ORIGINAL ARTICLE

Somatic instability of the DNA sequences encoding the polymorphic polyglutamine tract of the *AIB1* gene

P Dai, L-J C Wong

J Med Genet 2003;40:885–890

See end of article for
authors' affiliations

Correspondence to:
Dr L-J C Wong,
Department of Oncology,
Georgetown University
Medical Center, M4000,
3800 Reservoir Rd, NW,
Washington, DC 20007;
wonglj@georgetown.edu

Background: *AIB1* contains a polymorphic polyglutamine tract (poly Q) that is encoded by a trinucleotide CAG repeat. Previously there have been conflicting results regarding the effect of the poly Q tract length on breast cancer. Since poly Q is not encoded by a perfect CAG repeat, the heterozygous polymorphic alleles need to be resolved, to understand the exact DNA sequences encoding poly Q.

Methods: Poly Q encoding sequences of *AIB1* from 107 DNA samples, including breast cancer cell lines, sporadic primary breast tumours, and blood samples from *BRCA1/BRCA2* mutation carriers and the general population, were resolved by PCR/cloning followed by sequencing of each individual clone.

Results: 25 distinct poly Q encoding sequence patterns were found. More than two distinct sequence patterns were found in a significantly higher proportion of tumours and cell lines than that of the general population, suggesting somatic instability. A significantly higher proportion of cancer cell lines or primary breast tumours than that of the general population contained rare sequence patterns. The proportion of sporadic breast tumours having at least one allele ≤ 27 repeats is significantly higher than that in the blood of *BRCA1/BRCA2* mutation carrier breast cancer patients or the general population.

Conclusion: The poly Q encoding DNA sequences are somatically unstable in tumour tissues and cell lines. A missense mutation and a very short glutamine repeat in primary tumours suggests that *AIB1* activity may be modulated through poly Q, which in turn plays a role in the cotransactivation of gene expressions in breast cancers.

The amplified in breast cancer gene 1 (*AIB1*; Genbank Accession number: AL034418), also known as RAC3, TRAM-1, or ACTR, belongs to a family of nuclear receptor co-activators that include the steroid receptor co-activator SRC-1 and transcription intermediate factor TIF-2.^{1–4} These molecules directly bind to nuclear hormone receptors and stimulate hormone dependent transcriptional activation of the genes that have been implicated in the regulation of cell growth, development, differentiation, and homeostasis.^{2,4} The *AIB1* gene was found to be amplified and over expressed in breast and ovarian cancer cell lines and in about 0–10% of breast cancer biopsies.^{5–6} The *AIB1* protein interacts with the oestrogen receptor (OR) and enhances OR dependent transcription. Several characteristics of *AIB1* are similar to the androgen receptor (AR). For example, both *AIB1* and AR contain a polymorphic polyglutamine tract (poly Q). Both genes are involved in nuclear receptor mediated transactivation of gene expression. *AIB1* is amplified in breast tumours,^{5–6} and AR is amplified in prostate cancer.^{7–8} These observations are consistent with the hypothesis that over expression of the gene may confer a proliferative advantage during hormone deprivation and contribute to the development of cancer recurrence.⁸ Despite these similarities, the functional significance of the poly Q tract of *AIB1* in transactivation of gene expression and its role in the development of breast cancer has not been established.

The expansion of CAG repeats in proteins containing poly Q underlies a number of neurodegenerative diseases.⁹ The molecular mechanism of these proteins in these diseases is thought to be the formation of insoluble aggregates due to misfolded proteins that confer a gain in function. These novel toxic protein aggregates are ubiquitinated and found in neuronal intranuclear inclusions, leading to apoptosis and neurodegeneration.^{9–12} The poly Q tract in the AR gene is unique in that the large expansion of poly Q encoding CAG

repeat causes X-linked spinal bulbar muscular atrophy (SBMA, or Kennedy disease),¹³ and the polymorphic poly Q of AR also plays an important role in the hormone-dependent transactivation.^{14–15} In transfection experiments, the length of the CAG repeats inversely correlates with transcriptional activity of AR. Longer poly Q repeats in AR are associated with lower transcriptional activity.^{14–15} In contrast, a shorter CAG repeat in AR is associated with a higher risk of an aggressive prostate cancer phenotype characterised by extra-prostatic extension, distant metastases, or poor histological grade.¹⁶ Haiman *et al.* assessed the association between the poly Q repeat polymorphism in *AIB1* gene and breast cancer risk.¹⁷ They found that poly Q repeat genotype did not influence postmenopausal breast cancer risk among white women in the general population. However, a matched case control study of 448 *BRCA1/BRCA2* mutation carriers found that women with at least 28 poly Q repeats in *AIB1* had a higher risk of contracting breast cancer when compared to women who carried alleles with fewer poly Q repeats.¹⁸ These studies measured the poly Q repeat size by fragment length analysis without analysing the encoding DNA sequence. Our approach to understanding the association of polymorphic poly Q tract of the *AIB1* gene with breast cancer was to dissect the heterogeneous poly Q encoding sequences of individual alleles by cloning and sequencing. Our data demonstrates that the poly Q tract of *AIB1* is not only polymorphic in repeat length, but also polymorphic in its

Abbreviations: ACTR, activator of retinoid and thyroid receptors; *AIB1*, amplified in breast cancer gene 1; AR, androgen receptor; OR, oestrogen receptor; FAM, 6-carboxyfluorescein; poly Q, polymorphic polyglutamine tract; RAC3, receptor associated co-activator 3; TAMRA, 6-carboxytetramethyl rhodamine; TRAM-1, thyroid hormone receptor activator molecule 1

encoding sequence. The triplet CAG repeat is somatically unstable in cell lines and primary tumours despite the frequent interruption by CAA.

