

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
13 February 2003 (13.02.2003)

PCT

(10) International Publication Number
WO 03/012121 A2

- (51) International Patent Classification⁷: **C12Q** **Leena**; Cambridge, MA (US). **KEENE, Douglas, R.**; Cambridge, MA (US). **BRUNKEN, William, Joseph**; 34 Green Lodge Street, Canton, MA 02021 (US).
- (21) International Application Number: PCT/US02/23075
- (22) International Filing Date: 22 July 2002 (22.07.2002) (74) Agent: **MYERS, Louis**; Fish & Richardson P.C., 225 Franklin Street, Boston, MA 02110 (US).
- (25) Filing Language: English (81) Designated State (*national*): JP.
- (26) Publication Language: English (84) Designated States (*regional*): European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR).
- (30) Priority Data: 60/309,158 31 July 2001 (31.07.2001) US
- (71) Applicant: **THE GENERAL HOSPITAL CORPORATION** [US/US]; 55 Fruit Street, Boston, MA 02114 (US). **Published:**
— without international search report and to be republished upon receipt of that report
- (72) Inventors: **BURGESSON, Robert, Eugene**; 84 Naugus Avenue, Marblehead, MA 01945 (US). **KOCH, Manuel**; Cambridge, MA (US). **BRUCKNER-TUDERMAN**,
For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.



WO 03/012121 A2

(54) Title: COLLAGEN XXII, A NOVEL HUMAN COLLAGEN AND USES THEREOF

(57) Abstract: The invention provides isolated nucleic acids molecules, designated collagen XXII nucleic acid molecules, which encode a novel collagen. The invention also provides antisense nucleic acid molecules, recombinant expression vectors containing collagen XXII nucleic acid molecules, host cells into which the expression vectors have been introduced, and nonhuman transgenic animals in which a collagen XXII gene has been introduced or disrupted. The invention still further provides isolated collagen XXII proteins, fusion proteins, antigenic peptides and anti-collagen XXII antibodies. Diagnostic methods utilizing compositions of the invention are also provided.

5

COLLAGEN XXII, A NOVEL HUMAN COLLAGEN AND USES THEREOF

Related Applications

This application claims the benefit of U.S. Provisional Application Serial No. 60/309,158, filed July 31, 2001, the contents of which are incorporated by reference
10 in their entirety, including drawings.

Background of the Invention

Collagen is the most abundant protein in mammals, constituting a quarter of their total weight. Collagen provides the tensile strength in the connective tissues of all animals and is the major fibrous element of skin, bone, tendon, cartilage, blood
15 vessels, and teeth.

Collagens are classified into several types based on sequence identity and function. Types I, II, & III collagen molecules make up the main fibers of most animal extracellular structures. Type I forms about 90% of the body's collagen and is the primary component of bone, skin and tendons. Type II makes up the major fibers
20 of cartilage. Collagen fibers are arranged in rigid plates in bones, in parallel bundles in tendons, and in a dense meshwork in cartilage. Type I and lesser amounts of type III make up tendons and skin. Type IV collagen molecules make up very fine, unstriated fibers present in basal laminae. Over a dozen other collagen types are known but are less well characterized.

25 Collagen polypeptide chains are characterized by a core helical domain made up of repeating glycine-X-Y triplets and globular N-terminal and C-terminal domains. Three such chains are wound around one another in a superhelix to generate an individual ropelike collagen molecule.

5

Summary of the Invention

The present invention is based, in part, on the discovery of a novel collagen, referred to herein as collagen XXII. The nucleotide sequence of a cDNA encoding collagen XXII is shown in SEQ ID NO:1 and the corresponding amino acid sequence of collagen XXII polypeptide is shown in SEQ ID NO:2. In addition, the nucleotide sequences of the coding region are depicted in SEQ ID NO:3. An alternatively spliced variant of collagen XXII lacks nucleotides 3458-3518 of SEQ ID NO:1 and codes for a protein that lacks amino acids 1005-1024 of SEQ ID NO:2.

Accordingly, in one aspect, the invention features a nucleic acid molecule that encodes a collagen XXII protein or polypeptide, e.g., a biologically active portion of the collagen XXII protein. In one embodiment the isolated nucleic acid molecule encodes a polypeptide having the amino acid sequence of SEQ ID NO:2. In another embodiment, the isolated nucleic acid molecule encodes a polypeptide having the amino acid sequence of SEQ ID NO:2 but lacking amino acids 1005-1024 of SEQ ID NO:2. In other embodiments, the invention provides isolated collagen XXII nucleic acid molecules having the nucleotide sequence shown in SEQ ID NO:1, SEQ ID NO:3, or an alternatively spliced variant thereof. In still other embodiments, the invention provides nucleic acid molecules that are substantially identical (e.g., naturally occurring allelic variants) to the nucleotide sequence shown in SEQ ID NO:1, SEQ ID NO:3, or an alternatively spliced variant thereof. In other embodiments, the invention provides a nucleic acid molecule which hybridizes under a stringency condition described herein to a nucleic acid molecule comprising the nucleotide sequence of SEQ ID NO:1, SEQ ID NO:3, or an alternatively spliced variant thereof, wherein the nucleic acid encodes a full length collagen XXII protein or an active fragment thereof.

In a related aspect, the invention further provides nucleic acid constructs that include a collagen XXII nucleic acid molecule described herein. In certain embodiments, the nucleic acid molecules of the invention are operatively linked to native or heterologous regulatory sequences. Also included, are vectors and host cells containing the collagen XXII nucleic acid molecules of the invention e.g., vectors and

5 host cells suitable for producing collagen XXII nucleic acid molecules and polypeptides.

In another related aspect, the invention provides nucleic acid fragments suitable as primers or hybridization probes for the detection of collagen XXII-encoding nucleic acids.

10 In still another related aspect, isolated nucleic acid molecules that are antisense to a collagen XXII encoding nucleic acid molecule are provided.

In another aspect, the invention features, collagen XXII polypeptides, and biologically active or antigenic fragments thereof that are useful, e.g., as reagents or targets in assays applicable to treatment and diagnosis of collagen XXII-mediated or -
15 related disorders. In another embodiment, the invention provides collagen XXII polypeptides having a collagen XXII activity described herein. Preferred polypeptides are collagen XXII proteins including at least one collagen domain and, preferably, having a collagen XXII activity, e.g., a collagen XXII activity as described herein.

20 In other embodiments, the invention provides collagen XXII polypeptides, e.g., a collagen XXII polypeptide having the amino acid sequence shown in SEQ ID NO:2; an amino acid sequence that is substantially identical to the amino acid sequence shown in SEQ ID NO:2, e.g., an amino acid sequence comprising amino acids 1-1004 and 1025-1626 of SEQ ID NO:2; or an amino acid sequence encoded by
25 a nucleic acid molecule having a nucleotide sequence which hybridizes under a stringency condition described herein to a nucleic acid molecule comprising the nucleotide sequence of SEQ ID NO:1, SEQ ID NO:3, or an alternatively spliced variant thereof, wherein the nucleic acid encodes a full length collagen XXII protein or an active fragment thereof.

30 In a related aspect, the invention further provides nucleic acid constructs which include a collagen XXII nucleic acid molecule described herein.

In a related aspect, the invention provides collagen XXII polypeptides or fragments operatively linked to non-collagen XXII polypeptides to form fusion proteins.

5 In another aspect, the invention features antibodies and antigen-binding fragments thereof, that react with, or more preferably specifically bind collagen XXII polypeptides or fragments thereof, e.g., a collagen Gly-x-y core domain.

 In another aspect, the invention provides methods of screening for compounds that modulate the expression or activity of the collagen XXII polypeptides or nucleic
10 acids.

 In still another aspect, the invention provides a process for modulating collagen XXII polypeptide or nucleic acid expression or activity, e.g. using the screened compounds. In certain embodiments, the methods involve treatment of conditions related to aberrant activity or expression of the collagen XXII polypeptides
15 or nucleic acids, such as conditions involving aberrant or deficient tissue strength, e.g., in tissue boundaries, e.g., in myotendinous junctions; joints (e.g., diarthrodial joints, synovial joints), e.g., in the articular cartilage of joints; hair follicles; the ciliary body of the eye; heart valves; epithelial-basement membrane boundaries.

 The invention also provides assays for determining the activity of or the
20 presence or absence of collagen XXII polypeptides or nucleic acid molecules in a biological sample, including for disease diagnosis.

 In a further aspect, the invention provides methods for evaluating the efficacy of a treatment of a disorder, e.g., a collagen related disorder. The method includes: treating a subject, e.g., a patient or an animal, with a protocol under evaluation (e.g.,
25 treating a subject with one or more of: chemotherapy, radiation, and/or a compound identified using the methods described herein); and evaluating the expression of a collagen XXII nucleic acid or polypeptide before and after treatment. A change, e.g., a decrease or increase, in the level of a collagen XXII nucleic acid (e.g., mRNA) or polypeptide after treatment, relative to the level of expression before treatment, is
30 indicative of the efficacy of the treatment of the disorder. The level of collagen XXII nucleic acid or polypeptide expression can be detected by any method described herein.

 In a preferred embodiment, the evaluating step includes obtaining a sample (e.g., a tissue sample, e.g., a biopsy, or a fluid sample) from the subject, before and
35 after treatment and comparing the level of expressing of a collagen XXII nucleic acid (e.g., mRNA) or polypeptide before and after treatment.

5

In another aspect, the invention features a method of treating a subject. The method includes modulating the level of collagen XXII between a first tissue and a second tissue of the subject. The subject can be a human or a non-human animal, e.g., an animal model for a collagen related disorder, e.g., a connective tissue disorder. In
10 a preferred embodiment, the subject has or is at risk for a collagen XXII related disorder, e.g., a connective tissue disorder, e.g., a disorder of a joint, tendon, cartilage, basement membrane, or other tissue-tissue boundary. While not wanting to be bound by theory, the collagen XXII is believed to promote tensile strength and/or adherence of the two tissues. In one embodiment, collagen XXII is found in nature at the tissue-
15 tissue boundary. In another embodiment, collagen XXII is not normally found in nature at the tissue-tissue boundary.

The level of collagen XXII can be modulated by modulating, e.g., collagen XXII expression (e.g., modulating collagen XXII rate of transcription or mRNA stability), protein levels, or protein activity.

20 In a preferred embodiment, collagen XXII is increased, e.g., by administering an agent that increases the level of collagen XXII. An agent which increases the level of collagen XXII activity can be one or more of the following: a peptide or protein, e.g., a monoclonal antibody, which stabilizes or assists the binding of collagen XXII to a collagen XXII binding partner; a collagen XXII polypeptide or a functional
25 fragment or analog thereof; a nucleotide sequence encoding a collagen XXII polypeptide or functional fragment or analog thereof; an agent which increases collagen XXII nucleic acid expression; e.g., a small molecule which binds to the promoter region of collagen XXII. In a preferred embodiment, collagen XXII levels are increased by administering, e.g., introducing, a nucleotide sequence encoding a
30 collagen XXII polypeptide or functional fragment or analog thereof, into a particular cell, e.g., a connective tissue cell, in the subject. The nucleotide sequence can be a genomic sequence or a cDNA sequence. The nucleotide sequence can include: a collagen XXII coding region; a promoter sequence, e.g., a promoter sequence from a collagen XXII gene or from another gene; an enhancer sequence; untranslated
35 regulatory sequences, e.g., a 5' untranslated region (UTR), e.g., a 5'UTR from a collagen XXII gene or from another gene, a 3' UTR, e.g., a 3'UTR from a collagen

5 XXII gene or from another gene; a polyadenylation site; an insulator sequence. In another preferred embodiment, the level of collagen XXII protein is increased by increasing the level of expression of an endogenous collagen XXII gene, e.g., by increasing transcription of the collagen XXII gene or increasing collagen XXII mRNA stability. In a preferred embodiment, transcription of the collagen XXII gene
10 is increased by: altering the regulatory sequence of the endogenous collagen XXII gene, e.g., by the addition of a positive regulatory element (such as an enhancer or a DNA-binding site for a transcriptional activator); the deletion of a negative regulatory element (such as a DNA-binding site for a transcriptional repressor) and/or replacement of the endogenous regulatory sequence, or elements therein, with that of
15 another gene, thereby allowing the coding region of the collagen XXII gene to be transcribed more efficiently.

In some embodiments, the collagen XXII is increased in conjunction with increasing another collagen type, e.g., any one or more of collagen types I, II, III, or IV.

20 In a preferred embodiment, collagen XXII which is exogenous to one or two of the tissues is administered.

In another preferred embodiment, collagen XXII can be decreased by administering an agent that inhibits collagen XXII gene expression, mRNA stability, protein production levels and/or activity. An agent that inhibits collagen XXII can be
25 one or more of: a collagen XXII binding protein, e.g., a soluble collagen XXII binding protein that binds and inhibits a collagen XXII activity, e.g., helix forming activity, or inhibits the ability of collagen XXII to interact with a binding partner, e.g., a binding partner that binds to one or more of the non-helical interruption domains; an antibody that specifically binds to the collagen XXII protein, e.g., an antibody that disrupts
30 collagen XXII's ability to form a helix or bind to a binding partner; a mutated inactive collagen XXII or fragment thereof which binds to a collagen XXII but disrupts a collagen XXII activity, e.g., a collagen XXII fragment that is able to form a helix with an endogenous collagen XXII, but lacks the non-helical interruption domains and/or the N-terminal and/or C-terminal globular domains; a collagen XXII nucleic acid
35 molecule that can bind to a cellular collagen XXII nucleic acid sequence, e.g., mRNA, and inhibit expression of the protein, e.g., an antisense molecule or collagen XXII

5 ribozyme; an agent which decreases collagen XXII gene expression, e.g., a small molecule which binds the promoter of collagen XXII; an enzyme that breaks down collagen XXII, e.g., a collagenase, e.g., a collagen XXII specific collagenase. In another preferred embodiment, collagen XXII is inhibited by decreasing the level of expression of an endogenous collagen XXII gene, e.g., by decreasing transcription of
10 the collagen XXII gene. In a preferred embodiment, transcription of the collagen XXII gene can be decreased by: altering the regulatory sequences of the endogenous collagen XXII gene, e.g., by the addition of a negative regulatory sequence (such as a DNA-binding site for a transcriptional repressor), or by the removal of a positive regulatory sequence (such as an enhancer or a DNA-binding site for a transcriptional
15 activator). In another preferred embodiment, the antibody which binds collagen XXII is a monoclonal antibody, e.g., a humanized chimeric or human monoclonal antibody.

 In some embodiments, the collagen XXII is decreased in conjunction with decreasing another collagen type, e.g., any one or more of collagen types I, II, III, or IV.

20 The first tissue and the second tissue can be the same tissue type or they can be different tissue types. For example, the first and second tissue can be bone, connective tissue, (e.g., cartilage, tendon, ligament, basement membranes), muscle (e.g., muscle of the heart, eye, or hair follicle), or epithelial tissue. In a preferred embodiment, the first tissue is muscle and the second tissue is connective tissue, e.g.,
25 cartilage, tendon, or basement membrane tissue, bone, or another muscle tissue. In another preferred embodiment, the first tissue is bone and the second tissue is cartilage, ligament, muscle, or another bone tissue. In another preferred embodiment, the first tissue is connective tissue, e.g., a ligament, tendon, cartilage, basement membrane tissue, and the second tissue is a different type of connective tissue, e.g., if
30 the first tissue is basement membrane, the second tissue can be, e.g., a ligament, tendon, or cartilage. In yet another preferred embodiment, the first tissue is epithelial tissue and the second tissue is connective tissue, e.g., a basement membrane tissue. In preferred embodiments, the first and second tissue are present in the skin, heart, eye, joints (e.g., the articular cartilage of joints), or hair follicle of the subject.

35

5 In another aspect, the invention features a method of treating a subject who has had, is having, or will have a transplant. As used herein, a transplant encompasses the transfer of a non-artificial (biological) tissue from one area of the body to another, the transfer of a non-artificial tissue from one subject to another, or the transfer of an artificial or semi-artificial implantable device to the body of a
10 subject, e.g., a human or non-human animal. The method includes modulating the level of collagen XXII between a first tissue and a second transplanted tissue or implantable device. The transplanted tissue can be autologous, allogeneic, or xenogeneic.

 The level of collagen XXII can be modulated, e.g., increased or decreased, by
15 any of the methods described hereinabove. In some embodiments, the collagen XXII is increased in conjunction with increasing another collagen type, e.g., any one or more of collagen types I, II, III, or IV.

 In a preferred embodiment, collagen XXII which is exogenous to one or two of the tissues is administered.

20 If the transplanted tissue is a non-artificial tissue, the first tissue and the second tissue can be the same tissue type or they can be different tissue types. For example, the first and second tissue can be bone, connective tissue, (e.g., cartilage, tendon, ligament, basement membranes), muscle (e.g., muscle of the heart, eye, or hair follicle), or epithelia. In a preferred embodiment, the first tissue is muscle and the
25 second tissue is connective tissue, e.g., cartilage, tendon, or basement membrane tissue, bone, or another muscle tissue. In another preferred embodiment, the first tissue is bone and the second tissue is cartilage, ligament, muscle, or another bone tissue. In another preferred embodiment, the first tissue is connective tissue, e.g., a ligament, tendon, cartilage, basement membrane tissue, and the second tissue is a
30 different type of connective tissue, e.g., if the first tissue is basement membrane, the second tissue can be, e.g., a ligament, tendon, or cartilage. In yet another embodiment, the first tissue is epithelia and the second tissue is connective tissue, e.g., basement membrane tissue. In preferred embodiments, the transplanted tissue is an eye tissue, e.g., a lens; a heart tissue, e.g., a heart valve; a cartilage tissue, e.g., a tendon; a
35 muscle tissue; a hair follicle; an epithelial tissue, e.g., a skin graft.

5 In other preferred embodiments, the transplanted device is an artificial device, e.g., a non-biological or semi-biological device (e.g., a device containing biological and non-biological components). For example, the transplanted device can be an artificial joint, e.g., an artificial knee or hip; an artificial eye or portion thereof, e.g., lens; an artificial heart or portion thereof, e.g., a heart valve; an artificial hair follicle;
10 an artificial muscle; an artificial connective tissue, e.g., an artificial cartilage, tendon, ligament, or basement membrane tissue; an artificial bone tissue; an artificial tooth; a limb prosthesis.

In some embodiments, an implantable device described herein is contacted with, e.g., coated with, collagen XXII. The implantable device can be contacted with
15 collagen XXII before, during, and/or after the transplant into the subject.

In some embodiments, the collagen XXII is increased in conjunction with increasing another collagen type, e.g., one or more of collagen type I, II, III, or IV.

In another aspect, the invention features a method of promoting wound repair.
20 The method includes identifying a tissue in need of repair, e.g., a collagenous tissue, e.g., cornea, skin, bone, connective tissue, e.g., cartilage, ligament, tendon or basement membrane; and contacting the tissue with an agent that increases collagen XXII, e.g., an agent described hereinabove. The wound can be caused by trauma (e.g., a connective tissue or muscle tear), an incision, or surgery, e.g., transplant
25 surgery, e.g., heart transplant, lens transplant, joint replacement, hair follicle replacement, skin graft. In preferred embodiments, the wound is present in the eye, e.g., the wound is an incision in the lens, e.g., as a result of lens replacement surgery. In another embodiment, the wound is present in a heart tissue, e.g., a heart valve, e.g., as a result of a heart valve replacement or transplant. In yet another embodiment, the
30 wound is present in a connective tissue, e.g., in a cartilage, tendon, or ligament. In another embodiment, the wound is present in an epithelial tissue, e.g., in the dermis. The wound can also be present in a muscle tissue.

In a preferred embodiment, collagen XXII that is exogenous to the wound tissue is administered. In one embodiment, collagen XXII is normally present in the
35 wound tissue in nature. In another embodiment, collagen XXII is not present in the wound tissue in nature.

5 In a preferred embodiment, the tissue is an eye tissue. For example, the tissue, e.g., cornea, has been subjected to trauma, e.g., a surgical incision, cornea transplant surgery, LASIK flap reattachment, cataract surgery, laser surgery, keratoplasty, penetrating keratoplasty, posterior lamellar keratoplasty, refractive surgery, cornea reshaping, or treatment of corneal laceration.

10 In a preferred embodiment, the tissue is a connective tissue, e.g., cartilage, tendon, or ligament. In one embodiment, the tissue has been torn, stretched, or subjected to other trauma.

 In a preferred embodiment, the method further includes administration of an adjunctive therapy, e.g., administration of antibiotics. In another embodiment, the
15 collagen XXII is increased in conjunction with another collagen type, e.g., one or more of collagen type I, I, III, or IV.

 In another aspect, the invention provides methods for evaluating the efficacy of a therapeutic or prophylactic agent. The method includes: contacting a sample
20 with an agent (e.g., a compound identified using the methods described herein, a cytotoxic agent) and, evaluating the expression of collagen XXII nucleic acid or polypeptide in the sample before and after the contacting step. A change, e.g., a decrease or increase, in the level of collagen XXII nucleic acid (e.g., mRNA) or polypeptide in the sample obtained after the contacting step, relative to the level of
25 expression in the sample before the contacting step, is indicative of the efficacy of the agent. The level of collagen XXII nucleic acid or polypeptide expression can be detected by any method described herein. In a preferred embodiment, the sample includes cells obtained from muscle, tendon, cartilage, hair follicle, ocular ciliary body, or heart tissue.

30 In further aspect, the invention provides assays for determining the presence or absence of a genetic alteration in a collagen XXII polypeptide or nucleic acid molecule, including for disease diagnosis.

 In another aspect, the invention features a two dimensional array having a plurality of addresses, each address of the plurality being positionally distinguishable
35 from each other address of the plurality, and each address of the plurality having a unique capture probe, e.g., a nucleic acid or peptide sequence. At least one address of

5 the plurality has a capture probe that recognizes a collagen XXII molecule. In one embodiment, the capture probe is a nucleic acid, e.g., a probe complementary to a collagen XXII nucleic acid sequence. In another embodiment, the capture probe is a polypeptide, e.g., an antibody specific for collagen XXII polypeptides. Also featured is a method of analyzing a sample by contacting the sample to the aforementioned
10 array and detecting binding of the sample to the array.

In another aspect, the invention features a biomaterial containing collagen XXII, e.g., a recombinant collagen XXII as described herein, which biomaterial is biocompatible and can be, e.g., implanted in the human body. Examples of collagen XXII containing biomaterials of the invention include prostheses; bioengineered
15 tissue, e.g., for tissue repair, replacement, or to provide support of endogenous tissue where weakness exists; coatings, e.g., to improve biological functionality or biocompatibility of artificial implants; bio-compatible gelatins, gels, paste, strips, sponges, strings, fibers, tubes, and films, e.g., as carriers for drug delivery.

Other features and advantages of the invention will be apparent from the
20 following detailed description, and from the claims.

Description of the Drawings

Figure 1A-C shows the collagen XXII cDNA sequence (SEQ ID NO:1). The coding sequence of collagen XXII starts at nucleotide 448 (the start site ATG is
25 double underlined) and ends at nucleotide 5326 of SEQ ID NO:1. Nucleotides 3458-3518 of SEQ ID NO:1 (underlined) are absent in an alternatively spliced variant.

Figure 2A-B shows the collagen XXII amino acid sequence (SEQ ID NO:2). The shaded area highlights the G-X-Y repeats (triple helix domain). Non-triple helical interruptions are double underlined. A non-triple-helical alternatively spliced
30 interruption is bold and underlined. The predicted signal peptide is bolded.

5

Detailed Description

The human collagen XXII sequence (see SEQ ID NO:1, as recited in Example 1), which is approximately 6352 nucleotides long including untranslated regions, contains a predicted methionine-initiated coding sequence of about 4878 nucleotides. The coding sequence encodes a 1626 amino acid protein (see SEQ ID NO:2, as recited in Example 1). An alternatively spliced variant lacks nucleotides 3458-3518 of SEQ ID NO:1 and encodes an amino acid sequence that lacks amino acids 1005 –1024 of SEQ ID NO:2.

Human collagen XXII contains the following regions or other structural features: a predicted signal peptide, from about amino acid 1-27 of SEQ ID NO:2; a core helical domain rich in G-X-Y repeats from about amino acid 1479-1603 of SEQ ID NO:2; six non-helical interruptions of the core helical domain at about amino acids 853-864, 1004-1024 (present in alternatively spliced variant), 1097-1119, 1457-1495, 1550-1552, and 1572-1575 of SEQ ID NO:2

The collagen XXII protein is a member of the broad family of collagens. The term “family” when referring to the protein and nucleic acid molecules of the invention means two or more proteins or nucleic acid molecules having a common structural domain or motif. Collagens are characterized by a central domain rich G-X-Y triple repeats, which domain allows for the formation of collagen triple helix molecules made up of three collagen polypeptide chains wound around each other. Collagens typically have globular N-terminal and C-terminal domains, and the ability to form higher order collagen fibrils and/or fibers. Human collagen XXII described herein shares these characteristics of the collagen superfamily.

Type XXII collagens described herein likely comprise a subfamily of collagens. Type XXII collagens can be naturally or non-naturally occurring and can be from either the same or different species. For example, the collagen XXII subfamily can contain a first protein of human origin as well as other distinct proteins of human origin, or alternatively, can contain homologues of non-human origin, e.g., rat or mouse proteins. Members of the collagen type XXII subfamily can also have common functional characteristics.

5

As used herein, the term “core helical domain” is an amino acid sequence of at least 300 amino acids, preferably at least 500, 600, 700, 800, 900, or at least 1000 amino acids, that is capable of forming a helical structure. A core helical domain is typically rich in Glycine-X-Y repeats and is also rich in proline residues. A core
10 helical domain can have at least one non-helical sequence interruption. A non-helical sequence interruption is an amino acid sequence of between 3 and 100, preferably between about 3 and 70, more preferably between about 3 and 50, amino acid residues in length, that is present in the core helical domain but that is not predicted to form a helical structure. A collagen XXII described herein can have between one and 15,
15 preferably between 3 and 10, more preferably about 5 or 6 non-helical sequence interruptions that are interspersed throughout the core helical domain but that are not predicted to form a helical structure.

In a preferred embodiment collagen XXII polypeptide or protein has a “core helical domain” of between 500 and 1500, preferably between 700 and 1300, more
20 preferably about 1100 amino acid residues, having between 3 and 10, preferably about 5 or 6, non-helical sequence interruptions. The non-helical sequence interruptions can range in length from 3 to 100 amino acid residues, preferably between 3 and 50 amino acid residues.

To identify the presence of a “helical core domain” domain in a collagen XXII
25 protein sequence, and make the determination that a polypeptide or protein of interest has a particular profile, the amino acid sequence of the protein can be searched against the Pfam database of HMMs (e.g., the Pfam database, release 2.1) using the default parameters (http://www.sanger.ac.uk/Software/Pfam/HMM_search). For example, the hmmsf program, which is available as part of the HMMER package of
30 search programs, is a family specific default program for MILPAT0063 and a score of 15 is the default threshold score for determining a hit. Alternatively, the threshold score for determining a hit can be lowered (e.g., to 8 bits). A description of the Pfam database can be found in Sonhammer *et al.* (1997) *Proteins* 28(3):405-420 and a detailed description of HMMs can be found, for example, in Gribskov *et al.* (1990)
35 *Meth. Enzymol.* 183:146-159; Gribskov *et al.* (1987) *Proc. Natl. Acad. Sci. USA* 84:4355-4358; Krogh *et al.* (1994) *J. Mol. Biol.* 235:1501-1531; and Stultz *et*

- 5 *al.*(1993) *Protein Sci.* 2:305-314, the contents of which are incorporated herein by reference.

Collagen XXII is expressed primarily at the boundary of muscle and other tissue, e.g., connective tissue. In particular, collagen XXII is expressed in myo-
10 tendonous junctions; beneath the articular cartilage of joints; in the basement membrane of the outer root sheath at the base of hair follicles; in the ciliary body of the eye, e.g., where the muscle of the lens intersects the zonular fibers; in the heart, e.g., where the heart muscles intersect the cartilage ring in the region of the valves. As the collagen XXII polypeptides of the invention may modulate collagen XXII-
15 mediated activities, they may be useful as of for developing novel diagnostic and therapeutic agents for collagen XXII-mediated or related disorders, as described below.

As used herein, a "collagen XXII activity", "biological activity of collagen XXII" or "functional activity of collagen XXII", refers to an activity exerted by a
20 collagen XXII protein, polypeptide or nucleic acid molecule. For example, a collagen XXII activity can be an activity exerted by collagen XXII in a physiological milieu on, e.g., a collagen XXII-responsive cell or tissue, e.g., a cartilage tissue, a muscle tissue, a connective tissue, e.g., a tendon or other muscle-tissue junction in, e.g., a joint, the eye, the heart or the hair follicle. A collagen XXII activity can be
25 determined *in vivo* or *in vitro*. In one embodiment, a collagen XXII activity is a direct activity, such as an association with a collagen XXII binding partner. A "target molecule" or "binding partner" is a molecule with which a collagen XXII protein binds or interacts in nature.

A collagen XXII activity can also be an indirect activity, e.g., a cellular
30 signaling activity mediated by interaction of the collagen XXII protein with a collagen XXII receptor. The collagen XXII proteins of the present invention can have one or more of the following activities: (1) forms a helix, e.g., an interrupted helix; (2) forms a triple helix consisting of three collagen XXII polypeptide chains; (3) forms higher order collagen XXII fibrils and/or fibers; (4) forms at least 1, preferably
35 between 3 and 10 non-helical interruptions or domains within the core helical domain; (5) provides tensile strength and/or flexibility to tissue boundaries, e.g.,

5 muscle-connective tissue boundaries; (6) provides form to a tissue, e.g., during growth and/or development, (7) separates or anchors cell layers or tissue, e.g., at muscle-tissue junctions, dermal-epidermal junctions, basement membranes that separate epithelial and mesenchymal structures; (8) provides a filtration barrier between tissues; (9) provides transparency, e.g., in the eye; (10) is regulated, in part,
10 by alternative splicing; (11) is recognized by R34 antibody described herein.

Thus, the collagen XXII molecules can act as novel diagnostic targets and therapeutic agents for controlling collagen related disorders or conditions. Examples of such disorders include osteoarthritis; osteogenesis imperfecta; dystrophic epidermolysis bullosa; polymyositis and dermatomyositis ("PM/DM"). Further, as
15 collagen XXII is expressed in muscle-cartilage boundaries in heart valves, collagen XXII or functional fragments thereof may be useful in treating disorders of the heart valve or other regions of the heart, or in making, treating, or maintaining transplantable biological equivalents of the hearts of parts thereof. collagen XXII may serve to increase the success of such transplants, or increase the useful lifetime of
20 the heart and its valves, or increase the success of heart surgeries.

