FEBS Letters 581 (2007) 4668–4674

CpG/CpNpG motifs in the coding region are preferred sites for mutagenesis in the breast cancer susceptibility genes

Lydia W.T. Cheung[a,1], Yiu Fai Lee[b,1,2], Tuen Wai Ng[b], Wai Ki Ching[b], Ui Soon Khoo[c], Michael K.P. Ng[d], Alice S.T. Wong[a,*]

[a] School of Biological Sciences, University of Hong Kong, Hong Kong
[b] Department of Mathematics, University of Hong Kong, Hong Kong
[c] Department of Pathology, University of Hong Kong, Hong Kong
[d] Department of Mathematics, Hong Kong Baptist University, Hong Kong

Received 18 June 2007; revised 21 August 2007; accepted 26 August 2007

Available online 4 September 2007

Edited by Varda Rotter

Abstract **The range of *BRCA1/BRCA2* gene mutations is diverse and the mechanism accounting for this heterogeneity is obscure. To gain insight into the endogenous mutational mechanisms involved, we evaluated the association of specific sequences (i.e. CpG/CpNpG motifs, homonucleotides, short repeats) and mutations within the genes. We classified 1337 published mutations in *BRCA1* (1765 *BRCA2* mutations) for each specific sequence, and employed computer simulation combined with mathematical calculations to estimate the true underlying tendency of mutation occurrence. Interestingly, we found no mutational bias to homonucleotides and repeats in deletions/insertions and substitutions but striking bias to CpG/CpNpG in substitutions in both genes. This suggests that methylation-dependent DNA alterations would be a major mechanism for mutagenesis.**

Keywords: Breast cancer susceptibility gene; BRCA1; BRCA2; Gene mutation

1. Introduction

Breast and ovarian cancer are the leading causes of cancer death among women, with a lifetime risk of about 12.5% and 1.5%, respectively. *BRCA1* and *BRCA2* are the most important cancer susceptibility genes found in the diseases. Germline mutations of the *BRCA1/BRCA2* are responsible for 30–40% of familial breast cancer cases [1]. Familial ovarian cancer comprises at least 10% of cases.

BRCA1 (17q12-21) comprises over 70 kb of genomic DNA and contains 22 coding exons, which encodes a 7.8 kb transcript [2]. *BRCA2* (13q12-13) consists of 26 coding exons, which are transcribed into a 11–12 kb mRNA [3]. To date, more than 1337 different mutations in *BRCA1* (1765 *BRCA2* mutations) have been reported which are clearly associated with cancer susceptibility. These mutations and associated data have been listed in the Human Gene Mutation Database (HGMD) http://archive.uwcm.ac.uk/uwcm/mg/hgmd) and the Breast Cancer Information Core (BIC) in the National Human Genome Research Institute (http://research.nhgri.nih.gov/projects/bic/) [3]. The most prominent subclasses of these mutations are single base substitution, small deletion and insertion. The range of gene mutations is diverse and only a few recurrent mutations have been identified. The causes for the heterogeneity of molecular defects in these genes are not clear. However, there was evidence showing that specific nucleotide sequences (i.e. homonucleotides, short repeats, CpG/CpNpG motifs) could serve as common hotspots to mutagenesis in tumor suppressor genes, for example p53, retinoblastoma and neurofibromatosis (NF1) [4–9].

To gain insight into the endogenous mutational mechanisms involved, we would like to know whether the specific sequences significantly associate with mutations within the *BRCA1* and *BRCA2* genes. We developed a novel mathematical and computational approach which gives quantitative estimation of the association between simulated mutation events and specific sequences. This also excludes the possibility that specific sequences have been mutated by chance or because they are abundant in the genes.

2. Materials and methods

2.1. Data collection and mutation classification

BRCA1 and *BRCA2* gene mutations were retrieved from HGMD database at http://archive.uwcm.ac.uk/uwcm/mg/hgmd and BIC database at http://research.nhgri.nih.gov/projects/bic/ (last updated on November 2006). These entries comprise mutations with and without documented disease phenotype. We concluded from χ^2 test that the occurrence of the five specific sequences (HO, R(0), R(1), R(2) and CpG/CpNpG) (see below for annotation) did not differ between mutations with or without pathological significance. Thus, we included all the mutations listed for larger size of our dataset. We have 1337 (1765) different *BRCA1* (*BRCA2*) mutations including 764 (1135) substitutions, 415 (465) small deletions and 158 (165) insertions. All substitutions listed in the databases are non-synonymous which leads to amino acid replacement.

