

# Marsupial BRCA1: conserved regions in mammals and the potential effect of missense changes

Christina J Ramirez<sup>1,2</sup>, Melissa A Fleming<sup>1</sup>, John D Potter<sup>3</sup>, Gary K Ostrander<sup>4</sup> and Elaine A Ostrander<sup>\*1</sup>

<sup>1</sup>Divisions of Clinical Research and Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, USA; <sup>2</sup>Department of Molecular and Cellular Biology, University of Washington, Seattle, WA, 98195, USA; <sup>3</sup>Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, USA; <sup>4</sup>Departments of Biology and Comparative Medicine, Johns Hopkins University, Baltimore, MD 21218, USA

More than half of the reported missense changes in the breast cancer susceptibility protein BRCA1 occur in exon 11, but none has been clearly identified as disease associated and only 28 are designated 'probable' neutral polymorphisms. Previously, in a comparison of sequences from 57 eutherian mammal species, we found seven 'highly conserved regions' between amino acids 282 and 1103, and identified 38 missense changes as likely to disrupt gene function. These conserved regions were also present in birds and amphibians and included only six of the mutations predicted to affect function. In this new analysis, we hypothesized that using 37 ancestral sequences derived from the 57 GenBank sequences and including eight marsupial sequences would allow us to identify regions unique to mammals and refine our predictions of disease-associated missense changes. We identified 13 conserved regions, three of which appear to be unique to mammals, and 21 likely disease-associated missense changes, 11 of which occur in conserved regions. Seven regions identified in this analysis, including the three found only in mammalian sequences, and nine missense changes predicted to affect function are in the putative STAT1-interaction domain, suggesting that the role of STAT1 in immune response is important to mammary function. The reduction in the number of missense changes predicted to be disease associated and the identification of conserved regions specific to mammals can facilitate the further study of the role of missense changes in BRCA1-associated breast cancers.

*Oncogene* (2004) 23, 1780–1788. doi:10.1038/sj.onc.1207292

**Keywords:** breast cancer; comparative analysis; exon 11; gene evolution; missense; phylogenetics

## Introduction

Increased lifetime risks of breast (60–80%) and ovarian cancers (20–40%) are associated with mutations in the breast cancer susceptibility gene *BRCA1* (reviewed in Lee and Boyer, 2001). The Breast Cancer Information Core (BIC) is a central database of *BRCA1* and *BRCA2* mutations submitted voluntarily by researchers worldwide. Mutations are compiled from linkage, population-based, case-control, and hospital-based studies (Shen and Vadgama, 1999; Szabo *et al.*, 2000). Of the 1261 independent mutations reported in the BIC, only 57% are known to affect function, 80% of which are protein-truncation mutations; 28% are missense changes, but only 3.1% of these are confirmed as either disease associated or neutral polymorphisms. The effects of other missense changes have been difficult to characterize in large part because the function of the 1863 amino-acid (a.a.) BRCA1 protein is not yet fully understood (Venkitaraman, 2001). Functional and structural studies have focused on the N- and C-terminal domains, which include RING finger and transcription-activation domains, respectively. Over 90% of missense changes known to be disease associated occur in these domains (Monteiro *et al.*, 1996; Brzovic *et al.*, 2001; Vallon-Christersson *et al.*, 2001), but these regions encompass only 13% of the entire protein.

Exon 11 comprises over 60% of BRCA1 and includes a number of putative protein-interaction domains (Welsh and King, 2001). Research using the BRCA1- $\Delta$ 11 splice variant demonstrates that exon 11 is essential for BRCA1 function in genome stability, for tumor suppression, and at the G2-M cell cycle checkpoint (Huber *et al.*, 2001). Although more than half of the reported missense changes occur in exon 11, none has been clearly identified as disease associated and only 28 are designated probable neutral polymorphisms (i.e., unlikely to be disease associated). The difficulty in characterizing exon 11 mutations can be attributed to both the absence of functional assays for this region and the insufficient power of most case-control and family studies that underlie the BIC database (Castilla *et al.*, 1994; Friedman *et al.*, 1994; Simard *et al.*, 1994; Durocher *et al.*, 1996; Newman *et al.*, 1998; Malone *et al.*, 2000).

\*Correspondence: EA Ostrander, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., D4-100, Seattle, WA 98109-1024, USA; E-mail: eostrand@fhcrc.org

Received 6 August 2003; revised 14 October 2003; accepted 15 October 2003

One method of identifying functionally important regions of a protein is to locate highly conserved regions through a comparative analysis of homologous protein sequences from multiple organisms (see, for example, Koeberl *et al.*, 1990; Brown *et al.*, 2001; Ganesh *et al.*, 2001; Chapman *et al.*, 2003). Previously, we (Fleming *et al.*, 2003) performed a comparative analysis of exon 11 in 57 divergent eutherian mammal species whose sequences are available from GenBank. We found eight ‘highly conserved regions’ between amino acids 282 and 1152 defined as having greater than 80% amino-acid identity. We also identified 41 missense changes as likely to disrupt gene function because they (1) affect fixed sites; (2) are nonconservative substitutions at conservative sites; or (3) affect rapidly evolving or recently evolved residues in humans.

The occurrence of breast cancer has been reported in mammals besides humans, including primates, rodents, and carnivores (Hamilton, 1974; Casey *et al.*, 1979; Misdorp, 1996; Vail and MacEwen, 2000). The correlation of BRCA1 mutations with breast cancer in humans suggests that these mutations may disrupt functions unique to mammary tissue and thus, to mammals in general. However, our comparison of eutherian mammals did not identify any conserved regions that were not also conserved in birds and amphibians. To identify missense changes predicted to affect function, we minimized the loss of fixed and conservative sites due to species-specific neutral polymorphisms and sequencing errors by comparing 37 ‘ancestral sequences’, that is, sequences derived from the immediate ancestors of the 57 eutherian sequences. However, we did not use the ancestral sequences to identify the highly conserved regions, which could have prevented us from detecting other potentially functional regions.