MATERIALS AND METHODS

Samples and DNA preparation

A total of 107 DNA samples taken from 16 breast cancer cell lines (group A), 32 sporadic primary breast tumours (group B), 16 blood samples from familial breast cancer patients carrying *BRCA1/BRCA2* mutation (group C), and 43 blood specimens from normal individuals in the general population (group D), were included in this study. The 16 cell lines used were MDA-MB157, BT20, BT474, BT549, T47D, MCF7-P19, HBL100, MDA-MB436, MDA-MB134V, MDA-MB231N, MDA-MB435, MDA-MB361, MDA-MB468, ZR75-1, ZR75-30, BT483. The age of the 32 patients with sporadic breast cancer ranged from 26 to 81 years with a mean of 56.0 (14.3) years. The age of the 16 breast cancer patients carrying *BRCA1/BRCA2* mutations ranged from 35 to 76 years with a mean of 48.5 (12.3) years. The age of women from the general population ranged from 23 to 75 years with a mean of 54.5 (13.8) years. The B, C and D groups were matched for age, sex, and ethnicity (all are white), to enable comparison and statistical analysis. The tumour tissues were frozen in liquid nitrogen immediately after surgery and stored at -80°C until analysis. DNA was extracted from tumour tissues by proteinase K digestion, phenol chloroform extraction, and ethanol precipitation. DNA from blood lymphocytes and cell culture was extracted by the salting out method.¹⁹

Cloning and sequencing

The fragment containing poly Q was amplified by the forward primer F: 5' GTCTTATACTGGTGTATTG 3' and the reverse primer R: 5' CTGGGGGAAGCAGTCACATTAG 3', yielding a PCR product of 314 bp. The high fidelity amplification was carried out in a 30 μl reaction mixture containing 10 ng of genomic DNA, 0.2 μM of each primer, 1 $\times$ HF 2 PCR buffer, dNTP, and Advantage-HF 2 polymerase according to the manufacturer's recommendation (Clontech). After 1 min of initial denaturation at 94°C , the DNA was amplified with 30 cycles of 45 s at 95°C , 45 s at 55°C and 45 s at 72°C , followed by a final extension at 72°C for 5 min. The PCR products were purified and cloned into pCR2.1-TOPO (TOP10) vector (Invitrogen, Carlsbad, CA 92008) according to the manufacturer's protocol. At least six clones from each sample were picked for sequencing using BigDye termination sequencing kit and analysed on an ABI 377 DNA Sequencer (Perkin-Elmer, Foster City, CA, USA). Two primers, F and F2 (5' AGCAGGGTTTCTTAATGCTC 3') were used in sequencing for loading onto alternate lanes for easy tracking. The sequence results were analysed using sequence analysis software version 3.4.

Statistical analysis

Two sided Fisher's exact test was used to analyse the difference in allele distributions between the groups of samples.

RESULTS

Sequence patterns encoding Poly Q

Due to the heterozygous feature of *AIB1*, direct sequencing of the region encoding poly Q often produced overlapping sequences without clear distinction between the two heterozygous alleles. Cloning the PCR amplified region followed by sequencing of each individual clone allowed the exact poly Q encoding sequences to be determined. A total of 107 DNA samples and at least six single colonies per sample were analysed. Among the 697 clones that showed a clear single allele sequence, 25 distinct sequence patterns were identified

