A Human-specific Polymorphism in the Coding Region of the Aggrecan Gene

VARIABLE NUMBER OF TANDEM REPEATS PRODUCE A RANGE OF CORE PROTEIN SIZES IN THE GENERAL POPULATION

(Received for publication, December 2, 1996, and in revised form, March 13, 1997)

Kurt J. Doege‡, Silvija N. Coulter, Lauren M. Meek, Kirstin Maslen, and Jill G. Wood
From the Shriners Hospital for Children, Portland, Oregon 97210

Aggrecan, one of the major structural genes of cartilage, encodes a proteoglycan core protein composed of an extended central glycosaminoglycan-bearing domain, flanked by globular domains at each end. The central region consists of long stretches of repeating amino acids that serve as attachment sites for glycosaminoglycans such as chondroitin and keratan sulfate; the terminal globular domains interact with other cartilage components. The glycosaminoglycan attachment region is encoded in several species by a single large exon, within which are several different types of repeating sequences. Several species show within this exon a similar block of conserved repeats for attachment of chondroitin sulfate, but in humans this group of repeats is particularly well conserved. Examination of genomic DNA from a population of unrelated individuals by polymerase chain reaction or Southern blot assays shows this block of repeat sequences exists in multiple allelic forms, which differ by the number of repeats at this site in each allele. Thirteen different alleles have been identified, with repeat numbers ranging from 13 to 33. This is an unusual example of an expressed variable number of tandem repeat polymorphism. This polymorphism is apparently restricted to humans, of several species examined. This polymorphism results in individuals with differing length aggrecan core proteins, bearing different numbers of potential attachment sites for chondroitin sulfate. The possibility exists for a molecular understanding of biological variation in cartilage functional properties.

Aggrecan is the major proteoglycan of cartilage (1) and is expressed at high levels only in this tissue; low levels of aggrecan expression have been reported in notochord (2, 3) and calvaria (4). Aggrecan is composed of two types of structural elements, an extended central core and three flanking globular domains. Two of the globular domains are at the amino terminus, share homology with link protein (5), and one, at least (G1), binds strongly to hyaluronan and link protein, effectively localizing the molecule in the cartilage matrix. The third globular domain, G3, shares intriguing structural features with the selectin family of cell adhesion molecules (6) and may also provide important interactions with other cell or matrix components. The bulk of the molecule is made up of over 100 chondroitin sulfate chains attached to specific serine residues in the central portion of the core protein, the chondroitin-sulfate attachment region, or CS domain. These specific serine residues occur adjacent to glycine residues, and there has been considerable discussion as to what other features define a consensus attachment signal for chondroitin sulfate (7); these features include a nearby acidic amino acid as well as a nonpolar residue. It has not been determined whether all the suitable serine-glycine pairs are equally likely to be substituted with chondroitin sulfate. It is the tight packing of these highly hydrophilic, GAG-substituted molecules within the constraining collagen fibrillar network that provides the unique swelling pressure and resistance to compressive forces essential to cartilage function in articular joint surfaces.

Analysis of the cDNA sequences for rat (5) and human aggrecan (8) produced the observation that the serine-glycine pairs occurred in two distinct patterns of repeating sequences, designated the CS-1 and CS-2 repeat regions, and that the human cDNA contained a subregion of the CS-1 repeat that was remarkably conserved in 19 repeats of 19 amino acids each. The degree of conservation was that of identity for most of the repeats at the nucleic acid sequence level. Several restriction fragment length polymorphisms in the human aggrecan gene were identified for use in genetic linkage studies (9), and one, a HaeIII restriction fragment length polymorphism, was particularly polymorphic. The cDNA probe used in that case was pSA003, which encodes the CS domain including the highly repeated subregion. As the CS domain of aggrecan is encoded by a single large exon in the human gene (10, 11), and the highly repeated region within the CS domain is flanked by HaeIII sites, the possibility was suggested that this repeat region might constitute a rare VNTR within the structural coding region of a gene. The work reported here is a result of a study confirming this hypothesis, prompting the examination of the functional and biological consequences of this variation in a major component of cartilage.