Classification analysis was performed by localizing mutations to the cDNA sequences according to GenBank entries U14680 and U43746 for *BRCA1* and *BRCA2*, respectively, and then identifying the nucleotides flanking the mutation sites. The influence of CpG/CpNpG motifs, homonucleotides (HO) (e.g. GGGG) and short repeats, which include direct repeats (e.g. $(AG)_n$) and inverted repeats (e.g. AGGA), on each mutation type (substitution, insertion and deletion) was assessed. Our

[*] Corresponding author. Fax: +852 2559 9114.
E-mail address: awong1@hku.hk (A.S.T. Wong).

[1] These authors equally contributed.
[2] Present address: Department of Psychiatry, University of Hong Kong, Hong Kong.

0014-5793/$32.00 © 2007 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2007.08.061

classification scheme of short repeats is slightly different from that given in Rodenhiser et al. [9,10]. Since mutation sites were found to be only significantly associated within very short distance to the nearby repeats [11–13], repeats with more than two nucleotides in-between were not taken into account in this study. Therefore, we denoted the class of short repeats without any gaps by R(0) (e.g. CTCT), while those with single gap by R(1) (e.g. $\underline{\text{GTT}}\text{A}\underline{\text{GTT}}$) or two gaps by R(2) (e.g. $\underline{\text{AC}}\text{TG}\underline{\text{AC}}$). Also, a single mutation could be classified into more than one specific sequences (e.g. ACCA → ACTA as both HO and R(0)), as a mutation might be caused independently by more than one mechanism [14].

2.2. Computer simulation

Because the distribution of recorded mutations was highly non-uniform, and multiple mutations occurring at a same place could often be observed, we constructed a mutation model that could incorporate this phenomenon by assuming that the probability of having mutation would be greater at site with recorded mutations. Let the number of nucleotide positions which have mutation record be N_A and those without be N_B. Therefore, N_B is simply equal to the total length of the gene minus N_A. In this model, we assumed that $P = 1/(\alpha N_A + N_B)$ and αP are the probability for a mutation to take place at a position without or with mutation record respectively, where α ($\geqslant 1$) is the mutation weighting. In this study, α was taken as: 1, 10, 20 or 30. When α is equal to 1, the probability of a mutation to occur at any position within the gene is the same, which corresponds to a uniform distribution and we took this case as our natural starting point. As the value of α increase, the probability of mutation to occur at a recorded position increases. This mimics the non-uniform distribution of a real situation. Since a linear relation was found between simulated counts and α when $\alpha \leqslant 30$, we did not consider a value of $\alpha > 30$.

A position where the simulated mutation would occur was randomly selected with a built-in function RAND in Microsoft Excel, which produces random values between 0 and 1 in equal chance, referred to as interval [0, 1]. We modified the generation process with α to obtain random numbers in spatially non-uniform manner. We divided the interval [0, 1] into ($\alpha N_A + N_B$) subintervals of equal length and assigned α intervals to each nucleotide position with mutation record and one interval to each position without record. A random number from the interval [0, 1] was selected which indicated the site for simulation. And the chance to simulate mutation at site with record would increase with the value of α.

A substitution was simulated based on the nucleotide-to-nucleotide substitution rates estimated from the observed data in our dataset. The length and base composition of inserted/deleted nucleotide fragment for simulation were determined by the relative frequencies of nucleotide and fragment lengths from the dataset. For each mutation weighting α, we performed 200000 simulations for each mutation type.

2.3. Mathematical calculations

Equations were derived and applied to our simulated data. Let p and q be the probability of having a specific sequence at a position with and without recorded respectively. Close value of p and q implies that the mutation type is not biased to the specific sequence, whereas significant difference between the two values implicates bias. We then performed T simulations of each mutation type. Based on the assumption that the probability of having a mutation would be higher site with mutation recorded, the expected number of mutations associated with the specific pattern can be expressed as

$$\left(\frac{\alpha p N_A + q N_B}{\alpha N_A + N_B}\right) T$$

where α is the mutation weighting. Let C_α be the number of counts of the specific sequence that has occurred in the simulation when the mutation weighting is α. For each combination of mutation weighting α_1 and α_2, the values of p and q could be obtained by solving the following system of equations

$$\left(\frac{\alpha_1 p N_A + q N_B}{\alpha_1 N_A + N_B}\right) T = C_{\alpha_1}$$
$$\left(\frac{\alpha_2 p N_A + q N_B}{\alpha_2 N_A + N_B}\right) T = C_{\alpha_2}$$

