, 212–219.
41. Holmes, E. C., Zhang, L. Q., Simmonds, P., Ludlam, C. A. & Brown, A. J. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4835–4839.
42. Nielsen, R. & Huelsenbeck, J. P. (2002) *Pac. Symp. Biocomput.* 576–588.
43. Swanson, W. J. & Vacquier, V. D. (2002) *Nat. Rev. Genet.* **3**, 137–144.
44. Huber, L. J., Yang, T. W., Sarkisian, C. J., Master, S. R., Deng, C. X. & Chodosh, L. A. (2001) *Mol. Cell. Biol.* **21**, 4005–4015.
45. Venkitaraman, A. R. (2001) *J. Cell Sci.* **114**, 3591–3598.
46. Venkitaraman, A. R. (2002) *Cell* **108**, 171–182.
47. Barker, D. F., Almeida, E. R., Casey, G., Fain, P. R., Liao, S. Y., Masunaka, I., Noble, B., Kurosaki, T. & Anton-Culver, H. (1996) *Genet. Epidemiol.* **13**, 595–604.
48. Petersen, G. M., Parmigiani, G. & Thomas, D. (1998) *Am. J. Hum. Genet.* **62**, 1516–1524.
49. Durocher, F., Shattuck-Eidens, D., McClure, M., Labrie, F., Skolnick, M. H., Goldgar, D. E. & Simard, J. (1996) *Hum. Mol. Genet.* **5**, 835–842.
50. Dong, J., Chang-Claude, J., Wu, Y., Schumacher, V., Debatin, I., Tonin, P. & Royer-Pokora, B. (1998) *Hum. Genet.* **103**, 154–161.