(table 1). The size of poly Q repeats ranged from 17 to 29. The 17-repeat allele was from a primary tumour. The CAG repeat encoding the poly Q was frequently interrupted by CAA. Most commonly, there were a total of six CAA interruptions. The longest CAG stretch was 11 repeats. Of the 25 distinct sequence patterns, the four occurring more than 10 times were categorised as common sequence patterns, those that occurred less than 10 times were categorised as uncommon sequence patterns. Although there are substantial variations in the poly Q encoding sequence, certain characteristics exist. The poly Q encoding sequence is composed of three major parts. The first part is $(\text{CAG})_{3-6}\text{CAA}$ or $(\text{CAG})_3\text{CAA}(\text{CAG})_2\text{CAA}$, with $(\text{CAG})_6$, $(\text{CAG})_3$ and $(\text{CAG})_3\text{CAA}(\text{CAG})_2$ most common and always ends with CAA, except the one with the 17 repeats. The variations in this part include $(\text{CAG})_5\text{CAA}$, $(\text{CAG})_5(\text{CAA})_2$, and $(\text{CAG})_5\text{CAACAGCAA}$. The second part is $(\text{CAG})_{7-11}$, which is present in all the sequences except the one with 17 repeats where this part is completely lost. Finally, the last part is $(\text{CAACAG})_{2-3}(\text{CAACAGCAG})_2\text{CAA}$. The only variations in this part are $(\text{CAACAG})(\text{CAACAGCAG})_2\text{CAA}$ and $(\text{CAACAG})_4(\text{CAACAGCAG})_2\text{CAA}$. Unlike the CAG repeat expansion in the neurodegenerative diseases, the poly Q length in the *AIB1* gene remains quite stable, probably due to the frequent interruption by CAA.⁹ Both CAA and CAG encode for glutamine. Therefore, the poly Q repeats of the same size may be encoded by different sequences at the DNA level. A CGG was found in a primary tumour with a rare sequence of $(\text{CAG})_5\text{CGGCAA}(\text{CAG})_8(\text{CAACAG})_3(\text{CAACAGCAG})_2\text{CAA}$ (pattern #23). This primary tumour contained four distinct poly Q encoding sequences. The most prevalent sequence was $(\text{CAG})_6\text{CAA}(\text{CAG})_8(\text{CAACAG})_3(\text{CAACAGCAG})_2\text{CAA}$, a common 28 repeats, (pattern #2 in table 1) (four out of seven clones). The other two uncommon sequence patterns were 27 repeats, one with five CAG at the beginning (pattern #18), and the other one with seven CAG instead of 8 in the middle (pattern #10). Each of these uncommon sequence patterns was found in one out of seven clones sequenced. These results suggest high instability in this tumour. CGG encodes for arginine. Apparently this is the result of a point mutation from CAG to CGG. The effect of the interruption by an arginine residue in a stretch of glutamine will be investigated.

Instability of Poly Q encoding sequences in primary tumours and cell lines

Previous reports using fragment length analysis did not address sequence instability because the appearance of the "extra" alleles might have been confused with stutter bands. Without the determination of the exact sequence, it would be very difficult to assign the "extra" alleles. In addition, the same poly Q repeat length may be encoded by different DNA sequences that cannot be resolved by DNA fragment length analysis. During our initial studies using DNA fragment length analysis, several samples had consistently shown "extra peaks" not typical of stutter bands. This observation was confirmed by direct cloning and sequencing of the individual encoding sequences.

Table 2 lists a few examples of the sequence patterns and somatic instability in tumours and cell lines. The breast cancer cell line, MCF7 passage 72, showed three poly Q encoding sequence patterns. The 26 repeat allele was the original sequence pattern and was the predominant one in the later passages. The 25 and 24 repeat sequence patterns appeared to be derived from the original 26 repeat allele by deletion. Different passages of the same cell line, for example, MCF7 passage 19, also had three sequence patterns (patterns 1, 3, and 18). These results are consistent with the somatic instability of the poly Q encoding sequences of *AIB1* in