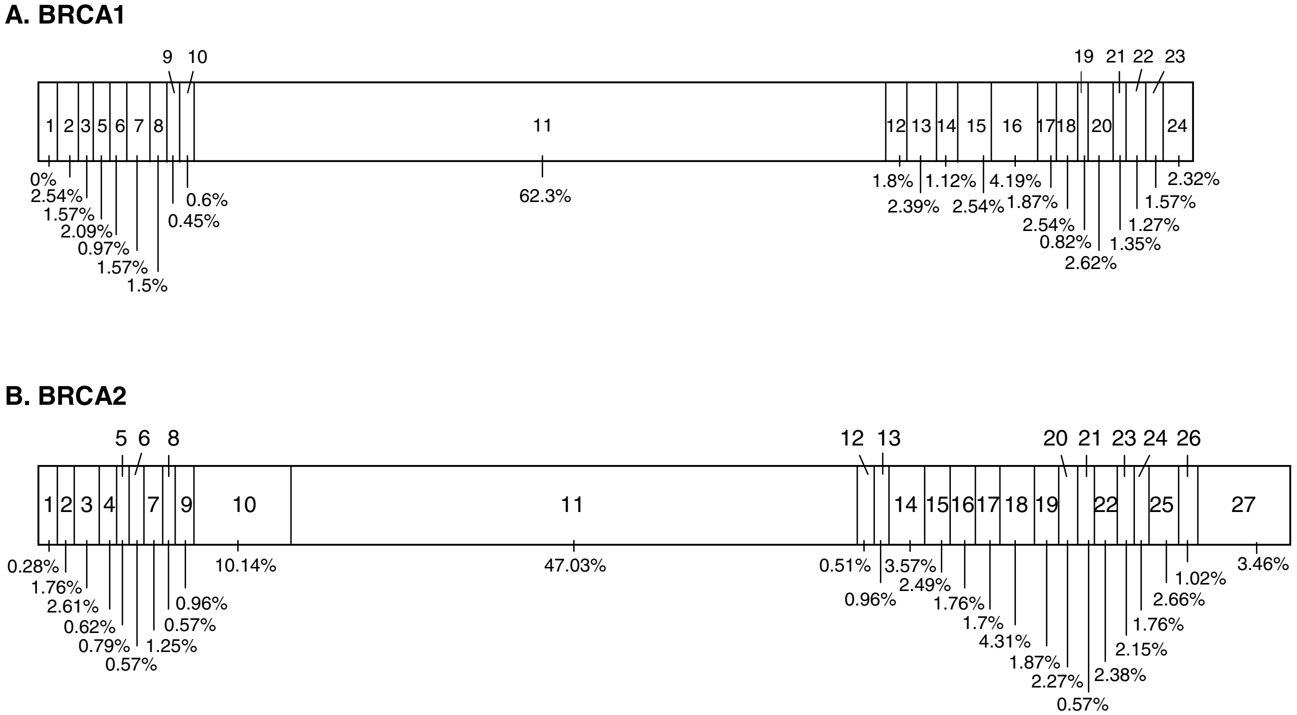

Fig. 1. Schematic of the distribution of mutations collected in the dataset for A, *BRCA1* and B, *BRCA2* genes. The mRNA sequence is drawn to scale for individual gene and the exons are represented as numbers that are either inside or above the boxes. Percentages of mutations in the dataset for each exon are shown below the boxes.

2.4. Statistical analysis

To show the significance of the difference between *p* and *q*, significance test for comparing proportions was performed. $P < 0.05$ was considered statistically significant.

3. Results

3.1. Mutation spectra of *BRCA1* and *BRCA2*

Fig. 1 is the schematic showing the distribution of mutations found in the two genes. The frequency of mutations increased with exon size, where 62.3% and 47.02% of all mutations were located in exon 11 of *BRCA1* and *BRCA2*, respectively (Fig. 1). Exon 4 of *BRCA1* was omitted as it was found to be an Alu element and cloning artifact during isolation. This exon is rarely expressed and its insertion introduces a premature stop codon [15].

3.2. Local DNA-sequence environment flanking short deletions and insertions

In an attempt to discern the nucleotide sequences flanking each mutation, nucleotide context around the mutation site was examined. Tables 1 and 2 show representative data randomly chosen from our analyses, with classified mutations for each specific sequence in the *BRCA1* and *BRCA2* genes, respectively. Table 3 shows the summarized data from our analysis. Our results showed that 86% deletions in *BRCA1* were associated with HO. Short repeats were predominant features in the sequence environment surrounding *BRCA1* small deletions (71% for R(0), 85% for R(1), 90% for R(2)). Seventy-two percent insertions occurred at the site of HO while 45% (R(0)), 67% (R(1)), 77% (R(2)) insertions were found at short repeats.

There were 465 deletions and 165 insertions in the dataset of the *BRCA2* gene. Eighty-four percent deletions and 86% insertions were found in HO. The frequencies of short repeats at deletion were 67%, 87% and 93%, and at insertions were 49%, 66% and 76% for R(0), R(1), and R(2), respectively.

There was no difference in the frequency of mutations per specific nucleotide sequence among the various exons in both the *BRCA1* and *BRCA2* genes (see Tables S1 and S2 in the Supplementary material).

