

# Cancer Epidemiology, Biomarkers & Prevention



## Short Tandem Repeat Polymorphism and Cancer Risk: Influence of Laboratory Analysis on Epidemiologic Findings

Nguyet Tran, Bupinder S. Bharaj, Eleftherios P. Diamandis, et al.

*Cancer Epidemiol Biomarkers Prev* 2004;13:2133-2140.

**Updated version** Access the most recent version of this article at:  
<http://cebp.aacrjournals.org/content/13/12/2133>

**Cited Articles** This article cites by 27 articles, 9 of which you can access for free at:  
<http://cebp.aacrjournals.org/content/13/12/2133.full.html#ref-list-1>

**Citing articles** This article has been cited by 5 HighWire-hosted articles. Access the articles at:  
<http://cebp.aacrjournals.org/content/13/12/2133.full.html#related-urls>

**E-mail alerts** [Sign up to receive free email-alerts](#) related to this article or journal.

**Reprints and Subscriptions** To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at [pubs@aacr.org](mailto:pubs@aacr.org).

**Permissions** To request permission to re-use all or part of this article, contact the AACR Publications Department at [permissions@aacr.org](mailto:permissions@aacr.org).

# Short Tandem Repeat Polymorphism and Cancer Risk: Influence of Laboratory Analysis on Epidemiologic Findings

Nguyet Tran,<sup>1</sup> Bupinder S. Bharaj,<sup>2</sup> Eleftherios P. Diamandis,<sup>2</sup> Mylinh Smith,<sup>3</sup> Benjamin D. L. Li,<sup>3</sup> and Herbert Yu<sup>1</sup>

<sup>1</sup>Department of Epidemiology and Public Health and Yale Cancer Center, Yale University School of Medicine, New Haven, Connecticut; <sup>2</sup>Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Toronto, Ontario, Canada; and <sup>3</sup>Department of Surgery and Feist-Weiller Cancer Center, Louisiana State University Health Sciences Center, Shreveport, Los Angeles, California

## Abstract

Short tandem repeats (STR) are common polymorphisms in the genome. The length of STR may influence gene transcription, exhibiting diverse phenotypes. Two STRs, one trinucleotide repeats in the androgen receptor (*AR*) gene and one dinucleotide repeats in the insulin-like growth factor-I (*IGF-I*) gene, have been studied for their role in cancer, and the results are conflicting. Although there are many reasons for inconsistent findings, laboratory issues are often overlooked. DNA sizing analysis is regularly used to determine the length of STR, but its analytic validity has not been evaluated in epidemiologic studies. To examine if sizing analysis can reliably determine dinucleotide STR, we compared the method with direct DNA sequencing in analyzing CA repeats in the *IGF-I* gene in a small case-control study. The study enrolled 75 breast cancer cases and 75 age- and race-matched controls. DNA was extracted from buffy

coats and was analyzed for CA repeats by both DNA sizing and direct sequencing. Our comparison indicated that these methods detected the same number of repeats in the short allele but not in the long allele. There was a substantial discrepancy between the methods in determining homozygous alleles. Although the two methods showed <10% of samples having an exact match on the number of repeats in both alleles, both techniques were able to detect a genotype-phenotype correlation and a racial disparity in the genotype. An association between breast cancer risk and *IGF-I* genotype was found in sequencing analysis but not in sizing analysis. Overall, the comparison suggests that laboratory analysis of dinucleotide STR may not be as reliable as originally thought. This unreliability in STR analysis may result in inconsistent study findings. (Cancer Epidemiol Biomarkers Prev 2004;13(12):2133–40)

## Introduction

To elucidate the mechanism underlying gene-environment interaction in carcinogenesis and an individual's susceptibility to cancer, two types of polymorphisms have been studied extensively in molecular epidemiology, single nucleotide polymorphisms and short tandem repeats (STR). STR, composed of multiple dinucleotide, trinucleotide, or tetranucleotide repeats, is a widely spread genetic variation in the genome. An example of the most investigated STR in cancer is the CAG repeats in the androgen receptor (*AR*) gene. Laboratory experiments show that the length of the repeats affects the activity of gene transcription (i.e., the longer the repeats the lower the transactivation; ref. 1). Based on the finding, it is speculated that prostate cancer risk may

vary with the length of CAG repeats due to the role of androgen receptor in the disease. This speculation has been supported by some epidemiologic studies in which longer CAG repeats are found to be associated with lower risk of prostate cancer. However, not all epidemiologic studies have found evidence in support of this relationship. In a recent article by Giovannucci (2), 12 published studies on this subject were discussed. One third of the studies showed an association between CAG repeats and prostate cancer risk, whereas two thirds did not find the association. The inconsistency has been attributed to the differences in study populations, which include patient age as well as the time and method of disease detection.

Whereas these variables are possible explanations for the inconsistency, other reasons may also contribute to the conflicting findings. One issue, which is often overlooked in epidemiologic studies, is the laboratory method used for STR analysis. STR is usually analyzed by PCR followed by DNA sizing or fragment analysis, which compares the molecular size of a PCR product with a DNA standard of known size through gel electrophoresis. Based on evaluations done under ideal laboratory conditions, this method is generally believed

Received 5/13/03; revised 6/7/04; accepted 6/15/04.

**Grant support:** National Cancer Institute Public Health Service grant CA80704.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

**Requests for reprints:** Herbert Yu, Department of Epidemiology and Public Health, Yale University School of Medicine, P.O. Box 208034 New Haven, CT 06520-8034. Phone: 203-785-5688; Fax: 203-785-6980. E-mail: Herbert.yu@yale.edu

Copyright © 2004 American Association for Cancer Research.

to be reliable. However, the actual analytic sensitivity and specificity of the method have not been rigorously evaluated in epidemiologic studies, especially for dinucleotide repeats. In a previous study, we found that CA dinucleotide repeats in the insulin-like growth factor-I (*IGF-I*) gene were associated with breast cancer risk and were related to IGF-I phenotype in the circulation (3), but these findings were not confirmed by other studies (4, 5). In search for possible reasons for the inconsistent findings, we noticed that different laboratory methods were used for STR analysis. Therefore, to determine if the difference in laboratory analysis is responsible for the discrepancy in study results, we compared these methods (i.e. direct DNA sequencing and DNA sizing analysis) with respect to their results of STR analysis in the *IGF-I* gene as well as the association of this polymorphism with breast cancer risk in a small case-control study.