The goal of this study was to determine whether there are other conserved regions of BRCA1 exon 11 specific to mammals. We performed comparative analyses of ancestral eutherian sequences reported in Fleming *et al.* (2003) and sequences from five orders of marsupial

mammals. The inclusion of marsupials makes the analyses more comprehensive of mammalian diversity, which should also refine our predictions of which missense changes are most likely to be disease associated in humans. Marsupials are similar to eutherians in terms of mammary structure and development (Findlay and Renfree, 1984; Knight, 1984; Tyndale-Biscoe *et al.*, 1984) and susceptibility to breast cancer (Straube and Callinan, 1980; Canfield *et al.*, 1990a, b), but the marsupial superorder of mammals diverged from eutherians much earlier than most eutherian orders diverged from one another (Kumar and Hedges, 1998; Table 1). Including the GenBank data four new marsupial sequences, and the eutherian ancestral sequences that we generated previously, this analysis encompasses 23 of 26 mammal orders and 130–170 million years of BRCA1 evolution.

## Results

### *Phylogeny of marsupial BRCA1 exon 11*

Marsupial sequences varied from 236 to 816 codons in length (Table 2) and extended over 843 amino-acid sites that correspond to human BRCA1 amino acids 282–1103 (~76% of exon 11). The positions of the novel marsupial sequences on the phylogenetic tree with

**Table 1** Vertebrate clades used in this analysis and the dates of their last common ancestor with humans

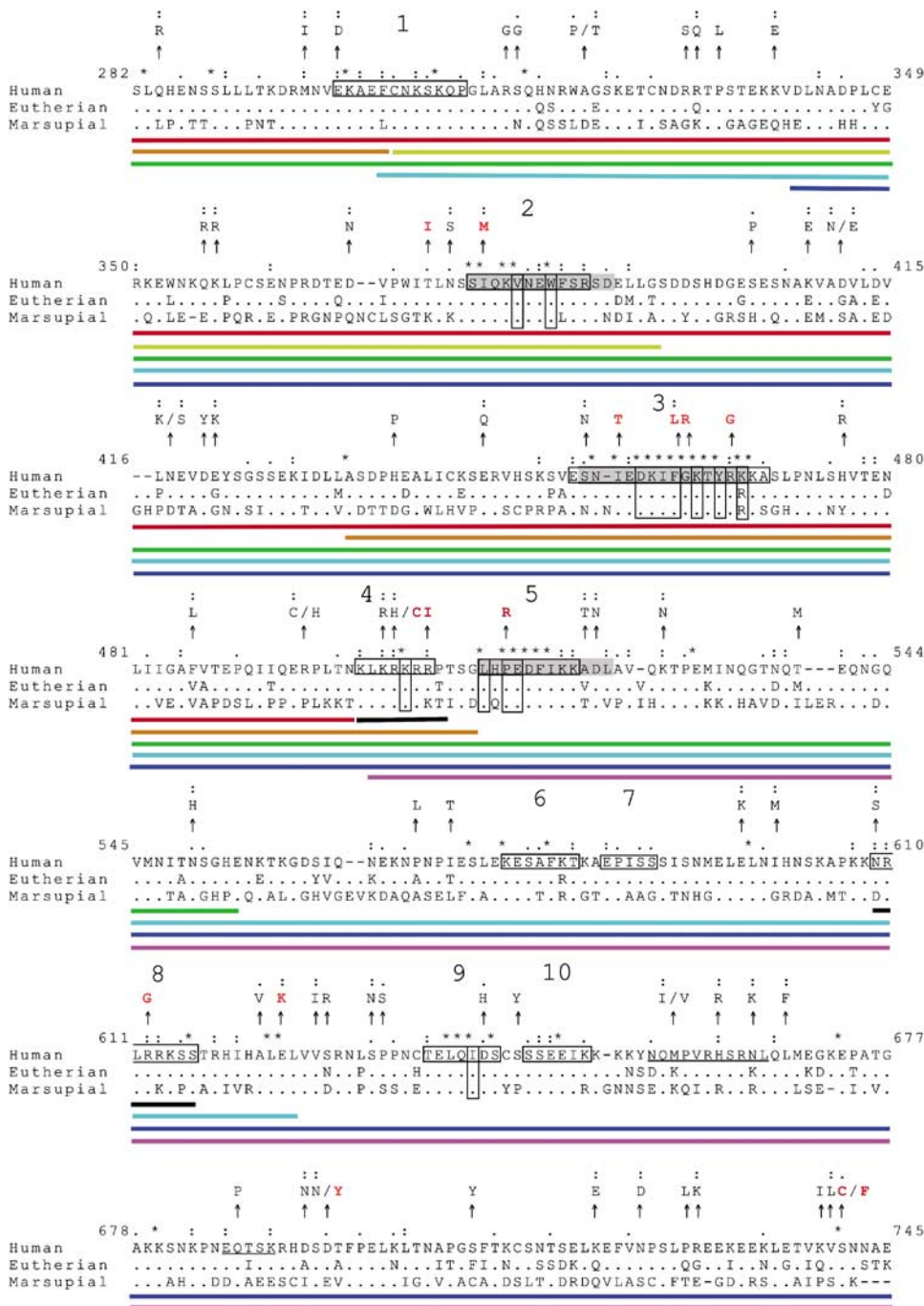
Clade	Members	Approximate divergence date (Myr)
Chordata		
Mammalia	Eutheria – placentals	5–130 Myr <sup>a,b</sup>
	Marsupalia – marsupials	175 Myr
Amniota	Aves – birds	310 Myr
Tetrapoda	Anura – frogs	360 Myr

<sup>a</sup>Approximate divergence dates relative to humans million of years ago (Kumar and Hedges, 1998). <sup>b</sup>Eutherian dates given as a range from 57 species.

**Table 2** Information for BRCA1 sequences used in analyses

Genbank Accession Number	Mammalian order	Species	Sequence length (a.a.)	% a.a. identity with human	% a.a. identity with marsupial ancestor	% a.a. identity with eutherian ancestor
AF284033	Diprotodontia	Red kangaroo ( <i>Macropus rufus</i> )	331	48.1	84.1	48.7
AY211953	Diprotodontia	Red kangaroo ( <i>Macropus rufus</i> ) <sup>a</sup>	239			
AY211954	Diprotodontia	Tree kangaroo ( <i>Dendrolagus matschiei</i> )	239	48.1	83.5	50.2
AY211955	Diprotodontia	Wallaroo ( <i>Macropus robustus</i> )	236	48.3	84.1	50.8
AF284031	Diprotodontia	Coarse-haired wombat ( <i>Vombatus ursinus</i> )	816	38.5	90.1	46.1
AF355795	Dasyuromorphia	Brush-tailed phasogale ( <i>Phascogale tapoatafa</i> )	792	40.0	80.3	43.2
AF355796	Peramelemorphia	Long-nosed bandicoot ( <i>Echymipera kalubu</i> )	806	37.0	79.0	43.5
AY211956	Didelphimorphia	Virginia opossum ( <i>Didelphis virginiana</i> )	236	45.4	78.9	47.1
AF355794	Paucituberculata	Silky shrew opossum ( <i>Caenolestes fuliginosus</i> )	794	39.1	76.7	42.3
		Marsupial average		43.1 ± 0.78	82.1 ± 0.73	46.5 ± 0.64
U14680	Primate	Human ( <i>Homo sapiens</i> )	822		42.3	80.2
AF355273	—	Chicken ( <i>Gallus gallus</i> )	796 <sup>b</sup>	21.1	24.9	23.3
AF416868	—	Frog ( <i>Xenopus laevi</i> )	627 <sup>b</sup>	14.1	13.8	15.4