3.3. Local DNA-sequence environment flanking substitutions

Sixty-seven percent of *BRCA1* substitutions were associated with HO. The percentage of short repeats were 66%, 87% and 92% for R(0), R(1) and R(2), respectively. In contrast, only 22% substitutions were detected in CpG/CpNpG motifs (Table 3). In addition, we also observed a preponderance of transitions (58%) over transversions (42%) (data not shown). Most of these transitions could be attributed to the hypermutability of the CpG dinucleotide to TG or CA. This type of transition mutations accounts for 34.2% of all substitutions and for 58.9% of transitions.

Table 1
Representative substitutions, deletions and insertions in the *BRCA1* gene

Exon	Mutation[a]	DNA sequence[a]	Predicted A.A change	Phenotype
A. Substitutions[b]				
(i) Substitutions associated with homonucleotides				
5	G331A	C**AAAA**GGAGCCT	Arg-Lys (R71K)	Breast and/or ovarian cancer
11	A1260T	TTCAG**AAA**GTTA	Lys-Term (K381X)	Breast cancer
(ii) Substitutions associated with repeats				
5	G259A	TT**TGC**A**TGC**TG	Cys-Tyr (C47Y)	Breast and/or ovarian cancer
7	A433G	ACAGC**TATA**AT	Tyr-Cys (Y105C)	Breast cancer
(iii) Substitutions associated with CpG/CpNpG motifs				
11	C1740T	CGGAG**CAG**AAT	Gln-Term (Q541X)	Breast cancer
11	C2715T	AAAG**CG**CCAG	Arg-Cys (R866C)	Breast and/or ovarian cancer
B. Deletions[c]				
(i) Deletions associated with homonucleotides				
11	916gelTT	GTTCTG**TTT**CAA		Breast cancer
16	5063delA	CAAC**AAAA**GAAT		Ovarian cancer
(ii) Deletions associated with repeats				
11	1477delAG	CAGT**AGAGAG**TAA		Ovarian cancer
12	4286delT	TCAG**AG**T**GA**CATT		Breast and/or ovarian cancer
C. Insertions[d]				
(i) Insertions associated with homonucleotides				
8	613insT	AACTC**TT**GAGG		Breast and/or ovarian cancer
11	3376insT	TAGA**TTT**AGGG		Ovarian cancer
(ii) Insertions associated with repeats				
11	3331insG	AGCA**GAGA**CTAG		Breast cancer
11	3768insA	ACT**TATA**CTAGT		Ovarian cancer

See the HGMD database for corresponding references.

[a] Mutation nomenclature is according to GenBank accession number U14680.

[b] The substituted nucleotide is underlined and the association with the specific sequence motifs are highlighted as bold.

[c] The deleted nucleotide is underlined and the association with the specific sequence motifs are highlighted as bold.

[d] The inserted nucleotide is underlined and the association with the specific sequence motifs are highlighted as bold.

Table 2
Representative substitutions, deletions and insertions in the *BRCA2* gene

Exon	Mutation [a]	DNA sequence[a]	Predicted A.A change	Phenotype
A. Substitutions[b]				
(i) Substitutions associated with homonucleotides				
3	G361T	TAATTCT**G**AACCTG	Glu-Term	Breast and/or ovarian cancer
11	A3058T	CAATT**A**AAAAAGA	Lys-Term	Breast cancer
(ii) Substitutions associated with repeats				
7	G809A	TATGTCT**TGGT**CAAGT	Trp-Term	Breast cancer
10	T1742C	AAAGTCT**ATAT**TCAG	Ile-Thr	Breast cancer
(iii) Substitutions associated with CpG/CpNpG motifs				
11	C6187T	GTC**CAG**GTATCAG	Gln-Term	Breast cancer
15	C7708T	TATG**CG**AATTAAGA	Arg-Term	Breast cancer
B. Deletions[c]				
(i) Deletions associated with homonucleotides				
3	432delA	AAGG**AAA**CCATCT		Breast and/or ovarian cancer
11	4084delAAA	GTG**AAAAAAA**TAAT		Breast cancer
(ii) Deletions associated with repeats				
2	277delAC	TTTAAG**ACAC**GCTG		Breast cancer
10	1900delA	AATT**TAAT**TGATA		Breast cancer
C. Insertions[d]				
(i) Insertions associated with homonucleotides				
11	3979insA	GG**AAAA**CTTCTGCA		Ovarian cancer
18	8299insTT	GTG**TTTTT**CTGACAT		Breast and/or ovarian cancer
(ii) Insertions associated with repeats				
7	767insAT	AAC**ATATAT**TTCTGAA		Breast and/or ovarian cancer
19	8664insA	GATG**GAAG**GAAAT		Breast cancer

See the HGMD database for corresponding references.
[a] Mutation nomenclature is according to GenBank accession number U43746.
[b] The substituted nucleotide is underlined and the association with the specific sequence motifs are highlighted as bold.
[c] The deleted nucleotide is underlined and the association with the specific sequence motifs are highlighted as bold.
[d] The inserted nucleotide is underlined and the association with the specific sequence motifs are highlighted as bold.

