The C-Propeptide in Collagen Proteostasis

by

Rasia (Chichi) Li

B.S. Chemistry
University of Virginia, 2016

SUBMITTED TO THE DEPARTMENT OF CHEMISTRY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN CHEMISTRY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MAY 2022

© 2022 Massachusetts Institute of Technology. All rights reserved.

Signature of Author: _____
Department of Chemistry
May 6, 2022

Certified by: _____
Matthew D. Shoulders
Associate Professor
Thesis Supervisor

Accepted by: _____
Adam P. Willard
Associate Professor
Graduate Officer

This doctoral thesis has been examined by a committee of the Department of Chemistry as follows:

Ronald T. Raines
Roger and Georges Firmenich Professor
Thesis Committee Chair

Matthew D. Shoulders
Associate Professor
Thesis Supervisor

Elizabeth M. Nolan
Ivan R. Cottrell Professor of Immunology
Thesis Committee Member

The C-Propeptide in Collagen Proteostasis

by

Rasia (Chichi) Li

Submitted to the Department of Chemistry on May 6, 2022 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Chemistry

ABSTRACT

Collagen folding is initiated at the C-terminal propeptide (C-Pro) domain, a globular domain wholly distinct from the collagen's characteristic triple helix. The C-Pro domain is responsible for assembling three collagen strands into the correct orientation and stoichiometry and nucleating folding of the most abundant protein in the human body. While the function of the C-Pro domain in guiding collagen assembly is well accepted, its role in proteostasis has only recently been appreciated. In Chapter 2, we demonstrate that the highly-conserved *N*-glycan in the collagen-I C-Pro domain is critical for maintaining collagen proteostasis under challenging conditions. Specifically, the *N*-glycan facilitates interaction between procollagen and ER lectin chaperones to ensure proper folding of misfolding-prone collagen variants or wild-type collagen under proteostatic stress. In Chapter 3, we present progress towards understanding the molecular mechanisms of collagen assembly. We previously showed that collagen-I C-Pro assembly patterns are guided by Ca^{2+}-dependent non-covalent assembly of all C-Pro trimers, followed by covalent immortalization by interchain disulfide bonds. While prior work focused on the C-Pro domains in isolation, Chapter 3 explores unanswered questions about collagen assembly using full-length procollagen constructs. We show that regions of procollagen beyond the C-Pro domain play an unexpected role in defining the ability of triple-helical domains to homotrimerize. These results also yield fresh insights into the still unknown molecular features that promote procollagen heterotrimerization. Collectively, the work described in this thesis advance our understanding of the critical roles of the C-Pro domain in collagen proteostasis.

Thesis Supervisor: Matthew D. Shoulders
Title: Associate Professor

ACKNOWLEDGEMENTS

Graduate school, and graduate school in a pandemic, has been an experience unlike any other. None of my achievements would have been possible without the following people:

Matt— Thank you for always believing in me and my science, even when I didn't. My admiration for your optimism, enthusiasm, tenacity, and love for science has only grown over the past six years. I am incredibly grateful to have had an advisor who truly cares for his students' well-being and tries to do his best by them.

Ron, Liz, and the other chemical biology professors— Thank you for always providing new perspectives on my work and for teaching me to appreciate a greater range of scientific innovation.

The broader scientific community, including, but not limited to, PIs and scientists, core facility staff, technical support staff, *etc*.— Thank you for providing ideas, feedback, reagents, access to equipment, technical services, and support. I greatly admire the depth of your expertise and your willingness and capability to share it.

The Shoulders lab, past and present— I am grateful to the Shoulders lab community above all else for making my six years here such a positive experience. A lot has changed over the years, but the people have remained brilliant, kind, giving, patient, and fun. Some of my best memories come from spontaneous lunch room conversations, and even on the worst days, I could count on someone being there to lighten the load.
To the O.G. Shoulders lab— Thank you for providing a nurturing environment for me to learn and ask questions. You all set great examples for how to be scientists without neglecting the things that matter in life.
To the new generation— I hope I've been able to pay forward what I've received. Your resilience through the past few years is inspiring, as is your commitment to improving our faults.

Team Collagen—Andrew, Madeline, Duc, Azade, Alex, Michelle, Agata, Kathryn, Seo-yeon, Sorin, Jingting, Sophia— You guys understand the struggles of this project better than anyone, yet I've always felt supported and encouraged by our discussions. I think the number of laughs we've shared far exceeds the number of deep sighs, and that speaks to the incredible spirit of this group.
Andrew, Madeline, and Duc— You taught me everything I know with such patience and spirit of giving. I will always be indebted to you for answering my many questions, even years after leaving the lab, and can only aspire to be as incredible as you all are.

Rebecca and Sam— I couldn't have done it without you guys. Thanks for always being there to provide commiseration, sympathy, distractions, laughs, and sometimes actual advice. There have been some tough moments over the years, and I'm proud of us for getting through them. Congrats to you both for achieving all that you have, and I'm excited to see what you accomplish next.

The biochem cohort, residents of 17 Pleasant Pl, 57 Warner, and other grad student friends— Thanks for always getting it.

My friends— Your texts, phone calls, Facetime conversations, Instagram memes, movie/TV nights, Boston explorations, holiday reunions, Zoom gatherings, visits, trips, and presence in my life have kept me going.

My family— Thanks for always ensuring that home is just a chaotic conference call away.

TABLE OF CONTENTS

Abstract 3
Acknowledgements 4
Table of Contents 5
List of Figures 7
List of Abbreviations 9
Chapter 1: Collagen Assembly and Proteostasis 12
Overview 13
The Fibrillar Collagens 13
The C-Pro Domain 15
The Triple-Helical Domain 24
The N-Pro Domain 33
ER Stress and the Unfolded Protein Response 35
Conclusions 36
References 38
Chapter 2: Collagen's Enigmatic, Highly Conserved *N*-Glycan has an Essential Proteostatic Function 50
Author Contributions 50
Acknowledgements 50
Abstract 51
Introduction 52
Results 55
Absence of the Conserved *N*-Glycan Impairs Secretion of Misfolding-Prone Collagen Variants. 55
Non-*N*-Glycosylated, Misfolding-Prone C-Proα1(I) Variants Induce ER Stress and are Retained as Intracellular Aggregates. 57
N-Glycosylation Promotes the Interaction of Misfolding-Prone C-Proα1(I) with the Lectin Chaperones. 60
Chemically Abrogating Access to the Calnexin/Calreticulin Chaperoning System Phenocopies Genetic Removal of the *N*-Glycan 60
Introduction of a Non-Native *N*-Glycosylation Sequon Elsewhere in C-Proα1(I) Restores Secretion of Misfolding-Prone, Non-*N*-Glycosylated C-Proα1(I). 62
The *N*-Glycan is Essential for Maintaining WT Collagen Proteostasis in the Context of ER Stress. 63
Discussion 65
Materials and Methods 68
Supporting Figures 73

Supporting Tables 75
References 76
Chapter 3: Refining Mechanistic Understanding of the Contributions of Collagen's Domains to Trimeric Assembly 80
Author Contributions 80
Abstract 81
Introduction 82
Results 86
Covalent linkage of C-Proα2(I) is insufficient to drive formation of pepsin-resistant proα2(I) triple helices. 86
Covalent linkage of C-Proα1(I) is not required for formation of pepsin-resistant proα1(I) homotrimers. 87
Swapping C-Pro domains does not alter assembly behavior of triple-helical domains. 88
Discussion 90
Materials and Methods 94
Supporting Figures 96
References 97
Chapter 4: Perspectives 99
References 103
Appendix A: Monitoring Assembly of the Collagen C-Pro Domain 106
Author Contributions 106
Introduction 107
Results and Discussion 108
Methods and Materials 112
References 114
Appendix B: Mapping Disulfide Bonds in the Collagen C-Pro Domain 115
Author Contributions 115
Introduction 116
Results and Discussion 117
Methods and Materials 121
References 123

LIST OF FIGURES

Chapter 1: Collagen Assembly and Proteostasis

Figure 1-1: Fibrillar collagen biosynthesis. 15

Figure 1-2: The fibrillar collagen C-Pro domain. 16

Figure 1-3: The chain recognition sequence within C-Proα1(III) may guide type-specific assembly.....18

Figure 1-4: The cysteine residues and Ca^{2+}-binding site in the C-Pro domain provide a possible mechanism for assembly 19

Figure 1-5: Potential functions of the collagen C-Pro *N*-glycan.....23

Figure 1-6: The structure of a collagen triple helix. 25

Figure 1-7: Collagen undergoes many post-translational modifications. 27

Figure 1-8: Hsp47 is involved in collagen secretion. 29

Figure 1-9: Misfolded procollagen is subject to multiple degradation pathways. 31

Figure 1-10: The N-Pro domains of the fibrillar collagens are highly varied. 33

Figure 1-11: The UPR responds to misfolded proteins. 35

Chapter 2: Collagen's Enigmatic, Highly Conserved *N*-Glycan has an Essential Proteostatic Function

Figure 2-1: The *N*-glycosylation motif within the collagen-I C-Pro domain is highly conserved. .53

Figure 2-2: Collagen's conserved *N*-glycan is required for folding, assembly, and secretion of misfolding-prone collagen variants. 56

Figure 2-3: Removal of the N-glycan from misfolding-prone but not wild-type C-Proα1(I) induces ER stress and intracellular aggregation. 59

Figure 2-4: Collagen's *N*-glycan addresses proteostasis defects by endowing access to the ER's lectin-based chaperoning system. 61

Figure 2-5: Collagen's *N*-glycan protects WT Colα1(I) against misfolding during ER stress. 64

Figure 2-6: Model for the context-dependent essentiality of collagen's conserved *N*-glycan.....67

Figure S2-1: Analyses of Proα1(I) and C-Proα1(I) expression. 73

Figure S2-2: Analyses of non-*N*-glycosylated, misfolding-prone C-Proα1(I) variants.....74

Chapter 3: Refining Mechanistic Understanding of the Contributions of Collagen's Domains to Trimeric Assembly

Figure 3-1: Fibrillar collagens form homo- or heterotrimers guided by the C-Pro domain. 82

Figure 3-2: S2C α2(I) forms disulfide-linked trimers but does not effectively form pepsin-resistant helices. 86

Figure 3-3: C2S α1(I) does not form disulfide-linked trimers yet is able to form pepsin-resistant triple helices. 87

Figure 3-4: Chimeric collagens behave consistently with their N-Pro and triple-helical domains, rather than their C-Pro domains. 88

Figure S3-1: Proα2(I) and S2C proα2(I) secrete poorly at 25 °C. 96

Appendix A: Monitoring Assembly of the Collagen C-Pro Domain

Figure A-1: *In vitro* refolding of the C-Pro domain shows disulfide-dependent reassembly.....109

Figure A-2: Diagonal SDS-PAGE allows for separation of the components of hetero-oligomers. 111

Appendix B: Mapping Disulfide Bonds in the Collagen C-Pro Domain

Figure B-1: Hypothesized disulfide linkages detected in all homo- and heterotrimeric species. 117

Figure B-2: OI variants within C-Proa1(I) are able to disrupt disulfide bonding. 119

LIST OF ABBREVIATIONS

ASPH	aspartyl-asparaginyl β-hydroxylase
ATF6	activating transcription factor 6
BBF2H7	BBF2 human homolog on chromosome 7
BiP	immunoglobulin binding protein
BSA	bovine serum albumin
BTP	bis-tris propane
CD	circular dichroism
cDNA	complimentary DNA
COPII	coat protein II
C-Pro	COOH-terminal propeptide
CRISPR	clustered regularly interspaced short palindromic repeats
CRS	chain recognition sequence
CRTAP	cartilage-associated protein
CST	castanospermine
C-telo	COOH-terminal telopeptide
CypB	cyclophilin B
DAPI	diamidino-2-phenylindole
DHFR	dihydrofolate
DMEM	Dulbecco's modified Eagle medium
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DTT	dithiothreitol
ECM	extracellular matrix
EDTA	ethylenediaminetetraacetic acid
EGTA	ethylene glycol tetraacetic acid
ER	endoplasmic reticulum
ERAD	ER-associated degradation
ERES	ER exit sites
ERGIC	ER-Golgi intermediate compartment
ERLAD	ER-to-lysosome-associated degradation
ERp29	ER resident protein 29
ERp57	ER resident protein 57
FAM134B	family with sequence similarity 134, member B
FBS	fetal bovine serum
FKBP	FK506-binding protein
Gal	galactose
GFP	green fluorescent protein
GGT	galactosylhydroxylysyl glucosyltransferase
GLT25D	galactosyltransferase
Glu	glucose
GM130	Golgi matrix protein 130

Grp94	glucose-regulated protein 94
GS	glucosidase
GSH	glutathione
GSSG	glutathione disulfide
GT	hydroxylysyl galactosyltransferases
h	hours
HA	hemagglutinin
HEK	human embryonic kidney
Hsp47	heat shock protein 47
HYOU1	hypoxia up-regulated protein 1
IAM	iodoacetamide
IP	immunoprecipitation
IRE1	inositol-requiring enzyme 1
kDa	kilodalton
LC3	microtubule-associated proteins 1A/1B light chain 3
LH	lysyl hydroxylase
MEM	minimum essential media
min	minutes
mRNA	messenger RNA
Ni-NTA	nickel nitrilotriacetic acid
N-Pro	NH_2-terminal propeptide
N-telo	NH_2-terminal telopeptide
OASIS	old astrocyte specifically induced substance
OI	osteogenesis imperfecta
P3H	prolyl-3-hydroxylase
P4H	prolyl-4-hydroxylase
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PDB	Protein Data Bank
PDI	protein disulfide isomerase
PDL	poly-D-lysine
PERK	protein kinase R-like endoplasmic reticulum kinase
PMSF	phenylmethylsulfonyl fluoride
PNGase F	peptide:N-glycosidase F
PPIase	peptidyl-prolyl *cis-trans* isomerase
PPII	polyproline II
PPT	preprotrypsin
qPCR	quantitative PCR
RFP	red fluorescent protein
RIPA	radioimmunoprecipitation assay buffer
RNA	ribonucleic acid
s	seconds

SDS	sodium dodecyl sulfate
SEC	size exclusion chromatography
TANGO1	transport and Golgi organization protein
TBS	Tris buffered saline
Tg	thapsigargin
TSP-1	thrombospondin-1
TSPN	thrombospondin N-terminal-like
UGGT	UDP-glucose:glycoprotein glucosyltransferase
UPR	unfolded protein response
VWC	von Willebrand factor type-C
WT	wild-type
YFP	yellow fluorescent protein

A NOTE ON COLLAGEN NOMENCLATURE USED HEREIN:

Gene names are written as COLXAY, where X is the collagen type in Arabic numerals and Y is the number of the chain in Arabic numerals.

Individual chains are written as αY(X), where X is the collagen type in Roman numerals and Y is the number of the chain in Arabic numerals.

“pro” preceding the chain name specifies procollagen, which includes both propeptides.

“N-Pro or “C-Pro” preceding the chain name specifies either the N-Pro or C-Pro domain of that chain alone, not fused to any other domain.

CHAPTER 1:

COLLAGEN ASSEMBLY AND PROTEOSTASIS

OVERVIEW

Collagen is the most abundant protein in the human body and, by one estimate, the second most abundant protein on the planet.[1] It serves as the molecular scaffold for animal life, comprising the primary proteinaceous component of connective tissue. Specifically, collagen is a fundamental component of the extracellular matrix (ECM) and is intimately involved in the ECM's many functions, ranging from providing structural support to regulating signal transduction to maintaining tissue homeostasis.[2]

Maintaining proper collagen biosynthesis is critical for many cells and tissues across human existence, yet collagen production is not a trivial process. Collagen monomers range from ~600 to ~3000 amino acids in length and fold into homo- or heterotrimeric molecules that are 75 to 425 nm in length.[3] They fold and assemble in the crowded endoplasmic reticulum (ER), where they interact with a large number of chaperones and modifying enzymes and are subject to extensive (and essential) co- and post-translational modifications.[4, 5] Assembly starts at the C-terminus, and cells often express multiple collagen strands simultaneously, meaning fully synthesized collagen strands must remain soluble and unfolded until they meet the correct partner strands.[6] Likely owing to collagen's unusual size and shape, it appears to utilize non-canonical secretion and degradation mechanisms that are still being actively investigated.[7-14] Secreted collagen molecules are subject to further processing, including cleavage of the propeptides, and modification, particularly formation of covalent crosslinks between triple helices, before they can assemble into their final supramolecular assemblies.[15] Remarkably, this entire process takes only ~30 minutes from translation to secretion.[16]

As a consequence of these complexities, collagen folding often goes awry. Under normal conditions, up to 20–50% of collagen is estimated to misfold,[17-22] and this problem is severely exacerbated in disease states. Dysregulated collagen production is linked to a variety of pathologies, from cancer to fibrosis to scurvy.[23] Those associated with genetic defects, usually in a collagen gene or collagen-related protein, have been collectively termed the "collagenopathies".[24] Disease-modifying treatments for the collagenopathies do not exist. Considerable effort is being directed towards understanding and targeting each part of the collagen production process with the goals of both understanding and potentially treating disease.[23]

THE FIBRILLAR COLLAGENS

There are twenty-eight different types of collagen, encoded by forty-four different genes. The excess of genes compared to collagen types reflects that some collagens are heterotrimers,

composed of multiple unique chains.[3] All collagens share the characteristic structural motif of a triple helix, formed by three polypeptide strands in polyproline II (PPII) helices twisted into a right-handed helix.[25] This triple-helical structure is enabled by a primary structure containing glycine at every third residue, in a repeating Gly-Xaa-Yaa triplet motif. Glycine, due to its small side chain, is the only amino acid that can sterically fit within the tightly packed interior of the triple helix. Xaa and Yaa can be any amino acid but are most commonly proline and 4(*R*)-hydroxyproline (Hyp), respectively, with Gly-Pro-Hyp being the most common triplet by far (10.5%).[26]

The various types of collagen are subdivided into classes based on the supramolecular structures they form. The fibrillar collagens (types I, II, III, V, XI, XXIV, and XXVII) form long, periodic fibrils that contribute to the structural scaffold of tissues such as bone, cartilage, and tendon (Table 1-1).[3] These collagens are the most abundant and most well-studied, particularly types I–III, and are the focus of this chapter.

Table 1-1: The fibrillar collagens and their assembly patterns.[3, 27-30]

Type	Strand	Homotrimeric form	Heterotrimeric form
type I	α1(I)	$[\alpha1(I)]_3$	$[\alpha1(I)]_2\alpha2(I)$
	α2(I)		$[\alpha1(I)]_2\alpha2(I)$
type II	α1(II)	$[\alpha1(II)]_3$	α1(XI)α2(XI)α3(XI)* α1(XI)α2(V)α3(XI)*
type III	α1(III)	$[\alpha1(III)]_3$	
type V	α1(V)	$[\alpha1(V)]_3$	α1(V)α2(V)α3(V) $[\alpha1(V)]_2\alpha2(V)$ α1(V)α1(XI)α2(V)
	α2(V)		α1(V)α2(V)α3(V) $[\alpha1(V)]_2\alpha2(V)$ $[\alpha1(XI)]_2\alpha2(V)$ α1(V)α1(XI)α2(V) α1(XI)α2(V)α3(XI)*
	α3(V)		α1(V)α2(V)α3(V)
type XI	α1(XI)		α1(XI)α2(XI)α3(XI)* $[\alpha1(XI)]_2\alpha2(V)$ α1(V)α1(XI)α2(V) α1(XI)α2(V)α3(XI)*
	α2(XI)		α1(XI)α2(XI)α3(XI)*
type XXIV	α1(XXIV)	$[\alpha1(XXIV)]_3$	
type XXVII	α1(XXVII)	$[\alpha1(XXVII)]_3$	
* α3(XI) has the same protein sequence as α1(II) with higher hydroxylysine-linked glycosylation content.[31, 32]			

The fibrillar collagens are synthesized as procollagen precursors, consisting of a single triple-helical domain bookended by short telopeptides and globular N- and C-terminal propeptide (N-Pro and C-Pro, respectively) domains.[30, 33] Upon co-translational insertion of the procollagen chain into the ER, the nascent collagen chains undergo extensive co- and post-translational modifications, in particular proline and lysine hydroxylation.[4] Folding and assembly into the procollagen trimer begins at the C-Pro domain, followed by zipper-like folding of the triple-helical domain.[34] Once the triple helix is properly formed, the individual strands are no longer able to be hydroxylated. The fully formed procollagen molecule is then secreted from the cell, where the propeptides are cleaved prior to incorporation of the collagen molecule into fibrils.[15] This chapter focuses on the process of fibrillar procollagen and assembly as it occurs in the ER, organized by the main domains of procollagen molecules (Fig. 1-1).

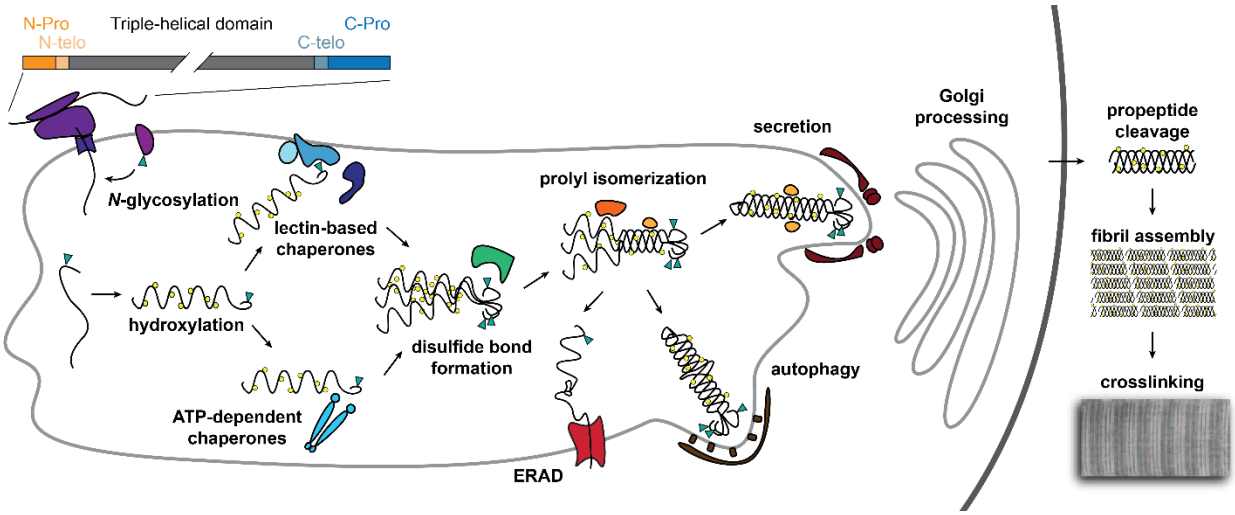


Figure 1-1: Fibrillar collagen biosynthesis.

Fibrillar collagen biosynthesis begins with translocation of individual strands into the ER. Each strand then undergoes a variety of co- and post-translational modifications prior to folding and assembly of the C-Pro domain. Once three C-Pro domains are properly assembled, the triple helix folds from C- to N-terminus. The resultant trimeric procollagen can be directed towards secretion or, if identified as somehow misfolded or misassembled, degradation. Secreted molecules undergo additional modifications during fibril assembly.

THE C-PRO DOMAIN

Early studies of collagen folding were confounded by the slow kinetics and low yield of properly folded collagen observed in *in vitro* renaturation of collagen purified from tissue.[35, 36] Both effects were attributed to the presence of some then-unknown mechanism of chain alignment *in vivo*, which was confirmed upon the discovery of the full procollagen precursor molecule.[37] The C-Pro domain was quickly shown to be responsible for overcoming the *in vitro* rate limiting step

of chain alignment, serving as the nucleation site for collagen folding. Other early stage functions, particularly register determination and chain selection, were also attributed to the C-Pro domain.[34]

The C-Pro domains of the fibrillar collagens are ~250 amino acids in length and ~30 kDa in size as monomers. In the last ten years, two high-resolution crystal structures of the collagen-III and collagen-I homotrimeric C-Pro trimers revealed that they adopt the overall appearance of a flower, with three petal regions, a base where the individual polypeptides come together, and a coiled-coil stalk.[38, 39] The crystal structures resolved discrepancies about the pattern of inter- and intrachain disulfide-linkages and revealed the presence of a Ca^{2+} ion bound at the interface of the monomers (Fig. 1-2). They also allowed for identification of specific interchain interactions proposed to be involved in folding and helped model the impact of disease-causing mutations that have been identified within the C-Pro domain.

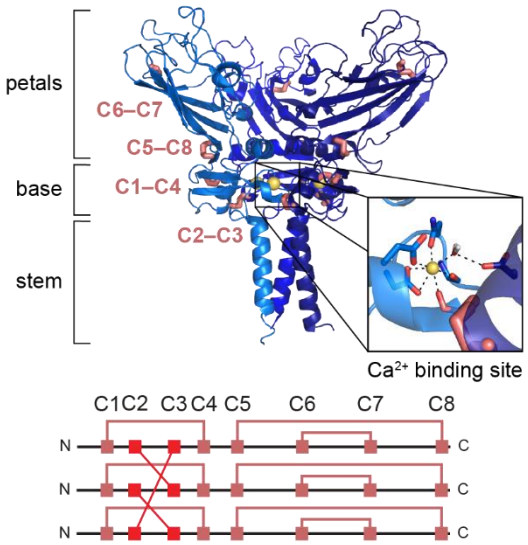


Figure 1-2: The fibrillar collagen C-Pro domain.
The C-Pro domain of homotrimeric collagen-I (PDB ID: 5K31) adopts a flower-like shape. Each monomer contains three intrachain disulfide bonds. At the interface between two monomers, there is a single interchain disulfide bond and a Ca^{2+} binding site.

The decades of research since the initial discovery of the C-Pro domain have largely focused on its role in governing trimeric assembly, and substantive progress has been made concerning specific interactions that govern chain selection as well as the role of its cysteines and Ca^{2+}-binding regions. More recently, there is a growing appreciation for the potential role of the C-Pro domain in collagen proteostasis, particularly in quality control. Major advances in these two areas are summarized below.

Chain-selective assembly. Many cells express multiple types of collagen simultaneously and, in the case of the heterotrimeric collagens, multiple strands of the same type.[40-42] Because trimerization of the C-Pro domain occurs prior to triple-helical folding, the C-Pro domain is generally believed to contain the molecular information that ensures proper chain selection. Chain selection can be broadly defined in two ways: (1) ensuring that each collagen strand only associates with other strands of the same type (with the exception of hybrid type-V and -XI molecules, which has led to the view that they are in fact a single type[43-45]), and (2) for heterotrimeric collagens, ensuring that collagen trimers of a given type have the correct stoichiometry.

One mechanism that contributes to ensuring type-specific chain selection was first identified by Bulleid and coworkers using a mini collagen system, which utilizes *COL3A1* and *COL1A2* genes with truncated triple-helical domains of ~190 amino acids expressed in a cell-free translation system. The mini-proα1(III) strand forms disulfide-linked, pepsin-resistant trimers, while the mini-proα2(I) strand does not form either. By swapping segments of the C-Pro domain between these two mini collagens, the researchers were able to identify a discontinuous 15 residue sequence within C-Proα1(III) that was sufficient to drive formation of pepsin-resistant trimers by mini-proα2(I).[46] This region, termed the chain recognition sequence (CRS), was noted to be fairly diverse amongst the fibrillar collagens while maintaining relative hydrophilicity (Fig. 1-3). It was therefore proposed to mediate initial interaction between C-Pro monomers of different collagen types. The crystal structure of C-Proα1(III) domain confirmed that the two segments of the CRS form salt bridges between monomers, and additional residues outside the CRS were identified as forming interchain interactions (Fig. 1-3).[38] The residues involved in these salt bridges are not conserved amongst the other fibrillar collagen strands (Fig. 1-3), and the crystal structure of the C-Proα1(I) homotrimer exhibited fewer interactions between strands in the same region, suggesting perhaps that the CRS is primarily important for type-III chain assembly.[39]

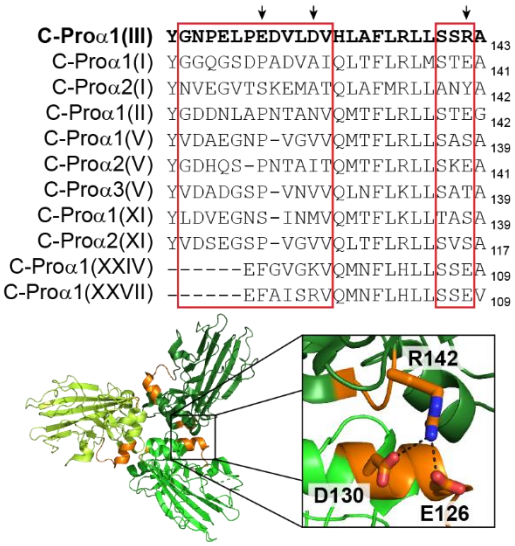


Figure 1-3: The chain recognition sequence within C-Proα1(III) may guide type-specific assembly.

The chain recognition sequence identified in C-Proα1(III) is shown to participate in two interchain salt bridges. The residues involved in these salt bridges are not conserved in the other fibrillar collagens, suggesting they may help ensure type-specific assembly for collagen-III (PDB ID: 4AE2).

Another intriguing difference between the crystal structures of C-Proα1(III) and C-Proα1(I) was a difference in symmetry of the overall trimer. An intrinsic asymmetry was observed in the C-Proα1(III) structure, particularly in the region between the two segments of the CRS, despite all three chains having the same amino acid sequence.[38] This observation was hypothesized to account for the evolution of heterotrimers. However, the C-Proα1(I) trimer was unexpectedly found to be fairly symmetric.[39] Whether this difference in symmetry alludes to any mechanism of chain selection, whether type-specific or stoichiometric, remains unknown. To date, no experiments have been performed to directly examine whether type-specific assembly can be disrupted by perturbation of the CRS or other proposed chain selection interactions. Thus, the degree to which these proposed molecular features are sufficient to control type-specific assembly remains unknown.