As collagen XXII is expressed in muscle-connective tissue boundaries in the eye, collagen XXII or functional fragments thereof may be useful in treating disorders of the eye, or in making, treating, or maintaining transplantable biological equivalents of the eye of parts thereof. Collagen XXII may serve to increase the
25 success of such transplants, e.g., lens replacements, or increase the useful lifetime of the lens and its attachments, or increase the success of eye surgeries,, e.g., to treat surgery induced endophthalmitis.

As collagen XXII is expressed in cartilage, collagen XXII or functional fragments thereof may be useful to treat cartilage-related disorders or conditions, or in
30 making, treating, or maintaining transplantable biological equivalents of the cartilage of parts thereof. Collagen XXII may serve to increase the success of such transplants, or increase the useful lifetime of the cartilage and its articular surface, or increase the success of, e.g., joint surgeries and cartilage replacement.

As collagen XXII is also expressed in tendon and myotendonous junctions, or
35 at junctions of tendons with cartilage or with perichondrium or with bone, collagen XXII or functional fragments thereof may be used to treat tendons or muscle disorders

5 or conditions, or to make, treat, or maintain transplantable biological equivalents of the tendons or muscles or of parts thereof. Collagen XXII may serve to increase the success of such transplants, or increase the useful lifetime of the tendons or the junctions of tendons with muscle or tendons with cartilage, or increase the success of joint surgeries involving tendons, or of tendon replacement or repair.

10 As collagen XXII is also expressed in the hair follicle, collagen XXII or functional fragments thereof may be used to treat hair related conditions, e.g., hair loss. In transplantable biological equivalents of the hair follicle or parts thereof, collagen XXII may serve to increase the success of such transplants, or increase the growth or stability of the hair follicle, or extend the anagen phase of the hair cycle.

15 Downregulation of collagen XXII may destabilize the integrity of the follicle in the dermis and result in hair loss.

The collagen XXII protein, fragments thereof, and derivatives and other variants of the sequence in SEQ ID NO:2 thereof are collectively referred to as

20 "polypeptides or proteins of the invention" or "collagen XXII polypeptides or proteins". Nucleic acid molecules encoding such polypeptides or proteins are collectively referred to as "nucleic acids of the invention" or "collagen XXII nucleic acids." Collagen XXII molecules refer to collagen XXII nucleic acids, polypeptides, and antibodies.

25 As used herein, the term "nucleic acid molecule" includes DNA molecules (e.g., a cDNA or genomic DNA), RNA molecules (e.g., an mRNA) and analogs of the DNA or RNA. A DNA or RNA analog can be synthesized from nucleotide analogs. The nucleic acid molecule can be single-stranded or double-stranded, but preferably is double-stranded DNA.

30 The term "isolated nucleic acid molecule" or "purified nucleic acid molecule" includes nucleic acid molecules that are separated from other nucleic acid molecules present in the natural source of the nucleic acid. For example, with regards to genomic DNA, the term "isolated" includes nucleic acid molecules which are separated from the chromosome with which the genomic DNA is naturally associated.

35 Preferably, an "isolated" nucleic acid is free of sequences which naturally flank the nucleic acid (i.e., sequences located at the 5' and/or 3' ends of the nucleic acid) in the

5 genomic DNA of the organism from which the nucleic acid is derived. For example, in various embodiments, the isolated nucleic acid molecule can contain less than about 5 kb, 4 kb, 3 kb, 2 kb, 1 kb, 0.5 kb or 0.1 kb of 5' and/or 3' nucleotide sequences which naturally flank the nucleic acid molecule in genomic DNA of the cell from which the nucleic acid is derived. Moreover, an "isolated" nucleic acid
10 molecule, such as a cDNA molecule, can be substantially free of other cellular material, or culture medium when produced by recombinant techniques, or substantially free of chemical precursors or other chemicals when chemically synthesized.

As used herein, the term "hybridizes under low stringency, medium
15 stringency, high stringency, or very high stringency conditions" describes conditions for hybridization and washing. Guidance for performing hybridization reactions can be found in *Current Protocols in Molecular Biology*, John Wiley & Sons, N.Y. (1989), 6.3.1-6.3.6, which is incorporated by reference. Aqueous and nonaqueous methods are described in that reference and either can be used. Specific hybridization
20 conditions referred to herein are as follows: 1) low stringency hybridization conditions in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by two washes in 0.2X SSC, 0.1% SDS at least at 50°C (the temperature of the washes can be increased to 55°C for low stringency conditions); 2) medium stringency hybridization conditions in 6X SSC at about 45°C, followed by one or more washes in
25 0.2X SSC, 0.1% SDS at 60°C; 3) high stringency hybridization conditions in 6X SSC at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 65°C; and preferably 4) very high stringency hybridization conditions are 0.5M sodium phosphate, 7% SDS at 65°C, followed by one or more washes at 0.2X SSC, 1% SDS at 65°C. Very high stringency conditions (4) are the preferred conditions and the ones
30 that should be used unless otherwise specified.

Preferably, an isolated nucleic acid molecule of the invention that hybridizes under a stringency condition described herein to the sequence of SEQ ID NO:1 or SEQ ID NO:3, corresponds to a naturally-occurring nucleic acid molecule.

5 As used herein, a "naturally-occurring" nucleic acid molecule refers to an RNA or DNA molecule having a nucleotide sequence that occurs in nature. For example a naturally occurring nucleic acid molecule can encode a natural protein.

 As used herein, the terms "gene" and "recombinant gene" refer to nucleic acid molecules which include at least an open reading frame encoding a collagen XXII
10 protein. The gene can optionally further include non-coding sequences, e.g., regulatory sequences and introns. Preferably, a gene encodes a mammalian collagen XXII protein or derivative thereof.

 An "isolated" or "purified" polypeptide or protein is substantially free of cellular material or other contaminating proteins from the cell or tissue source from
15 which the protein is derived, or substantially free from chemical precursors or other chemicals when chemically synthesized. "Substantially free" means that a preparation of collagen XXII protein is at least 10% pure. In a preferred embodiment, the preparation of collagen XXII protein has less than about 30%, 20%, 10% and more preferably 5% (by dry weight), of non-collagen XXII protein (also referred to
20 herein as a "contaminating protein"), or of chemical precursors or non-collagen XXII chemicals. When the collagen XXII protein or biologically active portion thereof is recombinantly produced, it is also preferably substantially free of culture medium, i.e., culture medium represents less than about 20%, more preferably less than about 10%, and most preferably less than about 5% of the volume of the protein preparation.
25 The invention includes isolated or purified preparations of at least 0.01, 0.1, 1.0, and 10 milligrams in dry weight.

 A "non-essential" amino acid residue is a residue that can be altered from the wild-type sequence of collagen XXII without abolishing or substantially altering a collagen XXII activity. Preferably the alteration does not substantially alter the
30 collagen XXII activity, e.g., the activity is at least 20%, 40%, 60%, 70% or 80% of wild-type. An "essential" amino acid residue is a residue that, when altered from the wild-type sequence of collagen XXII, results in abolishing a collagen XXII activity such that less than 20% of the wild-type activity is present. For example, conserved amino acid residues in collagen XXII are predicted to be particularly unamenable to
35 alteration.

5 A "conservative amino acid substitution" is one in which the amino acid residue is replaced with an amino acid residue having a similar side chain. Families of amino acid residues having similar side chains have been defined in the art. These families include amino acids with basic side chains (e.g., lysine, arginine, histidine), acidic side chains (e.g., aspartic acid, glutamic acid), uncharged polar side chains
10 (e.g., glycine, asparagine, glutamine, serine, threonine, tyrosine, cysteine), nonpolar side chains (e.g., alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan), beta-branched side chains (e.g., threonine, valine, isoleucine) and aromatic side chains (e.g., tyrosine, phenylalanine, tryptophan, histidine). Thus, a predicted nonessential amino acid residue in a collagen XXII
15 protein is preferably replaced with another amino acid residue from the same side chain family. Alternatively, in another embodiment, mutations can be introduced randomly along all or part of a collagen XXII coding sequence, such as by saturation mutagenesis, and the resultant mutants can be screened for collagen XXII biological activity to identify mutants that retain activity. Following mutagenesis of SEQ ID
20 NO:1 or SEQ ID NO:3, the encoded protein can be expressed recombinantly and the activity of the protein can be determined.

As used herein, a "biologically active portion" of a collagen XXII protein includes a fragment of a collagen XXII protein which participates in an interaction, e.g., an intramolecular or an inter-molecular interaction. An inter-molecular
25 interaction can be a specific binding interaction or an enzymatic interaction (e.g., the interaction can be transient and a covalent bond is formed or broken). An inter-molecular interaction can be between a collagen XXII molecule and a non-collagen XXII molecule or between a first collagen XXII molecule and a second and/or third collagen XXII molecule (e.g., a dimer or trimer interaction). Biologically active
30 portions of a collagen XXII protein include peptides comprising amino acid sequences sufficiently homologous to or derived from the amino acid sequence of the collagen XXII protein, e.g., the amino acid sequence shown in SEQ ID NO:2, which include less amino acids than the full length collagen XXII proteins, and exhibit at least one activity of a collagen XXII protein. Typically, biologically active portions
35 comprise a domain or motif with at least one activity of the collagen XXII protein, e.g., triple helix formation or fibril formation. A biologically active portion of a

5 collagen XXII protein can be a polypeptide which is, for example, 10, 25, 50, 100, 200 or more amino acids in length. Biologically active portions of a collagen XXII protein can be used as targets for developing agents which modulate a collagen XXII mediated activity, e.g., tensile strength.

Calculations of homology or sequence identity between sequences (the terms
10 are used interchangeably herein) are performed as follows.

To determine the percent identity of two amino acid sequences, or of two nucleic acid sequences, the sequences are aligned for optimal comparison purposes (e.g., gaps can be introduced in one or both of a first and a second amino acid or nucleic acid sequence for optimal alignment and non-homologous sequences can be
15 disregarded for comparison purposes). In a preferred embodiment, the length of a reference sequence aligned for comparison purposes is at least 30%, preferably at least 40%, more preferably at least 50%, 60%, and even more preferably at least 70%, 80%, 90%, 100% of the length of the reference sequence. The amino acid residues or nucleotides at corresponding amino acid positions or nucleotide positions are then
20 compared. When a position in the first sequence is occupied by the same amino acid residue or nucleotide as the corresponding position in the second sequence, then the molecules are identical at that position (as used herein amino acid or nucleic acid "identity" is equivalent to amino acid or nucleic acid "homology").

The percent identity between the two sequences is a function of the number of
25 identical positions shared by the sequences, taking into account the number of gaps, and the length of each gap, which need to be introduced for optimal alignment of the two sequences.

The comparison of sequences and determination of percent identity between two sequences can be accomplished using a mathematical algorithm. In a preferred
30 embodiment, the percent identity between two amino acid sequences is determined using the Needleman and Wunsch ((1970) *J. Mol. Biol.* 48:444-453) algorithm which has been incorporated into the GAP program in the GCG software package (available at <http://www.gcg.com>), using either a Blossum 62 matrix or a PAM250 matrix, and a gap weight of 16, 14, 12, 10, 8, 6, or 4 and a length weight of 1, 2, 3, 4, 5, or 6. In yet
35 another preferred embodiment, the percent identity between two nucleotide sequences is determined using the GAP program in the GCG software package (available at

5 <http://www.gcg.com>), using a NWSgapdna.CMP matrix and a gap weight of 40, 50, 60, 70, or 80 and a length weight of 1, 2, 3, 4, 5, or 6. A particularly preferred set of parameters (and the one that should be used unless otherwise specified) are a Blossum 62 scoring matrix with a gap penalty of 12, a gap extend penalty of 4, and a frameshift gap penalty of 5.

10 The percent identity between two amino acid or nucleotide sequences can be determined using the algorithm of E. Meyers and W. Miller ((1989) CABIOS, 4:11-17) which has been incorporated into the ALIGN program (version 2.0), using a PAM120 weight residue table, a gap length penalty of 12 and a gap penalty of 4.

The nucleic acid and protein sequences described herein can be used as a
15 "query sequence" to perform a search against public databases to, for example, identify other family members or related sequences. Such searches can be performed using the NBLAST and XBLAST programs (version 2.0) of Altschul, *et al.* (1990) *J. Mol. Biol.* 215:403-10. BLAST nucleotide searches can be performed with the NBLAST program, score = 100, wordlength = 12 to obtain nucleotide sequences
20 homologous to collagen XXII nucleic acid molecules of the invention. BLAST protein searches can be performed with the XBLAST program, score = 50, wordlength = 3 to obtain amino acid sequences homologous to collagen XXII protein molecules of the invention. To obtain gapped alignments for comparison purposes, Gapped BLAST can be utilized as described in Altschul *et al.*, (1997) *Nucleic Acids*
25 *Res.* 25:3389-3402. When utilizing BLAST and Gapped BLAST programs, the default parameters of the respective programs (e.g., XBLAST and NBLAST) can be used. See <http://www.ncbi.nlm.nih.gov>.

Particularly preferred collagen XXII polypeptides of the present invention have an amino acid sequence substantially identical to the amino acid sequence of
30 SEQ ID NO:2. In the context of an amino acid sequence, the term "substantially identical" is used herein to refer to a first amino acid that contains a sufficient or minimum number of amino acid residues that are i) identical to, or ii) conservative substitutions of aligned amino acid residues in a second amino acid sequence such that the first and second amino acid sequences can have a common structural domain
35 and/or common functional activity. For example, amino acid sequences that contain a common structural domain having at least about 60%, or 65% identity, likely 75%

5 identity, more likely 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% identity to SEQ ID NO:2 are termed substantially identical.

In the context of nucleotide sequence, the term "substantially identical" is used herein to refer to a first nucleic acid sequence that contains a sufficient or minimum number of nucleotides that are identical to aligned nucleotides in a second nucleic acid sequence such that the first and second nucleotide sequences encode a polypeptide having common functional activity, or encode a common structural polypeptide domain or a common functional polypeptide activity. For example, nucleotide sequences having at least about 60%, or 65% identity, likely 75% identity, more likely 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% identity to SEQ ID NO:1 or 3 are termed substantially identical.

"Misexpression or aberrant expression", as used herein, refers to a non-wildtype pattern of gene expression at the RNA or protein level. It includes: expression at non-wild type levels, i.e., over- or under-expression; a pattern of expression that differs from wild type in terms of the time or stage at which the gene is expressed, e.g., increased or decreased expression (as compared with wild type) at a predetermined developmental period or stage; a pattern of expression that differs from wild type in terms of altered, e.g., increased or decreased, expression (as compared with wild type) in a predetermined cell type or tissue type; a pattern of expression that differs from wild type in terms of the splicing size, translated amino acid sequence, post-translational modification, or biological activity of the expressed polypeptide; a pattern of expression that differs from wild type in terms of the effect of an environmental stimulus or extracellular stimulus on expression of the gene, e.g., a pattern of increased or decreased expression (as compared with wild type) in the presence of an increase or decrease in the strength of the stimulus.

30 "Subject," as used herein, refers to human and non-human animals. The term "non-human animals" of the invention includes all vertebrates, e.g., mammals, such as non-human primates (particularly higher primates), sheep, dog, rodent (e.g., mouse or rat), guinea pig, goat, pig, cat, rabbits, cow, and non-mammals, such as chickens, amphibians, reptiles, etc. In a preferred embodiment, the subject is a human. In another embodiment, the subject is an experimental animal or animal suitable as a disease model.

5 A "purified preparation of cells", as used herein, refers to an in vitro preparation of cells. In the case cells from multicellular organisms (e.g., plants and animals), a purified preparation of cells is a subset of cells obtained from the organism, not the entire intact organism. In the case of unicellular microorganisms (e.g., cultured cells and microbial cells), it consists of a preparation of at least 10%
10 and more preferably 50% of the subject cells.

Various aspects of the invention are described in further detail below.

Isolated Nucleic Acid Molecules

In one aspect, the invention provides, an isolated or purified, nucleic acid molecule that encodes a collagen XXII polypeptide described herein, e.g., a full-
15 length collagen XXII protein or a fragment thereof, e.g., a biologically active portion of collagen XXII protein. Also included is a nucleic acid fragment suitable for use as a hybridization probe, which can be used, e.g., to identify a nucleic acid molecule encoding a polypeptide of the invention, collagen XXII mRNA, and fragments suitable for use as primers, e.g., PCR primers for the amplification or mutation of
20 nucleic acid molecules.

In one embodiment, an isolated nucleic acid molecule of the invention includes the nucleotide sequence shown in SEQ ID NO:1, or a portion of any of these nucleotide sequences. In one embodiment, the nucleic acid molecule includes sequences encoding the human collagen XXII protein (i.e., "the coding region" of
25 SEQ ID NO:1, as shown in SEQ ID NO:3, and alternatively spliced variants thereof), as well as 5' untranslated sequences. Alternatively, the nucleic acid molecule can include only the coding region of SEQ ID NO:1 (e.g., SEQ ID NO:3 or an alternatively spliced variant thereof) and, e.g., no flanking sequences which normally accompany the subject sequence.

30 In another embodiment, an isolated nucleic acid molecule of the invention includes a nucleic acid molecule which is a complement of the nucleotide sequence shown in SEQ ID NO:1 or SEQ ID NO:3, or a portion of any of these nucleotide sequences. In other embodiments, the nucleic acid molecule of the invention is sufficiently complementary to the nucleotide sequence shown in SEQ ID NO:1 or

- 5 SEQ ID NO:3, such that it can hybridize (e.g., under a stringency condition described herein) to the nucleotide sequence shown in SEQ ID NO:1 or 3, thereby forming a stable duplex.

In one embodiment, an isolated nucleic acid molecule of the present invention includes a nucleotide sequence which is at least about: 60%, 65%, 70%, 75%, 80%,
10 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or more homologous to the entire length of the nucleotide sequence shown in SEQ ID NO:1 or SEQ ID NO:3, or a portion, preferably of the same length, of any of these nucleotide sequences.

Collagen XXII Nucleic Acid Fragments

- 15 A nucleic acid molecule of the invention can include only a portion of the nucleic acid sequence of SEQ ID NO:1 or 3. For example, such a nucleic acid molecule can include a fragment which can be used as a probe or primer or a fragment encoding a portion of a collagen XXII protein, e.g., an immunogenic or biologically active portion of a collagen XXII protein. A fragment can comprise those nucleotides
20 of SEQ ID NO:1, which encode a helix forming domain of human collagen XXII. The nucleotide sequence determined from the cloning of the collagen XXII gene allows for the generation of probes and primers designed for use in identifying and/or cloning other collagen XXII family members, or fragments thereof, as well as collagen XXII homologues, or fragments thereof, from other species.

- 25 In another embodiment, a nucleic acid includes a nucleotide sequence that includes part, or all, of the coding region and extends into either (or both) the 5' or 3' noncoding region. Other embodiments include a fragment which includes a nucleotide sequence encoding an amino acid fragment described herein. Nucleic acid fragments can encode a specific domain or site described herein or fragments thereof, particularly
30 fragments thereof which are at least 50 amino acids in length. Fragments also include nucleic acid sequences corresponding to specific amino acid sequences described above or fragments thereof. Nucleic acid fragments should not be construed as encompassing those fragments that may have been disclosed prior to the invention.

5 A nucleic acid fragment can include a sequence corresponding to a domain, region, or functional site described herein. A nucleic acid fragment can also include one or more domain, region, or functional site described herein. Thus, for example, a collagen XXII nucleic acid fragment can include a sequence corresponding to a core helical domain or a non-helical interruption.

10 Collagen XXII probes and primers are provided. Typically a probe/primer is an isolated or purified oligonucleotide. The oligonucleotide typically includes a region of nucleotide sequence that hybridizes under a stringency condition described herein to at least about 7, 12 or 15, preferably about 20 or 25, more preferably about 30, 35, 40, 45, 50, 55, 60, 65, or 75 consecutive nucleotides of a sense or antisense
15 sequence of SEQ ID NO:1 or SEQ ID NO:3, or of a naturally occurring allelic variant or mutant of SEQ ID NO:1 or SEQ ID NO:3. Preferably, an oligonucleotide is less than about 200, 150, 120, or 100 nucleotides in length.

 In one embodiment, the probe or primer is attached to a solid support, e.g., a solid support described herein.

20 One exemplary kit of primers includes a forward primer that anneals to the coding strand and a reverse primer that anneals to the non-coding strand. The forward primer can anneal to the start codon, e.g., the nucleic acid sequence encoding amino acid residue 1 of SEQ ID NO:2. The reverse primer can anneal to the ultimate codon, e.g., the codon immediately before the stop codon. In a preferred embodiment, the
25 annealing temperatures of the forward and reverse primers differ by no more than 5, 4, 3, or 2°C.

 In a preferred embodiment the nucleic acid is a probe which is at least 10, 12, 15, 18, 20 and less than 200, more preferably less than 100, or less than 50, nucleotides in length. It should be identical, or differ by 1, or 2, or less than 5 or 10
30 nucleotides, from a sequence disclosed herein. If alignment is needed for this comparison the sequences should be aligned for maximum homology. "Looped" out sequences from deletions or insertions, or mismatches, are considered differences.

 A probe or primer can be derived from the sense or anti-sense strand of a nucleic acid which encodes: a predicted signal peptide, from about amino acid 1-27 of
35 SEQ ID NO:2; a core helical domain rich in G-X-Y repeats from about amino acid 1479-1603 of SEQ ID NO:2; or any of six non-helical interruptions of the core helical

5 domain at about amino acids 853-864, 1004-1024 (present in alternatively spliced variant), 1097-1119, 1457-1495, 1550-1552, and 1572-1575 of SEQ ID NO:2

In another embodiment a set of primers is provided, e.g., primers suitable for use in a PCR, which can be used to amplify a selected region of a collagen XXII sequence, e.g., a domain, region, site or other sequence described herein. The primers
10 should be at least 5, 10, or 50 base pairs in length and less than 100, or less than 200, base pairs in length. The primers should be identical, or differs by one base from a sequence disclosed herein or from a naturally occurring variant. For example, primers suitable for amplifying all or a portion of any of the following regions are provided: a
15 domain rich in G-X-Y repeats from about amino acid 1479-1603 of SEQ ID NO:2; or any of six non-helical interruptions of the core helical domain at about amino acids 853-864, 1004-1024 (present in alternatively spliced variant), 1097-1119, 1457-1495, 1550-1552, and 1572-1575 of SEQ ID NO:2

A nucleic acid fragment can encode an epitope bearing region of a polypeptide
20 described herein.

A nucleic acid fragment encoding a "biologically active portion of a collagen XXII polypeptide" can be prepared by isolating a portion of the nucleotide sequence of SEQ ID NO:1 or 3, which encodes a polypeptide having a collagen XXII biological activity (e.g., the biological activities of the collagen XXII proteins are described
25 herein), expressing the encoded portion of the collagen XXII protein (e.g., by recombinant expression *in vitro*) and assessing the activity of the encoded portion of the collagen XXII protein. For example, a nucleic acid fragment encoding a biologically active portion of collagen XXII includes a core helical domain, e.g., amino acid residues about 1479-1606 of SEQ ID NO:2. A nucleic acid fragment
30 encoding a biologically active portion of a collagen XXII polypeptide, may comprise a nucleotide sequence which is greater than 300 or more nucleotides in length.

In preferred embodiments, a nucleic acid includes a nucleotide sequence which is about 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300 or more nucleotides in length and hybridizes under a stringency condition described herein to
35 a nucleic acid molecule of SEQ ID NO:1, or SEQ ID NO:3.

5 Collagen XXII Nucleic Acid Variants

The invention further encompasses nucleic acid molecules that differ from the nucleotide sequence shown in SEQ ID NO:1 or SEQ ID NO:3. Such differences can be due to degeneracy of the genetic code (and result in a nucleic acid which encodes the same collagen XXII proteins as those encoded by the nucleotide sequence
10 disclosed herein). In another embodiment, an isolated nucleic acid molecule of the invention has a nucleotide sequence encoding a protein having an amino acid sequence which differs, by at least 1, but less than 5, 10, 20, 50, or 100 amino acid residues that shown in SEQ ID NO:2. If alignment is needed for this comparison the sequences should be aligned for maximum homology. The encoded protein can differ
15 by no more than 5, 4, 3, 2, or 1 amino acid. "Looped" out sequences from deletions or insertions, or mismatches, are considered differences.

Nucleic acids of the inventor can be chosen for having codons, which are preferred, or non-preferred, for a particular expression system. E.g., the nucleic acid can be one in which at least one codon, at preferably at least 10%, or 20% of the
20 codons has been altered such that the sequence is optimized for expression in *E. coli*, yeast, human, insect, or CHO cells.

Nucleic acid variants can be naturally occurring, such as allelic variants (same locus), homologs (different locus), and orthologs (different organism) or can be non naturally occurring. Non-naturally occurring variants can be made by mutagenesis
25 techniques, including those applied to polynucleotides, cells, or organisms. The variants can contain nucleotide substitutions, deletions, inversions and insertions. Variation can occur in either or both the coding and non-coding regions. The variations can produce both conservative and non-conservative amino acid substitutions (as compared in the encoded product).

30 In a preferred embodiment, the nucleic acid differs from that of SEQ ID NO: 1 or 3, e.g., as follows: by at least one but less than 10, 20, 30, or 40 nucleotides; at least one but less than 1%, 5%, 10% or 20% of the nucleotides in the subject nucleic acid. The nucleic acid can differ by no more than 5, 4, 3, 2, or 1 nucleotide. If necessary for this analysis the sequences should be aligned for maximum homology. "Looped" out
35 sequences from deletions or insertions, or mismatches, are considered differences.

5 Orthologs, homologs, and allelic variants can be identified using methods known in the art. These variants comprise a nucleotide sequence encoding a polypeptide that is 50%, at least about 55%, typically at least about 70-75%, more typically at least about 80-85%, and most typically at least about 90-95% or more identical to the nucleotide sequence shown in SEQ ID NO:2 or a fragment of this sequence. Such nucleic acid
10 molecules can readily be identified as being able to hybridize under a stringency condition described herein, to the nucleotide sequence shown in SEQ ID NO 2 or a fragment of the sequence. Nucleic acid molecules corresponding to orthologs, homologs, and allelic variants of the collagen XXII cDNAs of the invention can further be isolated by mapping to the same chromosome or locus as the collagen XXII gene.

15 Preferred variants include those that are correlated with triple helix forming activity,

 Allelic variants of collagen XXII, e.g., human collagen XXII, include both functional and non-functional proteins. Functional allelic variants are naturally occurring amino acid sequence variants of the collagen XXII protein within a
20 population that maintain the ability to form a trimeric triple helix. Functional allelic variants will typically contain only conservative substitution of one or more amino acids of SEQ ID NO:2, or substitution, deletion or insertion of non-critical residues in non-critical regions of the protein. Non-functional allelic variants are naturally-occurring amino acid sequence variants of the collagen XXII, e.g., human collagen
25 XXII, protein within a population that do not have the ability to form a trimeric triple helix. Non-functional allelic variants will typically contain a non-conservative substitution, a deletion, or insertion, or premature truncation of the amino acid sequence of SEQ ID NO:2, or a substitution, insertion, or deletion in critical residues or critical regions of the protein.

30 Moreover, nucleic acid molecules encoding other collagen XXII family members and, thus, which have a nucleotide sequence which differs from the collagen XXII sequences of SEQ ID NO:1 or SEQ ID NO:3 are intended to be within the scope of the invention.

5 Antisense Nucleic Acid Molecules, Ribozymes and Modified collagen XXII Nucleic Acid Molecules

 In another aspect, the invention features, an isolated nucleic acid molecule which is antisense to collagen XXII. An "antisense" nucleic acid can include a nucleotide sequence which is complementary to a "sense" nucleic acid encoding a protein, e.g., complementary to the coding strand of a double-stranded cDNA molecule or complementary to an mRNA sequence. The antisense nucleic acid can be complementary to an entire collagen XXII coding strand, or to only a portion thereof (e.g., the coding region of human collagen XXII corresponding to SEQ ID NO:3). In another embodiment, the antisense nucleic acid molecule is antisense to a "noncoding region" of the coding strand of a nucleotide sequence encoding collagen XXII (e.g., the 5' and 3' untranslated regions).

 An antisense nucleic acid can be designed such that it is complementary to the entire coding region of collagen XXII mRNA, but more preferably is an oligonucleotide which is antisense to only a portion of the coding or noncoding region of collagen XXII mRNA. For example, the antisense oligonucleotide can be complementary to the region surrounding the translation start site of collagen XXII mRNA, e.g., between the -10 and +10 regions of the target gene nucleotide sequence of interest. An antisense oligonucleotide can be, for example, about 7, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, or more nucleotides in length.

 An antisense nucleic acid of the invention can be constructed using chemical synthesis and enzymatic ligation reactions using procedures known in the art. For example, an antisense nucleic acid (e.g., an antisense oligonucleotide) can be chemically synthesized using naturally occurring nucleotides or variously modified nucleotides designed to increase the biological stability of the molecules or to increase the physical stability of the duplex formed between the antisense and sense nucleic acids, e.g., phosphorothioate derivatives and acridine substituted nucleotides can be used. The antisense nucleic acid also can be produced biologically using an expression vector into which a nucleic acid has been subcloned in an antisense orientation (i.e., RNA transcribed from the inserted nucleic acid will be of an

5 antisense orientation to a target nucleic acid of interest, described further in the following subsection).

The antisense nucleic acid molecules of the invention are typically administered to a subject (e.g., by direct injection at a tissue site), or generated *in situ* such that they hybridize with or bind to cellular mRNA and/or genomic DNA
10 encoding a collagen XXII protein to thereby inhibit expression of the protein, e.g., by inhibiting transcription and/or translation. Alternatively, antisense nucleic acid molecules can be modified to target selected cells and then administered systemically. For systemic administration, antisense molecules can be modified such that they
15 specifically bind to receptors or antigens expressed on a selected cell surface, e.g., by linking the antisense nucleic acid molecules to peptides or antibodies which bind to cell surface receptors or antigens. The antisense nucleic acid molecules can also be delivered to cells using the vectors described herein. To achieve sufficient intracellular concentrations of the antisense molecules, vector constructs in which the antisense nucleic acid molecule is placed under the control of a strong pol II or pol III
20 promoter are preferred.

