

Recombinant Human Collagen and Biomimetic Variants Using a De Novo Gene Optimized for Modular Assembly

Sam Wei Polly Chan,[†] She-Pin Hung,[‡] Senthil Kumar Raman,[†] G. Wesley Hatfield,[§]
Richard H. Lathrop,^{§,||} Nancy A. Da Silva,^{*,†} and Szu-Wen Wang^{*,†}

Department of Chemical Engineering and Materials Science, The Computational Biology Research Laboratory, Institute for Genomics and Bioinformatics, and Department of Computer Science, University of California, Irvine, California 92697, and Verdezyne, Inc., Carlsbad, California 92008

Received January 16, 2010; Revised Manuscript Received April 9, 2010

A collagen-mimetic polymer that can be easily engineered with specific cell-responsive and mechanical properties would be of significant interest for fundamental cell-matrix studies and applications in regenerative medicine. However, oligonucleotide-based synthesis of full-length collagen has been encumbered by the characteristic glycine-X-Y sequence repetition, which promotes mismatched oligonucleotide hybridizations during de novo gene assembly. In this work, we report a novel, modular synthesis strategy that yields full-length human collagen III and specifically defined variants. We used a computational algorithm that applies codon degeneracy to design oligonucleotides that favor correct hybridizations while disrupting incorrect ones for gene synthesis. The resulting recombinant polymers were expressed in *Saccharomyces cerevisiae* engineered with prolyl-4-hydroxylase. Our modular approach enabled mixing-and-matching domains to fabricate different combinations of collagen variants that contained different secretion signals at the N-terminus and cysteine residues imbedded within the triple-helical domain at precisely defined locations. This work shows the flexibility of our strategy for designing and assembling specifically tailored biomimetic collagen polymers with re-engineered properties.

Introduction

Cells respond to specific biochemical and mechanical signals that exist within their extracellular matrix (ECM), and these signals regulate important cellular behaviors such as adhesion, proliferation, differentiation, and apoptosis.¹ The most abundant family of proteins in this matrix is collagen. Collagen dynamically controls the life cycle and function of cells through natural binding sites and, in turn, cells remodel the structure of the collagenous matrix.^{2,3} Dynamic cell–matrix associations in tissue development occur in the differentiation of cells such as osteoblasts⁴ and cardiac myocytes.^{5,6} During wound-healing, collagen can direct integrins to initiate a cascade which produces matrix metalloproteinases, thereby degrading the collagen matrix and enabling keratinocyte migration.⁷ The structure of collagen also imparts mechanical strength to tissues of the bone, tendon, and cartilage, while enabling the binding of other ECM proteins and soluble factors.^{8,9} Defects in collagen's structure or stability have been implicated in the development of various diseases, such as rheumatoid arthritis, varicose veins, osteogenesis imperfecta, and Ehlers-Danlos syndrome.^{8,10,11}

Because the ECM can impart biological effects on cells, its components and its mimics are often used as scaffolds in tissue engineering applications. Control over cellular functions can result from combining diverse properties including cell-interaction sites,^{12,13} their concentrations^{14,15} and spatial geometries,^{16–18} mechanical properties of the substrate material,^{19–21} and material degradation kinetics.^{22,23} Polymers that mimic the ECM, such as alginate,²⁴ polyethylene glycol,^{22,23} and polyacrylamide,²⁵

are often used. However, they are not natural matrices and can exhibit problems such as the need for cationic cross-linkers that interfere with cell adhesion, lack of biodegradation, or even degradation into toxic subunits. The ability to create polymers from natural building blocks that closely mimic matrix materials in structure and function, but with flexibility in properties beyond their natural scope, would be an attractive alternative in these applications.

Collagen is the most prevalent class of proteins in the ECM, and a collagen-based polymer material that can be easily modified and engineered for specific purposes would be of great interest. It would allow mixing-and-matching of specific functional sequences from the different families of collagens at exact locations, combinations, and densities, while keeping the overall size of the final polymeric product comparable to natural lengths. This capability would enable fundamental studies to elucidate the biochemical and biophysical principles that govern how cells and their microenvironments interact to yield growth and differentiation. Furthermore, such a platform would also be useful in tissue regeneration applications to generate tailored, artificial matrices which can be site-specifically modified for altering biological and physicochemical properties.

Our goal is to create a flexible platform that can be used to fabricate diverse biomimetic polymers based on native full-length collagen. Because the length of fibrillar human collagens is relatively large, typically near 1000 amino acids,¹⁰ successful synthesis strategies favor recombinant rather than chemical approaches. In a recombinant system, this means that changes are made directly in the DNA and should implement a bottom-up gene synthesis strategy for sequence flexibility.

However, the fabrication of genes that encode full-length, hydroxylated collagen-based polymers has been a significant challenge. The main difficulty is the prevalence of a trimeric repeating sequence with many GC-rich glycines (codons GGN,

* To whom correspondence should be addressed. Fax: 949-824-2541.

E-mail: wangsw@uci.edu (S.-W.W.); ndasilva@uci.edu (N.A.D.).

[†] Department of Chemical Engineering and Materials Science.

[‡] Verdezyne, Inc.

[§] Institute for Genomics and Bioinformatics.

^{||} Department of Computer Science.

where N is one of A, C, G, or T) and prolines (codons CCN). Collagen's characteristic repeating amino acid sequence of (glycine-X-Y)_N, where X is often proline and Y is often hydroxyproline, is repeated over 300 times in one strand of fibrillar collagen. Our analysis has shown that the native collagen III sequence has numerous amino acid repeats that are longer than expected for a typical protein of similar size, leading to numerous long direct nucleotide repeats in the DNA. The direct sequence repeats found in native collagen type III are shown in Figures S-1 (Supporting Information; amino acid repeats) and S-2 (Supporting Information; nucleotide repeats). These extensive GC-rich repeats can cause mismatched oligonucleotide hybridizations during gene synthesis. As a result, such an approach presents challenges in collagen gene synthesis.

The conventional method for producing synthetic genes is to assemble them from short overlapping oligonucleotide fragments in a PCR-based strategy.²⁶ However, the formation of incorrect hybridizations during synthesis yields erroneous final sequences in a population of different gene lengths and sequences.²⁷ These mismatched hybridizations are typically observed even in genes without repeating sequences.²⁷ The extensive repeats found in collagen create many possible incorrect hybridization sites. Hybridization melting temperature calculations by the method of Larsen et al.²⁷ indicate that the distributions of correct and incorrect hybridization melting temperatures substantially overlap when using the native DNA sequence of human collagen III. Its melting temperature gap is $\Delta T_m = 11.1$ °C, where $\Delta T_m = [\text{highest incorrect hybridization } T_m] - [\text{lowest correct hybridization } T_m]$. Therefore, attempts to create a collagen gene using standard PCR-assembly of oligonucleotides can result in a great many mismatched hybridizations, inhibiting gene assembly of the native DNA sequence.

Recombinant strategies for generating non-native collagen-like biopolymers have included the production of truncated regions within collagen,²⁸ tandem ligation of partial collagen regions,^{29,30} and collagen-like polymers that are one-fifth the size of the native length.^{31,32} Synthetic chemical methods for producing non-native sequences include generating short synthetic collagen peptides typically 24–63 amino acids in length.^{33–35} To generate longer polymers and because cysteine is uniquely absent from the triple-helical domains of fibrillar collagen, these short synthetic peptides can be linked through terminal cysteines and self-assembled into fibrillar-like structures.^{36,37} However, all of these strategies limit the location, identity, frequency, and combination of biologically relevant sites that can be presented in the polymer, while the shorter lengths affect the scope of nanoscale architecture and mechanical properties.

In this work, we report a novel synthesis strategy which yields full-length human collagen III protein and specifically defined variants. To remove mismatched hybridizations in gene synthesis, a computationally optimized DNA assembly (CODA) algorithm developed by Larsen et al.^{27,38} was used to design oligonucleotides and their overlaps which favor correct hybridizations and disrupt incorrect ones. This algorithm uses the degeneracy in the genetic code to design oligonucleotide sequences in which the minimum melting temperature for each correct hybridization is substantially higher than the maximum melting temperature for incorrect ones,^{27,38} that is, which have a negative ΔT_m as defined above. The algorithm also favors known organism preferences for codons and codon-pair usage. Although the resulting DNA sequence of the baseline human collagen III gene is non-native, the amino acid sequence of the final protein remains unchanged.

We have constructed a library of modules for human collagen III that has been optimized for yeast expression, is synthesized bottom-up entirely from oligonucleotides, and enables a truly modular assembly of full-length synthetic collagen. This modular library can be used to build collagen gene variants expressed in our recombinant *Saccharomyces cerevisiae* system, which carries the human prolyl-4-hydroxylase α and β genes. These enzymes are required for hydroxylating proline, which is necessary for imparting stability to the collagen polymer.³⁹ Our work demonstrates the full potential of using a modular, designed, codon-optimized approach for managing synthesis difficulties in recombinant biopolymers. We also show that full-length collagen and variants of collagen can be recombinantly fabricated using a bottom-up assembly strategy. The modular feature imparted to this synthetic gene allows tremendous flexibility for introducing site-specific, non-native sequences and unique chemical modification sites that will expand studies and applications of ECM-mimetic materials.