<sup>a</sup>Not used in analyses because 97.8% identical to longer red kangaroo sequence. <sup>b</sup>Exact length could not be determined because only small portions of these sequences could be aligned to the mammalian sequences.



**Figure 1** Alignment of BRCA1 amino acids 282–1103 from human and the eutherian (Fleming *et al.*, 2003) and marsupial ancestors derived from Bayesian inference. Residues in the ancestor sequences identical to the human are represented by dots. Marks above the human residues represent identical (\*), weakly conservative (·), and strongly conservative (:) sites in eight marsupial sequences and 37 eutherian ancestral sequences (Fleming *et al.*, 2003) based on the Gonnet PAM 250 matrix (Gonnet *et al.*, 1992). Boxed sites (vertical) represent amino acids that are fixed in 37 eutherian ancestral sequences, eight marsupial sequences, frog, and chicken sequences. In the human sequence, the shaded sequences show the five regions conserved  $\geq 60\%$  among 37 eutherian ancestral sequences, eight marsupial sequences, and frog and chicken sequences. Regions conserved among the eight marsupial and 37 eutherian ancestral sequences are outlined where conservation is  $\geq 80\%$  (numbered 1–13). Three regions conserved in marsupials, chicken, and frog are underlined. Arrows above the human sequence show missense changes found in the BIC database. Marks above the missense change represent weakly (·) or strongly conservative (:) changes. The 21 changes in red are predicted to affect function. Colored lines below the alignment indicate the approximate locations of sites that have been identified as protein-interaction domains (reviewed in Welch and King, 2001): c-Myc (175–303 and 433–511) and p300/CBP (1–303) and estrogen receptor (1–300) (orange line), P53 (224–500) (red line), SWI/SNF (260–553) (green line), pRB (304–394) (yellow line), VCP (303–625) (light blue line), RAD 50 and ZBRK1(341–748) (blue line), STAT1 (502–802) (pink line), RAD 51 and androgen receptor (758–1064) (olive green line), and BRAP2 and Importin alpha (interact with NLS 500–508 and 609–615) (black line)

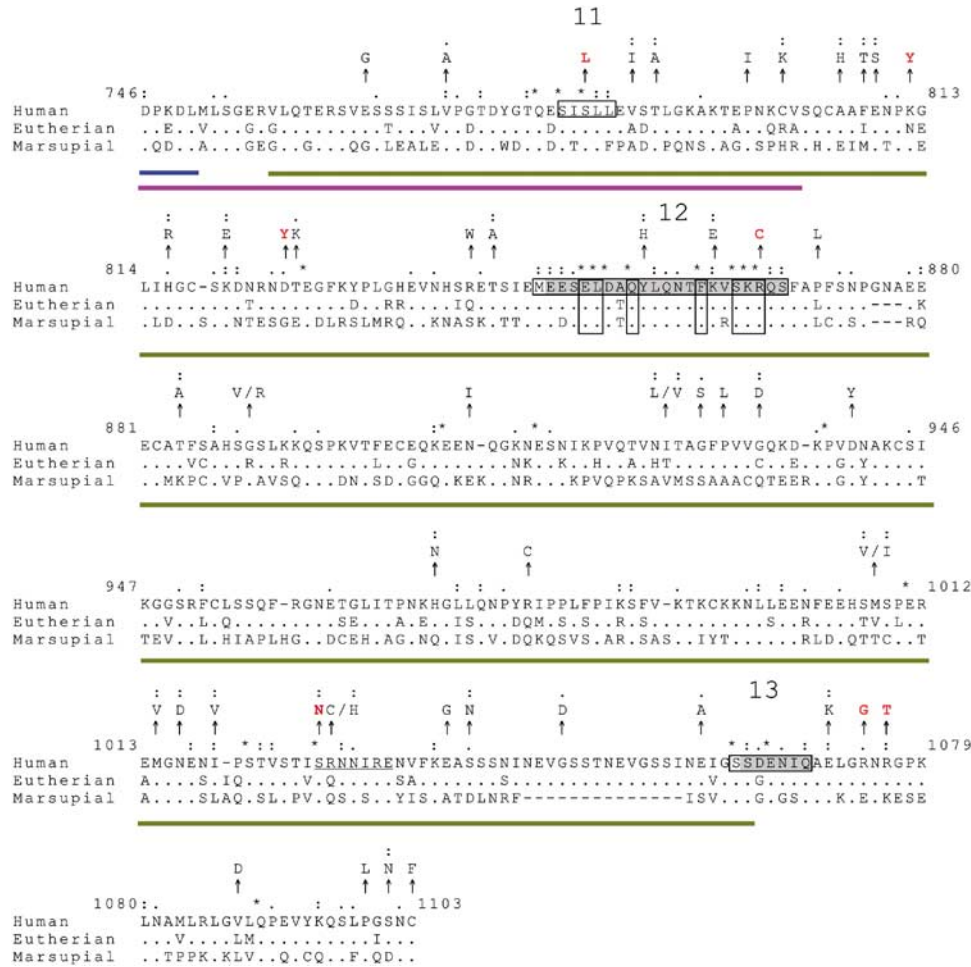


Figure 1 Continued.

respect to the GenBank sequences from closely related species confirm that the sequences are *BRCA1* orthologues and not pseudogenes (tree available at [http://www.fhcrc.org/science/dog\\_genome/supplemental\\_info/](http://www.fhcrc.org/science/dog_genome/supplemental_info/)). The novel tree kangaroo, red kangaroo, and wallaroo sequences form a well-supported clade that also includes the only other representative of the diprotodont order in this analysis, the wombat. Similarly, the two New World marsupial sequences in the analysis, from the Virginia and silky shrew opossums, form a clade for which we observe strong support.