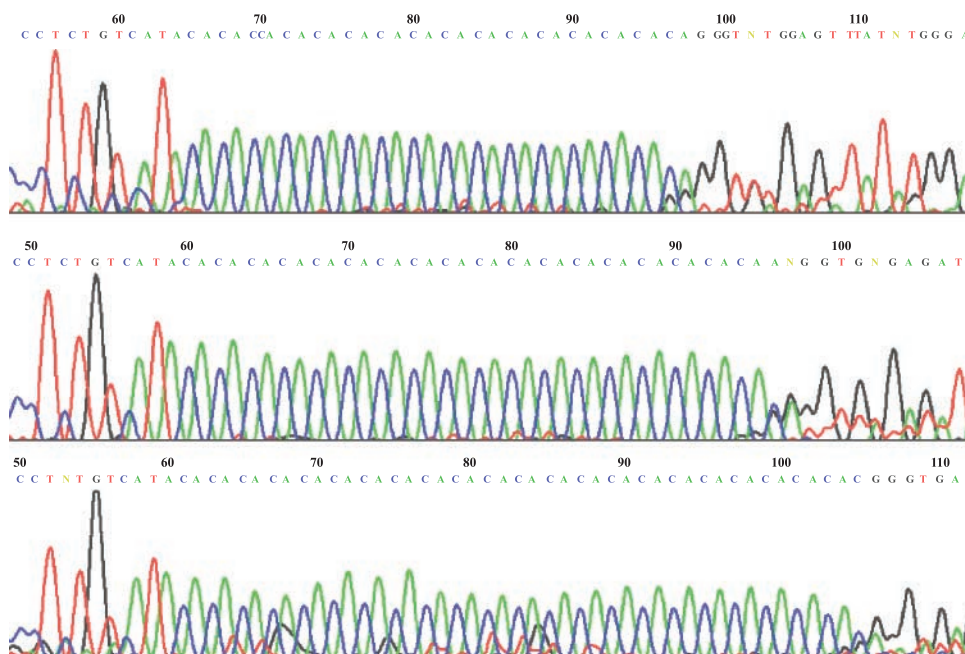
## Materials and Methods

**Study Subjects.** Between October 1998 and October 2000, women diagnosed with primary breast cancer in a breast clinic at Louisiana State University Health Sciences Center in Shreveport were recruited for a pilot case-control study, which was approved by the Institute Review Board at Louisiana State University Health Sciences Center. A total of 90 patients were enrolled in the study during the time period. For each enrolled patient, an age- ( $\pm 5$  years) and race-matched healthy volunteer was recruited as a control. The control women were identified either from healthy Louisiana State University Health Sciences Center employees or women who underwent a regular checkup at Louisiana State University Health Sciences Center with no evidence of breast cancer. All women enrolled in the study subsequently underwent an in-person interview with the use

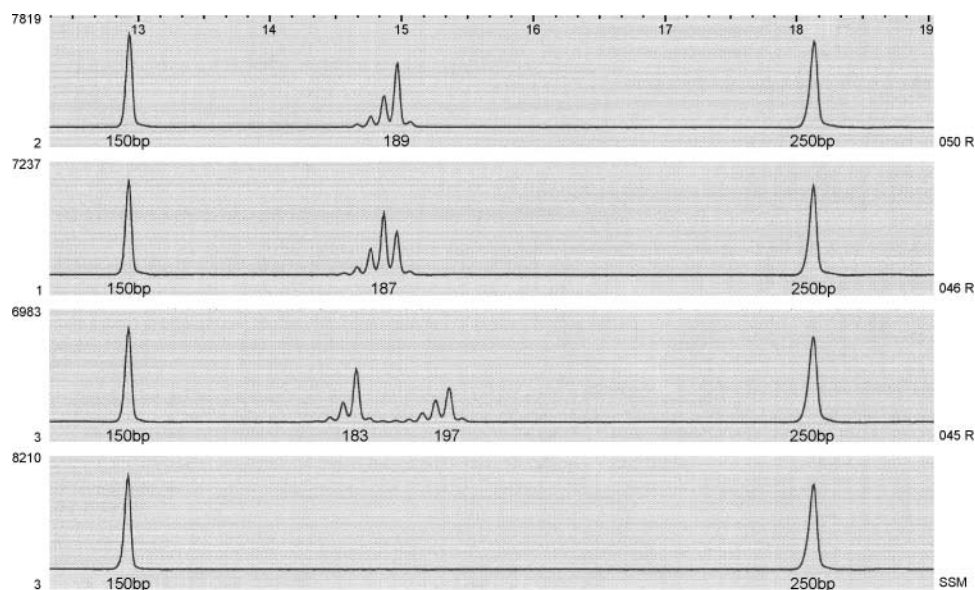
of a structured questionnaire, eliciting information on demographic features, menstrual and reproductive history, medical history, family history of cancer, and certain lifestyle factors.