Table 3
Frequency of homonucleotides, short repeats and CpG/CpNpG motifs associated with deletions,insertions and substitutions in *BRCA1* and *BRCA2* genes

Mutation type		HO	R(0)	R(1)	R(2)	CpG/CpNpG	Total
Substitutions	*BRCA1*	515 (67%)	503 (66%)	666 (87%)	704 (92%)	168 (22%)	764
	BRCA2	780 (69%)	782 (69%)	970 (85%)	1039 (92%)	225 (20%)	1135
Deletions	*BRCA1*	356 (86%)	293 (71%)	351 (85%)	375 (90%)	NA	415
	BRCA2	391 (84%)	312 (67%)	403 (87%)	431 (93%)	NA	465
Insertions	*BRCA1*	114 (72%)	71 (45%)	106 (67%)	121 (77%)	NA	158
	BRCA2	142 (86%)	81 (49%)	109 (66%)	125 (76%)	NA	165

NA, not applicable.

Similar to *BRCA1*, majority of substitutions (69%) in the *BRCA2* gene were associated with HO. The observed percentage of repeats were 69% for R(0), 85% for R(1) and 92% for R(2). Twenty percent were detected in CpG/CpNpG motifs (Table 3). An excess of transitional nucleotide changes was also observed in the *BRCA2* gene, with 61% transition over 39% transversion (data not shown).

Taken together, these observations showed that homonucleotides and repeats were most frequent for deletions/insertions and substitutions, whereas CpG/CpNpG motifs only had a moderate correlation. Although these results are consistent with many previous reports [9–11] based on percentage counts, they may not reflect the true underlying tendency of mutation occurrence since this tendency can be affected by a greater number of these sequences within the genes.

3.4. Simulations of mutations

To better define whether there was bias of mutation events towards certain specific sequences, we employed a computer simulation model which generated uneven distribution of mutations by incorporation of mutation weighting α. For each combination of mutation weighting α and mutation type, 200000 samples were generated and the proportion of simulated mutations associated with the specific sequences was

Table 4
Results of simulated mutations with varied mutation weighting α of the *BRCA1* gene

	Mutation weighting α	HO	R(0)	R(1)	R(2)	CpG/ CpNpG
SUB (764)	1	0.78	0.68	0.86	0.92	0.12
	10	0.75	0.68	0.87	0.92	0.17
	20	0.74	0.68	0.87	0.92	0.19
	30	0.73	0.68	0.87	0.92	0.20
	Chance in real dataset	0.67	0.66	0.87	0.92	0.22
DEL (415)	1	0.86	0.72	0.87	0.93	N.A
	10	0.85	0.72	0.87	0.93	N.A
	20	0.85	0.72	0.87	0.92	N.A
	30	0.85	0.72	0.87	0.93	N.A
	Chance in real dataset	0.86	0.71	0.85	0.90	N.A
INS (158)	1	0.64	0.52	0.74	0.84	N.A
	10	0.64	0.51	0.73	0.83	N.A
	20	0.63	0.50	0.72	0.82	N.A
	30	0.63	0.49	0.72	0.82	N.A
	Chance in real dataset	0.72	0.45	0.67	0.77	N.A

NA, not applicable.

calculated. Tables 4 and 5 give the proportion of mutations with increasing mutation weighting α for *BRCA1* and *BRCA2*, respectively. The chance of occurrence increases with the value and "1" represents maximal chance. An increasing trend of substitutions associated with CpG/CpNpG motifs could be clearly observed when the mutation weighting α was gradually increased for both the *BRCA1* gene (1.8-fold) and *BRCA2* gene (2-fold). This result suggests that these motifs are preferred sites for substitutions. It is noteworthy that there was also a moderate decreased chance of substitutions associated with the HO motifs. On the contrary, there was no significant change in the numbers of events in deletions and insertions associating with any specific sequences. The simulated counts in these cases remained steady even when the mutation weighting α increased.

Table 5
Results of simulated mutations with varied mutation weighting α of the *BRCA2* gene

	Mutation weighting α	HO	R(0)	R(1)	R(2)	CpG/ CpNpG
SUB (1135)	1	0.78	0.68	0.86	0.92	0.10
	10	0.76	0.69	0.86	0.92	0.14
	20	0.75	0.69	0.86	0.92	0.16
	30	0.75	0.69	0.86	0.92	0.17
	Chance in real dataset	0.69	0.69	0.85	0.92	0.20
DEL (465)	1	0.88	0.72	0.87	0.93	NA
	10	0.87	0.71	0.87	0.93	NA
	20	0.86	0.71	0.87	0.93	NA
	30	0.86	0.71	0.87	0.93	NA
	Chance in real dataset	0.84	0.67	0.87	0.93	NA
INS (165)	1	0.67	0.51	0.72	0.82	NA
	10	0.67	0.51	0.72	0.82	NA
	20	0.67	0.51	0.72	0.82	NA
	30	0.67	0.50	0.72	0.82	NA
	Chance in real dataset	0.86	0.49	0.66	0.76	NA

NA, not applicable.