Table 1 The sequence patterns encoding the poly Q tract					
Sample size		16	32	16	43
Common sequence patterns		(Gln)n	A*	B*	C* D*
1	(CAG) ₆ CAA(CAG) ₉ (CAACAG) ₃ (CAACAGCAG) ₂ CAA	29	11	13	8 26
2	(CAG) ₆ CAA(CAG) ₈ (CAACAG) ₃ (CAACAGCAG) ₂ CAA	28	8	20	11 20
3	(CAG) ₃ CAA(CAG) ₉ (CAACAG) ₃ (CAACAGCAG) ₂ CAA	26	4	10	3 14
4	(CAG) ₃ CAA(CAG) ₂ CAA(CAG) ₁₁ (CAACAG) ₂ (CAACAGCAG) ₂ CAA	29	3	5	1 7
Uncommon sequence patterns:					
5	(CAG) ₃ CAA(CAG) ₈ (CAACAG) ₃ (CAACAGCAG) ₂ CAA	25	1	6	1 1
6	(CAG) ₃ CAA(CAG) ₁₁ (CAACAG) ₃ (CAACAGCAG) ₂ CAA	28	—	—	— 1
7	(CAG) ₃ CAA(CAG) ₂ CAA(CAG) ₉ (CAACAG) ₂ (CAACAGCAG) ₂ CAA	27	—	1	— —
8	(CAG) ₃ CAA(CAG) ₂ CAA(CAG) ₁₀ (CAACAG) ₂ (CAACAGCAG) ₂ CAA	28	2	1	3 2
9	(CAG) ₃ CAA(CAG) ₂ CAA(CAG) ₁₁ (CAACAG)(CAACAGCAG) ₂ CAA	27	—	1	— —
10	(CAG) ₆ CAA(CAG) ₇ (CAACAG) ₃ (CAACAGCAG) ₂ CAA	27	2	3	— —
11	(CAG) ₆ CAA(CAG) ₇ (CAACAG) ₄ (CAACAGCAG) ₂ CAA	29	—	1	— —
12	(CAG) ₆ CAA(CAG) ₁₁ (CAACAG) ₂ (CAACAGCAG) ₂ CAA	29	—	1	— 1
13	(CAG) ₆ CAA(CAG) ₁₀ (CAACAG) ₂ (CAACAGCAG) ₂ CAA	28	1	3	— —
14	(CAG) ₆ CAA(CAG) ₉ (CAACAG) ₂ (CAACAGCAG) ₂ CAA	27	—	2	1 1
15	(CAG) ₆ CAA(CAG) ₈ (CAACAG) ₂ (CAACAGCAG) ₂ CAA	26	1	2	— —
16	(CAG) ₄ (CAACAG) ₂ (CAACAGCAG) ₂ CAA	17	—	1	— —
17	(CAG) ₅ CAA(CAG) ₉ (CAACAG) ₃ (CAACAGCAG) ₂ CAA	28	4	2	— 1
18	(CAG) ₅ CAA(CAG) ₈ (CAACAG) ₃ (CAACAGCAG) ₂ CAA	27	1	5	— 1
19	(CAG) ₅ CAA(CAG) ₇ (CAACAG) ₃ (CAACAGCAG) ₂ CAA	26	1	—	— —
20	(CAG) ₅ CAA(CAG) ₉ (CAACAG) ₂ (CAACAGCAG) ₂ CAA	26	—	1	— —
21	(CAG) ₅ (CAA) ₂ (CAG) ₈ (CAACAG) ₃ (CAACAGCAG) ₂ CAA	28	1	—	— —
22	(CAG) ₅ CAACAGCAA(CAG) ₇ (CAACAG) ₃ (CAACAGCAG) ₂ CAA	28	—	—	1 —
23	(CAG) ₅ CGGCAA(CAG) ₈ (CAACAG) ₃ (CAACAGCAG) ₂ CAA	28	—	1	— —
24†	(CAG) ₃ CAA(CAG) ₂ CAA(CAG) ₁₀ CAA(CAACAG) ₂ (CAACAGCAG) ₂ CAA	29	—	1	— —
25†	(CAG) ₃ CAA(CAG) ₂ CAA(CAG) ₇ (CAACAG) ₂ (CAACAGCAG) ₂ CAA	25	—	1	— —
Total number of distinct sequence patterns		13	21	8	11
*Group A, breast tumour cell lines; B, primary breast tumours; C, blood DNA from patients carrying mutations in BRCA1/BRCA2 genes; D, general population					
†the sequence patterns from the surrounding tissue of breast cancer.					

tumour cell lines. Different cell lines also revealed different degrees of instability and somatic alterations in the poly Q encoding region. MDA-MB-157 had six different sequence patterns (patterns 2, 10, 21, 19, 4, and 8). This cell line was known to have a high degree of genomic instability.

“Extra” sequence patterns were observed in primary tumours. In addition to the one described above, the DNA sequence encoding the poly Q tract of AIB1 in nine pairs of tumour and surrounding normal tissue were analysed. Five of them showed identical sequence patterns in the tumour and their surrounding normal tissues. Tumour tissue t106 was homozygous for the common 29 repeat sequence, and identical to its matched surrounding normal tissue n105. Three pairs, n107/t108, n141/t142, and n147/t148, were heterozygous, each with two distinct sequence patterns (table 2). Three distinct sequence patterns were identified for each of tumour t102 and its pair, n101 (table 2). These results suggested that the poly Q encoding sequence might not be stable in these tumours and their surrounding tissues. Four pairs showed different sequence patterns in the tumour and its surrounding normal tissue. The number of clones sequenced in pairs n113/t114 and n157/t158 was too small to conclude a difference in sequence patterns between tumour and normal tissue. Pair n103 and t104 clearly had distinct sequence patterns with only one in common (table 2). Patterns 1 and 3 of n103 were identical in poly Q repeat length but distinct in DNA sequence. Similarly, patterns 1 and 2 of t104 had the same poly Q repeat size of 28 but different encoding sequences. Specimen pair n153 and t154 also revealed differences in sequence patterns as well as instability (table 2). These results clearly indicate that the poly Q encoding region of AIB1 is somatically unstable in primary tumour tissue. The instability involves point mutations, small insertions and deletions. It is possible that somatic changes at the molecular level in the surrounding morphologically normal tissues precede the changes in

tumorigenic phenotype. Unfortunately, matched blood specimens were not available for analysis. Instability of the region encoding for poly Q is also strongly indicated in breast cancer cell lines (MCF7-P72 in table 2).