In yet another embodiment, the antisense nucleic acid molecule of the invention is an α -anomeric nucleic acid molecule. An α -anomeric nucleic acid molecule forms specific double-stranded hybrids with complementary RNA in which, contrary to the usual β -units, the strands run parallel to each other (Gaultier *et al.*
25 (1987) *Nucleic Acids. Res.* 15:6625-6641). The antisense nucleic acid molecule can also comprise a 2'-o-methylribonucleotide (Inoue *et al.* (1987) *Nucleic Acids Res.* 15:6131-6148) or a chimeric RNA-DNA analogue (Inoue *et al.* (1987) *FEBS Lett.* 215:327-330).

In still another embodiment, an antisense nucleic acid of the invention is a
30 ribozyme. A ribozyme having specificity for a collagen XXII-encoding nucleic acid can include one or more sequences complementary to the nucleotide sequence of a collagen XXII cDNA disclosed herein (i.e., SEQ ID NO:1 or SEQ ID NO:3), and a sequence having known catalytic sequence responsible for mRNA cleavage (see U.S. Pat. No. 5,093,246 or Haselhoff and Gerlach (1988) *Nature* 334:585-591). For
35 example, a derivative of a *Tetrahymena* L-19 IVS RNA can be constructed in which the nucleotide sequence of the active site is complementary to the nucleotide sequence

5 to be cleaved in a collagen XXII-encoding mRNA. See, e.g., Cech *et al.* U.S. Patent No. 4,987,071; and Cech *et al.* U.S. Patent No. 5,116,742. Alternatively, collagen XXII mRNA can be used to select a catalytic RNA having a specific ribonuclease activity from a pool of RNA molecules. See, e.g., Bartel, D. and Szostak, J.W. (1993) *Science* 261:1411-1418.

10 Collagen XXII gene expression can be inhibited by targeting nucleotide sequences complementary to the regulatory region of the collagen XXII (e.g., the collagen XXII promoter and/or enhancers) to form triple helical structures that prevent transcription of the collagen XXII gene in target cells. See generally, Helene, C. (1991) *Anticancer Drug Des.* 6:569-84; Helene, C. i (1992) *Ann. N.Y. Acad. Sci.* 660:27-36; and Maher, L.J. (1992) *Bioassays* 14:807-15. The potential sequences that
15 can be targeted for triple helix formation can be increased by creating a so-called "switchback" nucleic acid molecule. Switchback molecules are synthesized in an alternating 5'-3', 3'-5' manner, such that they base pair with first one strand of a duplex and then the other, eliminating the necessity for a sizeable stretch of either purines or
20 pyrimidines to be present on one strand of a duplex.

The invention also provides detectably labeled oligonucleotide primer and probe molecules. Typically, such labels are chemiluminescent, fluorescent, radioactive, or colorimetric.

A collagen XXII nucleic acid molecule can be modified at the base moiety,
25 sugar moiety or phosphate backbone to improve, e.g., the stability, hybridization, or solubility of the molecule. For non-limiting examples of synthetic oligonucleotides with modifications see Toulmé (2001) *Nature Biotech.* 19:17 and Faria *et al.* (2001) *Nature Biotech.* 19:40-44. Such phosphoramidite oligonucleotides can be effective antisense agents.

30 For example, the deoxyribose phosphate backbone of the nucleic acid molecules can be modified to generate peptide nucleic acids (see Hyrup B. *et al.* (1996) *Bioorganic & Medicinal Chemistry* 4: 5-23). As used herein, the terms "peptide nucleic acid" or "PNA" refers to a nucleic acid mimic, e.g., a DNA mimic, in which the deoxyribose phosphate backbone is replaced by a pseudopeptide backbone
35 and only the four natural nucleobases are retained. The neutral backbone of a PNA can allow for specific hybridization to DNA and RNA under conditions of low ionic

5 strength. The synthesis of PNA oligomers can be performed using standard solid phase peptide synthesis protocols as described in Hyrup B. *et al.* (1996) *supra* and Perry-O'Keefe *et al.* *Proc. Natl. Acad. Sci.* 93: 14670-675.

PNAs of collagen XXII nucleic acid molecules can be used in therapeutic and diagnostic applications. For example, PNAs can be used as antisense or antigene
10 agents for sequence-specific modulation of gene expression by, for example, inducing transcription or translation arrest or inhibiting replication. PNAs of collagen XXII nucleic acid molecules can also be used in the analysis of single base pair mutations in a gene; (e.g., by PNA-directed PCR clamping); as 'artificial restriction enzymes' when used in combination with other enzymes, (e.g., S1 nucleases (Hyrup B. *et al.*
15 (1996) *supra*)); or as probes or primers for DNA sequencing or hybridization (Hyrup B. *et al.* (1996) *supra*; Perry-O'Keefe *supra*).

In other embodiments, the oligonucleotide may include other appended groups such as peptides (e.g., for targeting host cell receptors *in vivo*), or agents facilitating transport across the cell membrane (see, e.g., Letsinger *et al.* (1989) *Proc. Natl. Acad. Sci. USA* 86:6553-6556; Lemaitre *et al.* (1987) *Proc. Natl. Acad. Sci. USA* 84:648-
20 652; PCT Publication No. W088/09810) or the blood-brain barrier (see, e.g., PCT Publication No. W089/10134). In addition, oligonucleotides can be modified with hybridization-triggered cleavage agents (see, e.g., Krol *et al.* (1988) *Bio-Techniques* 6:958-976) or intercalating agents. (see, e.g., Zon (1988) *Pharm. Res.* 5:539-549).
25 To this end, the oligonucleotide may be conjugated to another molecule, (e.g., a peptide, hybridization triggered cross-linking agent, transport agent, or hybridization-triggered cleavage agent).

The invention also includes molecular beacon oligonucleotide primer and probe molecules having at least one region which is complementary to a collagen
30 XXII nucleic acid of the invention, two complementary regions one having a fluorophore and one a quencher such that the molecular beacon is useful for quantitating the presence of the collagen XXII nucleic acid of the invention in a sample. Molecular beacon nucleic acids are described, for example, in Lizardi *et al.*, U.S. Patent No. 5,854,033; Nazarenko *et al.*, U.S. Patent No. 5,866,336, and Livak *et al.*, U.S. Patent 5,876,930.
35

5 Isolated collagen XXII Polypeptides

 In another aspect, the invention features, an isolated collagen XXII protein, or fragment, e.g., a biologically active portion, for use as immunogens or antigens to raise or test (or more generally to bind) anti-collagen XXII antibodies. collagen XXII protein can be isolated from cells or tissue sources using standard protein purification
10 techniques. collagen XXII protein or fragments thereof can be produced by recombinant DNA techniques or synthesized chemically.

 Polypeptides of the invention include those which arise as a result of the existence of multiple genes, alternative transcription events, alternative RNA splicing events, and alternative translational and post-translational events. The polypeptide
15 can be expressed in systems, e.g., cultured cells, which result in substantially the same post-translational modifications present when expressed the polypeptide is expressed in a native cell, or in systems which result in the alteration or omission of post-translational modifications, e.g., glycosylation or cleavage, present when expressed in a native cell.

20 In a preferred embodiment, a collagen XXII polypeptide has one or more of the following characteristics:

- (i) it has the ability to form a helix;
- (ii) it has the ability to form a triple helix consisting of three collagen XXII polypeptide chains;
- 25 (iii) it has the ability to form higher order collagen XXII fibrils and/or fibers;
- (iv) it has the ability to form at least 1, preferably between 3 and 10 non-helical interruptions or domains within the core helical domain;
- (v) it provides tensile strength and/or flexibility to tissue boundaries, e.g., muscle-connective tissue boundaries;
- 30 (vi) it can provide form to a tissue, e.g., during growth and/or development,
- (vii) it can separate or anchor cell layers or tissue, e.g., at muscle-tissue junctions, dermal-epidermal junctions, basement membranes that separate epithelial and mesenchymal structures;
- (viii) it can provide a filtration barrier between tissues;
- 35 (ix) it can provide transparency, e.g., in the eye;

5 (x) it is recognized by an R34 antibody described herein;

(xi) it has a molecular weight, e.g., a deduced molecular weight, preferably ignoring any contribution of post translational modifications, amino acid composition or other physical characteristic of SEQ ID NO:2;

10 (xii) it has an overall sequence similarity of at least 50%, preferably at least 60%, more preferably at least 70, 80, 90, or 95%, with a polypeptide a of SEQ ID NO:2 or an alternatively spliced variant thereof;

(xiii) it can be found in muscle-connective tissue boundaries, in joints, in basement membranes;

15 (xiv) it has a core helical domain which is preferably about 70%, 80%, 90% or 95% with amino acid residues about 1479-1606 of SEQ ID NO:2;

In a preferred embodiment the collagen XXII protein, or fragment thereof, differs from the corresponding sequence in SEQ ID:2. In one embodiment it differs by at least one but by less than 15, 10 or 5 amino acid residues. In another it differs from the corresponding sequence in SEQ ID NO:2 by at least one residue but less
20 than 20%, 15%, 10% or 5% of the residues in it differ from the corresponding sequence in SEQ ID NO:2. (If this comparison requires alignment the sequences should be aligned for maximum homology. "Looped" out sequences from deletions or insertions, or mismatches, are considered differences.) The differences are, preferably, differences or changes at a non essential residue or a conservative substitution. In a
25 preferred embodiment the differences are not in the core helical domain. In another preferred embodiment one or more differences are in the core helical domain or in the non-helical interruption domains.

Other embodiments include a protein that contain one or more changes in amino acid sequence, e.g., a change in an amino acid residue which is not essential for
30 activity. Such collagen XXII proteins differ in amino acid sequence from SEQ ID NO:2, yet retain biological activity.

In one embodiment, the protein includes an amino acid sequence at least about 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 98% or more homologous to SEQ ID NO:2 or an alternatively spliced variant thereof described herein.

35 In one embodiment, a biologically active portion of a collagen XXII protein includes a core helical domain and at least one non-helical interruption. Moreover,

5 other biologically active portions, in which other regions of the protein are deleted, can be prepared by recombinant techniques and evaluated for one or more of the functional activities of a native collagen XXII protein.

In a preferred embodiment, the collagen XXII protein has an amino acid sequence shown in SEQ ID NO:2. In other embodiments, the collagen XXII protein
10 is substantially identical to SEQ ID NO:2. In yet another embodiment, the collagen XXII protein is substantially identical to SEQ ID NO:2 and retains the functional activity of the protein of SEQ ID NO:2, as described in detail in the subsections above.

Collagen XXII Chimeric or Fusion Proteins

15 In another aspect, the invention provides collagen XXII chimeric or fusion proteins. As used herein, a collagen XXII "chimeric protein" or "fusion protein" includes a collagen XXII polypeptide linked to a non-collagen XXII polypeptide. A "non-collagen XXII polypeptide" refers to a polypeptide having an amino acid sequence corresponding to a protein which is not substantially homologous to the
20 collagen XXII protein, e.g., a protein which is different from the collagen XXII protein and which is derived from the same or a different organism. The collagen XXII polypeptide of the fusion protein can correspond to all or a portion e.g., a fragment described herein of a collagen XXII amino acid sequence. In a preferred embodiment, a collagen XXII fusion protein includes at least one (or two)
25 biologically active portion of a collagen XXII protein. The non-collagen XXII polypeptide can be fused to the N-terminus or C-terminus of the collagen XXII polypeptide.

The fusion protein can include a moiety which has a high affinity for a ligand. For example, the fusion protein can be a GST-collagen XXII fusion protein in which
30 the collagen XXII sequences are fused to the C-terminus of the GST sequences. Such fusion proteins can facilitate the purification of recombinant collagen XXII. Alternatively, the fusion protein can be a collagen XXII protein containing a heterologous signal sequence at its N-terminus. In certain host cells (e.g., mammalian

5 host cells), expression and/or secretion of collagen XXII can be increased through use of a heterologous signal sequence.

Fusion proteins can include all or a part of a serum protein, e.g., an IgG constant region, or human serum albumin.

10 The collagen XXII fusion proteins of the invention can be incorporated into pharmaceutical compositions and administered to a subject *in vivo*. The collagen XXII fusion proteins can be used to affect the bioavailability of a collagen XXII substrate. collagen XXII fusion proteins may be useful therapeutically for the treatment of disorders caused by, for example, (i) aberrant modification or mutation of a gene encoding a collagen XXII protein; (ii) mis-regulation of the collagen XXII
15 gene; and (iii) aberrant post-translational modification of a collagen XXII protein.

Moreover, the collagen XXII-fusion proteins of the invention can be used as immunogens to produce anti-collagen XXII antibodies in a subject, to purify collagen XXII ligands and in screening assays to identify molecules which inhibit the interaction of collagen XXII with a collagen XXII substrate.

20 Expression vectors are commercially available that already encode a fusion moiety (e.g., a GST polypeptide). A collagen XXII-encoding nucleic acid can be cloned into such an expression vector such that the fusion moiety is linked in-frame to the collagen XXII protein.

Variants of collagen XXII Proteins

25 In another aspect, the invention also features a variant of a collagen XXII polypeptide, e.g., which functions as an agonist (mimetics) or as an antagonist. Variants of the collagen XXII proteins can be generated by mutagenesis, e.g., discrete point mutation, the insertion or deletion of sequences or the truncation of a collagen XXII protein. An agonist of the collagen XXII proteins can retain substantially the
30 same, or a subset, of the biological activities of the naturally occurring form of a collagen XXII protein. An antagonist of a collagen XXII protein can inhibit one or more of the activities of the naturally occurring form of the collagen XXII protein by, for example, competitively modulating a collagen XXII-mediated activity of a collagen XXII protein. Thus, specific biological effects can be elicited by treatment

5 with a variant of limited function. Preferably, treatment of a subject with a variant having a subset of the biological activities of the naturally occurring form of the protein has fewer side effects in a subject relative to treatment with the naturally occurring form of the collagen XXII protein.

Variants of a collagen XXII protein can be identified by screening
10 combinatorial libraries of mutants, e.g., truncation mutants, of a collagen XXII protein for agonist or antagonist activity.

Libraries of fragments e.g., N terminal, C terminal, or internal fragments, of a collagen XXII protein coding sequence can be used to generate a variegated population of fragments for screening and subsequent selection of variants of a
15 collagen XXII protein. Variants in which a cysteine residues is added or deleted or in which a residue which is glycosylated is added or deleted are particularly preferred.

Methods for screening gene products of combinatorial libraries made by point mutations or truncation, and for screening cDNA libraries for gene products having a selected property are known in the art. Such methods are adaptable for rapid
20 screening of the gene libraries generated by combinatorial mutagenesis of collagen XXII proteins. Recursive ensemble mutagenesis (REM), a new technique which enhances the frequency of functional mutants in the libraries, can be used in combination with the screening assays to identify collagen XXII variants (Arkin and Yourvan (1992) *Proc. Natl. Acad. Sci. USA* 89:7811-7815; Delgrave *et al.* (1993)
25 *Protein Engineering* 6:327-331).

Cell based assays can be exploited to analyze a variegated collagen XXII library. For example, a library of expression vectors can be transfected into a cell line, e.g., a cell line, which ordinarily responds to collagen XXII in a substrate-dependent manner. The transfected cells are then contacted with collagen XXII and
30 the effect of the expression of the mutant on signaling by the collagen XXII substrate can be detected, e.g., by measuring collagen triple helix formation or collagenase sensitivity. Plasmid DNA can then be recovered from the cells which score for inhibition, or alternatively, potentiation of signaling by the collagen XXII substrate, and the individual clones further characterized.

35 In another aspect, the invention features a method of making a collagen XXII polypeptide, e.g., a peptide having a non-wild type activity, e.g., an antagonist,

agonist, or super agonist of a naturally occurring collagen XXII polypeptide, e.g., a naturally occurring collagen XXII polypeptide. The method includes: altering the sequence of a collagen XXII polypeptide, e.g., altering the sequence, e.g., by substitution or deletion of one or more residues of a non-conserved region, a domain or residue disclosed herein, and testing the altered polypeptide for the desired activity.

In another aspect, the invention features a method of making a fragment or analog of a collagen XXII polypeptide a biological activity of a naturally occurring collagen XXII polypeptide. The method includes: altering the sequence, e.g., by substitution or deletion of one or more residues, of a collagen XXII polypeptide, e.g., altering the sequence of a non-conserved region, or a domain or residue described herein, and testing the altered polypeptide for the desired activity.

Anti-collagen XXII Antibodies

In another aspect, the invention provides an anti-collagen XXII antibody, or a fragment thereof (e.g., an antigen-binding fragment thereof). The term "antibody" as used herein refers to an immunoglobulin molecule or immunologically active portion thereof, i.e., an antigen-binding portion. As used herein, the term "antibody" refers to a protein comprising at least one, and preferably two, heavy (H) chain variable regions (abbreviated herein as VH), and at least one and preferably two light (L) chain variable regions (abbreviated herein as VL). The VH and VL regions can be further subdivided into regions of hypervariability, termed "complementarity determining regions" ("CDR"), interspersed with regions that are more conserved, termed "framework regions" (FR). The extent of the framework region and CDR's has been precisely defined (see, Kabat, E.A., *et al.* (1991) *Sequences of Proteins of Immunological Interest, Fifth Edition*, U.S. Department of Health and Human Services, NIH Publication No. 91-3242, and Chothia, C. *et al.* (1987) *J. Mol. Biol.* 6:901-917, which are incorporated herein by reference). Each VH and VL is composed of three CDR's and four FRs, arranged from amino-terminus to carboxy-terminus in the following order: FR1, CDR1, FR2, CDR2, FR3, CDR3, FR4. The anti-collagen XXII antibody can further include a heavy and light chain constant region, to thereby form a heavy and light immunoglobulin chain, respectively. In one

5 embodiment, the antibody is a tetramer of two heavy immunoglobulin chains and two light immunoglobulin chains, wherein the heavy and light immunoglobulin chains are inter-connected by, e.g., disulfide bonds. The heavy chain constant region is comprised of three domains, CH1, CH2 and CH3. The light chain constant region is comprised of one domain, CL. The variable region of the heavy and light chains
10 contains a binding domain that interacts with an antigen. The constant regions of the antibodies typically mediate the binding of the antibody to host tissues or factors, including various cells of the immune system (e.g., effector cells) and the first component (C1q) of the classical complement system.

As used herein, the term "immunoglobulin" refers to a protein consisting of
15 one or more polypeptides substantially encoded by immunoglobulin genes. The recognized human immunoglobulin genes include the kappa, lambda, alpha (IgA1 and IgA2), gamma (IgG1, IgG2, IgG3, IgG4), delta, epsilon and mu constant region genes, as well as the myriad immunoglobulin variable region genes. Full-length immunoglobulin "light chains" (about 25 KDa or 214 amino acids) are encoded by a
20 variable region gene at the NH₂-terminus (about 110 amino acids) and a kappa or lambda constant region gene at the COOH--terminus. Full-length immunoglobulin "heavy chains" (about 50 KDa or 446 amino acids), are similarly encoded by a variable region gene (about 116 amino acids) and one of the other aforementioned constant region genes, e.g., gamma (encoding
25 about 330 amino acids).

The term "antigen-binding fragment" of an antibody (or simply "antibody portion," or "fragment"), as used herein, refers to one or more fragments of a full-length antibody that retain the ability to specifically bind to the antigen, e.g., collagen XXII polypeptide or fragment thereof. Examples of antigen-binding fragments of the
30 anti-collagen XXII antibody include, but are not limited to: (i) a Fab fragment, a monovalent fragment consisting of the VL, VH, CL and CH1 domains; (ii) a F(ab')₂ fragment, a bivalent fragment comprising two Fab fragments linked by a disulfide bridge at the hinge region; (iii) a Fd fragment consisting of the VH and CH1 domains; (iv) a Fv fragment consisting of the VL and VH domains of a single arm of an
35 antibody, (v) a dAb fragment (Ward *et al.*, (1989) *Nature* 341:544-546), which consists of a VH domain; and (vi) an isolated complementarity determining region

5 (CDR). Furthermore, although the two domains of the Fv fragment, VL and VH, are coded for by separate genes, they can be joined, using recombinant methods, by a synthetic linker that enables them to be made as a single protein chain in which the VL and VH regions pair to form monovalent molecules (known as single chain Fv (scFv); see *e.g.*, Bird *et al.* (1988) *Science* 242:423-426; and Huston *et al.* (1988) *Proc. Natl. Acad. Sci. USA* 85:5879-5883). Such single chain antibodies are also
10 encompassed within the term "antigen-binding fragment" of an antibody. These antibody fragments are obtained using conventional techniques known to those with skill in the art, and the fragments are screened for utility in the same manner as are intact antibodies.

15 The anti-collagen XXII antibody can be a polyclonal or a monoclonal antibody. In other embodiments, the antibody can be recombinantly produced, *e.g.*, produced by phage display or by combinatorial methods.

Phage display and combinatorial methods for generating anti-collagen XXII antibodies are known in the art (as described in, *e.g.*, Ladner *et al.* U.S. Patent No. 5,223,409; Kang *et al.* International Publication No. WO 92/18619; Dower *et al.* International Publication No. WO 91/17271; Winter *et al.* International Publication WO 92/20791; Markland *et al.* International Publication No. WO 92/15679; Breitling *et al.* International Publication WO 93/01288; McCafferty *et al.* International Publication No. WO 92/01047; Garrard *et al.* International Publication No. WO
20 92/09690; Ladner *et al.* International Publication No. WO 90/02809; Fuchs *et al.* (1991) *Bio/Technology* 9:1370-1372; Hay *et al.* (1992) *Hum Antibod Hybridomas* 3:81-85; Huse *et al.* (1989) *Science* 246:1275-1281; Griffiths *et al.* (1993) *EMBO J* 12:725-734; Hawkins *et al.* (1992) *J Mol Biol* 226:889-896; Clackson *et al.* (1991) *Nature* 352:624-628; Gram *et al.* (1992) *PNAS* 89:3576-3580; Garrad *et al.* (1991) *Bio/Technology* 9:1373-1377; Hoogenboom *et al.* (1991) *Nuc Acid Res* 19:4133-4137;
25 and Barbas *et al.* (1991) *PNAS* 88:7978-7982, the contents of all of which are incorporated by reference herein).

In one embodiment, the anti-collagen XXII antibody is a fully human antibody (*e.g.*, an antibody made in a mouse which has been genetically engineered to produce
35 an antibody from a human immunoglobulin sequence), or a non-human antibody, *e.g.*, a rodent (mouse or rat), goat, primate (*e.g.*, monkey), camel antibody. Preferably, the

5 non-human antibody is a rodent (mouse or rat antibody). Method of producing rodent antibodies are known in the art.

Human monoclonal antibodies can be generated using transgenic mice carrying the human immunoglobulin genes rather than the mouse system.

Splenocytes from these transgenic mice immunized with the antigen of interest are
10 used to produce hybridomas that secrete human mAbs with specific affinities for epitopes from a human protein (see, e.g., Wood et al. International Application WO 91/00906, Kucherlapati et al. PCT publication WO 91/10741; Lonberg et al. International Application WO 92/03918; Kay et al. International Application 92/03917; Lonberg, N. et al. 1994 *Nature* 368:856-859; Green, L.L. et al. 1994
15 *Nature Genet.* 7:13-21; Morrison, S.L. et al. 1994 *Proc. Natl. Acad. Sci. USA* 81:6851-6855; Bruggeman et al. 1993 *Year Immunol* 7:33-40; Tuaillon et al. 1993 *PNAS* 90:3720-3724; Bruggeman et al. 1991 *Eur J Immunol* 21:1323-1326).

An anti-collagen XXII antibody can be one in which the variable region, or a portion thereof, e.g., the CDR's, are generated in a non-human organism, e.g., a rat or
20 mouse. Chimeric, CDR-grafted, and humanized antibodies are within the invention. Antibodies generated in a non-human organism, e.g., a rat or mouse, and then modified, e.g., in the variable framework or constant region, to decrease antigenicity in a human are within the invention.

Chimeric antibodies can be produced by recombinant DNA techniques known
25 in the art. For example, a gene encoding the Fc constant region of a murine (or other species) monoclonal antibody molecule is digested with restriction enzymes to remove the region encoding the murine Fc, and the equivalent portion of a gene encoding a human Fc constant region is substituted (see Robinson et al., International Patent Publication PCT/US86/02269; Akira, et al., European Patent Application
30 184,187; Taniguchi, M., European Patent Application 171,496; Morrison et al., European Patent Application 173,494; Neuberger et al., International Application WO 86/01533; Cabilly et al. U.S. Patent No. 4,816,567; Cabilly et al., European Patent Application 125,023; Better et al. (1988 *Science* 240:1041-1043); Liu et al. (1987) *PNAS* 84:3439-3443; Liu et al., 1987, *J. Immunol.* 139:3521-3526; Sun et al. (1987)
35 *PNAS* 84:214-218; Nishimura et al., 1987, *Canc. Res.* 47:999-1005; Wood et al.

- 5 (1985) *Nature* 314:446-449; and Shiao et al., 1988, *J. Natl Cancer Inst.* 80:1553-1559).

A humanized or CDR-grafted antibody will have at least one or two but generally all three recipient CDR's (of heavy and or light immunoglobulin chains) replaced with a donor CDR. The antibody may be replaced with at least a portion of a non-human CDR or only some of the CDR's may be replaced with non-human CDR's. It is only necessary to replace the number of CDR's required for binding of the humanized antibody to a collagen XXII or a fragment thereof. Preferably, the donor will be a rodent antibody, e.g., a rat or mouse antibody, and the recipient will be a human framework or a human consensus framework. Typically, the immunoglobulin providing the CDR's is called the "donor" and the immunoglobulin providing the framework is called the "acceptor." In one embodiment, the donor immunoglobulin is a non-human (e.g., rodent). The acceptor framework is a naturally-occurring (e.g., a human) framework or a consensus framework, or a sequence about 85% or higher, preferably 90%, 95%, 99% or higher identical thereto.

20 As used herein, the term "consensus sequence" refers to the sequence formed from the most frequently occurring amino acids (or nucleotides) in a family of related sequences (See e.g., Winnaker, *From Genes to Clones* (Verlagsgesellschaft, Weinheim, Germany 1987). In a family of proteins, each position in the consensus sequence is occupied by the amino acid occurring most frequently at that position in the family. If two amino acids occur equally frequently, either can be included in the consensus sequence. A "consensus framework" refers to the framework region in the consensus immunoglobulin sequence.

An antibody can be humanized by methods known in the art. Humanized antibodies can be generated by replacing sequences of the Fv variable region which are not directly involved in antigen binding with equivalent sequences from human Fv variable regions. General methods for generating humanized antibodies are provided by Morrison, S. L., 1985, *Science* 229:1202-1207, by Oi et al., 1986, *BioTechniques* 4:214, and by Queen et al. US 5,585,089, US 5,693,761 and US 5,693,762, the contents of all of which are hereby incorporated by reference. Those methods include isolating, manipulating, and expressing the nucleic acid sequences that encode all or part of immunoglobulin Fv variable regions from at least one of a heavy or light

5 chain. Sources of such nucleic acid are well known to those skilled in the art and, for example, may be obtained from a hybridoma producing an antibody against a collagen XXII polypeptide or fragment thereof. The recombinant DNA encoding the humanized antibody, or fragment thereof, can then be cloned into an appropriate expression vector.

10 Humanized or CDR-grafted antibodies can be produced by CDR-grafting or CDR substitution, wherein one, two, or all CDR's of an immunoglobulin chain can be replaced. See e.g., U.S. Patent 5,225,539; Jones et al. 1986 *Nature* 321:552-525; Verhoeyan et al. 1988 *Science* 239:1534; Beidler et al. 1988 *J. Immunol.* 141:4053-4060; Winter US 5,225,539, the contents of all of which are hereby expressly
15 incorporated by reference. Winter describes a CDR-grafting method which may be used to prepare the humanized antibodies of the present invention (UK Patent Application GB 2188638A, filed on March 26, 1987; Winter US 5,225,539), the contents of which is expressly incorporated by reference.

Also within the scope of the invention are humanized antibodies in which
20 specific amino acids have been substituted, deleted or added. Preferred humanized antibodies have amino acid substitutions in the framework region, such as to improve binding to the antigen. For example, a humanized antibody will have framework residues identical to the donor framework residue or to another amino acid other than the recipient framework residue. To generate such antibodies, a selected, small
25 number of acceptor framework residues of the humanized immunoglobulin chain can be replaced by the corresponding donor amino acids. Preferred locations of the substitutions include amino acid residues adjacent to the CDR, or which are capable of interacting with a CDR (see e.g., US 5,585,089). Criteria for selecting amino acids from the donor are described in US 5,585,089, e.g., columns 12-16 of US 5,585,089,
30 the e.g., columns 12-16 of US 5,585,089, the contents of which are hereby incorporated by reference. Other techniques for humanizing antibodies are described in Padlan et al. EP 519596 A1, published on December 23, 1992.

A full-length collagen XXII protein or, antigenic peptide fragment of collagen XXII can be used as an immunogen or can be used to identify anti-collagen XXII
35 antibodies made with other immunogens, e.g., cells, membrane preparations, and the like. The antigenic peptide of collagen XXII should include at least 8 amino acid

5 residues of the amino acid sequence shown in SEQ ID NO:2 and encompasses an epitope of collagen XXII. Preferably, the antigenic peptide includes at least 10 amino acid residues, more preferably at least 15 amino acid residues, even more preferably at least 20 amino acid residues, and most preferably at least 30 amino acid residues.

Antibodies reactive with, or specific for, any of these regions, or other regions
10 or domains described herein are provided.

Antibodies which bind only native collagen XXII protein, only denatured or otherwise non-native collagen XXII protein, or which bind both, are within the invention. Antibodies with linear or conformational epitopes are within the invention. Conformational epitopes can sometimes be identified by identifying antibodies which
15 bind to native but not denatured collagen XXII protein.

Preferred epitopes encompassed by the antigenic peptide are regions of collagen XXII are located on the surface of the protein, e.g., hydrophilic regions, as well as regions with high antigenicity. For example, an Emini surface probability analysis of the human collagen XXII protein sequence can be used to indicate the
20 regions that have a particularly high probability of being localized to the surface of the collagen XXII protein and are thus likely to constitute surface residues useful for targeting antibody production.

The anti-collagen XXII antibody can be a single chain antibody. A single-chain antibody (scFV) may be engineered (see, for example, Colcher, D. *et al.* (1999) *Ann N Y Acad Sci* 880:263-80; and Reiter, Y. (1996) *Clin Cancer Res* 2:245-52). The
25 single chain antibody can be dimerized or multimerized to generate multivalent antibodies having specificities for different epitopes of the same target collagen XXII protein.