Meanwhile, the fundamental question of how collagen controls stoichiometric chain selection within a single type was historically considered in the context of collagen-I. Collagen-I exists naturally as a 2:1 α1(I):α2(I) heterotrimer, a ratio that is preserved even when the C-Pro domains are expressed independently of the rest of the domains.[39, 47] However, α1(I) homotrimers have been observed as a minority component of adult skin, with higher levels reported in disease

tissue.[48-55] Surprisingly, the α1(I) homotrimer is both more thermally stable and slower to denature than the heterotrimer,[56] raising questions about how heterotrimers are preferentially synthesized.

Although the crystal structure of the heterotrimeric collagen-I C-Pro domain has yet to be solved, modeling the C-Proα2(I) chain into the homotrimeric C-Proα1(I) structure led to the identification of several putative interchain salt bridges that are unique to the heterotrimer.[39] Mutagenic disruption of these interactions via substitution of one or more of the residues involved was shown to decrease incorporation of C-Proα2(I) into heterotrimers, although the data shown did not clarify whether this effect could have been due to lowered expression or misfolding of the variants.[39]

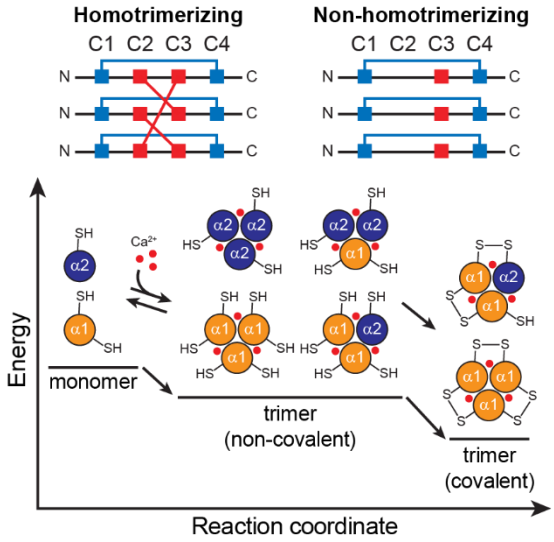


Figure 1-4: The cysteine residues and Ca^{2+}-binding site in the C-Pro domain provide a possible mechanism for assembly

The C-Pro domains of homotrimerizing collagens have 8 cysteines (the first four of which are shown), while non-homotrimerizing collagens are missing either C2 or C3. A model for stoichiometric collagen-I C-Pro assembly suggests that Ca^{2+} guides dynamic assembly of all species, while interchain disulfides immortalize specific biologically observed assemblies.

Another feature of the C-Pro domain that appears to be involved in chain selection is a network of cysteine residues. The C-Pro domain of all fibrillar collagens contain seven or eight highly conserved cysteine residues, numbered C1 through C8.[57] C1, C4, and C5–8 are present in all fibrillar C-Pro domains and form intrachain disulfide bonds. C2 and C3 form interchain disulfide bonds, but one or the other is missing in some collagen strands (Fig. 1-4).[38, 39] The strands that have the full complement of cysteines have all been observed to form homotrimers (with the exception of α2(XI)[27], although we have shown that C-Proα2(XI) forms disulfide-linked

homotrimers[47]), while those that are missing C2 or C3 have only been found in heterotrimers (Fig. 1-4).[47] In the case of α2(I), a seven-cysteine strand, it does not form homotrimers even when expressed in the absence of α(I).[58]

Initial studies using the mini collagen system examined the role of the missing C2 in α2(I) by substituting the existing serine residue with a cysteine. In this system, reintroduction of C2 did not induce formation of disulfide-linked trimers. In the decades since, any potential role of the Cys residues in informing collagen homo- versus heterotrimerization has been largely ignored.[59] However, our lab's recent work expressing the C-Pro domain alone, without any of the other collagen domains, in human cells demonstrated that the S2C substitution in C-Proα2(I) does in fact drive disulfide-linked trimer formation.[47] In fact, reintroduction of the missing cysteine in all non-homotrimerizing C-Pro domains allowed them to form stable, disulfide-linked homotrimers. Similarly, removal of C2 or C3 from homotrimerizing C-Pro domains abolished their ability to form disulfide-linked homotrimers.[47]

The discrepancy between our results and those obtained by Bulleid and co-workers could be related to contributions of the triple-helical and N-Pro domains to assembly or to other differences in the expression systems. Critically, we identified a specific, molecular explanation for why the mini collagen system did not yield disulfide-linked S2C α2(I) homotrimers.

The two C-Pro crystal structures revealed a previously unknown Ca^{2+} binding site found at the interface between monomers.[38, 39] The Ca^{2+} ion coordinates five residues on one chain and one residue on the neighboring chain via a water molecule (Fig. 1-2). Using sedimentation equilibration experiments, we made the remarkable, unexpected discovery that Ca^{2+} coordination actually promotes dynamic, non-covalent association of C-Proα2(I) into trimers.[47] We hypothesized that Ca^{2+}-mediated non-covalent assembly is necessary to template disulfide bond formation. The mini collagens were expressed in an *in vitro* translation system that required pre-translation treatment with Ca^{2+} chelators, obscuring the ability of S2C C-Proα2(I) to assemble into disulfide-linked homotrimers.[46, 59, 60]

Significantly, we observed dynamic, Ca^{2+}-mediated assembly of C2S C-Proα1(I) into homotrimers and the formation of 1:2 C-Proα1(I):C-Proα2(I) heterotrimers. Therefore, we proposed a model for C-Pro assembly wherein Ca^{2+}-mediated non-covalent assembly guides trimerization of the C-Pro domain into all possible stoichiometries, regardless of cysteine content. The formation of interchain disulfide bond formation then serves as a thermodynamic sink to immortalize appropriate species for triple-helix folding (Fig. 1-4).[47]

While this model accounts for the two observed species of collagen-I, the 2:1 α1(I):α2(I) heterotrimer and the α1(I) homotrimer, it leaves a key question unanswered: specifically, how is heterotrimer formation vastly preferred in biological settings? α1(I) homotrimers have more disulfide bonds than heterotrimers and in fact form a more stable triple helix than the heterotrimer, yet account for <5% of the collagen in normal skin.[61] Likely, there are additional molecular features in the collagen sequence that promote heterotrimer formation. Indeed, interchain disulfide formation appears to be unnecessary for triple helix formation in the mini collagen system, suggesting that other features likely contribute to chain-selective assembly .[60] This work sets the stage for further exploration of the fundamental mechanisms of trimer assembly. Of particular note, no attempt has been made to understand the mechanism of collagen assembly using full-length procollagen constructs, owing to the difficulties associated with genetic manipulation and biochemical analysis of such constructs.

Of note, chain selection of collagen-V appears to be perturbed in recombinant expression systems. Whereas endogenous collagen-V is typically found as a 2:1 α1(V):α2(V) heterotrimer, co-expression of these two chains in yeast, insect, and mammalian cells favored the α1(V) homotrimer.[62] This effect has not been reported with other heterotrimeric collagens, such as collagen-I. Expression of the triple-helical domain-binding chaperone, Hsp47 did not restore heterotrimer preference. This finding suggests that there may be specific cellular machineries that promote formation of heterotrimeric collagen-V, but the identify of these mechanisms remains unknown. The possibility that these mechanisms exist for other heterotrimeric collagens must be considered as well.

C-Pro Proteostasis. Correct folding of the C-Pro domain is critical for ensuring production of correct collagen molecules. Thus, even though the C-Pro domain is proteolytically removed from the functional collagen molecule, mutations within the C-Pro domain can cause significant impairment of collagen-containing tissues.[63] Approximately 6.5% of mutations within collagen-I genes that cause osteogenesis imperfecta (OI) are found in the C-Pro domain, and their phenotypes range from mild to lethal. This percentage is lower than what might be expected by random mutation rate, thus some have suggested that additional mutations are overlooked due to lack of phenotype (unlikely due to the C-Pro domain's critical function) or, more likely, embryonic lethality.[64]

Phenotypes observed due to C-Pro mutations may be caused by delayed folding and assembly, leading to overmodified triple helices, or impaired folding and assembly, leading to reduced collagen production and/or cellular dysfunction.[55, 63-69] Both of these effects can be

buffered by the cell's proteostasis network, a collection of chaperones and quality control enzymes that help maintain efficacious protein synthesis.[70, 71]

Studies of misfolding C-Pro variants have helped reveal how both mutant and wild-type collagen are engaged by the proteostasis network. Early studies of patient fibroblasts with OI-causing C-Pro domain mutations, including D1277H, W1312C, Δ1337–38, and a frameshift mutation, that disrupt chain association revealed that some mutant strands bind immunoglobulin binding protein (BiP), a key ER chaperone.[65, 66] Wild-type collagen has been shown to interact with protein disulfide isomerase (PDI), which assists with disulfide bond formation and may help retain monomeric C-Pro variants in the ER.[72-74]

The first detailed mapping of collagen-I C-Proα2(I) interactome in isolation (without the triple-helical domain)[69] revealed that a specific subset of the full-length collagen proteostasis network[75] interacts with the C-Pro domain.[75] The C-Pro domain interactors primarily fall within the categories of general ER chaperones, redox related proteins, lectin-based chaperones, and secretory pathway proteins. These ER proteostasis network interactions are expected for a globular protein that is disulfide-rich and *N*-glycosylated. Comparison of the wild-type C-Proα2(I) interactome with that of the OI-causing C1163R C-Proα2(I) variant demonstrated that the ER proteostasis network is able to differentially engage misfolding C-Pro variants. These results demonstrate that the C-Pro domain itself is able to engage with the ER proteostasis network to try and resolve folding defects, even in the absence of the triple-helical domain.

A potentially interesting feature of the C-Pro domain from a proteostasis perspective is a highly-conserved *N*-glycosylation site. *N*-Glycans can directly or indirectly promote protein folding.[76, 77] All fibrillar collagens have at least one putative *N*-glycosylation site within their C-Pro domains, with proα1(I) and proα2(I) known to contain exactly a single site that is occupied.[78-80] This site is highly conserved amongst the human fibrillar collagens and in the proα1(I) or equivalent strand of many species, suggesting a meaningful function.[80] However, early work in which the *N*-glycosylation site was genetically removed via site-directed mutagenesis demonstrated that non-*N*-glycosylated proα1(I) did not exhibit any functional difference compared to *N*-glycosylated proα1(I).[81] We recently demonstrated that, while removal of the proα1(I) *N*-glycan indeed does not impact well-folding proα1(I), its absence causes major proteostasis defects when collagen folding is challenged by either destabilizing mutations or even moderate ER stress.[80] This effect was due to the *N*-glycan promoting interaction of collagen with lectin-based chaperones to ensure proper folding and secretion under proteostatically challenged conditions (Fig. 1-5).

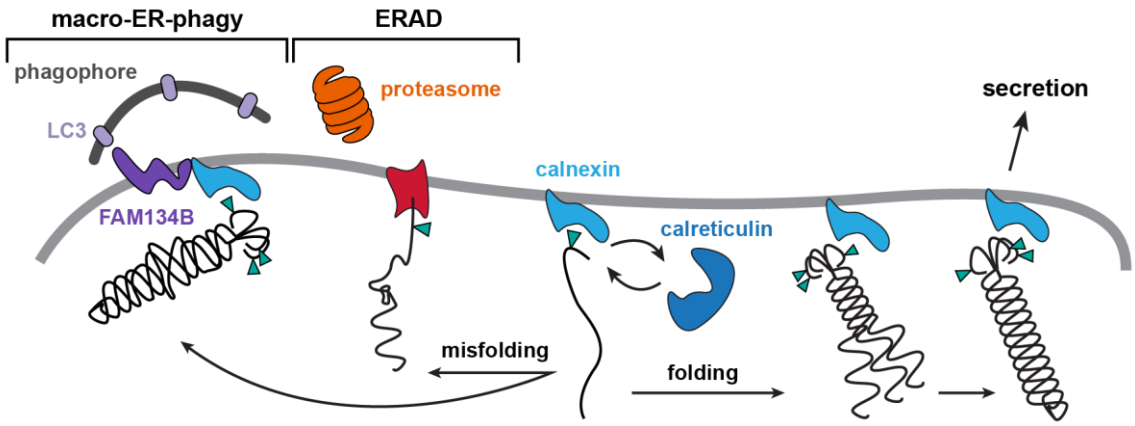


Figure 1-5: Potential functions of the collagen C-Pro *N*-glycan.
The collagen C-Pro *N*-glycan interacts with the ER lectin chaperones to assist with folding. Folding collagen is then directed to secretion. Misfolding collagen undergoes *N*-glycan processing that directs the protein to ERAD or macro-ER-phagy. Macro-ER-phagy has been shown to require *N*-glycan interaction with calnexin and FAM134B.

The C-Pro *N*-glycan and its interaction with the lectin chaperones is also noteworthy due to existing and emerging knowledge of the roles of the ER lectin-based machinery in quality control. Targeting of misfolded proteins to either ER-associated degradation (ERAD) or ER-to-lysosome-associated degradation (ERLAD) involves *N*-glycan processing.[82, 83] The latter process has been explicitly shown to require interaction between calnexin, a known C-Pro interactor,[80] and misfolded collagens via the collagen *N*-glycan (Fig. 1-5). Other calnexin-associated proteins, including collagen interactors calreticulin and ERp29, have also been suggested to be involved in collagen proteostasis and quality control.[75, 84] These recent findings point to the exciting possibility that the collagen C-Pro domain, and its *N*-glycan specifically, is central to ensuring that misfolded collagen is either refolded or targeted to a specific degradative pathway.

The interaction of the collagen C-Pro domain and calnexin is also intriguing in the context of evidence showing that procollagen associates with the ER membrane. Membrane association has been hypothesized to be advantageous for collagen assembly by helping co-localize three procollagen strands, though there is little evidence to date (Fig. 1-5).[85] Given that the entire collagen strand must be synthesized prior to folding from the C-terminus, free diffusion of unassociated strands would likely slow the process of assembly and potentially increase the chances of misfolding or aggregation. Recent work proposes the existence of collagenosomes", discrete bodies associated with the ER membrane where collagen folding may take place.[86] Microsome spreading has also suggested that procollagen strands are associated with the rough ER membrane via the C-Pro domain, although the exact nature of this association is not known.

Bächinger and co-workers suggested that the C-Pro domain may interact with specific membrane proteins or lipids.[85] Given that calnexin is a membrane-bound protein, its association with the C-Pro and its role in folding provides an appealing hypothesis for a mechanism of membrane-associated folding and assembly of the collagen trimer.[87]

More recently, procollagen containing OI-causing C-Pro domain variants were shown to mislocalize to the ER lumen, rather than the ER membrane, based on loss of colocalization with calnexin.[64] Why misfolded C-Pro domains do not associate with the membrane is unknown, especially given that calnexin is more likely to bind misfolded proteins. These same variants were shown to have delayed assembly, potentially due to the loss of spatial organization. These results again suggest that the C-Pro domain may serve as one important avenue for effective quality control of misfolding procollagens.

Summary. The C-Pro domain of collagen has long been appreciated for its role in nucleating triple helix folding, determining chain alignment and register, and ensuring correct alignment. While significant progress has been made towards understanding the exact molecular features that contribute to these functions, much remains unknown about how a relatively small domain governs such complex processes. Concurrently, a greater appreciation for the role of the C-Pro domain in collagen proteostasis, particularly quality control, has been gaining ground. Growing bodies of work suggest that ERLAD is closely tied to ER-to-Golgi transport, including the involvement of ER-phagy receptor FAM134B and various COPII-associated proteins in both processes.[88-94] The C-Pro *N*-glycan has already been demonstrated to be involved in ERLAD via FAM134B-calnexin interaction;[83] whether C-Pro-driven mechanisms are involved in ER-to-Golgi transport or other quality control pathways remains unknown. As the mechanisms behind the various pathways of collagen quality control and secretion continue to be studied, additional functions of the C-Pro domain may continue to be uncovered.

THE TRIPLE-HELICAL DOMAIN

Fibrillar collagens contain a single triple-helical domain that is ~1000 amino acids long and folds into a triple helix that is ~300 nm in length.[30] Key to the structure of the collagen triple helix is the inclusion of a glycine at every third residue in its primary structure, in repeating Gly-Xaa-Yaa triplets. Other than glycine, the amino acid content of collagen strands largely correlates with the stability conferred by each amino acid in collagen model peptides.[95] Approximately one-third of the non-glycine amino acids in the collagen triple-helical domain are proline residues, 40–50% of which are hydroxylated. Outside of glycine and proline, there is a relative abundance of charged residues, with ~40% of all Gly-Xaa-Yaa triplets containing at least one charged residue.

Negatively charged residues prefer the Xaa position, and positively charged residues prefer the Yaa position.[96] Fibrillar collagen triple helices are not known to contain cysteine or tryptophan residues.[26]

The cross-sectional view of a folded collagen triple helix shows that the obligate glycine residues are buried within the center of the helix. The other residues, particularly those in the Xaa position, are highly solvent-exposed (Fig. 1-6).[25] The three strands have a single amino acid stagger between neighboring strands, although the register of the three strands within heterotrimers is still not definitively known.[97-102] This arrangement allows for the formation of a hydrogen bond between the amine group of each glycine and the carboxyl group of the Xaa residue on the neighboring strand (Fig. 1-6). Charged residues also stabilize the triple helix by formation of intramolecular salt bridges as well as additional hydrogen bonds, in the case of Yaa-position arginines.[96, 103-105]

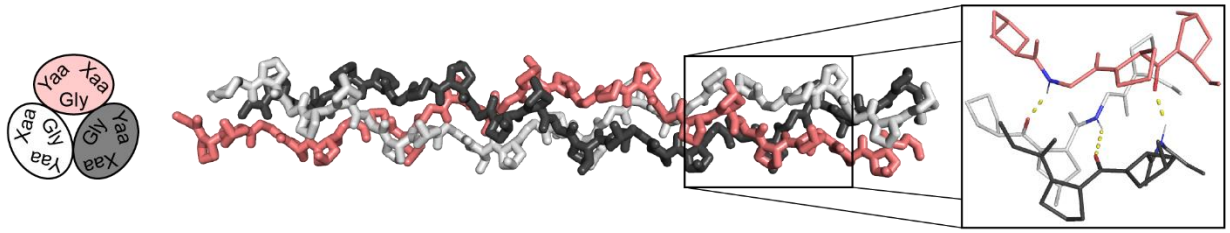


Figure 1-6: The structure of a collagen triple helix.
The collagen triple helix consists of three PPII helices wound into a triple helix. The cross-sectional view of the structure shows that the obligate glycine residues are buried inside the helix, while the Xaa and Yaa position amino acids are solvent exposed. Each Gly-Xaa-Yaa forms a hydrogen bond with each neighboring strand (PDI ID: 2G66).

The collagen triple helix is surprisingly unstable at physiological temperatures, with an equilibrium melting temperature several degrees below body temperature.[106] This instability has been observed across a variety of species with different physiological temperatures, largely correlating with hydroxyproline content, and altering physiological temperature appears to change collagen stability.[107-109] Uneven distribution of hydroxyprolines along the collagen strand also results in regions of varying stability.[110, 111] The advantages of the most abundant ECM protein being unstable are not understood, although it is hypothesized to be relevant for fibrillogenesis.[112] Interestingly, patients with a complete deficiency of the a2(I) chain due to homozygous mutations in *COL1A2*, synthesize only the a1(I) homotrimer, a more thermodynamically stable form of collagen-I. These patients exhibit severe OI[53-55] or Ehlers-Danlos syndrome (EDS),[113] suggesting that marginal stability may indeed be important for supramolecular assembly.[56, 114]

An interesting feature of the properly folded collagen triple helix is that it is resistant to commonly used proteases, including pepsin, trypsin, and chymotrypsin. This unique property is convenient in that it allows for straightforward isolation of collagen triple helices and can serve as an indirect assay for collagen folding when other biophysical methods cannot be used.[115]

The unusual amino acid content and structure of the collagen triple helix necessitates a unique ensemble of modifying enzymes, chaperones, and quality control factors for collagen folding. The expression and activity of these enzymes are often correlated with collagen expression, and disruptions to their function lead to altered collagen function.[4, 5] The following sections describe the process of triple-helical collagen folding in the ER and the proteins involved.

Modification of monomeric collagen strands. The triple-helical domain of procollagen strands is subject to extensive co- and post-translational modification prior to and during folding of the triple helix. In particular, proline and lysine residues are hydroxylated, and hydroxylysine residues can be further *O*-glycosylated.[116] The enzymes that catalyze these modifications can only associate with monomeric procollagen strands, meaning that modification must be complete before triple-helix folding.[4]

The unique structure of proline as an imino acid helps preorganize each collagen strand into PPII helices and therefore decreases the entropic cost for assembly of three PPII helices into a triple helix.[117] The majority of Yaa position prolines (~40–50% of all prolines) are hydroxylated in a 4*R* configuration (Fig. 1-7A), which confers additional, essential thermal stability to triple-helical structure via a stereoelectronic effect.[117, 118] 4*R*-Hydroxyproline favors the C^{γ}-*exo* conformation, rather than the C^{γ}-*endo* conformation preferred by unhydroxylated proline, which in turn favors the desired torsional angles for a Yaa-position amino acid. In addition, the C^{γ}-*exo* conformation allows for stabilizing $n \rightarrow \pi^*$ interactions.[117] Remarkably, inhibiting hydroxylation lowers the melting temperature of the collagen-I triple helix by ~15 °C,[119] a considerable difference, particularly given that collagen is only marginally stable at body temperature.[106]

Prolyl hydroxylation at the 3-position occurs at the Xaa position (Fig. 1-7A) and is much more rare. Initially, only a single 3(*S*)-hydroxyproline was identified in collagen-I and -II and none in type-III.[120-122] However, growing evidence, including unpublished mass spectrometry studies from our lab, suggests that there are additional Xaa sites that are 3-hydroxylated,[123-125] with a greater degree of heterogeneity than Yaa position 4-hydroxylation. The role of 3(*S*)-hydroxyproline is not fully defined, although it appears to be essential for type-IV collagen in embryonic development.[126] Recently, 4-hydroxylation was also identified at one Xaa position proline.[127]

Prolyl 4-hydroxylation and 3-hydroxylation are catalyzed by distinct prolyl-4-hydroxylases (P4Hs) and prolyl-3-hydroxylases (P3Hs), respectively (Fig. 1-7A). These enzymes are non-heme Fe(II), α-ketoglutarate-dependent dioxygenases, which require ascorbate as an essential co-factor. As a result, ascorbate is crucial for maintaining proper collagen biosynthesis, as demonstrated by scurvy, which is caused by ascorbate deficiency and results in collagen instability.[116]

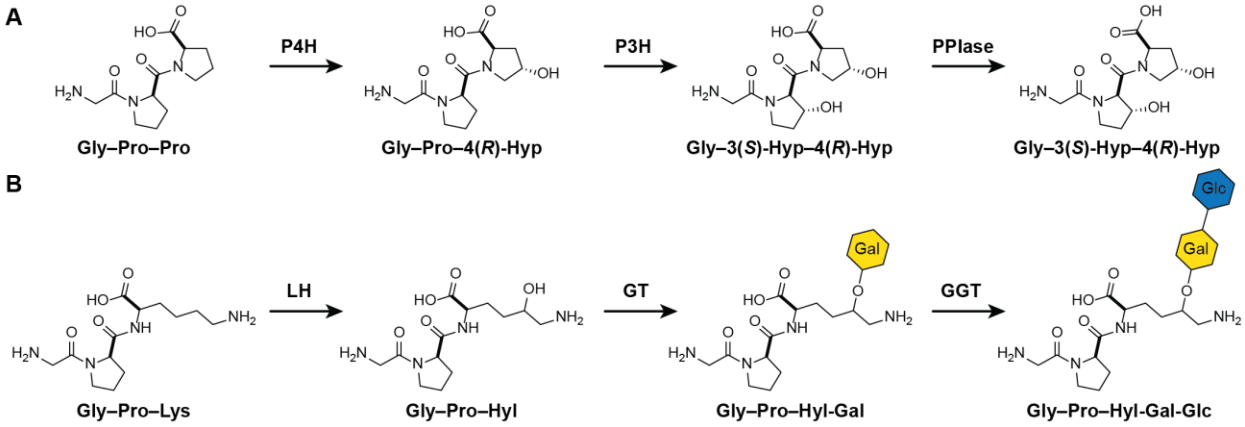


Figure 1-7: Collagen undergoes many post-translational modifications.
(A) Proline residues in the Yaa position are 4(*R)*-hydroxylated. Proline residues in the Xaa position can be 3(*S*)-hydroxylated. In order for the collagen triple helix to fold, *cis* peptide bonds must be isomerized to the *trans* configuration. (B) Lysine residues in the Yaa position can be hydroxylated. The resultant hydroxylysine residues can be further modified with sugars.

Three isoforms of P4H exist, with α(I) being the most common. Two P4H molecules assemble as the α subunits in an α2β2 tetramer, where the β subunit is PDI, another major collagen interactor previously mentioned.[128] Similarly, three isoforms of P3H exist, with P3H1 being the most well-characterized. P3H1 forms a complex with cartilage-associated protein (CRTAP) and cyclophilin B (CypB). While CRTAP is not enzymatically active, its presence is required for P3H1 activity. CypB is a well-known peptidyl-prolyl *cis-trans* isomerase (PPIase).[122] The P3H1-CRTAP-CypB complex has also been shown to have general molecular chaperone activity, separate from its P3H and PPIase activity.[129]

Lysine residues at Yaa positions and in both the N- and C-telopeptides are hydroxylated via a similar mechanism (Fig. 1-7B). Unlike prolyl hydroxylation, however, lysyl hydroxylation is much less consistent, with large variations across types and tissues.[130, 131] Lysine hydroxylation is primarily important for the formation of crosslinks in collagen fibrils.[131]

Lysyl hydroxylases (LHs) belong to the same enzymatic class as the prolyl hydroxylases and function as membrane-associated dimers.[132] The three LH isoforms have different tissue distributions and also appear to have different selectivities.[116, 131] LH1 primarily hydroxylates triple-helical lysines, while LH2 appears to primarily act on telopeptide lysines.[133, 134] LH3 appears to have selectivity for different types of collagen and is unique in that it also has glycosyltransferase activity.

Some hydroxylysine residues are further modified by the sequential addition of β-galactose and α-glucose (Fig. 1-7B). Lysyl glycosylation has been implicated in fibrillogenesis and intermolecular crosslinking.[135-137] Lysyl glycosylation is catalyzed by hydroxylysyl galactosyltransferases (GTs) and galactosylhydroxylysyl glucosyltransferase (GGTs), specifically GLT25D1, GLT25D2, and LH3. GLT25D1 and GLT25D2 appear to act as the primary GTs for collagens, while LH3 primarily functions as a GGT, with lesser GT activity.[138, 139] Interestingly, GLT25D1 inactivation led to increased collagen expression and intracellular collagen accumulation, suggesting a potential role in quality control or secretion.[140]

Zipper-like folding. Once the C-Pro domains are assembled, triple helix folding is initiated from the C-terminus. Elegant *in vitro* refolding studies performed with collagen-III revealed that folding takes place in two kinetic phases, with the first phase involving rapid folding of the ~75 most C-terminal residues, followed by a slow phase that is rate limited by *cis-trans* isomerization.[141] A similar mechanism was demonstrated with collagen-I *in cellulo*, again with *cis-trans* isomerization being the rate-limiting step.[142]

Cis-trans isomerization is particularly relevant to collagen folding due to the high proline content in the triple-helical domain. Proline residues, being imino acids, can form peptide bonds that are either *cis* or *trans*. All peptide bonds in folded collagen must be in the *trans* conformation, therefore any *cis* peptide bonds must be isomerized to the *trans* conformation for folding to proceed (Fig. 1-7A). This process is catalyzed by several PPIases, including CypB, as mentioned above. The P3H1-CRTAP-CypB complex has been shown to bind both monomeric chains, for its P3H activity, and folded collagen. Its binding to triple-helical collagen appears to have a high dissociation rate constant, thus the complex has been suggested to function at the unfolded-folded junction during triple helix folding.

The primary other PPIase known to interact with collagen is FKBP65, which functions both as a PPIase and a molecular chaperone. FKBP65 is able to bind to folded collagen and inhibit fibrillogenesis *in vitro*, though whether this occurs in the ER is not known.[143] Other members of the FKBP family of PPIases, specifically FKBP13, FKBP19, FKBP22, and FKBP23, have also been shown to interact with collagen. The different PPIases show different specificity towards

substrates containing hydroxyproline, suggesting that hydroxyproline content may modulate PPIase activity and thus folding kinetics.[144]

Maintaining proper folding kinetics is important for collagen function in the ECM. Delayed folding generally leads to overmodification of collagens, seen as a characteristic smear on SDS-PAGE gels. This effect primarily results from increased lysyl hydroxylation and hydroxylysyl glycosylation.[145] Overmodification of collagen helices can lead to defective fibrillogenesis and ECM function.[146, 147]

The folded triple helix. Once the collagen triple helix is folded, it is bound by Hsp47, a collagen-specific chaperone that recognizes Yaa position arginine residues.[148-150] Hsp47 is an ER-localized heat shock protein that primarily serves as a collagen chaperone, although it has also been shown to interact with other ECM proteins.[151-154] Unlike the collagen modifying enzymes, Hsp47 specifically binds folded, triple-helical collagen and not the individual monomers.[155, 156] Binding of Hsp47 to folded collagen is thought to not only stabilize properly folded helices but also prevent aggregation and fibrillogenesis in the ER.[157, 158] This binding is pH dependent, such that Hsp47 remains bound as folded procollagen is transported to the ER-Golgi intermediate compartment (ERGIC), where it is released upon exposure to lower pH and recycled back to the ER (Fig. 1-8).[159] However, this mechanism has recently been challenged based on live cell imaging showing that procollagen carriers that exhibit directional movement do not contain Hsp47. Instead, Hsp47 was shown to co-localize with procollagen only in the ER lumen, ERESs, and autophagic structures.[160]

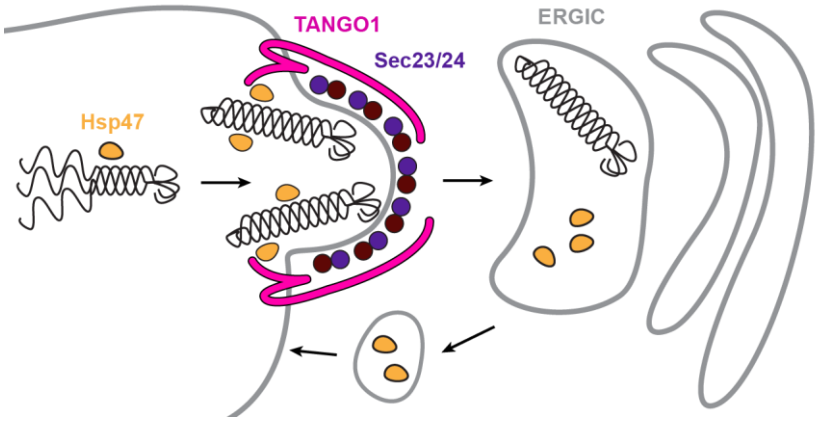


Figure 1-8: Hsp47 is involved in collagen secretion.