Experimental Methods

Modular Collagen (MCol) Gene Design and Assembly. Oligonucleotide sequences for gene synthesis were designed using the amino acid sequence of human procollagen type III (NCBI accession number NM_000090). The CODA algorithm²⁷ generated a list of overlapping oligonucleotides, each between 42 to 63 nucleotides (nts) long, which was optimized for gene self-assembly and expression in *S. cerevisiae*. Oligonucleotides were assembled into primary fragments (PF), primary fragments were assembled into secondary fragments (SF), and secondary fragments were assembled into full-length genes (FL). This optimized collagen III gene was designed to be modular and was designated "MCol". Sequences of individual oligonucleotides and the final MCol gene are given in Supporting Information. The distributions of correct and incorrect hybridization melting temperatures for MCol resulted in a melting temperature gap of $\Delta T_m = -7.7$ °C by the method of Larsen et al.,²⁷ which is an improvement of $\Delta\Delta T_m = -7.7$ °C – 11.1 °C = –18.8 °C on the melting temperature gap of $\Delta T_m = 11.1$ °C observed above when using the native DNA sequence. Figure S-3 (Supporting Information) gives the direct DNA repeats found in MCol, which are considerably fewer than for the native DNA sequence shown in Figure S-2. Both the melting temperature gap and the reduced direct DNA repeats of MCol indicate substantially improved gene assembly characteristics relative to the native DNA sequence.

The full-length MCol gene comprises three domains (front, middle, and back), as illustrated in Figure 1. Details of the modules are listed in Table 1. Two modular primary DNA fragments (FPF1, FPF2) were used for the front domain assembly, 12 modular primary DNA fragments (MPF1 to MPF12) were used for the middle domain assembly, and four modular primary DNA fragments (BPF1 to BPF4) were used for the back domain assembly. Overlapping sequences between primary fragments within the front, middle, and back regions were approximately 70, 90, and 70 nts, respectively.

Assembly of Primary Fragments (PF). Each modular DNA fragment was designed as a set of 10 or 12 oligonucleotides with overlapping sequences of 20–30 nts and named the "Primary Fragments" in Figure 1. The oligonucleotide sequences and final modular gene sequence are presented in Table S-1 and Figure S-4 (Supporting Information). All oligonucleotides were purchased from Integrated DNA Technologies (Corvall, IA) and assembled by polymerase extension. For primary fragments of the front and back domains, the oligonucleotides for each module were mixed together with final concentrations of 240 nM for the first and last oligonucleotides in each fragment and 40 nM for all other oligonucleotides. For primary fragments of the middle domain, the oligonucleotides in each module set were mixed together to 200 nM for the first and last oligonucleotides for each fragment and 20 nM for the other oligonucleotides. The oligonucleotides were primer-extended into a primary fragment using 2.5 U of PfuUltra II DNA

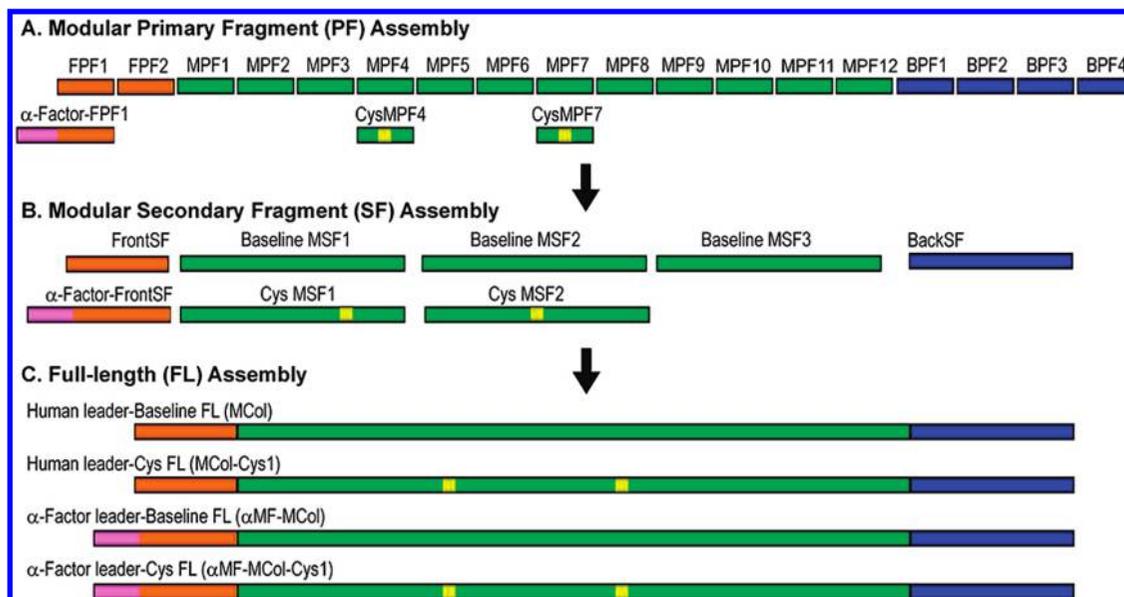


Figure 1. Modular strategy for fabricating native human collagen III and collagen-based variants. (A) Primary gene modules encoding the native amino acid sequence (or non-native modifications to sequence) are formed from PCR assembly of oligonucleotides. (B) Primary modules are PCR-assembled to form secondary fragments. (C) Secondary fragments are PCR-assembled to form the FL genes of native collagen and its variants. Green = native amino acid sequence of triple-helical region; yellow = sequence with non-native cysteine; pink = α -mating factor leader sequence; orange = N-propeptide region; blue = C-propeptide region.

Table 1. Names and Numbers of Oligonucleotides for Each Primary and Secondary Fragment Assembly^a

Collagen Gene Domain	Modular Primary Fragments		Modular Secondary Fragments
	Name	# of oligos	
Front (N-terminal propeptide)	FPP1 (α factor-FPF1)	12	FrontSF (α factor-FrontSF)
	FPE2	12	
Middle (Triple-helical region)	MPF1	12	MSF1 (Cys MSF1)
	MPF2	10	
	MPF3	10	
	MPF4 (Cys MPF4)	10	
	MPF5	10	
	MPF6	10	
	MPF7 (Cys MPF7)	10	MSF2 (Cys MSF2)
	MPF8	10	
	MPF9	10	
	MPF10	10	
	MPF11	10	
	MPF12	12	
Back (C-terminal propeptide)	BPF1	10	BackSF
	BPF2	10	
	BPF3	10	
	BPF4	12	

^a Names of the fragments in parentheses denote non-native variants of the human collagen III protein sequence. Sequences of these oligonucleotides are given in Supporting Information.

polymerase (Stratagene), 200 μ M dNTPs, and 1 \times PfuUltra reaction buffer in a final volume of 50 μ L. These primer extension and amplification reactions were performed in a MJ Research PTC-225 thermal cycler using the following protocol: 10 min denaturation step at 95 $^{\circ}$ C, followed by 30 cycles of 20 s at 95 $^{\circ}$ C, 30 s at 58 $^{\circ}$ C, and 15 s at 72 $^{\circ}$ C, with a final step of 5 min at 72 $^{\circ}$ C.

Primary fragments encoding non-native sequences were also synthesized. To assemble the collagen gene containing the yeast α -mating factor secretion signal, α -factor-FPF1 was generated which replaces the human collagen signal sequence with the yeast signal. We mixed 2 nM of the oligonucleotides encoding FPF1, without the human leader sequence but containing the yeast α -mating factor leader, with 200 nM of the first and last oligonucleotide primers of that fragment. These primary fragments were assembled from oligonucleotides using the primer-extension and amplification protocol described above, except with an annealing temperature of 62 $^{\circ}$ C (instead of 58 $^{\circ}$ C).

Assembled PF fragments were purified by the Qiagen PCR Purification Kit (Valencia, CA). A total of 20 ng of each assembled PF fragment were ligated with the pCR BluntII-TOPO vector (Invitrogen, Carlsbad, CA) following manufacturer instructions. Colonies containing PF

inserted into the vector were selected from LB-Kan plates and sequenced. Clones containing the correct sequence were used in subsequent secondary fragment assembly.