There is no evidence in marsupials that any sites in this region of *BRCA1* exon 11 are evolving more rapidly than predicted by neutral theory. None of the marsupial *BRCA1* exon 11 sites in the alignment of eight marsupial species had posterior probabilities  $\geq 0.5$  of being under positive selection. The average and highest posterior probabilities reported were 0.1738 and 0.2272, respectively.

#### Conserved regions

Based on pairwise analysis, there is an average of 43.1% ( $\pm 0.78\%$  s.e.) amino-acid identity between marsupial and human *BRCA1* exon 11 sequences. When homo-

logous amino-acid sites were labeled as fixed, weakly, or strongly conservative, or nonconservative based on the alignment of eight marsupial and 37 eutherian ancestral sequences, 67 (7.95%) of the 843 amino-acid sites are fixed and an additional 189 (22.4%) are conservative. When all eight marsupial sequences were aligned with 37 eutherian ancestral, chicken, and frog sequences, only 21 fixed sites remain. Fixed and conservative residues in mammalian exon 11 are not randomly distributed across the sequence; most are associated with other fixed ( $Z=8.27$ ,  $P<0.001$ ) or conservative sites ( $Z=3.86$ ,  $P<0.001$ ).

In all, 13 highly conserved regions were identified among the eight marsupial and 37 eutherian ancestral sequences ('mammal alignment' boxed in Figure 1) between human amino acids 282 and 1103 following the criteria that we developed earlier (Fleming *et al.*, 2003). Five of these (#1, 2, 3, 5, 12) expanded six of the conserved regions previously identified in an alignment of 57 eutherian mammal sequences (Fleming *et al.*, 2003). A seventh region conserved in the eutherian alignment, adjacent to region #1 from this analysis, was not conserved in marsupials. Seven of the highly conserved regions in the mammal alignment (#4 and 6–11) were between codons 501 and 786. When chicken

and frog sequences were included ('three-class alignment'), five regions (#2, 3, 5, 12, 13) were conserved among all three classes at  $\geq 60\%$  (shaded in Figure 1).

To determine the percent conservation levels not likely to occur by chance among the various amino-acid sequences, we calculated the mean  $\pm$  s.d. for the percent of fixed and conservative sites in each five amino-acid window across the alignments. For the mammal alignment, the mean  $\pm$  s.d. percent conservation is  $30 \pm 25\%$ , supporting the selection of a  $\geq 80$  percent conservation criterion to define highly conserved regions in mammals (Figure 2a). In the three-class alignment, the percent conservation is  $13 \pm 20\%$ , which includes five highly conserved regions that were previously identified using the  $\geq 60\%$  criterion (Figure 2a). The remaining eight regions (#1, 4, 6, 7, 8, 9, 10, and 11) are also conserved to varying degrees in the chicken and frog sequences (Figure 2b and c). Regions #4 and 11 are highly conserved in the chicken ( $21 \pm 23\%$ ) but not the frog sequences ( $16 \pm 22\%$ ), region #9 is conserved in the frog but not chicken sequences, and regions #1 and 10 are conserved in both, leaving three regions that may be considered mammal specific (#6, 7, and 8; Figure 2a–c).

To determine if eutherian mammals have lost any regions that are conserved among marsupials, chicken, and frog sequences, we calculated the mean  $\pm$  s.d. in an alignment of eight marsupial sequences and frog and chicken sequences ('noneutherian alignment'; Figure 2d). The mean  $\pm$  s.d. was  $22 \pm 25\%$ , and we identified three additional conserved regions among the noneutherian sequences (underlined in Figure 1).

### Missense changes

There are 130 missense changes at 118 sites reported in the BIC database between amino acids 282 and 1103. Missense changes were designated conservative or nonconservative based on their effects in the human sequence. The missense changes are randomly distributed across the alignment of 45 mammalian sequences ( $\chi^2 = 1.5$ , d.f. = 1,  $P > 0.1$ ). There is no difference in the distribution of conservative and nonconservative missense changes relative to fixed, conservative, or nonconservative sites ( $\chi^2 = 1.03$ , d.f. = 1,  $P > 0.1$ ).

The distribution of missense changes was compared to that of fixed and conservative amino-acid sites. Changes predicted to affect protein function were either (1) nonconservative missense changes at fixed or conservative sites or (2) conservative missense changes at fixed sites. Based on the levels of site conservation in marsupial and eutherian ancestral sequences, 21 of the 130 missense changes are predicted to affect function (red in Figure 1). In all, 11 occur at 10 sites that are fixed in the mammal sequences (Fleming *et al.*, 2003). Three of the 11 are also at fixed sites in frog and chicken. The remaining 10 are nonconservative changes that occur in sites that are conservative in mammals. In all, 13 missense changes predicted to affect function in Fleming *et al.* (2003) occur at sites that are not as well conserved in the marsupial sequences, and thus are no longer predicted to be deleterious.