In a preferred embodiment the antibody has effector function and/or can fix
30 complement. In other embodiments the antibody does not recruit effector cells; or fix complement.

In a preferred embodiment, the antibody has reduced or no ability to bind an Fc receptor. For example, it is a isotype or subtype, fragment or other mutant, which does not support binding to an Fc receptor, e.g., it has a mutagenized or deleted Fc
35 receptor binding region.

5 In a preferred embodiment, an anti-collagen XXII antibody alters (e.g., increases or decreases) a collagen XXII activity, e.g., triple helix forming activity or tensile strength providing activity, of a collagen XXII polypeptide.

 The antibody can be coupled to a toxin, e.g., a polypeptide toxin, e.g., ricin or diphtheria toxin or active fragment hereof, or a radioactive nucleus, or imaging agent,
10 e.g. a radioactive, enzymatic, or other, e.g., imaging agent, e.g., a NMR contrast agent. Labels which produce detectable radioactive emissions or fluorescence are preferred.

 An anti-collagen XXII antibody (e.g., monoclonal antibody) can be used to isolate collagen XXII by standard techniques, such as affinity chromatography or
15 immunoprecipitation. Moreover, an anti-collagen XXII antibody can be used to detect collagen XXII protein (e.g., in a cellular lysate or cell supernatant) in order to evaluate the abundance and pattern of expression of the protein. Anti-collagen XXII antibodies can be used diagnostically to monitor protein levels in tissue as part of a clinical testing procedure, e.g., to determine the efficacy of a given treatment regimen.
20 Detection can be facilitated by coupling (i.e., physically linking) the antibody to a detectable substance (i.e., antibody labelling). Examples of detectable substances include various enzymes, prosthetic groups, fluorescent materials, luminescent materials, bioluminescent materials, and radioactive materials. Examples of suitable enzymes include horseradish peroxidase, alkaline phosphatase, β -galactosidase, or
25 acetylcholinesterase; examples of suitable prosthetic group complexes include streptavidin/biotin and avidin/biotin; examples of suitable fluorescent materials include umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride or phycoerythrin; an example of a luminescent material includes luminol; examples of bioluminescent materials include
30 luciferase, luciferin, and aequorin, and examples of suitable radioactive material include ^{125}I , ^{131}I , ^{35}S or ^3H .

 The invention also includes a nucleic acid which encodes an anti-collagen XXII antibody, e.g., an anti-collagen XXII antibody described herein. Also included are vectors which include the nucleic acid and cells transformed with the nucleic acid,
35 particularly cells which are useful for producing an antibody, e.g., mammalian cells, e.g. CHO or lymphatic cells.

5 The invention also includes cell lines, e.g., hybridomas, which make an anti-collagen XXII antibody, e.g., and antibody described herein, and method of using said cells to make a collagen XXII antibody.

 In a preferred embodiment, the antibody is R34 antibody described herein.

Recombinant Expression Vectors, Host Cells and Genetically Engineered Cells

10 In another aspect, the invention includes, vectors, preferably expression vectors, containing a nucleic acid encoding a polypeptide described herein. As used herein, the term "vector" refers to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked and can include a plasmid, cosmid or viral vector. The vector can be capable of autonomous replication or it can integrate
15 into a host DNA. Viral vectors include, e.g., replication defective retroviruses, adenoviruses and adeno-associated viruses.

 A vector can include a collagen XXII nucleic acid in a form suitable for expression of the nucleic acid in a host cell. Preferably the recombinant expression vector includes one or more regulatory sequences operatively linked to the nucleic
20 acid sequence to be expressed. The term "regulatory sequence" includes promoters, enhancers and other expression control elements (e.g., polyadenylation signals). Regulatory sequences include those which direct constitutive expression of a nucleotide sequence, as well as tissue-specific regulatory and/or inducible sequences. The design of the expression vector can depend on such factors as the choice of the
25 host cell to be transformed, the level of expression of protein desired, and the like. The expression vectors of the invention can be introduced into host cells to thereby produce proteins or polypeptides, including fusion proteins or polypeptides, encoded by nucleic acids as described herein (e.g., collagen XXII proteins, mutant forms of collagen XXII proteins, fusion proteins, and the like).

30 The recombinant expression vectors of the invention can be designed for expression of collagen XXII proteins in prokaryotic or eukaryotic cells. For example, polypeptides of the invention can be expressed in *E. coli*, insect cells (e.g., using baculovirus expression vectors), yeast cells or mammalian cells. Suitable host cells are discussed further in Goeddel, (1990) *Gene Expression Technology: Methods in*

5 *Enzymology* 185, Academic Press, San Diego, CA. Alternatively, the recombinant expression vector can be transcribed and translated *in vitro*, for example using T7 promoter regulatory sequences and T7 polymerase.

Expression of proteins in prokaryotes is most often carried out in *E. coli* with vectors containing constitutive or inducible promoters directing the expression of
10 either fusion or non-fusion proteins. Fusion vectors add a number of amino acids to a protein encoded therein, usually to the amino terminus of the recombinant protein. Such fusion vectors typically serve three purposes: 1) to increase expression of recombinant protein; 2) to increase the solubility of the recombinant protein; and 3) to aid in the purification of the recombinant protein by acting as a ligand in affinity
15 purification. Often, a proteolytic cleavage site is introduced at the junction of the fusion moiety and the recombinant protein to enable separation of the recombinant protein from the fusion moiety subsequent to purification of the fusion protein. Such enzymes, and their cognate recognition sequences, include Factor Xa, thrombin and enterokinase. Typical fusion expression vectors include pGEX (Pharmacia Biotech
20 Inc; Smith, D.B. and Johnson, K.S. (1988) *Gene* 67:31-40), pMAL (New England Biolabs, Beverly, MA) and pRIT5 (Pharmacia, Piscataway, NJ) which fuse glutathione S-transferase (GST), maltose E binding protein, or protein A, respectively, to the target recombinant protein.

Purified fusion proteins can be used in collagen XXII activity assays, (e.g.,
25 direct assays or competitive assays described in detail below), or to generate antibodies specific for collagen XXII proteins. In a preferred embodiment, a fusion protein expressed in a retroviral expression vector of the present invention can be used to infect bone marrow cells which are subsequently transplanted into irradiated recipients. The pathology of the subject recipient is then examined after sufficient
30 time has passed (e.g., six weeks).

To maximize recombinant protein expression in *E. coli* is to express the protein in a host bacteria with an impaired capacity to proteolytically cleave the recombinant protein (Gottesman, S., (1990) *Gene Expression Technology: Methods in Enzymology* 185, Academic Press, San Diego, California 119-128). Another strategy
35 is to alter the nucleic acid sequence of the nucleic acid to be inserted into an expression vector so that the individual codons for each amino acid are those

5 preferentially utilized in *E. coli* (Wada *et al.*, (1992) *Nucleic Acids Res.* 20:2111-2118). Such alteration of nucleic acid sequences of the invention can be carried out by standard DNA synthesis techniques.

The collagen XXII expression vector can be a yeast expression vector, a vector for expression in insect cells, e.g., a baculovirus expression vector or a vector
10 suitable for expression in mammalian cells.

When used in mammalian cells, the expression vector's control functions can be provided by viral regulatory elements. For example, commonly used promoters are derived from polyoma, Adenovirus 2, cytomegalovirus and Simian Virus 40.

In another embodiment, the promoter is an inducible promoter, e.g., a
15 promoter regulated by a steroid hormone, by a polypeptide hormone (e.g., by means of a signal transduction pathway), or by a heterologous polypeptide (e.g., the tetracycline-inducible systems, "Tet-On" and "Tet-Off"; see, e.g., Clontech Inc., CA, Gossen and Bujard (1992) *Proc. Natl. Acad. Sci. USA* 89:5547, and Paillard (1989) *Human Gene Therapy* 9:983).

20 In another embodiment, the recombinant mammalian expression vector is capable of directing expression of the nucleic acid preferentially in a particular cell type (e.g., tissue-specific regulatory elements are used to express the nucleic acid). Non-limiting examples of suitable tissue-specific promoters include the albumin promoter (liver-specific; Pinkert *et al.* (1987) *Genes Dev.* 1:268-277), lymphoid-specific promoters (Calame and Eaton (1988) *Adv. Immunol.* 43:235-275), in
25 particular promoters of T cell receptors (Winoto and Baltimore (1989) *EMBO J.* 8:729-733) and immunoglobulins (Banerji *et al.* (1983) *Cell* 33:729-740; Queen and Baltimore (1983) *Cell* 33:741-748), neuron-specific promoters (e.g., the neurofilament promoter; Byrne and Ruddle (1989) *Proc. Natl. Acad. Sci. USA*
30 86:5473-5477), pancreas-specific promoters (Edlund *et al.* (1985) *Science* 230:912-916), and mammary gland-specific promoters (e.g., milk whey promoter; U.S. Patent No. 4,873,316 and European Application Publication No. 264,166).

Developmentally-regulated promoters are also encompassed, for example, the murine hox promoters (Kessel and Gruss (1990) *Science* 249:374-379) and the α -fetoprotein
35 promoter (Campes and Tilghman (1989) *Genes Dev.* 3:537-546).

5 The invention further provides a recombinant expression vector comprising a DNA molecule of the invention cloned into the expression vector in an antisense orientation. Regulatory sequences (e.g., viral promoters and/or enhancers) operatively linked to a nucleic acid cloned in the antisense orientation can be chosen which direct the constitutive, tissue specific or cell type specific expression of
10 antisense RNA in a variety of cell types. The antisense expression vector can be in the form of a recombinant plasmid, phagemid or attenuated virus.

 Another aspect the invention provides a host cell which includes a nucleic acid molecule described herein, e.g., a collagen XXII nucleic acid molecule within a recombinant expression vector or a collagen XXII nucleic acid molecule containing
15 sequences which allow it to homologously recombine into a specific site of the host cell's genome. The terms "host cell" and "recombinant host cell" are used interchangeably herein. Such terms refer not only to the particular subject cell but to the progeny or potential progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or environmental influences,
20 such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term as used herein.

 A host cell can be any prokaryotic or eukaryotic cell. For example, a collagen XXII protein can be expressed in bacterial cells (such as *E. coli*), insect cells, yeast or mammalian cells (such as Chinese hamster ovary cells (CHO) or COS cells (African
25 green monkey kidney cells CV-1 origin SV40 cells; Gluzman (1981) *Cell* 23:175-182)). Other suitable host cells are known to those skilled in the art.

 Vector DNA can be introduced into host cells via conventional transformation or transfection techniques. As used herein, the terms "transformation" and "transfection" are intended to refer to a variety of art-recognized techniques for
30 introducing foreign nucleic acid (e.g., DNA) into a host cell, including calcium phosphate or calcium chloride co-precipitation, DEAE-dextran-mediated transfection, lipofection, or electroporation.

 A host cell of the invention can be used to produce (i.e., express) a collagen XXII protein. Accordingly, the invention further provides methods for producing a
35 collagen XXII protein using the host cells of the invention. In one embodiment, the method includes culturing the host cell of the invention (into which a recombinant

5 expression vector encoding a collagen XXII protein has been introduced) in a suitable medium such that a collagen XXII protein is produced. In another embodiment, the method further includes isolating a collagen XXII protein from the medium or the host cell.

10 In another aspect, the invention features, a cell or purified preparation of cells which include a collagen XXII transgene, or which otherwise misexpress collagen XXII. The cell preparation can consist of human or non-human cells, e.g., rodent cells, e.g., mouse or rat cells, rabbit cells, or pig cells. In preferred embodiments, the cell or cells include a collagen XXII transgene, e.g., a heterologous form of a collagen XXII, e.g., a gene derived from humans (in the case of a non-human cell). The
15 collagen XXII transgene can be misexpressed, e.g., overexpressed or underexpressed. In other preferred embodiments, the cell or cells include a gene that mis-expresses an endogenous collagen XXII, e.g., a gene the expression of which is disrupted, e.g., a knockout. Such cells can serve as a model for studying disorders that are related to mutated or mis-expressed collagen XXII alleles or for use in drug screening.

20 In another aspect, the invention features, a human cell, a cartilage or connective tissue cell, transformed with nucleic acid which encodes a subject collagen XXII polypeptide.

Also provided are cells, preferably human cells, e.g., muscle or connective tissue or fibroblast cells, in which an endogenous collagen XXII is under the control
25 of a regulatory sequence that does not normally control the expression of the endogenous collagen XXII gene. The expression characteristics of an endogenous gene within a cell, e.g., a cell line or microorganism, can be modified by inserting a heterologous DNA regulatory element into the genome of the cell such that the inserted regulatory element is operably linked to the endogenous collagen XXII gene.
30 For example, an endogenous collagen XXII gene which is "transcriptionally silent," e.g., not normally expressed, or expressed only at very low levels, may be activated by inserting a regulatory element which is capable of promoting the expression of a normally expressed gene product in that cell. Techniques such as targeted homologous recombinations, can be used to insert the heterologous DNA as described
35 in, e.g., Chappel, US 5,272,071; WO 91/06667, published in May 16, 1991.

5 In a preferred embodiment, recombinant cells described herein can be used for replacement therapy in a subject. For example, a nucleic acid encoding a collagen XXII polypeptide operably linked to an inducible promoter (e.g., a steroid hormone receptor-regulated promoter) is introduced into a human or nonhuman, e.g., mammalian, e.g., porcine recombinant cell. The cell is cultivated and encapsulated in
10 a biocompatible material, such as poly-lysine alginate, and subsequently implanted into the subject. See, e.g., Lanza (1996) *Nat. Biotechnol.* 14:1107; Joki *et al.* (2001) *Nat. Biotechnol.* 19:35; and U.S. Patent No. 5,876,742. Production of collagen XXII polypeptide can be regulated in the subject by administering an agent (e.g., a steroid hormone) to the subject. In another preferred embodiment, the implanted
15 recombinant cells express and secrete an antibody specific for a collagen XXII polypeptide. The antibody can be any antibody or any antibody derivative described herein.

Transgenic Animals

 The invention provides non-human transgenic animals. Such animals are
20 useful for studying the function and/or activity of a collagen XXII protein and for identifying and/or evaluating modulators of collagen XXII activity. As used herein, a "transgenic animal" is a non-human animal, preferably a mammal, more preferably a rodent such as a rat or mouse, in which one or more of the cells of the animal includes a transgene. Other examples of transgenic animals include non-human primates,
25 sheep, dogs, cows, goats, chickens, amphibians, and the like. A transgene is exogenous DNA or a rearrangement, e.g., a deletion of endogenous chromosomal DNA, which preferably is integrated into or occurs in the genome of the cells of a transgenic animal. A transgene can direct the expression of an encoded gene product in one or more cell types or tissues of the transgenic animal, other transgenes, e.g., a
30 knockout, reduce expression. Thus, a transgenic animal can be one in which an endogenous collagen XXII gene has been altered by, e.g., by homologous recombination between the endogenous gene and an exogenous DNA molecule introduced into a cell of the animal, e.g., an embryonic cell of the animal, prior to development of the animal.

5 Intronic sequences and polyadenylation signals can also be included in the transgene to increase the efficiency of expression of the transgene. A tissue-specific regulatory sequence(s) can be operably linked to a transgene of the invention to direct expression of a collagen XXII protein to particular cells. A transgenic founder animal can be identified based upon the presence of a collagen XXII transgene in its genome
10 and/or expression of collagen XXII mRNA in tissues or cells of the animals. A transgenic founder animal can then be used to breed additional animals carrying the transgene. Moreover, transgenic animals carrying a transgene encoding a collagen XXII protein can further be bred to other transgenic animals carrying other transgenes.

15 collagen XXII proteins or polypeptides can be expressed in transgenic animals or plants, e.g., a nucleic acid encoding the protein or polypeptide can be introduced into the genome of an animal. In preferred embodiments the nucleic acid is placed under the control of a tissue specific promoter, e.g., a milk or egg specific promoter, and recovered from the milk or eggs produced by the animal. Suitable animals are
20 mice, pigs, cows, goats, and sheep.

The invention also includes a population of cells from a transgenic animal, as discussed, e.g., below.

Uses

The nucleic acid molecules, proteins, protein homologues, and antibodies
25 described herein can be used in one or more of the following methods: a) screening assays; b) predictive medicine (e.g., diagnostic assays, prognostic assays, monitoring clinical trials, and pharmacogenetics); c) methods of treatment (e.g., therapeutic and prophylactic); and d) biomaterials.

The isolated nucleic acid molecules of the invention can be used, for example,
30 to express a collagen XXII protein (e.g., via a recombinant expression vector in a host cell in gene therapy applications), to detect a collagen XXII mRNA (e.g., in a biological sample) or a genetic alteration in a collagen XXII gene, and to modulate collagen XXII activity, as described further below. The collagen XXII proteins can be used to treat disorders characterized by insufficient or excessive production of a

5 collagen XXII substrate or production of collagen XXII inhibitors. In addition, the collagen XXII proteins can be used to screen for naturally occurring collagen XXII substrates, to screen for drugs or compounds which modulate collagen XXII activity, as well as to treat disorders characterized by insufficient or excessive production of collagen XXII protein or production of collagen XXII protein forms which have
10 decreased, aberrant or unwanted activity compared to collagen XXII wild type protein. Moreover, the anti-collagen XXII antibodies of the invention can be used to detect and isolate collagen XXII proteins, regulate the bioavailability of collagen XXII proteins, and modulate collagen XXII activity.

A method of evaluating a compound for the ability to interact with, e.g., bind,
15 a subject collagen XXII polypeptide is provided. The method includes: contacting the compound with the subject collagen XXII polypeptide; and evaluating ability of the compound to interact with, e.g., to bind or form a complex with the subject collagen XXII polypeptide. This method can be performed in vitro, e.g., in a cell free system, or in vivo, e.g., in a two-hybrid interaction trap assay. This method can be
20 used to identify naturally occurring molecules that interact with subject collagen XXII polypeptide. It can also be used to find natural or synthetic inhibitors of subject collagen XXII polypeptide. Screening methods are discussed in more detail below.

Screening Assays

The invention provides methods (also referred to herein as "screening assays")
25 for identifying modulators, i.e., candidate or test compounds or agents (e.g., proteins, peptides, peptidomimetics, peptoids, small molecules or other drugs) which bind to collagen XXII proteins, have a stimulatory or inhibitory effect on, for example, collagen XXII expression or collagen XXII activity, or have a stimulatory or inhibitory effect on, for example, the expression or activity of a collagen XXII
30 substrate. Compounds thus identified can be used to modulate the activity of target gene products (e.g., collagen XXII genes) in a therapeutic protocol, to elaborate the biological function of the target gene product, or to identify compounds that disrupt normal target gene interactions.

5 In one embodiment, the invention provides assays for screening candidate or test compounds which are substrates of a collagen XXII protein or polypeptide or a biologically active portion thereof. In another embodiment, the invention provides assays for screening candidate or test compounds that bind to or modulate an activity of a collagen XXII protein or polypeptide or a biologically active portion thereof.

10 In one embodiment, an activity of a collagen XXII protein can be assayed as follows. Triple helix formation can be assayed by: incubating type XXII collagen polypeptides together under conditions that promote helix formation; separating the incubation mixture and a control mixture on a non-reducing PAGE gel; and comparing the incubation mixture to the control mixture. The mobility of collagen
15 XXII dimers and trimers (detected by protein staining or antibody binding) will be distinguishable from the mobility of uncomplexed collagen XXII polypeptide chains on a PAGE gel. Collagen expression can be assayed by sensitivity to collagenase by methods known in the art.

 The test compounds of the present invention can be obtained using any of the
20 numerous approaches in combinatorial library methods known in the art, including: biological libraries; peptoid libraries (libraries of molecules having the functionalities of peptides, but with a novel, non-peptide backbone which are resistant to enzymatic degradation but which nevertheless remain bioactive; see, e.g., Zuckermann, R.N. *et al.* (1994) *J. Med. Chem.* 37:2678-85); spatially addressable parallel solid phase or
25 solution phase libraries; synthetic library methods requiring deconvolution; the 'one-bead one-compound' library method; and synthetic library methods using affinity chromatography selection. The biological library and peptoid library approaches are limited to peptide libraries, while the other four approaches are applicable to peptide, non-peptide oligomer or small molecule libraries of compounds (Lam (1997)
30 *Anticancer Drug Des.* 12:145).

 Examples of methods for the synthesis of molecular libraries can be found in the art, for example in: DeWitt *et al.* (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90:6909; Erb *et al.* (1994) *Proc. Natl. Acad. Sci. USA* 91:11422; Zuckermann *et al.* (1994). *J. Med. Chem.* 37:2678; Cho *et al.* (1993) *Science* 261:1303; Carrell *et al.* (1994)
35 *Angew. Chem. Int. Ed. Engl.* 33:2059; Carrell *et al.* (1994) *Angew. Chem. Int. Ed. Engl.* 33:2061; and Gallop *et al.* (1994) *J. Med. Chem.* 37:1233.

5 Libraries of compounds may be presented in solution (e.g., Houghten (1992) *Biotechniques* 13:412-421), or on beads (Lam (1991) *Nature* 354:82-84), chips (Fodor (1993) *Nature* 364:555-556), bacteria (Ladner, U.S. Patent No. 5,223,409), spores (Ladner U.S. Patent No. 5,223,409), plasmids (Cull *et al.* (1992) *Proc Natl Acad Sci USA* 89:1865-1869) or on phage (Scott and Smith (1990) *Science* 249:386-390; Devlin (1990) *Science* 249:404-406; Cwirla *et al.* (1990) *Proc. Natl. Acad. Sci.* 87:6378-6382; Felici (1991) *J. Mol. Biol.* 222:301-310; Ladner *supra.*).

In one embodiment, an assay is a cell-based assay in which a cell which expresses a collagen XXII protein or biologically active portion thereof is contacted with a test compound, and the ability of the test compound to modulate collagen XXII activity is determined. Determining the ability of the test compound to modulate collagen XXII activity can be accomplished by monitoring, for example, triple helix formation. The cell, for example, can be of mammalian origin, e.g., human.

The ability of the test compound to modulate collagen XXII binding to a compound, e.g., a collagen XXII substrate, or to bind to collagen XXII can also be evaluated. This can be accomplished, for example, by coupling the compound, e.g., the substrate, with a radioisotope or enzymatic label such that binding of the compound, e.g., the substrate, to collagen XXII can be determined by detecting the labeled compound, e.g., substrate, in a complex. Alternatively, collagen XXII could be coupled with a radioisotope or enzymatic label to monitor the ability of a test compound to modulate collagen XXII binding to a collagen XXII substrate in a complex. For example, compounds (e.g., collagen XXII substrates) can be labeled with ^{125}I , ^{35}S , ^{14}C , or ^3H , either directly or indirectly, and the radioisotope detected by direct counting of radioemission or by scintillation counting. Alternatively, compounds can be enzymatically labeled with, for example, horseradish peroxidase, alkaline phosphatase, or luciferase, and the enzymatic label detected by determination of conversion of an appropriate substrate to product.

The ability of a compound (e.g., a collagen XXII substrate) to interact with collagen XXII with or without the labeling of any of the interactants can be evaluated. For example, a microphysiometer can be used to detect the interaction of a compound with collagen XXII without the labeling of either the compound or the collagen XXII. McConnell, H. M. *et al.* (1992) *Science* 257:1906-1912. As used herein, a

5 "microphysiometer" (e.g., Cytosensor) is an analytical instrument that measures the rate at which a cell acidifies its environment using a light-addressable potentiometric sensor (LAPS). Changes in this acidification rate can be used as an indicator of the interaction between a compound and collagen XXII.

10 In yet another embodiment, a cell-free assay is provided in which a collagen XXII protein or biologically active portion thereof is contacted with a test compound and the ability of the test compound to bind to the collagen XXII protein or biologically active portion thereof is evaluated. Preferred biologically active portions of the collagen XXII proteins to be used in assays of the present invention include fragments which participate in interactions with non-collagen XXII molecules, e.g.,
15 fragments with high surface probability scores.

Cell-free assays involve preparing a reaction mixture of the target gene protein and the test compound under conditions and for a time sufficient to allow the two components to interact and bind, thus forming a complex that can be removed and/or detected.

20 The interaction between two molecules can also be detected, e.g., using fluorescence energy transfer (FET) (see, for example, Lakowicz *et al.*, U.S. Patent No. 5,631,169; Stavrianopoulos, *et al.*, U.S. Patent No. 4,868,103). A fluorophore label on the first, 'donor' molecule is selected such that its emitted fluorescent energy will be absorbed by a fluorescent label on a second, 'acceptor' molecule, which in
25 turn is able to fluoresce due to the absorbed energy. Alternately, the 'donor' protein molecule may simply utilize the natural fluorescent energy of tryptophan residues. Labels are chosen that emit different wavelengths of light, such that the 'acceptor' molecule label may be differentiated from that of the 'donor'. Since the efficiency of energy transfer between the labels is related to the distance separating the molecules,
30 the spatial relationship between the molecules can be assessed. In a situation in which binding occurs between the molecules, the fluorescent emission of the 'acceptor' molecule label in the assay should be maximal. An FET binding event can be conveniently measured through standard fluorometric detection means well known in the art (e.g., using a fluorimeter).

35 In another embodiment, determining the ability of the collagen XXII protein to bind to a target molecule can be accomplished using real-time Biomolecular

5 Interaction Analysis (BIA) (see, e.g., Sjolander, S. and Urbaniczky, C. (1991) *Anal. Chem.* 63:2338-2345 and Szabo *et al.* (1995) *Curr. Opin. Struct. Biol.* 5:699-705). "Surface plasmon resonance" or "BIA" detects biospecific interactions in real time, without labeling any of the interactants (e.g., BIAcore). Changes in the mass at the binding surface (indicative of a binding event) result in alterations of the refractive
10 index of light near the surface (the optical phenomenon of surface plasmon resonance (SPR)), resulting in a detectable signal which can be used as an indication of real-time reactions between biological molecules.

In one embodiment, the target gene product or the test substance is anchored onto a solid phase. The target gene product/test compound complexes anchored on
15 the solid phase can be detected at the end of the reaction. Preferably, the target gene product can be anchored onto a solid surface, and the test compound, (which is not anchored), can be labeled, either directly or indirectly, with detectable labels discussed herein.

It may be desirable to immobilize either collagen XXII, an anti-collagen XXII
20 antibody or its target molecule to facilitate separation of complexed from uncomplexed forms of one or both of the proteins, as well as to accommodate automation of the assay. Binding of a test compound to a collagen XXII protein, or interaction of a collagen XXII protein with a target molecule in the presence and absence of a candidate compound, can be accomplished in any vessel suitable for
25 containing the reactants. Examples of such vessels include microtiter plates, test tubes, and micro-centrifuge tubes. In one embodiment, a fusion protein can be provided which adds a domain that allows one or both of the proteins to be bound to a matrix. For example, glutathione-S-transferase/collagen XXII fusion proteins or glutathione-S-transferase/target fusion proteins can be adsorbed onto glutathione
30 sepharose beads (Sigma Chemical, St. Louis, MO) or glutathione derivatized microtiter plates, which are then combined with the test compound or the test compound and either the non-adsorbed target protein or collagen XXII protein, and the mixture incubated under conditions conducive to complex formation (e.g., at physiological conditions for salt and pH). Following incubation, the beads or
35 microtiter plate wells are washed to remove any unbound components, the matrix immobilized in the case of beads, complex determined either directly or indirectly, for

5 example, as described above. Alternatively, the complexes can be dissociated from the matrix, and the level of collagen XXII binding or activity determined using standard techniques.

Other techniques for immobilizing either a collagen XXII protein or a target molecule on matrices include using conjugation of biotin and streptavidin.

10 Biotinylated collagen XXII protein or target molecules can be prepared from biotin-NHS (N-hydroxy-succinimide) using techniques known in the art (e.g., biotinylation kit, Pierce Chemicals, Rockford, IL), and immobilized in the wells of streptavidin-coated 96 well plates (Pierce Chemical).

In order to conduct the assay, the non-immobilized component is added to the
15 coated surface containing the anchored component. After the reaction is complete, unreacted components are removed (e.g., by washing) under conditions such that any complexes formed will remain immobilized on the solid surface. The detection of complexes anchored on the solid surface can be accomplished in a number of ways. Where the previously non-immobilized component is pre-labeled, the detection of
20 label immobilized on the surface indicates that complexes were formed. Where the previously non-immobilized component is not pre-labeled, an indirect label can be used to detect complexes anchored on the surface; e.g., using a labeled antibody specific for the immobilized component (the antibody, in turn, can be directly labeled or indirectly labeled with, e.g., a labeled anti-Ig antibody).

25 In one embodiment, this assay is performed utilizing antibodies reactive with collagen XXII protein or target molecules but which do not interfere with binding of the collagen XXII protein to its target molecule. Such antibodies can be derivatized to the wells of the plate, and unbound target or collagen XXII protein trapped in the wells by antibody conjugation. Methods for detecting such complexes, in addition to
30 those described above for the GST-immobilized complexes, include immunodetection of complexes using antibodies reactive with the collagen XXII protein or target molecule, as well as enzyme-linked assays which rely on detecting an enzymatic activity associated with the collagen XXII protein or target molecule.

Alternatively, cell free assays can be conducted in a liquid phase. In such an
35 assay, the reaction products are separated from unreacted components, by any of a number of standard techniques, including but not limited to: differential centrifugation

5 (see, for example, Rivas, G., and Minton, A.P., (1993) *Trends Biochem Sci* 18:284-7); chromatography (gel filtration chromatography, ion-exchange chromatography); electrophoresis (see, e.g., Ausubel, F. *et al.*, eds. *Current Protocols in Molecular Biology* 1999, J. Wiley: New York.); and immunoprecipitation (see, for example, Ausubel, F. *et al.*, eds. (1999) *Current Protocols in Molecular Biology*, J. Wiley: New
10 York). Such resins and chromatographic techniques are known to one skilled in the art (see, e.g., Heegaard, N.H., (1998) *J Mol Recognit* 11:141-8; Hage, D.S., and Tweed, S.A. (1997) *J Chromatogr B Biomed Sci Appl.* 699:499-525). Further, fluorescence energy transfer may also be conveniently utilized, as described herein, to detect binding without further purification of the complex from solution.