Homozygous, heterozygous, and “extra” alleles

The total number of poly Q encoding sequences was 223, more than 214 for 107 individuals if each had two alleles (table 1). This was because some of the samples had more than two distinct sequences encoding for poly Q, particularly in cell lines and primary tumours. Twenty five samples had only one sequence pattern encoding for poly Q, presumably homozygous. Fifty nine samples had two sequence patterns encoding for poly Q and 23 specimens, mostly breast cancer cell lines and primary tumours, had more than two distinct sequence patterns, indicating that somatic alterations occurred (table 3a). Distinct sequence patterns of the same poly Q repeat size are indistinguishable by Genescan, but can be resolved by cloning and sequencing. Therefore, an individual with a single size allele is counted as heterozygous if the alleles are encoded by two distinct sequence patterns. Table 3a shows the distribution of the number of cases that were homozygous or heterozygous, or had more than two distinct encoding sequences among the four groups (breast tumour cell lines, primary breast tumours, patients with BRCA1/BRCA2 mutations and the general population). Pairwise comparison by the two sided Fisher’s exact test revealed that the proportion of the cell lines (group A) and primary tumours (group B) with more than two poly Q encoding sequence patterns was significantly higher than that of the general population (group D) (p = 0.01 and 0.0002, respectively). Since the blood specimens of the matched tumour cell lines and primary tumours were unavailable, their germline genotype status could not be determined. However, “extra” alleles rarely occurred in blood samples. Thus, these “extra” sequence patterns found in primary

Table 2 Instability of the poly Q encoding sequence

Sequence pattern								Clone no	Size
MCF7-P72	1	(CAG) ₃	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	5/7	26
	2	(CAG) ₃	CAA	(CAG) ₇	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	1/7	24
	3	(CAG) ₃	CAA	(CAG) ₈	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	1/7	25
n101	1	(CAG) ₆	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	1/3	29
	2	(CAG) ₆	CAA	(CAG) ₈	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	1/3	28
	3	(CAG) ₃	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	1/3	26
t102	1	(CAG) ₃	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	5/15	26
	2	(CAG) ₆	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	7/15	29
	3	(CAG) ₆	CAA	(CAG) ₈	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	3/15	28
n103	1	(CAG) ₃ CAA(CAG) ₂	CAA	(CAG) ₁₁	(CAACAG) ₂	(CAACAGCAG) ₂	CAA	2/4	29
	2	(CAG) ₃ CAA(CAG) ₂	CAA	(CAG) ₈	(CAACAG) ₂	(CAACAGCAG) ₂	CAA	1/4	25
	3	(CAG) ₃ CAA(CAG) ₂	CAA	(CAG) ₁₀ CAA	(CAACAG) ₂	(CAACAGCAG) ₂	CAA	1/4	29
t104	1	(CAG) ₆	CAA	(CAG) ₈	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	3/7	28
	2	(CAG) ₃ CAA(CAG) ₂	CAA	(CAG) ₁₀	(CAACAG) ₂	(CAACAGCAG) ₂	CAA	1/7	28
	3	(CAG) ₃ CAA(CAG) ₂	CAA	(CAG) ₁₁	(CAACAG) ₂	(CAACAGCAG) ₂	CAA	1/7	29
n105	4	(CAG) ₅	CAA	(CAG) ₈	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	2/7	27
	1	(CAG) ₆	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	6/6	29
	1	(CAG) ₆	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	7/7	29
n107	1	(CAG) ₃	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	1/4	26
t108	2	(CAG) ₆	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	3/4	29
	1	(CAG) ₃	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	2/3	26
	2	(CAG) ₆	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	1/3	29
n113	1	(CAG) ₆	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	2/2	29
t114	1	(CAG) ₆	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	2/2	29
	2	(CAG) ₆	CAA	(CAG) ₁₀	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	2/2	30
	1	(CAG) ₃	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	1/6	26
n141	2	(CAG) ₆	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	5/6	29
	1	(CAG) ₃	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	4/6	26
	2	(CAG) ₆	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	2/6	29
t142	1	(CAG) ₆	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	2/5	28
	2	(CAG) ₃	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	3/5	26
	1	(CAG) ₆	CAA	(CAG) ₈	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	3/6	28
t148	2	(CAG) ₃	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	3/6	26
	1	(CAG) ₆	CAA	(CAG) ₈	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	3/6	28
	2	(CAG) ₃	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	1/6	26
n153	3	(CAG) ₃ CAACAG	CAA	(CAG) ₈	(CAACAG) ₂	(CAACAGCAG) ₂	CAA	1/6	26
	4	(CAG) ₆	CAA	(CAG) ₆	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	1/6	26
	1	(CAG) ₆	CAA	(CAG) ₈	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	3/7	28
t154	2	(CAG) ₃	CAA	(CAG) ₈	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	2/7	25
	3	(CAG) ₃	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	2/7	26
	1	(CAG) ₆	CAA	(CAG) ₉	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	1/4	29
n157	2	(CAG) ₆	CAA	(CAG) ₈	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	3/4	28
	1	(CAG) ₆	CAA	(CAG) ₈	(CAACAG) ₃	(CAACAGCAG) ₂	CAA	4/4	28

n, surrounding normal tissue (odd number); t, tumour tissue (even number).