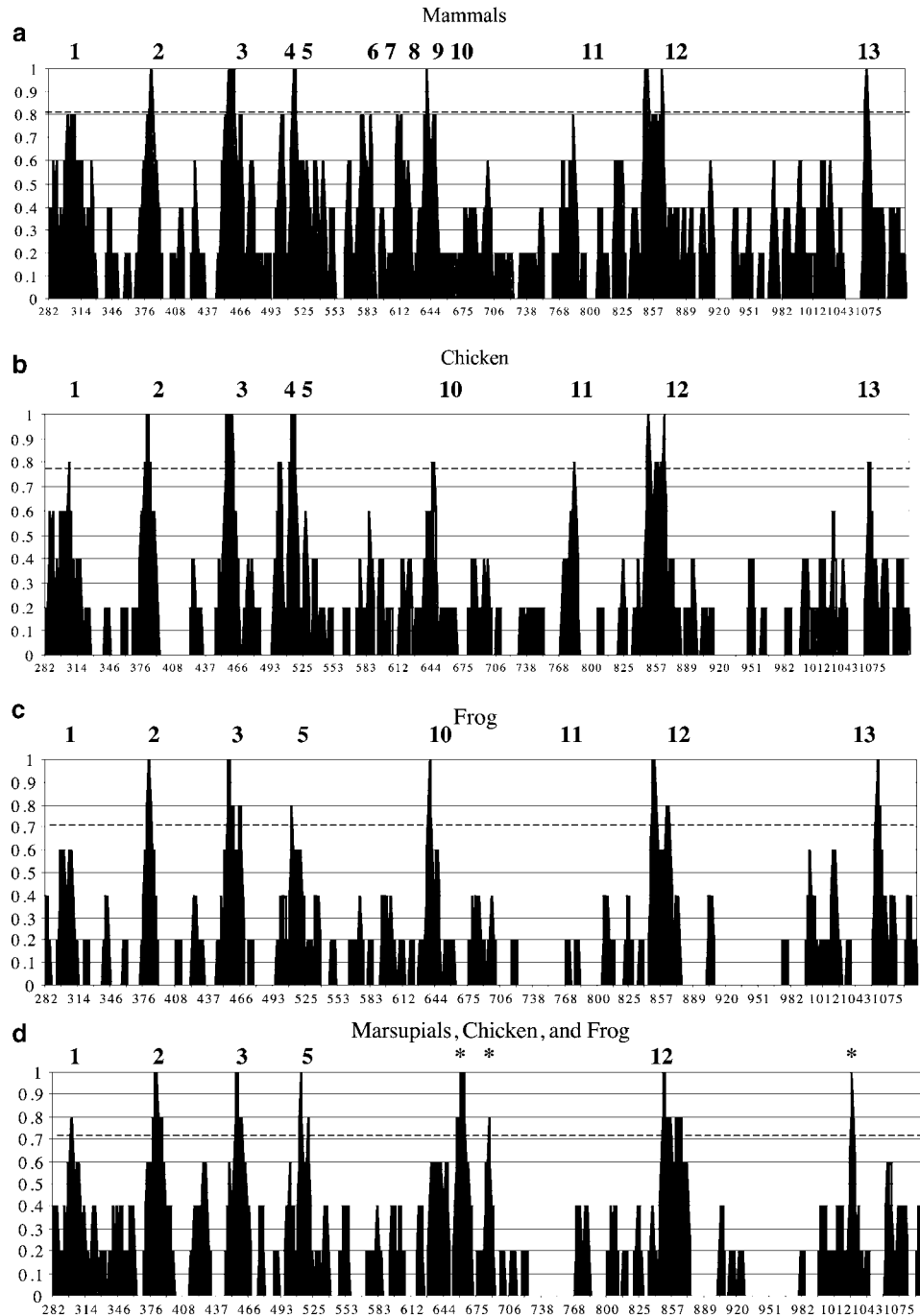
Half of the missense changes (11 of 21) predicted to affect function occur in highly conserved regions in the mammal alignment, and seven occur in highly conserved regions in the three-class alignment with four missense changes in region #3 alone (Figure 1). Another three changes occur in regions newly identified in this analysis (#4 and 11) that were also conserved in chicken but not frog (Figures 1 and 2). Region #4 overlaps a nuclear localization site (NLS), and the 11th missense change in a highly conserved region predicted to affect function occurs in region #8, which overlaps a second NLS that is conserved in mammals but not in chicken or frog (Figures 1 and 2). One missense change predicted to affect function occurs at a fixed site in a region that is well conserved in noneutherians, but not eutherian mammals.

### Discussion

The addition of eight marsupial BRCA1 sequences to our (Fleming *et al.*, 2003) 37 eutherian ancestral BRCA1 sequences reduced the number of missense changes predicted to affect function due to changes in site conservation between human amino acids 282 and 1103 from 34 to 21 and expanded six highly conserved regions. In all, 11 additional conserved regions were identified: two conserved in all three classes, one in mammals and frog, two in mammals and chicken, three in noneutherian species, and three only in mammals. Positively selected sites in BRCA1 exon 11 have been identified in previous studies using eutherian mammals (Huttley *et al.*, 2000; Fleming *et al.*, 2003). We (Fleming *et al.*, 2003) found three positively selected sites in the RAD51-interaction domain in eutherian BRCA1, but none was identified in marsupial BRCA1. This suggests that the RAD51 domain may serve a unique function in eutherians, which has benefited from diversifying selection of amino acids at specific sites.

The eutherian ancestral sequences allowed us to detect highly conserved regions in mammals other than those previously identified in the 57 eutherian sequences by eliminating substitutions that arose independently in single species and thus could not be distinguished from sequencing errors or species-specific neutral polymorphisms. In all, 13 conserved regions were identified among eight marsupial and 37 eutherian ancestral exon 11 sequences, compared to the seven found among 57 eutherian sequences between amino acids 282 and 1103 (Fleming *et al.*, 2003). Eight of the 13 were not identified in the eutherian alignment; however, all but one of the seven eutherian conserved regions are conserved in the mammal alignment. The addition of marsupials to the analysis does not change the five regions of conservation across eutherian mammals, birds, and amphibians (Fleming *et al.*, 2003), confirming the probable functional significance of these regions in all three classes.

Four (#1, 2, 3, and 5) of the eight regions conserved in all three classes (Figure 2) occur in the 5' portion of the sequence (amino acids 282–553), which overlaps several



**Figure 2** Results of sliding window analyses of the percent conservation across sites 282–1103 of exon 11 within a sliding window of five amino acids for (a) the alignment of 37 eutherian ancestral sequences and eight marsupial sequences (mammal sequences), (b) the alignment of the mammal and chicken sequences, (c) the alignment of the mammal and frog sequences, and (d) the alignment of the marsupial, frog, and chicken sequences. The dotted lines indicate the percent conservation that is two standard deviations above the mean value for each alignment. Values that exceed this line show a much higher percent conservation than would be expected by chance. Numbers above the lines indicate the approximate positions of highly conserved regions as defined by the criteria of Fleming *et al.* (2003) and shown in Figure 1

protein-interaction domains hypothesized to be involved in transcription (reviewed in Welsh and King, 2001). Nine missense changes (six within these conserved regions) predicted to affect function also occur in this portion of exon 11. Amino acids in the 5' portion of the sequence may be more conserved due to the density of interaction domains. Another two conserved regions in

the three-class alignment and five missense changes (one in a conserved region) predicted to affect function occur in the RAD51-interaction domain, which is implicated in double-strand break repair.