To assemble collagen with two non-native cysteines at amino acid positions 452 and 748, two additional primary fragments, CysMPF4 and CysMPF7, were generated via site-directed mutagenesis of the MCol primary fragments. For fragment CysMPF4, which contains an Ala-to-Cys mutation at site 452, 300 nM oligonucleotides [Col3-M a452c F' (GAC CAA GGG GAG AAC GTG GAG AAT GTG GCA TAC CCG GAG TAC CCG GTG CTA AAG) and Col3-M a452c R' (CTT TAG CAC CGG GTA CTC CGG GTA TGC CAC ATT CTC CAC GTT CTC CCC TTG GTC)] were mixed with 50 μ g of the DNA template and MPF4 ligated into a pCR BluntII-TOPO vector (Invitrogen, CA). For CysMPF7, which has a Gly-to-Cys mutation at site 748, 300 nM oligonucleotides [Col3-M g748c F' (CAA AAG GTG ATA AGG GTG AAC CTG GTT GTC CTG GCG CAG ACG GTG TTC CCG G) and Col3-M g748c R' (CAA AAG GTG ATA AGG GTG AAC CTG GTT GTC CTG GCG CAG ACG GTG TTC CCG G)] were mixed with 50 μ g of the DNA template and MPF7 ligated into a pCR BluntII-TOPO vector. A PCR reaction with 2.5 U of PfuUltra II DNA polymerase, 200 μ M dNTPs, and 1 \times PfuUltra reaction buffer in a final volume of 50 μ L was performed using a 10 min denaturation step at 95 $^{\circ}$ C, followed by 16 cycles of 20 s at 95 $^{\circ}$ C, 30 s at 55 $^{\circ}$ C, and 5 min at 72 $^{\circ}$ C, with a final step of 5 min at 72 $^{\circ}$ C. The PCR solutions were incubated with 20 U of Dpn I (New England Biolabs, Ipswich, MA) at 37 $^{\circ}$ C for 2 h and transformed into *Escherichia coli* DH5 α cells plated on LB-Kan. Colonies with correct CysMPF4 and CysMPF7 modules were identified by sequencing.

The primary fragments of correct sequence were amplified by mixing 200 nM each of the first and last oligonucleotides for each PF with 50 ng of the respective pCR BluntII-TOPO vector (containing the PF insert), 2.5 U of PfuUltra II DNA polymerase, 200 μ M dNTPs, and 1 \times PfuUltra reaction buffer in a final volume of 50 μ L. The PCR reaction was run using 10 min at 95 $^{\circ}$ C, followed by 25 cycles of 20 s at 95 $^{\circ}$ C, 30 s at 62 $^{\circ}$ C, and 15 s at 72 $^{\circ}$ C, with a final step of 5 min at 72 $^{\circ}$ C. Purified PF DNA fragments (Qiagen PCR Purification Kit) were used for subsequent secondary fragment assembly.

Assembly of Secondary Fragments (SF). We assembled secondary fragments from the primary fragments. For the N-terminal (front) domain, the FPF1 primary fragments containing either the human leader

Table 2. Summary of Genes and Corresponding Gene Products Used in Generating Recombinant Collagen and Its Variants^a

plasmid	gene	protein expressed by gene
YE α COL3A1	COL3A1 (from ATCC)	Human collagen III with 303 amino acid EDS deletion
YE μ COL	Modular human collagen III (optimized gene)	Native human collagen III alpha 1 chain
YE α MCOL	Modular collagen III with yeast alpha mating factor leader sequence	Native human collagen III alpha 1 chain with yeast secretion signal
YE μ COLCys1	Modular collagen III with two non-native cysteines in the triple-helical region	Human collagen III with two non-native cysteines in the triple-helical region
YE δ	Yeast delta sequence in place of COL3A1 gene (Control plasmid)	Nonfunctional control peptide (20 amino acids)
pddUB-CODAp4Ha2	CODA-optimized human hydroxylase α subunit (Integration vector)	Native human prolyl hydroxylase, α subunit
pddUB-CODAp4Hb	CODA-optimized human hydroxylase β subunit (Integration vector)	Native human prolyl hydroxylase, β subunit

^a All genes were cloned under the yeast GAL1 promoter.

or the α -factor leader sequences were PCR-assembled with FPF2 to obtain FrontSF and α -factor-FrontSF, respectively. To generate the secondary fragments for the triple-helical (middle) domain, primary fragments were mixed as described in Table 1: MPF1, MPF2, MPF3, MPF4, and MPF5 were used to assemble baseline MSF1; MPF5, MPF6, MPF7, MPF8, and MPF9 were used to assemble baseline MSF2; MPF9, MPF10, MPF11, and MPF12 were used to assemble baseline MSF3; MPF1, MPF2, MPF3, CysMPF4, and MPF5 were used to assemble Cys MSF1; and MPF5, MPF6, CysMPF7, MPF8, and MPF9 were used to assemble Cys MSF2. To generate the C-terminal (back) domain, BPF1, BPF2, BPF3, and BPF4 were used to assemble BackSF.

To assemble these secondary fragments, 2 nM of each primary fragment was mixed together with 200 nM of the first and the last oligos for each secondary fragment. The primary fragments were primer extended into a secondary DNA fragment with 2.5U of PfuUltra II DNA polymerase, 200 μ M dNTPs, and 1 \times PfuUltra reaction buffer in a final volume of 50 μ L. Reactions were performed using 10 min at 95 $^{\circ}$ C, followed by 30 cycles of 20 s at 95 $^{\circ}$ C, 30 s at 68 $^{\circ}$ C, and 45 s at 72 $^{\circ}$ C, with a final step of 5 min at 72 $^{\circ}$ C. Each assembled SF DNA fragment was column purified and ligated into the pCR BluntII-TOPO vector using the procedure described for the primary fragments. Clones were sequenced and plasmids containing correct SF DNA sequences were used for subsequent FL gene assembly.

Assembly of Full-Length Genes (FL). The full-length collagen and collagen variant genes were assembled from the secondary fragments. Secondary fragments of correct sequence were amplified from plasmids and purified according to the protocol described above for the primary fragments, except that the extension time was increased to 45 s at 72 $^{\circ}$ C. Purified FrontSF, Baseline MSF1, Baseline MSF2, Baseline MSF3, and BackSF were mixed to assemble the Baseline FL collagen gene with the human leader sequence (MCol). The FrontSF, Cys MSF1, Cys MSF2, Baseline SF3, and BackSF were mixed to assemble the gene encoding the collagen variant gene containing the two non-native cysteines and the human leader sequence (MCol-Cys1). The α -factor-FrontSF, Baseline MSF1, Baseline MSF2, Baseline MSF3, and BackSF were mixed to assemble the Baseline FL collagen gene with the α -mating factor leader sequence (α MF-MCol). The α -factor-FrontSF, Cys MSF1, Cys MSF2, Baseline MSF3, and BackSF were mixed to assemble the cysteine variant gene with the α -mating factor leader (α MF-MCol-Cys1).

These FL genes were assembled by mixing 2 nM of each secondary fragment with 200 nM of the 5' and 3' PCR gene primers. The secondary fragments were primer extended into a FL gene using 2.5 U of Stratagene PfuUltra II DNA polymerase, 200 μ M dNTPs, and 1 \times PfuUltra reaction buffer in a final volume of 50 μ L. These reactions were performed in a thermal cycler using a 10 min denaturation step at 95 $^{\circ}$ C, followed by 30 cycles of 20 s at 95 $^{\circ}$ C, 30 s at 68 $^{\circ}$ C, and 1.5 min at 72 $^{\circ}$ C, with a final step of 5 min at 72 $^{\circ}$ C. Assembled FL DNA was column purified and ligated into the pCR BluntII-TOPO vector using the protocol described for primary fragments. Clones were sequenced and plasmids containing correct FL DNA sequences were used for subcloning into the yeast expression vector.

Strains and Vectors. Haploid *S. cerevisiae* strain INVSc1 (Invitrogen) has the genotype (*MATa his3 Δ 1 leu2 trp1-289 ura3-52*). *E. coli* strain XL1-Blue (Stratagene) was used for plasmid maintenance and amplification. Human prolyl-4-hydroxylase α -subunit cDNA (p4Ha2, BC035813), human PDI cDNA (p4Hb, BC029617), and human procollagen (III) cDNA (COL3A1, BC028178 with Ehlers-Danlos Syndrome [EDS]) were obtained from the American Type Culture Collection (ATCC, Manassas, VA) on pCMV-Sport6 holding vectors. The collagen from ATCC (COL3A1 with EDS) has 303 amino acids truncated from the normal collagen III alpha chain, starting at amino acid 843. Although there are two isoforms for human prolyl-4-hydroxylase α , no kinetic differences have been found between the two isoforms.⁴⁰ We chose type 2 (p4Ha2) for our system because it had already been successfully expressed to hydroxylate a human collagen III fragment in *S. cerevisiae*.²⁸ Plasmid YEplac195⁴¹ is a 2 μ -based vector with a *URA3* selection marker. Plasmid pRY121⁴² carries the *GAL1-10* promoter sequence. Integrating vector pd δ UB(BsrGI) is similar to pd δ UB⁴³ but has a unique BsrGI site within each of the two δ sequences.

Plasmid and Strain Construction. Restriction enzymes and DNA ligase were from New England Biolabs. KOD Hot Start DNA Polymerase was from Novagen. The *GAL1-10* bidirectional promoter sequence was obtained by PCR using plasmid pRY121 as template and primers GalPFBamHI and GalPREcoRI (Table S-2, Supporting Information), digested with BamHI and EcoRI, and cloned into pBlueScriptII SK(-) (Stratagene) to create pBS-GAL1p. pBS-GAL1p-MluI was constructed in the same manner using primers GALPF and GALPRMluI, and has a MluI restriction site replacing the EcoRI site.