EXPERIMENTAL PROCEDURES

DNA Preparation—DNA samples were obtained for the 40 large family founder couples from the CEPH (Center Etude Polymorphisme Humaine) (12) collection, as well as for three of the extended families including grandparents and offspring. Where available, lymphoblast cell stocks for these individuals were obtained from the NIGMS mutant cell repository (maintained by the Coriell Institute for Medical Research, Camden, NJ) and expanded in culture as directed by the supplier. Cells were grown in RPMI 1640 media containing 15% fetal

* This work was supported by Shriners Hospitals for Children. 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.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) U76615.

‡ To whom correspondence should be addressed: Shriners Hospital for Children, 3101 SW Sam Jackson Park Rd., Portland, OR 97210. Tel.: 503-221-3438; Fax: 503-221-3451; E-mail: kjd@shcc.org.

1 The abbreviations used are: CS, chondroitin sulfate; GAG, glycosaminoglycan; VNTR, variable number of tandem repeats; CEPH, Center Etude Polymorphisme Humaine.
bovine serum (Sigma), 1% t-glutamine, and 1% penicillin/streptomycin. DNA was prepared from the cultured cells using the Puregene Genomic DNA Isolation Kit (Gentra Systems, Inc.). DNA (non-CEPH) was also prepared from whole tissues, either surgical discards including human placenta, other cultured cell lines, as well as bovine liver, also by the Puregene Kit method. DNA for three extended CEPH families was provided by Dr. Michael Litt, Oregon Health Sciences University; several other DNA samples were kindly given by Mirta Machado, Shriners Hospital for Children, Portland.

PCR Assay—The primers used in the human genomic PCR assay were located at positions 2888–2911 of the human aggrecan cDNA sequence (8) for the upstream sense primer (5′-TAGAGGGCTCTCCGTCGCTCTGGAGTTC-3′), and for the downstream, antisense primer positions 4115–4092 (5′-AGGTCCCTACCGCAGAGGTAGAA-3′). The bovine genomic PCR reaction used primers from the two published cDNA fragments (13, 14); the upstream primer was located at position 1304–1324 of the 5′ bovine cDNA (14) (5′-AGGTGCCCTCTGCGGGGTGAGG-3′), and the downstream, antisense primer at position 57–57 of the 3′ bovine cDNA fragment (13) (5′-GGTGTCCTCCGCGCTCAGGTC-3′). The 50-μl PCR reaction typically included 50 pmol of each of the sense and antisense primers, 100 ng of genomic DNA, 0.2 mM dNTPs, 0.1% Triton X-100, 2.5 units of Taq DNA polymerase, 10 mM Tris-HCl, pH 8.3, 1.5 mM MgCl2, and 50 mM potassium chloride. The annealing temperature for both the human and bovine reactions was 67 °C, and 29 cycles of amplification were carried out. PCR products were subcloned by ligation into the T-tailed vector pCR2.1 (Invitrogen).

Southern Analysis—2 μg of genomic DNA was cut with HaeIII and fractionated by agarose gel electrophoresis. The DNA in the gel was denatured, neutralized, and transferred to derivatized nylon membrane (ZetaProbe GT, Bio-Rad) by capillary action using standard methods (15). The membrane was baked for 2 h at 80 °C in a vacuum oven, pre-washed at 65 °C for 1 h in 0.1 × SSC (0.15 M sodium chloride, 0.001 M sodium citrate, pH 7.0), 0.5% SDS, then hybridized in Church buffer (0.5 M sodium phosphate, pH 7.2, 7% SDS, 1 mM EDTA, 1% bovine serum albumin) at 65 °C for 16 h. The probe used was a HaeIII fragment of the cDNA clone pSA003 (8) encoding the human VNTR region. The blots were washed at 65 °C in 0.1 × SSC, 1% SDS, following a prewash at room temperature in 2 × SSC, 5% SDS. DNA size markers were co-electrophoresed and probed in the same hybridization with 32P-labeled marker DNA. DNA probes were labeled using random primer extension with Klenow and [α-32P]dCTP according to the kit manufacturer (Amersham Corp.).