**Specimen Analysis.** A 7-mL blood sample was collected from each study subject in a heparinized vacutainer tube. Within 2 hours of collection, the samples were centrifuged at 2,000 rpm for 15 minutes at 4°C. Plasma and buffy coats were collected after centrifugation, and the specimens were stored at  $-80^{\circ}\text{C}$  until analysis. Genomic DNA was extracted from buffy coats using a commercial DNA extraction kit (Gentra Systems, MN). Of the 90 matched cases and controls, 75 cases and 75 controls had DNA samples available for the study. PCR was done to amplify the promoter region of the *IGF-I* gene. The primer sequences were 5'-GCT AGC CAG CTG GTG TTA TT-3' (forward) and 5'-ACC ACT CTG GGA GAA GGG TA-3' (reverse). The PCR reaction was carried out in a 50  $\mu\text{L}$  solution containing 5  $\mu\text{L}$  of 10 $\times$  PCR buffer with 200 mmol/L Tris-HCL and 500 mmol/L KCl, 5  $\mu\text{L}$  of 10 mmol/L deoxynucleotide triphosphate mixture, 1.5  $\mu\text{L}$  of 50 mmol/L  $\text{MgCl}_2$ , 2.5  $\mu\text{L}$  each of primers with 0.34 to 0.45 nmol/L oligonucleotides, 0.5  $\mu\text{L}$  (2.5 units) Taq polymerase (Life Technologies, MD), 5  $\mu\text{L}$  sample with 5 to 10 ng genomic DNA, and 28  $\mu\text{L}$  autoclaved distilled water. The PCR procedure started with sample denaturing at 94°C for 5 minutes, followed by 30 cycles of denaturing at 94°C for 30 seconds, annealing at 63°C for 35 seconds, and extension at 72°C for 35 seconds, and a final extension at 72°C for 5 minutes.

The number of CA repeats in the IGF-I promoter was determined by two different methods. One was PCR-based direct DNA sequencing using the ABI Prism 377 DNA sequencer with forward primers (Applied Biosystems, CA). The number of CA repeats on each allele was determined directly based on the sequencing results (Fig. 1). For the short allele, the count of repeats started



**Figure 1.** Results of direct DNA sequencing analysis on three different samples. *Top*, sequencing from sample A with 15 and 16 CA repeats; *middle*, sequencing from sample B with 18 and 19 CA repeats; *bottom*, sequencing from sample C with 22 and 23 CA repeats.



**Figure 2.** Results of DNA sizing analysis on three different samples. *Top*, sizing analysis from sample D, which is homozygous with 189 bp; *upper middle*, sizing analysis from sample E, which is homozygous with 187 bp; *lower middle*, sizing analysis from sample F, which is heterozygous with 183 and 197 bp; *bottom*, molecular size standards without DNA samples.

from the 5' end and stopped at the 3' end where the last CA did not overlap with other bases. For the long allele, the count began at the same position as the short allele but ended at the CA repeat that overlapped with other bases. The second method used to determine the CA polymorphism was DNA sizing analysis utilizing the MicroGene Blaster automated sequencer (Visible Genetics, Ontario). In the sizing analysis, the forward primers in PCR were labeled with fluorescent dye Cy5.5 at the 5' end, and the fluorescence was captured by a detector after laser excitation during electrophoresis (6). Two molecular weight standards, one 150 bp and one 250 bp, were included in the sequencing gel. The distances between the peaks of a PCR product and two molecular weight standards were used to determine the size of the PCR product (Fig. 2). Based on the size, the number of CA repeats was estimated. The sizing analysis was done in a service lab at the University of California, Irvine. The sizing analysis was carried out at Mount Sinai Hospital, University of Toronto. Both laboratories analyzed the samples without information on the specimens and knowledge of the study. The two laboratories did the analysis at different times during the study, and the numbers of samples being tested were also different. Sequencing analysis was done initially for 106 samples, and sizing analysis was done later for 150 samples. Of these samples, 94 samples were analyzed by both methods. Plasma concentrations of IGF-I were measured to determine the phenotype of this molecule in the circulation; the measurement was done with the use of commercial immunoassay kits (Diagnostic Systems Laboratories, TX).

**Statistical Analysis.** The number of CA repeats in each allele and the average number of repeats in both alleles,  $(\text{allele 1} + \text{allele 2})/2$ , were compared between the two analytic methods using the paired Student's *t* test. In the paired analysis, we assume that the longer alleles determined by sequencing analysis were identical to the longer alleles decided by sizing analysis. We also assume that the average number of repeats bears no

difference between homozygous and heterozygous alleles; however, this assumption may disguise the difference between the two genotypes. The distributions of CA repeats in each allele were also compared between cases and controls using the Wilcoxon rank sum test. The association of CA repeats with breast cancer risk was examined by calculating odds ratio (OR) and their 95% confidence interval (CI) using the unconditional logistic regression model. In the regression analysis, CA repeats were analyzed as categorical variables, which included (a) homozygous versus heterozygous alleles (only from the DNA sizing analysis), (b) presence of the 19 repeats in either allele versus absence of the 19 repeats in both alleles, (c) presence of the 18 repeats in either allele versus absence of the 18 repeats in both alleles, (d)  $<17$  CA repeats versus  $\geq 17$  CA repeats in the short allele, (e)  $\geq 20$  CA repeats versus  $<20$  CA repeats in the long allele, and (f) three groups of average repeats in both alleles. For sizing analysis, the groups of average repeats were  $\leq 17$ , 17.5, and  $\geq 18$ . For sequencing analysis, the groups were  $\leq 18$ , 18.5, and  $\geq 19$  repeats. These classifications were based on the tertile distribution of the repeats in the control group. In multivariate analysis, the model was adjusted for age, race, body mass index, menopausal status, age at menarche, and age at first live birth. Spearman correlation coefficients were calculated to assess the correlation between IGF-I genotype and phenotype.

## Results

**DNA Sizing Analysis.** The results of sizing analysis showed that 59.3% of the samples (89 of 150) were homozygous and 40.7% were heterozygous (61 of 150). The median CA repeats were 18 for homozygous and 17 (short allele) and 19 (long allele) for heterozygous, ranging from 15 and 23 (Table 1). The distributions of CA repeats were not substantially different between cases and controls both in homozygous and heterozygous samples (Fig. 3A and B).