A summary of plasmids, genes, and the corresponding collagen variant is presented in Table 2. A COL3A1 gene fragment flanked by EcoRV and BstEII restriction sites was obtained by PCR using pCMV-Sport6-Col3A1 as template and primers COL3A1FEcoRVPmeI and COL3A1RBstEII (Table S-2, Supporting Information). The fragment was inserted into pCMV-sport 6-COL3A1 using EcoRV and BstEII to remove unwanted restriction sites preceding the collagen gene. The COL3A1 gene was then excised using EcoRV and HindIII and inserted into pBS-Gal1p digested with the same enzymes. The GAL1p-COL3A1 cassette was excised using SpeI and XbaI and inserted into YEplac195 at an XbaI site. The *ADH2* terminator was obtained from plasmid pADH2-sfp⁴⁴ using XbaI and KpnI and inserted downstream of COL3A1 to obtain YE α COL3A1. The control plasmid YE δ was created by replacing the COL3A1 gene in YE α COL3A1 with a 346 bp yeast delta sequence flanked by PmeI/NotI sites from pd δ UB(BsrGI). Plasmids carrying the synthetic modular collagen gene with native and yeast signals (alpha mating factor) and the collagen variant were constructed by replacing the COL3A1 gene in YE α COL3A1 at the PmeI and NotI sites with the new genes, resulting in YE μ COL, YE α MCOL, and YE μ COLCys1. The final plasmid map for the collagen genes is shown in Figure S-5A (Supporting Information).

The CODAp4Ha and CODAp4Hb genes were sequentially integrated⁴⁵ into the δ sequences of yeast strain INVSc1. Each integration vector (pddUB-CODAp4Ha and pddUB-CODAp4Hb; see Figures S-5B

and S-5C and Supporting Information) was linearized at the two BsrGI sites and transformed into INVSc1. Integrants were selected on uracil-deficient SDC(A,T) plates. Counter-selection in SDC-5FOA medium was used to select for excision of the reusable *URA3* blaster cassette.⁴⁶ YEpCOL3A1, YEpMCOL, YEp α MCOL, or YEpMCys1 was then transformed into the p4Ha- and p4Hb-containing strain. All yeast transformations, including integrative transformations, were done by a modified lithium acetate method, as described.⁴⁷

Media and Cultivation. Selective SDC(A,T) and SGC(A,T) media contain casamino acids (5 g/L, Difco), yeast nitrogen base without amino acids (6.7 g/L, Difco), adenine sulfate (20 mg/L), tryptophan (20 mg/L), and glucose (20 g/L) or galactose (20 g/L), respectively. SDC-5FOA medium was supplemented with uracil (20 mg/L) and 5-fluoroorotic acid monohydrate (1 g/L, Zymo Research). For plates, Bacto-agar (20 g/L, Difco) was added. LB (Luria–Bertani) medium was used for *E. coli* cultivation⁴⁸ and ampicillin (100 mg/L) was added for selection of plasmid-bearing cells.

Cells were inoculated from -80°C stock cells, grown overnight in 16×125 mm culture tubes containing 5 mL SDC (A,T) medium, and used to inoculate (1% v/v) 2.8 L Fernbach flasks containing 1 L of SGC(A,T) induction medium. The yeast were cultivated at 250 rpm and 30°C (Series 25, New Brunswick), and OD₆₀₀ was monitored. After 18–22 h, cells were harvested by centrifugation (RC28S, Sorvall) and frozen at -80°C .

Protein Polymer Purification. Yeast cells were thawed and resuspended at 0.1 g wet cell/mL in Tris-buffered saline (TBS, 20 mM Tris, 500 mM NaCl, pH 7.5) with 1 mM phenylmethylsulfonyl fluoride (Pierce), 2 U/mL DNase (NEB), and 1 U/mL RNase (NEB). The culture was lysed in a French Press (Thermo Scientific) at 16000 psi for four cycles. Ethylenediaminetetraacetic acid (EDTA; 1 mM) was added to the lysate and the supernatant was collected by ultracentrifugation (Optima LE80K, Beckman) for 1 h 15 min at $100000 \times g$ and 4°C .

Hydrochloric acid (4 M) was added to the soluble fraction of the cell lysate to lower the pH to 2. Pepsin (20 mg/mL in 4 M HCl; Mallinckrodt) was added to a final concentration of 0.2 mg/mL and kept at 4°C overnight to digest the background protein and collagen propeptides. Insoluble protein from the pepsin digestion was removed by centrifugation for 1 h 15 min at $100000 \times g$ and 4°C . The pH of the soluble fraction was raised to 7.5 with NaOH to inactivate the protease.

Protein solutions were dialyzed in 40 mM sodium acetate (pH 4.5) and applied to a cation exchange column (Sephacrose XL) in an FPLC (AKTA System, GE Healthcare). Protein polymers were eluted using a linear NaCl gradient. Fractions containing the purified protein were pooled and concentrated using a stirred-cell concentrator with a 100 kDa MWCO Biomax polyethersulfone membrane (Millipore).

Protein Gel Analyses. An aliquot of the French-pressed cell lysate (200 μL) was centrifuged at $16000 \times g$ (Microfuge 18, Beckman Coulter). Denaturing gel electrophoresis (SDS-PAGE) sample buffer containing the reducing agent dithiothreitol (6 \times) was added to the supernatant and cell pellet, and samples were separated in a 7 or 12% SDS-PAGE gel.⁴⁹ Protein polymers were visualized by staining with Imperial Protein Stain (Pierce). For native, nondenaturing gel electrophoresis, purified collagen samples were mixed with blue native PAGE (BN-PAGE) sample buffer, resolved in a 4–16% polyacrylamide gradient gel (Invitrogen) using BN-PAGE running buffer, and stained according to published protocols.^{50,51}

For Western blot analysis of cell lysates, samples were denatured and run on an SDS-PAGE gel. The bands on this gel were electrotransferred to a nylon membrane (positively charged, Roche). The membrane was blocked with 0.5% nonfat milk in TBS and probed with an antibody raised against the C-terminal region of human collagen III (sc-8781, Santa Cruz Biotechnology, 1:1000 dilution). The membrane was then washed and incubated with a secondary detection antibody conjugated with alkaline phosphatase (sc-2771, Santa Cruz Biotechnology). Hybridization was detected with a chromogenic kit (Roche).

Proline Hydroxylation Analysis. Collagen samples were dialyzed with water and sent to the Biopolymer Laboratory core facility at UCLA for amino acid analysis. Samples were hydrolyzed in 6 M hydrochloric acid for 22 h at 110°C . Hydrolyzates were derivatized with a fluorescent amino-reactive probe. Amino acids were quantified by reverse phase (RP)-HPLC with a fluorescence detector. Percent hydroxylation was defined as (moles of hydroxylated proline)/(total moles of hydroxylated proline + unhydroxylated proline).

Results

Correct Assembly of Genes that Encode Native Human Collagen III and Rationally Designed Variants were Confirmed. In collagen, the problem of mismatched oligonucleotide hybridization in gene assembly is significant, due to the Gly-X-Y repeating sequence and the longer direct amino acid repeats of the triple-helical region (Supporting Information, Figure S-1). Based on melting temperature calculations by the method of Larsen et al.,²⁷ the melting temperature gap between the distributions of correct and incorrect hybridizations for oligonucleotides encoding the native sequence is $\Delta T_m = 11.1^{\circ}\text{C}$. This substantial overlap facilitates incorrect hybridizations and encumbers full-length assembly of the gene. To solve this problem, we assembled the collagen III gene using the CODA algorithm,²⁷ which applies degeneracy and frequency in yeast codons,⁵² favors known yeast preferences for codons and codon-pair usage,⁵³ and computes favorable melting temperatures for oligonucleotide hybridization by making synonymous codon substitutions while the amino acid sequence remains fixed.^{27,38,54}

We designed the synthetic human collagen III gene in a modular fashion to facilitate the introduction of non-native sites. When this approach is used, making changes at specific locations in the sequence should require only the redesign and resynthesis of the modules in which the changes are specified. Incorporation of the new modules into the gene is accomplished by PCR assembly together with the remaining modules, following the process described above for the initial gene synthesis. We have also introduced several unique restriction sites into the gene design, which facilitates traditional cloning strategies.

We divided the MCol gene encoding native type III human collagen into several fragments and domains, as described in the Experimental Methods (Table 1 and Figure 1): the N-propeptide (front), C-propeptide (back), and 12 modules in the (Gly-X-Y)_N triple-helical domain (middle). To demonstrate the feasibility of synthesizing collagen variants using this modular approach, we also designed and fabricated two modules which introduced cysteine mutations into the middle of the triple-helical region (in modules 4 and 7) and one module which incorporated the yeast α -mating factor leader sequence in the N-propeptide, all of which are non-native modifications to human collagen III.