Table 6
The probabilities p and q calculated from simulated data

			HO	R(0)	R(1)	R(2)	CpG/CpNpG
SUB	*BRCA1*	p	0.72[a]	0.68	0.87	0.92	0.22[a]
		q	0.79[a]	0.68	0.87	0.92	0.11[a]
	BRCA2	p	0.73[a]	0.69	0.86	0.92	0.19[a]
		q	0.79[a]	0.68	0.86	0.92	0.09[a]
DEL	*BRCA1*	p	0.85	0.72	0.87	0.92	NA
		q	0.85	0.72	0.88	0.93	NA
	BRCA2	p	0.83	0.70	0.87	0.93	NA
		q	0.88	0.72	0.87	0.93	NA
INS	*BRCA1*	p	0.61	0.46	0.69	0.79	NA
		q	0.64	0.52	0.74	0.84	NA
	BRCA2	p	0.66	0.49	0.70	0.80	NA
		q	0.67	0.51	0.72	0.82	NA

NA, not applicable.
[a]Significant difference between p and q.

3.5. Demonstration of relationship by mathematical models

To further ascertain the relationship among simulated mutations and surrounding nucleotide environment, we employed mathematical equations and compared the value of p and q obtained (Table 6), which can reflect the likelihood of mutational bias towards surrounding nucleotides compositions. To show the significance of the difference between p and q, significance test for comparing proportions was performed. We noticed that in both *BRCA1* and *BRCA2*, p was significantly larger than q in CpG/CpNpG ($P < 0.005$). This suggested that CpG/CpNpG occurred more frequently than expected. On the other hand, p was smaller than q in HO ($P < 0.005$). No significant difference could be seen between the p and q values for the other mutation types, indicating essentially no bias between these mutation events and the neighboring nucleotides.

4. Discussion

BRCA1/*BRCA2* gene mutations are extensively studied because of their importance in breast and/or ovarian cancer. However, the mechanism(s) responsible for their mutagenesis is largely unknown. Although a previous study has assessed the significance of specific sequences at mutational sites in *BRCA1* [10], the present study distinguishes itself in several key aspects. First, the sample size of our datasets is greatly expanded that allows statistically robust analysis. Our evaluation was based on 1337 independent *BRCA1* mutations compared with 74 cases from the previous study [10]. Second, to our knowledge, this is the first study to analyze the nucleotides sequences flanking mutations of the *BRCA2* gene. Third, we adopted a stricter set of rules for classifying the mutations. Finally, while the previous study was based on percentage counts, we employed computational method with simulation model to test whether these results reflect the true underlying tendency of mutation occurrence.

One major finding of this study is that although there was only a moderate correlation between substitutions and CpG/CpNpG motifs, these mutational events were statistically biased to the CpG/CpNpG motifs of both the *BRCA1* and

BRCA2 genes. The hypermutability of CpG/CpNpG motifs has largely been attributed to spontaneous deamination of 5-methylcytosines to thymine. We also observed a preponderance of transitions over transversions, suggesting a strong bias towards TG or CA deamination products in the genes. These data are coincident with in vitro evidence demonstrating DNA methylation in the *BRCA1* gene as a possible mechanism to generate mutations at CpG/CpNpG motifs [10]. In fact, exogenous chemical carcinogens have also been found to be causative factors in methylation-mediated mutagenesis of the CpG/CpNpG motifs [16,17], reflecting the motifs as mutational hotspot causing human genetic disease and cancer [18–20]. For example, mutations within CpG/CpNpG sites represent 45–50% of p53 point mutations in colorectal cancer [21], 20–25% of all Rb mutations [22] and 13–17% of all APC mutations in colon cancer [23].

Since it has previously been reported that sequence slippage, which occur in homonucleotides and short repeats, as another possible means to generate substitution mutations [24,25], we also assessed the frequency of mutations occurring at these particular sequence elements. We found a downward (negative) mutation bias for substitutions at homonucleotides/repeats was noted in the *BRCA1* and *BRCA2* genes, suggesting that homonucleotides and short repeats are not the major contributor to the generation of substitutions in the *BRCA1* and *BRCA2* genes. One explanation might be that the mutational bias towards next neighbors is conditional. The mutability of a nucleotide has been found to be different within different combinations of specific dinucleotides [26]. Of 16 combinations, only half of them were shown to be favorable for substitutions to occur [26]. This selectivity might be explicable by reading frame-sensitive DNA-repair bias and DNA thermodynamic stability, which in turn introduces mutation bias only at specific codon position and DNA strand [13]. There is also contradictory evidence showing only subtle contribution of repetitive sequences surrounding substitution sites [13].