DNA Sequencing—The two smallest alleles, 13 and 18, were sequenced after subcloning PCR-generated DNA into a T-tailed vector, using M13 sequencing primers flanking the cloning site. Cycle sequencing using Taq polymerase (ABI kit) and an ABI 383 automated DNA sequencer were used. Allele 18 was too large to sequence completely across the clone in this manner, but 15 of the 18 repeats were sequenced. The bovine VNTR product, approximately 1.6 kilobase pairs, required a nested deletion strategy, as the sequence was too repetitive for internal priming and too long to sequence across from the ends. Nested deletions in pUC18 were constructed using exonuclease III and mung bean nuclease (Stratagene kit) and sequenced as above. Nested deletions in pUC18 were constructed using exonuclease III and mung bean nuclease (Stratagene kit) and sequenced as above.

RESULTS

A PCR assay was designed to examine the size of the highly repeated region in various individuals. This strategy made use of the fact that the repeat region was part of a large exon, so that total genomic DNA could be used for analysis. Primers were chosen at positions just outside the highly repeated domain, so that the size of the product was predicted to be 33 nucleotides greater than the repeat region, which in turn was determined by the number of repeats of the 57 nucleotides encoding each 19 amino acid unit. DNA samples were obtained from a number of sources, including common cell lines and surgical discards, but the major resource used for this study was the CEPH collection’s 40 founder couples. These are the founders of large, well studied families for which lymphoblast cell cultures, as well as purified DNA in some cases, are available from the NIGMS cell repository. Upon screening DNA of 72 of these unrelated individuals, as well as 22 additional samples, a range of PCR product sizes was observed. Fig. 1 displays the first 10 of these alleles that were identified and that includes the extremes of the range; three additional alleles (not shown) have been observed making a total of 13 alleles to date.

Several lines of evidence support the conclusion that these PCR products arise from a polymorphic repeat region in the CS domain of aggrecan. First, the bands differ in size by a base pair count that is an integral number of the base pairs in a repeat unit, or 57 bases. In this respect, it was noted that the band derived from pSA003, the allele that was published as a cDNA sequence (8), was exactly one repeat larger than predicted. Apparently the large number of identical repeats in this sequence allowed the alignment to “slip” by one repeat during the computerized assembly of the random, sonicated fragment sequences in the original sequencing. This is a difficult problem to circumvent if a repeat region is long and the repeats are very similar. Second, the same pattern, adjusted for restriction site versus primer location, was obtained by genomic Southern analysis. Total genomic DNA was digested with HaeIII, electrophoretically fractionated, blotted to nylon after agarose gel electrophoresis, and hybridized with a probe specific to the VNTR region. CEPH pedigree DNA sample NIGMS Repository numbers for corresponding alleles were as given in A. Lane 1, alleles 33/27; lane 2, alleles 32/28; lane 3, alleles 29/27; lane 4, alleles 28/13. A, Southern blot assay. Total human genomic DNA was cut with HaeIII, blotted to nylon after agarose gel electrophoresis, and hybridized with a probe specific to the VNTR region. CEPH pedigree DNA sample NIGMS Repository numbers for corresponding alleles were as given in A. Lane 1, alleles 33/27; lane 2, alleles 32/28; lane 3, alleles 29/27; lane 4, alleles 28/13. B, Southern blot assay. Total human genomic DNA was cut with HaeIII, blotted to nylon after agarose gel electrophoresis, and hybridized with a probe specific to the VNTR region. CEPH pedigree DNA sample NIGMS Repository numbers for corresponding alleles were as given in A. Lane 1, alleles 33/27; lane 2, alleles 32/28; lane 3, alleles 29/27; lane 4, alleles 28/13.
generations (data not shown). Finally, as discussed below, two of the alleles have been cloned and sequenced and conclusively identified as variants of the published repeat sequence.