As shown in Figure 2, we successfully assembled the 12 gene modules encoding the native collagen III triple-helical region, the N- and C-propeptide domains, and the three modules encoding non-native sequences within the protein sequence. Each of these modules comprises 10–12 oligonucleotides. The CODA-optimization of oligonucleotide sequence and PCR assembly of each of the module fragments yielded a distinct band of correct length and sequence. We note that mishybridizations of oligonucleotides were not observed; instead, the infrequent sequence errors were point mutations and single base-pair deletions, which are errors associated with oligonucleotide synthesis and PCR. Primary modules with correct sequences were used for subsequent gene assembly. By assembling the individual modules of a full-length gene together via PCR assembly, the collagen genes encoding the amino acid sequence

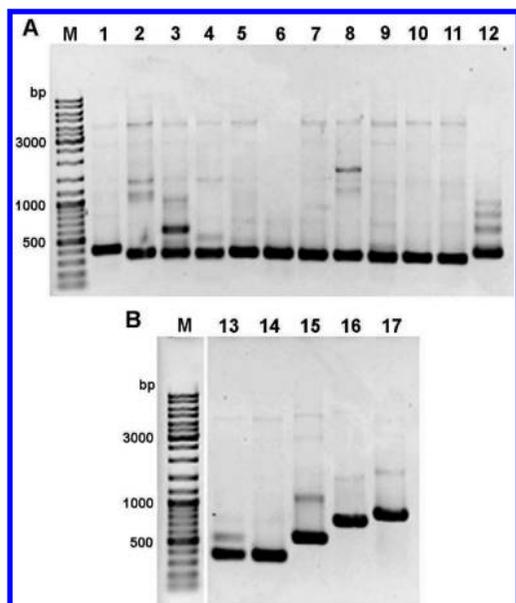


Figure 2. PCR assembly of computationally optimized gene modules for human collagen III. (A) Lanes 1–12 show the 12 individual gene modules comprising the (Gly-X-Y)_N triple-helical regions that encode for native human collagen III, as described in Figure 1. (B) Lanes 13–17 show the gene modules encoding the variant regions and the N- and C-termini: Lane 13, cysteine variant of module 4 (CysMPF4); Lane 14, cysteine variant of module 7 (CysMPF7); Lane 15, N-terminal propeptide with human leader sequence (FrontSF); Lane 16, N-terminal propeptide with yeast α -mating factor leader sequence (α -factor-FrontSF); Lane 17, C-terminal propeptide (BackSF).

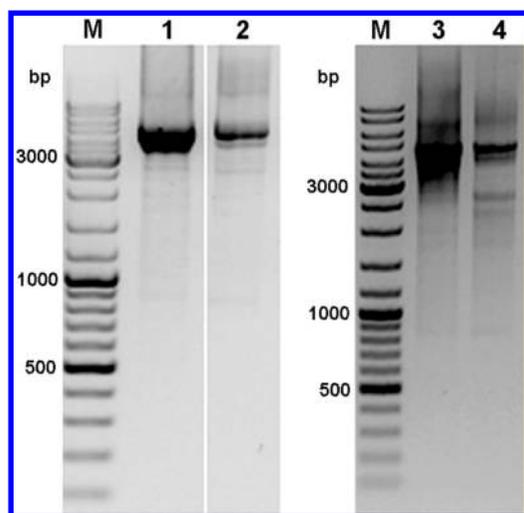


Figure 3. Final de novo genes encoding full-length native human collagen and three variants, assembled from modules as described in Figures 1 and 2. Lane 1, human collagen III, baseline with human leader sequence (4398 bp, MCol gene); Lane 2, human collagen III, baseline with yeast α -mating factor leader sequence (4596 bp, α MF-MCol gene); Lane 3, human collagen III with two non-native cysteines and human leader sequence (4398 bp, MCol-Cys1 gene); Lane 4, human collagen III with two non-native cysteines and yeast α -mating factor leader sequence (4596 bp, α MF-MCol-Cys1 gene); Lane M, size marker.

for native human collagen III (MCol) and its variants (MCol-Cys1, α MF-MCol, α MF-MCol-Cys1) were fabricated (Figure 3). The final gene sequences were confirmed, and collagen and variant genes were cloned into the YEp expression plasmid containing the *GALI* promoter, as described in Figure S-5 (Supporting Information) and in Experimental Methods.

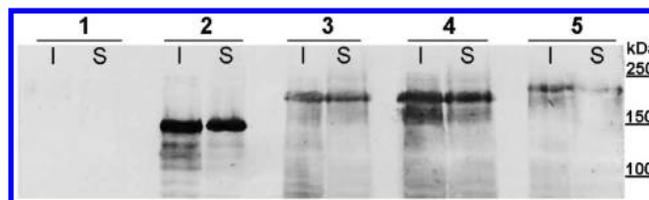


Figure 4. Western blot of human collagen III proteins and its variants. All constructs were cloned into a yeast episomal plasmid (YEp) and expressed in yeast strain InvSc1 with integrated prolyl-4-hydroxylase α and β subunits (strain InvSc1p4H). Lane 1, plasmid without collagen III, negative control (YEpdelta); Lane 2, native human collagen III (with EDS) gene (YEpCol3A1); Lane 3, modular gene encoding human collagen III (YEpMCol); Lane 4, modular gene encoding human collagen III with non-native cysteines (YEpMColCys1); Lane 5, modular gene encoding human collagen III with yeast α -mating factor leader sequence (YEp α MCol); I = insoluble fraction; S = soluble fraction.

S. cerevisiae Engineered with Human Prolyl-4-hydroxylase Expresses Recombinant Hydroxylated Human Collagen III and Its Tailored, Non-Native Variants.

One limitation in generating stable, recombinant collagen is the need to hydroxylate prolines in the protein. In *Pichia pastoris* and *S. cerevisiae*, this can be performed by incorporating the hydroxylase genes into the yeast genome, and we have therefore chosen yeast as the recombinant expression system in our biopolymers platform.^{28,55} We selected the yeast *S. cerevisiae* due to the ease of genetic manipulations and the availability of both integrative and autonomous plasmid systems. As described in Experimental Methods, we have integrated the CODA-optimized genes for both the alpha and beta subunits of human prolyl-4-hydroxylase into the genome of *S. cerevisiae* strain InvSc1 (InvSc1p4H). The Western blot in Figure 4 shows bands at the molecular weight for collagen for all samples containing a collagen gene or variant, including expression of Col3A1 (lanes 2I, 2S), the synthetic modular collagen III gene (lanes 3I, 3S), the variant gene containing non-native cysteines (lanes 4I, 4S), and the modular gene with the yeast leader sequence (lanes 5I, 5S). All of these collagen proteins were present in both the soluble and the insoluble fractions. In contrast, lanes containing the lysate from the yeast host (no collagen gene; lanes 1I, 1S) show no collagen product.

The collagen gene with the native human leader sequence gave consistently higher expression than the gene containing the leader sequence from yeast α -mating factor. Furthermore, expression levels of the cysteine variant were comparable to those obtained from the synthetic modular collagen gene. It appears that the incorporation of cysteine, at the level of frequency in this variant, does not inhibit expression of the protein. This is a promising result for generating variants that could be covalently modified at specific sites.

We note that the product of the COL3A1 gene is shorter than the collagen obtained from the modular gene, which is expected since COL3A1 contains an Ehlers-Danlos syndrome (EDS) deletion. We used this collagen as a positive control for protein expression because it was the only form of the COL3A1 gene that was commercially available (ATCC). Its use does not affect our final conclusions, as this gene simply confirms collagen expression in yeast. This is already a known result,^{28,39} and we do not use this mutant sequence as a basis for subsequent gene synthesis.

The activity of the hydroxylase was confirmed by amino acid analysis, although levels were lower than expected, approximately 0.5% proline hydroxylation in purified samples of collagen when produced in InvSC1p4H. This low level of

hydroxylation suggests an imbalance between collagen and prolyl hydroxylase expression and may also reflect instability of the integrated hydroxylase genes and inefficient translocation of the P4H β -subunit into the endoplasmic reticulum. Manipulation of strain, copy number, and secretion signal increased hydroxylation 10-fold to approximately 5% (see Discussion and Supporting Information). In contrast, control samples in which YEpCol3A1 was transformed into *S. cerevisiae* strain InvSC1 in the absence of α and β prolyl-4-hydroxylases yielded no detectable hydroxylation.

Our overall collagen yields in the InvSC1p4H are greater than those published in the literature for other *Saccharomyces* systems. After full purification, we obtained an average of $5.1 \pm 0.4 \mu\text{g}$ collagen/mg of total protein for COL3A1, MCol, and MColCys1. Prior studies on producing full-length collagen in *S. cerevisiae* reported a yield of $1.6 \mu\text{g}$ full-length collagen/mg protein.³⁹ We note that little optimization of yield was performed for our investigations, and it is likely that overall amounts of collagen can be significantly increased with strain improvement and fermentation studies.