15 In a preferred embodiment, the assay includes contacting the collagen XXII protein or biologically active portion thereof with a known compound which binds collagen XXII to form an assay mixture, contacting the assay mixture with a test compound, and determining the ability of the test compound to interact with a collagen XXII protein, wherein determining the ability of the test compound to
20 interact with a collagen XXII protein includes determining the ability of the test compound to preferentially bind to collagen XXII or biologically active portion thereof, or to modulate the activity of a target molecule, as compared to the known compound.

The target gene products of the invention can, *in vivo*, interact with one or
25 more cellular or extracellular macromolecules, such as proteins. For the purposes of this discussion, such cellular and extracellular macromolecules are referred to herein as "binding partners." Compounds that disrupt such interactions can be useful in regulating the activity of the target gene product. Such compounds can include, but are not limited to molecules such as antibodies, peptides, and small molecules. The
30 preferred target genes/products for use in this embodiment are the collagen XXII genes herein identified. In an alternative embodiment, the invention provides methods for determining the ability of the test compound to modulate the activity of a collagen XXII protein through modulation of the activity of a downstream effector of a collagen XXII target molecule. For example, the activity of the effector molecule
35 on an appropriate target can be determined, or the binding of the effector to an appropriate target can be determined, as previously described.

5 To identify compounds that interfere with the interaction between the target gene product and its cellular or extracellular binding partner(s), a reaction mixture containing the target gene product and the binding partner is prepared, under conditions and for a time sufficient, to allow the two products to form complex. In order to test an inhibitory agent, the reaction mixture is provided in the presence and
10 absence of the test compound. The test compound can be initially included in the reaction mixture, or can be added at a time subsequent to the addition of the target gene and its cellular or extracellular binding partner. Control reaction mixtures are incubated without the test compound or with a placebo. The formation of any complexes between the target gene product and the cellular or extracellular binding
15 partner is then detected. The formation of a complex in the control reaction, but not in the reaction mixture containing the test compound, indicates that the compound interferes with the interaction of the target gene product and the interactive binding partner. Additionally, complex formation within reaction mixtures containing the test compound and normal target gene product can also be compared to complex
20 formation within reaction mixtures containing the test compound and mutant target gene product. This comparison can be important in those cases wherein it is desirable to identify compounds that disrupt interactions of mutant but not normal target gene products.

 These assays can be conducted in a heterogeneous or homogeneous format.
25 Heterogeneous assays involve anchoring either the target gene product or the binding partner onto a solid phase, and detecting complexes anchored on the solid phase at the end of the reaction. In homogeneous assays, the entire reaction is carried out in a liquid phase. In either approach, the order of addition of reactants can be varied to obtain different information about the compounds being tested. For example, test
30 compounds that interfere with the interaction between the target gene products and the binding partners, e.g., by competition, can be identified by conducting the reaction in the presence of the test substance. Alternatively, test compounds that disrupt preformed complexes, e.g., compounds with higher binding constants that displace one of the components from the complex, can be tested by adding the test compound
35 to the reaction mixture after complexes have been formed. The various formats are briefly described below.

5 In a heterogeneous assay system, either the target gene product or the interactive cellular or extracellular binding partner, is anchored onto a solid surface (e.g., a microtiter plate), while the non-anchored species is labeled, either directly or indirectly. The anchored species can be immobilized by non-covalent or covalent attachments. Alternatively, an immobilized antibody specific for the species to be
10 anchored can be used to anchor the species to the solid surface.

 In order to conduct the assay, the partner of the immobilized species is exposed to the coated surface with or without the test compound. After the reaction is complete, unreacted components are removed (e.g., by washing) and any complexes formed will remain immobilized on the solid surface. Where the non-immobilized
15 species is pre-labeled, the detection of label immobilized on the surface indicates that complexes were formed. Where the non-immobilized species is not pre-labeled, an indirect label can be used to detect complexes anchored on the surface; e.g., using a labeled antibody specific for the initially non-immobilized species (the antibody, in turn, can be directly labeled or indirectly labeled with, e.g., a labeled anti-Ig
20 antibody). Depending upon the order of addition of reaction components, test compounds that inhibit complex formation or that disrupt preformed complexes can be detected.

 Alternatively, the reaction can be conducted in a liquid phase in the presence or absence of the test compound, the reaction products separated from unreacted
25 components, and complexes detected; e.g., using an immobilized antibody specific for one of the binding components to anchor any complexes formed in solution, and a labeled antibody specific for the other partner to detect anchored complexes. Again, depending upon the order of addition of reactants to the liquid phase, test compounds that inhibit complex or that disrupt preformed complexes can be identified.

30 In an alternate embodiment of the invention, a homogeneous assay can be used. For example, a preformed complex of the target gene product and the interactive cellular or extracellular binding partner product is prepared in that either the target gene products or their binding partners are labeled, but the signal generated by the label is quenched due to complex formation (see, e.g., U.S. Patent
35 No. 4,109,496 that utilizes this approach for immunoassays). The addition of a test substance that competes with and displaces one of the species from the preformed

5 complex will result in the generation of a signal above background. In this way, test substances that disrupt target gene product-binding partner interaction can be identified.

In yet another aspect, the collagen XXII proteins can be used as "bait proteins" in a two-hybrid assay or three-hybrid assay (see, e.g., U.S. Patent No. 5,283,317; 10 Zervos *et al.* (1993) *Cell* 72:223-232; Madura *et al.* (1993) *J. Biol. Chem.* 268:12046-12054; Bartel *et al.* (1993) *Biotechniques* 14:920-924; Iwabuchi *et al.* (1993) *Oncogene* 8:1693-1696; and Brent WO94/10300), to identify other proteins, which bind to or interact with collagen XXII ("collagen XXII-binding proteins" or "collagen XXII-bp") and are involved in collagen XXII activity. Such collagen XXII-bps can 15 be activators or inhibitors of signals by the collagen XXII proteins or collagen XXII targets as, for example, downstream elements of a collagen XXII-mediated signaling pathway.

The two-hybrid system is based on the modular nature of most transcription factors, which consist of separable DNA-binding and activation domains. Briefly, the 20 assay utilizes two different DNA constructs. In one construct, the gene that codes for a collagen XXII protein is fused to a gene encoding the DNA binding domain of a known transcription factor (e.g., GAL-4). In the other construct, a DNA sequence, from a library of DNA sequences, that encodes an unidentified protein ("prey" or "sample") is fused to a gene that codes for the activation domain of the known 25 transcription factor. (Alternatively the: collagen XXII protein can be the fused to the activator domain.) If the "bait" and the "prey" proteins are able to interact, *in vivo*, forming a collagen XXII-dependent complex, the DNA-binding and activation domains of the transcription factor are brought into close proximity. This proximity allows transcription of a reporter gene (e.g., lacZ) which is operably linked to a 30 transcriptional regulatory site responsive to the transcription factor. Expression of the reporter gene can be detected and cell colonies containing the functional transcription factor can be isolated and used to obtain the cloned gene which encodes the protein which interacts with the collagen XXII protein.

In another embodiment, modulators of collagen XXII expression are 35 identified. For example, a cell or cell free mixture is contacted with a candidate compound and the expression of collagen XXII mRNA or protein evaluated relative

5 to the level of expression of collagen XXII mRNA or protein in the absence of the candidate compound. When expression of collagen XXII mRNA or protein is greater in the presence of the candidate compound than in its absence, the candidate compound is identified as a stimulator of collagen XXII mRNA or protein expression. Alternatively, when expression of collagen XXII mRNA or protein is less
10 (statistically significantly less) in the presence of the candidate compound than in its absence, the candidate compound is identified as an inhibitor of collagen XXII mRNA or protein expression. The level of collagen XXII mRNA or protein expression can be determined by methods described herein for detecting collagen XXII mRNA or protein.

15 In another aspect, the invention pertains to a combination of two or more of the assays described herein. For example, a modulating agent can be identified using a cell-based or a cell free assay, and the ability of the agent to modulate the activity of a collagen XXII protein can be confirmed *in vivo*, e.g., in an animal such as an animal model for a connective tissue or joint disorder.

20 This invention further pertains to novel agents identified by the above-described screening assays. Accordingly, it is within the scope of this invention to further use an agent identified as described herein (e.g., a collagen XXII modulating agent, an antisense collagen XXII nucleic acid molecule, a collagen XXII-specific antibody, or a collagen XXII-binding partner) in an appropriate animal model to
25 determine the efficacy, toxicity, side effects, or mechanism of action, of treatment with such an agent. Furthermore, novel agents identified by the above-described screening assays can be used for treatments as described herein.

Detection Assays

30 Portions or fragments of the nucleic acid sequences identified herein can be used as polynucleotide reagents. For example, these sequences can be used to: (i) map their respective genes on a chromosome e.g., to locate gene regions associated with genetic disease or to associate collagen XXII with a disease; (ii) identify an individual from a minute biological sample (tissue typing); and (iii) aid in forensic

- 5 identification of a biological sample. These applications are described in the subsections below.

Chromosome Mapping

The collagen XXII nucleotide sequences or portions thereof can be used to map the location of the collagen XXII genes on a chromosome. This process is called
10 chromosome mapping. Chromosome mapping is useful in correlating the collagen XXII sequences with genes associated with disease.

Briefly, collagen XXII genes can be mapped to chromosomes by preparing PCR primers (preferably 15-25 bp in length) from the collagen XXII nucleotide sequences. These primers can then be used for PCR screening of somatic cell hybrids
15 containing individual human chromosomes. Only those hybrids containing the human gene corresponding to the collagen XXII sequences will yield an amplified fragment.

A panel of somatic cell hybrids in which each cell line contains either a single human chromosome or a small number of human chromosomes, and a full set of mouse chromosomes, can allow easy mapping of individual genes to specific human
20 chromosomes. (D'Eustachio P. *et al.* (1983) *Science* 220:919-924).

Other mapping strategies e.g., in situ hybridization (described in Fan, Y. *et al.* (1990) *Proc. Natl. Acad. Sci. USA*, 87:6223-27), pre-screening with labeled flow-sorted chromosomes, and pre-selection by hybridization to chromosome specific cDNA libraries can be used to map collagen XXII to a chromosomal location.

25 Fluorescence *in situ* hybridization (FISH) of a DNA sequence to a metaphase chromosomal spread can further be used to provide a precise chromosomal location in one step. The FISH technique can be used with a DNA sequence as short as 500 or 600 bases. However, clones larger than 1,000 bases have a higher likelihood of binding to a unique chromosomal location with sufficient signal intensity for simple
30 detection. Preferably 1,000 bases, and more preferably 2,000 bases will suffice to get good results at a reasonable amount of time. For a review of this technique, see Verma *et al.*, Human Chromosomes: A Manual of Basic Techniques ((1988) Pergamon Press, New York).

5 Reagents for chromosome mapping can be used individually to mark a single chromosome or a single site on that chromosome, or panels of reagents can be used for marking multiple sites and/or multiple chromosomes. Reagents corresponding to noncoding regions of the genes actually are preferred for mapping purposes. Coding sequences are more likely to be conserved within gene families, thus increasing the
10 chance of cross hybridizations during chromosomal mapping.

 Once a sequence has been mapped to a precise chromosomal location, the physical position of the sequence on the chromosome can be correlated with genetic map data. (Such data are found, for example, in V. McKusick, Mendelian Inheritance in Man, available on-line through Johns Hopkins University Welch Medical Library).
15 The relationship between a gene and a disease, mapped to the same chromosomal region, can then be identified through linkage analysis (co-inheritance of physically adjacent genes), described in, for example, Egeland, J. *et al.* (1987) *Nature*, 325:783-787.

 Moreover, differences in the DNA sequences between individuals affected and
20 unaffected with a disease associated with the collagen XXII gene, can be determined. If a mutation is observed in some or all of the affected individuals but not in any unaffected individuals, then the mutation is likely to be the causative agent of the particular disease. Comparison of affected and unaffected individuals generally involves first looking for structural alterations in the chromosomes, such as deletions
25 or translocations that are visible from chromosome spreads or detectable using PCR based on that DNA sequence. Ultimately, complete sequencing of genes from several individuals can be performed to confirm the presence of a mutation and to distinguish mutations from polymorphisms.

Tissue Typing

30 collagen XXII sequences can be used to identify individuals from biological samples using, e.g., restriction fragment length polymorphism (RFLP). In this technique, an individual's genomic DNA is digested with one or more restriction enzymes, the fragments separated, e.g., in a Southern blot, and probed to yield bands

5 for identification. The sequences of the present invention are useful as additional DNA markers for RFLP (described in U.S. Patent 5,272,057).

Furthermore, the sequences of the present invention can also be used to determine the actual base-by-base DNA sequence of selected portions of an individual's genome. Thus, the collagen XXII nucleotide sequences described herein
10 can be used to prepare two PCR primers from the 5' and 3' ends of the sequences. These primers can then be used to amplify an individual's DNA and subsequently sequence it. Panels of corresponding DNA sequences from individuals, prepared in this manner, can provide unique individual identifications, as each individual will have a unique set of such DNA sequences due to allelic differences.

15 Allelic variation occurs to some degree in the coding regions of these sequences, and to a greater degree in the noncoding regions. Each of the sequences described herein can, to some degree, be used as a standard against which DNA from an individual can be compared for identification purposes. Because greater numbers of polymorphisms occur in the noncoding regions, fewer sequences are necessary to
20 differentiate individuals. The noncoding sequences of SEQ ID NO:1 can provide positive individual identification with a panel of perhaps 10 to 1,000 primers which each yield a noncoding amplified sequence of 100 bases. If predicted coding sequences, such as those in SEQ ID NO:3 are used, a more appropriate number of primers for positive individual identification would be 500-2,000.

25 If a panel of reagents from collagen XXII nucleotide sequences described herein is used to generate a unique identification database for an individual, those same reagents can later be used to identify tissue from that individual. Using the unique identification database, positive identification of the individual, living or dead, can be made from extremely small tissue samples.

30 Use of Partial collagen XXII Sequences in Forensic Biology

DNA-based identification techniques can also be used in forensic biology. To make such an identification, PCR technology can be used to amplify DNA sequences taken from very small biological samples such as tissues, e.g., hair or skin, or body fluids, e.g., blood, saliva, or semen found at a crime scene. The amplified sequence

5 can then be compared to a standard, thereby allowing identification of the origin of the biological sample.

The sequences of the present invention can be used to provide polynucleotide reagents, e.g., PCR primers, targeted to specific loci in the human genome, which can enhance the reliability of DNA-based forensic identifications by, for example,
10 providing another "identification marker" (i.e. another DNA sequence that is unique to a particular individual). As mentioned above, actual base sequence information can be used for identification as an accurate alternative to patterns formed by restriction enzyme generated fragments. Sequences targeted to noncoding regions of SEQ ID NO:1 (e.g., fragments derived from the noncoding regions of SEQ ID NO:1 having a
15 length of at least 20 bases, preferably at least 30 bases) are particularly appropriate for this use.

The collagen XXII nucleotide sequences described herein can further be used to provide polynucleotide reagents, e.g., labeled or labelable probes which can be used in, for example, an *in situ* hybridization technique, to identify a specific tissue.
20 This can be very useful in cases where a forensic pathologist is presented with a tissue of unknown origin. Panels of such collagen XXII probes can be used to identify tissue by species and/or by organ type.

In a similar fashion, these reagents, e.g., collagen XXII primers or probes can be used to screen tissue culture for contamination (i.e. screen for the presence of a
25 mixture of different types of cells in a culture).

Predictive Medicine

The present invention also pertains to the field of predictive medicine in which diagnostic assays, prognostic assays, and monitoring clinical trials are used for prognostic (predictive) purposes to thereby treat an individual.

30 Generally, the invention provides, a method of determining if a subject is at risk for a disorder related to a lesion in or the misexpression of a gene which encodes collagen XXII.

Such disorders include, e.g., a disorder associated with the misexpression of collagen XXII gene; a disorder of the connective tissue system.

5 The method includes one or more of the following:

detecting, in a tissue of the subject, the presence or absence of a mutation which affects the expression of the collagen XXII gene, or detecting the presence or absence of a mutation in a region which controls the expression of the gene, e.g., a mutation in the 5' control region;

10 detecting, in a tissue of the subject, the presence or absence of a mutation which alters the structure of the collagen XXII gene;

detecting, in a tissue of the subject, the misexpression of the collagen XXII gene, at the mRNA level, e.g., detecting a non-wild type level of a mRNA ;

15 detecting, in a tissue of the subject, the misexpression of the gene, at the protein level, e.g., detecting a non-wild type level of a collagen XXII polypeptide.

In preferred embodiments the method includes: ascertaining the existence of at least one of: a deletion of one or more nucleotides from the collagen XXII gene; an insertion of one or more nucleotides into the gene, a point mutation, e.g., a substitution of one or more nucleotides of the gene, a gross chromosomal
20 rearrangement of the gene, e.g., a translocation, inversion, or deletion.

For example, detecting the genetic lesion can include: (i) providing a probe/primer including an oligonucleotide containing a region of nucleotide sequence which hybridizes to a sense or antisense sequence from SEQ ID NO:1, or naturally occurring mutants thereof or 5' or 3' flanking sequences naturally associated with the
25 collagen XXII gene; (ii) exposing the probe/primer to nucleic acid of the tissue; and detecting, by hybridization, e.g., *in situ* hybridization, of the probe/primer to the nucleic acid, the presence or absence of the genetic lesion.

In preferred embodiments detecting the misexpression includes ascertaining the existence of at least one of: an alteration in the level of a messenger RNA
30 transcript of the collagen XXII gene; the presence of a non-wild type splicing pattern of a messenger RNA transcript of the gene; or a non-wild type level of collagen XXII.

Methods of the invention can be used prenatally or to determine if a subject's offspring will be at risk for a disorder.

In preferred embodiments the method includes determining the structure of a
35 collagen XXII gene, an abnormal structure being indicative of risk for the disorder.

5 In preferred embodiments the method includes contacting a sample from the subject with an antibody to the collagen XXII protein or a nucleic acid, which hybridizes specifically with the gene. These and other embodiments are discussed below.

Diagnostic and Prognostic Assays

10 Diagnostic and prognostic assays of the invention include method for assessing the expression level of collagen XXII molecules and for identifying variations and mutations in the sequence of collagen XXII molecules.

Expression Monitoring and Profiling. The presence, level, or absence of collagen XXII protein or nucleic acid in a biological sample can be evaluated by
15 obtaining a biological sample from a test subject and contacting the biological sample with a compound or an agent capable of detecting collagen XXII protein or nucleic acid (e.g., mRNA, genomic DNA) that encodes collagen XXII protein such that the presence of collagen XXII protein or nucleic acid is detected in the biological sample. The term "biological sample" includes tissues, cells and biological fluids isolated
20 from a subject, as well as tissues, cells and fluids present within a subject. A preferred biological sample is serum. The level of expression of the collagen XXII gene can be measured in a number of ways, including, but not limited to: measuring the mRNA encoded by the collagen XXII genes; measuring the amount of protein encoded by the collagen XXII genes; or measuring the activity of the protein encoded
25 by the collagen XXII genes.

 The level of mRNA corresponding to the collagen XXII gene in a cell can be determined both by *in situ* and by *in vitro* formats.

 The isolated mRNA can be used in hybridization or amplification assays that include, but are not limited to, Southern or Northern analyses, polymerase chain
30 reaction analyses and probe arrays. One preferred diagnostic method for the detection of mRNA levels involves contacting the isolated mRNA with a nucleic acid molecule (probe) that can hybridize to the mRNA encoded by the gene being detected. The nucleic acid probe can be, for example, a full-length collagen XXII nucleic acid, such as the nucleic acid of SEQ ID NO:1, or a portion thereof, such as an oligonucleotide

5 of at least 7, 15, 30, 50, 100, 250 or 500 nucleotides in length and sufficient to specifically hybridize under stringent conditions to collagen XXII mRNA or genomic DNA. The probe can be disposed on an address of an array, e.g., an array described below. Other suitable probes for use in the diagnostic assays are described herein.

10 In one format, mRNA (or cDNA) is immobilized on a surface and contacted with the probes, for example by running the isolated mRNA on an agarose gel and transferring the mRNA from the gel to a membrane, such as nitrocellulose. In an alternative format, the probes are immobilized on a surface and the mRNA (or cDNA) is contacted with the probes, for example, in a two-dimensional gene chip array described below. A skilled artisan can adapt known mRNA detection methods for use
15 in detecting the level of mRNA encoded by the collagen XXII genes.

The level of mRNA in a sample that is encoded by one of collagen XXII can be evaluated with nucleic acid amplification, e.g., by rtPCR (Mullis (1987) U.S. Patent No. 4,683,202), ligase chain reaction (Barany (1991) *Proc. Natl. Acad. Sci. USA* 88:189-193), self sustained sequence replication (Guatelli *et al.*, (1990) *Proc. Natl. Acad. Sci. USA* 87:1874-1878), transcriptional amplification system (Kwoh *et al.*, (1989), *Proc. Natl. Acad. Sci. USA* 86:1173-1177), Q-Beta Replicase (Lizardi *et al.*, (1988) *Bio/Technology* 6:1197), rolling circle replication (Lizardi *et al.*, U.S. Patent No. 5,854,033) or any other nucleic acid amplification method, followed by the detection of the amplified molecules using techniques known in the art. As used
25 herein, amplification primers are defined as being a pair of nucleic acid molecules that can anneal to 5' or 3' regions of a gene (plus and minus strands, respectively, or vice-versa) and contain a short region in between. In general, amplification primers are from about 10 to 30 nucleotides in length and flank a region from about 50 to 200 nucleotides in length. Under appropriate conditions and with appropriate reagents,
30 such primers permit the amplification of a nucleic acid molecule comprising the nucleotide sequence flanked by the primers.

For *in situ* methods, a cell or tissue sample can be prepared/processed and immobilized on a support, typically a glass slide, and then contacted with a probe that can hybridize to mRNA that encodes the collagen XXII gene being analyzed.

35 In another embodiment, the methods further contacting a control sample with a compound or agent capable of detecting collagen XXII mRNA, or genomic DNA,

5 and comparing the presence of collagen XXII mRNA or genomic DNA in the control sample with the presence of collagen XXII mRNA or genomic DNA in the test sample. In still another embodiment, serial analysis of gene expression, as described in U.S. Patent No. 5,695,937, is used to detect collagen XXII transcript levels.

10 A variety of methods can be used to determine the level of protein encoded by collagen XXII. In general, these methods include contacting an agent that selectively binds to the protein, such as an antibody with a sample, to evaluate the level of protein in the sample. In a preferred embodiment, the antibody bears a detectable label. Antibodies can be polyclonal, or more preferably, monoclonal. An intact antibody, or a fragment thereof (e.g., Fab or F(ab')₂) can be used. The term "labeled", with regard
15 to the probe or antibody, is intended to encompass direct labeling of the probe or antibody by coupling (i.e., physically linking) a detectable substance to the probe or antibody, as well as indirect labeling of the probe or antibody by reactivity with a detectable substance. Examples of detectable substances are provided herein.

The detection methods can be used to detect collagen XXII protein in a
20 biological sample *in vitro* as well as *in vivo*. *In vitro* techniques for detection of collagen XXII protein include enzyme linked immunosorbent assays (ELISAs), immunoprecipitations, immunofluorescence, enzyme immunoassay (EIA), radioimmunoassay (RIA), and Western blot analysis. *In vivo* techniques for detection of collagen XXII protein include introducing into a subject a labeled anti-collagen
25 XXII antibody. For example, the antibody can be labeled with a radioactive marker whose presence and location in a subject can be detected by standard imaging techniques. In another embodiment, the sample is labeled, e.g., biotinylated and then contacted to the antibody, e.g., an anti-collagen XXII antibody positioned on an antibody array (as described below). The sample can be detected, e.g., with avidin
30 coupled to a fluorescent label.

In another embodiment, the methods further include contacting the control sample with a compound or agent capable of detecting collagen XXII protein, and comparing the presence of collagen XXII protein in the control sample with the presence of collagen XXII protein in the test sample.

35 The invention also includes kits for detecting the presence of collagen XXII in a biological sample. For example, the kit can include a compound or agent capable of

5 detecting collagen XXII protein or mRNA in a biological sample; and a standard.
The compound or agent can be packaged in a suitable container. The kit can further
comprise instructions for using the kit to detect collagen XXII protein or nucleic acid.

For antibody-based kits, the kit can include: (1) a first antibody (e.g., attached
to a solid support) which binds to a polypeptide corresponding to a marker of the
10 invention; and, optionally, (2) a second, different antibody which binds to either the
polypeptide or the first antibody and is conjugated to a detectable agent.

For oligonucleotide-based kits, the kit can include: (1) an oligonucleotide, e.g.,
a detectably labeled oligonucleotide, which hybridizes to a nucleic acid sequence
encoding a polypeptide corresponding to a marker of the invention or (2) a pair of
15 primers useful for amplifying a nucleic acid molecule corresponding to a marker of
the invention. The kit can also include a buffering agent, a preservative, or a protein
stabilizing agent. The kit can also include components necessary for detecting the
detectable agent (e.g., an enzyme or a substrate). The kit can also contain a control
sample or a series of control samples which can be assayed and compared to the test
20 sample contained. Each component of the kit can be enclosed within an individual
container and all of the various containers can be within a single package, along with
instructions for interpreting the results of the assays performed using the kit.

The diagnostic methods described herein can identify subjects having, or at
risk of developing, a disease or disorder associated with misexpressed or aberrant or
25 unwanted collagen XXII expression or activity. As used herein, the term "unwanted"
includes an unwanted phenomenon involved in a biological response such as
connective tissue formation and maintenance.

In one embodiment, a disease or disorder associated with aberrant or unwanted
collagen XXII expression or activity is identified. A test sample is obtained from a
30 subject and collagen XXII protein or nucleic acid (e.g., mRNA or genomic DNA) is
evaluated, wherein the level, e.g., the presence or absence, of collagen XXII protein
or nucleic acid is diagnostic for a subject having or at risk of developing a disease or
disorder associated with aberrant or unwanted collagen XXII expression or activity.
As used herein, a "test sample" refers to a biological sample obtained from a subject
35 of interest, including a biological fluid (e.g., serum), cell sample, or tissue.

5 The prognostic assays described herein can be used to determine whether a subject can be administered an agent (e.g., an agonist, antagonist, peptidomimetic, protein, peptide, nucleic acid, small molecule, or other drug candidate) to treat a disease or disorder associated with aberrant or unwanted collagen XXII expression or activity.

10 In another aspect, the invention features a computer medium having a plurality of digitally encoded data records. Each data record includes a value representing the level of expression of collagen XXII in a sample, and a descriptor of the sample. The descriptor of the sample can be an identifier of the sample, a subject from which the sample was derived (e.g., a patient), a diagnosis, or a treatment (e.g., a preferred
15 treatment). In a preferred embodiment, the data record further includes values representing the level of expression of genes other than collagen XXII (e.g., other genes associated with a collagen XXII-disorder, or other genes on an array). The data record can be structured as a table, e.g., a table that is part of a database such as a relational database (e.g., a SQL database of the Oracle or Sybase database
20 environments).

 Also featured is a method of evaluating a sample. The method includes providing a sample, e.g., from the subject, and determining a gene expression profile of the sample, wherein the profile includes a value representing the level of collagen XXII expression. The method can further include comparing the value or the profile
25 (i.e., multiple values) to a reference value or reference profile. The gene expression profile of the sample can be obtained by any of the methods described herein (e.g., by providing a nucleic acid from the sample and contacting the nucleic acid to an array). The method can be used to diagnose a connective tissue disorder in a subject wherein an increase or decrease in collagen XXII expression is an indication that the subject
30 has or is disposed to having a connective tissue disorder. The method can be used to monitor a treatment for a connective tissue disorder in a subject. For example, the gene expression profile can be determined for a sample from a subject undergoing treatment. The profile can be compared to a reference profile or to a profile obtained from the subject prior to treatment or prior to onset of the disorder (see, e.g., Golub *et al.* (1999) *Science* 286:531).
35

5 In yet another aspect, the invention features a method of evaluating a test compound (see also, "Screening Assays", above). The method includes providing a cell and a test compound; contacting the test compound to the cell; obtaining a subject expression profile for the contacted cell; and comparing the subject expression profile to one or more reference profiles. The profiles include a value representing the level
10 of collagen XXII expression. In a preferred embodiment, the subject expression profile is compared to a target profile, e.g., a profile for a normal cell or for desired condition of a cell. The test compound is evaluated favorably if the subject expression profile is more similar to the target profile than an expression profile obtained from an uncontacted cell.

15 In another aspect, the invention features, a method of evaluating a subject. The method includes: a) obtaining a sample from a subject, e.g., from a caregiver, e.g., a caregiver who obtains the sample from the subject; b) determining a subject expression profile for the sample. Optionally, the method further includes either or
20 both of steps: c) comparing the subject expression profile to one or more reference expression profiles; and d) selecting the reference profile most similar to the subject reference profile. The subject expression profile and the reference profiles include a value representing the level of collagen XXII expression. A variety of routine statistical measures can be used to compare two reference profiles. One possible
25 metric is the length of the distance vector that is the difference between the two profiles. Each of the subject and reference profile is represented as a multi-dimensional vector, wherein each dimension is a value in the profile.

 The method can further include transmitting a result to a caregiver. The result can be the subject expression profile, a result of a comparison of the subject expression profile with another profile, a most similar reference profile, or a
30 descriptor of any of the aforementioned. The result can be transmitted across a computer network, e.g., the result can be in the form of a computer transmission, e.g., a computer data signal embedded in a carrier wave.

 Also featured is a computer medium having executable code for effecting the following steps: receive a subject expression profile; access a database of reference
35 expression profiles; and either i) select a matching reference profile most similar to the subject expression profile or ii) determine at least one comparison score for the

- 5 similarity of the subject expression profile to at least one reference profile. The subject expression profile, and the reference expression profiles each include a value representing the level of collagen XXII expression.

Arrays and Uses Thereof

In another aspect, the invention features an array that includes a substrate
10 having a plurality of addresses. At least one address of the plurality includes a capture probe that binds specifically to a collagen XXII molecule (e.g., a collagen XXII nucleic acid or a collagen XXII polypeptide). The array can have a density of at least than 10, 50, 100, 200, 500, 1,000, 2,000, or 10,000 or more addresses/cm², and ranges between. In a preferred embodiment, the plurality of addresses includes at
15 least 10, 100, 500, 1,000, 5,000, 10,000, 50,000 addresses. In a preferred embodiment, the plurality of addresses includes equal to or less than 10, 100, 500, 1,000, 5,000, 10,000, or 50,000 addresses. The substrate can be a two-dimensional substrate such as a glass slide, a wafer (e.g., silica or plastic), a mass spectroscopy plate, or a three-dimensional substrate such as a gel pad. Addresses in addition to
20 address of the plurality can be disposed on the array.