The description and frequency of these alleles in the studied population are presented in Table I. The numbers of repeats in these 13 alleles range from 33 to 13; the allele cloned as pSA003 corresponded to 22 repeats in this assay, which includes the published 19 repeats (8), the missing “slipped” repeat, and two additional repeats resulting from a change in the defined start of a repeat (see below), making it one of the smaller and also less common forms of the gene. The most common alleles in the general population are those of 28, 27, or 26 repeats that make up 87% observed alleles and represent a normal distribution with a mean and mode of 27 repeats. The CEPH founders used in this study consist of 49 Utah, 17 French, 4 Venezuelan, and 2 Amish individuals; the allele distribution in the two largest groups, the Utah and French families, showed the same mean allele size, 27, but somewhat different rankings of the different alleles. Allele 26 showed the most divergence, as it was 16% in the Utah families and 35% in the French families, although this sample group is quite small. No data are yet available from other ethnic groups.

The generality of the VNTR phenomenon among species was examined by attempting to demonstrate its occurrence in the bovine system. The rat (5), mouse (16), and chicken (17) aggrecan cDNA sequences have been reported, and although a recognizable variant of the human repeat region can be seen in all cases, the sequences are poorly conserved as a group for each species relative to the level of the human repeats. Partial bovine cDNA sequences have been published (13, 14), but these non-overlapping clones do not include the CS domain sequence where the repeat region would be expected. The human and bovine sequences both share related KS hexapeptide repeats however (14), suggesting the CS repeats might also be similar, at least in level of conservation. Accordingly, the bovine CS-encoding region was cloned by PCR amplification from genomic DNA, using primer sequences from the published sequence flanking the gap region. The sequence determined from this PCR clone indeed contains a series of conserved repeats similar in size and number to the human conserved region, but the degree of conservation is lower in the bovine case (Fig. 2). The rat repeat region sequence is presented for comparison and shows repeats that are conserved as a group even less well than the bovine. It is possibly due to this lower degree of sequence conservation that there has been no detectable polymorphism in approximately 20 bovine genomic DNA samples that have been examined by PCR (data not shown). These samples were obtained from different breeds and geographical regions in an effort to overcome the effects of agricultural inbreeding, but it may be that cattle from sufficiently isolated or distantly removed populations would still show some variation at this locus. At least it can be said that the repeated region in bovine is not polymorphic for the practical purpose of serving as an experimental system. Concurrently with this work, a bovine aggrecan cDNA clone encoding the CS domain was obtained in similar fashion by another group and sequenced, and the repeat region in that clone appears virtually identical to that reported here.2

The two shortest human alleles, 13 and 18, were cloned and sequenced, and the repeat structure compared with the previously sequenced allele 22 (Fig. 3). This comparison confirms that these alleles result from varying numbers of repeats and not from interpolation of some other sequence or amplification of some other unrelated sequence. The fine structure of these repeats can be examined because the near-identical repeats do show some variation, particularly in the penultimate codon, which may encode either threonine or alanine (Fig. 2). Even where the amino acid sequences are identical, types of repeats can be distinguished by a third-base change in codon 17 of the 19-codon repeat, permitting four classes of repeats to be distinguished on the basis of these two adjacent codons: ACC ACT,

### TABLE I

| Observed allele repeat number | PCR product size | Frequency | % |
|-----------------------------|-----------------|-----------|---|
| 32                          | 1915            | 1         | 0.5 |
| 32                          | 1855            | 3         | 1.6 |
| 29                          | 1687            | 4         | 2.1 |
| 28                          | 1630            | 45        | 23.9 |
| 27                          | 1573            | 84        | 44.7 |
| 26                          | 1516            | 34        | 18.1 |
| 25                          | 1459            | 4         | 2.1 |
| 22                          | 1288            | 7         | 3.7 |
| 21                          | 1231            | 1         | 0.5 |
| 20                          | 1174            | 1         | 0.5 |
| 19                          | 1117            | 1         | 0.5 |
| 18                          | 1069            | 2         | 1.1 |
| 13                          | 775             | 1         | 0.5 |