Purified Collagen and Its Polymeric Variants are Trimeric and Consistent with Triple-Helical Assembly. Figure 5A shows that when pepsin is added to the crude cell lysates, most of the background yeast protein is digested and bands remain that correspond to the triple-helical region of collagen or the collagen variants. Samples can be further purified on a Sepharose XL cation-exchange column (Figure 5B) to remove the proteolytically degraded proteins and fragments.

Previous work has established that the enzymes pepsin and trypsin can be used as probes for triple helicity since they cleave the N- and C-termini of fibrillar collagen while leaving the (Gly-X-Y)_N triple-helical domain intact.⁵⁶ When these Gly-X-Y regions are in monomeric form, however, they are prone to proteolytic degradation.⁵⁷ Therefore, the presence of distinct bands at the expected molecular weight of the triple-helical region after pepsin digestion (Figure 5A) suggests that the collagen structure produced in our yeast expression system is consistent with triple-helicity. The trimeric nature of this assembly is further supported by nondenaturing blue native polyacrylamide gel electrophoresis (Figure 6), which yields bands at molecular weights corresponding to three times the size of a single polymeric strand. Expected sizes of a trimer are approximately 190 and 270 kDa for collagen III with EDS and for modular collagen III, respectively.

Discussion

Our goal was to create a flexible gene synthesis strategy that enables the fabrication of specifically tailored recombinant variants of full-length collagen-based polymers. Such a system would expand the scope of the material and cell-interaction properties of collagen and artificial ECM. We have shown that a novel, modular gene strategy enables bottom-up gene synthesis, and these genes can be used to produce non-native collagen variants. The advantages of a bottom-up gene assembly approach are that significant, multiple changes can be designed that span the entire gene, can be any distance from each other, and are not limited by the natural sequences of a gene template.

In previous methods for producing collagen-based polymers, restriction sites and enzymatic ligation were used to introduce alterations in the gene. These traditional strategies for creating collagen-mimetic biopolymers, however, place limits on allowable sequences, locations of changes, the scope of sequence variability and biologically relevant sites, and the lengths of

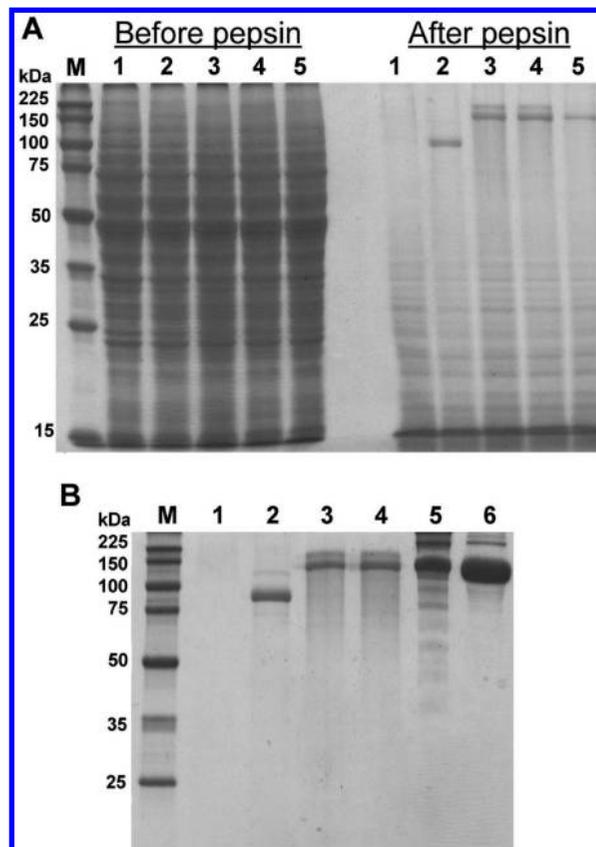


Figure 5. SDS-PAGE gels of recombinant human collagen III and its variants which are produced in InvSC1p4H. (A) Right and left sides of gel show soluble fractions of crude cell lysate without and with 200 $\mu\text{g}/\text{mL}$ pepsin, respectively. Lane 1, no collagen III, negative control (YEpdelta); Lane 2, human collagen III (with 303-amino acid EDS deletion; YEpCol3A1); Lane 3, modular human collagen III (YEpMCol); Lane 4, modular human collagen III with non-native cysteines (YEpMColCys1); Lane 5, modular human collagen III with yeast α -mating factor leader sequence (YEp α MCol); Lane M, molecular weight marker. (B) Purified samples of collagen III and its variants. Lanes 1–4, same as part A, but purified by pepsin-digestion followed by cation-exchange column. Lane 5, commercially obtained native human collagen III from placenta; Lane 6, commercially obtained human collagen III, which is produced in *Pichia pastoris*; Lane M, molecular weight marker.

the final polymeric product. For example, the use of naturally occurring restriction sites to introduce non-native sequences into a collagen gene is limited by the location of the restriction sites.⁵⁸ In another strategy, polymers can be made by ligating repeating oligonucleotide cassettes together in tandem. However, due to the repeating nature of the subunits, multimer ligation of such cassettes is not easily controlled. A range of DNA lengths is typically obtained with a maximum length of only ~ 1000 bp;^{31,32} this is one-third the length of the native gene encoding the helical, polymeric region of collagen. Furthermore, from these collagen-like gene fragments, protein products of only 18³¹ and 22 kDa³² were reported, smaller than the ~ 100 kDa of mature collagen III. An analogous strategy for biopolymer synthesis ligates copies of a 234-amino acid region of collagen II in tandem to obtain a final length corresponding to mature collagen.²⁹ Finally, site-directed mutagenesis has been performed on fibrillar collagen, but this approach limits the number of changes to single mutations or double mutations that are within only a few base pair proximity.^{59,60}

In this work, we have shown the feasibility of using a bottom-up approach for flexible introduction of foreign sites within

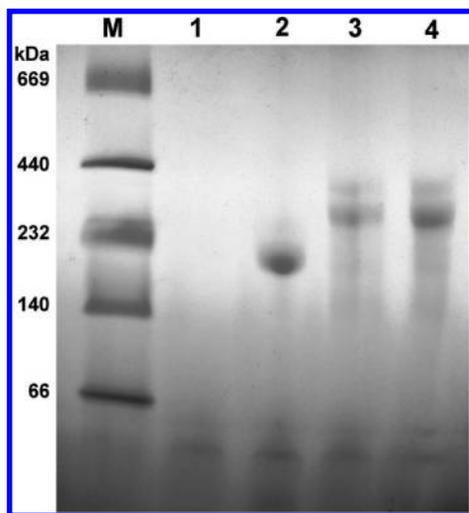


Figure 6. Blue native PAGE gel of purified collagen III samples which are produced in InvSc1p4H. Lane M, molecular weight marker; Lane 1, no collagen III, negative control (YEpdelta); Lane 2, native human collagen III (with EDS) (YECol3A1); Lane 3, modular human collagen III (YEpMCol); Lane 4, modular human collagen III with non-native cysteines (YEpMColCys1).

collagen. By combining computational optimization with a modular strategy, different combinations of non-native sequences (cysteines and yeast α -mating factor leader sequence) were placed in defined locations, at defined frequencies, and at large separations throughout the polymer. Our other work has also shown that several native integrin-binding sites can be removed from the collagen III sequence and replaced with a sequence from collagen I.⁵⁴ These results demonstrate that specific non-native sequences, locations, and their combinations within the full-length collagen biopolymer can be engineered. The flexibility and scope of modular strategies in general have been successfully implemented in other recombinant protein-based biopolymers, including elastin,^{61,62} elastin-like polypeptides,⁶³ and silk-like proteins.⁶⁴

One complexity of producing fibrillar collagen and its biomimetic forms originates from the need to post-translationally modify the primary sequence. The (Gly-X-Y)_N sequence of collagen enables assembly of a helix consisting of three polypeptide chains. X and Y can be almost any amino acid, but proline and hydroxyproline, respectively, are preferred.⁶⁵ Proline hydroxylation is essential for the stability of fibrillar collagen^{66,67} and the triple-helical, three-dimensional structure is required for correct function and cell interactions.^{8,68} Attempts to produce human collagen in bacterial cell systems have been limited because of their inability to hydroxylate selectively the collagen polypeptide at the Y-position.^{32,69,70} Expression in mammalian cells, although giving adequate post-translational modification, usually requires relatively long growth times and results in low protein yields.²⁹ Recombinant collagen can also be produced in other systems, including insect⁷¹ and plant cells.⁷²

Yeast systems are a particularly good choice for recombinant human collagen production because they can be engineered to hydroxylate prolines and can yield high expression levels of collagen. Prior studies have shown that *P. pastoris*^{55,73,74} and *S. cerevisiae*^{28,39,75} can produce hydroxylated human collagen types I, II, and III.^{39,55,73,76} In fact, commercially available recombinant human collagen III uses a *P. pastoris* expression system⁷⁷ based on cDNA. Prior studies have reported a range of 25–42% proline hydroxylation of recombinant collagens produced in *S. cerevisiae*^{28,39,55,75} and 44.2–47.2% in *P.*

pastoris.^{55,73} In contrast, fibrillar human collagens from native tissues have been reported to yield 42–54% hydroxylation, with variations dependent on tissue source, collagen type, and extraction method.^{55,78,79} Although our levels of approximately 0.5% proline hydroxylation in *S. cerevisiae* strain InvSC1p4H are low relative to previously reported levels for this yeast, we observed 10-fold increases (to 4–6% hydroxylation) by altering the *S. cerevisiae* strain, hydroxylase gene copy numbers, and the plasmid system (see Supporting Information). Hydroxylation levels in the InvSC1p4H strain do not appear to be dependent on the source of the collagen gene, as these values are consistent regardless of whether the collagen gene is the native (COL3A1) or the modular collagen (MCol). Ongoing studies are focused on further increasing hydroxylation by redesigning the recombinant yeast systems.