tumours and cell lines must have occurred somatically or in vitro during cell culturing. Some cell lines and primary tumours had multiple “extra” poly Q encoding sequences, but none of the blood DNA of the *BRCA1/BRCA2* mutation carrier patients or the general population had more than one “extra” poly Q encoding sequences. The extra encoding sequences are not due to PCR artefacts (see discussion). These results suggest that the poly Q encoding DNA sequences are unstable in tumour cell lines and primary tumours. Eighty four percent of the “extra” sequences in breast tumours and cell lines were shorter than their parental alleles, suggesting that somatic instability tends to result in shorter repeat size. This

observation was also consistent with the previous report that a shorter CAG repeat in the androgen receptor gene was correlated with higher transcriptional activity and a higher risk of more aggressive prostate cancer phenotype.^{14–16}

Common and uncommon sequence patterns

The distribution of the number of cases with common and uncommon poly Q encoding sequence patterns among the four groups of samples is shown in table 3b. Among the 21 uncommon sequence patterns, pattern 5 was present in nine specimens while ten sequence patterns occurred only once (table 1). Overall, the total uncommon sequence patterns

Table 3a Distribution of distinct poly Q sequence patterns

Number of sequence patterns	A*		B*		C*		D*	
	No	%	No	%	No	%	No	%
1, homozygous	2	12.5	4	12.5	6	37.5	13	30.2
2, heterozygous	9	56.2	15	46.9	7	43.8	28	65.1
3	2	12.5	6	18.8	3	18.8	2	4.7
>3	3	18.8	7	21.9	0	0	0	0
Total	16	100	32	100	16	100	43	100

The p values for two sided Fisher’s exact test for distribution of >2 and ≤2 sequence patterns in groups A and B, A and C, A and D, B and C, B and D, and C and D are 0.75, 0.69, 0.01, 0.2, 0.0002, and 0.12, respectively.
*Group A, breast tumour cell lines; B, primary breast tumours; C, blood DNA from patients carrying mutations in *BRCA1/BRCA2* genes; D, general population.

Table 3b Distribution of common and uncommon poly Q encoding sequence patterns

	A*	B*	C*	D*
Common	4	12	9	35
Uncommon	12	20	7	8
Uncommon/total	0.75	0.63	0.44	0.18

The p values for two sided Fisher's exact test for distribution of common and uncommon sequences in groups A and B, A and C, A and D, B and C, B and D, and C and D are 0.52, 0.15, 0.0001, 0.36, 0.0002, and 0.09, respectively.

*Group A, breast tumour cell lines; B, primary breast tumours; C, blood DNA from patients carrying mutations in *BRCA1/BRCA2* genes; D, general population.

occurred 59 times. Eighty percent of the uncommon sequence patterns represent less than 25% of the clones analysed in each specific specimen, suggesting that most of the uncommon sequences are not the predominant species. Statistical analysis revealed that the proportion of breast cancer cell lines and primary tumours with at least one uncommon sequence pattern is significantly higher than that of the general population ($p = 0.0001$ and <0.0002 respectively). These data suggest that most of the uncommon alleles may have arisen from somatic mutations due to genomic instability in the tumour cell lines and primary tumours.

Poly Q repeat size

The relationship between the poly Q repeat length and the risk of developing breast cancer has been an important question. Table 3c summarises the distribution of the number of cases that had at least one poly Q tract with ≤ 27 repeats in each group. Statistical analysis revealed that the proportion of sporadic breast cancer cases having at least one small size allele (≤ 27 repeats) is significantly higher than that of women with *BRCA1/BRCA2* mutations or of the general population ($p = 0.006$ and 0.005 respectively). This result is consistent with the observation that the women with *BRCA1/BRCA2* mutations are at higher risk of breast cancer if they carry alleles with at least 28 poly Q at the *AIB1* gene.¹⁸ The data also suggests that shorter poly Q encoding sequence patterns are associated with sporadic breast cancer, either by predisposition or somatic instability. If only the two predominant poly Q encoding sequence patterns from each individual, excluding the "extra" sequences, were considered, the significant difference between Groups B and C, and Groups B and D still exists.

DISCUSSION

This study provides the first insight into the DNA sequences encoding the poly Q tract of the *AIB1* gene. Previous studies^{17,18} using DNA fragment length analysis could not distinguish heterozygous alleles having the same poly Q repeat size encoded by different sequence patterns. In

Table 3c Distribution of poly Q repeat length

Number of Gln	A*	B*	C*	D*
≥ 28	10	10	12	28
≤ 27	6	22	4	15
≤ 27 /total	0.38	0.69	0.25	0.35

The p values for two sided Fisher's exact test for distribution of allele size ≥ 28 and ≤ 27 in groups A and B, A and C, A and D, B and C, B and D, and C and D are 0.06, 0.70, 1.00, 0.006, 0.005, and 0.55, respectively.