FIG. 2. Comparison of VNTR-like sequences in different species. The corresponding conserved repeated sequences within the CS domain of human (8), rat (5), and bovine aggrecan are compared, with shading indicating residues conserved in over half of the repeats of the block. Outside of these blocks the sequences are considerably divergent. Gaps are introduced to optimize the alignment of the shorter human repeats with the longer rat and bovine repeats, which contain one or two additional residues. The bovine DNA sequence, from which this translated sequence was deduced, was determined independently by Dr. Tom Hering.2

1.8

2 Dr. Tom Hering, Case Western Reserve University, personal communication. This sequence is now available in GenBank (accession number U76815).
epithelial mucin cloned from breast and pancreatic tumor cells (22), consists of 20 to 100 repeats of a 20-amino acid sequence. This range in repeat number is much greater than seen in the aggrecan expressed VNTR repeats; the number of different alleles is also much higher for this mucin. The functional consequences of the length variation has not been explored for the mucins, and in fact the properties of mucins may be rather insensitive to variations in mucin core protein lengths. The p57KIP2 gene, an inhibitor of cyclin-dependent kinase, has a proline-alanine repeat domain of 40 hexanucleotide repeats, which are polymorphic in that 7–15% of individuals have 12-nucleotide deletions. The GC-rich character of these short repeats recalls the triplet repeat type of mutations (23). There are numerous cases of variation in numbers of repeated sequences in related proteins, and perhaps not surprisingly these are often carbohydrate-bearing domains. These variants occur in different members of multigene families, as for the trout apolysialoglycoproteins (24), or as evolutionary variants in closely related species, as for the involucrin gene (25). These examples may indicate tolerance for such core protein length variation but may also reflect compensatory changes in other tissue components. The aggrecan gene variation is unique, as the variation in repeats occurs at a polymorphic locus within a single population, as for mucins; yet unlike mucins the molecule is a primary functional component of a stable tissue, interacts with several other components of that tissue, and may contribute to a subtle tissue architecture, in addition to its bulk properties. There are two major aspects that may be affected by this variation in aggrecan repeat length (Fig. 4). Most obvious is the chondroitin-sulfate content. Each repeat contains two possible attachment points for CS, so that extreme range alleles 33 and 13 might vary by as many as 40 CS chains per core protein monomer. It is not known whether all serine-glycine pairs in the molecule are equally likely to receive CS chains, but if full substitution is assumed, there would be a range of 172 to 132 CS chains between these two alleles, respectively, representing a 30% variation. It is not clear what the consequences of such a difference in GAG content might mean functionally for cartilage, assuming all other variables in this complex system are unchanged; it is of course possible that chain length or percentage of sites substituted may differ in these individuals and thus compensate for acceptor site number. Differences in sulfation produce profound changes in cartilage, as seen in the brachymorphic mouse (26), and the recently identified diastrophic dysplasia gene as a sulfate transporter (27). The chondroitin sulfate chains of aggrecan undergo changes in size and number with aging, and these changes have been implicated in the onset of osteoarthritis (28). It is possible that individuals expressing extreme repeat number alleles for the aggrecan expressed VNTR possess cartilage with unusual physical properties or that may show compensatory changes in aggrecan expression level, CS chain length, CS attachment site utilization, or sulfation pattern.

The length of the core protein varies directly with repeat number, and this length variation also may lead to changes in cartilage function. The CS-bearing part of aggrecan is thought to attain an extended rod-like structure, so that more repeats should correlate with a longer core protein. The range in the CS domain would be 1538 to 1158 amino acids (allele 33 to allele 18), of which 589 to 209 are in the CS repeat domain; this range predicts a CS region length difference of 33%. This difference in core protein length may affect some critical spacing or packing aspects of cartilage. Aggrecan is anchored at the amino terminus by the hyaluronan/link protein binding of the G1 domain, but the carboxyl-terminal G3 domain also is likely to be involved in binding to some component of cartilage. The G3