Although we use an expression system which has precedence in the literature, the unique aspects of our work are the use of a completely de novo synthetic gene to express a fibrillar human collagen and more importantly, the demonstration that variants built by the flexibility of this gene can indeed also be expressed in yeast. All prior yeast expression reports that we are aware of have used collagen cloned from cDNA as the starting basis for their genes.^{28,39,55,59,73,75} Our new bottom-up ability to make and express variants starting from custom-ordered oligonucleotides opens up vast possibilities in fundamental collagen studies and in synthetic biology applications which were not previously possible with collagen genes from cDNA or short collagen-like peptides.

In addition to developing the technology to create a de novo synthetic collagen gene, we have applied the technology to demonstrate that mutants of collagen which contain multiple cysteine (Cys) residues in defined places can be recombinantly fabricated in a single round of assembly. Cysteines are prevalent in the C-propeptide domains of fibrillar collagens and are speculated to bind the collagen trimers together for initiating triple-helical formation.^{57,80} However, they are absent from the internal portion of the triple-helical regions of all normal fibrillar human collagens, with Cys mutations resulting in disease pathologies.^{81–83}

Our modular approach in synthesizing recombinant collagen enables the ability to make specific full-length variants of interest, thereby expanding the scope of potential collagen-based materials. Placing multiple Cys into prescribed locations within the fibrillar region allows significant opportunities for expanding matrix properties by covalent incorporation of molecules. In fact, others in the field have noted that “[collagen’s] site-specific covalent modification is not feasible”,³⁶ leading to approaches using short synthetic collagen-like peptides as an alternative for producing synthetic matrices. With our novel approach and the ability to introduce multiple non-native thiols into precisely defined locations of native collagen, we enable multiple site-specific covalent changes.

Non-natural functionalities in collagen could include the introduction of covalent cross-linkers for increasing mechanical strength, altering the number of proteolytic sites for tuning degradation kinetics, and combining cell interaction sites from different classes of collagen to define cellular microenvironments. This design flexibility is valuable for applications in tissue engineering and drug delivery and also enables novel mechanistic studies of cell–matrix interactions. It will enable the rational creation of new scaffolds that closely mimic natural ECM proteins for designing appropriate microenvironments for regenerating tissues. This technology also has potential in

medicine, as the effects of natural mutations resulting in abnormal matrix pathologies can be investigated in molecular detail.

Acknowledgment. We gratefully acknowledge Merce Dalmau and Devan Huynh for assistance with the FPLC. We also thank Margaret Condron and David Teplow at the UCLA Biopolymer Laboratory core facility for assistance with amino acid analysis. The PDB structure of the collagen-like peptide used in the Table of Contents graphic is 1QSU.⁸⁴ This work was supported by the National Science Foundation (DMR-0706669) and Verdezyne, Inc. (previously CODA Genomics, Inc.). Materials described in this report are available for distribution according to the terms of a Materials Transfer Agreement of the Regents of the University of California.

Supporting Information Available. Analyses of amino acid and DNA repeating sequences, lists of oligonucleotide sequences, and additional experimental protocols for molecular cloning are presented. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References and Notes