Hsp47 binds to folded collagen and is purportedly recognized by TANGO1. TANGO1 helps to assemble enlarged COP-II vesicles that are able to transport collagen while Hsp47 is still bound. Hsp47 is released by the lower pH in the Golgi and is recycled back to the ER.

Hsp47 is also directly implicated in ER-to-Golgi collagen transport. Due to collagen's unusual size, it utilizes a non-canonical pathway for ER-to-Golgi transport. Traditional COPII vesicles for ER-to-Golgi transport are <90 nm in diameter, compared to the ~300 nm procollagen molecule. The ER cargo receptor, TANGO1, has been shown to assemble enlarged COPII vesicles at ER exit sites (ERES) to accommodate collagen and other large proteins.[7, 12] The mechanics and generality of this process are still under investigation. TANGO1 physically interacts with Hsp47, potentially serving as a mechanism for directing properly folded collagen to COPII vesicles (Fig. 1-8).[10]

The misfolded triple helix. Up to 20–50% of wild-type procollagen has been shown to misfold and be degraded basally. When collagen proteostasis is disrupted, as is the case in disease, the misfolded or misassembled fraction is often even higher.[17-22]

Substitution of collagen glycine residues with any other residue is highly disruptive to collagen structure and folding and is the most common cause of collagenopathies, especially OI. The phenotype of autosomal dominant, OI-causing glycine substitutions depends on the nature of the substitution and its location within the triple-helical domain. Substitutions with charged or branched residues tend to be the most severe, as are those found in major ligand binding regions.[146, 161, 162] In addition, glycine substitutions closer to the C-terminus, where it is most likely to disturb triple-helix folding, as well as those located within relatively unstable, proline-poor regions tend to result in more severe phenotypes.[163, 164]

In the past two decades, a new paradigm has emerged for the collagenopathies, where autosomal recessive mutations in collagen-related proteins have been shown to present many of the same disease phenotypes as mutations in collagen itself. While 85–90% of OI cases are still caused by mutations within *COL1A1* or *COL1A2*, mutations in at least 17 other genes have now been associated with OI. These include many of the genes encoding for the collagen chaperones and modifying enzymes mentioned in this chapter, such as all members of the P3H1-CRTAP-CypB complex, Hsp47, and FKBP65, among others.[165] Mutations in these proteins lead to many of the same biosynthetic defects as traditional OI, such as overmodification, delayed or lowered secretion, and increased misfolding and intracellular accumulation.

Whether in disease or under basal conditions, misfolded or misassembled collagens must be targeted towards ER degradation pathways. Collagen has been shown to utilize several different pathways for degradation (Fig. 1-9). Major progress in identification of these pathways has been made in recent years, but there remain many open questions about how misfolded collagens are recognized and targeted towards each.

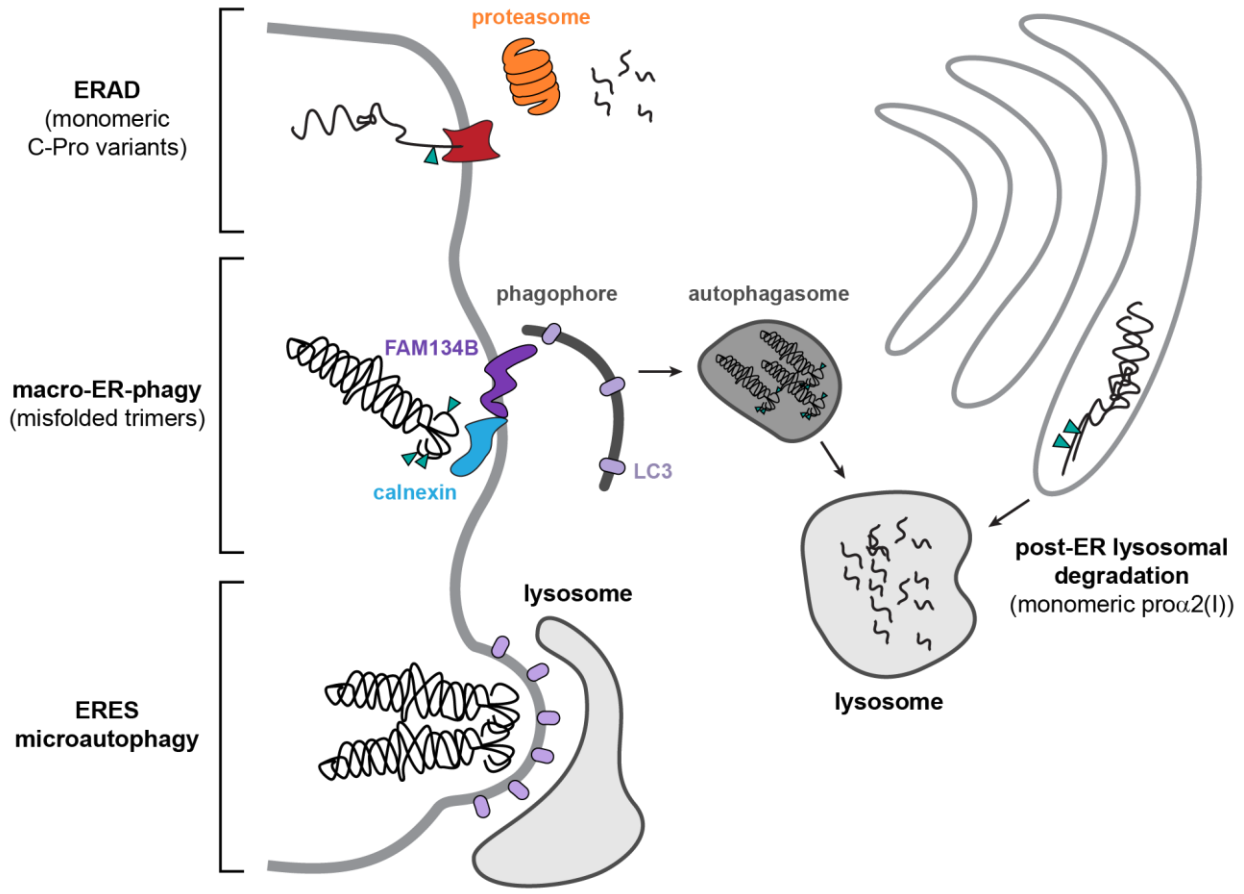


Figure 1-9: Misfolded procollagen is subject to multiple degradation pathways.
Misfolded monomeric collagen has been shown to undergo two types of degradation: ERAD, where the protein is translocated to the cytosol for degradation by the proteasome, and a post-ER lysosomal degradation pathway. Misfolded trimeric collagen has been shown to undergo macro-ER-phagy, guided by calnexin and FAM134B, or a non-canonical microautophagy pathway.

The prevailing model for procollagen degradation suggests that misfolded monomers are directed to ERAD, while misfolded trimers are degraded by one of several known lysosomal pathways, collectively called ERLAD (Fig. 1-9).[66, 166] This model arose from experiments performed using either proα1(I) with a truncated C-Pro domain that does not trimerize or proα1(I) with a glycine-to-arginine substitution that does trimerize. The former was shown to be degraded by classic ERAD, where the misfolded collagen strand is translocated out of the ER and degraded by the proteasome.[167] The latter, as well as misfolded collagen resulting from Hsp47 disruption, was instead cleared via LC3-dependent autophagy. That different types of misfolded collagen transit distinct degradation pathways is hypothesized to be due to the different structures adopted by misfolded collagen monomers and trimers. Misfolded monomers are more likely to exhibit large

hydrophobic surfaces, similar to misfolded globular proteins, that engage classic ER chaperones and quality control factors. As mentioned previously, collagen is an *N*-glycoprotein, thus the *N*-glycan and lectin chaperones are likely to be involved in recognition, as is BiP and PDI. Misfolded trimers, on the other hand, tend to form aggregates that are distinct from misfolded globular proteins and may not have hydrophobic surfaces. Thus, a different mechanism is necessary to identify and target them for degradation.[8]

Interestingly, monomeric proα2(I) that is not complexed with proα1(I) appears to be degraded by a separate pathway from misfolding proα1(I) monomers (Fig. 1-9). This pathway has not been fully explored, but it appears to be a lysosomal pathway that occurs after proα2(I) leaves the ER and enters the Golgi.[58] While Golgi quality control has been identified, it is not as well understood as ER quality control pathways.[168] Why some collagens are subject to Golgi quality control rather than ER quality control is unknown.

Two additional lysosomal degradation pathways for procollagen have been identified in recent years. Both fall under the ERLAD umbrella in that they target misfolded procollagens in the ER and direct them to lysosomes. The first of these pathways has been termed ERES micro-autophagy and has been shown to target a trimerizing glycine-to-cysteine variant of proα2(I) fused to GFP.[8, 9] Accumulation of this variant at ERES leads to the recruitment of both COPII- and autophagy-related proteins, followed by engulfment by lysosomes (Fig. 1-9). This pathway differs from traditional macroautophagy in that it does not involve formation of a phagophore and is insensitive to bafilomycin A1, which inhibits autophagosome-lysosome fusion.

The second recently discovered collagen degradation pathway has been termed macro-ER-phagy. This pathway, as described in relation to the collagen C-Pro *N*-glycan, involves the recognition of misfolded procollagen by calnexin, which then acts in conjunction with the ER-phagy receptor FAM134B to induce autophagosome formation and target misfolded proteins to the lysosome (Fig. 1-9).[82, 83, 169, 170]

Substantial progress in understanding ER-to-Golgi transport and ER degradation pathways has been made in recent years, with many of the findings involving procollagen. Many questions remained unanswered, particularly how properly folded and misfolded collagens are distinguished and targeted to the appropriate pathways. Forthcoming progress is likely to reveal the close coupling of secretion and degradation pathways and elucidate the specific molecular mechanisms behind them.

THE N-PRO DOMAIN

The N-Pro domain of collagen is the least well-characterized of the fibrillar collagen domains and the most varied amongst the different types. Broadly, the fibrillar collagen N-Pro domains fall into four classes: (1) a von Willebrand factor type-C (VWC) domain followed by a short triple helix, (2) a thrombospondin N-terminal-like (TSPN) domain followed by a short triple helix, (3) a TSPN domain without a short triple helix, and (4) a short triple helix only (Fig. 1-10A).[3, 33, 171] While the structure of a full N-Pro domain has not been solved, both the VWC domain of collagen-II (Fig. 1-10B) and the N-terminal domain of thrombospondin-1 (TSP-1) (Fig. 1-10C) have known crystal structures.[172, 173]

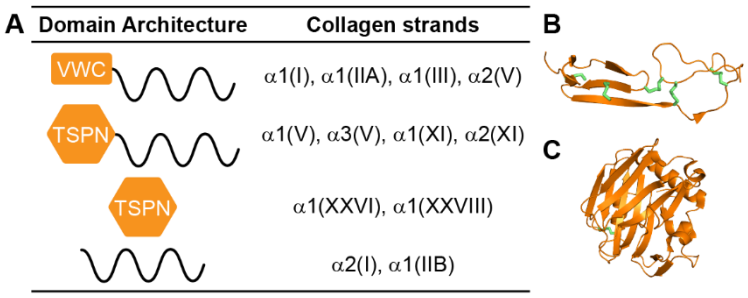


Figure 1-10: The N-Pro domains of the fibrillar collagens are highly varied.
(A) There are four broad structural categories for the fibrillar collagen N-Pro domains. (B) The von Willebrand factor type-C domain. (PDB ID: 5NIR) (C) The thrombospondin N-terminal domain. (PDB ID: 2ERF)

The proα1(IIA) VWC domain can be subdivided into three subdomains, an N-terminal β-hairpin, followed by a three-stranded antiparallel β-sheet, and a third unstructured region.[172] The VWC domain contains ten highly conserved cysteine residues that form five intrachain disulfide bonds, consistent with biophysical data showing that the N-Pro domain of collagen-I is monomeric (Fig. 1-10B).[174, 175] The N-terminal domain of TSP-1, meanwhile, has a β-sandwich structure with a single intrachain disulfide bond (Fig. 1-10C).[173] How either of these domains look in a trimeric molecule and how heterotrimeric assemblies with different N-Pro domains interact remain unknown.

Our lab recently discovered an unusual post-translational modification of aspartyl hydroxylation within the VWC domain of the N-Pro domain of proα1(I). This modification appears to be catalyzed by aspartyl-asparaginyl β-hydroxylase (ASPH), which robustly interacts with collagen-I.[75] The potential function of this modification is unknown, but mutations in ASPH have been found to cause Traboulsi syndrome, an ophthalmological disorder with some phenotypic

overlap with Marfan syndrome, a well-characterized connective tissue disorder, including facial dysmorphism.[176, 177] ASPH has also been implicated in osteogenic differentiation and senescence of bone marrow mesenchymal stem cells, supporting potential importance of the aspartyl hydroxylation site in the collagen-I N-Pro.[178]

The function of the collagen N-Pro domain remains unclear. Deletion of either the entire N-Pro domain or exon 2, which encodes the VWC domain, from proα1(I) did not affect collagen folding or secretion in Mov13 mouse fibroblasts that express proα2(I).[171, 179] It did, however, reduce secretion from Chinese hamster lung cells.[179] Even more surprisingly, deletion of exon 2 from mice resulted in no significant phenotype.[180] The mice appeared to be normal and healthy, and dermal fibroblasts isolated from them exhibited normal collagen secretion. The only notable differences between the exon 2 deleted mice and wild-type mice were a slight irregularity in fibril contour and a lower than expected generation of homozygous mutants. Cleavage of N-Proα1(II) and N-Proα1(III) has previously been suggested to regulate fibril formation,[181] but evidence is limited. Based on these results, the authors proposed that the N-Pro domain of collagen-I may have a specific role in embryonic development. They also noted that the amino acid sequence of the mouse and rat VWC domains are relatively divergent, perhaps suggesting that these findings are not broadly applicable to other species.

The N-Pro domains of various collagens have been suggested to operate as feedback regulators of collagen synthesis.[171, 182] Peptides derived from the N-Proα1(I) and N-Proα1(III) were shown to inhibit procollagen translation and synthesis in cell free, *in vitro* systems.[183, 184] In addition, cytosolic expression of N-Proα1(I) itself or an N-Pro fusion protein exhibited reduced collagen production.[185, 186] While direct evidence for cellular uptake of collagen N-Pro has not been reported, cleavage of the collagen-I N-Pro has been shown to occur intracellularly,[187] perhaps suggesting a mechanism by which the N-Pro can directly function as a translation inhibitor. Arguably more compellingly, N-Proα1(IIA) and N-Proα1(III) have been shown to bind TGFβ1 and BMP-2, both of which serve to regulate embryonic development and differentiation.[172, 188, 189] This role is particularly compelling for the N-Pro domain of collagen-II. Collagen-II is known to have two isoforms generated from alternative splicing. Collagen-IIA contains all exons, while collagen-IIB lacks exon 2, which encodes the VWC region of the N-Pro domain. Collagen-IIA is primarily expressed by chondroprogenitor cells, while collagen-IIB is primarily expressed by differentiated chondrocytes.[190] Thus, a role for the N-Proα1(IIA) in development and chondrogenesis merits further consideration.

Overall, functions of the fibrillar collagen N-Pro domains may not be generalizable. The N-Pro domains of different collagen strands are structurally diverse, and existing evidence suggests that they function differently in different cell lines. The generally normal phenotype exhibited by the collagen-I exon 2-deleted mouse model is certainly a surprising result and suggests that further work is needed to understand the function of the collagen N-Pro domains.

ER STRESS AND THE UNFOLDED PROTEIN RESPONSE

In the past few decades, there has been a new appreciation and understanding of how the cell responds to protein misfolding in the ER. The unfolded protein response (UPR) has been characterized as a highly conserved pathway that responds to protein misfolding stress in the ER. The UPR consists of three transmembrane sensors, IRE1, ATF6, and PERK, which bind to BiP under basal conditions. When BiP is titrated away by misfolded proteins, these sensors dimerize and generate downstream transcription factors that upregulate chaperones and protein quality control factors to mitigate stress. While initial UPR activation is cytoprotective, chronic UPR activation leads to apoptosis (Fig. 1-11).[23, 24, 70, 71]

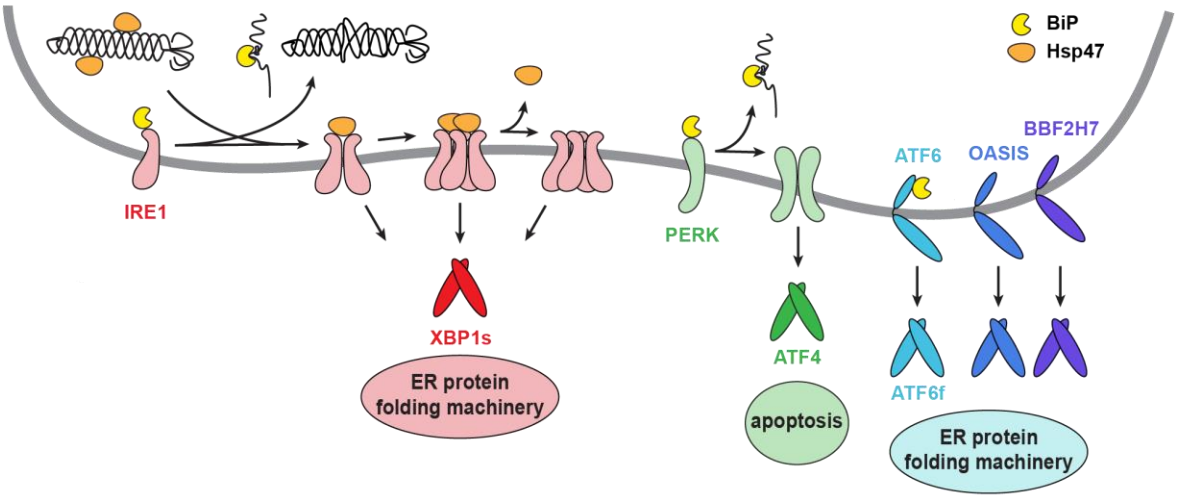


Figure 1-11: The UPR responds to misfolded proteins.
The UPR consists of three canonical stress sensors: IRE1, PERK, and ATF6. These are inactive when BiP is bound but become active when BiP is titrated away. Activation results in the formation of transcription factors that enhance ER proteostasis or signal for apoptosis. OASIS and BBF2H7 are additional stress sensors in the ATF6 family that have been implicated in connective tissue development. Hsp47 has been shown to bind IRE1 in its activated form, suggesting a model for sensing misfolding of collagen that does not involve BiP.

The interplay between collagen proteostasis and the UPR has been the subject of much recent consideration and debate.[24] Given that the classical method of UPR activation is the

titration of BiP away from the stress sensors and that BiP does not recognize all forms of misfolded collagen, particularly not misfolded triple-helical domains,[65] it is unclear whether the UPR is effectively induced by misfolded collagens.[23, 24] Recently, the collagen chaperone Hsp47 was shown to interact with IRE1, helping to displace BiP and activate downstream signaling.[191] This finding suggests an additional model by which free Hsp47 that is not bound to folded procollagen might activate the UPR in response to accumulation of misfolded collagens that are not recognized by BiP (Fig. 1-11). However, this idea has not been directly examined. In addition, the novel ER stress transducers OASIS and BBF2H7 have been shown to be involved in osteogenesis and chondrogenesis, respectively (Fig. 1-11).[192, 193] Mutations in OASIS also lead to OI,[194, 195] suggesting interplay between stress sensors and collagen proteostasis.

Chronic UPR activation has been suggested to contribute to pathology for a variety of protein folding disorders, including the collagenopathies. However, evidence for chronic UPR activation by misfolding collagen variants is weak overall. At best, there may be mutation-, collagen type-, and tissue-specific cases where chronic UPR activation does contribute to pathology.[24] That said, there is evidence suggesting that modulating ER proteostasis, via activating individual pathways within the UPR or through the use of chemical chaperones, may help ameliorate the impact of collagen mutations on collagen biosynthesis and function.[196-200]

Interestingly, a high-throughput assay for collagen secretion has suggested that the cytosolic proteostasis network also has a role in collagen production. Pan-Hsp90 inhibitors were shown to selectively reduce collagen secretion from multiple cell lines. This effect was not seen with knockdown of the ER's resident Hsp90 isoform, Grp94, suggesting that cytosolic Hsp90a and Hsp90b or mitochondrial TRAP1 are involved in collagen secretion.[201] One hypothesis for this effect is that the cytosolic Hsp90 isoforms may be required for assembly of collagen-specific transport vesicles.

Cytosolic Hsp90s are targets of the cytosolic stress response, the heat shock response (HSR). Collagen-specific chaperone Hsp47 is also an HSR target.[154] While there is some evidence that certain methods of activating the HSR also increase collagen expression levels,[202, 203] the mechanisms of this effect have not been explored. Tools for stress-independent modulation of the HSR may prove useful in determining whether additional aspects of the cytosolic proteostasis network are involved in collagen production.[70]

CONCLUSIONS

Collagen biosynthesis is an exceedingly complex process and, in many ways, unique from the folding of most globular proteins. Decades of research have addressed basic topics of

collagen production, from structure to assembly mechanism to posttranslational modification, among others. Yet, there remain many unanswered questions about the specifics of the collagen biosynthetic process and growing evidence of how complex it can be.

In this thesis, we discuss newly discovered or newly appreciated contributions of the collagen C-Pro domain to collagen assembly and proteostasis. In Chapter 2, we describe the discovery that the C-Proα1(I) *N*-glycan is critical for maintaining proper collagen biosynthesis when folding is challenged. These findings are particularly pertinent in the face of newly discovered pathways for collagen quality control. In Chapter 3, we explore the mechanism of collagen homo- and heterotrimerization, providing new insights that suggest the importance of factors outside the C-Pro domain in the chain selection process. Finally, we conclude with additional perspectives on the field and discussion of the challenges associated with biochemical collagen research.

REFERENCES

1. Bar-On, Y. M., Milo, R., The global mass and average rate of rubisco. *Proc. Natl. Acad. Sci. U. S. A.* 116, 4738-4743 (2019).
2. Mouw, J. K., Ou, G., Weaver, V. M., Extracellular matrix assembly: a multiscale deconstruction. *Nat. Rev. Mol. Cell Biol.* 15, 771-785 (2014).
3. Ricard-Blum, S., The collagen family. *Cold Spring. Harb. Perspect. Biol.* 3, a004978 (2011).
4. Ishikawa, Y., Bachinger, H. P., A molecular ensemble in the rER for procollagen maturation. *Biochim. Biophys. Acta* 1833, 2479-2491 (2013).
5. Ito, S., Nagata, K., Quality control of procollagen in cells. *Annu. Rev. Biochem.* 90, 631-658 (2021).
6. Boudko, S. P., Engel, J., Bachinger, H. P., The crucial role of trimerization domains in collagen folding. *Int. J. Biochem. Cell Biol.* 44, 21-32 (2012).
7. Malhotra, V., Erlmann, P., The pathway of collagen secretion. *Annu. Rev. Cell. Dev. Biol.* 31, 109-124 (2015).
8. Omari, S., Makareeva, E., Roberts-Pilgrim, A., Mirigian, L., Jarnik, M., Ott, C., Lippincott-Schwartz, J., Leikin, S., Noncanonical autophagy at ER exit sites regulates procollagen turnover. *Proc. Natl. Acad. Sci. U.S.A.* 115, E10099-E10108 (2018).
9. Omari, S., Makareeva, E., Roberts-Pilgrim, A., Mirigian, L., Jarnik, M., Leikin, S., Lippincott-Schwartz, J., Noncanonical autophagy at ER exit sites in osteoblasts. *J. Bone Miner. Res.* 34, 230-230 (2019).
10. Ishikawa, Y., Ito, S., Nagata, K., Sakai, L. Y., Bachinger, H. P., Intracellular mechanisms of molecular recognition and sorting for transport of large extracellular matrix molecules. *Proc. Natl. Acad. Sci. U. S. A.* 113, E6036-E6044 (2016).
11. McCaughey, J., Stephens, D. J., ER-to-Golgi Transport: A Sizeable Problem. *Trends Cell Biol.* 29, 940-953 (2019).
12. Saito, K., Chen, M., Bard, F., Chen, S., Zhou, H., Woodley, D., Polischuk, R., Schekman, R., Malhotra, V., TANGO1 facilitates cargo loading at endoplasmic reticulum exit sites. *Cell* 136, 891-902 (2009).
13. McCaughey, J., Stevenson, N. L., Cross, S., Stephens, D. J., ER-to-Golgi trafficking of procollagen in the absence of large carriers. *J. Cell Biol.* 218, 929-948 (2019).
14. Raote, I., Saxena, S., Campelo, F., Malhotra, V., TANGO1 marshals the early secretory pathway for cargo export. *Biochim Biophys Acta Biomembr* 1863, 183700 (2021).
15. Canty, E. G., Kadler, K. E., Procollagen trafficking, processing and fibrillogenesis. *J. Cell Sci.* 118, 1341-1353 (2005).
16. Morris, N. P., Fessler, L. I., Weinstock, A., Fessler, J. H., Procollagen assembly and secretion in embryonic chick bone. *J. Biol. Chem.* 250, 5719-5726 (1975).
17. Berg, R. A., Schwartz, M. L., Crystal, R. G., Regulation of the production of secretory proteins: intracellular degradation of newly synthesized "defective" collagen. *Proc. Natl. Acad. Sci. U. S. A.* 77, 4746-4750 (1980).
18. Rennard, S. I., Stier, L. E., Crystal, R. G., Intracellular degradation of newly synthesized collagen. *J. Invest. Dermatol.* 79 Suppl 1, 77-82 (1982).
19. Bienkowski, R. S., Intracellular degradation of newly synthesized collagen. *Rev. Biol. Cellular* 21, 423-443 (1989).
20. Bienkowski, R. S., Collagen degradation in human lung fibroblasts: extent of degradation, role of lysosomal proteases, and evaluation of an alternate hypothesis. *J. Cell. Physiol.* 121, 152-158 (1984).
21. Bienkowski, R. S., Baum, B. J., Crystal, R. G., Fibroblasts degrade newly synthesized collagen within the cell before secretion. *Nature* 276, 413-416 (1978).

22. Bienkowski, R. S., Curran, S. F., Berg, R. A., Kinetics of intracellular degradation of newly synthesized collagen. *Biochemistry* 25, 2455-2459 (1986).
23. Wong, M. Y., Shoulders, M. D., Targeting defective proteostasis in the collagenopathies. *Curr. Opin. Chem. Biol.* 50, 80-88 (2019).
24. Bateman, J. F., Shoulders, M. D., Lamande, S. R., Collagen misfolding mutations: the contribution of the unfolded protein response to the molecular pathology. *Connect. Tissue Res.* 63, 210-227 (2022).
25. Bella, J., Eaton, M., Brodsky, B., Berman, H. M., Crystal and molecular structure of a collagen-like peptide at 1.9 A resolution. *Science* 266, 75-81 (1994).
26. Ramshaw, J. A., Shah, N. K., Brodsky, B., Gly-X-Y tripeptide frequencies in collagen: a context for host-guest triple-helical peptides. *J. Struct. Biol.* 122, 86-91 (1998).
27. Nallanthighal, S., Heiserman, J. P., Cheon, D. J., Collagen Type XI Alpha 1 (COL11A1): A Novel Biomarker and a Key Player in Cancer. *Cancers (Basel)* 13 (2021).
28. Hoffman, G. G., Branam, A. M., Huang, G., Pelegri, F., Cole, W. G., Wenstrup, R. M., Greenspan, D. S., Characterization of the six zebrafish clade B fibrillar procollagen genes, with evidence for evolutionarily conserved alternative splicing within the pro-alpha1(V) C-propeptide. *Matrix Biol.* 29, 261-275 (2010).
29. Mak, K. M., Png, C. Y., Lee, D. J., Type V Collagen in Health, Disease, and Fibrosis. *Anat Rec (Hoboken)* 299, 613-629 (2016).
30. Bella, J., Hulmes, D. J. S., Fibrillar collagens. *Subcell. Biochem.* 82, 457-490 (2017).
31. Burgeson, R. E., Hollister, D. W., Collagen heterogeneity in human cartilage: identification of several new collagen chains. *Biochem. Biophys. Res. Commun.* 87, 1124-1131 (1979).
32. Furuto, D. K., Miller, E. J., Different levels of glycosylation contribute to the heterogeneity of alpha 1(II) collagen chains derived from a transplantable rat chondrosarcoma. *Arch. Biochem. Biophys.* 226, 604-611 (1983).
33. Exposito, J. Y., Valcourt, U., Cluzel, C., Lethias, C., The fibrillar collagen family. *Int. J. Mol. Sci.* 11, 407-426 (2010).
34. Engel, J., Prockop, D. J., The zipper-like folding of collagen triple helices and the effects of mutations that disrupt the zipper. *Annu. Rev. Biophys. Biophys. Chem.* 20, 137-152 (1991).
35. Tkocz, C., Kühn, K., The formation of triple-helical collagen molecules from alpha-1 or alpha-2 polypeptide chains. *Eur. J. Biochem.* 7, 454-462 (1969).
36. Beier, G., Engel, J., The renaturation of soluble collagen. Products formed at different temperatures. *Biochemistry* 5, 2744-2755 (1966).
37. Bellamy, G., Bornstein, P., Evidence for procollagen, a biosynthetic precursors of collagen. *Proc. Natl. Acad. Sci. U. S. A.* 68, 1138-1142 (1971).
38. Bourhis, J. M., Mariano, N., Zhao, Y., Harlos, K., Exposito, J. Y., Jones, E. Y., Moali, C., Aghajari, N., Hulmes, D. J., Structural basis of fibrillar collagen trimerization and related genetic disorders. *Nat. Struct. Mol. Biol.* 19, 1031-1036 (2012).
39. Sharma, U., Carrique, L., Vadon-Le Goff, S., Mariano, N., Georges, R. N., Delolme, F., Koivunen, P., Myllyharju, J., Moali, C., Aghajari, N., Hulmes, D. J., Structural basis of homo- and heterotrimerization of collagen I. *Nat. Commun.* 8, 14671-14681 (2017).
40. Crouch, E., Bornstein, P., Collagen synthesis by human amniotic fluid cells in culture: characterization of a procollagen with three identical proalpha1(I) chains. *Biochemistry* 17, 5499-5509 (1978).
41. Gay, S., Martin, G. R., Muller, P. K., Timpl, R., Kuhn, K., Simultaneous synthesis of types I and III collagen by fibroblasts in culture. *Proc. Natl. Acad. Sci. U. S. A.* 73, 4037-4040 (1976).
42. Sandberg, M. M., Hirvonen, H. E., Elima, K. J., Vuorio, E. I., Co-expression of collagens II and XI and alternative splicing of exon 2 of collagen II in several developing human tissues. *Biochem. J.* 294 (Pt 2), 595-602 (1993).