### Table: Sequence analysis of repeat structure in three allelic VNTR regions

| Repeats | Allele 13 | Allele 18 | Allele 22 | Type |
|---------|-----------|-----------|-----------|------|
| 1       | ACC GCT   | ACC GCT   | ACC GCT   | I    |
| 2       | ACT GCT   | ACT GCT   | ACT GCT   | II   |
| 3       | ACT ACT   | ACT ACT   | ACT ACT   | III  |
| 4       | ACT ACT   | ACT ACT   | ACT ACT   | IV   |
| 5       | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 6       | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 7       | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 8       | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 9       | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 10      | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 11      | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 12      | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 13      | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 14      | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 15      | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 16      | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 17      | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 18      | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 19      | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 20      | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 21      | ACT ACT   | ACT ACT   | ACT ACT   |      |
| 22      | ACT ACT   | ACT ACT   | ACT ACT   |      |

**Fig. 3. Sequence analysis of repeat structure in three allelic VNTR regions.** The VNTR regions of alleles 13 and 18 were cloned as PCR products and sequenced from each end; the allele 18 sequence was incomplete in the central portion, indicated by *dashed line* for three of the repeats. The 57-nucleotide repeats were identical except for two diagnostic codons, numbers 17 and 18 out of 19 total, which permit identification of repeat type. The sequence of codons 17 and 18 is shown for each repeat in alleles 13, 18, and the previously determined allele 22 (8), which contains one additional repeat thus far not sequenced, again shown as a *dashed line*. Lines delineate blocks of repeats of the same type that occur in similar positions in each allele; exceptions to these blocks of homogeneous repeats are within *boxes*.

ACC GCT, ACT GCT, and ACT ACT. On this basis, alleles 13 and 18 are rather similar, differing primarily by an expansion of the number of type III repeats in allele 18. It should be noted that three repeats in the middle of allele 18 could not be conclusively identified, due to the technical difficulties of sequencing such repeated regions; a deletion strategy will be required to sequence the longer alleles. Allele 22 shows several changes in the repeat pattern, including an expansion of type I repeats, a loss of type III repeats, and the introduction of a proline-alanine repeat domain of 40 hexanucleotide repeats, which are polymorphic in that 7–15% of individuals have 12-nucleotide deletions. The GC-rich character of these short repeats recalls the triplet repeat type of mutations (23). There are numerous cases of variation in numbers of repeated sequences in related proteins, and perhaps not surprisingly these are often carbohydrate-bearing domains. These variants occur in different members of multigene families, as for the trout apolysialoglycoproteins (24), or as evolutionary variants in closely related species, as for the involucrin gene (25). These examples may indicate tolerance for such core protein length variation but may also reflect compensatory changes in other tissue components. The aggrecan gene variation is unique, as the variation in repeats occurs at a polymorphic locus within a single population, as for mucins; yet unlike mucins the molecule is a primary functional component of a stable tissue, interacts with several other components of that tissue, and may contribute to a subtle tissue architecture, in addition to its bulk properties. There are two major aspects that may be affected by this variation in aggrecan repeat length (Fig. 4). Most obvious is the chondroitin-sulfate content. Each repeat contains two possible attachment points for CS, so that extreme range alleles 33 and 13 might vary by as many as 40 CS chains per core protein monomer. It is not known whether all serine-glycine pairs in the molecule are equally likely to receive CS chains, but if full substitution is assumed, there would be a range of 172 to 132 CS chains between these two alleles, respectively, representing a 30% variation. It is not clear what the consequences of such a difference in GAG content might mean functionally for cartilage, assuming all other variables in this complex system are unchanged; it is of course possible that chain length or percentage of sites substituted may differ in these individuals and thus compensate for acceptor site number. Differences in sulfation produce profound changes in cartilage, as seen in the brachymorphic mouse (26), and the recently identified diastrophic dysplasia gene as a sulfate transporter (27). The chondroitin sulfate chains of aggrecan undergo changes in size and number with aging, and these changes have been implicated in the onset of osteoarthritis (28). It is possible that individuals expressing extreme repeat number alleles for the aggrecan expressed VNTR possess cartilage with unusual physical properties or that may show compensatory changes in aggrecan expression level, CS chain length, CS attachment site utilization, or sulfation pattern.