*Group A, breast tumour cell lines; B, primary breast tumours; C, blood DNA from patients carrying mutations in *BRCA1/BRCA2* genes; D, general population.

addition, it was difficult to distinguish a stutter band from low frequency of true sequence patterns that result from sequence instability. Cloning and sequencing of the individual poly Q encoding DNA region allows the identification of distinct sequence patterns of the same repeat size, as well as underrepresented "extra" sequences derived from somatic instability. It is clear that the poly Q encoding sequence is unstable in breast cancer cell lines and primary tumours, not so much by large expansion or reduction of the poly Q repeat length, but by point mutations, small deletions or insertions of one or two trinucleotide repeats. The lack of large expansion is probably due to the frequent interruption by CAA. Higher sequence variability was associated with tumour cell lines (13 distinct sequence patterns in 16 cell lines) and primary tumours (21 sequence patterns in 32 tumours) than in the germline of general population (11 distinct sequence patterns in 43 individuals from the general population) (table 1). Several lines of evidence support that the presence of "extra" poly Q encoding sequences in an individual specimen is real rather than a PCR or sequencing artefact. First, high fidelity DNA polymerase was used in cloning. If the cloning and sequencing procedures were repeated on the same sample, the same sequence patterns were reproduced, indicating that the uncommon sequences were present in the original DNA specimens and not due to PCR or cloning errors. Second, the samples containing homozygous alleles consistently gave the same single sequence pattern despite repetitive cloning and PCR sequencing, suggesting that "extra" sequence patterns do not simply arise from in vitro PCR manipulation. Third, repeat cloning and sequencing confirmed the presence of three distinct sequence patterns in two specimens from the general population, supporting that the results were not PCR artefacts. The presence of three sequence patterns in the general population could be due to slippage during somatic DNA replication followed by clonal expansion of the haematopoietic cells. In these two cases, the "extra" allele represents 16.7% (1/6) of the clones sequenced.

Genescan analysis of cell lines and primary tumours (data not shown) confirmed that the scored allele size corresponded to the prevalent sequence patterns in the samples. "Extra" sequence patterns either could not be detected by Genescan or were detected as stutter bands. Using cloning/sequencing methods, the exact sequence of the "extra" bands was clearly revealed.

It is worth noting that the proportion of specimens having ≤ 27 glutamine repeats was significantly higher in the sporadic primary breast tumour group than that in the familial *BRCA1/BRCA2* mutation carrier group. This observation leads us to speculate that the molecular pathway leading to breast cancer in the familial and the sporadic form might be different. It has been reported that women with *BRCA1/BRCA2* mutations have a higher risk of contracting breast cancer if they carry alleles with at least a 28 or longer glutamine stretch at *AIB1*.¹⁸ It is possible that in women with *BRCA1/BRCA2* mutations, *AIB1* acts through the tumour suppressor/DNA repair pathway by the mechanism of reduction or loss of *AIB1* function. This is evidenced by the loss of heterozygosity of *AIB1* in some tumours.²⁰ In the sporadic cases, *AIB1* may act through the oncogenic transcription activation pathway by the mechanism of gain of function. This is evidenced by the amplification of *AIB1* in some breast cancers. In addition, shorter glutamine repeats may have higher transactivation activities, similar to AR. The polyglutamine stretch of *AIB1* in the histone acetylase domain contributes to the interaction with histone substrate. Many transcription factors interact with the promoter regions through the poly Q stretch.²¹ Evidently, the length of the poly Q may affect the protein-protein interaction, through which the gene expression in breast cancer is modulated. Although

the change in poly Q length may be subtle, such a subtle effect may be amplified through other yet unidentified mechanisms. Furthermore, the length of poly Q may alter the stability of *AIB1* or its potency to enhance hormone action through nuclear receptors, thus affecting not only the susceptibility to breast cancers but also the sensitivity to hormones.²² Whether the "extra" poly Q encoding sequence patterns arising somatically simply reflect the genomic instability or have a functional effect, will require further investigation.

Two rare sequences are worth noting. One contains an unusually short repeat of 17 Q that may have altered cotransactivating activity. Another mutation (table 1, sequence pattern 23), a CAG to CGG change, results in the substitution of glutamine with arginine. The interruption of the poly Q tract with a highly basic charged amino acid, arginine, may cause a significant change in protein-protein interactions. We are currently investigating the transcriptional co-activation activity of these mutant *AIB1* genes in a cotransfection system containing the expression plasmids of oestrogen receptor and a reporter, luciferase, under the control of the oestrogen responsive element.

Somatic instability in breast cancer cell lines and primary tumours was clearly indicated by the presence of "extra" alleles and the differences in poly Q encoding sequences in the tumour and its surrounding normal tissue. It was speculated that gene amplification might be associated with somatic instability due to frequent DNA replication and possible slippage.

Results from this study have important implications for the role of *AIB1* in tumorigenesis. It is possible that shorter poly Q binds tighter to transcription factors for transactivation of gene expression. Therefore, even if the rare sequence is present at a low percentage, it probably would still be able to exert a significant effect. In addition to the modulation of *AIB1* activity through protein-protein interactions involving poly Q, *AIB1* is amplified in some breast tumours. Thus, *AIB1* activity can be regulated both quantitatively and qualitatively at the level of gene copy number, transcription, translation, stability of mRNA and protein, and protein function itself. In conclusion, the poly Q encoding sequences of *AIB1* have been dissected and characterised, for the first time, in tumour cell lines, primary tumours, blood samples of patients with *BRCA1/BRCA2* mutations, and blood samples of the general population. Our results suggest that the DNA sequence encoding the poly Q region of the *AIB1* gene is unstable in breast tumours and cell lines. Both oncogenic and growth suppressing roles of *AIB1* are possibly contributing to its role in tumorigenesis of breast cancer.²⁰

ACKNOWLEDGMENTS

The authors thank Dr Robert Clarke for valuable discussion and the Biostatistics Shared Resource, Georgetown University, for executing all statistic analysis. We would also like to acknowledge the Histopathology and Tissue Bank Resources at the Lombardi Cancer Center, Georgetown University, for supplying the tumour tissues.