**Table 1. Distribution of CA repeats in the IGF-I gene among all subjects**

No. CA repeats	Homozygous, n (%)	Heterozygous, n (%)	
		Short allele	Long allele
<i>Sizing analysis (n = 150)</i>			
15	1 (1.1)	14 (23.0)	
16	2 (2.3)	14 (23.0)	
17	32 (36.0)	16 (26.2)	2 (3.3)
18	47 (52.8)	15 (24.6)	16 (26.2)
19	4 (4.5)	1 (1.6)	17 (27.9)
20	2 (2.3)	1 (1.6)	18 (29.5)
21	1 (1.1)		4 (6.6)
22			3 (4.9)
23			1 (1.6)
<i>Sequencing analysis (n = 106)</i>			
15		12 (11.3)	
16		7 (6.6)	1 (0.9)
17		32 (30.2)	1 (0.9)
18		49 (46.2)	9 (8.5)
19		5 (4.7)	51 (48.1)
20		1 (0.9)	30 (28.3)
21			10 (9.4)
22			3 (2.8)
23			1 (0.9)

**DNA Sequencing Analysis.** Using direct DNA sequencing, we found no samples having homozygous alleles. The range of CA repeats determined by sequencing was between 15 and 23, which was similar to the range determined by sizing analysis. The median numbers of CA repeats were 18 for the short allele and 19 for the long allele (Table 1). Cases and controls did not show significant differences in the distribution of CA repeats (Fig. 3C and D). In comparison of sequencing with sizing analysis, it seems that the discrepancy in CA distribution between cases and controls is more evident in sequencing data than in sizing data (Fig. 3).

**Method Comparison.** Tables 2 and 3 show the results of method comparison. A total of 94 DNA samples were analyzed by both sequencing and sizing methods. These analyses generated similar results on the short allele. The average numbers of CA repeats in the short allele were 17.2 by direct sequencing and 17.3 by sizing analysis ( $P = 0.287$ ). For the long allele, the results were quite different. On average, sequencing results were one CA repeat longer than sizing results, 19.4 versus 18.3, respectively ( $P < 0.001$ ). Because the comparisons were done among all subjects, including those with homozygous alleles determined by sizing analysis, we further compared the methods among heterozygous women only. The results were similar: no difference in the short allele ( $P = 0.661$ ), but significant discrepancy in the long allele ( $P < 0.001$ ). The average numbers of repeats in both alleles were also significantly different between the methods. This was probably due to the difference in the long allele.

Table 3 shows the difference of CA repeats between alleles determined by both methods. Of the samples that were homozygous by sizing analysis, 29 samples had one-repeat difference between alleles and 30 samples had two-repeat difference when the samples were analyzed by direct sequencing. Among the heterozygous samples, there were 12 samples that had the same allelic difference

in CA repeats between the two methods, including 5 samples with two-repeat difference, 3 with three-repeat difference, 3 with four-repeat difference, and 1 with five-repeat difference. Of the 12 samples that had the same allelic difference in CA repeats, only 8 had exactly the same number of repeats by both methods. Thus, the absolute agreement between the two methods was only 8.5%. Overall, direct sequencing tended to detect more allelic difference in CA repeats than sizing analysis.

**Genotype Versus Phenotype.** The correlation between IGF-I genotype and phenotype was examined among all control subjects whose samples were analyzed by both methods. The results of the analysis are shown in Table 4. IGF-I phenotype in the circulation was correlated with IGF-I genotype in the long or both alleles but not in the short allele alone. Longer repeats in long or both alleles were correlated with lower levels of circulating IGF-I; this correlation was seen consistently in both sequencing and sizing analyses.

**Disease Association.** The association between breast cancer risk and IGF-I genotype was analyzed using categorical variables described earlier. The analysis showed that breast cancer risk was not associated with any categories of IGF-I genotype determined by sizing analysis (Table 5). However, when using the genotype data from direct sequencing, women with 19 CA repeats in either allele had a 2-fold increase in risk for breast cancer compared with those without 19 repeats in either of the alleles. Moreover, this finding did not change substantially when potential confounding variables and risk factors were adjusted in the analysis, including age, race, body mass index, menopausal status, age at menarche, and age at first live birth. Further adjusting for circulating levels of IGF-I also did not change the results (data not shown).

**Genotype and Race.** Over 60% of the study subjects were African American women. Thus, racial differences in relation to IGF-I genotype were examined. The results suggested that African Americans had fewer numbers of CA repeats than their Caucasian counterparts; over 50% of African Americans were in the category of short average repeats compared with <13% of Caucasians in the same category ( $P < 0.01$ ). When comparing the median number of repeats in the short allele, African Americans had 17 repeats, whereas Caucasians had 18 repeats ( $P < 0.01$ ; Table 6). Also, fewer Blacks than Whites had 18 CA repeats in either allele ( $P < 0.05$ ). These results were the same regardless of which genotyping method was used. However, there was a method difference when comparing 19 repeats between races. Using sizing analysis, the distribution of 19 repeats was similar between racial groups, but in sequencing analysis, more Whites than Blacks had 19 alleles and the racial difference was significant ( $P = 0.01$ ; Table 6). The substantial change in 19 repeats among Whites from 9.4% in sizing analysis to 65.2% in sequencing analysis may be partially due to the method in detecting homozygous alleles. Sizing analysis detected 53.2% of Blacks but 84.4% of Whites having homozygous alleles; the difference was significant ( $P = 0.004$ ). This racial difference in the distribution of 19 repeats could have a confounding effect on the analysis of breast cancer in relation to IGF-I genotype. Thus, the interpretation of the association should be cautious.