In a preferred embodiment, at least one address of the plurality includes a nucleic acid capture probe that hybridizes specifically to a collagen XXII nucleic acid, e.g., the sense or anti-sense strand. In one preferred embodiment, a subset of addresses of the plurality of addresses has a nucleic acid capture probe for collagen
25 XXII. Each address of the subset can include a capture probe that hybridizes to a different region of a collagen XXII nucleic acid. In another preferred embodiment, addresses of the subset include a capture probe for a collagen XXII nucleic acid. Each address of the subset is unique, overlapping, and complementary to a different variant of collagen XXII (e.g., an allelic variant, or all possible hypothetical variants).
30 The array can be used to sequence collagen XXII by hybridization (see, e.g., U.S. Patent No. 5,695,940).

An array can be generated by various methods, e.g., by photolithographic methods (see, e.g., U.S. Patent Nos. 5,143,854; 5,510,270; and 5,527,681), mechanical methods (e.g., directed-flow methods as described in U.S. Patent No.

5 5,384,261), pin-based methods (e.g., as described in U.S. Pat. No. 5,288,514), and bead-based techniques (e.g., as described in PCT US/93/04145).

In another preferred embodiment, at least one address of the plurality includes a polypeptide capture probe that binds specifically to a collagen XXII polypeptide or fragment thereof. The polypeptide can be a naturally-occurring interaction partner of collagen XXII polypeptide. Preferably, the polypeptide is an antibody, e.g., an
10 antibody described herein (see "Anti-collagen XXII Antibodies," above), such as a monoclonal antibody or a single-chain antibody.

In another aspect, the invention features a method of analyzing the expression of collagen XXII. The method includes providing an array as described above;
15 contacting the array with a sample and detecting binding of a collagen XXII-molecule (e.g., nucleic acid or polypeptide) to the array. In a preferred embodiment, the array is a nucleic acid array. Optionally the method further includes amplifying nucleic acid from the sample prior or during contact with the array.

In another embodiment, the array can be used to assay gene expression in a tissue to ascertain tissue specificity of genes in the array, particularly the expression of collagen XXII. If a sufficient number of diverse samples is analyzed, clustering (e.g., hierarchical clustering, k-means clustering, Bayesian clustering and the like) can be used to identify other genes which are co-regulated with collagen XXII. For example, the array can be used for the quantitation of the expression of multiple
25 genes. Thus, not only tissue specificity, but also the level of expression of a battery of genes in the tissue is ascertained. Quantitative data can be used to group (e.g., cluster) genes on the basis of their tissue expression *per se* and level of expression in that tissue.

For example, array analysis of gene expression can be used to assess the effect of cell-cell interactions on collagen XXII expression. A first tissue can be perturbed and nucleic acid from a second tissue that interacts with the first tissue can be analyzed. In this context, the effect of one cell type on another cell type in response to a biological stimulus can be determined, e.g., to monitor the effect of cell-cell interaction at the level of gene expression.

35 In another embodiment, cells are contacted with a therapeutic agent. The expression profile of the cells is determined using the array, and the expression profile

5 is compared to the profile of like cells not contacted with the agent. For example, the assay can be used to determine or analyze the molecular basis of an undesirable effect of the therapeutic agent. If an agent is administered therapeutically to treat one cell type but has an undesirable effect on another cell type, the invention provides an assay to determine the molecular basis of the undesirable effect and thus provides the
10 opportunity to co-administer a counteracting agent or otherwise treat the undesired effect. Similarly, even within a single cell type, undesirable biological effects can be determined at the molecular level. Thus, the effects of an agent on expression of other than the target gene can be ascertained and counteracted.

In another embodiment, the array can be used to monitor expression of one or
15 more genes in the array with respect to time. For example, samples obtained from different time points can be probed with the array. Such analysis can identify and/or characterize the development of a collagen XXII-associated disease or disorder; and processes, such as a cellular transformation associated with a collagen XXII-associated disease or disorder. The method can also evaluate the treatment and/or
20 progression of a collagen XXII-associated disease or disorder

The array is also useful for ascertaining differential expression patterns of one or more genes in normal and abnormal cells. This provides a battery of genes (*e.g.*, including collagen XXII) that could serve as a molecular target for diagnosis or therapeutic intervention.

25 In another aspect, the invention features an array having a plurality of addresses. Each address of the plurality includes a unique polypeptide. At least one address of the plurality has disposed thereon a collagen XXII polypeptide or fragment thereof. Methods of producing polypeptide arrays are described in the art, *e.g.*, in De Wildt *et al.* (2000). *Nature Biotech.* 18, 989-994; Lueking *et al.* (1999). *Anal.*
30 *Biochem.* 270, 103-111; Ge, H. (2000). *Nucleic Acids Res.* 28, e3, I-VII; MacBeath, G., and Schreiber, S.L. (2000). *Science* 289, 1760-1763; and WO 99/51773A1. In a preferred embodiment, each addresses of the plurality has disposed thereon a polypeptide at least 60, 70, 80, 85, 90, 95 or 99 % identical to a collagen XXII polypeptide or fragment thereof. For example, multiple variants of a collagen XXII
35 polypeptide (*e.g.*, encoded by allelic variants, site-directed mutants, random mutants,

5 or combinatorial mutants) can be disposed at individual addresses of the plurality.
Addresses in addition to the address of the plurality can be disposed on the array.

The polypeptide array can be used to detect a collagen XXII binding compound, e.g., an antibody in a sample from a subject with specificity for a collagen XXII polypeptide or the presence of a collagen XXII-binding protein or ligand.

10 The array is also useful for ascertaining the effect of the expression of a gene on the expression of other genes in the same cell or in different cells (*e.g.*, ascertaining the effect of collagen XXII expression on the expression of other genes). This provides, for example, for a selection of alternate molecular targets for therapeutic intervention if the ultimate or downstream target cannot be regulated.

15 In another aspect, the invention features a method of analyzing a plurality of probes. The method is useful, e.g., for analyzing gene expression. The method includes: providing a two dimensional array having a plurality of addresses, each address of the plurality being positionally distinguishable from each other address of the plurality having a unique capture probe, e.g., wherein the capture probes are from
20 a cell or subject which express collagen XXII or from a cell or subject in which a collagen XXII mediated response has been elicited, e.g., by contact of the cell with collagen XXII nucleic acid or protein, or administration to the cell or subject collagen XXII nucleic acid or protein; providing a two dimensional array having a plurality of addresses, each address of the plurality being positionally distinguishable from each
25 other address of the plurality, and each address of the plurality having a unique capture probe, e.g., wherein the capture probes are from a cell or subject which does not express collagen XXII (or does not express as highly as in the case of the collagen XXII positive plurality of capture probes) or from a cell or subject which in which a collagen XXII mediated response has not been elicited (or has been elicited to a lesser
30 extent than in the first sample); contacting the array with one or more inquiry probes (which is preferably other than a collagen XXII nucleic acid, polypeptide, or antibody), and thereby evaluating the plurality of capture probes. Binding, e.g., in the case of a nucleic acid, hybridization with a capture probe at an address of the plurality, is detected, e.g., by signal generated from a label attached to the nucleic
35 acid, polypeptide, or antibody.

5 In another aspect, the invention features a method of analyzing a plurality of probes or a sample. The method is useful, e.g., for analyzing gene expression. The method includes: providing a two dimensional array having a plurality of addresses, each address of the plurality being positionally distinguishable from each other address of the plurality having a unique capture probe, contacting the array with a first
10 sample from a cell or subject which express or mis-express collagen XXII or from a cell or subject in which a collagen XXII-mediated response has been elicited, e.g., by contact of the cell with collagen XXII nucleic acid or protein, or administration to the cell or subject collagen XXII nucleic acid or protein; providing a two dimensional array having a plurality of addresses, each address of the plurality being positionally
15 distinguishable from each other address of the plurality, and each address of the plurality having a unique capture probe, and contacting the array with a second sample from a cell or subject which does not express collagen XXII (or does not express as highly as in the case of the collagen XXII positive plurality of capture probes) or from a cell or subject which in which a collagen XXII mediated response
20 has not been elicited (or has been elicited to a lesser extent than in the first sample); and comparing the binding of the first sample with the binding of the second sample. Binding, e.g., in the case of a nucleic acid, hybridization with a capture probe at an address of the plurality, is detected, e.g., by signal generated from a label attached to the nucleic acid, polypeptide, or antibody. The same array can be used for both
25 samples or different arrays can be used. If different arrays are used the plurality of addresses with capture probes should be present on both arrays.

 In another aspect, the invention features a method of analyzing collagen XXII, e.g., analyzing structure, function, or relatedness to other nucleic acid or amino acid sequences. The method includes: providing a collagen XXII nucleic acid or amino
30 acid sequence; comparing the collagen XXII sequence with one or more preferably a plurality of sequences from a collection of sequences, e.g., a nucleic acid or protein sequence database; to thereby analyze collagen XXII.

5 Detection of Variations or Mutations

 The methods of the invention can also be used to detect genetic alterations in a collagen XXII gene, thereby determining if a subject with the altered gene is at risk for a disorder characterized by misregulation in collagen XXII protein activity or nucleic acid expression, such as a connective tissue disorder. In preferred
10 embodiments, the methods include detecting, in a sample from the subject, the presence or absence of a genetic alteration characterized by at least one of an alteration affecting the integrity of a gene encoding a collagen XXII-protein, or the mis-expression of the collagen XXII gene. For example, such genetic alterations can be detected by ascertaining the existence of at least one of 1) a deletion of one or
15 more nucleotides from a collagen XXII gene; 2) an addition of one or more nucleotides to a collagen XXII gene; 3) a substitution of one or more nucleotides of a collagen XXII gene; 4) a chromosomal rearrangement of a collagen XXII gene; 5) an alteration in the level of a messenger RNA transcript of a collagen XXII gene; 6) aberrant modification of a collagen XXII gene, such as of the methylation pattern of
20 the genomic DNA; 7) the presence of a non-wild type splicing pattern of a messenger RNA transcript of a collagen XXII gene; 8) a non-wild type level of a collagen XXII-protein; 9) allelic loss of a collagen XXII gene, and 10) inappropriate post-translational modification of a collagen XXII-protein.

 An alteration can be detected without a probe/primer in a polymerase chain
25 reaction, such as anchor PCR or RACE PCR, or, alternatively, in a ligation chain reaction (LCR), the latter of which can be particularly useful for detecting point mutations in the collagen XXII-gene. This method can include the steps of collecting a sample of cells from a subject, isolating nucleic acid (e.g., genomic, mRNA or both) from the sample, contacting the nucleic acid sample with one or more primers which
30 specifically hybridize to a collagen XXII gene under conditions such that hybridization and amplification of the collagen XXII-gene (if present) occurs, and detecting the presence or absence of an amplification product, or detecting the size of the amplification product and comparing the length to a control sample. It is anticipated that PCR and/or LCR may be desirable to use as a preliminary
35 amplification step in conjunction with any of the techniques used for detecting

5 mutations described herein. Alternatively, other amplification methods described herein or known in the art can be used.

In another embodiment, mutations in a collagen XXII gene from a sample cell can be identified by detecting alterations in restriction enzyme cleavage patterns. For example, sample and control DNA is isolated, amplified (optionally), digested with
10 one or more restriction endonucleases, and fragment length sizes are determined, e.g., by gel electrophoresis and compared. Differences in fragment length sizes between sample and control DNA indicates mutations in the sample DNA. Moreover, the use of sequence specific ribozymes (see, for example, U.S. Patent No. 5,498,531) can be used to score for the presence of specific mutations by development or loss of a
15 ribozyme cleavage site.

In other embodiments, genetic mutations in collagen XXII can be identified by hybridizing a sample and control nucleic acids, e.g., DNA or RNA, two-dimensional arrays, e.g., chip based arrays. Such arrays include a plurality of addresses, each of which is positionally distinguishable from the other. A different probe is located at
20 each address of the plurality. A probe can be complementary to a region of a collagen XXII nucleic acid or a putative variant (e.g., allelic variant) thereof. A probe can have one or more mismatches to a region of a collagen XXII nucleic acid (e.g., a destabilizing mismatch). The arrays can have a high density of addresses, e.g., can contain hundreds or thousands of oligonucleotides probes (Cronin, M.T. *et al.* (1996)
25 *Human Mutation* 7: 244-255; Kozal, M.J. *et al.* (1996) *Nature Medicine* 2: 753-759). For example, genetic mutations in collagen XXII can be identified in two-dimensional arrays containing light-generated DNA probes as described in Cronin, M.T. *et al. supra*. Briefly, a first hybridization array of probes can be used to scan through long stretches of DNA in a sample and control to identify base changes between the
30 sequences by making linear arrays of sequential overlapping probes. This step allows the identification of point mutations. This step is followed by a second hybridization array that allows the characterization of specific mutations by using smaller, specialized probe arrays complementary to all variants or mutations detected. Each mutation array is composed of parallel probe sets, one complementary to the wild-
35 type gene and the other complementary to the mutant gene.

5 In yet another embodiment, any of a variety of sequencing reactions known in the art can be used to directly sequence the collagen XXII gene and detect mutations by comparing the sequence of the sample collagen XXII with the corresponding wild-type (control) sequence. Automated sequencing procedures can be utilized when performing the diagnostic assays ((1995) *Biotechniques* 19:448), including
10 sequencing by mass spectrometry.

 Other methods for detecting mutations in the collagen XXII gene include methods in which protection from cleavage agents is used to detect mismatched bases in RNA/RNA or RNA/DNA heteroduplexes (Myers *et al.* (1985) *Science* 230:1242; Cotton *et al.* (1988) *Proc. Natl Acad Sci USA* 85:4397; Saleeba *et al.* (1992) *Methods*
15 *Enzymol.* 217:286-295).

 In still another embodiment, the mismatch cleavage reaction employs one or more proteins that recognize mismatched base pairs in double-stranded DNA (so called "DNA mismatch repair" enzymes) in defined systems for detecting and mapping point mutations in collagen XXII cDNAs obtained from samples of cells.
20 For example, the mutY enzyme of *E. coli* cleaves A at G/A mismatches and the thymidine DNA glycosylase from HeLa cells cleaves T at G/T mismatches (Hsu *et al.* (1994) *Carcinogenesis* 15:1657-1662; U.S. Patent No. 5,459,039).

 In other embodiments, alterations in electrophoretic mobility will be used to identify mutations in collagen XXII genes. For example, single strand conformation
25 polymorphism (SSCP) may be used to detect differences in electrophoretic mobility between mutant and wild type nucleic acids (Orita *et al.* (1989) *Proc Natl. Acad. Sci USA*: 86:2766, see also Cotton (1993) *Mutat. Res.* 285:125-144; and Hayashi (1992) *Genet. Anal. Tech. Appl.* 9:73-79). Single-stranded DNA fragments of sample and control collagen XXII nucleic acids will be denatured and allowed to renature. The
30 secondary structure of single-stranded nucleic acids varies according to sequence, the resulting alteration in electrophoretic mobility enables the detection of even a single base change. The DNA fragments may be labeled or detected with labeled probes. The sensitivity of the assay may be enhanced by using RNA (rather than DNA), in which the secondary structure is more sensitive to a change in sequence. In a
35 preferred embodiment, the subject method utilizes heteroduplex analysis to separate

5 double stranded heteroduplex molecules on the basis of changes in electrophoretic mobility (Keen *et al.* (1991) *Trends Genet* 7:5).

In yet another embodiment, the movement of mutant or wild-type fragments in polyacrylamide gels containing a gradient of denaturant is assayed using denaturing gradient gel electrophoresis (DGGE) (Myers *et al.* (1985) *Nature* 313:495). When
10 DGGE is used as the method of analysis, DNA will be modified to insure that it does not completely denature, for example by adding a GC clamp of approximately 40 bp of high-melting GC-rich DNA by PCR. In a further embodiment, a temperature gradient is used in place of a denaturing gradient to identify differences in the mobility of control and sample DNA (Rosenbaum and Reissner (1987) *Biophys Chem*
15 265:12753).

Examples of other techniques for detecting point mutations include, but are not limited to, selective oligonucleotide hybridization, selective amplification, or selective primer extension (Saiki *et al.* (1986) *Nature* 324:163); Saiki *et al.* (1989) *Proc. Natl Acad. Sci USA* 86:6230). A further method of detecting point mutations is
20 the chemical ligation of oligonucleotides as described in Xu *et al.* ((2001) *Nature Biotechnol.* 19:148). Adjacent oligonucleotides, one of which selectively anneals to the query site, are ligated together if the nucleotide at the query site of the sample nucleic acid is complementary to the query oligonucleotide; ligation can be monitored, e.g., by fluorescent dyes coupled to the oligonucleotides.

25 Alternatively, allele specific amplification technology that depends on selective PCR amplification may be used in conjunction with the instant invention. Oligonucleotides used as primers for specific amplification may carry the mutation of interest in the center of the molecule (so that amplification depends on differential hybridization) (Gibbs *et al.* (1989) *Nucleic Acids Res.* 17:2437-2448) or at the
30 extreme 3' end of one primer where, under appropriate conditions, mismatch can prevent, or reduce polymerase extension (Prossner (1993) *Tibtech* 11:238). In addition it may be desirable to introduce a novel restriction site in the region of the mutation to create cleavage-based detection (Gasparini *et al.* (1992) *Mol. Cell Probes* 6:1). It is anticipated that in certain embodiments amplification may also be
35 performed using Taq ligase for amplification (Barany (1991) *Proc. Natl. Acad. Sci USA* 88:189). In such cases, ligation will occur only if there is a perfect match at the

5 3' end of the 5' sequence making it possible to detect the presence of a known mutation at a specific site by looking for the presence or absence of amplification.

In another aspect, the invention features a set of oligonucleotides. The set includes a plurality of oligonucleotides, each of which is at least partially complementary (e.g., at least 50%, 60%, 70%, 80%, 90%, 92%, 95%, 97%, 98%, or
10 99% complementary) to a collagen XXII nucleic acid.

In a preferred embodiment the set includes a first and a second oligonucleotide. The first and second oligonucleotide can hybridize to the same or to different locations of SEQ ID NO:1 or the complement of SEQ ID NO:1. Different locations can be different but overlapping, or non-overlapping on the same strand.
15 The first and second oligonucleotide can hybridize to sites on the same or on different strands.

The set can be useful, e.g., for identifying SNP's, or identifying specific alleles of collagen XXII. In a preferred embodiment, each oligonucleotide of the set has a different nucleotide at an interrogation position. In one embodiment, the set
20 includes two oligonucleotides, each complementary to a different allele at a locus, e.g., a biallelic or polymorphic locus.

In another embodiment, the set includes four oligonucleotides, each having a different nucleotide (e.g., adenine, guanine, cytosine, or thymidine) at the interrogation position. The interrogation position can be a SNP or the site of a
25 mutation. In another preferred embodiment, the oligonucleotides of the plurality are identical in sequence to one another (except for differences in length). The oligonucleotides can be provided with differential labels, such that an oligonucleotide that hybridizes to one allele provides a signal that is distinguishable from an oligonucleotide that hybridizes to a second allele. In still another embodiment, at
30 least one of the oligonucleotides of the set has a nucleotide change at a position in addition to a query position, e.g., a destabilizing mutation to decrease the T_m of the oligonucleotide. In another embodiment, at least one oligonucleotide of the set has a non-natural nucleotide, e.g., inosine. In a preferred embodiment, the oligonucleotides are attached to a solid support, e.g., to different addresses of an array or to different
35 beads or nanoparticles.

5 In a preferred embodiment the set of oligo nucleotides can be used to specifically amplify, e.g., by PCR, or detect, a collagen XXII nucleic acid.

The methods described herein may be performed, for example, by utilizing pre-packaged diagnostic kits comprising at least one probe nucleic acid or antibody reagent described herein, which may be conveniently used, e.g., in clinical settings to
10 diagnose patients exhibiting symptoms or family history of a disease or illness involving a collagen XXII gene.

Use of collagen XXII Molecules as Surrogate Markers

The collagen XXII molecules of the invention are also useful as markers of disorders or disease states, as markers for precursors of disease states, as markers for
15 predisposition of disease states, as markers of drug activity, or as markers of the pharmacogenomic profile of a subject. Using the methods described herein, the presence, absence and/or quantity of the collagen XXII molecules of the invention may be detected, and may be correlated with one or more biological states in vivo. For example, the collagen XXII molecules of the invention may serve as surrogate
20 markers for one or more disorders or disease states or for conditions leading up to disease states. As used herein, a "surrogate marker" is an objective biochemical marker which correlates with the absence or presence of a disease or disorder, or with the progression of a disease or disorder (e.g., with the presence or absence of a tumor). The presence or quantity of such markers is independent of the disease.
25 Therefore, these markers may serve to indicate whether a particular course of treatment is effective in lessening a disease state or disorder. Surrogate markers are of particular use when the presence or extent of a disease state or disorder is difficult to assess through standard methodologies (e.g., early stage tumors), or when an assessment of disease progression is desired before a potentially dangerous clinical
30 endpoint is reached (e.g., an assessment of cardiovascular disease may be made using cholesterol levels as a surrogate marker, and an analysis of HIV infection may be made using HIV RNA levels as a surrogate marker, well in advance of the undesirable clinical outcomes of myocardial infarction or fully-developed AIDS). Examples of

5 the use of surrogate markers in the art include: Koomen *et al.* (2000) *J. Mass. Spectrom.* 35: 258-264; and James (1994) *AIDS Treatment News Archive* 209.

The collagen XXII molecules of the invention are also useful as pharmacodynamic markers. As used herein, a "pharmacodynamic marker" is an objective biochemical marker which correlates specifically with drug effects. The presence or quantity of a pharmacodynamic marker is not related to the disease state or disorder for which the drug is being administered; therefore, the presence or quantity of the marker is indicative of the presence or activity of the drug in a subject. For example, a pharmacodynamic marker may be indicative of the concentration of the drug in a biological tissue, in that the marker is either expressed or transcribed or not expressed or transcribed in that tissue in relationship to the level of the drug. In this fashion, the distribution or uptake of the drug may be monitored by the pharmacodynamic marker. Similarly, the presence or quantity of the pharmacodynamic marker may be related to the presence or quantity of the metabolic product of a drug, such that the presence or quantity of the marker is indicative of the relative breakdown rate of the drug in vivo. Pharmacodynamic markers are of particular use in increasing the sensitivity of detection of drug effects, particularly when the drug is administered in low doses. Since even a small amount of a drug may be sufficient to activate multiple rounds of marker (e.g., a collagen XXII marker) transcription or expression, the amplified marker may be in a quantity which is more readily detectable than the drug itself. Also, the marker may be more easily detected due to the nature of the marker itself; for example, using the methods described herein, anti-collagen XXII antibodies may be employed in an immune-based detection system for a collagen XXII protein marker, or collagen XXII-specific radiolabeled probes may be used to detect a collagen XXII mRNA marker. Furthermore, the use of a pharmacodynamic marker may offer mechanism-based prediction of risk due to drug treatment beyond the range of possible direct observations. Examples of the use of pharmacodynamic markers in the art include: Matsuda *et al.* US 6,033,862; Hattis *et al.* (1991) *Env. Health Perspect.* 90: 229-238; Schentag (1999) *Am. J. Health-Syst. Pharm.* 56 Suppl. 3: S21-S24; and Nicolau (1999) *Am. J. Health-Syst. Pharm.* 56 Suppl. 3: S16-S20.

5 The collagen XXII molecules of the invention are also useful as pharmacogenomic markers. As used herein, a "pharmacogenomic marker" is an objective biochemical marker which correlates with a specific clinical drug response or susceptibility in a subject (see, e.g., McLeod *et al.* (1999) *Eur. J. Cancer* 35:1650-1652). The presence or quantity of the pharmacogenomic marker is related to the
10 predicted response of the subject to a specific drug or class of drugs prior to administration of the drug. By assessing the presence or quantity of one or more pharmacogenomic markers in a subject, a drug therapy which is most appropriate for the subject, or which is predicted to have a greater degree of success, may be selected. For example, based on the presence or quantity of RNA, or protein (e.g., collagen
15 XXII protein or RNA) for specific tumor markers in a subject, a drug or course of treatment may be selected that is optimized for the treatment of the specific tumor likely to be present in the subject. Similarly, the presence or absence of a specific sequence mutation in collagen XXII DNA may correlate collagen XXII drug response. The use of pharmacogenomic markers therefore permits the application of
20 the most appropriate treatment for each subject without having to administer the therapy.

Pharmaceutical Compositions

 The nucleic acid and polypeptides, fragments thereof, as well as anti-collagen XXII antibodies (also referred to herein as "active compounds") of the invention can
25 be incorporated into pharmaceutical compositions. Such compositions typically include the nucleic acid molecule, protein, or antibody and a pharmaceutically acceptable carrier. As used herein the language "pharmaceutically acceptable carrier" includes solvents, dispersion media, coatings, antibacterial and antifungal agents, isotonic and absorption delaying agents, and the like, compatible with pharmaceutical
30 administration. Supplementary active compounds can also be incorporated into the compositions.

 A pharmaceutical composition is formulated to be compatible with its intended route of administration. Examples of routes of administration include parenteral, e.g., intravenous, intradermal, subcutaneous, oral (e.g., inhalation),

5 transdermal (topical), transmucosal, and rectal administration. Solutions or suspensions used for parenteral, intradermal, or subcutaneous application can include the following components: a sterile diluent such as water for injection, saline solution, fixed oils, polyethylene glycols, glycerine, propylene glycol or other synthetic solvents; antibacterial agents such as benzyl alcohol or methyl parabens; antioxidants
10 such as ascorbic acid or sodium bisulfite; chelating agents such as ethylenediaminetetraacetic acid; buffers such as acetates, citrates or phosphates and agents for the adjustment of tonicity such as sodium chloride or dextrose. pH can be adjusted with acids or bases, such as hydrochloric acid or sodium hydroxide. The parenteral preparation can be enclosed in ampoules, disposable syringes or multiple
15 dose vials made of glass or plastic.

Pharmaceutical compositions suitable for injectable use include sterile aqueous solutions (where water soluble) or dispersions and sterile powders for the extemporaneous preparation of sterile injectable solutions or dispersion. For intravenous administration, suitable carriers include physiological saline,
20 bacteriostatic water, Cremophor EL™ (BASF, Parsippany, NJ) or phosphate buffered saline (PBS). In all cases, the composition must be sterile and should be fluid to the extent that easy syringability exists. It should be stable under the conditions of manufacture and storage and must be preserved against the contaminating action of microorganisms such as bacteria and fungi. The carrier can be a solvent or dispersion
25 medium containing, for example, water, ethanol, polyol (for example, glycerol, propylene glycol, and liquid polyethylene glycol, and the like), and suitable mixtures thereof. The proper fluidity can be maintained, for example, by the use of a coating such as lecithin, by the maintenance of the required particle size in the case of dispersion and by the use of surfactants. Prevention of the action of microorganisms
30 can be achieved by various antibacterial and antifungal agents, for example, parabens, chlorobutanol, phenol, ascorbic acid, thimerosal, and the like. In many cases, it will be preferable to include isotonic agents, for example, sugars, polyalcohols such as manitol, sorbitol, sodium chloride in the composition. Prolonged absorption of the injectable compositions can be brought about by including in the composition an
35 agent which delays absorption, for example, aluminum monostearate and gelatin.

5 Sterile injectable solutions can be prepared by incorporating the active compound in the required amount in an appropriate solvent with one or a combination of ingredients enumerated above, as required, followed by filtered sterilization. Generally, dispersions are prepared by incorporating the active compound into a sterile vehicle which contains a basic dispersion medium and the required other
10 ingredients from those enumerated above. In the case of sterile powders for the preparation of sterile injectable solutions, the preferred methods of preparation are vacuum drying and freeze-drying which yields a powder of the active ingredient plus any additional desired ingredient from a previously sterile-filtered solution thereof.

Oral compositions generally include an inert diluent or an edible carrier. For
15 the purpose of oral therapeutic administration, the active compound can be incorporated with excipients and used in the form of tablets, troches, or capsules, e.g., gelatin capsules. Oral compositions can also be prepared using a fluid carrier for use as a mouthwash. Pharmaceutically compatible binding agents, and/or adjuvant materials can be included as part of the composition. The tablets, pills, capsules,
20 troches and the like can contain any of the following ingredients, or compounds of a similar nature: a binder such as microcrystalline cellulose, gum tragacanth or gelatin; an excipient such as starch or lactose, a disintegrating agent such as alginic acid, Primogel, or corn starch; a lubricant such as magnesium stearate or Sterotes; a glidant such as colloidal silicon dioxide; a sweetening agent such as sucrose or saccharin; or a
25 flavoring agent such as peppermint, methyl salicylate, or orange flavoring.

For administration by inhalation, the compounds are delivered in the form of an aerosol spray from pressured container or dispenser which contains a suitable propellant, e.g., a gas such as carbon dioxide, or a nebulizer.

Systemic administration can also be by transmucosal or transdermal means.
30 For transmucosal or transdermal administration, penetrants appropriate to the barrier to be permeated are used in the formulation. Such penetrants are generally known in the art, and include, for example, for transmucosal administration, detergents, bile salts, and fusidic acid derivatives. Transmucosal administration can be accomplished through the use of nasal sprays or suppositories. For transdermal administration, the
35 active compounds are formulated into ointments, salves, gels, or creams as generally known in the art.

5 The compounds can also be prepared in the form of suppositories (e.g., with conventional suppository bases such as cocoa butter and other glycerides) or retention enemas for rectal delivery.

 In one embodiment, the active compounds are prepared with carriers that will protect the compound against rapid elimination from the body, such as a controlled
10 release formulation, including implants and microencapsulated delivery systems. Biodegradable, biocompatible polymers can be used, such as ethylene vinyl acetate, polyanhydrides, polyglycolic acid, collagen, polyorthoesters, and polylactic acid. Methods for preparation of such formulations will be apparent to those skilled in the art. The materials can also be obtained commercially from Alza Corporation and
15 Nova Pharmaceuticals, Inc. Liposomal suspensions (including liposomes targeted to infected cells with monoclonal antibodies to viral antigens) can also be used as pharmaceutically acceptable carriers. These can be prepared according to methods known to those skilled in the art, for example, as described in U.S. Patent No. 4,522,811.

20 It is advantageous to formulate oral or parenteral compositions in dosage unit form for ease of administration and uniformity of dosage. Dosage unit form as used herein refers to physically discrete units suited as unitary dosages for the subject to be treated; each unit containing a predetermined quantity of active compound calculated to produce the desired therapeutic effect in association with the required
25 pharmaceutical carrier.