43. Wu, J. J., Weis, M. A., Kim, L. S., Carter, B. G., Eyre, D. R., Differences in chain usage and cross-linking specificities of cartilage type V/XI collagen isoforms with age and tissue. *J. Biol. Chem.* 284, 5539-5545 (2009).
44. Mayne, R., Brewton, R. G., Mayne, P. M., Baker, J. R., Isolation and characterization of the chains of type V/type XI collagen present in bovine vitreous. *J. Biol. Chem.* 268, 9381-9386 (1993).
45. Kleman, J. P., Hartmann, D. J., Ramirez, F., van der Rest, M., The human rhabdomyosarcoma cell line A204 lays down a highly insoluble matrix composed mainly of alpha 1 type-XI and alpha 2 type-V collagen chains. *Eur. J. Biochem.* 210, 329-335 (1992).
46. Lees, J. F., Tasab, M., Bulleid, N. J., Identification of the molecular recognition sequence which determines the type-specific assembly of procollagen. *EMBO J.* 16, 908-916 (1997).
47. DiChiara, A. S., Li, R. C., Suen, P. H., Hosseini, A. S., Taylor, R. J., Weickhardt, A. F., Malhotra, D., McCaslin, D. R., Shoulders, M. D., A cysteine-based molecular code informs collagen C-propeptide assembly. *Nat. Commun.* 9, 4206 (2018).
48. Uitto, J., Collagen polymorphism: Isolation and partial characterization of α1(I)-trimer molecules in normal human skin. *Arch. Biochem. Biophys.* 192, 371-379 (1979).
49. Jimenez, S. A., Bashey, R. I., Benditt, M., Yankowski, R., Identification of collagen α1(I) trimer in embryonic chick tendons and calvaria. *Biochem. Biophys. Res. Commun.* 78, 1354-1361 (1977).
50. Makareeva, E., Han, S., Vera, J. C., Sackett, D. L., Holmbeck, K., Phillips, C. L., Visse, R., Nagase, H., Leikin, S., Carcinomas contain a matrix metalloproteinase-resistant isoform of type I collagen exerting selective support to invasion. *Cancer Res.* 70, 4366-4374 (2010).
51. Moro, L., Smith, B. D., Identification of collagen α1(I) trimer and normal type I collagen in a polyoma virus-induced mouse tumor. *Arch. Biochem. Biophys.* 182, 33-41 (1977).
52. Bailey, A. J., Sims, T. J., Knott, L., Phenotypic expression of osteoblast collagen in osteoarthritic bone: Production of type I homotrimer. *Int. J. Biochem. Cell Biol.* 34, 176-182 (2002).
53. Nicholls, A. C., Osse, G., Schloon, H. G., Lenard, H. G., Deak, S., Myers, J. C., Prockop, D. J., Weigel, W. R., Fryer, P., Pope, F. M., The clinical features of homozygous alpha 2(I) collagen deficient osteogenesis imperfecta. *J. Med. Genet.* 21, 257-262 (1984).
54. Pihlajaniemi, T., Dickson, L. A., Pope, F. M., Korhonen, V. R., Nicholls, A., Prockop, D. J., Myers, J. C., Osteogenesis imperfecta: cloning of a pro-alpha 2(I) collagen gene with a frameshift mutation. *J. Biol. Chem.* 259, 12941-12944 (1984).
55. Pace, J. M., Wiese, M., Drenguis, A. S., Kuznetsova, N., Leikin, S., Schwarze, U., Chen, D., Mooney, S. H., Unger, S., Byers, P. H., Defective C-propeptides of the proα2(I) chain of type I procollagen impede molecular assembly and result in osteogenesis imperfecta. *J. Biol. Chem.* 283, 16061-16067 (2008).
56. Kuznetsova, N. V., McBride, D. J., Leikin, S., Changes in thermal stability and microunfolding pattern of collagen helix resulting from the loss of alpha2(I) chain in osteogenesis imperfecta murine. *J. Mol. Biol.* 331, 191-200 (2003).
57. Dion, A. S., Myers, J. C., COOH-terminal propeptides of the major human procollagens. Structural, functional and genetic comparisons. *J. Mol. Biol.* 193, 127-143 (1987).
58. Gotkin, M. G., Ripley, C. R., Lamande, S. R., Bateman, J. F., Bienkowski, R. S., Intracellular trafficking and degradation of unassociated proalpha2 chains of collagen type I. *Exp. Cell Res.* 296, 307-316 (2004).
59. Lees, J. F., Bulleid, N. J., The role of cysteine residues in the folding and association of the COOH-terminal propeptide of type-I and type-III procollagen. *J. Biol. Chem.* 269, 24354-24360 (1994).

60. Bulleid, N. J., Wilson, R., Lees, J. F., Type-III procollagen assembly in semi-intact cells: chain association, nucleation and triple-helix folding do not require formation of inter-chain disulphide bonds but triple-helix nucleation does require hydroxylation. *Biochem. J.* 317 (Pt 1), 195-202 (1996).
61. Uitto, J., Collagen polymorphism: isolation and partial characterization of alpha 1(I)-trimer molecules in normal human skin. *Arch. Biochem. Biophys.* 192, 371-379 (1979).
62. Roulet, M., Valkkila, M., Chanut-Delalande, H., Hamalainen, E. R., Kessler, E., Ala-Kokko, L., Mannikko, M., Bonod-Bidaud, C., Ruggiero, F., The collagen V homotrimer [alpha1(V)](3) production is unexpectedly favored over the heterotrimer [alpha1(V)](2)alpha2(V) in recombinant expression systems. *J. Biomed. Biotechnol.* 2010, 376927 (2010).
63. Symoens, S., Hulmes, D. J. S., Bourhis, J. M., Coucke, P. J., De Paepe, A., Malfait, F., Type I procollagen C-propeptide defects: study of genotype–phenotype correlation and predictive role of crystal structure. *Hum. Mutat.* 35, 1330-1341 (2014).
64. Barnes, A. M., Ashok, A., Makareeva, E. N., Brusel, M., Cabral, W. A., Weis, M., Moali, C., Bettler, E., Eyre, D. R., Cassella, J. P., Leikin, S., Hulmes, D. J. S., Kessler, E., Marini, J. C., COL1A1 C-propeptide mutations cause ER mislocalization of procollagen and impair C-terminal procollagen processing. *Biochim. Biophys. Acta Mol. Basis Dis.* 1865, 2210-2223 (2019).
65. Chessler, S. D., Wallis, G. A., Byers, P. H., Mutations in the carboxyl-terminal propeptide of the proα1(I) chain of type I collagen results in defective chain association and produce lethal osteogenesis imperfecta. *J. Biol. Chem.* 268, 18218-18225 (1993).
66. Lamandé, S. R., Chessler, S. D., Golub, S. B., Byers, P. H., Chan, D., Cole, W. G., Sillence, D. O., Bateman, J. F., Endoplasmic reticulum-mediated quality-control of type-I collagen production by cells from osteogenesis imperfecta patients with mutations in the pro-alpha-1(I) chain carboxyl-terminal propeptide which impair subunit assembly. *J. Biol. Chem.* 270, 8642-8649 (1995).
67. Pace, J. M., Chitayat, D., Atkinson, M., Wilcox, W. R., Schwarze, U., Byers, P. H., A single amino acid substitution (D1441Y) in the carboxyl-terminal propeptide of the proα(1) chain of type I collagen results in a lethal variant of osteogenesis imperfecta with features of dense bone diseases. *J. Med. Genet.* 39, 23-29 (2002).
68. Pace, J. M., Kuslich, C. D., Willing, M. C., Byers, P. H., Disruption of one intra-chain disulphide bond in the carboxyl-terminal propeptide of the proα1(I) chain of type I procollagen permits slow assembly and secretion of overmodified, but stable procollagen trimers and results in mild osteogenesis imperfecta. *J. Med. Genet.* 38, 443-449 (2001).
69. Doan, N. D., Hosseini, A. S., Bikovtseva, A. A., Huang, M. S., DiChiara, A. S., Papa, L. J., 3rd, Koller, A., Shoulders, M. D., Elucidation of proteostasis defects caused by osteogenesis imperfecta mutations in the collagen-α2(I) C-propeptide domain. *J. Biol. Chem.* 295, 9959-9973 (2020).
70. Sebastian, R. M., Shoulders, M. D., Chemical biology framework to illuminate proteostasis. *Annu. Rev. Biochem.* 89, 529-555 (2020).
71. Wong, M. Y., DiChiara, A. S., Suen, P. H., Chen, K., Doan, N. D., Shoulders, M. D., Adapting secretory proteostasis and function through the unfolded protein response. *Curr. Top. Microbiol. Immunol.* 414, 1-25 (2018).
72. Koivu, J., Myllyla, R., Interchain disulfide bond formation in types I and II procollagen. Evidence for a protein disulfide isomerase catalyzing bond formation. *J. Biol. Chem.* 262, 6159-6164 (1987).
73. Wilson, R., Lees, J. F., Bulleid, N. J., Protein disulfide isomerase acts as a molecular chaperone during the assembly of procollagen. *J. Biol. Chem.* 273, 9637-9643 (1998).

74. Bottomley, M. J., Batten, M. R., Lumb, R. A., Bulleid, N. J., Quality control in the endoplasmic reticulum: PDI mediates the ER retention of unassembled procollagen C-propeptides. *Curr. Biol.* 11, 1114-1118 (2001).
75. DiChiara, A. S., Taylor, R. J., Wong, M. Y., Doan, N. D., Rosario, A. M., Shoulders, M. D., Mapping and exploring the collagen-I proteostasis network. *ACS Chem. Biol.* 11, 1408-1421 (2016).
76. Hebert, D. N., Lamriben, L., Powers, E. T., Kelly, J. W., The intrinsic and extrinsic effects of N-linked glycans on glycoproteostasis. *Nat. Chem. Biol.* 10, 902-910 (2014).
77. Helenius, A., Aebi, M., Roles of N-linked glycans in the endoplasmic reticulum. *Annu. Rev. Biochem.* 73, 1019-1049 (2004).
78. Clark, C. C., The distribution and initial characterization of oligosaccharide units on the COOH-terminal propeptide extensions of the pro-α1 and pro-α2 chains of type I procollagen. *J. Biol. Chem.* 254, 10798-10802 (1979).
79. Olsen, B. R., Guzman, N. A., Engel, J., Condit, C., Aase, S., Purification and characterization of a peptide from the carboxy-terminal region of chick tendon procollagen type I. *Biochemistry* 16, 3030-3036 (1977).
80. Li, R. C., Wong, M. Y., DiChiara, A. S., Hosseini, A. S., Shoulders, M. D., Collagen's enigmatic, highly conserved *N*-glycan has an essential proteostatic function. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2026608118 (2021).
81. Lamandé, S. R., Bateman, J. F., The type I collagen proα1(I) COOH-terminal propeptide N-linked oligosaccharide: functional analysis by site-directed mutagenesis. *J. Biol. Chem.* 270, 17858-17865 (1995).
82. Forrester, A., De Leonibus, C., Grumati, P., Fasana, E., Piemontese, M., Staiano, L., Fregno, I., Raimondi, A., Marazza, A., Bruno, G., Iavazzo, M., Intartaglia, D., Seczynska, M., van Anken, E., Conte, I., De Matteis, M. A., Dikic, I., Molinari, M., Settembre, C., A selective ER-phagy exerts procollagen quality control via a calnexin-FAM134B complex. *EMBO J.* 38, e99847 (2019).
83. Fregno, I., Fasana, E., Solda, T., Galli, C., Molinari, M., *N*-Glycan processing selects ERAD-resistant misfolded proteins for ER-to-lysosome-associated degradation. *EMBO J.* 40, e107240 (2021).
84. Van Duyn Graham, L., Sweetwyne, M. T., Pallero, M. A., Murphy-Ullrich, J. E., Intracellular calreticulin regulates multiple steps in fibrillar collagen expression, trafficking, and processing into the extracellular matrix. *J. Biol. Chem.* 285, 7067-7078 (2010).
85. Beck, K., Boswell, B. A., Ridgway, C. C., Bachinger, H. P., Triple helix formation of procollagen type I can occur at the rough endoplasmic reticulum membrane. *J. Biol. Chem.* 271, 21566-21573 (1996).
86. Stefanovic, B., Stefanovic, L., Manojlovic, Z., Imaging of type I procollagen biosynthesis in cells reveals biogenesis in highly organized bodies; Collagenosomes. *Matrix Biol Plus* 12, 100076 (2021).
87. McLaughlin, S. H., Bulleid, N. J., Molecular recognition in procollagen chain assembly. *Matrix Biol.* 16, 369-377 (1998).
88. Cui, Y. X., Parashar, S., Zahoor, M., Needham, P. G., Mari, M., Zhu, M., Chen, S. L., Ho, H. C., Reggiori, F., Farhan, H., Brodsky, J. L., Ferro-Novick, S., A COPII subunit acts with an autophagy receptor to target endoplasmic reticulum for degradation. *Science* 365, 53-60 (2019).
89. Davis, S., Wang, J., Zhu, M., Stahmer, K., Lakshminarayan, R., Ghassemian, M., Jiang, Y., Miller, E. A., Ferro-Novick, S., Sec24 phosphorylation regulates autophagosome abundance during nutrient deprivation. *eLife* 5, e21167 (2016).
90. Li, Z., Huang, W., Wang, W., Multifaceted roles of COPII subunits in autophagy. *Biochim. Biophys. Acta Mol. Cell. Res.* 1867, 118627 (2020).

91. Shima, T., Kirisako, H., Nakatogawa, H., COPII vesicles contribute to autophagosomal membranes. *J. Cell Biol.* 218, 1503-1510 (2019).
92. Graef, M., Friedman, J. R., Graham, C., Babu, M., Nunnari, J., ER exit sites are physical and functional core autophagosome biogenesis components. *Mol. Biol. Cell* 24, 2918-2931 (2013).
93. Ge, L., Melville, D., Zhang, M., Schekman, R., The ER-Golgi intermediate compartment is a key membrane source for the LC3 lipidation step of autophagosome biogenesis. *eLife* 2, e00947 (2013).
94. Ge, L., Zhang, M., Kenny, S. J., Liu, D., Maeda, M., Saito, K., Mathur, A., Xu, K., Schekman, R., Remodeling of ER-exit sites initiates a membrane supply pathway for autophagosome biogenesis. *EMBO Rep* 18, 1586-1603 (2017).
95. Persikov, A. V., Ramshaw, J. A., Kirkpatrick, A., Brodsky, B., Amino acid propensities for the collagen triple-helix. *Biochemistry* 39, 14960-14967 (2000).
96. Salem, G., Traub, W., Conformational implications of amino acid sequence regularities in collagen. *FEBS Lett.* 51, 94-99 (1975).
97. Jalan, A. A., Sammon, D., Hartgerink, J. D., Brear, P., Stott, K., Hamaia, S. W., Hunter, E. J., Walker, D. R., Leitinger, B., Farndale, R. W., Chain alignment of collagen I deciphered using computationally designed heterotrimers. *Nat. Chem. Biol.* 16, 423–429 (2020).
98. Traub, W., Fietzek, P. P., Contribution of the alpha2 chain to the molecular stability of collagen. *FEBS Lett.* 68, 245-249 (1976).
99. Malone, J. P., Alvares, K., Veis, A., Structure and assembly of the heterotrimeric and homotrimeric C-propeptides of type I collagen: significance of the alpha2(I) chain. *Biochemistry* 44, 15269-15279 (2005).
100. Piez, K. A., Trus, B. L., Sequence regularities and packing of collagen molecules. *J. Mol. Biol.* 122, 419-432 (1978).
101. Bender, E., Silver, F. H., Hayashi, K., Trelstad, R. L., Type I collagen segment long spacing banding patterns. Evidence that the alpha 2 chain is in the reference or A position. *J. Biol. Chem.* 257, 9653-9657 (1982).
102. Orgel, J. P., Irving, T. C., Miller, A., Wess, T. J., Microfibrillar structure of type I collagen in situ. *Proc. Natl. Acad. Sci. U. S. A.* 103, 9001-9005 (2006).
103. Katz, E. P., David, C. W., Energetics of intrachain salt-linkage formation in collagen. *Biopolymers* 29, 791-798 (1990).
104. Venugopal, M. G., Ramshaw, J. A., Braswell, E., Zhu, D., Brodsky, B., Electrostatic interactions in collagen-like triple-helical peptides. *Biochemistry* 33, 7948-7956 (1994).
105. Yang, W., Chan, V. C., Kirkpatrick, A., Ramshaw, J. A., Brodsky, B., Gly-Pro-Arg confers stability similar to Gly-Pro-Hyp in the collagen triple-helix of host-guest peptides. *J. Biol. Chem.* 272, 28837-28840 (1997).
106. Leikina, E., Mertts, M. V., Kuznetsova, N., Leikin, S., Type I collagen is thermally unstable at body temperature. *Proc. Natl. Acad. Sci. U. S. A.* 99, 1314-1318 (2002).
107. Privalov, P. L., Stability of proteins. Proteins which do not present a single cooperative system. *Adv. Protein Chem.* 35, 1-104 (1982).
108. Duan, R., Konno, K., Zhang, J. J., Wang, S. J., Yuan, C. H., Different Thermostability of Collagens from Scale of Carp (Cyprinus Carpio) in Winter and Summer. *J. Food Biochem.* 34, 1275-1287 (2010).
109. Fujii, K. K., Taga, Y., Takagi, Y. K., Masuda, R., Hattori, S., Koide, T., The Thermal Stability of the Collagen Triple Helix Is Tuned According to the Environmental Temperature. *Int. J. Mol. Sci.* 23 (2022).
110. Makareeva, E., Cabral, W. A., Marini, J. C., Leikin, S., Molecular mechanism of alpha 1(I)-osteogenesis imperfecta/Ehlers-Danlos syndrome: unfolding of an N-anchor domain at the N-terminal end of the type I collagen triple helix. *J. Biol. Chem.* 281, 6463-6470 (2006).

111. Ryhanen, L., Zaragoza, E. J., Uitto, J., Conformational stability of type I collagen triple helix: evidence for temporary and local relaxation of the protein conformation using a proteolytic probe. *Arch. Biochem. Biophys.* 223, 562-571 (1983).
112. Kadler, K. E., Hojima, Y., Prockop, D. J., Assembly of type I collagen fibrils de novo. Between 37 and 41 degrees C the process is limited by micro-unfolding of monomers. *J. Biol. Chem.* 263, 10517-10523 (1988).
113. Malfait, F., Symoens, S., Coucke, P., Nunes, L., De Almeida, S., De Paepe, A., Total absence of the alpha2(I) chain of collagen type I causes a rare form of Ehlers-Danlos syndrome with hypermobility and propensity to cardiac valvular problems. *J. Med. Genet.* 43, e36 (2006).
114. Chang, S. W., Shefelbine, S. J., Buehler, M. J., Structural and mechanical differences between collagen homo- and heterotrimers: relevance for the molecular origin of brittle bone disease. *Biophys. J.* 102, 640-648 (2012).
115. Bruckner, P., Prockop, D. J., Proteolytic enzymes as probes for the triple-helical conformation of procollagen. *Anal. Biochem.* 110, 360-368 (1981).
116. Salo, A. M., Myllyharju, J., Prolyl and lysyl hydroxylases in collagen synthesis. *Exp. Dermatol.* 30, 38-49 (2021).
117. Shoulders, M. D., Raines, R. T., Collagen structure and stability. *Annu. Rev. Biochem.* 78, 929-958 (2009).
118. Holmgren, S. K., Taylor, K. M., Bretscher, L. E., Raines, R. T., Code for collagen's stability deciphered. *Nature* 392, 666-667 (1998).
119. Berg, R. A., Prockop, D. J., The thermal transition of a non-hydroxylated form of collagen. Evidence for a role for hydroxyproline in stabilizing the triple-helix of collagen. *Biochem. Biophys. Res. Commun.* 52, 115-120 (1973).
120. Fietzek, P. P., Rexrodt, F. W., Wendt, P., Stark, M., Kuhn, K., The covalent structure of collagen. Amino-acid sequence of peptide 1-CB6-C2. *Eur. J. Biochem.* 30, 163-168 (1972).
121. Tryggvason, K., Risteli, J., Kivirikko, K. I., Separation of prolyl 3-hydroxylase and 4-hydroxylase activities and the 4-hydroxyproline requirement for synthesis of 3-hydroxyproline. *Biochem. Biophys. Res. Commun.* 76, 275-281 (1976).
122. Morello, R., Bertin, T. K., Chen, Y. Q., Hicks, J., Tonachini, L., Monticone, M., Castagnola, P., Rauch, F., Glorieux, F. H., Vranka, J., Bachinger, H. P., Pace, J. M., Schwarze, U., Byers, P. H., Weis, M., Fernandes, R. J., Eyre, D. R., Yao, Z. Q., Boyce, B. F., Lee, B., CRTAP is required for prolyl 3-hydroxylation and mutations cause recessive osteogenesis imperfecta. *Cell* 127, 291-304 (2006).
123. Weis, M. A., Hudson, D. M., Kim, L., Scott, M., Wu, J. J., Eyre, D. R., Location of 3-hydroxyproline residues in collagen types I, II, III, and V/XI implies a role in fibril supramolecular assembly. *J. Biol. Chem.* 285, 2580-2590 (2010).
124. Montgomery, N. T., Zientek, K. D., Pokidysheva, E. N., Bachinger, H. P., Post-translational modification of type IV collagen with 3-hydroxyproline affects its interactions with glycoprotein VI and nidogens 1 and 2. *J. Biol. Chem.* 293, 5987-5999 (2018).
125. Kirchner, M., Deng, H., Xu, Y., Heterogeneity in proline hydroxylation of fibrillar collagens observed by mass spectrometry. *PLoS One* 16, e0250544 (2021).
126. Pokidysheva, E., Boudko, S., Vranka, J., Zientek, K., Maddox, K., Moser, M., Fassler, R., Ware, J., Bachinger, H. P., Biological role of prolyl 3-hydroxylation in type IV collagen. *Proc. Natl. Acad. Sci. U. S. A.* 111, 161-166 (2014).
127. van Huizen, N. A., Burgers, P. C., Saintmont, F., Brocorens, P., Gerbaux, P., Stingl, C., Dekker, L. J. M., JNM, I. J., Luider, T. M., Identification of 4-Hydroxyproline at the Xaa Position in Collagen by Mass Spectrometry. *J. Proteome Res.* 18, 2045-2051 (2019).
128. Gorres, K. L., Raines, R. T., Prolyl 4-hydroxylase. *Crit. Rev. Biochem. Mol. Biol.* 45, 106-124 (2010).

129. Ishikawa, Y., Wirz, J., Vranka, J. A., Nagata, K., Bachinger, H. P., Biochemical characterization of the prolyl 3-hydroxylase 1.cartilage-associated protein.cyclophilin B complex. *J. Biol. Chem.* 284, 17641-17647 (2009).
130. Uzawa, K., Yeowell, H. N., Yamamoto, K., Mochida, Y., Tanzawa, H., Yamauchi, M., Lysine hydroxylation of collagen in a fibroblast cell culture system. *Biochem. Biophys. Res. Commun.* 305, 484-487 (2003).
131. Yamauchi, M., Sricholpech, M., Lysine post-translational modifications of collagen. *Essays Biochem.* 52, 113-133 (2012).
132. Suokas, M., Myllyla, R., Kellokumpu, S., A single C-terminal peptide segment mediates both membrane association and localization of lysyl hydroxylase in the endoplasmic reticulum. *J. Biol. Chem.* 275, 17863-17868 (2000).
133. Uzawa, K., Grzesik, W. J., Nishiura, T., Kuznetsov, S. A., Robey, P. G., Brenner, D. A., Yamauchi, M., Differential expression of human lysyl hydroxylase genes, lysine hydroxylation, and cross-linking of type I collagen during osteoblastic differentiation in vitro. *J. Bone Miner. Res.* 14, 1272-1280 (1999).
134. van der Slot, A. J., Zuurmond, A. M., Bardoel, A. F., Wijmenga, C., Pruijs, H. E., Sillence, D. O., Brinckmann, J., Abraham, D. J., Black, C. M., Verzijl, N., DeGroot, J., Hanemaaijer, R., TeKoppele, J. M., Huizinga, T. W., Bank, R. A., Identification of PLOD2 as telopeptide lysyl hydroxylase, an important enzyme in fibrosis. *J. Biol. Chem.* 278, 40967-40972 (2003).
135. Amudeswari, S., Liang, J. N., Chakrabarti, B., Polar-apolar characteristics and fibrillogenesis of glycosylated collagen. *Coll. Relat. Res.* 7, 215-223 (1987).
136. Notbohm, H., Nokelainen, M., Myllyharju, J., Fietzek, P. P., Muller, P. K., Kivirikko, K. I., Recombinant human type II collagens with low and high levels of hydroxylysine and its glycosylated forms show marked differences in fibrillogenesis in vitro. *J. Biol. Chem.* 274, 8988-8992 (1999).
137. Eyre, D. R., Glimcher, M. J., Analysis of a crosslinked peptide from calf bone collagen: evidence that hydroxylysyl glycoside participates in the crosslink. *Biochem. Biophys. Res. Commun.* 52, 663-671 (1973).
138. Schegg, B., Hulsmeier, A. J., Rutschmann, C., Maag, C., Hennet, T., Core glycosylation of collagen is initiated by two beta(1-O)galactosyltransferases. *Mol. Cell. Biol.* 29, 943-952 (2009).
139. Sricholpech, M., Perdivara, I., Nagaoka, H., Yokoyama, M., Tomer, K. B., Yamauchi, M., Lysyl hydroxylase 3 glucosylates galactosylhydroxylysine residues in type I collagen in osteoblast culture. *J. Biol. Chem.* 286, 8846-8856 (2011).
140. Baumann, S., Hennet, T., Collagen Accumulation in Osteosarcoma Cells lacking GLT25D1 Collagen Galactosyltransferase. *J. Biol. Chem.* 291, 18514-18524 (2016).
141. Bachinger, H. P., Bruckner, P., Timpl, R., Prockop, D. J., Engel, J., Folding mechanism of the triple helix in type-III collagen and type-III pN-collagen. Role of disulfide bridges and peptide bond isomerization. *Eur. J. Biochem.* 106, 619-632 (1980).
142. Bruckner, P., Eikenberry, E. F., Prockop, D. J., Formation of the triple helix of type I procollagen in cellulo. A kinetic model based on cis-trans isomerization of peptide bonds. *Eur. J. Biochem.* 118, 607-613 (1981).
143. Ishikawa, Y., Vranka, J., Wirz, J., Nagata, K., Bachinger, H. P., The rough endoplasmic reticulum-resident FK506-binding protein FKBP65 is a molecular chaperone that interacts with collagens. *J. Biol. Chem.* 283, 31584-31590 (2008).
144. Ishikawa, Y., Mizuno, K., Bachinger, H. P., Ziploc-ing the structure 2.0: Endoplasmic reticulum-resident peptidyl prolyl isomerases show different activities toward hydroxyproline. *J. Biol. Chem.* 292, 9273-9282 (2017).