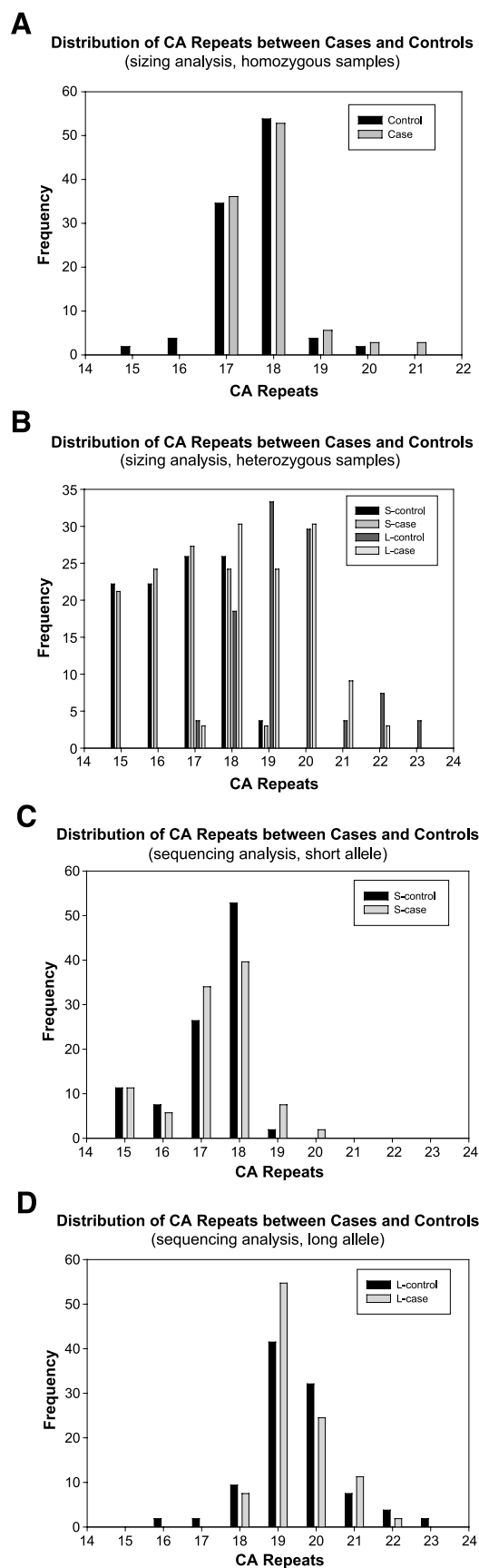
## Discussion

In this study, IGF-I genotype was found to be associated with IGF-I phenotype; higher numbers of CA dinucleotide repeats in the promoter region of the *IGF-I* gene were correlated with lower levels of IGF-I in the circulation. This correlation suggests that the length of CA repeats may adversely affect the activity of IGF-I transcription, and the direction of the influence is in agreement with what has been found for this type of polymorphism in the epidermal growth factor receptor (*EGFR*) gene, which is also involved in mitogenic signaling (7). In addition to its association with phenotype, IGF-I genotype was also different between racial groups; African Americans had shorter CA repeats than Caucasians. Furthermore, the study indicated that this polymorphism might be associated with breast cancer risk, although the association was not seen consistently between laboratory methods used for genotyping.

Our method comparison indicated that genotyping results could substantially differ when different laboratory methods were used to determine STR. Interestingly, the discrepancy in genotyping exists mainly in one allele. The study did not show any difference in the number of CA repeats in the short allele but substantial discrepancy in the long allele. On average, sequencing method reported one repeat longer than sizing analysis. This discrepancy seems to suggest that direct DNA sequencing is more sensitive than sizing analysis in detecting polymorphic dinucleotide STR. However, the results may also indicate that the sequencing method is over sensitive, which makes it incapable of identifying homozygous alleles. Our comparison showed that sizing analysis found nearly 60% of the samples to be homozygous, whereas using the strategy described in the Materials and Methods section to determine the number of repeats, sequencing analysis did not detect a single sample with homozygous alleles. It is known that the sequencing method is not perfect for STR analysis because errors may occur during DNA synthesis causing certain uncertainties. Whereas sequencing analysis has limitations in determining homozygous samples, DNA sizing analysis also has problems in distinguishing heterozygous specimens. Sizing analysis may miss certain heterozygous alleles because small peaks sometimes are difficult to detect or electrophoretic mobility may shift due to some unknown reasons.

Although both methods have limitations, our further comparison of the results and review of literature reveal several interesting observations. First, when comparing the genotyping results of the short allele, the two methods showed no difference; sequencing data matched with the sizing results, suggesting that the sequencing method may be reliable in analyzing the short allele. With regard to the long allele, although sequencing data tended to have one or two repeats longer than the sizing data, the

**Figure 3.** Distributions of CA repeats between cases and controls. **A.** Distributions of sizing analysis on homozygous samples. **B.** Distributions of sizing analysis on heterozygous samples. **C.** Distributions of sequencing analysis on the short alleles. **D.** Distributions of sequencing analysis on the long alleles.



**Table 2. Comparison of CA repeats between sequencing and sizing analyses**

Category	Sequence*	Sizing†	Difference‡	P§
<i>All subjects (n = 94)</i>				
Short allele	17.2	17.3	-0.04	0.287
Long allele	19.4	18.3	1.12	<0.001
Both allele	18.3	17.8	0.54	<0.001
<i>Heterozygous subjects (n = 35)</i>				
Short allele	16.5	16.5	-0.03	0.661
Long allele	19.8	19.3	0.54	<0.001
Both allele	18.1	17.9	0.26	0.002

\*Sequence: average number of CA repeats determined by sequencing analysis.

†Sizing: average number of CA repeats determined by sizing analysis.

‡Difference: average number of difference in CA repeats between sequencing analysis and sizing analysis.

§P, paired Student's *t* test.

|| Both allele: average number of CA repeats in both allele, (allele 1 + allele 2)/2.

discrepancies were quite consistent without exceptions. Furthermore, both methods showed that the number of CA repeats in the long allele was inversely correlated with IGF-I levels in the circulation, suggesting that both methods are capable of detecting the correlation between genotype and phenotype. The possible link between IGF-I genotype in CA repeats and IGF-I phenotype in the circulation was also indicated in another larger study involving more than 400 women; the study found that IGF-I concentrations were 179 ng/mL in women with 18 of 19 genotype, 173 in 19 of 19, 169 in 19 of 20, 166 in 19 of 21, 149 in 20 of 20, and 153 in others (*P* for trend = 0.08; ref. 4). This trend could be more evident if the analysis had excluded the last category "others" because it contained subjects who had alleles shorter than 18. Although not all studies found this genotype-phenotype association, different methods used to determine circulating levels of IGF-I and diverse study populations could affect the observation.