The putative STAT1 interaction domain (a.a. 502–802) (Ouchi *et al.*, 2000) overlaps seven of the highly conserved regions in the mammal alignment not

previously identified (Fleming *et al.*, 2003) (#4 and 6–11), which include the mammal-specific regions, and nine of the missense changes predicted to affect function. STAT1 is a member of a family of signal transducers and transcriptional activators and is required for response to interferon gamma, which is involved in innate and adaptive immune response (reviewed in Stark *et al.*, 1998). Two of the conserved regions found in the noneutherian alignment and three regions that were conserved in mammals and either frog or chicken also occur within the STAT1 interaction domain. This pattern of loss of regions highly conserved among some distantly related taxa but not others is consistent with changes of function under different directional rather than disruptive (positive) selection pressures in the different classes and superorders. The concentration of apparently mammal-specific regions in the STAT1-interaction domain suggests that STAT1 may be important to mammary function and missense changes in this domain could be involved in breast cancer. Three missense changes, but only one predicted to affect function (R612G), are found in two of these mammal-specific regions (#6 and 8).

Regions 4 and 8 also perfectly overlap the two BRCA1 NLSs (a.a. 500–508 and 609–615) and three missense changes predicted to affect function occur in these regions (R504C, R507I, and R612G). Aberrant localization of BRCA1 has been found in breast cancer cell lines with NLS mutations, suggesting that the localization to the nucleus is important for the normal function of the protein (Chen *et al.*, 1996). In mammals, BARD1 interacts with the RING domain of BRCA1 and explains the nuclear localization of BRCA1 exon 11 splice variants (Fabbro *et al.*, 2002). However, efficient localization of BRCA1 to the nucleus through NLSs may be required for normal function in mammals; therefore, mutations at these sites are likely to be associated with breast cancers and merit further examination.

Combined with functional and structural data, comparative analyses of proteins can provide an important resource for identifying functionally important regions of incompletely characterized proteins, like BRCA1. The inclusion of divergent marsupial mammal sequences reduced the number of missense changes predicted to affect function by almost 40%, indicating a probable improvement over our previous predictions (Fleming *et al.*, 2003). Half of the missense changes predicted to affect function in this study occur in identified conserved regions, making them particularly good candidates for further studies of the relationship between missense changes and breast cancer. The remaining 10 missense changes may have site-specific functions or occur in regions conserved under less stringent criteria. While this study shows that the STAT1 and NLS regions of BRCA1 exon 11 appear to be important to both mammalian and nonmammalian BRCA1 function, the sequencing of other exons in additional eutherian, marsupial, monotreme, and non-mammalian species may be useful for identifying other mammal-specific conserved regions and could be helpful

in identifying regions evolving differently between mammals and nonmammals. While further evaluation is needed, the 21 missense changes prioritized for functional analyses will hopefully improve the efficiency of disease-associated BRCA1 mutational screening.

## Materials and methods

### *BRCA1 sequencing and sequence alignment*

We sequenced a portion of exon 11 that had not been previously sequenced from four marsupial species: tree kangaroo (*Dendrolagus matschiei*), red kangaroo (*Macropus rufus*), wallaroo (*Macropus robustus*), and Virginia opossum (*Didelphis virginiana*). DNA samples were extracted from liver tissue or lymphocyte nuclei from blood collected in Anti-coagulant Acid Citrate Dextrose blood collection tubes (Becton Dickinson, Franklin Lakes, NJ, USA) (Bell *et al.*, 1981).

Partial exon 11 sequences for all marsupials were amplified using Polymerase Chain Reaction (PCR). Degenerate primers 5'AGCAGCATCCAGAAGGTGAAYGARTGGTT3' and 5'YTTCTTGATAAARTCCTCAAGATGAAG3' were designed using CODEHOP (Rose *et al.*, 1998) based on eutherian exon 11 sequences from GenBank. PCR amplification of BRCA1 nucleotides (nt) 846–3309 was performed in 12.0  $\mu$ l reactions containing 100 ng genomic DNA, 25.6 mM (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 107.2 mM Tris-Cl (pH 8.8), 0.016% Tween-20, 1.5 mM MgCl<sub>2</sub>, 0.2 mM each of dNTPs, 0.25 U T4 *Taq* polymerase, 12.5 mM reverse primer, and 12.5 mM  $\gamma$ -<sup>32</sup>P-labeled forward primer. The forward primer was labeled with [ $\gamma$ -<sup>32</sup>P]ATP in a 1.0  $\mu$ l reaction of 12.5 mM forward primer, 0.3 U T4 PNK, and 1  $\times$  T4 kinase reaction buffer (Roche, Indianapolis, IN, USA). PCR reactions were performed in a Perkin-Elmer (Boston, MA, USA) 9600 thermocycler under the following cycling conditions: initial denaturation of 95°C for 1 min, 35 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 45 s, and a final extension of 72°C for 7 min.

Radiolabeled PCR products were separated on a 4% polyacrylamide gel and visualized using autoradiography. Bands of the expected size (500–600 nt) for each of the marsupial species were cut from the gel, and the DNA was eluted in 50–100  $\mu$ l sterile water for 15 min at room temperature.

A nonradioactive 50  $\mu$ l PCR reaction was then performed using 5  $\mu$ l of the radioactive PCR product described above to obtain the material for DNA sequencing. Reactions included 25.6 mM (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 107.2 mM Tris-Cl (pH 8.8), 0.016% Tween-20, 1.5 mM MgCl<sub>2</sub>, 0.2 mM each of dNTPs, 0.5 U T4 *Taq* polymerase, 1.0 mM reverse primer, and 1.0 mM forward primer. PCR reactions were carried out as described at about a 58°C annealing temperature.

The PCR products were quantified on a 1.0% agarose gel using a low-mass DNA ladder (GIBCO BRL, Carlsbad, CA, USA). Unincorporated PCR primers and dNTPs were neutralized prior to sequencing using 15 ng of product, 10 U of exonuclease I, and 1 U of shrimp alkaline phosphatase with a 30 min incubation at 37°C and a 15 min deactivation at 80°C.