The length of the core protein varies directly with repeat number, and this length variation also may lead to changes in cartilage function. The CS-bearing part of aggrecan is thought to attain an extended rod-like structure, so that more repeats should correlate with a longer core protein. The range in the CS domain would be 1538 to 1158 amino acids (allele 33 to allele 18), of which 589 to 209 are in the CS repeat domain; this range predicts a CS region length difference of 33%. This difference in core protein length may affect some critical spacing or packing aspects of cartilage. Aggrecan is anchored at the amino terminus by the hyaluronan/link protein binding of the G1 domain, but the carboxyl-terminal G3 domain also is likely to be involved in binding to some component of cartilage. The G3
domain of aggrecan has been shown to have lectin-like activity (29), and the similar domain of the related aggrecan family member versican has been reported to bind a variety of matrix molecules, most notably tenascin-R (30) and heparin sulfate (31). While the GAG attachment domain of versican does not show the same type of conserved repeat sequence as aggrecan and is unlikely to vary in repeat number, it does undergo alternative splicing, producing molecules with varied spacing between G1 and G3 domains (32, 33). Variation in spacing between binding domains might be expected to have disruptive effects on tissue architecture, if the ligands are relatively fixed in position. Such disruption might be greater if two alleles are expressed together in the tissue; it is currently not known if both alleles of an individual heterozygous for aggrecan are expressed together and/or at an equal level in cartilage.

The structure of the proteins expressed from these various polymorphic alleles has not been investigated; different alleles must be expressed in different individuals, however, as several alleles have been observed in the homozygous condition; these include alleles 28, 27, 26, and 22 (data not shown). The low frequency of most of the alleles (Table I) is probably the reason that homozygous individuals for the extreme forms of the gene have not been observed; it is also likely that other allelic forms exist in the population and would be observed in a larger sampling. We have analyzed the allelic distribution pattern for this gene in an independent population similar in size to that reported here, with very similar observed frequencies of these same alleles.3 There are numerous unanswered questions about the expression of such alleles, including whether both alleles in a heterozygote are expressed, whether both appear in equivalent amount and pattern in the matrix, as well as the questions regarding the glycosaminoglycan components raised above. Since these alleles occur widely in the population, the effects of the polymorphisms are obviously tolerated, and any proposed consequences would fall by necessity within the range of functional parameters seen in the population. One such potential consequence of different alleles might be a subtle variation in cartilage function affecting the resistance of an individual’s articular cartilage to the onset of joint degeneration or osteoarthritis. These conditions occur widely with aging but to different degrees of severity and ages of onset in the population. The etiology of these disorders is undoubtedly complex and is under intense investigation, but a genetic component is suspected in families with early onset osteoarthritis, and in some cases has been demonstrated (see Ref. 34, and references therein). In such a widespread disease, the type of variation seen at the polymorphic aggrecan locus may provide an explanation for cartilage with a common but subtle reduction in functionality. Support for this hypothesis has recently been obtained in a population study linking one of the aggrecan VNTR alleles to bilateral hand osteoarthritis.3

It is perhaps surprising that this type of aggrecan polymorphism only appears in humans, of the species so far examined. The occurrence of the polymorphism in other species has only been directly tested in cows, but the level of conservation of the repeats is insufficient in other species to support unequal crossing over and generation of different repeat numbers. The polymorphism depends on a high degree of sequence conservation, leading to the question of why the sequences are so highly conserved in humans alone of the species so far examined. The actual repeat sequence is somewhat different in the various species (Fig. 2). The human repeat is unique in conserving the two closely spaced serine-glycine pairs in the middle of the repeat; the bovine and rat repeats have lost one of the middle serine-glycine pairs in favor of an additional one at the beginning, which is uniformly a proline-glycine pair in human repeats. Until more is known about the utilization of particular serine-glycine pairs as CS acceptor sequences, as well as details of cartilage ultrastructure, it is difficult to speculate on the consequences of these different repeat structures.