Second, we observed a distinct racial difference in IGF-I genotype; African Americans had fewer repeats than Caucasians. This racial variation was seen consistently with both genotyping methods, suggesting that these methods are comparable when assessing the geno-

**Table 3. Comparison of allele difference between sequencing and sizing analyses**

Allele difference by sizing analysis	Allele difference by sequencing analysis					
	1	2	3	4	5	8
Homozygous	29	30				
Heterozygous 2 CA repeats	2	5	14			
Heterozygous 3 CA repeats			3	2	1	
Heterozygous 4 CA repeats				3	3	
Heterozygous 5 CA repeats					1	
Heterozygous 7 CA repeats						1

NOTE: Allele difference in the number of CA repeats.

**Table 4. Correlation between IGF-I genotype and phenotype among control subjects**

Correlation with IGF-I levels (ng/mL)	IGF genotype		
	Short allele	Long allele	Both allele*
<i>Sizing analysis (n = 49)</i>			
Correlation coefficient†	-0.23	-0.31	-0.37
<i>P</i>	0.112	0.032	0.010
<i>Sequencing analysis (n = 49)</i>			
Correlation coefficient†	-0.14	-0.35	-0.36
<i>P</i>	0.341	0.014	0.011

\*Both allele: average number of CA repeats in both allele, (allele 1 + allele 2)/2.

†Spearman correlation.

type in relation to important covariables, such as racial ethnicity. The racial discrepancy in IGF-I genotype has been similarly observed in other studies. Jernstrom et al. (8) found that more Blacks than Whites had CA repeats  $\leq 18$  in the *IGF-I* gene (24.3% versus 6.6%). The same study also showed that Blacks had higher plasma levels of IGF-I than Whites. Given the inverse correlation between IGF-I genotype and phenotype, higher circulating IGF-I in Blacks than in Whites should be expected. A similar racial disparity in IGF-I genotype was also seen in another study, which compared the CA polymorphism between 646 White and 165 Black premenopausal women (9). Takacs et al. found that 43.5% of Blacks, but only 9.5% of Whites, had CA repeat size of 190 bp or smaller ( $\leq 18$  repeats). In that study, the most common size of CA repeats among Caucasians was 192 bp (19 repeats), of which 63.5% of Whites had this size whereas only 29.4% of Blacks had the same genotype.

Finally, in our study, sequencing results showed that women with 19 CA repeats in either allele were at higher risk for breast cancer compared with those without 19 repeats in either allele. However, the same association was not seen when the polymorphism was determined by sizing analysis. Using the same sizing analysis technique, two other studies also found no association between IGF-I genotype and breast cancer risk (4, 5), although one of the studies did show an insignificantly elevated risk (OR, 1.41; 95% CI, 0.91-2.18). One might consider that our findings of the association between IGF-I genotype and breast cancer risk were due to chance or to the unreliable sequencing method. However, several lines of evidence suggest that IGF-I genotype may have important biological implications.

Vaessen et al. (10) reported that in a study of 900 Dutch men and women (60% females), individuals with 192-bp genotype (19 repeats) of IGF-I had 2.7 cm higher body length than those with different genotypes. Furthermore, in a sample of 150 individuals randomly selected from the study population, the researchers found that serum levels of IGF-I were varied by its genotype; individuals with the genotype of 192 bp had 18% higher serum levels of IGF-I than those without the genotype (10). The same research group further showed in a different study an association between IGF-I genotype and birth weight; it was found that individuals without the 192 genotype had 215 g lower birth weight than those with the genotype (11). Thus, both studies suggest that the STR polymorphism

**Table 5. Association between IGF-I genotype and breast cancer risk**

Genotype	Sequencing analysis		Sizing analysis	
	OR* (95% CI)	OR† (95% CI)	OR* (95% CI)	OR† (95% CI)
CA 18 (no)‡	1.00	1.00	1.00	1.00
CA 18 (yes)	0.54 (0.25-1.17)	0.54 (0.22-1.30)	1.13 (0.59-2.15)	1.13 (0.53-2.43)
CA 19 (no)§	1.00	1.00	1.00	1.00
CA 19 (yes)	2.15 (0.99-4.68)	2.51 (1.03-6.13)	1.17 (0.47-2.90)	1.19 (0.44-3.19)
CA 17 (no)¶	1.00	1.00	1.00	1.00
CA 17 (yes)	0.88 (0.33-2.38)	0.65 (0.20-2.07)	1.85 (0.53-2.64)	1.00 (0.41-2.45)
CA 20 (no)*	1.00	1.00	1.00	1.00
CA 20 (yes)	0.73 (0.34-1.59)	0.82 (0.34-1.92)	1.53 (0.68-3.47)	1.26 (0.50-3.18)
Both allele (short)**	1.00	1.00	1.00	1.00
Both allele (medium)	0.73 (0.29-1.82)	0.80 (0.27-2.38)	0.71 (0.34-1.48)	0.81 (0.35-1.83)
Both allele (long)	0.72 (0.28-1.83)	0.83 (0.29-2.35)	1.49 (0.61-3.63)	1.39 (0.52-3.74)
Homozygous			1.00	1.00
Heterozygous			1.77 (0.91-3.43)	1.58 (0.76-3.32)

\*OR derived from unconditional logistic regression models without adjustment.

†OR derived from unconditional logistic regression models with adjustment for age, race, body mass index, menopause, age at menarche, age at first live birth.

‡CA 18: CA repeat number is 18 in either alleles.

§CA 19: CA repeat number is 19 in either alleles.

¶CA 17: CA repeat number is <17 in short allele.

\*CA 20: CA repeat number is ≥20 in long allele.