The enzyme-treated PCR product was divided into two separate reactions for the forward and reverse primers used in the previous PCR amplifications. Sequencing reactions were carried out using the Big Dye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems, Foster City, CA, USA) according to the manufacturer's instructions. Unincorporated dye terminators were removed from the sequencing

product by filtration through Sephadex G-50 resin using Millipore Multiscreen HV plates. Samples were separated on the ABI PRISM 3700 DNA Analyzer (Foster City, CA, USA), and the data were analysed with the Applied Biosystems ABI PRISM 3700 Collection software, version 2.0. Chromatogram files were transferred to a Unix workstation and reanalysed using the base-calling software Phred, version 0.000925 ([www.phrap.org](http://www.phrap.org)) and the assembly program Phrap, version 0.990329 ([www.phrap.org](http://www.phrap.org)) and viewed using the program Consed, version 11.0 ([www.phrap.org](http://www.phrap.org)).

We obtained partial exon 11 nucleotide sequences of four additional marsupial species plus red kangaroo, humans, and representatives of two other vertebrate classes, Amphibia (frog) and Aves (chicken) from GenBank (Table 2) and 37 eutherian ancestral sequences from Fleming *et al.* (2003). Nucleotide sequences were aligned using ClustalX, version 1.81 (Jeanmougin *et al.*, 1998) and adjusted by eye in MacClade, version 4.03 (Maddison and Maddison, 2001). Phylogenetic analyses were based on the nucleotide sequences and the levels of site conservation across taxa were determined using the amino-acid translation. The alignment of ancestral eutherian, frog, and chicken sequences with marsupial sequences was also performed using ClustalX and based on amino-acid translations. Alignments are available online at [http://www.fhrc.org/science/dog\\_genome/supplemental\\_info/](http://www.fhrc.org/science/dog_genome/supplemental_info/).

#### Phylogenetic analysis of marsupial BRCA1 exon 11

All analyses were based on sequences that started at human codon 282 and ended at or before codon 1103. The determination of amino-acid conservation levels where sequence was not available for all taxa was based on the sequences of the remaining taxa.

We derived a phylogenetic tree of marsupial sequences with the human sequence as an outgroup using maximum-likelihood methods. Modeltest, version 3.04 (Posada and Crandall, 1998), was used to determine the substitution model that best fits these data: a general time-reversible substitution model with invariant sites and a gamma distribution. Maximum-likelihood analysis using a heuristic search with 10 repetitions was run using PAUP\* (Swofford, 1998). A total of 500 bootstrap replicates were run to assess support for the branches.

MrBayes, version 2.01 (Huelsenbeck and Ronquist, 2001), a Bayesian analysis program, was used to find positively selected sites in marsupials. Residues at a site that are evolving more rapidly than predicted by neutral theory (i.e. posterior probability >0.50) may demonstrate positive (diversifying) selection at that site. We generated a phylogenetic tree using humans as the outgroup and specifying a general time-reversible substitution model with estimated base frequencies. Codons were used as sites and adjacent sites were assumed to have correlated rates. Omega, the ratio of nonsynonymous to synonymous substitutions, was varied according to the Nielsen and Yang (1998) model of among-site rate variation. Markov chain Monte Carlo sampling was started from random trees for four simultaneous chains with uniform prior probability distributions for tree topologies and the rate matrix and a Dirichlet prior of 4.0 for base frequencies. The analysis was run for 100 000 generations and trees were saved every 100

generations. The likelihood values converged around 20 000 generations, so the first 200 trees were discarded as 'burn in.' The consensus tree was viewed in PAUP\*, version 4.0 beta (Swofford, 1998).

MrBayes, version 2.01, was also used to estimate the sequence of the ancestor of the eight marsupial sequences ('marsupial ancestor') included for heuristic purposes in Figure 1. To determine the tree topology, an initial analysis was run using the above parameters, except that the rate variation for each codon position was estimated separately. The marsupial ancestor's sequence was generated by a second analysis, which was run with the clades constrained to maintain the same tree topology as the initial tree.

#### Conserved regions

Homologous amino-acid sites were categorized according to the Gonnet PAM 250 substitution matrix (Gonnet *et al.*, 1992). Sites were labeled as fixed (having a single residue in all sequences), conservative (including conservative substitutions only), or nonconservative (including at least one nonconservative substitution). Conservative sites were further described as 'weakly' or 'strongly' conservative based on matrix values >0 and ≤0.5 or values >0.5, respectively. Two-tailed, one-sample Runs' tests (Siegel, 1956) were used to determine whether residues that were fixed or conservative were associated with one another or randomly distributed.

Regions of the protein with high levels of amino-acid conservation across taxa ('highly conserved regions') were identified using a sliding window of five amino acids in length as described previously (Fleming *et al.*, 2003). To compare percent conservation of regions between alignments (e.g. mammals alone versus mammals and chicken), we calculated the mean ± s.d. for the percent of fixed and conservative sites in each five amino-acid window across the alignment. Windows that included the sequence with conservation levels greater than two standard deviations above the mean were considered highly conserved.

#### Missense changes

Missense changes reported in the BIC database between codons 282 and 1103 of BRCA1 were identified. The missense changes were designated conservative or nonconservative based on the Gonnet PAM 250 matrix. The association of missense changes (conservative or nonconservative) with fixed or conservative amino-acid sites in the alignment of eight marsupial and 37 eutherian ancestral sequences was tested using  $\chi^2$  with correction for continuity ( $P \leq 0.05$ ) (Sokal and Rohlf, 1995).

#### Acknowledgements

We thank Dr John AW Kirsch of the University of Wisconsin Zoological Museum for the Virginia opossum tissue sample, Dr Tammie Bettinger of the Cleveland Zoo for red kangaroo and wallaroo blood samples, and Dr Janis Joslin of the Woodland Park Zoo for tree kangaroo blood samples. This work was supported by Grant T32-HG00035 to CJ Ramirez, Grant K05 CA-90754-01 to EA Ostrander, and Grant U24 CA-78164 to JD Potter.

#### References

Bell G, Karam J and Rutter W. (1981). *Proc. Natl. Acad. Sci.*, **78**, 5759–5763.

Brown JR, Douady CJ, Italia MJ, Marshall WE and Stanhope MJ. (2001). *Nat. Genet.*, **28**, 281–285.

