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

P Dai, L-J C Wong

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.

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
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-F19, 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 statistical analysis. During our initial studies using DNA fragment length analysis. During our initial studies using DNA fragment length analysis, sequences that cannot be resolved by DNA fragment length analysis. 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.

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 (CAG)3-6CAA or (CAG)5CAA(CAG)2CAA, with (CAG)5CAA(CAG)2CAA (pattern #23). 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. CCG 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
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 matching 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 pairing normal tissue. 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 sequences was significantly higher than the proportion of those with fewer than two poly Q encoding sequences. 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 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 that 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
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>
<thead>
<tr>
<th>Table 2</th>
<th>Instability of the poly Q encoding sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence pattern</td>
<td>Clone no</td>
</tr>
<tr>
<td>MCF7-P72</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>n101</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>t102</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>n103</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>t104</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3a</th>
<th>Distribution of distinct poly Q sequence patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sequence patterns</td>
<td>A*</td>
</tr>
<tr>
<td>1, homozygous</td>
<td>2</td>
</tr>
<tr>
<td>2, heterozygous</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>&gt;3</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
</tr>
</tbody>
</table>

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.
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 = 0.006 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. The data also suggest 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** using DNA fragment length analysis could not distinguish heterozygous alleles having the same poly Q repeat size encoded by different sequence patterns. In 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. Highly 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, suggesting 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. 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. 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. 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
coding 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.23

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
1 Voegel JJ, Heine MJ, Zechel C, et al. TIF2, a 160 kDa transcriptional
mediator for the ligand-dependent activation function AF-2 of nuclear
2 Takeshita A, Cardona GR, Kabuchi N, et al. TRAM-1, a novel 160-kDa
thyroid hormone receptor activator molecule, exhibits distinct properties from
3 Onate SA, Tsai SY, Tsai MJ, et al. Sequence and characterization of a
coactivator for the steroid hormone receptor superfamily. Science
4 Li H, Games PJ, Chen JD, BAC3, a steroid/nuclear receptor-associated
coactivator that is related SRC-1 and TIF-2. Proc Nat Acad Sci U S A
1997;94:8479–84.
5 Anzick SL, Ronanen J, Walker RL, et al. AIB1, a steroid receptor coactivator
from microdissected DNA: isolation of genes within amplified region at
underlying local recurrence of prostate carcinoma during androgen
receptor gene and progression of human prostate cancer. Nat Genet
9 Paulson HL. Protein fate in neurodegenerative diseases: polyglutamine
10 Saudou F, Finkbeiner S, Devys D, et al. Huntington acts in the nucleus to induce
apoptosis but death does not correlate with the formation of intranuclear
11 Koehly BT, Zahgbi HY. The CAG/polyglutamine tract diseases: gene products
12 Klement IA, Skinner PJ, Kaytor MD, et al. Atoxin-1 nuclear localization and
coregulation: role in polyglutamine-induced disease in SCAl transgenic mice.
13 La Spada AR, Wilsan EM, Lubahn DB, et al. Androgen receptor gene
mutations in X-linked spinal and bulbar muscular atrophy. Nature
14 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 (CA9)n-expanded neuroepitheliomas. Hum Mol
15 Chamberlain NL, Driver ED, Miesfeld RL. The length and location of CAG
trinucleotide repeats in the androgen receptor N-terminal domain affect
16 Giovannucci E, Stampfer MJ, Krithivas K, et al. The CAG repeat within the
androgen receptor gene and its relationship to prostate cancer. Proc Natl
17 Haiman CA, Hankinson SE, Spiegelman D, et al. Polymorphic repeat in AIB1
18 Rebbeck TR, Wang Y, Kantoff PW, et al. Modification of BRCA1- and BRCA2-
associated breast cancer risk by AIB1 genotype and reproductive history.
19 Lahiri D, Nurnberger JJ. A rapid non-enzymatic method for the preparation
of HMW DNA from blood for RFLP studies. Nucleic Acids Res
20 Shibata A, Hayashi Y, Imai T, et al. Somatic gene alteration of AIB1 gene in
21 Gerber HP, Seipel K, Georgiev O, et al. Transcriptional activation mediated by
polyglutamylated transcription factor X in prostate cancer cells. Science
1999;283:808–11.
22 Hayashi Y, Yamamoto S, Ohnari S, et al. Polyglutamylated transcriptional
Somatic instability of the DNA sequences encoding the polymorphic polyglutamine tract of the AIB1 gene

P Dai and L-J C Wong

doi: 10.1136/jmg.40.12.885

Updated information and services can be found at:
http://jmg.bmj.com/content/40/12/885

These include:

**References**
This article cites 22 articles, 8 of which you can access for free at:
http://jmg.bmj.com/content/40/12/885#BIBL

**Email alerting service**
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

**Topic Collections**
Articles on similar topics can be found in the following collections

- Breast cancer (239)
- Molecular genetics (1254)
- Genetic screening / counselling (886)

**Notes**