 Toxicity and therapeutic efficacy of such compounds can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, *e.g.*, for determining the LD₅₀ (the dose lethal to 50% of the population) and the ED₅₀ (the dose therapeutically effective in 50% of the population). The dose ratio between toxic
30 and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD₅₀/ED₅₀. Compounds which exhibit high therapeutic indices are preferred. While compounds that exhibit toxic side effects may be used, care should be taken to design a delivery system that targets such compounds to the site of affected tissue in order to minimize potential damage to uninfected cells and, thereby, reduce side
35 effects.

5 The data obtained from the cell culture assays and animal studies can be used
in formulating a range of dosage for use in humans. The dosage of such compounds
lies preferably within a range of circulating concentrations that include the ED50 with
little or no toxicity. The dosage may vary within this range depending upon the
dosage form employed and the route of administration utilized. For any compound
10 used in the method of the invention, the therapeutically effective dose can be
estimated initially from cell culture assays. A dose may be formulated in animal
models to achieve a circulating plasma concentration range that includes the IC50
(i.e., the concentration of the test compound which achieves a half-maximal inhibition
of symptoms) as determined in cell culture. Such information can be used to more
15 accurately determine useful doses in humans. Levels in plasma may be measured, for
example, by high performance liquid chromatography.

As defined herein, a therapeutically effective amount of protein or polypeptide
(i.e., an effective dosage) ranges from about 0.001 to 30 mg/kg body weight,
preferably about 0.01 to 25 mg/kg body weight, more preferably about 0.1 to 20
20 mg/kg body weight, and even more preferably about 1 to 10 mg/kg, 2 to 9 mg/kg, 3 to
8 mg/kg, 4 to 7 mg/kg, or 5 to 6 mg/kg body weight. The protein or polypeptide can
be administered one time per week for between about 1 to 10 weeks, preferably
between 2 to 8 weeks, more preferably between about 3 to 7 weeks, and even more
preferably for about 4, 5, or 6 weeks. The skilled artisan will appreciate that certain
25 factors may influence the dosage and timing required to effectively treat a subject,
including but not limited to the severity of the disease or disorder, previous
treatments, the general health and/or age of the subject, and other diseases present.
Moreover, treatment of a subject with a therapeutically effective amount of a protein,
polypeptide, or antibody can include a single treatment or, preferably, can include a
30 series of treatments.

For antibodies, the preferred dosage is 0.1 mg/kg of body weight (generally 10
mg/kg to 20 mg/kg). If the antibody is to act in the brain, a dosage of 50 mg/kg to
100 mg/kg is usually appropriate. Generally, partially human antibodies and fully
human antibodies have a longer half-life within the human body than other antibodies.
35 Accordingly, lower dosages and less frequent administration is often possible.
Modifications such as lipidation can be used to stabilize antibodies and to enhance

5 uptake and tissue penetration (e.g., into the brain). A method for lipidation of antibodies is described by Cruikshank *et al.* ((1997) *J. Acquired Immune Deficiency Syndromes and Human Retrovirology* 14:193).

The present invention encompasses agents which modulate expression or activity. An agent may, for example, be a small molecule. For example, such small
10 molecules include, but are not limited to, peptides, peptidomimetics (e.g., peptoids), amino acids, amino acid analogs, polynucleotides, polynucleotide analogs, nucleotides, nucleotide analogs, organic or inorganic compounds (i.e., including heteroorganic and organometallic compounds) having a molecular weight less than about 10,000 grams per mole, organic or inorganic compounds having a molecular
15 weight less than about 5,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 1,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 500 grams per mole, and salts, esters, and other pharmaceutically acceptable forms of such compounds.

Exemplary doses include milligram or microgram amounts of the small
20 molecule per kilogram of subject or sample weight (e.g., about 1 microgram per kilogram to about 500 milligrams per kilogram, about 100 micrograms per kilogram to about 5 milligrams per kilogram, or about 1 microgram per kilogram to about 50 micrograms per kilogram. It is furthermore understood that appropriate doses of a small molecule depend upon the potency of the small molecule with respect to the
25 expression or activity to be modulated. When one or more of these small molecules is to be administered to an animal (e.g., a human) in order to modulate expression or activity of a polypeptide or nucleic acid of the invention, a physician, veterinarian, or researcher may, for example, prescribe a relatively low dose at first, subsequently increasing the dose until an appropriate response is obtained. In addition, it is
30 understood that the specific dose level for any particular animal subject will depend upon a variety of factors including the activity of the specific compound employed, the age, body weight, general health, gender, and diet of the subject, the time of administration, the route of administration, the rate of excretion, any drug combination, and the degree of expression or activity to be modulated.

35 An antibody (or fragment thereof) may be conjugated to a therapeutic moiety such as a cytotoxin, a therapeutic agent or a radioactive ion. A cytotoxin or cytotoxic

5 agent includes any agent that is detrimental to cells. Examples include taxol, cytochalasin B, gramicidin D, ethidium bromide, emetine, mitomycin, etoposide, tenoposide, vincristine, vinblastine, colchicin, doxorubicin, daunorubicin, dihydroxy anthracin dione, mitoxantrone, mithramycin, actinomycin D, 1-dehydrotestosterone, glucocorticoids, procaine, tetracaine, lidocaine, propranolol, puromycin, 10 maytansinoids, *e.g.*, maytansinol (see US Patent No. 5,208,020), CC-1065 (see US Patent Nos. 5,475,092, 5,585,499, 5,846,545) and analogs or homologs thereof. Therapeutic agents include, but are not limited to, antimetabolites (*e.g.*, methotrexate, 6-mercaptopurine, 6-thioguanine, cytarabine, 5-fluorouracil decarbazine), alkylating agents (*e.g.*, mechlorethamine, thioepa chlorambucil, CC-1065, melphalan, 15 carmustine (BSNU) and lomustine (CCNU), cyclophosphamide, busulfan, dibromomannitol, streptozotocin, mitomycin C, and cis-dichlorodiamine platinum (II) (DDP) cisplatin), anthracyclines (*e.g.*, daunorubicin (formerly daunomycin) and doxorubicin), antibiotics (*e.g.*, dactinomycin (formerly actinomycin), bleomycin, mithramycin, and anthramycin (AMC)), and anti-mitotic agents (*e.g.*, vincristine, 20 vinblastine, taxol and maytansinoids). Radioactive ions include, but are not limited to iodine, yttrium and praseodymium.

The conjugates of the invention can be used for modifying a given biological response, the drug moiety is not to be construed as limited to classical chemical therapeutic agents. For example, the drug moiety may be a protein or polypeptide 25 possessing a desired biological activity. Such proteins may include, for example, a toxin such as abrin, ricin A, pseudomonas exotoxin, or diphtheria toxin; a protein such as tumor necrosis factor, α -interferon, β -interferon, nerve growth factor, platelet derived growth factor, tissue plasminogen activator; or, biological response modifiers such as, for example, lymphokines, interleukin-1 ("IL-1"), interleukin-2 ("IL-2"), 30 interleukin-6 ("IL-6"), granulocyte macrophage colony stimulating factor ("GM-CSF"), granulocyte colony stimulating factor ("G-CSF"), or other growth factors. Alternatively, an antibody can be conjugated to a second antibody to form an antibody heteroconjugate as described by Segal in U.S. Patent No. 4,676,980.

The nucleic acid molecules of the invention can be inserted into vectors and 35 used as gene therapy vectors. Gene therapy vectors can be delivered to a subject by, for example, intravenous injection, local administration (see U.S. Patent 5,328,470) or

5 by stereotactic injection (see e.g., Chen *et al.* (1994) *Proc. Natl. Acad. Sci. USA*
91:3054-3057). The pharmaceutical preparation of the gene therapy vector can
include the gene therapy vector in an acceptable diluent, or can comprise a slow
release matrix in which the gene delivery vehicle is imbedded. Alternatively, where
the complete gene delivery vector can be produced intact from recombinant cells, e.g.,
10 retroviral vectors, the pharmaceutical preparation can include one or more cells which
produce the gene delivery system.

The pharmaceutical compositions can be included in a container, pack, or
dispenser together with instructions for administration.

Methods of Treatment

15 The present invention provides for both prophylactic and therapeutic methods
of treating a subject at risk of (or susceptible to) a disorder or having a disorder
associated with, e.g., insufficient, aberrant, or unwanted collagen XXII expression or
activity. As used herein, the term "treatment" is defined as the application or
administration of a therapeutic agent to a patient, or application or administration of a
20 therapeutic agent to an isolated tissue or cell line from a patient, who has a disease, a
symptom of disease or a predisposition toward a disease, with the purpose to cure,
heal, alleviate, relieve, alter, remedy, ameliorate, improve or affect the disease, the
symptoms of disease or the predisposition toward disease. A therapeutic agent
includes, but is not limited to, small molecules, peptides, antibodies, ribozymes and
25 antisense oligonucleotides.

With regards to both prophylactic and therapeutic methods of treatment, such
treatments may be specifically tailored or modified, based on knowledge obtained
from the field of pharmacogenomics. "Pharmacogenomics", as used herein, refers to
the application of genomics technologies such as gene sequencing, statistical genetics,
30 and gene expression analysis to drugs in clinical development and on the market.
More specifically, the term refers the study of how a patient's genes determine his or
her response to a drug (e.g., a patient's "drug response phenotype", or "drug response
genotype".) Thus, another aspect of the invention provides methods for tailoring an
individual's prophylactic or therapeutic treatment with either the collagen XXII

5 molecules of the present invention or collagen XXII modulators according to that individual's drug response genotype. Pharmacogenomics allows a clinician or physician to target prophylactic or therapeutic treatments to patients who will most benefit from the treatment and to avoid treatment of patients who will experience toxic drug-related side effects.

10 In one aspect, the invention provides a method for preventing in a subject, a disease or condition associated with an aberrant or unwanted collagen XXII expression or activity, by administering to the subject a collagen XXII or an agent which modulates collagen XXII expression or at least one collagen XXII activity. Subjects at risk for a disease which is caused or contributed to by aberrant or
15 unwanted collagen XXII expression or activity can be identified by, for example, any or a combination of diagnostic or prognostic assays as described herein. Administration of a prophylactic agent can occur prior to the manifestation of symptoms characteristic of the collagen XXII aberrance, such that a disease or disorder is prevented or, alternatively, delayed in its progression. Depending on the
20 type of collagen XXII aberrance, for example, a collagen XXII, collagen XXII agonist or collagen XXII antagonist agent can be used for treating the subject. The appropriate agent can be determined based on screening assays described herein.

It is possible that some collagen XXII disorders can be caused, at least in part, by an abnormal level of gene product, or by the presence of a gene product exhibiting
25 abnormal activity. As such, the reduction in the level and/or activity of such gene products would bring about the amelioration of disorder symptoms.. In addition, the collagen XXII molecules described herein can serve to increase the success of transplants, e.g., transplants of connective tissue, e.g., cartilage, tendon, lens, heart valves, or hair follicles.

30 As discussed, successful treatment of collagen XXII disorders can be brought about by techniques that serve to inhibit the expression or activity of target gene products. For example, compounds, e.g., an agent identified using an assays described above, that proves to exhibit negative modulatory activity, can be used in accordance with the invention to prevent and/or ameliorate symptoms of collagen
35 XXII disorders. Such molecules can include, but are not limited to peptides, phosphopeptides, small organic or inorganic molecules, or antibodies (including, for

5 example, polyclonal, monoclonal, humanized, anti-idiotypic, chimeric or single chain antibodies, and Fab, F(ab')₂ and Fab expression library fragments, scFV molecules, and epitope-binding fragments thereof).

Further, antisense and ribozyme molecules that inhibit expression of the target gene can also be used in accordance with the invention to reduce the level of target
10 gene expression, thus effectively reducing the level of target gene activity. Still further, triple helix molecules can be utilized in reducing the level of target gene activity. Antisense, ribozyme and triple helix molecules are discussed above.

It is possible that the use of antisense, ribozyme, and/or triple helix molecules to reduce or inhibit mutant gene expression can also reduce or inhibit the transcription
15 (triple helix) and/or translation (antisense, ribozyme) of mRNA produced by normal target gene alleles, such that the concentration of normal target gene product present can be lower than is necessary for a normal phenotype. In such cases, nucleic acid molecules that encode and express target gene polypeptides exhibiting normal target gene activity can be introduced into cells via gene therapy method. Alternatively, in
20 instances in that the target gene encodes an extracellular protein, it can be preferable to co-administer normal target gene protein into the cell or tissue in order to maintain the requisite level of cellular or tissue target gene activity.

Another method by which nucleic acid molecules may be utilized in treating or preventing a disease characterized by collagen XXII expression is through the use
25 of aptamer molecules specific for collagen XXII protein. Aptamers are nucleic acid molecules having a tertiary structure which permits them to specifically bind to protein ligands (see, e.g., Osborne, *et al.* (1997) *Curr. Opin. Chem Biol.* 1: 5-9; and Patel, D.J. (1997) *Curr Opin Chem Biol* 1:32-46). Since nucleic acid molecules may in many cases be more conveniently introduced into target cells than therapeutic
30 protein molecules may be, aptamers offer a method by which collagen XXII protein activity may be specifically decreased without the introduction of drugs or other molecules which may have pluripotent effects.

Antibodies can be generated that are both specific for target gene product and that reduce target gene product activity. Such antibodies may, therefore, by
35 administered in instances whereby negative modulatory techniques are appropriate for

5 the treatment of collagen XXII disorders. For a description of antibodies, see the Antibody section above.

In circumstances wherein injection of an animal or a human subject with a collagen XXII protein or epitope for stimulating antibody production is harmful to the subject, it is possible to generate an immune response against collagen XXII through
10 the use of anti-idiotypic antibodies (see, for example, Herlyn, D. (1999) *Ann Med* 31:66-78; and Bhattacharya-Chatterjee, M., and Foon, K.A. (1998) *Cancer Treat Res.* 94:51-68). If an anti-idiotypic antibody is introduced into a mammal or human subject, it should stimulate the production of anti-anti-idiotypic antibodies, which should be specific to the collagen XXII protein. Vaccines directed to a disease
15 characterized by collagen XXII expression may also be generated in this fashion.

In instances where the target antigen is intracellular and whole antibodies are used, internalizing antibodies may be preferred. Lipofectin or liposomes can be used to deliver the antibody or a fragment of the Fab region that binds to the target antigen into cells. Where fragments of the antibody are used, the smallest inhibitory fragment
20 that binds to the target antigen is preferred. For example, peptides having an amino acid sequence corresponding to the Fv region of the antibody can be used. Alternatively, single chain neutralizing antibodies that bind to intracellular target antigens can also be administered. Such single chain antibodies can be administered, for example, by expressing nucleotide sequences encoding single-chain antibodies
25 within the target cell population (see e.g., Marasco *et al.* (1993) *Proc. Natl. Acad. Sci. USA* 90:7889-7893).

The identified compounds that inhibit target gene expression, synthesis and/or activity can be administered to a patient at therapeutically effective doses to prevent, treat or ameliorate collagen XXII disorders. A therapeutically effective dose refers to
30 that amount of the compound sufficient to result in amelioration of symptoms of the disorders. Toxicity and therapeutic efficacy of such compounds can be determined by standard pharmaceutical procedures as described above.

The data obtained from the cell culture assays and animal studies can be used in formulating a range of dosage for use in humans. The dosage of such compounds
35 lies preferably within a range of circulating concentrations that include the ED₅₀ with little or no toxicity. The dosage can vary within this range depending upon the dosage

5 form employed and the route of administration utilized. For any compound used in the method of the invention, the therapeutically effective dose can be estimated initially from cell culture assays. A dose can be formulated in animal models to achieve a circulating plasma concentration range that includes the IC_{50} (i.e., the concentration of the test compound that achieves a half-maximal inhibition of

10 symptoms) as determined in cell culture. Such information can be used to more accurately determine useful doses in humans. Levels in plasma can be measured, for example, by high performance liquid chromatography.

Another example of determination of effective dose for an individual is the ability to directly assay levels of “free” and “bound” compound in the serum of the test subject.

15 Such assays may utilize antibody mimics and/or “biosensors” that have been created through molecular imprinting techniques. The compound which is able to modulate collagen XXII activity is used as a template, or “imprinting molecule”, to spatially organize polymerizable monomers prior to their polymerization with catalytic reagents. The subsequent removal of the imprinted molecule leaves a polymer matrix

20 which contains a repeated “negative image” of the compound and is able to selectively rebind the molecule under biological assay conditions. A detailed review of this technique can be seen in Ansell, R. J. *et al* (1996) *Current Opinion in Biotechnology* 7:89-94 and in Shea, K.J. (1994) *Trends in Polymer Science* 2:166-173. Such “imprinted” affinity matrixes are amenable to ligand-binding assays,

25 whereby the immobilized monoclonal antibody component is replaced by an appropriately imprinted matrix. An example of the use of such matrixes in this way can be seen in Vlatakis, G. *et al* (1993) *Nature* 361:645-647. Through the use of isotope-labeling, the “free” concentration of compound which modulates the expression or activity of collagen XXII can be readily monitored and used in

30 calculations of IC_{50} .

Such “imprinted” affinity matrixes can also be designed to include fluorescent groups whose photon-emitting properties measurably change upon local and selective binding of target compound. These changes can be readily assayed in real time using appropriate fiberoptic devices, in turn allowing the dose in a test subject to be quickly

35 optimized based on its individual IC_{50} . An rudimentary example of such a “biosensor” is discussed in Kriz, D. *et al* (1995) *Analytical Chemistry* 67:2142-2144.

5 Another aspect of the invention pertains to methods of modulating collagen XXII expression or activity for therapeutic purposes, e.g., for use in transplants of, e.g., connective tissue, e.g., cartilage, tendon, lens, heart valves, or hair follicles. Accordingly, in an exemplary embodiment, the modulatory method of the invention involves contacting a cell with a collagen XXII or agent that modulates one or more
10 of the activities of collagen XXII protein activity associated with the cell. An agent that modulates collagen XXII protein activity can be an agent as described herein, such as a nucleic acid or a protein, a naturally-occurring target molecule of a collagen XXII protein (e.g., a collagen XXII substrate or receptor), a collagen XXII antibody, a collagen XXII agonist or antagonist, a peptidomimetic of a collagen XXII agonist or
15 antagonist, or other small molecule.

 In one embodiment, the agent stimulates one or more collagen XXII activities. Examples of such stimulatory agents include active collagen XXII protein and a nucleic acid molecule encoding collagen XXII. In another embodiment, the agent inhibits one or more collagen XXII activities. Examples of such inhibitory
20 agents include antisense collagen XXII nucleic acid molecules, anti-collagen XXII antibodies, and collagen XXII inhibitors. These modulatory methods can be performed *in vitro* (e.g., by culturing the cell with the agent) or, alternatively, *in vivo* (e.g., by administering the agent to a subject). As such, the present invention provides methods of treating an individual afflicted with a disease or disorder characterized by
25 aberrant or unwanted expression or activity of a collagen XXII protein or nucleic acid molecule. In one embodiment, the method involves administering an agent (e.g., an agent identified by a screening assay described herein), or combination of agents that modulates (e.g., up regulates or down regulates) collagen XXII expression or activity. In another embodiment, the method involves administering a collagen XXII protein or
30 nucleic acid molecule as therapy to compensate for reduced, aberrant, or unwanted collagen XXII expression or activity.

 Stimulation of collagen XXII activity is desirable in situations in which collagen XXII is abnormally downregulated and/or in which increased collagen XXII activity is likely to have a beneficial effect. For example, stimulation of collagen
35 XXII activity is desirable in situations in which a collagen XXII is downregulated and/or in which increased collagen XXII activity is likely to have a beneficial effect.

- 5 Likewise, inhibition of collagen XXII activity is desirable in situations in which collagen XXII is abnormally upregulated and/or in which decreased collagen XXII activity is likely to have a beneficial effect.

Pharmacogenomics

- 10 The collagen XXII molecules of the present invention, as well as agents, or modulators which have a stimulatory or inhibitory effect on collagen XXII activity (e.g., collagen XXII gene expression) as identified by a screening assay described herein can be administered to individuals to treat (prophylactically or therapeutically) collagen XXII associated, e.g., connective tissue disorders, associated with aberrant or
15 unwanted collagen XXII activity. In conjunction with such treatment, pharmacogenomics (i.e., the study of the relationship between an individual's genotype and that individual's response to a foreign compound or drug) may be considered. Differences in metabolism of therapeutics can lead to severe toxicity or therapeutic failure by altering the relation between dose and blood concentration of
20 the pharmacologically active drug. Thus, a physician or clinician may consider applying knowledge obtained in relevant pharmacogenomics studies in determining whether to administer a collagen XXII molecule or collagen XXII modulator as well as tailoring the dosage and/or therapeutic regimen of treatment with a collagen XXII molecule or collagen XXII modulator.

- 25 Pharmacogenomics deals with clinically significant hereditary variations in the response to drugs due to altered drug disposition and abnormal action in affected persons. See, for example, Eichelbaum, M. *et al.* (1996) *Clin. Exp. Pharmacol. Physiol.* 23:983-985 and Linder, M.W. *et al.* (1997) *Clin. Chem.* 43:254-266. In general, two types of pharmacogenetic conditions can be differentiated. Genetic
30 conditions transmitted as a single factor altering the way drugs act on the body (altered drug action) or genetic conditions transmitted as single factors altering the way the body acts on drugs (altered drug metabolism). These pharmacogenetic conditions can occur either as rare genetic defects or as naturally-occurring polymorphisms. For example, glucose-6-phosphate dehydrogenase deficiency

5 (G6PD) is a common inherited enzymopathy in which the main clinical complication is haemolysis after ingestion of oxidant drugs (anti-malarials, sulfonamides, analgesics, nitrofurans) and consumption of fava beans.

One pharmacogenomics approach to identifying genes that predict drug response, known as "a genome-wide association", relies primarily on a high-
10 resolution map of the human genome consisting of already known gene-related markers (e.g., a "bi-allelic" gene marker map which consists of 60,000-100,000 polymorphic or variable sites on the human genome, each of which has two variants.) Such a high-resolution genetic map can be compared to a map of the genome of each of a statistically significant number of patients taking part in a Phase II/III drug trial to
15 identify markers associated with a particular observed drug response or side effect. Alternatively, such a high resolution map can be generated from a combination of some ten-million known single nucleotide polymorphisms (SNPs) in the human genome. As used herein, a "SNP" is a common alteration that occurs in a single nucleotide base in a stretch of DNA. For example, a SNP may occur once per every
20 1000 bases of DNA. A SNP may be involved in a disease process, however, the vast majority may not be disease-associated. Given a genetic map based on the occurrence of such SNPs, individuals can be grouped into genetic categories depending on a particular pattern of SNPs in their individual genome. In such a manner, treatment regimens can be tailored to groups of genetically similar individuals, taking into
25 account traits that may be common among such genetically similar individuals.

Alternatively, a method termed the "candidate gene approach," can be utilized to identify genes that predict drug response. According to this method, if a gene that encodes a drug's target is known (e.g., a collagen XXII protein of the present invention), all common variants of that gene can be fairly easily identified in the
30 population and it can be determined if having one version of the gene versus another is associated with a particular drug response.

Alternatively, a method termed the "gene expression profiling," can be utilized to identify genes that predict drug response. For example, the gene expression of an animal dosed with a drug (e.g., a collagen XXII molecule or collagen XXII modulator
35 of the present invention) can give an indication whether gene pathways related to toxicity have been turned on.

5 Information generated from more than one of the above pharmacogenomics approaches can be used to determine appropriate dosage and treatment regimens for prophylactic or therapeutic treatment of an individual. This knowledge, when applied to dosing or drug selection, can avoid adverse reactions or therapeutic failure and thus enhance therapeutic or prophylactic efficiency when treating a subject with a collagen
10 XXII molecule or collagen XXII modulator, such as a modulator identified by one of the exemplary screening assays described herein.

 The present invention further provides methods for identifying new agents, or combinations, that are based on identifying agents that modulate the activity of one or more of the gene products encoded by one or more of the collagen XXII genes of the
15 present invention, wherein these products may be associated with resistance of the cells to a therapeutic agent. Specifically, the activity of the proteins encoded by the collagen XXII genes of the present invention can be used as a basis for identifying agents for overcoming agent resistance. By blocking the activity of one or more of the resistance proteins, target cells, e.g., human cells, will become sensitive to
20 treatment with an agent that the unmodified target cells were resistant to.

 Monitoring the influence of agents (e.g., drugs) on the expression or activity of a collagen XXII protein can be applied in clinical trials. For example, the effectiveness of an agent determined by a screening assay as described herein to increase collagen XXII gene expression, protein levels, or upregulate collagen XXII
25 activity, can be monitored in clinical trials of subjects exhibiting decreased collagen XXII gene expression, protein levels, or downregulated collagen XXII activity. Alternatively, the effectiveness of an agent determined by a screening assay to decrease collagen XXII gene expression, protein levels, or downregulate collagen XXII activity, can be monitored in clinical trials of subjects exhibiting increased
30 collagen XXII gene expression, protein levels, or upregulated collagen XXII activity. In such clinical trials, the expression or activity of a collagen XXII gene, and preferably, other genes that have been implicated in, for example, a collagen XXII-associated disorder can be used as a "read out" or markers of the phenotype of a particular cell.

5 Collagen XXII Informatics

 The sequence of a collagen XXII molecule is provided in a variety of media to facilitate use thereof. A sequence can be provided as a manufacture, other than an isolated nucleic acid or amino acid molecule, which contains a collagen XXII. Such a manufacture can provide a nucleotide or amino acid sequence, e.g., an open reading
10 frame, in a form which allows examination of the manufacture using means not directly applicable to examining the nucleotide or amino acid sequences, or a subset thereof, as they exists in nature or in purified form. The sequence information can include, but is not limited to, collagen XXII full-length nucleotide and/or amino acid sequences, partial nucleotide and/or amino acid sequences, polymorphic sequences
15 including single nucleotide polymorphisms (SNPs), epitope sequence, and the like. In a preferred embodiment, the manufacture is a machine-readable medium, e.g., a magnetic, optical, chemical or mechanical information storage device.

 As used herein, "machine-readable media" refers to any medium that can be read and accessed directly by a machine, e.g., a digital computer or analogue
20 computer. Non-limiting examples of a computer include a desktop PC, laptop, mainframe, server (e.g., a web server, network server, or server farm), handheld digital assistant, pager, mobile telephone, and the like. The computer can be stand-alone or connected to a communications network, e.g., a local area network (such as a VPN or intranet), a wide area network (e.g., an Extranet or the Internet), or a
25 telephone network (e.g., a wireless, DSL, or ISDN network). Machine-readable media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM, ROM, EPROM, EEPROM, flash memory, and the like; and hybrids of these categories such as magnetic/optical
30 storage media.

 A variety of data storage structures are available to a skilled artisan for creating a machine-readable medium having recorded thereon a nucleotide or amino acid sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition,
35 a variety of data processor programs and formats can be used to store the nucleotide

5 sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and Microsoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. The skilled artisan can readily adapt any number of
10 data processor structuring formats (e.g., text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

In a preferred embodiment, the sequence information is stored in a relational database (such as Sybase or Oracle). The database can have a first table for storing
15 sequence (nucleic acid and/or amino acid sequence) information. The sequence information can be stored in one field (e.g., a first column) of a table row and an identifier for the sequence can be store in another field (e.g., a second column) of the table row. The database can have a second table, e.g., storing annotations. The second table can have a field for the sequence identifier, a field for a descriptor or
20 annotation text (e.g., the descriptor can refer to a functionality of the sequence, a field for the initial position in the sequence to which the annotation refers, and a field for the ultimate position in the sequence to which the annotation refers. Non-limiting examples for annotation to nucleic acid sequences include polymorphisms (e.g., SNP's) translational regulatory sites and splice junctions. Non-limiting examples for
25 annotations to amino acid sequence include polypeptide domains, e.g., a domain described herein; active sites and other functional amino acids; and modification sites.

By providing the nucleotide or amino acid sequences of the invention in computer readable form, the skilled artisan can routinely access the sequence information for a variety of purposes. For example, one skilled in the art can use the
30 nucleotide or amino acid sequences of the invention in computer readable form to compare a target sequence or target structural motif with the sequence information stored within the data storage means. A search is used to identify fragments or regions of the sequences of the invention which match a particular target sequence or target motif. The search can be a BLAST search or other routine sequence
35 comparison, e.g., a search described herein.

5 Thus, in one aspect, the invention features a method of analyzing collagen XXII, e.g., analyzing structure, function, or relatedness to one or more other nucleic acid or amino acid sequences. The method includes: providing a collagen XXII nucleic acid or amino acid sequence; comparing the collagen XXII sequence with a
10 collection of sequences, e.g., a nucleic acid or protein sequence database to thereby analyze collagen XXII. The method can be performed in a machine, e.g., a computer, or manually by a skilled artisan.

 The method can include evaluating the sequence identity between a collagen XXII sequence and a database sequence. The method can be performed by accessing
15 the database at a second site, e.g., over the Internet.

 As used herein, a "target sequence" can be any DNA or amino acid sequence of six or more nucleotides or two or more amino acids. A skilled artisan can readily recognize that the longer a target sequence is, the less likely a target sequence will be present as a random occurrence in the database. Typical sequence lengths of a target
20 sequence are from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues. However, it is well recognized that commercially important fragments, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

 Computer software is publicly available which allows a skilled artisan to
25 access sequence information provided in a computer readable medium for analysis and comparison to other sequences. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are and can be used in the computer-based systems of the present invention. Examples of such software include, but are not limited to, MacPattern (EMBL),
30 BLASTN and BLASTX (NCBI).

 Thus, the invention features a method of making a computer readable record of a sequence of a collagen XXII sequence which includes recording the sequence on a computer readable matrix. In a preferred embodiment the record includes one or more of the following: identification of an ORF; identification of a domain, region, or
35 site; identification of the start of transcription; identification of the transcription

5 terminator; the full length amino acid sequence of the protein, or a mature form thereof; the 5' end of the translated region.