\*\*Both allele: average number of CA repeats in both alleles, (allele 1 + allele 2)/2.

in the IGF-I promoter, particularly the 19 CA repeats, not only affects IGF-I phenotype in the circulation but also has subsequent impacts on infant development and growth. A study of selected children born at low gestational age by Arends et al. (12) found that another STR in the *IGF-I* gene, located between exon 2 and 3, was associated with some interesting phenotypic features; children carrying the allele of 191 bp (wild-type 189 bp) had substantially lower levels of circulating IGF-I and a smaller head circumference than those not carrying this allele. Because the two polymorphic regions are very close to one another, the authors suspect that functional variant genotypes may exist in the region between IGF-I promoter and exon 3 (12).

It is interesting to note that the finding of IGF-I genotype in relation to anthropometric features at birth supports the associations between breast cancer and birth characteristics. A number of large epidemiologic investigations have shown that birth weight is associated with breast cancer risk. Women with higher birth weight have an elevated risk for breast cancer, and the relationship seems to be dose dependent (13-16). Two recent studies further indicate that body length and head circumference at birth are also associated with breast cancer risk later in life; women who have a longer body length or a larger head circumference at birth have increased risk for breast cancer (17, 18). Because these anthropometric features are believed to be the consequence of strong growth stimulation, mitogenic growth factor IGF-I is suspected to play a role in these relationships. The role of IGF-I in cancer has been suggested by a large body of evidence both from laboratory experiments and epidemiologic observations (19). Several epidemiologic studies have shown that high IGF-I in adult circulation is a risk factor for breast cancer, especially among premenopausal women (20-25). Given these apparently interrelated relationships, it is important to determine if the association between fetal anthropometrics and breast cancer

risk is due to long-term exposure to IGF-I at relatively high levels or to strong mitogenic stimulation involving IGF-I during early life development and growth.

As discussed earlier, our study showed different STR results when using direct DNA sequencing versus DNA sizing analysis. Interestingly, the discrepancy in STR results does not affect the finding of a correlation between IGF-I genotype and phenotype as well as the association of genotype with race. However, a difference in the relationship with breast cancer risk was detected. It is possible to argue that sequencing data are not reliable for STR genotyping, yet the association between IGF-I genotype and breast cancer risk seems to be supported by circumstantial evidence presented above. The difference in STR genotyping exists not only between the methods of sequencing and sizing analysis but also within the method of sizing analysis itself. With the use of the same primers, some studies found that the PCR product of 192 bp was the most common genotype (10, 26, 27), whereas others including ours showed that 189 bp was the most frequent IGF-I genotype (28). This inconsistency is probably due to the use of different DNA standards. Our method comparison also indicates that sizing analysis may be less sensitive in identifying repeat differences between alleles because there is only one heterozygous subject in our study who has one CA repeat difference between alleles; the majority of heterozygous samples had at least two-repeat difference between alleles. The lack of sensitivity of sizing analysis may result in misclassifications not only on the status of zygosity but also on the distributions of CA repeats, both of which may diminish the ability of a study in detecting an association between IGF-I genotype and breast cancer risk.

Although there are a number of interesting findings in the study, the observations are based on a very small sample size. Multiple testing may further reduce the study power. Thus, the study results should be interpreted cautiously. In summary, the study showed some evidence that IGF-I genotype in the promoter region



**Table 6. IGF-I genotype by race among control subjects**

Genotype*	Black	White	P†
<i>Sizing analysis</i>			
CA 18 (no)	32 (68.1)	7 (21.9)	
CA 18 (yes)	15 (31.9)	25 (78.1)	<0.001
CA 19 (no)	39 (83.0)	29 (90.6)	
CA 19 (yes)	8 (17.0)	3 (9.4)	0.335
CA 17 (no)	34 (72.3)	30 (93.8)	
CA 17 (yes)	13 (27.7)	2 (6.2)	0.017
CA 20 (no)	37 (78.7)	29 (90.6)	
CA 20 (yes)	10 (21.3)	3 (9.4)	0.161
Both allele (short)	24 (51.1)	4 (12.5)	
Both allele (medium)	15 (31.9)	23 (71.9)	
Both allele (long)	8 (17.0)	5 (15.6)	<0.001
Homozygous	25 (53.2)	27 (84.4)	
Heterozygous	22 (46.8)	5 (15.6)	0.004
<i>Sequencing analysis</i>			
CA 18 (no)	15 (50.0)	5 (21.7)	
CA 18 (yes)	15 (50.0)	18 (78.3)	0.035
CA 19 (no)	22 (73.3)	8 (34.8)	
CA 19 (yes)	8 (26.7)	15 (65.2)	0.005
CA 17 (no)	22 (73.3)	21 (91.3)	
CA 17 (yes)	8 (26.7)	2 (8.7)	0.098
CA 20 (no)	14 (46.7)	15 (65.2)	
CA 20 (yes)	16 (53.3)	8 (34.8)	0.179
Both allele (short)	15 (50.0)	3 (13.0)	
Both allele (medium)	5 (16.7)	13 (56.5)	
Both allele (long)	10 (33.3)	7 (30.4)	0.003
<i>Sizing analysis</i>			
Short allele	17	18	<0.001
Long allele	18	18	0.669
Both allele	17	18	0.008
<i>Sequencing analysis</i>			
Short allele	17	18	0.001
Long allele	20	19	0.397
Both allele	18.3	18.5	0.208

NOTE: Values are *n* (%) or median, where applicable.

\* $\chi^2$  test used for categorical data and Wilcoxon rank sum test used for continuous variable.

†See Table 5 for category definition.

containing CA repeats was associated with IGF-I phenotype in the circulation and was different between racial groups. A possible link between IGF-I genotype and breast cancer risk was suggested when the genotype was determined by DNA sequencing, but a similar result was not found when DNA sizing analysis was used for genotyping. Although DNA sizing analysis is the method of choice in analyzing STR polymorphism, our method comparison indicates that this method may not be as reliable as originally thought in detecting small allelic difference. Because this weakness in methodology is likely to introduce errors in the classification of genotype, misclassification bias should be considered a possible explanation for many inconsistent findings in molecular epidemiologic studies.