Apart from substitutions, the vast majority of *BRCA1* and *BRCA2* mutations comprise short deletions and insertions. It has been suggested that these two types of mutation share very similar generative mechanisms mediated by homonucleotides and repeats [27]. Although, we observed homonucleotides and repeats were commonly associated with deletions and insertions, the computer simulation with mathematical analysis indicated that *BRCA1* and *BRCA2* deletions and insertions were not caused by a general bias towards homonucleotides/repeats. We obtained the same conclusion even if we have taken into account of other factors. For example, we have considered other possibilities of repeats with more than two nucleotides in-between (R(n); $n > 2$) [28]. As n increased (e.g. $n = 3$, 4, and 5), there was an increase in mutations associated with repeat elements (95%, 96%, and 97%) but no difference was seen between the p and q values ($p = q = 0.95$, 0.96, 0.97). Moreover, artifacts may appear with inclusion of the dinucleotide repeat due to the abundance and high error frequency of this repeat unit in one or more repeat lengths. To address this problem, we repeated our analysis excluding the dinucleotide repeats. Again, this did not alter the results obtained (see Table S3 in the supplementary material). Thus, our results suggest that the mutagenesis might be due to a previously unrecognized or less common mechanism. Slippage repair may not be the sole mechanism by which deletions and insertions occur. In fact, it has previously been shown that sequence slippage accounts for only about 30–50% of the observed mutations [29]. However, despite the importance of this information, little is known about the alternative mechanisms of mutation. Quasi-palindromic sequences [30,31], symmetrics elements [27], and "knot" (inversion of inverted repeats) [32] have been shown to be involved in the mechanism of mutagenesis in certain circumstances.

In summary, this study significantly extends the analysis of BRCA mutations over existing data and methods to allow comprehensive and unbiased identification of specific nucleotide sequences that are associated with BRCA mutagenesis. The great advantage of the new algorithm used in this study is that it allows both the observed and selective frequencies of mutational events to be considered. Therefore, it can be argued against the possibility that mutations have occurred in a specific nucleotide sequence simply by chance due to that sequence relative abundance in the gene rather than because of a causal connection. In our analyses, we have not only offered additional support for the known patterns, which indicates that our strategy is a valid approach, but also unraveled new (or unusual) mechanisms that could have been overlooked by traditional methods. Further studies on the precise mechanisms are undoubtedly an important area of investigation. The new algorithm presented is general and may present a new paradigm for understanding the mutational events for other genes. Our results also have significant implications for genetic screening. At present, due to the large size of the gene, the extremely heterogeneous mutations, and the lack of high efficiency screening methods, the molecular diagnosis for *BRCA1* and *BRCA2* mutations has been onerous. Since our data show that CpG/CpNpG motifs are significantly associated with *BRCA1* and *BRCA2* mutations, the search for an unknown mutation could be predicted to some degree based on the analysis of such motifs within the sequence.

Acknowledgements: This work was supported by Hong Kong Government RGC Grants HKU7484/04M and Science Faculty Collaborative Seed Grant (to A.S.T.W).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2007.08.061.

References

[1] Lux, M.P., Fasching, P.A. and Beckmann, M.W. (2006) Hereditary breast and ovarian cancer: review and future perspectives. J. Mol. Med. 84, 16–28.

[2] Chen, Y., Farmer, A.A., Chen, C.F., Jones, D.C., Chen, P.L. and Lee, W.H. (1996) BRCA1 is a 220-kDa nuclear phosphoprotein that is expressed and phosphorylated in a cell cycle-dependent manner. Cancer Res. 56, 3168–3172.

[3] Bertwistle, D., Swift, S., Marston, N.J., Jackson, L.E., Crossland, S., Crompton, M.R., Marshall, C.J. and Ashworth, A. (1997) Nuclear location and cell cycle regulation of the BRCA2 protein. Cancer Res. 57, 5485–5488.

[4] Lander, E.S. et al. (2001) Initial sequencing and analysis of the human genome. Nature 409, 860–921.

[5] Jego, N., Thomas, G. and Hamelin, R. (1993) Short direct repeats flanking deletions, and duplicating insertions in p53 gene in human cancers. Oncogene 8, 209–213.

[6] Redston, M.S., Caldas, C., Seymour, A.B., Hruban, R.H., da Costa, L., Yeo, C.J. and Kern, S.E. (1994) p53 mutations in pancreatic carcinoma and evidence of common involvement of homocopolymer tracts in DNA microdeletions. Cancer Res. 54, 3025–3033.