145. Bateman, J. F., Mascara, T., Chan, D., Cole, W. G., Abnormal type I collagen metabolism by cultured fibroblasts in lethal perinatal osteogenesis imperfecta. *Biochem. J.* 217, 103-115 (1984).
146. Forlino, A., Cabral, W. A., Barnes, A. M., Marini, J. C., New perspectives on osteogenesis imperfecta. *Nat. Rev. Endocrin.* 7, 540-557 (2011).
147. Torre-Blanco, A., Adachi, E., Hojima, Y., Wootton, J. A., Minor, R. R., Prockop, D. J., Temperature-induced post-translational over-modification of type I procollagen. Effects of over-modification of the protein on the rate of cleavage by procollagen N-proteinase and on self-assembly of collagen into fibrils. *J. Biol. Chem.* 267, 2650-2655 (1992).
148. Koide, T., Nishikawa, Y., Asada, S., Yamazaki, C. M., Takahara, Y., Homma, D. L., Otaka, A., Ohtani, K., Wakamiya, N., Nagata, K., Kitagawa, K., Specific recognition of the collagen triple helix by chaperone HSP47. II. The HSP47-binding structural motif in collagens and related proteins. *J. Biol. Chem.* 281, 11177-11185 (2006).
149. Koide, T., Takahara, Y., Asada, S., Nagata, K., Xaa-Arg-Gly triplets in the collagen triple helix are dominant binding sites for the molecular chaperone HSP47. *J. Biol. Chem.* 277, 6178-6182 (2002).
150. Widmer, C., Gebauer, J. M., Brunstein, E., Rosenbaum, S., Zaucke, F., Drogemuller, C., Leeb, T., Baumann, U., Molecular basis for the action of the collagen-specific chaperone Hsp47/SERPINH1 and its structure-specific client recognition. *Proc. Natl. Acad. Sci. U. S. A.* 109, 13243-13247 (2012).
151. Ishikawa, Y., Rubin, K., Bachinger, H. P., Kalamajski, S., The endoplasmic reticulum-resident collagen chaperone Hsp47 interacts with and promotes the secretion of decorin, fibromodulin, and lumican. *J. Biol. Chem.* 293, 13707-13716 (2018).
152. Hirayoshi, K., Kudo, H., Takechi, H., Nakai, A., Iwamatsu, A., Yamada, K. M., Nagata, K., HSP47: a tissue-specific, transformation-sensitive, collagen-binding heat shock protein of chicken embryo fibroblasts. *Mol. Cell. Biol.* 11, 4036-4044 (1991).
153. Ito, S., Nagata, K., Roles of the endoplasmic reticulum-resident, collagen-specific molecular chaperone Hsp47 in vertebrate cells and human disease. *J. Biol. Chem.* 294, 2133-2141 (2019).
154. Nagata, K., Saga, S., Yamada, K. M., A major collagen-binding protein of chick embryo fibroblasts is a novel heat shock protein. *J. Cell Biol.* 103, 223-229 (1986).
155. Ono, T., Miyazaki, T., Ishida, Y., Uehata, M., Nagata, K., Direct in vitro and in vivo evidence for interaction between Hsp47 protein and collagen triple helix. *J. Biol. Chem.* 287, 6810-6818 (2012).
156. Tasab, M., Batten, M. R., Bulleid, N. J., Hsp47: a molecular chaperone that interacts with and stabilizes correctly-folded procollagen. *EMBO J.* 19, 2204-2211 (2000).
157. Ishida, Y., Kubota, H., Yamamoto, A., Kitamura, A., Bachinger, H. P., Nagata, K., Type I collagen in Hsp47-null cells is aggregated in endoplasmic reticulum and deficient in N-propeptide processing and fibrillogenesis. *Mol. Biol. Cell* 17, 2346-2355 (2006).
158. Thomson, C. A., Ananthanarayanan, V. S., Structure-function studies on hsp47: pH-dependent inhibition of collagen fibril formation in vitro. *Biochem. J.* 349 Pt 3, 877-883 (2000).
159. Satoh, M., Hirayoshi, K., Yokota, S., Hosokawa, N., Nagata, K., Intracellular interaction of collagen-specific stress protein HSP47 with newly synthesized procollagen. *J. Cell Biol.* 133, 469-483 (1996).
160. Omari, S., Makareeva, E., Gorrell, L., Jarnik, M., Lippincott-Schwartz, J., Leikin, S., Mechanisms of procollagen and HSP47 sorting during ER-to-Golgi trafficking. *Matrix Biol.* 93, 79-94 (2020).
161. Forlino, A., Marini, J. C., Osteogenesis imperfecta. *Lancet* 387, 1657-1671 (2016).
162. Marini, J. C., Forlino, A., Cabral, W. A., Barnes, A. M., San Antonio, J. D., Milgrom, S., Hyland, J. C., Korkko, J., Prockop, D. J., De Paepe, A., Coucke, P., Symoens, S.,

Glorieux, F. H., Roughley, P. J., Lund, A. M., Kurila-Svahn, K., Hartikka, H., Cohn, D. H., Krakow, D., Mottes, M., Schwarze, U., Chen, D., Yang, K., Kuslich, C., Troendle, J., Dalgleish, R., Byers, P. H., Consortium for osteogenesis imperfecta mutations in the helical domain of type I collagen: Regions rich in lethal mutations align with collagen binding sites for integrins and proteoglycans. *Hum. Mut.* 28, 209-221 (2007).

163. Makareeva, E., Mertz, E. L., Kuznetsova, N. V., Sutter, M. B., DeRidder, A. M., Cabral, W. A., Barnes, A. M., McBride, D. J., Marini, J. C., Leikin, S., Structural heterogeneity of type I collagen triple helix and its role in osteogenesis imperfecta. *J. Biol. Chem.* 283, 4787-4798 (2008).
164. Bateman, J. F., Moeller, I., Hannagan, M., Chan, D., Cole, W. G., Characterization of three osteogenesis imperfecta collagen alpha 1(I) glycine to serine mutations demonstrating a position-dependent gradient of phenotypic severity. *Biochem. J.* 288 (Pt 1), 131-135 (1992).
165. Marini, J. C., Dang Do, A. N., "Osteogenesis Imperfecta" in Endotext*,* Feingold, K. R., Anawalt, B., Boyce, A., Chrousos, G., de Herder, W. W., Dhatariya, K., Dungan, K., Hershman, J. M., Hofland, J., Kalra, S., Kaltsas, G., Koch, C., Kopp, P., Korbonits, M., Kovacs, C. S., Kuohung, W., Laferrere, B., Levy, M., McGee, E. A., McLachlan, R., Morley, J. E., New, M., Purnell, J., Sahay, R., Singer, F., Sperling, M. A., Stratakis, C. A., Trence, D. L., Wilson, D. P., Eds. (South Dartmouth (MA), 2000).
166. Ishida, Y., Yamamoto, A., Kitamura, A., Lamandé, S. R., Yoshimori, T., Bateman, J. F., Kubota, H., Nagata, K., Autophagic elimination of misfolded procollagen aggregates in the endoplasmic reticulum as a means of cell protection. *Mol. Biol. Cell* 20, 2744-2754 (2009).
167. Fitzgerald, J., Lamandé, S. R., Bateman, J. F., Proteasomal degradation of unassembled mutant type I collagen pro-α1(I) chains. *J. Biol. Chem.* 274, 27392-27398 (1999).
168. Sun, Z., Brodsky, J. L., Protein quality control in the secretory pathway. *J. Cell Biol.* 218, 3171-3187 (2019).
169. Reggio, A., Buonomo, V., Berkane, R., Bhaskara, R. M., Tellechea, M., Peluso, I., Polishchuk, E., Di Lorenzo, G., Cirillo, C., Esposito, M., Hussain, A., Huebner, A. K., Hubner, C. A., Settembre, C., Hummer, G., Grumati, P., Stolz, A., Role of FAM134 paralogues in endoplasmic reticulum remodeling, ER-phagy, and collagen quality control. *EMBO Rep.* 22, e52289 (2021).
170. Cinque, L., De Leonibus, C., Iavazzo, M., Krahmer, N., Intartaglia, D., Salierno, F. G., De Cegli, R., Di Malta, C., Svelto, M., Lanzara, C., Maddaluno, M., Wanderlingh, L. G., Huebner, A. K., Cesana, M., Bonn, F., Polishchuk, E., Hubner, C. A., Conte, I., Dikic, I., Mann, M., Ballabio, A., Sacco, F., Grumati, P., Settembre, C., MiT/TFE factors control ER-phagy via transcriptional regulation of FAM134B. *EMBO J.* 39, e105696 (2020).
171. Bornstein, P., The NH(2)-terminal propeptides of fibrillar collagens: highly conserved domains with poorly understood functions. *Matrix Biol.* 21, 217-226 (2002).
172. Xu, E. R., Blythe, E. E., Fischer, G., Hyvonen, M., Structural analyses of von Willebrand factor C domains of collagen 2A and CCN3 reveal an alternative mode of binding to bone morphogenetic protein-2. *J. Biol. Chem.* 292, 12516-12527 (2017).
173. Tan, K., Duquette, M., Liu, J. H., Zhang, R., Joachimiak, A., Wang, J. H., Lawler, J., The structures of the thrombospondin-1 N-terminal domain and its complex with a synthetic pentameric heparin. *Structure* 14, 33-42 (2006).
174. Engel, J., Bruckner, P., Becker, U., Timpl, R., Rutschmann, B., Physical properties of the amino-terminal precursor-specific portion of type I procollagen. *Biochemistry* 16, 4026-4033 (1977).
175. Misenheimer, T. M., Huwiler, K. G., Annis, D. S., Mosher, D. F., Physical characterization of the procollagen module of human thrombospondin 1 expressed in insect cells. *J. Biol. Chem.* 275, 40938-40945 (2000).

176. Patel, N., Khan, A. O., Mansour, A., Mohamed, J. Y., Al-Assiri, A., Haddad, R., Jia, X., Xiong, Y., Megarbane, A., Traboulsi, E. I., Alkuraya, F. S., Mutations in ASPH cause facial dysmorphism, lens dislocation, anterior-segment abnormalities, and spontaneous filtering blebs, or Traboulsi syndrome. *Am. J. Hum. Genet.* 94, 755-759 (2014).
177. Kulkarni, N., Lloyd, I. C., Ashworth, J., Biswas, S., Black, G. C. M., Clayton-Smith, J., Consortium, N. B., Traboulsi syndrome due to ASPH mutation: an under-recognised cause of ectopia lentis. *Clin. Dysmorphol.* 28, 184-189 (2019).
178. Peng, H., Guo, Q., Xiao, Y., Su, T., Jiang, T. J., Guo, L. J., Wang, M., ASPH Regulates Osteogenic Differentiation and Cellular Senescence of BMSCs. *Front Cell Dev Biol* 8, 872 (2020).
179. Lee, S. T., Lee, S., Peters, D. P., Hoffman, G. G., Stacey, A., Greenspan, D. S., Deletion of the pro-alpha 1(I) N-propeptide affects secretion of type I collagen in Chinese hamster lung cells but not in Mov-13 mouse cells. *J. Biol. Chem.* 267, 24126-24133 (1992).
180. Bornstein, P., Walsh, V., Tullis, J., Stainbrook, E., Bateman, J. F., Hormuzdi, S. G., The globular domain of the proalpha 1(I) N-propeptide is not required for secretion, processing by procollagen N-proteinase, or fibrillogenesis of type I collagen in mice. *J. Biol. Chem.* 277, 2605-2613 (2002).
181. Fleischmajer, R., Timpl, R., Tuderman, L., Raisher, L., Wiestner, M., Perlish, J. S., Graves, P. N., Ultrastructural identification of extension aminopropeptides of type I and III collagens in human skin. *Proc. Natl. Acad. Sci. U. S. A.* 78, 7360-7364 (1981).
182. Wiestner, M., Krieg, T., Horlein, D., Glanville, R. W., Fietzek, P., Muller, P. K., Inhibiting effect of procollagen peptides on collagen biosynthesis in fibroblast cultures. *J. Biol. Chem.* 254, 7016-7023 (1979).
183. Paglia, L., Wilczek, J., de Leon, L. D., Martin, G. R., Horlein, D., Muller, P., Inhibition of procollagen cell-free synthesis by amino-terminal extension peptides. *Biochemistry* 18, 5030-5034 (1979).
184. Horlein, D., McPherson, J., Goh, S. H., Bornstein, P., Regulation of protein synthesis: translational control by procollagen-derived fragments. *Proc. Natl. Acad. Sci. U. S. A.* 78, 6163-6167 (1981).
185. Fouser, L., Sage, E. H., Clark, J., Bornstein, P., Feedback regulation of collagen gene expression: a Trojan horse approach. *Proc. Natl. Acad. Sci. U. S. A.* 88, 10158-10162 (1991).
186. Oganesian, A., Au, S., Horst, J. A., Holzhausen, L. C., Macy, A. J., Pace, J. M., Bornstein, P., The NH2-terminal propeptide of type I procollagen acts intracellularly to modulate cell function. *J. Biol. Chem.* 281, 38507-38518 (2006).
187. Stevenson, N. L., Bergen, D. J. M., Lu, Y., Prada-Sanchez, M. E., Kadler, K. E., Hammond, C. L., Stephens, D. J., Giantin is required for intracellular N-terminal processing of type I procollagen. *J. Cell Biol.* 220 (2021).
188. Brisson, B. K., Stewart, D. C., Burgwin, C., Chenoweth, D., Wells, R. G., Adams, S. L., Volk, S. W., Cysteine-rich domain of type III collagen N-propeptide inhibits fibroblast activation by attenuating TGFb signaling. *Matrix Biol.* 10.1016/j.matbio.2022.03.004 (2022).
189. Zhu, Y., Oganesian, A., Keene, D. R., Sandell, L. J., Type IIA procollagen containing the cysteine-rich amino propeptide is deposited in the extracellular matrix of prechondrogenic tissue and binds to TGF-beta1 and BMP-2. *J. Cell Biol.* 144, 1069-1080 (1999).
190. McAlinden, A., Smith, T. A., Sandell, L. J., Ficheux, D., Parry, D. A., Hulmes, D. J., Alpha-helical coiled-coil oligomerization domains are almost ubiquitous in the collagen superfamily. *J. Biol. Chem.* 278, 42200-42207 (2003).
191. Sepulveda, D., Rojas-Rivera, D., Rodriguez, D. A., Groenendyk, J., Kohler, A., Lebeaupin, C., Ito, S., Urra, H., Carreras-Sureda, A., Hazari, Y., Vasseur-Cognet, M., Ali, M. M. U., Chevet, E., Campos, G., Godoy, P., Vaisar, T., Bailly-Maitre, B., Nagata, K., Michalak, M.,

Sierralta, J., Hetz, C., Interactome Screening Identifies the ER Luminal Chaperone Hsp47 as a Regulator of the Unfolded Protein Response Transducer IRE1alpha. *Mol. Cell* 69, 238-252 e237 (2018).
192. Ishikawa, T., Toyama, T., Nakamura, Y., Tamada, K., Shimizu, H., Ninagawa, S., Okada, T., Kamei, Y., Ishikawa-Fujiwara, T., Todo, T., Aoyama, E., Takigawa, M., Harada, A., Mori, K., UPR transducer BBF2H7 allows export of type II collagen in a cargo- and developmental stage-specific manner. *J. Cell Biol.* 216, 1761-1774 (2017).
193. Murakami, T., Saito, A., Hino, S., Kondo, S., Kanemoto, S., Chihara, K., Sekiya, H., Tsumagari, K., Ochiai, K., Yoshinaga, K., Saitoh, M., Nishimura, R., Yoneda, T., Kou, I., Furuichi, T., Ikegawa, S., Ikawa, M., Okabe, M., Wanaka, A., Imaizumi, K., Signalling mediated by the endoplasmic reticulum stress transducer OASIS is involved in bone formation. *Nat. Cell Biol.* 11, 1205-1211 (2009).
194. Keller, R. B., Tran, T. T., Pyott, S. M., Pepin, M. G., Savarirayan, R., McGillivray, G., Nickerson, D. A., Bamshad, M. J., Byers, P. H., Monoallelic and biallelic CREB3L1 variant causes mild and severe osteogenesis imperfecta, respectively. *Genet. Med.* 20, 411-419 (2018).
195. Symoens, S., Malfait, F., D'Hondt, S., Callewaert, B., Dheedene, A., Steyaert, W., Bachinger, H. P., De Paepe, A., Kayserili, H., Coucke, P. J., Deficiency for the ER-stress transducer OASIS causes severe recessive osteogenesis imperfecta in humans. *Orphanet J. Rare Dis.* 8, 154 (2013).
196. Besio, R., Iula, G., Garibaldi, N., Cipolla, L., Sabbioneda, S., Biggiogera, M., Marini, J. C., Rossi, A., Forlino, A., 4-PBA ameliorates cellular homeostasis in fibroblasts from osteogenesis imperfecta patients by enhancing autophagy and stimulating protein secretion. *Biochim. Biophys. Acta Mol. Basis. Dis.* 1864, 1642-1652 (2018).
197. Takeyari, S., Kubota, T., Ohata, Y., Fujiwara, M., Kitaoka, T., Taga, Y., Mizuno, K., Ozono, K., 4-Phenylbutyric acid enhances the mineralization of osteogenesis imperfecta iPSC-derived osteoblasts. *J. Biol. Chem.* 296, 100027 (2021).
198. Wang, D., Mohammad, M., Wang, Y., Tan, R., Murray, L. S., Ricardo, S., Dagher, H., van Agtmael, T., Savige, J., The Chemical Chaperone, PBA, Reduces ER Stress and Autophagy and Increases Collagen IV alpha5 Expression in Cultured Fibroblasts From Men With X-Linked Alport Syndrome and Missense Mutations. *Kidney Int Rep* 2, 739-748 (2017).
199. DiChiara, A. S., Doan, N. D., Bikovtseva, A. A., Rowly, L., Butty, V. L., Weis, M. E., Eyre, D. R., Lamandé, S. R., Bateman, J. F., Shoulders, M. D., XBP1s-mediated endoplasmic reticulum proteostasis network enhancement can selectively improve folding and secretion of an osteogenesis imperfecta-causing collagen-I variant. *bioRxiv* 10.1101/2021.04.15.439909 (2021).
200. Garibaldi, N., Contento, B. M., Babini, G., Morini, J., Siciliani, S., Biggiogera, M., Raspanti, M., Marini, J. C., Rossi, A., Forlino, A., Besio, R., Targeting cellular stress in vitro improves osteoblast homeostasis, matrix collagen content and mineralization in two murine models of osteogenesis imperfecta. *Matrix Biol.* 98, 1-20 (2021).
201. Wong, M. Y., Doan, N. D., DiChiara, A. S., Papa, L. J., 3rd, Cheah, J. H., Soule, C. K., Watson, N., Hulleman, J. D., Shoulders, M. D., A High-Throughput Assay for Collagen Secretion Suggests an Unanticipated Role for Hsp90 in Collagen Production. *Biochemistry* 57, 2814-2827 (2018).
202. Dams, S. D., de Liefde-van Beest, M., Nuijs, A. M., Oomens, C. W., Baaijens, F. P., Heat shocks enhance procollagen type I and III expression in fibroblasts in ex vivo human skin. *Skin Res. Technol.* 17, 167-180 (2011).
203. Nishiwaki, H., Ito, M., Negishi, S., Sobue, S., Ichihara, M., Ohno, K., Molecular hydrogen upregulates heat shock response and collagen biosynthesis, and downregulates cell cycles: meta-analyses of gene expression profiles. *Free Radic. Res.* 52, 434-445 (2018).

CHAPTER 2:

Collagen's Enigmatic, Highly Conserved *N*-Glycan has an Essential Proteostatic Function

This chapter has been adapted from:

Li, R.C., Wong, M.Y., DiChiara, A.S., Hosseini, A.S., Shoulders, M.D. Collagen's enigmatic, highly conserved *N*-glycan has an essential proteostatic function. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2026608118 (2021).

AUTHOR CONTRIBUTIONS

R.C.L., M.Y.W., A.S.D., and M.D.S. designed research; R.C.L., M.Y.W., A.S.D., A.S.H., and M.D.S. performed research; R.C.L., M.Y.W., A.S.D., and A.S.H. contributed new reagents/analytic tools; R.C.L., M.Y.W., A.S.D., A.S.H., and M.D.S. analyzed data, and R.C.L., M.Y.W., and M.D.S. wrote the paper.

ACKNOWLEDGEMENTS

The authors thank Prof. Robert Sauer (MIT) for experimental suggestions and Dr. Charles Whittaker (MIT) for helpful discussions regarding generation of sequence alignments. This work was supported by the National Institutes of Health (NIH Grant 1R01AR071443) and a Research Grant from the G. Harold and Leila Y. Mathers Foundation (both to M.D.S.). R.C.L. and M.Y.W. were supported by National Science Foundation Graduate Research Fellowships. A.S.D. was supported by a National Institutes of Health Ruth L. Kirschstein Predoctoral Fellowship (1F31AR067615). The work was also supported in part by the MIT Center for Environmental Health Sciences through NIH Grant P30-ES002109.

ABSTRACT

Intracellular collagen folding begins at the protein's C-terminal propeptide (C-Pro) domain, which initiates triple-helix assembly and defines the composition and chain register of fibrillar collagen trimers. The C-Pro domain is later proteolytically cleaved and excreted from the body, while the mature triple helix is incorporated into the extracellular matrix. The collagen C-Pro domain possesses a single *N*-glycosylation site that is widely conserved in all the fibrillar collagens across humans and diverse other species. Given that the C-Pro domain is removed once collagen folding is complete, the *N*-glycan might be presumed to be important for folding. Surprisingly, however, there is no difference in the folding and secretion of *N*-glycosylated versus non-*N*-glycosylated collagen type-I, leaving the function of the *N*-glycan unclear. We hypothesized that the collagen *N*-glycan might have a context-dependent function, specifically, that it could be required to promote collagen folding only when proteostasis is challenged. We show that removal of the *N*-glycan from misfolding-prone C-Pro domain variants does indeed cause serious collagen and ER proteostasis defects. The *N*-glycan promotes folding and secretion of destabilized C-Pro variants by providing access to the ER's lectin-based chaperone machinery. Finally, we show that the C-Pro *N*-glycan is actually critical for the folding and secretion of even wild-type collagen under ER stress conditions. Such stress is commonly incurred during development, wound healing, and other processes where collagen production plays a key role. Collectively, these results establish an essential, context-dependent function for collagen's previously enigmatic *N*-glycan, wherein the carbohydrate moiety buffers collagen folding against proteostatic challenge.

INTRODUCTION

As the molecular scaffold for animal life, the various types of collagen perform diverse structural and biological functions.[1] Fibrillar collagens, the most abundant collagen class, consist of N- and C-propeptide (N-Pro and C-Pro, respectively) domains sandwiching a lengthy, continuous triple-helical domain (Fig. 2-1A).[2] Both the N-Pro and C-Pro domains are proteolytically cleaved upon fibrillar procollagen secretion, such that only the triple-helical domain is deposited in the mature extracellular matrix.[3] The function of the N-Pro domain remains largely unclear. The C-Pro domain, in contrast, has a well-established role in templating triple helix assembly, controlling the composition and stoichiometry of triple helices, and setting the register of the triple helix.[4-8]

The collagen-I C-Pro domain (~30 kDa in its monomeric form) is globular, cysteine-rich, calcium-binding, and *N*-glycosylated on a single asparagine residue (Fig. 2-1A; N1365 in *COL1A1* or N1267 in *COL1A2*). *N*-glycosylation of the C-Pro domain was first observed nearly fifty years ago.[9, 10] A recent crystal structure of the human collagen-I homotrimer established that the assembled C-Pro adopts a structure analogous to a flower.[7] Mapping a model *N*-glycan onto the crystal structure of the Colα1(I) C-Pro homotrimer suggests that the *N*-glycan is positioned on the outer face of the β-sheets that form each “petal” and is likely directed out into the solvent (Fig. 2-1*B*). This *N*-glycosylation site is conserved across nearly all fibrillar collagen types and domains of life (Figs. 2-1C and 2-1D). Even in instances where the specific site is not conserved, an *N*-glycosylation sequon is still present at least once in the C-Pro domain.

The energetic cost and high conservation of C-Pro *N*-glycosylation[11] argues for an essential biological function. Outside the cell, *N*-glycans can mediate protein–protein interactions, stabilize and/or organize extracellular matrices, and provide protection against proteases and other enzymes.[12] However, the *N*-glycan-containing C-Pro domain is cleaved from procollagen prior to extracellular matrix assembly, indicating that the *N*-glycan is unlikely to have a significant extracellular function.

N-Glycans can also have critical functions inside the cell, largely by providing access to the extensive lectin-based proteostasis network in the endoplasmic reticulum (ER) and thereby extrinsically enhancing protein folding and quality control.[13-15] In particular, the lectin chaperones calreticulin and calnexin are known to interact with collagen-I.[16, 17] Knockdown of calreticulin can reduce collagen secretion by mouse embryonic fibroblasts, and calnexin appears to be involved in directing misassembled collagen to ER-phagy.[18, 19] *N*-Glycosylation can also intrinsically modulate protein folding and stability via specific interactions with local amino acids and/or via entropic effects.[20-23]

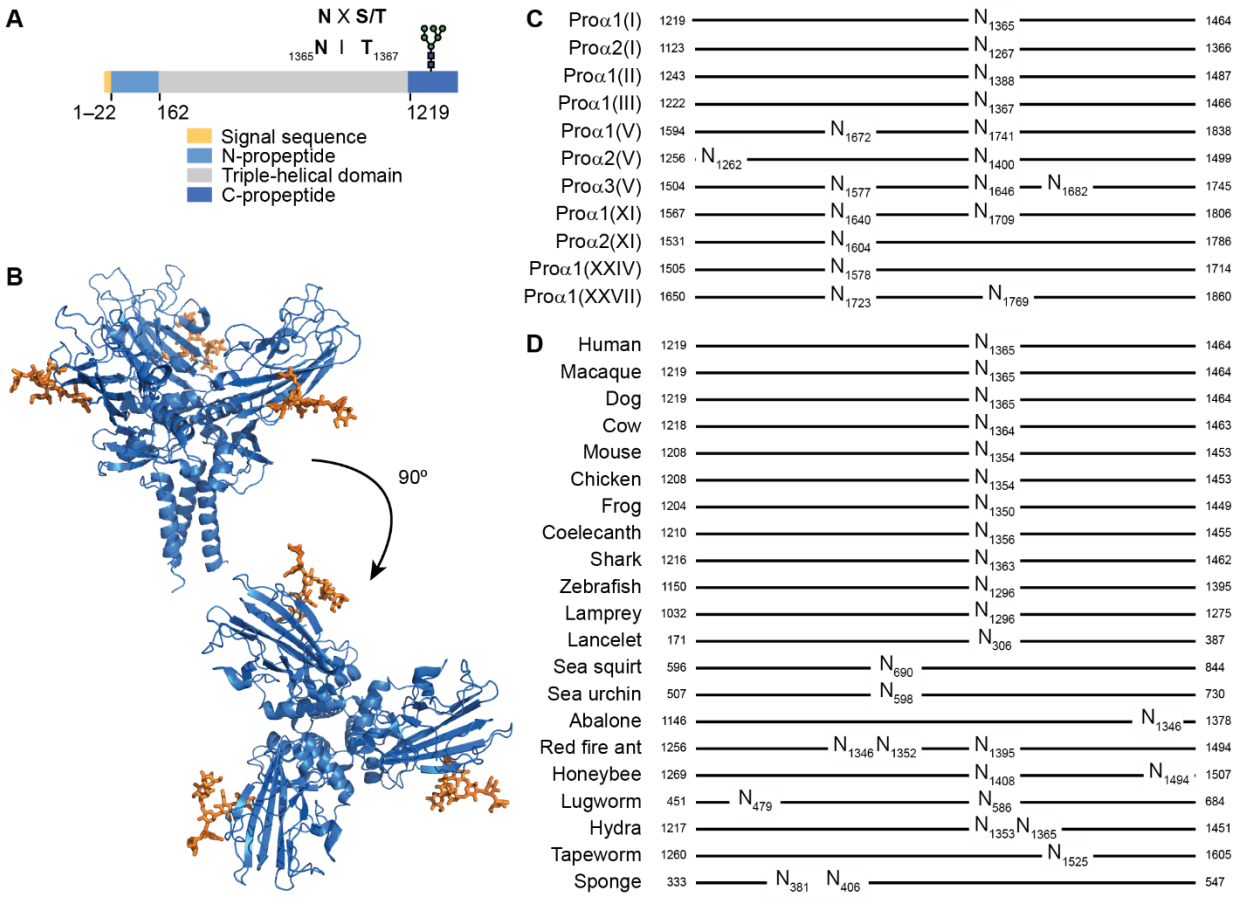


Figure 2-1: The *N*-glycosylation motif within the collagen-I C-Pro domain is highly conserved.

(A) Schematic of the Colα1(I) sequence showing the *N*-glycosylation sequon located within the C-terminal propeptide (C-Pro) domain. (B) Two views of a representative *N*-glycan (from PDB ID: 1S4P) modeled onto the structure of the Colα1(I) C-Pro homotrimer (PDB ID: 5K31). (C) Alignment of the C-Pro amino acid sequences for all human fibrillar collagens illustrates the strong conservation of the *N*-glycosylation sequon. (D) Alignment of C-Proα1(I) (or equivalent) amino acid sequences from diverse species again illustrates the near universal conservation of *N*-glycosylation within the C-Pro domain.

Given that (i) the collagen *N*-glycan is widely conserved within the C-Pro domain, (ii) the C-Pro domain is only present in nascent procollagen molecules, not mature fibrillar collagens, and (iii) collagen interacts with the lectin chaperones, it seems reasonable to assume that the *N*-glycan plays an essential role in folding the C-Pro domain and thereby templating collagen triple-helix assembly. It was quite surprising, then, when nearly thirty years ago, a non-*N*-glycosylated variant of Colα1(I) was shown to fold properly, maintain appropriate hydroxylation, and secrete with kinetics similar to the wild-type (WT) *N*-glycosylated protein.[24]

We are left with a decades-old enigma. *N*-Glycosylation of fibrillar collagen C-Pro domains is nearly universally conserved and therefore almost certainly functional, but cleavage of the C-Pro prior to deposition of collagen triple helices means that the *N*-glycan must not function within the extracellular matrix. Meanwhile, the *N*-glycan apparently does not meaningfully contribute to intracellular procollagen folding. What, then, is the reason for *N*-glycosylation of the most abundant protein in the human body?

We hypothesized that the function of the *N*-glycan might be apparent only in the right proteostatic context. In particular, we speculated that *N*-glycan-mediated access to the ER's lectin-based chaperone network could be essential when collagen folding is challenged, such as in the context of misfolding-prone C-Pro domain mutations and/or ER stress. To test this hypothesis, we generated non-*N*-glycosylated collagen variants via genetic substitution of the native acceptor asparagine with glutamine. We then expressed either *N*-glycosylated or non-*N*-glycosylated versions of osteogenesis imperfecta (OI)-causing, misfolding-prone Colα1(I) variants in human cells and assayed the secretion of both the full-length protein and the isolated C-Pro domain. We observed that misfolding-prone C-Pro variants were highly sensitive to *N*-glycan removal. Non-*N*-glycosylated variants proved defective in folding, assembling, and secreting to the extracellular milieu, instead forming insoluble intracellular aggregates. Chemical induction of low levels of ER stress revealed a similarly essential role for the *N*-glycan in the folding and secretion of even WT Colα1(I) under conditions of proteostatic challenge. Cumulatively, these observations address the *N*-glycan enigma, unveiling the context-dependent essentiality of collagen's conserved *N*-glycan to enable proper folding and trafficking of collagen in challenging proteostasis environments.