- Brzovic PS, Meza JE, King MC and Klevit RE. (2001). *J. Biol. Chem.*, **276**, 41399–41406.
- Canfield PJ, Hartley WJ and Reddacliff GL. (1990a). *J. Comp. Pathol.*, **103**, 147–158.
- Canfield PJ, Hartley WJ and Reddacliff GL. (1990b). *J. Comp. Pathol.*, **103**, 135–146.
- Casey HW, Giles RC and Kwapien RP. (1979). *Recent Results Cancer Res.*, **66**, 129–160.
- Castilla LH, Couch FJ, Erdos MR, Hoskins KF, Calzone K, Collins FS and Weber BL. (1994). *Nat. Genet.*, **8**, 387–391.
- Chapman MA, Charchar FJ, Kinston S, Bird CP, Grafham D, Rogers J, Grutzner F, Marshall Graves JA, Green AR and Gottgens B. (2003). *Genomics*, **81**, 249–259.
- Chen CF, Li S, Chen Y, Chen PL, Sharp ZD and Lee WH. (1996). *J. Biol. Chem.*, **271**, 32863–32868.
- Durocher F, Shattuck-Eidens D, McClure M, Labrie F, Skolnick MH, Goldgar DE and Simard J. (1996). *Hum. Mol. Genet.*, **5**, 835–842.
- Fabbro M, Rodriguez JA, Baer R and Henderson BR. (2002). *J. Biol. Chem.*, **277**, 21315–21324.
- Findlay L and Renfree M. (1984). *Physiological Strategies in Lactation*, Vol. 51. Peak M, Vernon RG, Knight CH (eds). Academic Press: London, pp. 403–432.
- Fleming MA, Ostrander G, Ramirez CJ, Potter J and Ostrander EA. (2003). *Proc. Natl. Acad. Sci.*, **100**, 1151–1156.
- Friedman LS, Ostermeyer EA, Szabo CI, Dowd P, Lynch ED, Rowell SE and King MC. (1994). *Nat. Genet.*, **8**, 399–404.
- Ganesh S, Agarwala KL, Amano K, Suzuki T, Delgado-Escueta AV and Yamakawa K. (2001). *Biochem. Biophys. Res. Commun.*, **283**, 1046–1053.
- Gonnet GH, Cohen MA and Benner SA. (1992). *Science*, **256**, 1443–1445.
- Hamilton JM. (1974). *Adv. Cancer Res.*, **19**, 1–45.
- Huber LJ, Yang TW, Sarkisian CJ, Master SR, Deng CX and Chodosh LA. (2001). *Mol. Cell. Biol.*, **21**, 4005–4015.
- Huelsenbeck J and Ronquist F. (2001). *Bioinformatics*, **17**, 754–755.
- Huttley GA, Easteal S, Southey MC, Tesoriero A, Giles GG, McCredie MR, Hopper JL and Venter DJ. (2000). *Nat. Genet.*, **25**, 410–413.
- Jeanmougin F, Thompson JD, Gouy M, Higgins DG and Gibson TJ. (1998). *Trends Biochem. Sci.*, **23**, 403–405.
- Knight CH. (1984). *Physiological Strategies in Lactation*, Vol. 51. Peak M, Vernon RG, Knight CH (eds). Academic Press: London, pp. 147–170.
- Koerberl DD, Bottema CD, Ketterling RP, Bridge PJ, Lillicrap DP and Sommer SS. (1990). *Am. J. Hum. Genet.*, **47**, 202–217.
- Kumar S and Hedges SB. (1998). *Nature*, **392**, 917–920.
- Lee WH and Boyer TG. (2001). *Lancet*, **358** (Suppl), S5.
- Maddison D and Maddison W. (2001). *MacClade 4: Analysis of phylogeny and character evolution*. Sinauer Associates: Sunderland.
- Malone KE, Daling JR, Neal C, Suter NM, O'Brien C, Cushing-Haugen K, Jonasdottir TJ, Thompson JD and Ostrander EA. (2000). *Cancer*, **88**, 1393–1402.
- Misdorp W. (1996). *Vet. Q.*, **18**, 32–36.
- Monteiro AN, August A and Hanafusa H. (1996). *Proc. Natl. Acad. Sci. USA*, **93**, 13595–13599.
- Newman B, Mu H, Butler LM, Millikan RC, Moorman PG and King MC. (1998). *JAMA*, **279**, 915–921.
- Nielsen R and Yang Z. (1998). *Genetics*, **148**, 929–936.
- Ouchi T, Lee SW, Ouchi M, Aaronson SA and Horvath CM. (2000). *Proc. Natl. Acad. Sci. USA*, **97**, 5208–5213.
- Posada D and Crandall K. (1998). *Bioinformatics*, **14**, 817–818.
- Rose TM, Schultz ER, Henikoff JG, Pietrokovski S, McCallum CM and Henikoff S. (1998). *Nucleic Acids Res.*, **26**, 1628–1635.
- Shen D and Vadgama JV. (1999). *Oncol Res.*, **11**, 63–69.
- Siegel S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill: New York.
- Simard J, Tonin P, Durocher F, Morgan K, Rommens J, Gringras S, Samson C, Leblanc JF, Belanger C, Dion F, Liu Q, Skolnick M, Goldgar DE, Shattuck-Eidens D, Labrie F and Narod SA. (1994). *Nat. Genet.*, **8**, 392–398.
- Sokal RR and Rohlf FJ. (1995). *Biometry: The Principles Practice of Statistics in Biological Research*, 3rd edn. WH Freeman Company: New York.
- Stark GR, Kerr IM, Williams BR, Silverman RH and Schreiber RD. (1998). *Annu. Rev. Biochem.*, **67**, 227–264.
- Straube EF and Callinan RB. (1980). *J. Comp. Pathol.*, **90**, 495–497.
- Swofford D. (1998). *PAUP\*. Phylogenetic Analysis Using Parsimony (\* and other Methods)*. Sinauer Associates: Sunderland.
- Szabo C, Masiello A, Ryan JF and Brody LC. (2000). *Hum. Mutat.*, **16**, 123–131.
- Tyndale-Biscoe CH, Stewart F and Hinds LA. (1984). *Physiological Strategies in Lactation*, Vol. 51. Peak M, Vernon RG, Knight CH (eds). Academic Press: London, pp. 389–401.
- Vail DM and MacEwen EG. (2000). *Cancer Invest.*, **18**, 781–792.
- Vallon-Christersson J, Cayanan C, Haraldsson K, Loman N, Bergthorsson JT, Brondum-Nielsen K, Gerdes AM, Moller P, Kristoffersson U, Olsson H, Borg A and Monteiro AN. (2001). *Hum. Mol. Genet.*, **10**, 353–360.
- Venkitaraman AR. (2001). *J. Cell Sci.*, **114**, 3591–3598.
- Welch PL and King MC. (2001). *Hum. Mol. Genet.*, **10**, 705–713.