[7] Canning, S. and Dryja, T.P. (1989) Short, direct repeats at the breakpoints of deletions of the retinoblastoma gene. Proc. Natl. Acad. Sci. USA 86, 5044–5048.

[8] Mancini, D., Singh, S., Ainsworth, P. and Rodenhiser, D. (1997) Constitutively methylated CpG dinucleotides as mutation hot spots in the retinoblastoma gene (RB1). Am. J. Hum. Genet. 61, 80–87.

[9] Rodenhiser, D.I., Andrews, J.D., Mancini, D.N., Jung, J.H. and Singh, S.M. (1997) Homonucleotide tracts, short repeats and CpG/CpNpG motifs are frequent sites for heterogeneous mutations in the neurofibromatosis type 1 (NF1) tumour-suppressor gene. Mutat. Res. 373, 185–195.

[10] 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.

[11] Kondrashov, A.S. and Rogozin, I.B. (2004) Context of deletions and insertions in human coding sequences. Hum. Mutat. 23, 177–185.

[12] Schmucker, B. and Krawczak, M. (1997) Meiotic microdeletion breakpoints in the BRCA1 gene are significantly associated with symmetric DNA-sequence elements. Am. J. Hum. Genet. 61, 1454–1456.

[13] Krawczak, M., Ball, E.V. and Cooper, D.N. (1998) Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am. J. Hum. Genet. 63, 474–488.

[14] Todorova, A. and Danieli, G.A. (1997) Large majority of single-nucleotide mutations along the dystrophin gene can be explained by more than one mechanism of mutagenesis. Hum. Mutat. 9, 537–547.

[15] 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.

[16] Hu, W., Feng, Z. and Tang, M.S. (2003) Preferential carcinogen-DNA adduct formation at codons 12 and 14 in the human K-ras gene and their possible mechanisms. Biochemistry 42, 10012–10023.

[17] Denissenko, M.F., Chen, J.X., Tang, M.S. and Pfeifer, G.P. (1997) Cytosine methylation determines hot spots of DNA damage in the human P53 gene. Proc. Natl. Acad. Sci. USA 94, 3893–3898.

[18] Skandalis, A., Ford, B.N. and Glickman, B.W. (1994) Strand bias in mutation involving 5-methylcytosine deamination in the human hprt gene. Mutat. Res. 314, 21–26.

[19] Cooper, D.N. and Youssoufian, H. (1988) The CpG dinucleotide and human genetic disease. Hum. Genet. 78, 151–155.

[20] Ollila, J., Lappalainen, I. and Vihinen, M. (1996) Sequence specificity in CpG mutation hotspots. FEBS Lett. 396, 119–122.

[21] Greenblatt, M.S., Bennett, W.P., Hollstein, M. and Harris, C.C. (1994) Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. Cancer Res. 54, 4855–4878.

[22] Cowell, J.K., Smith, T. and Bia, B. (1994) Frequent constitutional C to T mutations in CGA-arginine codons in the RB1 gene produce premature stop codons in patients with bilateral (hereditary) retinoblastoma. Eur. J. Hum. Genet. 2, 281–290.

[23] Nagase, H. and Nakamura, Y. (1993) Mutations of the APC (adenomatous polyposis coli) gene. Hum. Mutat. 2, 425–434.

[24] Kunkel, T.A. (1990) Misalignment-mediated DNA synthesis errors. Biochemistry 29, 8003–80011.

[25] Meuth, M. (1989) The molecular basis of mutations induced by deoxyribonucleoside triphosphate pool imbalances in mammalian cells. Exp. Cell Res. 181, 305–316.

[26] Cooper, D.N. and Krawczak, M. (1990) The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. Hum. Genet. 85, 55–74.

[27] Cooper, D.N. and Krawczak, M. (1991) Mechanisms of insertional mutagenesis in human genes causing genetic disease. Hum. Genet. 87, 409–415.

[28] Krawczak, M. and Cooper, D.N. (1991) Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. Hum. Genet. 86, 425–441.

[29] Taylor, M.S., Ponting, C.P. and Copley, R.R. (2004) Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. Genome Res. 14, 555–566.

[30] Wang, F.J. and Ripley, L.S. (1994) DNA sequence effects on single base deletions arising during DNA polymerization in vitro by *Escherichia coli* Klenow fragment polymerase. Genetics 136, 709–719.

[31] Ripley, L.S. (1982) Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. Proc. Natl. Acad. Sci. USA 79, 4128–4132.

[32] Chuzhanova, N.A., Anassis, E.J., Ball, E.V., Krawczak, M. and Cooper, D.N. (2003) Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum. Mutat. 21, 28–44.