RESULTS

Absence of the Conserved *N*-Glycan Impairs Secretion of Misfolding-Prone Collagen Variants. A number of mutations in the C-Pro domain of collagen-I are known to promote misfolding, with many leading to the debilitating disease osteogenesis imperfecta (OI).[25-29] For example, disruption of the disulfide bond network in the C-Pro domain by the C1299W amino acid substitution leads to over-modification of triple helices and delayed collagen-I secretion, resulting in mild OI.[30] We hypothesized that the collagen-I *N*-glycan could be critical for assisting the intracellular folding of such misfolding-prone variants.

To test this hypothesis, we created replication-incompetent adenoviruses encoding HA-tagged WT Colα1(I) or misfolding-prone C1299W Colα1(I), with or without an N1365Q substitution to prevent *N*-glycosylation (Fig. 2-2A). The HA tag, located at the Colα1(I) N-terminus, allows for selective immunoprecipitation and quantification of the ectopically expressed HA-Colα1(I) separate from endogenously expressed Colα1(I). We previously showed that introduction of an HA epitope does not disrupt normal collagen-I folding, modification, or secretion.[16] We used these viruses to transduce primary fibroblasts that also natively express WT collagen-I and then performed pulse-chase experiments to quantify secretion of the HA-tagged Colα1(I) variants in this relevant cellular context. We found that removal of the *N*-glycosylation sequon from WT Colα1(I) did not alter its secretion, fully consistent with previous data.[24] In contrast, removal of the *N*-glycosylation sequon from the C1299W variant led to a significant and substantial secretion defect (Fig. 2-2B and Fig. S2-1A). These results suggest that the *N*-glycan plays a unique role in C1299W Colα1(I) folding that is not critical for successful WT Colα1(I) folding.

We next sought to ascertain whether the secretion defect observed upon removal of the *N*-glycan from C1299W was generalizable to other misfolding-prone C-Pro variants. Collagen genes are large, highly repetitive, GC-rich, and not readily amenable to genetic manipulation.[17] Moreover, mechanistic analyses are challenging in the context of the full-length protein. Hence, for further studies, we used a more tractable system in which the C-Proα1(I) domain was expressed alone, in the absence of the N-Pro and triple-helical domains. We previously showed that expression of the C-Pro domain alone recapitulates the behavior and proteostasis defects associated with various C-Pro domain mutations in collagen-I.[31] In addition, both we and others have shown that the C-Proα1(I) domain folds and assembles correctly in HEK293 cells and is robustly secreted into the media.[6, 7, 32] We confirmed that WT C-Proα1(I) expressed alone was *N*-glycosylated and that the N1365Q substitution eliminated *N*-glycosylation, based on the

equivalent shifts in electrophoretic mobility caused by PNGase F digestion and the N1365Q substitution (Fig. S2-1B).

We generated HA antibody epitope-tagged constructs encoding WT C-Proα1(I), C1299W C-Proα1(I), and four additional OI-causing C-Proα1(I) variants: G1272V, W1275R, A1286D, and D1277H,[26, 27] all with and without the N1365Q substitution. Pulse-chase experiments revealed that, consistent with our full-length Colα1(I) data (Fig. 2-2B), removal of the *N*-glycosylation sequon from WT C-Proα1(I) had no impact on its secretion (Fig. 2-2C and Fig. S2-1C). In contrast, removal of the *N*-glycosylation sequon from misfolding-prone C1299W C-Proα1(I) resulted in a significant and substantial secretion defect. Three of the four additional OI-causing C-Proα1(I) variants (G1272V, W1275R, and A1286D) also exhibited a significant and substantial decrease in secretion upon removal of the *N*-glycan (Fig. 2-2C and Fig. S2-1C), consistent with the notion that the *N*-glycan is somehow critical for ensuring the successful folding of misfolding-prone collagen variants.

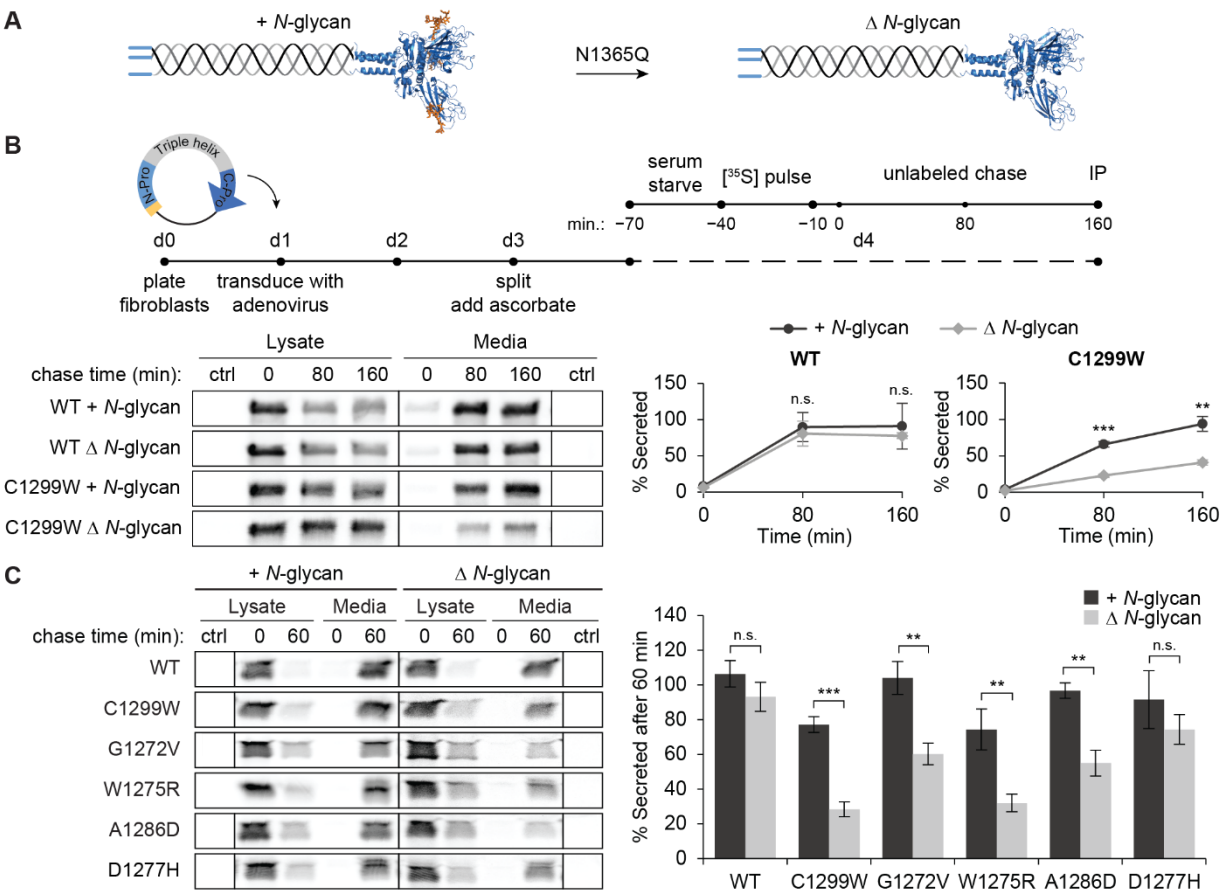


Figure 2-2: Collagen's conserved *N*-glycan is required for folding, assembly, and secretion of misfolding-prone collagen variants.

(A) Substitution of the *N*-glycan acceptor residue, N1365, with a glutamine prevents Colα1(I) from being *N*-glycosylated. (B) Pulse-chase experiments reveal that removal of the *N*-glycan from WT and C1299W Colα1(I) significantly and substantially reduces secretion of the latter ($p_{80\ min} < 0.001$, $p_{160\ min} = 0.001$) but not the former ($p_{80\ min} = 0.593$, $p_{160\ min} = 0.512$). Representative autoradiographs are shown. (C) Pulse-chase experiments conducted with C-Proα1(I) expressed alone recapitulates the secretion patterns exhibited by full-length Colα1(I). Removal of the *N*-glycan significantly reduces secretion of C1299W C-Proα1(I) ($p_{C1299W} < 0.001$) but not WT C-Proα1(I) ($p_{WT} = 0.113$). This effect is consistently observed for other misfolding-prone C-Proα1(I) variants ($p_{G1272V} = 0.003$, $p_{W1275R} = 0.005$, $p_{A1286D} = 0.001$) but not for the assembly-deficient D1277H C-Proα1(I) variant ($p_{D1277H} = 0.185$). Representative autoradiographs are shown.

Notably, removal of the *N*-glycan from the OI-causing D1277H C-Proα1(I) variant did not significantly alter its secretion (Fig. 2-2C and Fig. S2-1C). Unlike C1299W, G1272V, W1275R, and A1286D C-Proα1(I), which are all misfolding-prone but nonetheless capable of forming properly assembled trimers, D1277H C-Proα1(I) is known to be assembly-defective. The D1277H substitution permanently disrupts the Ca^{2+}-binding site in the C-Pro domain that is critical for proper C-Pro assembly.[6, 7] Immunoblots of secreted C-Proα1(I) variants confirmed that all the variants except for D1277H assembled into the expected disulfide-linked trimers (Fig. S2-1D). D1277H C-Proα1(I) largely failed to assemble properly, instead forming a complex mixture of monomers, disulfide-linked dimers and trimers, and larger disulfide-linked aggregates (Fig. S2-1D). Compellingly, these results suggest that, while the presence of the *N*-glycan substantially assists misfolding-prone C-Proα1(I) variants in navigating challenging protein-folding landscapes, it does not greatly promote the secretion of irrevocably misassembling structures, such as those induced by the D1277H substitution in C-Proα1(I).

Two additional features of the immunoblots (Figs. S2-1D and S2-1E) are noteworthy. First, the secretion defect observed in pulse-chase experiments for the non-*N*-glycosylated, misfolding-prone C-Proα1(I) variants was even more striking, with virtually no protein detected in media immunoblots (Fig. S2-1D). We speculated that this observation may be attributable to long-term ER stress caused by production of non-*N*-glycosylated misfolding-prone C-Proα1(I) variants reducing protein synthesis over time and/or extensive aggregation of those same variants.[33] Consistent with this notion, whereas the non-*N*-glycosylated, misfolding-prone C-Proα1(I) variants were readily observed in pulse-chase experiments, they were barely detectable by immunoblotting of the soluble intracellular fraction at steady-state (Fig. S2-1E).

Non-*N*-Glycosylated, Misfolding-Prone C-Proα1(I) Variants Induce ER Stress and are Retained as Intracellular Aggregates. We hypothesized that decreased secretion upon removal of the *N*-glycan could be attributed to exacerbated misfolding of C-Proα1(I) variants

already rendered misfolding-prone by disease-causing mutations. These severely misfolding non-*N*-glycosylated variants may cause ER stress, leading to induction of the unfolded protein response (UPR). To test this hypothesis, we used qPCR to assess mRNA expression levels of UPR target genes.[34] We observed that expression of the non-*N*-glycosylated, misfolding-prone C-Proα1(I) variants induced a substantial increase in mRNA levels of ER stress-responsive transcripts, including those encoding BiP, HYOU1, and Grp94 (Fig. 2-3A and Fig. S2-2A). Notably, none of the misfolding-prone variants induced the UPR when the *N*-glycan was present. Thus, the *N*-glycan appears to be essential to mitigate ER stress-inducing misfolding caused by these mutations within C-Proα1(I).

To evaluate whether the non-*N*-glycosylated, misfolding-prone C-Proα1(I) variants were being targeted to degradation, we treated cells expressing *N*-glycosylated or non-*N*-glycosylated WT or C1299W C-Proα1(I) with either MG-132, a proteasome inhibitor, or bafilomycin A1, an autophagy inhibitor. Neither inhibitor treatment rescued secreted or intracellular steady-state levels of the non-*N*-glycosylated C1299W C-Proα1(I) (Figs. S2-2B and S2-2C).

We next performed a more comprehensive pulse-chase experiment to monitor the intracellular retention, secretion, and degradation of *N*-glycosylated versus non-*N*-glycosylated WT and C1299W C-Proα1(I). We found that secretion of the non-*N*-glycosylated C1299W C-Proα1(I) plateaued at ~50% after 80 minutes. Even after 160 minutes, ~40% of the non-*N*-glycosylated C1299W variant was still retained intracellularly (Fig. 2-3B). These findings were recapitulated in a much longer, 8-hour chase: non-*N*-glycosylated WT C-Proα1(I) was exclusively present in the media after 8 hours (Fig. S2-2D), whereas large quantities of non-*N*-glycosylated C1299W C-Proα1(I) were still retained in cells. Treatment with MG-132 in the pulse-chase experiment had no substantial effect on secretion and only very modestly increased intracellular protein levels (Fig. 2-3B). Together, these results indicate that the reduced secretion of C1299W C-Proα1(I) following removal of the *N*-glycan can largely be attributed to sustained intracellular retention rather than degradation.

We next used confocal microscopy to visualize the status of the intracellularly retained protein. We co-stained the HA-tagged C-Proα1(I) variants with antibodies against the ER marker protein disulfide isomerase (PDI) or the Golgi marker GM130. We observed diffuse staining of WT C-Proα1(I), both with and without the *N*-glycan, that was primarily localized to the Golgi (Fig. 2-3C), consistent with previous observations of well-behaved, robustly secreted WT collagen.[18, 31, 35] *N*-Glycosylated C1299W C-Proα1(I) stained diffusely within both the Golgi and the ER, consistent with the longer ER residence time and excessive post-translational modification

previously observed with this variant[30] and indicative of the protein successfully transiting the secretory pathway. In contrast, the non-*N*-glycosylated C1299W C-Proα1(I) exhibited punctate staining, suggestive of the formation of extensive intracellular aggregates.

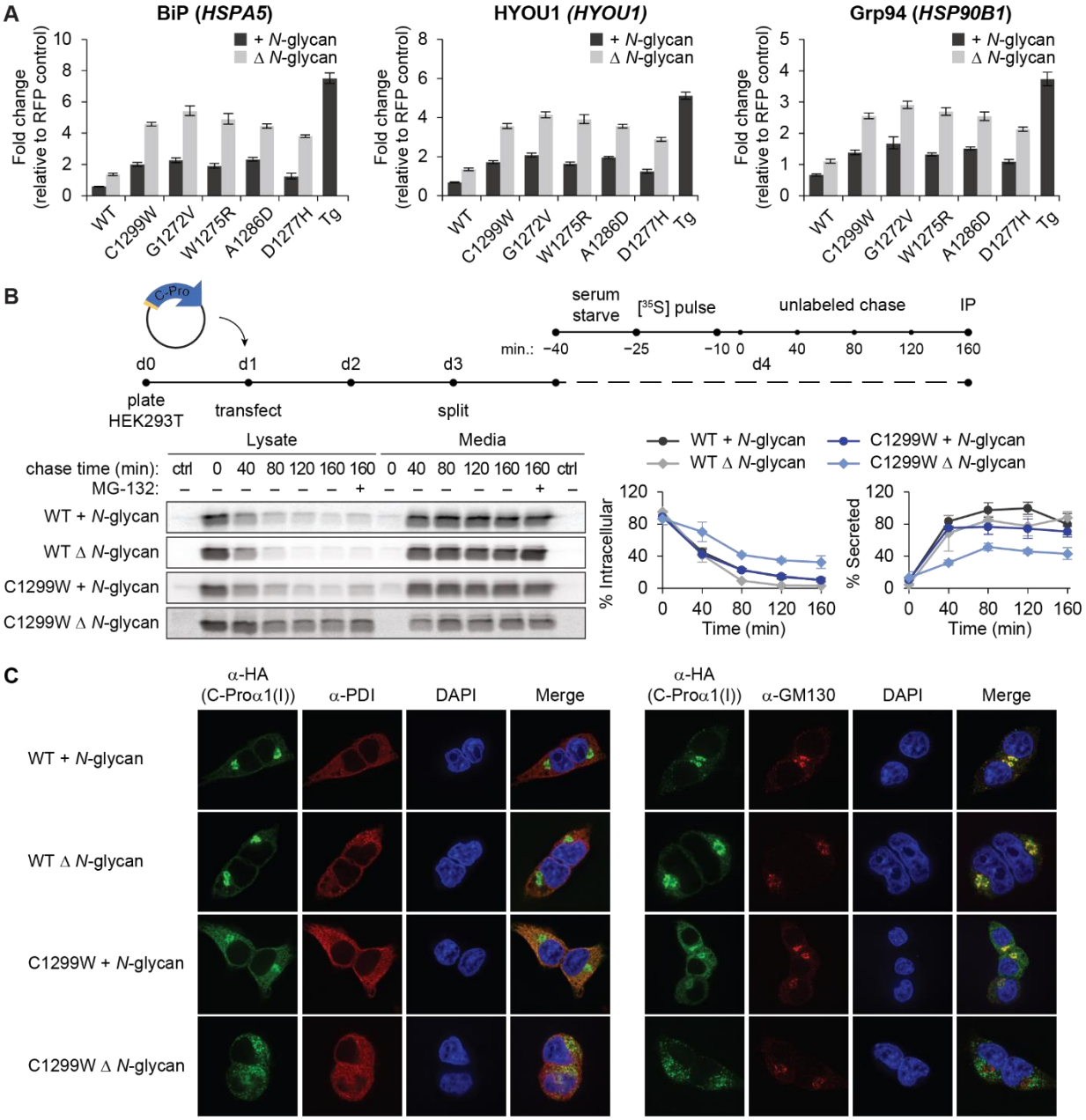


Figure 2-3: Removal of the N-glycan from misfolding-prone but not wild-type C-Proα1(I) induces ER stress and intracellular aggregation.

(A) mRNA expression levels of the unfolded protein response target genes HSPA5, HYOU1, and HSP90B1 show the induction of ER stress upon expression of non-N-glycosylated, misfolding-prone C-Proα1(I) variants (see also Fig. S2-2A). Thapsigargin (Tg) treatment was used as a positive control for ER stress induction. (B) Extended pulse-chase experiments demonstrate the

sustained secretion defect exhibited for non-*N*-glycosylated C1299W C-Proα1(I) and corresponding intracellular retention. Treatment with the proteasome inhibitor MG-132 did not substantively rescue either secreted or intracellular protein levels. Representative autoradiographs are shown. (C) Representative confocal microscopy images show the formation of intracellular aggregates in cells expressing non-*N*-glycosylated C1299W C-Proα1(I).

***N*-Glycosylation Promotes the Interaction of Misfolding-Prone C-Proα1(I) with the Lectin Chaperones.** Formation of intracellular aggregates, coupled with induction of ER stress, strongly suggests that misfolding of these C-Proα1(I) variants is badly exacerbated by the absence of the *N*-glycan. We hypothesized that such misfolding could stem from a failure to effectively engage the ER's lectin-based chaperone machinery – the calnexin/calreticulin cycle (Fig. 2-4A). Nascent *N*-glycoproteins carry a 14-residue precursor oligosaccharide, which is processed in the ER by cleavage of the two terminal glucose residues by glucosidases-I and -II. Calnexin and calreticulin bind the resultant 12-residue sugar on *N*-glycoproteins and, together with ERp57, ERp29, and CypB, assist client protein folding. Upon release from calnexin or calreticulin, the final glucose residue is cleaved by glucosidase-II, and the folded glycoprotein is then transported to the Golgi for secretion. Misfolded proteins can either re-engage with calnexin and calreticulin after re-addition of the final glucose residue by UDP-glucose:glycoprotein glucosyltransferase 1 (UGGT1), or be targeted to quality control.[15, 36, 37]

To evaluate whether and how C-Proα1(I) interacts with the lectin-based chaperone system, we immunoprecipitated HA-tagged WT and C1299W C-Proα1(I) from cell lysates and probed for calnexin, calreticulin, and other related co-chaperones or enzymes (Fig. 2-4B). We observed that all of the interactors we probed for interacted strongly with the *N*-glycosylated, misfolding-prone C1299W C-Proα1(I). In contrast, they only minimally or undetectably engaged WT C-Proα1(I). Removal of the *N*-glycan abolished or strongly reduced interactions between C1299W C-Proα1(I) and components of the lectin-based chaperone system, suggesting that the *N*-glycan is critical for the engagement of misfolding C-Proα1(I) in this system.

Chemically Abrogating Access to the Calnexin/Calreticulin Chaperoning System Phenocopies Genetic Removal of the *N*-Glycan. The observations in Fig. 2-4B suggest that (i) the lectin chaperones specifically engage misfolding-prone C-Proα1(I) variants and (ii) productive interactions between C-Proα1(I) and these chaperones require the presence of the *N*-glycan. It follows that the secretion defect observed for non-*N*-glycosylated, misfolding-prone C-Proα1(I) variants (Fig. 2-2C) could be attributed to an inability of these proteins to fold in the absence of assistance from the calnexin-/calreticulin-based chaperone machinery.

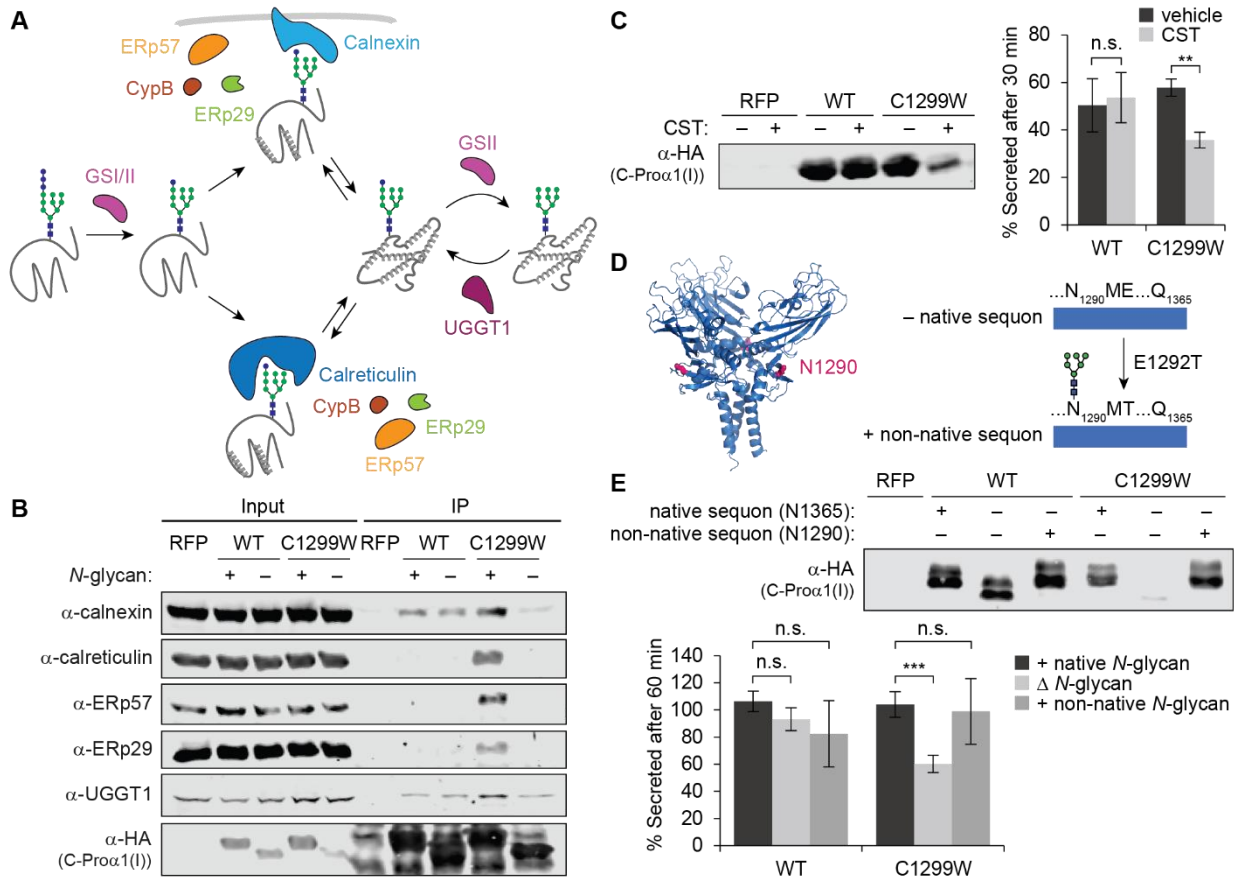


Figure 2-4: Collagen's *N*-glycan addresses proteostasis defects by endowing access to the ER's lectin-based chaperoning system.

(A) Schematic of the calnexin-calreticulin cycle and related co-chaperones or enzymes. (B) Immunoprecipitation (IP) of *N*-glycosylated and non-*N*-glycosylated WT and C1299W C-Proα1(I) revealed that misfolding-prone C1299W C-Proα1(I) interacted much more extensively with the lectin-based chaperone system than WT C-Proα1(I) did. These interactions required the presence of an *N*-glycan on C-Proα1(I). (C) Treating cells with castanospermine (CST), which blocks *N*-glycan-mediated access to the lectin chaperones, impaired secretion of *N*-glycosylated C1299W, chemically phenocopying the secretion defect caused by genetic removal of the *N*-glycan. This effect was observed both by immunoblotting of the media (*Left*) and by pulse-chase (*Right,* ** indicates $p < 0.005$). Representative autoradiographs are provided in the *SI Appendix*, Fig. S2-3A. (D) A non-native *N*-glycosylation sequon can be introduced at C-Proα1(I) residue 1290 via the substitution E1292T. (E) Secretion of the otherwise non-*N*-glycosylated WT C-Proα1(I) was not significantly affected by introduction of the non-native *N*-glycosylation sequon, while secretion of the otherwise non-*N*-glycosylated C1299W C-Proα1(I) was restored to the level of the natively *N*-glycosylated C1299W variant. This effect was observed both by immunoblotting (*Top*) and by pulse-chase (*Bottom*, *** indicates $p < 0.001$). Representative autoradiographs are provided in the Fig. S2-3B.

To differentiate the effects of *N*-glycan-mediated access to the lectin chaperones from any intrinsic stabilization afforded by the *N*-glycan, we treated cells with castanospermine to chemically abrogate access to the calnexin/calreticulin cycle. Castanospermine inhibits

glucosidases-I and -II, preventing removal of the terminal glucose residues from the *N*-glycan and thereby ensuring that *N*-glycosylated proteins cannot access these chaperones via the normal sugar–lectin interaction.[38-40] We observed that, upon treatment with castanospermine, the formerly well-behaved *N*-glycosylated C1299W C-Proα1(I) variant now exhibited a significant secretion defect observable both by pulse-chase experiments and immunoblotting (Fig. 2-4C and Fig. S2-3A), chemically phenocopying the same defect we had previously observed for genetically non-*N*-glycosylated C1299W C-Proα1(I) (Fig. 2-2C). Meanwhile, castanospermine treatment did not affect the secretion of WT C-Proα1(I). Thus, the *N*-glycan is essential for the misfolding-prone C1299W variant specifically because it allows access to the calnexin/calreticulin cycle.

Introduction of a Non-Native *N*-Glycosylation Sequon Elsewhere in C-Proα1(I) Restores Secretion of Misfolding-Prone, Non-*N*-Glycosylated C-Proα1(I). Our data indicate that the *N*-glycan allows misfolding-prone C-Proα1(I) variants to interact with the ER's lectin-based chaperone system, thereby promoting proper folding, assembly, and secretion. We next asked whether this effect was specific to the *N*-glycan located at the native sequon, or whether an *N*-glycan elsewhere within C-Proα1(I) could play a similar role in maintaining proper proteostasis. Notably, not all collagens have an *N*-glycosylation sequon at the equivalent of the N1365 site in Colα1(I). Several instead have a sequon at a different location in the C-Pro domain (Figs. 2-1C and 2-1D).

We introduced a non-native *N*-glycosylation sequon at residue 1290 in otherwise non-*N*-glycosylated WT and C1299W C-Proα1(I). We chose position 1290 because (i) it corresponds to the second most common site for *N*-glycosylation sequons in the fibrillar collagens (Fig. 2-1C, see for example N1672 in Colα1(V)), (ii) it is already an asparagine in the WT C-Proα1(I) sequence, and (iii) it is located in a solvent-exposed loop in the "base" region of the C-Pro domain, where an *N*-glycan likely would not sterically obstruct C-Proα1(I) folding (Fig. 2-4D). We used site-directed mutagenesis to introduce an E1292T substitution within the WT Asn-Met-Glu sequence at position 1290, thereby creating a non-native *N*-glycosylation sequon of Asn-Met-Thr at position 1290 in C-Proα1(I) (Fig. 2-4D).

Based on electrophoretic mobility and PNGase F digests (Fig. S2-1B), we determined that the non-native sequon was successfully *N*-glycosylated when the E1292T/N1365Q variant was expressed in HEK293T cells. Both immunoblotting and pulse-chase experiments (Fig. 2-4E and Fig. S2-3B) revealed that WT C-Proα1(I) with the non-native sequon secreted at a similar level as both the native *N*-glycosylated and non-*N*-glycosylated WT C-Proα1(I) variants. Remarkably, introduction of the non-native sequon in otherwise non-*N*-glycosylated C1299W C-Proα1(I) was

sufficient to restore secretion to similar levels as natively *N*-glycosylated C1299W (Fig. 2-4E). This result is consistent with the notion that the key requirement for solving proteostasis defects in the C-Proα1(I) domain is access to the calnexin/calreticulin cycle via an *N*-glycan; the specific location of the *N*-glycan is flexible.