In another aspect, the invention features, a method of analyzing a sequence. The method includes: providing a collagen XXII sequence, or record, in machine-readable form; comparing a second sequence to the collagen XXII sequence; thereby
10 analyzing a sequence. Comparison can include comparing to sequences for sequence identity or determining if one sequence is included within the other, e.g., determining if the collagen XXII sequence includes a sequence being compared. In a preferred embodiment the collagen XXII or second sequence is stored on a first computer, e.g., at a first site and the comparison is performed, read, or recorded on a second
15 computer, e.g., at a second site. E.g., the collagen XXII or second sequence can be stored in a public or proprietary database in one computer, and the results of the comparison performed, read, or recorded on a second computer. In a preferred embodiment the record includes one or more of the following: identification of an ORF; identification of a domain, region, or site; identification of the start of
20 transcription; identification of the transcription terminator; the full length amino acid sequence of the protein, or a mature form thereof; the 5' end of the translated region.

In another aspect, the invention provides a machine-readable medium for holding instructions for performing a method for determining whether a subject has a collagen XXII-associated disease or disorder or a pre-disposition to a collagen XXII-associated disease or disorder, wherein the method comprises the steps of determining
25 collagen XXII sequence information associated with the subject and based on the collagen XXII sequence information, determining whether the subject has a collagen XXII-associated disease or disorder or a pre-disposition to a collagen XXII-associated disease or disorder and/or recommending a particular treatment for the disease,
30 disorder or pre-disease condition.

The invention further provides in an electronic system and/or in a network, a method for determining whether a subject has a collagen XXII-associated disease or disorder or a pre-disposition to a disease associated with a collagen XXII wherein the method comprises the steps of determining collagen XXII sequence information
35 associated with the subject, and based on the collagen XXII sequence information, determining whether the subject has a collagen XXII-associated disease or disorder or

5 a pre-disposition to a collagen XXII-associated disease or disorder, and/or recommending a particular treatment for the disease, disorder or pre-disease condition. In a preferred embodiment, the method further includes the step of receiving information, e.g., phenotypic or genotypic information, associated with the subject and/or acquiring from a network phenotypic information associated with the
10 subject. The information can be stored in a database, e.g., a relational database. In another embodiment, the method further includes accessing the database, e.g., for records relating to other subjects, comparing the collagen XXII sequence of the subject to the collagen XXII sequences in the database to thereby determine whether the subject as a collagen XXII-associated disease or disorder, or a pre-disposition for
15 such.

The present invention also provides in a network, a method for determining whether a subject has a collagen XXII associated disease or disorder or a pre-disposition to a collagen XXII-associated disease or disorder associated with collagen XXII, said method comprising the steps of receiving collagen XXII sequence
20 information from the subject and/or information related thereto, receiving phenotypic information associated with the subject, acquiring information from the network corresponding to collagen XXII and/or corresponding to a collagen XXII-associated disease or disorder (e.g., a connective tissue disorder, and based on one or more of the phenotypic information, the collagen XXII information (e.g., sequence information
25 and/or information related thereto), and the acquired information, determining whether the subject has a collagen XXII-associated disease or disorder or a pre-disposition to a collagen XXII-associated disease or disorder. The method may further comprise the step of recommending a particular treatment for the disease, disorder or pre-disease condition.

30 The present invention also provides a method for determining whether a subject has a collagen XXII -associated disease or disorder or a pre-disposition to a collagen XXII-associated disease or disorder, said method comprising the steps of receiving information related to collagen XXII (e.g., sequence information and/or information related thereto), receiving phenotypic information associated with the
35 subject, acquiring information from the network related to collagen XXII and/or related to a collagen XXII-associated disease or disorder, and based on one or more of

5 the phenotypic information, the collagen XXII information, and the acquired information, determining whether the subject has a collagen XXII-associated disease or disorder or a pre-disposition to a collagen XXII-associated disease or disorder. The method may further comprise the step of recommending a particular treatment for the disease, disorder or pre-disease condition.

10 This invention is further illustrated by the following examples that should not be construed as limiting. The contents of all references, patents and published patent applications cited throughout this application are incorporated herein by reference.

EXAMPLES

Example 1: Identification and Characterization of Human collagen XXII cDNA

15 The human collagen XXII sequence is shown in Fig. 1; SEQ ID NO:1. The nucleic acid sequence includes an initiation codon (ATG) that is double underlined. The region between and inclusive of the initiation codon and the termination codon is a methionine-initiated coding sequence of about 4878 nucleotides, including the termination codon. The coding sequence encodes a 1626 amino acid protein (SEQ ID
20 NO:2), which is shown as SEQ ID NO:2 (Figure 2).

Two splice variants of collagen XXII have been identified. One variant lacks the nucleotides from nucleotide 3458-3518 of SEQ ID NO:1, and the corresponding protein thereby lacks amino acid residues from amino acid 1005-1024 of SEQ ID
25 NO:2. The collagen XXII protein migrates as a 160 kD polypeptide on SDS PAGE.

Example 2: Anti-collagen XXII antibody R34

The 5' non-triple helical domain was expressed in 293 cells and was purified using an incorporated His tag. The recombinant domain was injected into a rabbit to produce antibodies. The rabbit serum was affinity purified on the recombinant
30 fragment and the antibody product is referred to as R34. R34 recognizes the recombinant fragment from XXII but not a similar product from a related collagen, XXIV. R34 does not recognize the His tag. R34 recognizes a band of the appropriate mobility on SDS-PAGE of an extract of muscle

5 Full-length recombinant collagen XXII was made in 293 cells and show the expected electrophoretic mobility after disulfide bond reduction (above). Without reduction, the product shows mobilities of dimers and trimers, as expected for a product that associates into a trimeric triple helix. The recombinant molecule is collagenase sensitive as expected.

10 Example 3: Tissue Distribution of collagen XXII by Immunofluorescence Analysis

Immunofluorescent localization of collagen XXII shows deposition in the following locations: myotendinous junctions; in joints, e.g., immediately beneath the articular cartilage of joints; in the basement membrane, e.g., of the outer root sheath at the base of the hair follicles; in the ciliary body of the eye, e.g., where the muscle of
15 the lens intersects the zonular fibers; in the heart, e.g., where the heart muscles intersect the cartilage ring in the region of the valves.

Equivalents

Those skilled in the art will recognize, or be able to ascertain using no more
20 than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims.

What is claimed is:

1. An isolated nucleic acid molecule comprising a nucleotide sequence at least 80% identical to the nucleotide sequence of SEQ ID NO:1 or SEQ ID NO:3 or an alternatively spliced variant thereof.

5 2. The nucleic acid of claim 1, wherein the nucleotide sequence at least 85% identical to the nucleotide sequence of SEQ ID NO:1 or SEQ ID NO:3 or an alternatively spliced variant thereof.

3. The nucleic acid of claim 1, wherein the nucleotide sequence at least 90% identical to the nucleotide sequence of SEQ ID NO:1 or SEQ ID NO:3 or an alternatively
10 spliced variant thereof.

4. The nucleic acid of claim 1, wherein the nucleotide sequence at least 95% identical to the nucleotide sequence of SEQ ID NO:1 or SEQ ID NO:3 or an alternatively spliced variant thereof.

5. The nucleic acid of claim 1, wherein the nucleotide sequence at least 98%
15 identical to the nucleotide sequence of SEQ ID NO:1 or SEQ ID NO:3 or an alternatively spliced variant thereof.

6. The nucleic acid of claim 1, wherein the nucleotide sequence at least 99% identical to the nucleotide sequence of SEQ ID NO:1 or SEQ ID NO:3 or an alternatively spliced variant thereof.

20 7. The nucleic acid of claim 1, wherein the alternatively spliced variant lacks nucleotides 3458-3518 of SEQ ID NO:1.

8. An isolated nucleic acid comprising the nucleotide sequence of SEQ ID NO:1 or SEQ ID NO:3.

9. An isolated nucleic acid consisting of the sequence of SEQ ID NO:1 or SEQ
25 ID NO:3, or its complement.

10. An isolated nucleic acid that encodes a polypeptide having an amino acid sequence which is at least 80% identical to the amino acid sequence of SEQ ID NO:2 or a functional fragment thereof.

11. The nucleic acid of claim 10, wherein the polypeptide has an amino acid sequence at least 85% identical to the amino acid sequence of SEQ ID NO:2 or a functional fragment thereof.

12. The nucleic acid of claim 10, wherein the polypeptide has an amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:2 or a functional fragment thereof.

13. The nucleic acid of claim 10, wherein the polypeptide has an amino acid sequence at least 95% identical to the amino acid sequence of SEQ ID NO:2 or a functional fragment thereof.

14. The nucleic acid of claim 10, wherein the polypeptide has an amino acid sequence at least 98% identical to the amino acid sequence of SEQ ID NO:2 or a functional fragment thereof.

15. The nucleic acid of claim 10, wherein the polypeptide or functional fragment thereof has one or more of the following activities:

- (a) has the ability to form a triple helix;
- (b) has the ability to form higher order collagen fibrils or fibers;
- (c) provides tensile strength or flexibility to muscle-connective tissue boundaries;
- (d) is recognized by R34 antibody; and
- (e) is collagenase sensitive.

16. The nucleic acid of claim 10, wherein the polypeptide or functional fragment thereof comprises one or more of: residues 1-27 of SEQ ID NO:2; residues 1-1478 of SEQ ID NO:2; residues 1479-1603 of SEQ ID NO:2; and residues 1604-1626 of SEQ ID NO:2.

17. An isolated nucleic acid comprising a nucleotide sequence that hybridizes under highly stringent conditions to a nucleotide sequence consisting of SEQ ID NO:1 or 3, or its complement.

5 18. The nucleic acid of claim 17, wherein the nucleic acid encodes a polypeptide comprising one or more of the following activities:

- (a) has the ability to form a triple helix;
- (b) has the ability to form higher order collagen fibrils or fibers;
- (c) provides tensile strength or flexibility to muscle-connective tissue boundaries;
- 10 (d) is recognized by R34 antibody; and
- (e) is collagenase sensitive.

19. An isolated nucleic acid comprising a nucleotide sequence that hybridizes under highly stringent conditions to a nucleotide sequence that encodes a polypeptide having the amino acid sequence of SEQ ID NO:2, or its complement.

15 20. The nucleic acid of claim 19, wherein the nucleic acid encodes a polypeptide comprising one or more of the following activities:

- (a) has the ability to form a triple helix;
- (b) has the ability to form higher order collagen fibrils or fibers;
- 20 (c) provides tensile strength or flexibility to muscle-connective tissue boundaries;
- (d) is recognized by R34 antibody; and
- (e) is collagenase sensitive.

21. An oligonucleotide probe at least 90% identical to a nucleotide sequence consisting of between 25 and 200 contiguous nucleotides of SEQ ID NO:1 or SEQ ID NO:3.

25 22. The probe of claim 21, wherein the probe is at least 50 nucleotides in length.

23. The probe of claim 21, wherein the probe is at least 100 nucleotides in length.

24. The probe of claim 21, wherein the probe is 100 % identical to said nucleotide sequence.

25. The probe of claim 22, wherein the probe is 100 % identical to said nucleotide sequence.

26. The probe of claim 23, wherein the probe is 100 % identical to said nucleotide sequence.

5 27. An isolated probe or primer that hybridizes under highly stringent conditions to the sense or antisense nucleic acid sequence of SEQ ID NO:1 or SEQ ID NO:3, wherein the isolated probe or primer is at least 50 nucleotides in length.

28. The nucleic acid molecule of any one of claims 1, 8, 10 or 17 further comprising a nucleic acid sequence encoding a heterologous amino acid sequence .

10 29. A vector comprising the nucleic acid molecule of any one of claims 1, 8, 10 or 17.

30. A host cell harboring the vector of claim 29.

31. The host cell of claim 30, wherein the host cell is a mammalian cell.

15 32. A method for detecting the presence of a nucleic acid molecule in a sample, the method comprising the steps of:

- a) contacting the sample with the probe of claim 21; and
- b) determining whether the probe binds to a nucleic acid molecule in the sample.

33. The method of claim 32, wherein the sample comprises mRNA molecules.

20 34. A method for detecting the presence of a nucleic acid molecule in a sample, the method comprising the steps of:

- a) contacting the sample with the probe of claim 27; and
- b) determining whether the probe binds to a nucleic acid molecule in the sample.

25 35. The method of claim 34, wherein the sample comprises mRNA molecules.

collagen XXII cDNA

GGGGCTGGCCCCGAGGCTGCGGCGTTCTCCCCAAGGAAGTGTCTTCTCCGCTTTCCCTGTTCTTCT
GTTTCTCACACACTTTTCTATCTCATTCTGTAACTTTCAAGCCTTTCCTTCTAACTGTATGTATTT
AGTTACTTGTTTTCAAGCTGGTTCCCTCTCTGGCCCCCTGGCCTGGGGAAAGCCTCCACACTTAC
TGCGGGTCTTGTTTAGAGTCTGAGTTTGTGAGATTATTTGGGGCAGAGTGGGCGAGTGGCTGACA
GGTGACCCCCAGGAGGAGGATTCCCTGGGGCTGGTGTCTTCTCCAGCTGCTGCTTCCAGTGGGCC
TGGGCCCAGGACTGGACCTCCGCTGGCACCCCTGAGTGCCTCCCTGCCAGGCCATGCTGCTGTAG
ACCCTAACAGCGTCTCTTCCTGGCCAAGAGAAGCCTGTCCCCAAGAACAGGAGAGCCATGGCCGG
CCTCCGAGGGAAACGCTGTGGCTGGCCTCCTCTGGATGCTGCTGCTGTGGAGTGGGGGCGGCGGCT
GCCAGGCTCAGCGGGCAGGTTGCAAAGTGTCCACTACGATCTGGTCTTCCTCCTGGACACCTCC
TCCAGCGTGGGCAAGGAGGACTTTGAGAAGGTCCGGCAGTGGGTGGCCAACCTGGTGGACACCTT
CGAGGTGGGCCCCGACCGCACCCGTGTGGGGGTCTGTGCGCTACAGCGACCGGCCACCACGGCCT
TCGAGTTGGGACTCTTTGGCTCGCAGGAGGAGGTCAAGGCGGCTGCCCGGCGTCTCGCCTACCAC
GGGGGCAACACCAACACGGGAGACGCGCTCCGCTACATCACGGCCCGCAGCTTCTCCCCACACGC
CGGCGGCCCGCCCCAGGGACCGCGCCTACAAGCAGGTGGCCATCCTGCTCACCGACGGCCGCAGCC
AGGACCTGGTGCTGGACGCCGCGGCGGCAGCCCACCGCGCTGGCATCCGCATCTTTGCCGTGGGC
GTGGGCGAGGCACTCAAGGAGGAGCTGGAGGAGATCGCCTCAGAGCCCAAGTCCGCCCCACGTCTT
CCACGTGTCCGACTTCAATGCCATCGACAAGATCCGGGGCAAGCTGCGGCGCCGTCTTTGTGAAA
ATGTGCTCTGTCTAGCGTTTCGTGTAGAAGGAGATCGCTTTAAGCACACCAATGGAGGAACCAAG
GAAATCACAGGTTTTGACCTGATGGATTTGTTTCAGTGTGAAGGAAATCTTGGGGAAGAGAGAGAA
TGGAGCTCAGAGTTCCATGTACGGATGGGATCCTTCCCTGTGGTGCAAAGTACTGAGGATGTGT
TCCCCCAAGGTTTACCTGATGAGTACGCCTTTGTCAACCTTCCGGTTCAGGAAAACCTCTCGG
AAGGAAGACTGGTATATCTGGCAGGTCATCGACCAGTACGGCATCCACAGGTCTCCATCCGGCT
GGATGGTGAAAACAAGGCAGTCGAGTACAACGCTGTGGGTGCCATGAAAGATGCTGTCAGGGTGG
TCTTCCGAGGTTCTCGGGTCAATGACCTCTTTGACCGGGACTGGCACAAGATGGCCCTGAGCATC
CAGGCCCAGAACGTCTCCCTGCACATTGACTGTGCGCTGGTGCAGACACTACCCATCGAGGAACG
GGAGAACATTGACATCCAGGGCAAGACTGTGATTGGCAAGCGCCTCTACGACAGTGTGCCCATTG
ACTTTGACCTACAGCGGATTGTGATCTATTGTGACTCGAGACACGCAGAATTGGAGACTTGTGT
GATATCCCCTCGGGTCCGTGCCAGGTGACCGTGGTGACAGAGCCTCCACCTCCACCCCCACCCCA
GCGGCCCTCCACCCAGGCAGTGAACAGATTGGGTTTTTGAAGACCATCAACTGCTCCTGCCCAG
CTGGAGAGAAGGGTGAAATGGGAGTTGCTGGCCCCATGGGGCTCCCTGGTCCAAAGGGAGACATA
GGAGCCATTGGGCCGGTTGGCGCTCCTGGACCTAAGGGAGAGAAAGGTGATGTGGGCATAGGACC
TTTTGGCCAAGGGGAAAAGGGTGAAAAGGGTTCCTGGGCCTGCCCGGCCCCCCTGGGAGAGACG
GCAGCAAAGGCATGAGAGGGGAGCCAGGAGAGCTGGGAGAGCCGGGGCTGCCGGGTGAGGTCTGGC

FIGURE 1A

ATGCGGGGGCCCCAAGGACCACCTGGACTCCCCGGACCTCCTGGACGTGTCGGAGCTCCTGGTCT
CCAAGGAGAACGAGGTGAAAAGGGAACTCGAGGAGAAAAGGGAGAGCGAGGCCTGGATGGATTCC
CTGGGAAGCCTGGGGACACAGGACAGCAGGGCAGGCCCGGCCCTTCTGGTGTGGCAGGACCCAG
GGAGAAAAGGGTGACGTGGGACCTGCGGGGCCACCTGGTGTACCAGGCTCAGTGGTGCAGCAAGA
GGGCTTGAAAAGGGGAACAGGGAGCTCCAGGACCCAGAGGTACCAAGGCGCCCCCGGTCTCCAG
GAGCTCGGGGTCCAATAGGCCCAGAAGGCAGGGATGGACCTCCTGGTTTGCAAGGTCTCCGAGGG
AAGAAAGGTGACATGGGACCACCTGGAATCCCTGGATTGCTGGGGCTGCAGGGCCCTCCAGGACC
CCCTGGTGTCCCAGGCCCCCCCTGGACCGGGAGGTTCTCCGGGTTTGCTGGAGAGATCGGCTTCC
CGGGAAAGCCTGGACCTCCTGGGCCCACGGGACCCCTGGAAAGGACGGGGCCAAATGGACCACCA
GGTCCGCCAGGAACCAAGGGAGAACCAGGAGAAAGAGGGGAAGATGGTCTGCCTGGAAAACCAGG
CCTTCGGGGAGAAATTGGGGAGCAGGGCCTGGCAGGCCGACCTGGAGAGAAGGGAGAAGCAGGCC
TCCCAGGGGCTCCAGGCTTCCCAGGTGTGAGAGGAGAGAAAGGAGACCAGGGAGAAAAAGGTGAA
CTGGGACTTCCAGGACTGAAAGGTGACCGAGGTGAAAAGGGTGAAGCTGGTCTGCAGGCCCTCC
CGGGTTACCTGGAACATACATCCCTGTTACACCCACATCCACGGATGCCCCGAGAACAAGGGCCCA
AAGGAGAGAAGGGCGATCCAGGCCTGCCTGGGGAACCGGGACTGCAGGGCCGTCCTGGAGAATTG
GGGCCTCAGGGACCCACTGGACCACCGGGTGCCAAGGGACAGGAAGGTGCACATGGGGCTCCTGG
AGCAGCTGGAAACCCCGGTGCTCCCGGACATGTGGTGCCCCCGGTCCCAGTGGCCCTCCAGGAA
GTGTGGGTGCTCCCGGCCTCAGAGGCACCCAGGGAAAGATGGGGAGCGTGCTGAGAAGGGTGCA
GCGGGGAAGAAGGCAGCCCAGGGCCAGTTGGTCCCAGGGGAGATCCTGGTGCTCCTGGGCTCCC
TGGGCCGCCCCGAAAAGGGAAGGATGGAGAGCCGGGACTCCGTGGATCACCTGGACTCCCTGGAC
CCCTAGGAACCA~~aggctgcttgccgaaaaagtcagaggggtcagaaaaactgtgcactgggagggcaa~~
~~tgtgtta~~AGGGGGATCGAGGAGCTCCTGGGATCCCTGGTTCTCCTGGCAGCCGTGGTGACCCAGG
CATTGGGGTTGCTGGCCCTCCTGGCCCTTCCGGACCACCAGGAGACAAAGGATCCCCGGGATCAC
GAGGCTTACCTGGATTCCCTGGCCCCCAGGGCCCAGCCGGCCGGGACGGTGCACCAGGAAATCCA
GGAGAAAGAGGGCCTCCTGGCAAGCCGGGCCTCTCTTCACTACTGTCTCCAGGGGACATAAATCT
CTTGGCTAAGGATGTGTGCAATGACTGCCCTCCTGGCCCCCAGGCCTCCCTGGTCTACCAGGTT
TTAAAGGGGACAAAGGTGTCCCAGGAAAGCCAGGGAGAGAAGGCACAGAAGGGAAAAAGGGAGAG
GCTGGGCCTCCAGGCCTACCAGGGCCCCCAGGAATAGCTGGACCACAGGGAAGTCAAGGAGAACG
TGGTGCAGATGGTGAGGTTGGGCAGAAAGGTGATCAGGGTCATCCTGGAGTTCCAGGTTTCATGG
GGCCCCCAGGGAACCCCGGGCCACCAGGGGCAGATGGAATTGCAGGAGCTGCTGGACCACCAGGA
ATCCAAGGGTCACCTGGGAAAGAAGGCCCTCCTGGCCCCCAAGGCCATCTGGATTACCCGGAAT
CCCAGGAGAAGAAGGCAAAGAGGGCAGAGATGGAAAGCCGGGTCCCCCTGGAGAGCCGGGCAAAG
CAGGAGAGCCAGGTCTACCAGGACCAGAGGGTGCCCGAGGCCCACCTGGCTTCAAGGGACACACA
GGCGATTCTGGTGCACCCGGTCCCCGGGGAGAGTCTGGTGCCATGGGGCTTCTGGTGCAGGAAGG
GTTACCAGGAAAAGATGGTGACACTGGACCCACTGGGCCACAGGGTCCCCAAGGACCAAGGGGCC

FIGURE 1B

CACCGGGCAAGAATGGATCACCGGGATCTCCAGGAGAGCCTGGCCCTTCAGGAACCCCTGGCCAG
AAAGGAAGCAAAGGGGAAAATGGCAGCCCAGGACTTCCTGGCTTCCTGGGTCCCCGTGGGCCTCC
GGGAGAACCAGGAGAGAAAGGAGTCCCAGGCAAGGAGGGGGTCCCTGGGAAGCCTGGAGAGCCTG
GATTCAAAGGAGAAAGGGGAGATCCTGGGATCAAAGGTGACAAAGGACCTCCTGGTGAAAAGGC
CAGCCTGGGGACCCTGGAATCCCAGGCCACAAAGGCCACACAGGCCTGATGGGTCCCCAAGGACT
ACCTGGGGAGAATGGACCAGTTGGACCCCCAGGGCCTCCAGGCCAGCCGGGATTTCCAGGACTGA
GGGGGGAGTCTCCATCCATGGAAACCCTGCGTCGGCTTATTCAAGAAGAGCTGGGGAAGCAGCTT
GAAACCAGACTCGCCTACCTCCTGGCCCAGATGCCCCCGGCGTACATGAAGTCATCTCAAGGCAG
ACCTGGGCCCCCAGGGCCCCCTGGAAAAGATGGGCTTCAGGCCGGGCGGCCCCATGGGGGAGC
CAGGTCGTCCTGGGCAGGGGGGTCTGGAAGGACCCTCTGGACCCATAGGTCCCAAAGGTGAGCGA
GGAGCCAAAGGTGACCCAGGTGCACCTGGAGTTGGCCTCCGAGGCGAGATGGGACCCCTGGAAT
CCCAGGTCAACCCGGGGAACCTGGCTATGCTAAAGATGGACTTCCTGGGATCCCTGGCCCTCAAG
GGGAGACAGGACCAGCTGGACATCCTGGCCTCCCAGGACCTCCCGGTCCCCCAGGCCAATGTGAC
CCTTCCCAGTGTGCCTACTTCGCCAGCCTTGCTGCCCCGGCGGGTAATGTGAAGGGTCCCTAAAG
GACTCTGGAAAGCCAGAAGACTGCAGTGGATTTCTGAAACTTGAAGTCAAGAGCCCAGTGGGAAGC
CAGAGGTCTTGAAAGACTTCAGCCATGTGTTCTTTTTTTTTTTCTTTCTTTTATCGTTTGCTTTT
TGTTTTATTTTTCTTGAGAGACCTCAAAATTATTAAATCCAACAGACGCTGCCGGTCGGTCAGATT
ATTATTAATATTATTGTTGTTGTTAATTATTATTATTATTTTCATATGCTGATGCTTTGTGAGTTC
TTTTCCACTCCTTTAAAGTTGGGAAAACCTTGATTCGTGGGGCAGGAGATTGTTTCTTCATTCTTC
TGACAGCCCCCATCTGACGCGTAACTGCCATTTTAAGGAAACTCTTGGTGCTACAAAACCCTGA
CCAGACACTTGGCAAATTTACCTCTTTCTTCAAAAAGAAAACTTTAAGAAAATGAGCCAATGGGC
TTCATTCTCAGTCATGCCCGGAGATCACCCAGGAGAAATAATACAAACACCACCACTGTCCAGAG
AGAGTAAAGAAGCAGAAAGAGAAAGAATTTGCAACCATGAGGAATGTTCCACCTCCCGACGGGA
CGTGCAATTTGGAAAACACAGAATCAGCCCTCAGGGTGCACCTCAGCCACCTCAGTGCTCTAAGCT
CACAGAAGTGAAATAATGTCTGTGGGTGGCAATGGCTTTGTGGGATCATATGTCTTGGCCAAAG
ATGGGAAAACCTATGTTGAAGAGGCAGCCCTTGAGTGTTAATTTGTCTTCTAAACTGTGTAAGGC
CCCTTCAAGTTCCTCTTGTTGGTTTCAATTATATTAATTATAAAACAAGTGGATGTGGTGACCAT
CCACTTGTTGTTTCCCTAATGATGGGCAGTTGGCCAGGGCACTGACCAGAGCTGGGAAATTTGTAT
CTCCAAGGCGGCTCTGTCTCTGAAATAAATGGCATCAAGTGCATGTGTGTATGCGACATGCCCTG
CCTGAACAGGTGCTCAATAAATCCAAGTTTCCTTCTCTTGAAAAAA

FIGURE 1C

Collagen 22 amino acid sequence

MAGLRGNAVAGLLWMLLLWSGGGGCQAQRAGCKSVHYDLVFLLDTSSSVGKEDFEKVRQW 60
VANLVDTFEVGPDRTRVGVVRYSDRPTTAFELGLFGSQEEVKAAARRLAYHGGNTNTGDA 120
LRYITARFSFPHAGGRPRDRAYKQVAILLTDGRSQDLVLDAAAAAHRAGIRIFAVGVGEA 180
LKEELEEEIASPKSAHV FHVSDFN AIDKIRGKLRRRLCENVLCPSVRVEGDRFKHTNGGT 240
KEITGFDLMDLFSVKEILGKRENGAQSSYVRMGSPFPVQSTEDVFPQGLPDEYAFVTTFR 300
FRKTSRKEDWYIWQVIDQYGIPQVSIRLDGENKAVEYNAV GAMKDAVRVVFRGSRVNDLF 360
DRDWHKMALSIQAQNVSLHIDCALVQTLPIEERENIDIQKTVIGKRLYDSVPIDFDLQR 420
IVIYCDSRHAELETCCDIPSGPCQVTVVTEPPPPPPPPQRPPTPGSEQIGFLLKINCSCPA 480
GEKGEMGVAGPMGLPGPKGDTGAI GPVGA PGPKGEKGDVGIGPFGOGEEKGEKGSGLGLPGP 540
PGRDGSKGMRGEPGELGEPGLPGEVGMRGPOGPPGLPGPPGRVGAPGLOGERGEKGTGE 600
KGERGLDGFPGKPGDTGOQGRPGPSGVAGPOGEKGDVGAPPGVPVGSVVOQEGLKGEQG 660
APGPRGHOGAPGPPGARGPIGPEGRDGPPGLQGLRGKKGDMGPPGIPGILLGLQGPPGPPG 720
VPGPPGPGGSPGLPGEIGFPGKPGPPGPTGPPGKDGPNGPPGPPGPGTKGEPGERGEDGLPG 780
KPLGRGEIGEGLAGRPGEKGEAGLPGAPGFPVGRGEKGDGEGKELGLPGLKGDRGEKG 840
EAGPAGPPGLPGTTSLFTPHPRMPGEOGPKGEKGDGGLPGEPGLQGRPGELGPOGPTGPP 900
GAKGOEGAHGAPGAAGNPGAPGHVGAPGPSGPPGSGVAPGLRGTPGKDGERGEKGAAGEE 960
GSPGPVGPGRDGPAGPLPGPPGKGDGEPLRGSPGLPGPLGTTAAACGKVRGSENCALGG 1020
QCVKGDRGAPGIPGSPGSRGDPGIGVAGPPGPGSPGPGDKGSPGSRGLPGFPGPOGPAGRD 1080
GAPGNPGERGPPGKPLSSLLSPGDINLLAKDVCNDCPPGPPGLPGLPGFKGDKGVPGKP 1140
GREGTEGKKGEAGPPGLPGPPGIAGPOGSQGERCADGEVQOKGDQGHGVPVPGFMGPPGNP 1200
GPPGADGIAGAAGPPGIQGS PGKEGPPGPQGPSGLPGIPGEEGKEGRDGKPGPPGEPGKA 1260
GEPGLPGPEGARGPPGFKGHTGDSGAPGPRGESGAMGLPGQEGLPKDGDTGPTGPOGPQ 1320
GPRGPPGKNGSPGSPGEPGSGTPGQKGSKGENGSPGLPGFLGPRGPPGEPGEKGVPGKE 1380
GVPGKPGEPGFKGERGDPGKGDGPPGKGOPGDPGIPGHKGHTGLMGPOGLPGENGVPV 1440
GPPGPPGQPGEPGLRGESPSMETLRLRIOELGKOLETRLAYLLAOMPPAYMKSSOGRPG 1500
PPGPPGKDGLPGRAGPMGEPRPGOGGLEGPGPIGPKGERGAKGDPGAPGVGLRGEMGP 1560
PGIPGQPGEPGYAKDGLPGIPGPOGETGPAGHPGLPGPPGPPGQCDPSQCAYFASLAARP 1620
GNVKGP*

FIGURE 2