## References

- Gerber HP, Seipel K, Georgiev O, et al. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* 1994;263:808–11.
- Giovannucci E. Is the androgen receptor CAG repeat length significant for prostate cancer? *Cancer Epidemiol Biomarkers Prev* 2002 Oct;11:985–6.

- Yu H, Li BD, Smith M, Shi R, Berkel HJ, Kato I. Polymorphic CA repeats in the IGF-I gene and breast cancer. *Breast Cancer Res Treat* 2001 Nov;70:117–22.
- Missmer SA, Haiman CA, Hunter DJ, et al. A sequence repeat in the insulin-like growth factor-1 gene and risk of breast cancer. *Int J Cancer* 2002 Jul 20;100:332–6.
- DeLellis K, Ingles S, Kolonel L, et al. IGF1 genotype, mean plasma level and breast cancer risk in the Hawaii/Los Angeles multiethnic cohort. *Br J Cancer* 2003 Jan 27;88:277–82.
- Bharaj BS, Vassilikos EJ, Diamandis EP. Rapid and accurate determination of (CAG)*n* repeats in the androgen receptor gene using polymerase chain reaction and automated fragment analysis. *Clin Biochem* 1999 Jul;32:327–32.
- Gebhardt F, Zanker KS, Brandt B. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J Biol Chem* 1999 May 7;274:13176–80.
- Jernstrom H, Chu W, Vesprini D, et al. Genetic factors related to racial variation in plasma levels of insulin-like growth factor-1: implications for premenopausal breast cancer risk. *Mol Genet Metab* 2001 Feb;72:144–54.
- Takacs I, Koller DL, Peacock M, et al. Sibling pair linkage and association studies between bone mineral density and the insulin-like growth factor I gene locus. *J Clin Endocrinol Metab* 1999 Dec;84:467–71.
- Vaessen N, Heutink P, Janssen JA, et al. A polymorphism in the gene for IGF-I: functional properties and risk for type 2 diabetes and myocardial infarction. *Diabetes* 2001 Mar;50:637–42.
- Vaessen N, Janssen JA, Heutink P, et al. Association between genetic variation in the gene for insulin-like growth factor-I and low birthweight. *Lancet* 2002 Mar 23;359:1036–7.
- Arends N, Johnston L, Hokken-Koelega A, et al. Polymorphism in the IGF-I gene: clinical relevance for short children born small for gestational age (SGA). *J Clin Endocrinol Metab* 2002 Jun;87:2720.
- Michels KB, Trichopoulos D, Robins JM, et al. Birthweight as a risk factor for breast cancer. *Lancet* 1996 Dec 7;348:1542–6.
- Sanderson M, Williams MA, Malone KE, et al. Perinatal factors and risk of breast cancer. *Epidemiology* 1996 Jan;7:34–7.
- Innes K, Byers T, Schymura M. Birth characteristics and subsequent risk for breast cancer in very young women. *Am J Epidemiol* 2000 Dec 15;152:1121–8.
- Stavola BL, Hardy R, Kuh D, Silva IS, Wadsworth M, Swerdlow AJ. Birthweight, childhood growth and risk of breast cancer in a British cohort. *Br J Cancer* 2000 Oct;83:964–8.
- Vatten LJ, Romundstad PR, Trichopoulos D, Skjaerven R. Pre-eclampsia in pregnancy and subsequent risk for breast cancer. *Br J Cancer* 2002 Oct 21;87:971–3.
- McCormack VA, dos Santos Silva I, De Stavola BL, Mohsen R, Leon DA, Lithell HO. Fetal growth and subsequent risk of breast cancer: results from long term follow up of Swedish cohort. *BMJ* 2003 Feb 1;326:248.
- Yu H, Rohan T. Role of the insulin-like growth factor family in cancer development and progression. *J Natl Cancer Inst* 2000 Sep 20;92:1472–89.
- Hankinson SE, Willett WC, Colditz GA, et al. Circulating concentrations of insulin-like growth factor-I and risk of breast cancer. *Lancet* 1998 May 9;351:1393–6.
- Toniolo P, Bruening PF, Akhmedkhanov A, et al. Serum insulin-like growth factor-I and breast cancer. *Int J Cancer* 2000 Dec 1;88:828–32.
- Byrne C, Colditz GA, Willett WC, Speizer FE, Pollak M, Hankinson SE. Plasma insulin-like growth factor (IGF) I, IGF-binding protein 3, and mammographic density. *Cancer Res* 2000 Jul 15;60:3744–8.
- Muti P, Quattrin T, Grant BJ, et al. Fasting glucose is a risk factor for breast cancer: a prospective study. *Cancer Epidemiol Biomarkers Prev* 2002 Nov;11:1361–8.
- Krajcik RA, Borofsky ND, Massardo S, Orentreich N. Insulin-like growth factor I (IGF-I), IGF-binding proteins, and breast cancer. *Cancer Epidemiol Biomarkers Prev* 2002 Dec;11:1566–73.
- Yu H, Jin F, Shu XO, et al. Insulin-like growth factors and breast cancer risk in Chinese women. *Cancer Epidemiol Biomarkers Prev* 2002 Aug;11:705–12.
- Rosen CJ, Kurland ES, Vereault D, et al. Association between serum insulin growth factor-I (IGF-I) and a simple sequence repeat in IGF-I gene: implications for genetic studies of bone mineral density. *J Clin Endocrinol Metab* 1998 Jul;83:2286–90.
- Takacs I, Koller DL, Peacock M, et al. Sibling pair linkage and association studies between bone mineral density and the insulin-like growth factor I gene locus. *J Clin Endocrinol Metab* 1999 Dec;84:467–71.
- Sun G, Gagnon J, Chagnon YC, et al. Association and linkage between an insulin-like growth factor-1 gene polymorphism and fat free mass in the HERITAGE Family Study. *Int J Obes Relat Metab Disord* 1999 Sep;23:929–35.