The *N*-Glycan is Essential for Maintaining WT Collagen Proteostasis in the Context of ER Stress. Collectively, our results demonstrate that collagen *N*-glycosylation is required to preserve proper folding and secretion when misfolding-prone mutations occur in disease. They do not, however, establish a functional role for the *N*-glycan in WT collagen. The demonstrated need for the *N*-glycan when collagen misfolding is exacerbated does, however, suggest a hypothesis that we found quite compelling – specifically, the proper folding and secretion of even WT collagen might be heavily reliant on the presence of the *N*-glycan in settings where general ER proteostasis is impaired. In this regard, collagen-producing tissues often experience physiological ER stress and chronic UPR activation owing to the high secretory output of ECM proteins that accompanies development and extracellular matrix remodeling processes.[41-50]

To test this hypothesis, we first used immunoblotting to assess secretion of WT Colα1(I) with or without the *N*-glycan in either the absence or presence of a low level of ER stress, induced by treatment with thapsigargin (Tg), a Ca^{2+} homeostasis disruptor and well-known ER stress-causing UPR activator. ER stress induction was confirmed by the Tg-induced increase in the protein and mRNA transcript levels for Grp94, BiP, and other UPR-responsive genes (Figs. S2-4A and S2-4B). We found that secretion of both *N*-glycosylated and non-*N*-glycosylated Colα1(I) was decreased by Tg treatment. However, non-*N*-glycosylated Colα1(I) was much more sensitive to ER stress than *N*-glycosylated Colα1(I). Only ~15% of non-*N*-glycosylated Colα1(I) was secreted under Tg treatment compared to DMSO treatment, whereas *N*-glycosylated Colα1(I) was able to maintain nearly 50% of basal secretion (Fig. 2-5A). This observation was confirmed in pulse-chase experiments, in which non-*N*-glycosylated WT Colα1(I) again proved more sensitive to induced ER stress (Fig. 2-5B and Fig. S2-4C). Thus, the collagen-I *N*-glycan protects against WT collagen misfolding when proteostasis is challenged by ER stress.

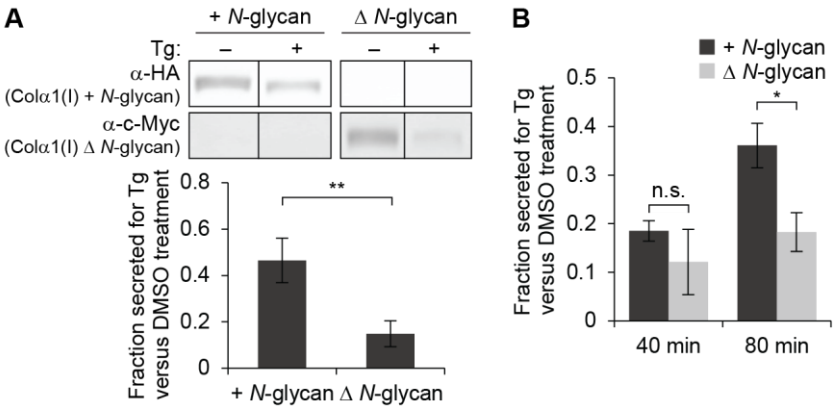


Figure 2-5: Collagen's *N*-glycan protects WT Colα1(I) against misfolding during ER stress.

Primary fibroblasts were transduced with adenoviruses encoding wild-type, full-length Colα1(I) with or without the *N*-glycosylation sequon and then treated with thapsigargin (Tg; 750 nM) to induce ER stress. (A) Immunoblotting of media indicated that secretion of both *N*-glycosylated and non-*N*-glycosylated Colα1(I) was impaired by Tg treatment, with the non-*N*-glycosylated Colα1(I) displaying a much larger defect (*Top*). Quantification from three biological replicates is shown (*Bottom*), alongside a representative immunoblot. The upregulation of BiP and Grp94 in cell lysates confirmed the unfolded protein response induction caused by Tg treatment (Fig. S2-4*B*). (B) Pulse-chase analyses of *N*-glycosylated versus non-*N*-glycosylated Colα1(I) likewise confirmed a significant difference in the fraction of Colα1(I) secreted during Tg treatment relative to DMSO treatment at 80 min (* indicates $p < 0.05$).

DISCUSSION

The high conservation of the *N*-glycosylation sequon in the C-Pro domain of the fibrillar collagens has been noted for decades,[9, 10] yet previous attempts to establish a function for the *N*-glycan proved inconclusive. Lamandé and Bateman revealed that genetic removal of the *N*-glycan attachment site does not alter assembly, secretion, propeptide processing, or matrix deposition of WT collagen-I in normal, *in vitro* model systems.[24] Separately, the cleaved C-Pro domain has been proposed to function as an extracellular signal to regulate collagen synthesis[51-54] or to promote C-Pro uptake and degradation,[55] suggesting that the *N*-glycan could theoretically have a role in C-Pro recognition by a receptor. However, the portion of the C-Pro domain that is important for these signaling processes does not include the *N*-glycan.[52, 53]

Here, we demonstrate that the collagen-I *N*-glycan has an essential intracellular role in maintaining collagen folding and secretion under conditions of proteostatic challenge, such as those incurred during disease or ER stress. We first confirmed that the *N*-glycan is dispensable for folding and secretion of WT collagen-I, fully consistent with previous results from Lamandé and Bateman.[24] We then used OI-causing variants of collagen-I, specifically those located within the C-Pro domain, as a model for conditions under which collagen folding is challenged. These variants exhibit delayed folding but are ultimately able to trimerize and pass cellular quality control checkpoints required for secretion.[26, 30] We discovered that removal of the *N*-glycan from these misfolding-prone Colα1(I) variants leads to dramatically decreased secretion of both the full-length protein and the C-Pro domain when expressed alone. In contrast, secretion of the irrevocably assembly-defective D1277H variant does not benefit as significantly from the presence of the *N*-glycan. Notably, we observed activation of ER stress pathways and formation of intracellular aggregates upon expression of non-*N*-glycosylated, misfolding-prone C-Proα1(I) variants.

We hypothesized that the secretion defect observed for the non-*N*-glycosylated, misfolding-prone C-Proα1(I) variants could be attributed to either intrinsic destabilization upon loss of the *N*-glycan or exacerbated misfolding due to the inability of non-*N*-glycosylated variants to access lectin-based chaperones in the ER. Immunoprecipitation experiments revealed that misfolding-prone C1299W C-Proα1(I), but not WT, associated extensively with the ER's lectin-based chaperone machinery in an *N*-glycan dependent manner. Chemical inhibition of lectin chaperone association using castanospermine fully phenocopied the secretion defect we observed for non-*N*-glycosylated C1299W C-Proα1(I). Taken together, these data indicate that misfolding-prone C-Proα1(I) variants depend heavily on *N*-glycan-mediated access to the lectin chaperones in order to fold, assemble, and pass quality control checkpoints required for secretion.

We further found that the specific location of the *N*-glycan is not critical for this assistance, merely its presence.

The intracellular C-Proα1(I) aggregates we observed in the absence of the *N*-glycan were not readily cleared by either ER-associated degradation or autophagy, even though both pathways are reported to have roles in collagen-I quality control.[19, 56-58] Calnexin has notably been found to interact with FAM134B, an ER-phagy receptor, thereby targeting misfolding procollagen to an autophagic fate that can be inhibited by bafilomycin A1.[19] Such clearance would likely be slowed or eliminated by removal of the *N*-glycan, which reduces binding to calnexin. Separately, another recent report of autophagic collagen clearance that is bafilomycin A1-insensitive suggests that non-canonical mechanisms may be involved.[56] Regardless, clearance by any of these or other pathways is very slow for the variants studied here, requiring >8 hours after synthesis of non-*N*-glycosylated, misfolding-prone C-Proα1(I).

Misfolding-prone collagen-I variants are found only in a very small fraction of the human population, suggesting that the highly conserved *N*-glycan is unlikely to function only in this context. Indeed, we find that the *N*-glycan is also critical for proper folding and secretion of WT collagen-I under conditions of mild ER stress. ER stress and UPR activation occur in various commonly encountered physiological contexts, including differentiation of collagen-producing cells such as osteoblasts, chondrocytes, and fibroblasts.[42-45] Chronic activation of the UPR has also been observed in connection with processes that require synthesis of large amounts of matrix proteins, such as development[41] and wound healing.[46, 47, 50] The *N*-glycan thus appears to be crucial for maintaining WT collagen proteostasis and secretion during these critical, yet commonplace, circumstances.

Our data point to the *N*-glycan as a key structural motif responsible for enabling collagen folding under challenging proteostasis conditions, adding to the repertoire of critical functions for collagen's C-Pro domain. Under normal conditions, the highly conserved *N*-glycan on collagen-I is dispensable because the protein can fold and secrete efficiently without relying on lectin chaperones. However, when protein folding is challenged, whether by disease-associated mutations or ER stress, the *N*-glycan becomes essential for preventing collagen misfolding by providing access to the lectin chaperone network in the ER (Fig. 2-6).

Absence of the *N*-glycan not only exacerbates folding defects but also disrupts collagen quality control, leading to extensive intracellular aggregation. These observations highlight the remarkable efficacy of collagen proteostasis, wherein multiple systems cooperate to safeguard the production of a critical protein under changing conditions. Moreover, they suggest the potential of modulating the levels and/or activities of the ER's lectin chaperone network as a strategy to

resolve dysregulated collagen proteostasis in the collagenopathies.[29] In conclusion, our results reveal a fundamental, context-dependent role for collagen's highly conserved *N*-glycan; its implications for cellular signaling, development, and collagen misfolding diseases provide ample subject for future work.

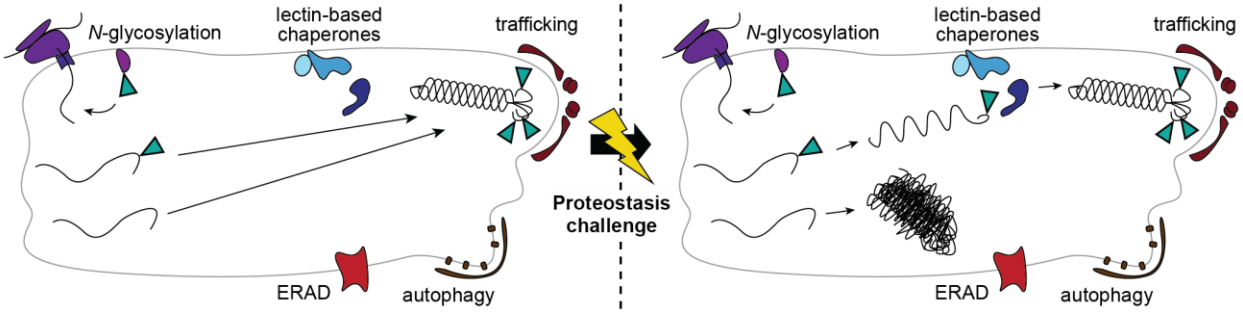


Figure 2-6: Model for the context-dependent essentiality of collagen's conserved *N*-glycan.

Under normal conditions, collagen folds and secretes properly without significant aid from *N*-glycan-mediated access to the lectin chaperones. When folding is challenged, whether by misfolding-prone mutations or physiological ER stress, the *N*-glycan is essential to allow collagen to associate with lectin chaperones in the ER to promote proper folding, assembly, and secretion. The inability to access lectin chaperones leads to aggregation and impaired secretion.

MATERIALS AND METHODS

Cell Lines and Reagents. HEK293T cells (ATCC) were cultured in complete DMEM media (Corning) supplemented with 10% fetal bovine serum (FBS), 100 IU penicillin, 100 μg/mL streptomycin, and 2 mM L-glutamine (Corning). Healthy dermal fibroblasts (GM05294; Coriell Cell Repositories) were cultured in complete MEM (Corning) supplemented with 15% FBS, 100 IU penicillin, 100 μg/mL streptomycin, and 2 mM L-glutamine. Procollagen-I expression and hydroxylation was induced in primary fibroblasts using 200 μM sodium ascorbate (Amresco) during expression of triple-helical domain-containing procollagen constructs. Immunoblots were probed with the following primary antibodies, as indicated: anti-HA (Santa Cruz Biotechnology, 1:200; ThermoFisher 26183, 1:10000), anti-β-actin (Sigma A1978, 1:2000), anti-GFP (GeneTex GTX-113617, 1:1000), anti-LC3B (Abcam ab192890, 1:1000), anti-calnexin (Enzo ADI-SPA-860, 1:1000), anti-calreticulin (ThermoFisher PA3-9000, 1:1000), anti-ERp57 (Cell Signaling Technology 2881, 1:1000), anti-UGGT1 (Abcam ab124879, 1:10000), anti-ERp29 (Cell Signaling Technology, 1:1000), anti-c-Myc (Sigma M4439, 1:5000), and anti-KDEL (Enzo, 1:1000).

Sequence Alignments. Procollagen amino acid sequences were aligned using the default settings of ClustalOmega (https://www.ebi.ac.uk/Tools/msa/clustalo/). For the indicated organisms, the sequences of proα1(I) or the closest available analogue were aligned. The accession numbers used are listed in Table S2-1.

Vector Construction. Site-directed mutagenesis was performed on previously described PPT.HA.proα1(I).pENTR1A[16] and blasticidin-resistant versions of PPT.HA.C-Proα1(I).pcDNA3.1[6] plasmids using the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent). As required, procollagen pENTR1a vectors were recombined into pAd.CMV.V5.DEST Gateway destination vectors using LR Clonase II (Invitrogen). c-Myc-tagged versions of full-length N1365Q constructs were generated by insertion of appropriate procollagen genes into a PPT.c-Myc.pENTR1a vector, generated by ligating the c-Myc epitope tag into pENTR1a using the NotI and XbaI (New England Biolabs) sites for restriction cloning. The DHFR.YFP.HA.pLenti.CMV.PuroDEST vector was cloned from previously described DHFR.YFP.pENTR1A[34] using LR Clonase II recombination.

Full-Length Proα1(I) Adenovirus Production. Adenoviral destination vectors encoding HA.proα1(I), HA.proα1(I)C1299W, HA.proα1(I)N1365Q, HA.proα1(I)C1299W/N1365Q, c-Myc.proα1(I)N1365Q, or c-Myc.proα1(I)C1299W/N1365Q were used to produce adenovirus according to the manufacturer's instructions (Life Technologies ViraPower Adenoviral Expression System). Viruses were subsequently amplified in HEK293A cells (Thermo Fisher) and titrated using GM05294 primary fibroblasts to identify an appropriate volume to use for experiments,

defined as the minimum viral volume that provided consistent signal for all constructs at both the protein and transcript level.

C-Proα1(I) Experiments. HEK293T cells were plated at a density of 7.5 × 10^5 cells/well in 6-well plates the day before transfection. Cells were transfected with the appropriate vectors using Lipofectamine 3000 (Thermo Fisher Scientific), according to the manufacturer's instructions. Media was changed to fresh DMEM the following day, after which media and lysates were harvested for analysis two days post-transfection. For *N*-glycan analyses, media was digested with PNGase F (New England Biolabs) according to the manufacturer's instructions.

For small molecule treatment, transfected cells were split into 6-well plates at a density of 1.5 × 10^6 cells/well one day post-transfection. The following day, media was changed to fresh DMEM containing DMSO or the appropriate small molecule. Cells were treated with 10 μM MG-132 (Enzo Life Sciences) for 16 h, 100 nM bafilomycin A1 (Alfa Aesar) for 16 h, or 1 μM castanospermine (Sigma) for 6 h. Media and lysates were then harvested for analysis.

Immunoblotting. To obtain cell lysates, cells were trypsinized, washed with phosphate-buffered saline (PBS), and lysed in Triton-X lysis buffer (50 mM Tris-HCl, pH 7.5, 150 mM sodium chloride, 1 mM EDTA, 1.5 mM magnesium chloride, 1% Triton X-100, 1.5 mM phenylmethylsulfonyl fluoride (PMSF; Sigma), and protease inhibitor (Pierce)). Lysates were cleared by centrifugation at 21,100 × *g* and 4 °C for 15 min. Total protein in the supernatant was quantified using the bicinchoninic acid assay (Pierce) or Bradford assay (Bio-Rad), and equal amounts of total protein were analyzed for each sample. For analyses under reducing conditions, cell lysates or media were denatured by boiling in 6× Laemmli buffer (300 mM Tris, pH 6.8, 15% glycerol, 6% SDS, and 10% (w/v) bromophenol blue) supplemented with 167 mM 1,4-dithiothreitol (DTT), separated on homemade 4/8% or 4/12% Tris SDS−PAGE gels, and transferred to nitrocellulose membranes. For analyses under non-reducing conditions, media samples were boiled in Laemmli buffer without DTT prior to SDS-PAGE gel separation. Owing to cross-reactivity between antibodies, samples of medium from primary fibroblasts for full-length procollagen-I blots were run on duplicate gels, transferred, and then probed separately for either mouse anti-HA or mouse anti-c-Myc. Blots were imaged after incubation with appropriate primary and 800CW or 680LT secondary antibodies (LI-COR) by scanning on an Odyssey infrared imager (LI-COR), followed by quantification as required using ImageJ. Each experiment was performed in biological triplicate, with representative blots shown.

Full-Length Proα1(I) Pulse-Chase Experiments. GM05294 primary human fibroblasts were plated at a density of 5.6 × 10^6 cells/plate in 150-mm dishes the day before transduction. After allowing cells to adhere overnight, cells received fresh media and were transduced with

HA.proα1(I) variant-encoding adenoviruses or untransduced as a control. The next day, cells received fresh media. Two days post-transduction, cells were trypsinized and plated on poly-D-lysine (PDL; Sigma)-coated 6-well plates at a density of 2.5×10^5 cells/well. For experiments involving Tg treatment, cells were allowed to adhere for approximately 8 h, then the media was spiked with DMSO or Tg to a final concentration of 750 nM. After incubating overnight, cells were serum starved for 30 min in DMEM without Met- and Cys- (Corning, Gibco), then pulsed with 100 μCi/mL of ^{35}S-labeled Met- and Cys-containing media (Perkin Elmer) for 30 min. DMSO or Tg were included during the starve and pulse steps as appropriate. Cells were then washed three times with complete MEM to remove excess radiolabel prior to incubating for the indicated chase times. Media was harvested directly, while cells were lysed in RIPA buffer (25 mM Tris at pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton, 0.5% sodium deoxycholate, 0.1% sodium dodecyl sulfate, 1.5 mM PMSF, and protease inhibitor tablets) and then spun at $21.1 \times 10^3 \times g$ at 4 °C for 15 min to pellet cell debris. Samples were then immunoprecipitated using 30 μL of HA-antibody agarose beads (Sigma) overnight, mixing end-over-end at 4 °C. The following day, beads were washed three times with RIPA buffer for 10 min end-over-end at 4 °C, washed once with PBS for 10 min end-over-end at 4 °C, and then dried with a 30.5-gauge needle. Samples were eluted by boiling in 6× Laemmli buffer containing 167 mM DTT for 10 min, spun at 21.1×10^3 *g* at 4 °C for 5 min, and then separated on 4/8% Tris-Gly SDS-PAGE gels. Gels were dried on a gel slab dryer at 60 °C for 120 min, developed with a GE Healthcare PhosphoScreen, and imaged on a Typhoon FLA 7000. All experiments were performed in biological triplicate, and gels were quantified using ImageQuant TL. Statistical analyses were performed using unpaired Student *t*-tests.

C-Proα1(I) Pulse-Chase Experiments. HEK293T cells were plated at a density of 5.0×10^6 cells/dish in 100-mm dishes the day before transfection. Cells received fresh media and were then transfected with C-Proα1(I)-encoding plasmids or an RFP-encoding plasmid as a control using Lipofectamine 3000, according to the manufacturer's instructions. The next day, cells were trypsinized and plated on PDL (Sigma)-coated 6-well plates at a density of 1.7×10^6 cells/well. For experiments involving castanospermine treatment, cells received fresh media containing either 1 μM castanospermine or vehicle the following day, 4 h prior to serum starvation. Cells were serum-starved for 15 min in DMEM without Met- and Cys- (Corning, Gibco), then pulsed with 100 μCi/mL of ^{35}S-labeled Met- and Cys-containing media (Perkin Elmer) for 15 min. Castanospermine was included during the starve and pulse steps as appropriate. Cells were then washed three times with complete DMEM to remove excess radiolabel prior to incubating for the indicated chase times. Where indicated, 15 μM of the proteasome inhibitor MG-132 was added

at the start of the chase. Media was harvested directly, while lysates were harvested in RIPA buffer and then spun at 21.1 × 10^3 *g* at 4 °C for 15 min to pellet cell debris. Samples were then immunoprecipitated using 30 μL of HA-antibody agarose beads (Sigma) overnight, mixing end-over-end at 4 °C. The following day, beads were washed three times with RIPA buffer for 10 min end-over-end at 4 °C, washed once with PBS for 10 min end-over-end at 4 °C, and then dried with a 30.5-gauge needle. Samples were then eluted by boiling in 6× Laemmli buffer containing 167 mM DTT for 10 min, spun at 21.1 × 10^3 g and 4 °C for 5 min, and then separated on 4/12% Tris-Gly SDS-PAGE gels. Gels were dried on a gel slab dryer at 60 °C for 120 min, developed with a GE Healthcare PhosphoScreen, and imaged on a Typhoon FLA 7000. All experiments were performed in biological triplicate, and gels were quantified using ImageQuant TL. Statistical analyses were performed using unpaired Student *t*-tests.

RNA Purification and qPCR. HEK293T cells were harvested by trypsinization and washed with PBS. Primary patient fibroblasts were rinsed with PBS, then harvested in the plate. RNA was isolated using the E.Z.N.A. Total RNA Kit I (Omega Bio-Tek), according to the manufacturer's instructions. RNA concentrations were normalized to 1 μg total RNA prior to cDNA synthesis using a High Capacity cDNA Reverse Transcription Kit (Applied Biosciences). cDNA was amplified in a Light Cycler 480 II Real Time PCR Instrument in the MIT BioMicro Center using the KAPA SYBR FAST qPCR Master Mix (Kapa Biosystems) and appropriate primers. Transcript levels were normalized to the housekeeping gene *RPLP2*, and fold-change was calculated relative to the RFP transfected or untransduced control. All measurements were performed in biological and technical triplicate.

Immunofluorescence and confocal microscopy. HEK293T cells were transfected as described above. The following day, 1.0 × 10^5 cells were transferred to PDL-coated cover slips housed in 24-well plates. After incubating for at least 24 h, culture media was removed and the cover slips were washed with PBS. The cells were then fixed in 3.7% formaldehyde (Mallinckrodt) overnight at 4 °C. After removing the formaldehyde, the cover slips were washed with PBS and then incubated with 0.1% Triton X-100 for 30 min to permeabilize the cells. Cover slips were then incubated overnight in a blocking solution of 5% bovine serum albumin in TBS with 0.01% sodium azide. Cells were labelled overnight with the anti-HA antibody (Abcam ab9110; 1:100) in 5% BSA at 4 °C, followed by either the anti-PDI (Abcam ab2792; 1:100) or anti-GM130 antibody (Sigma G7295; 1:500). Between antibodies, the cover slips were washed three times with TBS. Both secondary antibodies, Alexa Fluor 568-conjugated anti-mouse and Alexa Fluor 488-conjugated anti-rabbit (Invitrogen; 1:1000), were applied simultaneously in the 5% BSA solution for 2 h at room temperature. The cover slips were then washed five times with TBS and labelled with DAPI

(Invitrogen) for 15 min at room temperature, followed by one wash with TBS and two with water. Cover slips were mounted on microscope slides using ProLong Gold Antifade Mountant (Invitrogen). Images were acquired at the W. M. Keck Microscopy Facility at the Whitehead Institute on a Zeiss AxioVert200M microscope with a 100× oil immersion objective with a Yokogawa CSU-22 spinning disk confocal head with a Borealis modification (Spectral Applied Research/Andor) and a Hamamatsu ORCA-ER CCD camera. The MetaMorph software package (Molecular Devices) was used for image acquisition. The excitation lasers used to capture the images were 405 nm, 488 nm, and 561 nm. Image processing was performed using ImageJ.

Immunoprecipitations. HEK293T cells expressing C-Proα1(I) variants were trypsinized from 100-mm dishes, washed with PBS, then resuspended in 10 mL of PBS. Cells were crosslinked with 150 μM dithiobis(succinimidyl propionate) by mixing end-over-end for 30 min at room temperature. Crosslinking was quenched by addition of Tris at pH 8.0 to a final concentration of 100 mM for 10 min end-over-end. Cell lysates were prepared as described above using RIPA buffer, and 1 mg of total protein was immunoprecipitated using 60 μL of HA-antibody agarose beads overnight, mixing end-over-end at 4 °C. The following day, samples were centrifuged for 5 min at 200 × g and 4 °C. The supernatant was removed, and then the beads were washed with RIPA buffer three times followed by PBS once. After the third wash, samples were eluted in 40 μL of 300 mM Tris at pH 6.8 and 6% SDS (w/v) by boiling for 10 min, followed by separation on homemade 4/12% Tris SDS−PAGE gels and immunoblotting as described above.

Full-Length Proα1(I) Immunoblotting Experiment. GM05294 primary human fibroblasts were plated at a density of 3.5×10^5 cells/plate in 60-mm dishes the day before transduction. After allowing cells to adhere overnight, cells received fresh media and were transduced with adenoviruses encoding HA.proα1(I) and c-Myc.proα1(I)N1365Q. The next day, cells received fresh media. Two days post-transduction, cells were split 1:4 into 12-well plates. Endogenous procollagen-I expression was induced by treatment with 200 μM sodium ascorbate the following day, and cells were treated with DMSO or 750 nM thapsigargin (Tg). After 24 h, media and cells were collected for analysis.

SUPPORTING FIGURES

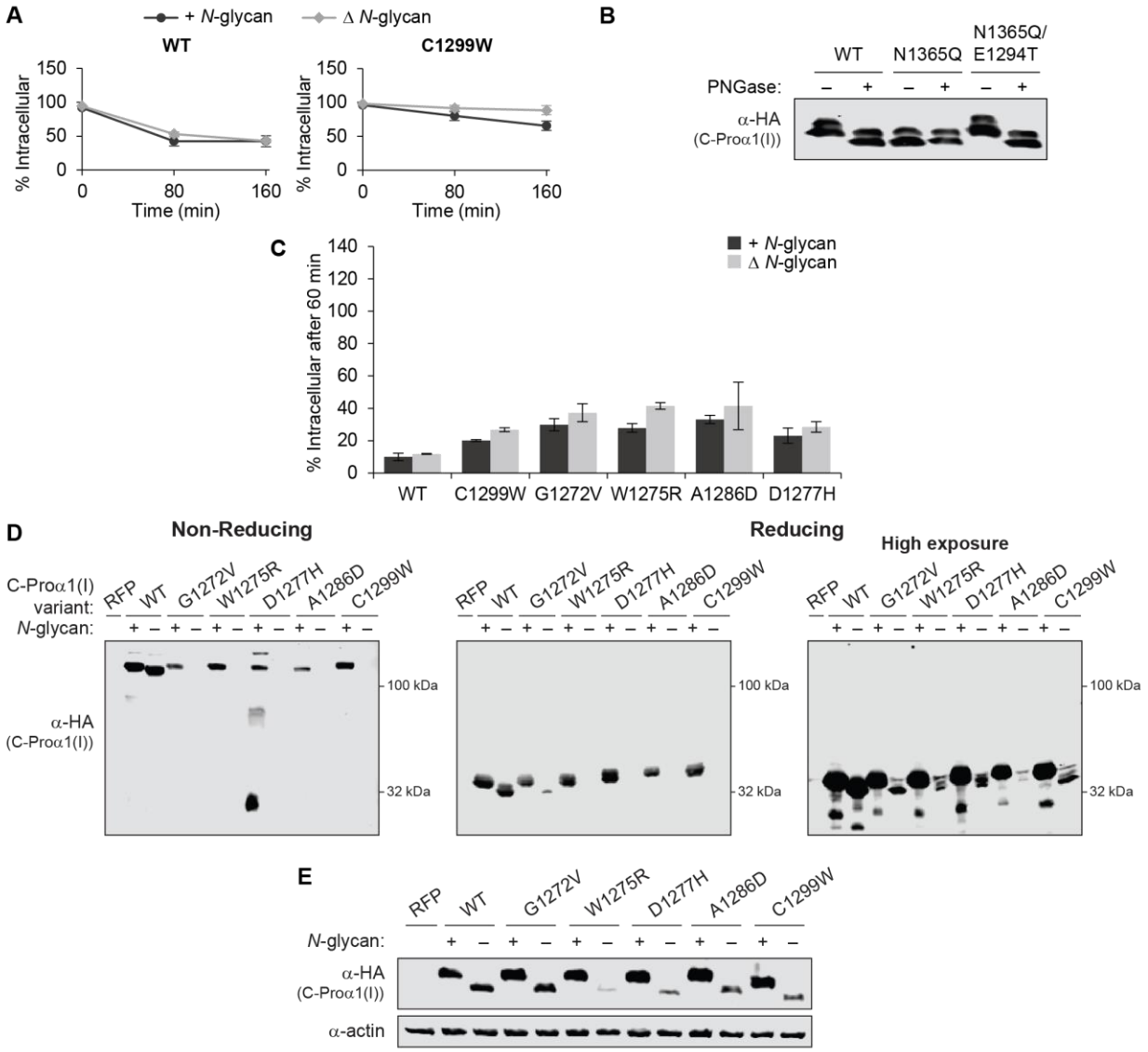


Figure S2-1: Analyses of Proα1(I) and C-Proα1(I) expression.

(A) Quantification of the lysate fraction corresponding to the representative autoradiographs shown in Fig. 2-2B. (B) Digestion of WT and N1365Q/E1294T C-Proα1(I) with PNGase F resulted in a shift in electrophoretic mobility that corresponds to *N*-glycan removal and matched the electrophoretic mobility of the N1365Q variant. (C) Quantification of the lysate fraction corresponding to the representative autoradiographs shown in Fig. 2-2C. (D) Immunoblots of secreted C-Proα1(I) variants with and without the *N*-glycan under non-reducing (*Left*) and reducing conditions (*Right*) demonstrate: (i) assembly of all *N*-glycosylated variants into disulfide-linked trimers, with the exception of D1277H and (ii) the secretion defect resulting from removal of the *N*-glycan from misfolding-prone, C-Proα1(I) variants, which is even more striking when analyzed at steady state (note the high exposure media immunoblot under reducing conditions, in which some low level of secretion can be observed). (E) Steady-state immunoblots of the intracellular fraction for C-Proα1(I) variants with and without the *N*-glycan reveal very low levels of soluble intracellular protein detectable by immunoblotting at steady state.