Authors' affiliations

P Dai, L-J C Wong, Department of Oncology, Georgetown University Medical Center, Washington, DC 20057

This study is supported by DOD Breast Cancer Research Program DAMD17-01-1-0257.

REFERENCES

- Voegel JJ, Heine MJS, Zechel C, *et al*. TIF2, a 160 kDa transcriptional mediator for the ligand-dependent activation function AF-2 of nuclear receptor. *EMBO J* 1996;**15**:3667-75.
- Takeshita A, Cardona GR, Koibuchi N, *et al*. TRAM-1, a novel 160-kDa thyroid hormone receptor activator molecule, exhibits distinct properties from steroid receptor coactivator-1. *J Biol Chem* 1997;**272**:27629-34.
- Onate SA, Tsai SY, Tsai MJ, *et al*. Sequence and characterization of a coactivator for the steroid hormone receptor superfamily. *Science* 1995;**270**:1354-7.
- Li H, Gomes PJ, Chen JD. RAC3, a steroid/nuclear receptor-associated coactivator that is related SRC-1 and TIF-2. *Proc Natl Acad Sci U S A* 1997;**94**:8479-84.
- Anzick SL, Kononen J, Walker RL, *et al*. AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer. *Science* 1997;**277**:965-8.
- Guan XY, Xu J, Anzick SL, *et al*. Hybrid selection of transcribed sequences from microdissected DNA: isolation of genes within amplified region at 20q11-q13.2 in breast cancer. *Cancer Res* 1996;**56**:3446-50.
- Koivisto P, Hyytinen E, Palmberg C, *et al*. Analysis of genetic changes underlying local recurrence of prostate carcinoma during androgen deprivation therapy. *Am J Pathol* 1995;**147**:1608-14.
- Visakorpi T, Hyytinen E, Koivisto P, *et al*. In vivo amplification of the androgen receptor gene and progression of human prostate cancer. *Nat Genet* 1995;**9**:401-6.
- Paulson HL. Protein fate in neurodegenerative proteinopathies: polyglutamine diseases join the (Mis)fold. *Am J Hum Genet* 1999;**64**:339-45.
- Saudou F, Finkbeiner S, Devys D, *et al*. Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear inclusions. *Cell* 1998;**95**:55-66.
- Koshy BT, Zoghbi HY. The CAG/polyglutamine tract diseases: gene products and molecular pathogenesis. *Brain Pathol* 1997;**7**:927-42.
- Klement IA, Skinner PJ, Kaytor MD, *et al*. Ataxin-1 nuclear localization and aggregation: role in polyglutamine-induced disease in SCA1 transgenic mice. *Cell* 1998;**95**:41-53.
- La Spada AR, Wilson EM, Lubahn DB, *et al*. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* 1991;**352**:77-9.
- Kazemi-Esfarjani P, Trifiro MA, Pinsky L. Evidence for a repressive function of the long polyglutamine tract in the human androgen receptor: possible pathogenetic relevance for the (CAG)n-expanded neuronopathies. *Hum Mol Genet* 1995;**4**:523-7.
- Chamberlain NL, Driver ED, Miesfeld RL. The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function. *Nucleic Acids Res* 1994;**22**:3183-6.
- Giovannucci E, Stampfer MJ, Krithivas K, *et al*. The CAG repeat within the androgen receptor gene and its relationship to prostate cancer. *Proc Natl Acad Sci U S A* 1997;**94**:3320-3.
- Haiman CA, Hankinson SE, Spiegelman D, *et al*. Polymorphic repeat in AIB1 does not alter breast cancer risk. *Breast Cancer Res* 2000;**2**:378-85.
- Rebbeck TR, Wang Y, Kantoff PW, *et al*. Modification of BRCA1- and BRCA2-associated breast cancer risk by AIB1 genotype and reproductive history. *Cancer Res* 2001;**61**:5420-4.
- Lahiri D, Nurnberger JJ. A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies. *Nucleic Acids Res* 1991;**19**:5444.
- Shibata A, Hayashi Y, Imai T, *et al*. Somatic gene alteration of AIB1 gene in patients with breast cancer. *Endocr J* 2001;**48**:199-204.
- Gerber HP, Seipel K, Georgiev O, *et al*. transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* 1994;**263**:808-11.
- Hayashi Y, Yamamoto S, Ohmori S, *et al*. Polymorphism of homopolymeric glutamines in coactivators for nuclear hormone receptors. *Endocr J* 1999;**46**:279-84.