Sequencing of the three shortest alleles confirms that they have arisen by some mechanism that expands or contracts blocks of repeated sequences, probably unequal exchange during meiosis or mitosis. There are several points of divergence among these three sequences, so the evolution has been complex and probably these three alleles are distantly related in the allelic “family.” The other alleles require sequencing to clarify the relationship. It is also possible that there may be multiple alleles of the same repeat number that have arisen independently and show a different pattern of amplified repeats. This has not been observed and would require screening by extensive sequencing. The median allele size, 25 repeats, is smaller than the mean of 27 repeats; the smaller repeat number alleles are thus over-represented. The smaller repeat alleles are probably more stable than the larger repeat alleles, which may be more likely to recombine; in addition, the large alleles may be more difficult to detect using the PCR assay. Finally, the sample size may not be large enough to provide a true representation of allele frequencies.

Acknowledgment—We gratefully acknowledge Stephanie Willis for performance of automated DNA sequencing.

3 W. E. Horton, Jr., R. Balakir, P. Precht, C. Plato, J. D. Tobin, M. Lethbridge-Cejku, L. Meek, and K. Doege, submitted for publication.
REFERENCES

1. Wight, T. N., Heinegard, D. K., and Hascall, V. C. (1991) in Cell Biology of the Extracellular Matrix (Hay, E. D., ed) pp. 45–78, Plenum Publishing Corp., New York.

2. Stirpe, N. S., and Goetinck, P. F. (1988) Development 107, 22–33.

3. Domowicz, M., Li, H., Hennig, A., Henry, J., Vertel, B. M., and Schwartz, N. B. (1995) Dev. Biol. 171, 655–664.

4. Wong, M., Lawton, T., Goetinck, P. F., Kuhn, J. L., Goldstein, S. A., and Bonadio, J. (1992) J. Biol. Chem. 267, 5592–5598.

5. Domowicz, M., Li, H., Hennig, A., Henry, J., Vertel, B. M., and Schwartz, N. B. (1995) Dev. Biol. 171, 655–664.

6. Ross, G. A., Shearer, K. B., Kuhn, J. L., and Goldstein, S. A. (1992) J. Biol. Chem. 267, 5592–5598.

7. Ross, G. A., Shearer, K. B., Kuhn, J. L., and Goldstein, S. A. (1992) J. Biol. Chem. 267, 5592–5598.

8. Doege, K. J., Sasaki, M., Kimura, T., and Yamada, Y. (1991) J. Biol. Chem. 266, 894–902.

9. Ross, G. A., Shearer, K. B., Kuhn, J. L., and Goldstein, S. A. (1992) J. Biol. Chem. 267, 5592–5598.

10. Doege, K., Sasaki, M., and Yamada, Y. (1990) Biochem. Soc. Trans. 18, 206–202.

11. Valhmu, W. B., Palmer, G. D., Rivers, P. A., Ebara, S., Cheng, J.-F., Fischer, S., and Ratcliffe, A. (1995) Biochem. J. 309, 535–542.

12. Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.-M., and White, R. (1989) Genomics 6, 575–577.

13. Oldberg, A., Antonsson, P., and Heinegard, D. (1987) Biochem. J. 243, 255–259.

14. Antonsson, P., Heinegard, D., and Oldberg, A. (1989) J. Biol. Chem. 264, 16170–16173.

15. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd Ed., pp. 9.31–9.58, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

16. Wexler, E., Deak, P., Erhardt, P., Coulter, S. N., Fulpo, C., Horvath, P., Doege, K. J., and Glant, T. T. (1994) Genomics 22, 364–371.
A Human-specific Polymorphism in the Coding Region of the Aggrecan Gene: VARIABLE NUMBER OF TANDEM REPEATS PRODUCE A RANGE OF CORE PROTEIN SIZES IN THE GENERAL POPULATION
Kurt J. Doege, Silvija N. Coulter, Lauren M. Meek, Kirstin Maslen and Jill G. Wood

J. Biol. Chem. 1997, 272:13974-13979.
doi: 10.1074/jbc.272.21.13974

Access the most updated version of this article at http://www.jbc.org/content/272/21/13974

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 32 references, 17 of which can be accessed free at http://www.jbc.org/content/272/21/13974.full.html#ref-list-1