- Stevens, M. M.; George, J. H. *Science* **2005**, *310*, 1135–1138.
- Stopak, D.; Wessells, N. K.; Harris, A. K. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 2804–2808.
- Sawhney, R. K.; Howard, J. *J. Cell Biol.* **2002**, *157*, 1083–1091.
- Mizuno, M.; Fujisawa, R.; Kuboki, Y. *J. Cell. Physiol.* **2000**, *184*, 207–213.
- Petrov, V. V.; Lijnen, P. J.; Fagard, R. H. *J. Hypertens.* **2002**, *20*, S29–S29.
- Eghbali, M.; Tomek, R.; Woods, C.; Bhambi, B. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 795–799.
- Saarialhokere, U. K.; Kovacs, S. O.; Pentland, A. P.; Olerud, J. E.; Welgus, H. G.; Parks, W. C. *J. Clin. Invest.* **1993**, *92*, 2858–2866.
- Gelse, K.; Poschl, E.; Aigner, T. *Adv. Drug Delivery Rev.* **2003**, *55*, 1531–1546.
- Li, S.-T. In *Biomaterials: Principles and Applications*; Park, J. B., Bronzino, J. D., Eds.; CRC Press/Taylor & Francis: Boca Raton, FL, 2002; pp 117–140.
- Kadler, K. *Protein Profile* **1994**, *1*, 519–638.
- Sansilvestri-Morel, P.; Rupin, A.; Jullien, N. D.; Lembrez, N.; Mestries-Dubois, P.; Fabiani, J. N.; Verbeuren, T. J. *J. Vasc. Res.* **2005**, *42*, 388–398.
- Benoit, D. S.; Anseth, K. S. *Biomaterials* **2005**, *26*, 5209–20.
- Lutolf, M. P.; Weber, F. E.; Schmoekel, H. G.; Schense, J. C.; Kohler, T.; Muller, R.; Hubbell, J. A. *Nat. Biotechnol.* **2003**, *21*, 513–518.
- Palecek, S. P.; Loftus, J. C.; Ginsberg, M. H.; Lauffenburger, D. A.; Horwitz, A. F. *Nature* **1997**, *385*, 537–540.
- Silva, G. A.; Czeisler, C.; Niece, K. L.; Beniash, E.; Harrington, D. A.; Kessler, J. A.; Stupp, S. I. *Science* **2004**, *303*, 1352–1355.
- Maheshwari, G.; Brown, G.; Lauffenburger, D. A.; Wells, A.; Griffith, L. G. *J. Cell Sci.* **2000**, *113* (Pt 10), 1677–1686.
- Andersson, A. S.; Backhed, F.; von Euler, A.; Richter-Dahlfors, A.; Sutherland, D.; Kasemo, B. *Biomaterials* **2003**, *24*, 3427–3436.
- Dike, L. E.; Chen, C. S.; Mrksich, M.; Tien, J.; Whitesides, G. M.; Ingber, D. E. *In Vitro Cell. Dev. Biol.: Anim.* **1999**, *35*, 441–448.
- Wong, J. Y.; Velasco, A.; Rajagopalan, P.; Pham, Q. *Langmuir* **2003**, *19*, 1908–1913.
- Gray, D. S.; Tien, J.; Chen, C. S. *J. Biomed. Mater. Res.* **2003**, *66*, 605–614.
- Lo, C. M.; Wang, H. B.; Dembo, M.; Wang, Y. L. *Biophys. J.* **2000**, *79*, 144–152.
- Lutolf, M. P.; Lauer-Fields, J. L.; Schmoekel, H. G.; Metters, A. T.; Weber, F. E.; Fields, G. B.; Hubbell, J. A. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 5413–5418.
- Halstenberg, S.; Panitch, A.; Rizzi, S.; Hall, H.; Hubbell, J. A. *Biomacromolecules* **2002**, *3*, 710–723.
- Rowley, J. A.; Madlambayan, G.; Mooney, D. J. *Biomaterials* **1999**, *20*, 45–53.
- Peyton, S. R.; Putnam, A. J. *J. Cell. Physiol.* **2005**, *204*, 198–209.
- Stemmer, W. P. C.; Cramer, A.; Ha, K. D.; Brennan, T. M.; Heyneker, H. L. *Gene* **1995**, *164*, 49–53.
- Larsen, L. S. Z.; Wassman, C. D.; Hatfield, G. W.; Lathrop, R. H. *Int. J. Bioinf. Res. Appl.* **2008**, *4*, 324–336.
- Vaughan, P. R.; Galanis, M.; Richards, K. M.; Tebb, T. A.; Ramshaw, J. A.; Werkmeister, J. A. *DNA Cell Biol.* **1998**, *17*, 511–518.
- Majsterek, I.; McAdams, E.; Adachi, E.; Dhume, S. T.; Fertala, A. *Protein Sci.* **2003**, *12*, 2063–2072.
- Fertala, A. Recombinant Collagen-like Proteins. U.S. Pat. Appl. #2004/0115771 A1, 2004.
- Vaughan, P. R.; Galanis, M.; Ramshaw, J. A.; Werkmeister, J. A. Method for Producing, in Yeast, a Hydroxylated Triple Helical Protein, and Yeast Host Cells Useful in Said Method. U.S. Pat. Appl. 2003/0064436 A1, 2003.
- Goldberg, I.; Salerno, A. J.; Patterson, T.; Williams, J. I. *Gene* **1989**, *80*, 305–314.
- Brodsky, B.; Ramshaw, J. A. *Matrix Biol.* **1997**, *15*, 545–54.
- Minond, D.; Lauer-Fields, J. L.; Nagase, H.; Fields, G. B. *Biochemistry* **2004**, *43*, 11474–11481.
- Raynal, N.; Hamaia, S. W.; Siljander, P. R. M.; Maddox, B.; Peachey, A. R.; Fernandez, R.; Foley, L. J.; Slatter, D. A.; Jarvis, G. E.; Farndale, R. W. *J. Biol. Chem.* **2006**, *281*, 3821–3831.
- Kotch, F. W.; Raines, R. T. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 3028–3033.
- Yamazaki, C. M.; Kadoya, Y.; Hozumi, K.; Okano-Kosugi, H.; Asada, S.; Kitagawa, K.; Nomizu, M.; Koide, T. *Biomaterials* **2010**, *31*, 1925–34.
- Lathrop, R. H.; Hatfield, G. W. Method for Producing a Synthetic Gene or Other DNA Sequence. U.S. Pat. #7,262,031, 2004.
- Olsen, D. R.; Leigh, S. D.; Chang, R.; McMullin, H.; Ong, W.; Tai, E.; Chisholm, G.; Birk, D. E.; Berg, R. A.; Hitzeman, R. A.; Toman, P. D. *J. Biol. Chem.* **2001**, *276*, 24038–24043.
- Vuori, K.; Pihlajaniemi, T.; Marttila, M.; Kivirikko, K. I. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 7467–7470.
- Gietz, R. D.; Sugino, A. *Gene* **1988**, *74*, 527–534.
- West, R. W.; Yocum, R. R.; Ptashne, M. *Mol. Cell. Biol.* **1984**, *4*, 2467–2478.
- Lee, K. M.; Da Silva, N. A. *Yeast* **2005**, *22*, 431–440.
- Ching, C. G. Sequential Cloned Gene Integration: Enhancements in *Saccharomyces cerevisiae*, Extension to Polyploid Yeast Strains, and Application to Polyketide Production. Ph.D. Thesis, University of California, Irvine, Irvine, CA, 2005.
- Lee, F. W. F.; Da Silva, N. A. *Biotechnol. Prog.* **1997**, *13*, 368–373.
- Alani, E.; Cao, L.; Kleckner, N. *Genetics* **1987**, *116*, 541–545.
- Gietz, R. D.; Woods, R. A. *Biotechniques* **2001**, *30*, 816–820.
- Bertani, G. *J. Bacteriol.* **1951**, *62*, 293–300.
- Laemmli, U. K. *Nature* **1970**, *227*, 680–685.
- Wittig, I.; Braun, H. P.; Schagger, H. *Nat. Protoc.* **2006**, *1*, 418–428.
- Schagger, H. *Nat. Protoc.* **2006**, *1*, 16–22.
- Sharp, P. M.; Cowe, E. *Yeast* **1991**, *7*, 657–678.
- Gutman, G.; Hatfield, G. W. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 3699–3703.
- Wang, S.-W.; Da Silva, N. A.; Hung, S.-P.; Lathrop, R. H. Collagen-Like Polypeptides and Encoding Polynucleotides. U.S. Pat. Appl #20090291474 A1, 2009.
- Vuorela, A.; Myllyharju, J.; Nissi, R.; Pihlajaniemi, T.; Kivirikko, K. I. *EMBO J.* **1997**, *16*, 6702–6712.
- Bruckner, P.; Prockop, D. J. *Anal. Biochem.* **1981**, *110*, 360–368.
- Bachinger, H. P.; Fessler, L. I.; Timpl, R.; Fessler, J. H. *J. Biol. Chem.* **1981**, *256*, 3193–3199.
- Berg, R. A.; Toman, P. D.; Wallace, D. G. Mutated Recombinant Collagens. U.S. Pat. #6653450, 2003.
- Williams, K. E.; Olsen, D. R. *Matrix Biol.* **2009**, *28*, 373–379.
- Chung, H. J.; Jensen, D. A.; Gawron, K.; Steplewski, A.; Fertala, A. *J. Mol. Biol.* **2009**, *390*, 306–318.
- Mithieux, S. M.; Tu, Y. D.; Korkmaz, E.; Braet, F.; Weiss, A. S. *Biomaterials* **2009**, *30*, 431–435.
- Wise, S. G.; Weiss, A. S. *Int. J. Biochem. Cell Biol.* **2009**, *41*, 494–497.
- MacEwan, S. R.; Chilkoti, A. *Biopolymers* **2010**, *94*, 60–77.
- Kluge, J. A.; Rabotyagova, U.; Leisk, G. G.; Kaplan, D. L. *Trends Biotechnol.* **2008**, *26*, 244–251.
- Persikov, A. V.; Ramshaw, J. A. M.; Kirkpatrick, A.; Brodsky, B. *Biochemistry* **2000**, *39*, 14960–14967.
- Bella, J.; Eaton, M.; Brodsky, B.; Berman, H. M. *Science* **1994**, *266*, 75–81.
- Persikov, A. V.; Ramshaw, J. A. M.; Brodsky, B. *Biopolymers* **2000**, *55*, 436–450.
- Emsley, J.; Knight, C. G.; Farndale, R. W.; Barnes, M. J. *J. Mol. Biol.* **2004**, *335*, 1019–1028.

- (69) Buechter, D. D.; Paoletta, D. N.; Leslie, B. S.; Brown, M. S.; Mehos, K. A.; Gruskin, E. A. *J. Biol. Chem.* **2003**, *278*, 645–650.
- (70) Cappello, J.; Ferrari, F. High Molecular Weight Collagen-like Protein Polymers. U.S. Pat. #5496712, 1996.
- (71) Lamberg, A.; Helaakoski, T.; Myllyharju, J.; Peltonen, S.; Notbohm, H.; Pihlajaniemi, T.; Kivirikko, K. I. *J. Biol. Chem.* **1996**, *271*, 11988–11995.
- (72) Perret, S.; Merle, C.; Bernocco, S.; Berland, P.; Garrone, R.; Hulmes, D. J. S.; Theisen, M.; Ruggiero, F. *J. Biol. Chem.* **2001**, *276*, 43693–43698.
- (73) Nokelainen, M.; Tu, H.; Vuorela, A.; Notbohm, H.; Kivirikko, K. I.; Myllyharju, J. *Yeast* **2001**, *18*, 797–806.
- (74) Myllyharju, J.; Nokelainen, M.; Vuorela, A.; Kivirikko, K. I. *Biochem. Soc. Trans.* **2000**, *28*, 353–357.
- (75) Toman, P. D.; Chisholm, G.; McMullin, H.; Giere, L. M.; Olsen, D. R.; Kovach, R. J.; Leigh, S. D.; Fong, B. E.; Chang, R.; Daniels, G. A.; Berg, R. A.; Hitzeman, R. A. *J. Biol. Chem.* **2000**, *275*, 23303–23309.
- (76) Olsen, D.; Yang, C.; Bodo, M.; Chang, R.; Leigh, S.; Baez, J.; Carmichael, D.; Perala, M.; Hamalainen, E.-R.; Jarvinen, M.; Polarek, J. *Adv. Drug Delivery Rev.* **2003**, *55*, 1547–1567.
- (77) Baez, J.; Olsen, D.; Polarek, J. W. *Appl. Microbiol. Biotechnol.* **2005**, *69*, 245–252.
- (78) Uitto, J. *Arch. Biochem. Biophys.* **1979**, *192*, 371–379.
- (79) Chung, E.; Miller, E. J. *Science* **1974**, *183*, 1200–1201.
- (80) Koide, T.; Nagata, K. *Collagen* **2005**, *247*, 85–114.
- (81) Malfait, F.; Symoens, S.; De Backer, J.; Hermans-Le, T.; Sakalihasan, N.; Lapiere, C. M.; Coucke, P.; De Paepe, A. *Hum. Mutat.* **2007**, *28*, 387–395.
- (82) Cabral, W. A.; Makareeva, E.; Letocha, A. D.; Scribanu, N.; Fertala, A.; Steplewski, A.; Keene, D. R.; Persikov, A. V.; Leikin, S.; Marini, J. C. *Hum. Mutat.* **2007**, *28*, 396–405.
- (83) Lund, A. M.; Joensen, F.; Christensen, E.; Duno, M.; Skovby, F.; Schwartz, M. *Clin. Genet.* **2008**, *73*, 97–101.
- (84) Kramer, R. Z.; Venugopal, M. G.; Bella, J.; Mayville, P.; Brodsky, B.; Berman, H. M. *J. Mol. Biol.* **2000**, *301*, 1191–1205.

BM100052Y