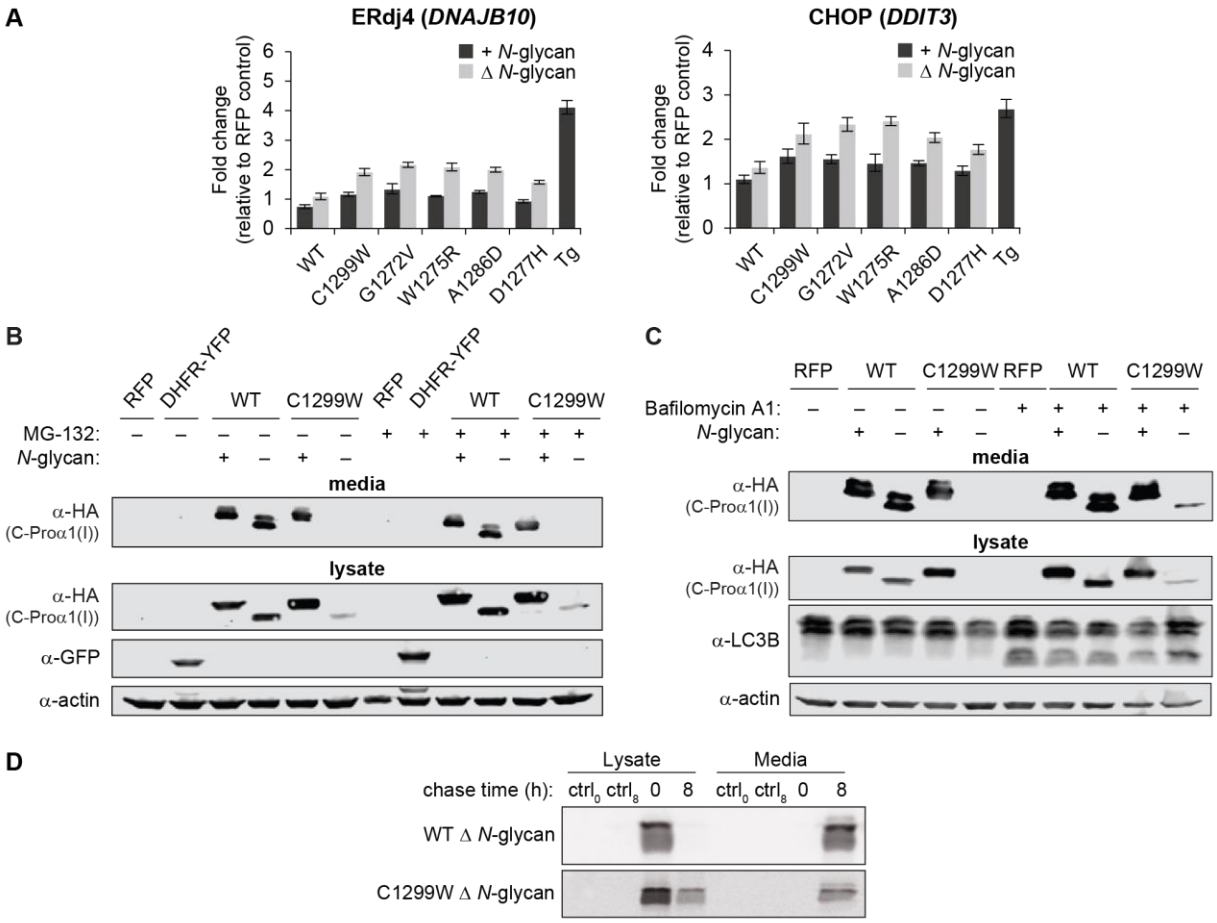


Figure S2-2: Analyses of non-*N*-glycosylated, misfolding-prone C-Proα1(I) variants.

(A) mRNA expression levels of the unfolded protein response-induced transcripts *DNAJB10* and *DDIT3* further highlight induction of ER stress when non-*N*-glycosylated, misfolding-prone C-Proα1(I) variants are expressed. Thapsigargin (Tg) treatment was used as a positive control for ER stress induction. (B) Treatment of cells expressing WT or C1299W C-Proα1(I) with or without the *N*-glycan with MG-132, a proteasome inhibitor, suggests that none of the variants were extensively degraded by the proteasome. DHFR.YFP represents a destabilizing domain variant of *Escherichia coli* dihydrofolate reductase fused to YFP. It serves as a positive control, because successful proteasome inhibition is known to increase intracellular levels of DHFR.YFP.[59] (C) Treatment of cells expressing WT or C1299W C-Proα1(I) with or without the *N*-glycan with bafilomycin A1, an autophagy inhibitor, suggests that the secretion defect exhibited by the non-*N*-glycosylated C1299W C-Proα(I) cannot be fully accounted for by autophagic degradation. Inhibition of autophagy using bafilomycin A1 decreases LC3B lipidation, resulting in the appearance of lower molecular weight bands in Western blots.[60] (D) Representative ^{35}S gel images from an 8 h pulse-chase indicate sustained intracellular retention of non-*N*-glycosylated C1299W C-Proα1(I).

SUPPORTING TABLES

Table S2-1. Accession numbers for procollagen amino acid sequences of the indicated species.

Species	Accession number
Human	P02452
Macaque	H9Z595
Dog	F1Q3I5
Cow	P02453
Mouse	P11087
Chicken	P02457
Frog	Q802B5
Coelecanth	XP_005992246.1
Shark	XP_007908904.1
Zebrafish	UPI00038359B8
Lamprey	B3Y964
Lancelet	Q53UC0
Tunicate	Q53UC1
Sea urchin	Q26052
Abalone	O97405
Red fire ant	XP_011176175.1
Honeybee	UPI000252BA34
Mollusk	UPI00084B9720
Annelid	P90679
Hydra	T2MHG7
Tapeworm	A0A0X3PV25
Sponge	P18856

REFERENCES

1. Mouw, J. K., Ou, G., Weaver, V. M., Extracellular matrix assembly: a multiscale deconstruction. *Nat. Rev. Mol. Cell Biol.* 15, 771-785 (2014).
2. Ricard-Blum, S., The collagen family. *Cold Spring. Harb. Perspect. Biol.* 3, a004978 (2011).
3. Gelse, K., Pöschl, E., Aigner, T., Collagens—structure, function, and biosynthesis. *Adv. Drug Deliv. Rev.* 55, 1531-1546 (2003).
4. Doege, K. J., Fessler, J. H., Folding of carboxyl domain and assembly of procollagen I. *J. Biol. Chem.* 261, 8924-8935 (1986).
5. Boudko, S. P., Engel, J., Bachinger, H. P., The crucial role of trimerization domains in collagen folding. *Int. J. Biochem. Cell Biol.* 44, 21-32 (2012).
6. DiChiara, A. S., Li, R. C., Suen, P. H., Hosseini, A. S., Taylor, R. J., Weickhardt, A. F., Malhotra, D., McCaslin, D. R., Shoulders, M. D., A cysteine-based molecular code informs collagen C-propeptide assembly. *Nat. Commun.* 9, 4206 (2018).
7. Sharma, U., Carrique, L., Vadon-Le Goff, S., Mariano, N., Georges, R. N., Delolme, F., Koivunen, P., Myllyharju, J., Moali, C., Aghajari, N., Hulmes, D. J., Structural basis of homo- and heterotrimerization of collagen I. *Nat. Commun.* 8, 14671-14681 (2017).
8. Jalan, A. A., Sammon, D., Hartgerink, J. D., Brear, P., Stott, K., Hamaia, S. W., Hunter, E. J., Walker, D. R., Leitinger, B., Farndale, R. W., Chain alignment of collagen I deciphered using computationally designed heterotrimers. *Nat. Chem. Biol.* 16, 423–429 (2020).
9. Clark, C. C., The distribution and initial characterization of oligosaccharide units on the COOH-terminal propeptide extensions of the pro-α1 and pro-α2 chains of type I procollagen. *J. Biol. Chem.* 254, 10798-10802 (1979).
10. Olsen, B. R., Guzman, N. A., Engel, J., Condit, C., Aase, S., Purification and characterization of a peptide from the carboxy-terminal region of chick tendon procollagen type I. *Biochemistry* 16, 3030-3036 (1977).
11. Kornfeld, R., Kornfeld, S., Assembly of asparagine-linked oligosaccharides. *Annu. Rev. Biochem.* 54, 631-664 (1985).
12. Varki, A., Biological roles of glycans. *Glycobiology* 27, 3-49 (2017).
13. Helenius, A., Aebi, M., Intracellular functions of N-linked glycans. *Science* 291, 2364-2369 (2001).
14. Helenius, A., Aebi, M., Roles of N-linked glycans in the endoplasmic reticulum. *Annu. Rev. Biochem.* 73, 1019-1049 (2004).
15. Caramelo, J. J., Parodi, A. J., Getting in and out from calnexin/calreticulin cycles. *J. Biol. Chem.* 283, 10221-10225 (2008).
16. DiChiara, A. S., Taylor, R. J., Wong, M. Y., Doan, N. D., Rosario, A. M., Shoulders, M. D., Mapping and exploring the collagen-I proteostasis network. *ACS Chem. Biol.* 11, 1408-1421 (2016).
17. Doan, N.-D., DiChiara, A. S., Del Rosario, A. M., Schiavoni, R. P., Shoulders, M. D., "Mass spectrometry-based proteomics to define intracellular collagen interactomes" in Collagen. Methods in Molecular Biology, Sagi, I., Afratis, N. A., Eds. (Humana Press, New York, NY, 2019), vol. 1944, pp. 95-114.
18. Van Duyn Graham, L., Sweetwyne, M. T., Pallero, M. A., Murphy-Ullrich, J. E., Intracellular calreticulin regulates multiple steps in fibrillar collagen expression, trafficking, and processing into the extracellular matrix. *J. Biol. Chem.* 285, 7067-7078 (2010).
19. Forrester, A., De Leonibus, C., Grumati, P., Fasana, E., Piemontese, M., Staiano, L., Fregno, I., Raimondi, A., Marazza, A., Bruno, G., Iavazzo, M., Intartaglia, D., Seczynska, M., van Anken, E., Conte, I., De Matteis, M. A., Dikic, I., Molinari, M., Settembre, C., A selective ER-phagy exerts procollagen quality control via a calnexin-FAM134B complex. *EMBO J.* 38, e99847 (2019).

20. Hanson, S. R., Culyba, E. K., Hsu, T.-L., Wong, C.-H., Kelly, J. W., Powers, E. T., The core trisaccharide of an N-linked glycoprotein intrinsically accelerates folding and enhances stability. *Proc. Natl. Acad. Sci. U.S.A.* 106, 3131-3136 (2009).
21. Hebert, D. N., Lamriben, L., Powers, E. T., Kelly, J. W., The intrinsic and extrinsic effects of N-linked glycans on glycoproteostasis. *Nat. Chem. Biol.* 10, 902-910 (2014).
22. O'Connor, S. E., Imperiali, B., A molecular basis for glycosylation-induced conformational switching. *Chem. Biol.* 5, 427-437 (1998).
23. Wang, C., Eufemi, M., Turano, C., Giartosio, A., Influence of the carbohydrate moiety on the stability of glycoproteins. *Biochemistry* 35, 7299-7307 (1996).
24. Lamandé, S. R., Bateman, J. F., The type I collagen proα1(I) COOH-terminal propeptide N-linked oligosaccharide: functional analysis by site-directed mutagenesis. *J. Biol. Chem.* 270, 17858-17865 (1995).
25. Barnes, A. M., Ashok, A., Makareeva, E. N., Brusel, M., Cabral, W. A., Weis, M., Moali, C., Bettler, E., Eyre, D. R., Cassella, J. P., Leikin, S., Hulmes, D. J. S., Kessler, E., Marini, J. C., COL1A1 C-propeptide mutations cause ER mislocalization of procollagen and impair C-terminal procollagen processing. *Biochim. Biophys. Acta Mol. Basis Dis.* 1865, 2210-2223 (2019).
26. Symoens, S., Hulmes, D. J. S., Bourhis, J. M., Coucke, P. J., De Paepe, A., Malfait, F., Type I procollagen C-propeptide defects: study of genotype–phenotype correlation and predictive role of crystal structure. *Hum. Mutat.* 35, 1330-1341 (2014).
27. Chessler, S. D., Wallis, G. A., Byers, P. H., Mutations in the carboxyl-terminal propeptide of the proα1(I) chain of type I collagen results in defective chain association and produce lethal osteogenesis imperfecta. *J. Biol. Chem.* 268, 18218-18225 (1993).
28. Pace, J. M., Wiese, M., Drenguis, A. S., Kuznetsova, N., Leikin, S., Schwarze, U., Chen, D., Mooney, S. H., Unger, S., Byers, P. H., Defective C-propeptides of the proα2(I) chain of type I procollagen impede molecular assembly and result in osteogenesis imperfecta. *J. Biol. Chem.* 283, 16061-16067 (2008).
29. Wong, M. Y., Shoulders, M. D., Targeting defective proteostasis in the collagenopathies. *Curr. Opin. Chem. Biol.* 50, 80-88 (2019).
30. Pace, J. M., Kuslich, C. D., Willing, M. C., Byers, P. H., Disruption of one intra-chain disulphide bond in the carboxyl-terminal propeptide of the proα1(I) chain of type I procollagen permits slow assembly and secretion of overmodified, but stable procollagen trimers and results in mild osteogenesis imperfecta. *J. Med. Genet.* 38, 443-449 (2001).
31. Doan, N. D., Hosseini, A. S., Bikovtseva, A. A., Huang, M. S., DiChiara, A. S., Papa, L. J., 3rd, Koller, A., Shoulders, M. D., Elucidation of proteostasis defects caused by osteogenesis imperfecta mutations in the collagen-α2(I) C-propeptide domain. *J. Biol. Chem.* 295, 9959-9973 (2020).
32. Seo, W. Y., Kim, J. H., Baek, D. S., Kim, S. J., Kang, S., Yang, W. S., Song, J. A., Lee, M. S., Kim, S., Kim, Y. S., Production of recombinant human procollagen type I C-terminal propeptide and establishment of a sandwich ELISA for quantification. *Sci. Rep.* 7, 15946 (2017).
33. Wong, M. Y., DiChiara, A. S., Suen, P. H., Chen, K., Doan, N. D., Shoulders, M. D., Adapting secretory proteostasis and function through the unfolded protein response. *Curr. Top. Microbiol. Immunol.* 414, 1-25 (2018).
34. Shoulders, M. D., Ryno, L. M., Genereux, J. C., Moresco, J. J., Tu, P. G., Wu, C., Yates, J. R., III, Su, A. I., Kelly, J. W., Wiseman, R. L., Stress-independent activation of XBP1s and/or ATF6 reveals three functionally diverse ER proteostasis environments. *Cell Rep.* 3, 1279-1292 (2013).
35. Nabavi, N., Pustylnik, S., Harrison, R. E., Rab GTPase mediated procollagen trafficking in ascorbic acid stimulated osteoblasts. *PLoS One* 7, e46265 (2012).

36. Ferris, S. P., Kodali, V. K., Kaufman, R. J., Glycoprotein folding and quality-control mechanisms in protein-folding diseases. *Dis. Model. Mech.* 7, 331-341 (2014).
37. Hebert, D. N., Foellmer, B., Helenius, A., Glucose trimming and reglucosylation determine glycoprotein association with calnexin in the endoplasmic reticulum. *Cell* 81, 425-433 (1995).
38. Elbein, A. D., Glycosidase inhibitors: inhibitors of N-linked oligosaccharide processing. *FASEB J.* 5, 3055-3063 (1991).
39. Keller, S. H., Lindstrom, J., Taylor, P., Inhibition of glucose trimming with castanospermine reduces calnexin association and promotes proteasome degradation of the α-subunit of the nicotinic acetylcholine receptor. *J. Biol. Chem.* 273, 17064-17072 (1998).
40. Saul, R., Molyneux, R. J., Elbein, A. D., Studies on the mechanism of castanospermine inhibition of α- and β-glucosidases. *Arch. Biochem. Biophys.* 230, 668-675 (1984).
41. Horiuchi, K., Tohmonda, T., Morioka, H., The unfolded protein response in skeletal development and homeostasis. *Cell. Mol. Life Sci.* 73, 2851-2869 (2016).
42. Saito, A., Ochiai, K., Kondo, S., Tsumagari, K., Murakami, T., Cavener, D. R., Imaizumi, K., Endoplasmic reticulum stress response mediated by the PERK-eIF2α-ATF4 pathway is involved in osteoblast differentiation induced by BMP2. *J. Biol. Chem.* 286, 4809-4818 (2011).
43. Matsuzaki, S., Hiratsuka, T., Taniguchi, M., Shingaki, K., Kubo, T., Kiya, K., Fujiwara, T., Kanazawa, S., Kanematsu, R., Maeda, T., Takamura, H., Yamada, K., Miyoshi, K., Hosokawa, K., Tohyama, M., Katayama, T., Physiological ER stress mediates the differentiation of fibroblasts. *PLoS One* 10, e0123578 (2015).
44. Hisanaga, S., Miyake, M., Taniuchi, S., Oyadomari, M., Morimoto, M., Sato, R., Hirose, J., Mizuta, H., Oyadomari, S., PERK-mediated translational control is required for collagen secretion in chondrocytes. *Sci. Rep.* 8, 773 (2018).
45. Yang, X., Matsuda, K., Bialek, P., Jacquot, S., Masuoka, H. C., Schinke, T., Li, L., Brancorsini, S., Sassone-Corsi, P., Townes, T. M., Hanauer, A., Karsenty, G., ATF4 is a substrate of RSK2 and an essential regulator of osteoblast biology; implication for Coffin-Lowry Syndrome. *Cell* 117, 387-398 (2004).
46. Tran, M. C., Wang, Z., Barbhaiya, N. J., Butler, P. D., Spiotto, M., Fattakhov, E., Rashidi, V., Longaker, M. T., Koong, A., Yang, G. P., Inhibition of the unfolded protein response delays wound healing by decreasing collagen. *J. Am. Coll. Surg.* 213, S90 (2011).
47. Wang, Z., Butler, P., Ly, D., Spiotto, M., Koong, A., Yang, G., Activation of the unfolded protein response in wound healing. *J. Surg. Res.* 158, 209 (2010).
48. Ishikawa, T., Toyama, T., Nakamura, Y., Tamada, K., Shimizu, H., Ninagawa, S., Okada, T., Kamei, Y., Ishikawa-Fujiwara, T., Todo, T., Aoyama, E., Takigawa, M., Harada, A., Mori, K., UPR transducer BBF2H7 allows export of type II collagen in a cargo- and developmental stage-specific manner. *J. Cell Biol.* 216, 1761-1774 (2017).
49. Ishikawa, T., Okada, T., Ishikawa-Fujiwara, T., Todo, T., Kamei, Y., Shigenobu, S., Tanaka, M., Saito, T. L., Yoshimura, J., Morishita, S., Toyoda, A., Sakaki, Y., Taniguchi, Y., Takeda, S., Mori, K., ATF6α/β-mediated adjustment of ER chaperone levels is essential for development of the notochord in medaka fish. *Mol. Biol. Cell* 24, 1387-1395 (2013).
50. Bachar-Wikstrom, E., Manchanda, M., Bansal, R., Karlsson, M., Kelly-Pettersson, P., Skoldenberg, O., Wikstrom, J. D., Endoplasmic reticulum stress in human chronic wound healing: Rescue by 4-phenylbutyrate. *Int. Wound J.* 10.1111/iwj.13525 (2020).
51. Aycock, R. S., Raghow, R., Stricklin, G. P., Seyer, J. M., Kang, A. H., Post-transcriptional inhibition of collagen and fibronectin synthesis by a synthetic homolog of a portion of the carboxyl-terminal propeptide of human type I collagen. *J. Biol. Chem.* 261, 14355-14360 (1986).

52. Katayama, K., Seyer, J. M., Raghow, R., Kang, A. H., Regulation of extracellular matrix production by chemically synthesized subfragments of type I collagen carboxy propeptide. *Biochemistry* 30, 7097-7104 (1991).
53. Katayama, K., Armendariz-Borunda, J., Raghow, R., Kang, A. H., Seyer, J. M., A pentapeptide from type I procollagen promotes extracellular matrix production. *J. Biol. Chem.* 268, 9941-9944 (1993).
54. Davies, D., Tuckwell, D. S., Calderwood, D. A., Weston, S. A., Takigawa, M., Humphries, M. J., Molecular characterisation of integrin-procollagen C-propeptide interactions. *Eur. J. Biochem.* 246, 274-282 (1997).
55. Smedsrød, B., Melkko, J., Risteli, L., Risteli, J., Circulating C-terminal propeptide of type I procollagen is cleared mainly via the mannose receptor in liver endothelial cells. *Biochem. J.* 271, 345-350 (1990).
56. Omari, S., Makareeva, E., Roberts-Pilgrim, A., Mirigian, L., Jarnik, M., Ott, C., Lippincott-Schwartz, J., Leikin, S., Noncanonical autophagy at ER exit sites regulates procollagen turnover. *Proc. Natl. Acad. Sci. U.S.A.* 115, E10099-E10108 (2018).
57. Ishida, Y., Yamamoto, A., Kitamura, A., Lamandé, S. R., Yoshimori, T., Bateman, J. F., Kubota, H., Nagata, K., Autophagic elimination of misfolded procollagen aggregates in the endoplasmic reticulum as a means of cell protection. *Mol. Biol. Cell* 20, 2744-2754 (2009).
58. Fitzgerald, J., Lamandé, S. R., Bateman, J. F., Proteasomal degradation of unassembled mutant type I collagen pro-α1(I) chains. *J. Biol. Chem.* 274, 27392-27398 (1999).
59. Iwamoto, M., Björklund, T., Lundberg, C., Kirik, D., Wandless, T. J., A general chemical method to regulate protein stability in the mammalian central nervous system. *Chem. Biol.* 17, 981-988 (2010).
60. Kabeya, Y., Mizushima, N., Ueno, T., Yamamoto, A., Kirisako, T., Noda, T., Kominami, E., Ohsumi, Y., Yoshimori, T., LC3, a mammalian homologue of yeast Apg8p, is localized in autophagosome membranes after processing. *EMBO J.* 19, 5720-5728 (2000).

CHAPTER 3:

REFINING MECHANISTIC UNDERSTANDING OF THE CONTRIBUTIONS OF COLLAGEN'S DOMAINS TO TRIMERIC ASSEMBLY

AUTHOR CONTRIBUTIONS

Rasia C. Li, Andrew S. DiChiara, and Matthew D. Shoulders designed research; R.C.L. performed research; R.C.L. and A.S.D. contributed reagents.

ABSTRACT

The fibrillar collagens are large, complex, multi-domain, oligomeric proteins. Although they make up the majority of the protein in the human body, the molecular features that govern their folding and assembly are not well-understood. We previously proposed a model of collagen assembly using data obtained from expression and purification of collagen-I C-Pro domains in the absence of other domains. This model suggested two key features of C-Pro assembly: First, we showed that Ca^{2+} mediates the non-covalent assembly of collagen-I C-Pro domains in all possible stoichiometries. Second, data collected using C-Pro domains in isolation suggested that the presence or absence of interchain disulfide-forming cysteine residues is likely key to template stable triple helix formation. Here, we explore the applicability of this model to full-length collagen, as all prior work was conducted with either C-Pro domains in isolation or with engineered mini-collagen constructs. We find that the ability of a collagen strand to stably homotrimerize is not solely reliant on the cysteine code and is unexpectedly influenced by factors outside the C-Pro domain. We also confirm that the absence of an interchain disulfide-forming cysteine in non-homotrimerizing collagen strands is critical to prevent formation of covalently-linked, unstable trimers and to maintain availability of the monomer for heterotrimer formation. Collectively, these results advance our understanding of collagen assembly while raising additional questions regarding how Nature engineers the correct assembly of such a complex protein.

INTRODUCTION

The fibrillar collagens are the most abundant class of proteins in the human body, providing the structural scaffold of our connective tissues.[1] They are characterized by a single triple-helical domain bookended at the N- and C-termini by short telopeptides (N-telo and C-telo) and globular propeptide domains (N-Pro and C-Pro) (Fig. 3-1A).[2] The propeptide domains are proteolytically cleaved prior to supramolecular assembly of individual triple helices into fibrils, while the telopeptides form covalent, inter-helix crosslinks to stabilize the fibrils.[3]

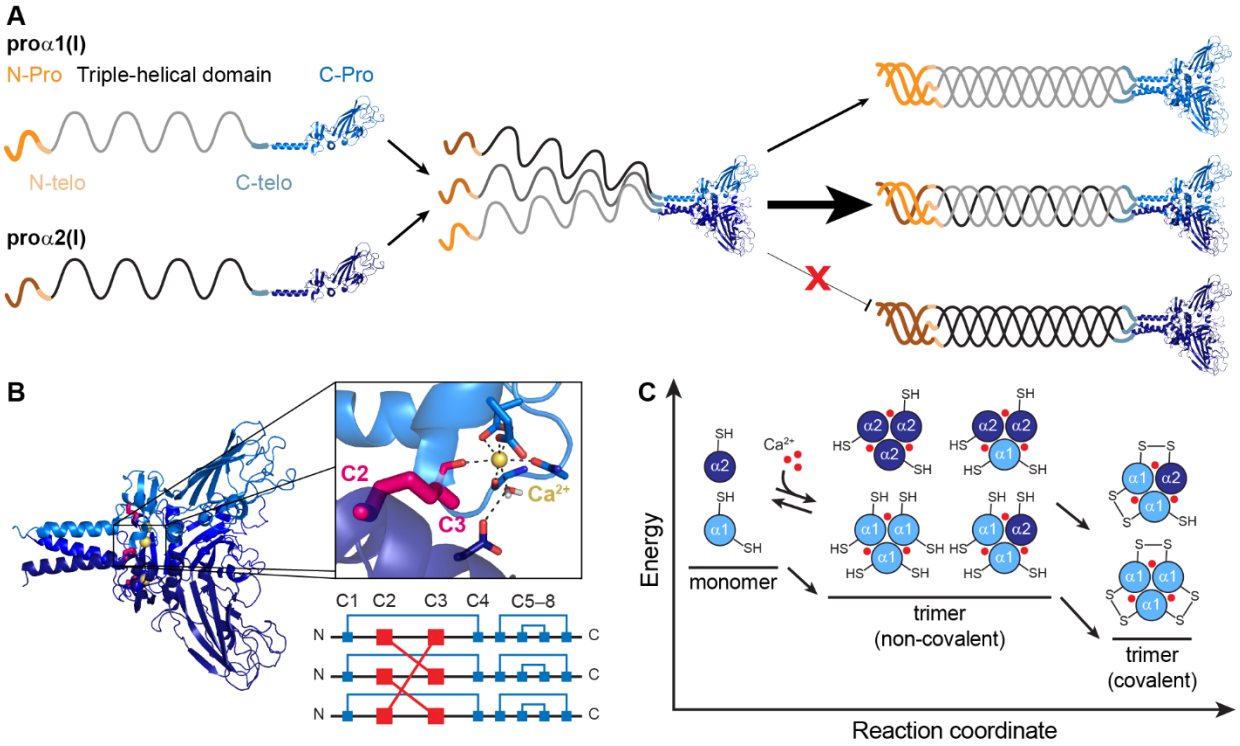


Figure 3-1: Fibrillar collagens form homo- or heterotrimers guided by the C-Pro domain. (A) Collagen folding begins at the globular C-terminal propeptide (C-Pro) domain, followed by zipper-like folding of the triple-helical domain. Collagen-I preferentially forms 2:1 proα1(I):proα2(I) heterotrimers. Proα1(I) homotrimers are a minor product, while proα2(I) homotrimers have never been observed. (B) A crystal structure of the C-Proα1(I) homotrimer shows that Ca^{2+} binds at the interface between monomers and that C2 and C3 form interchain disulfide bonds. The schematic below shows the disulfide bonding pattern of the eight cysteines in the C-Proα1(I) homotrimer. (C) Our previously proposed model for collagen assembly describes transient, Ca^{2+}-mediated, non-covalent association of all stoichiometries of C-Pro domains, followed by covalent immortalization of the species that are able to form disulfide bonds between all three subunits.

There are seven types of fibrillar collagens, three of which form heterotrimers, resulting in a total of eleven different human fibrillar collagen strands.[2] One of the most remarkable features of collagen biosynthesis is the formation of trimeric molecules that are the correct type and correct

stoichiometry, particularly when multiple fibrillar collagens are expressed in a single cell. For example, collagen-I is almost always found as a 2:1 α1(I):α2(I) heterotrimer, despite the ability of the α1(I) strand to form highly stable homotrimers.

This process of chain selection is generally attributed to the C-Pro domain, as the fibrillar collagens begin folding at their C-termini.[4] The C-Pro domains are thought to assemble into a disulfide-linked trimer first, after which zipper-like folding of the triple helix takes place (Fig 3-1A).[5] The question of how the ~30 kDa (as a monomer) globular C-Pro domain governs assembly of the much larger and highly repetitive ~100 kDa triple-helical region into the desired trimeric products has intrigued researchers since its discovery, and various primary structural features have been implicated in the process of chain selection.[6, 7]

Our previous work identified roles for two features of the fibrillar collagen C-Pro domains in governing trimeric assembly patterns.[8] First, we showed that Ca^2 binding between C-Pro monomers is critical for templating non-covalent assembly. Second, we found that the highly conserved network of cysteine residues within the C-Pro domain predicts what stoichiometric forms of collagen homo- and heterotrimers exist biologically – a so-called cysteine code.

The ability of the collagen C-Pro domain to bind Ca^{2+} was unappreciated until the publication of crystal structures of C-Proα1(III) and C-Proα1(I) homotrimers.[7, 9] Both of these structures revealed a highly conserved Ca^{2+}-binding site located between C-Pro monomers involving five residues from one strand and one residue from its neighbor (Fig. 3-1B). Our work demonstrated, unexpectedly and significantly, that the presence of Ca^{2+} promotes non-covalent, trimeric association of both C-Proα1(I) and C-Proα2(I) into all possible stoichiometries, even in the absence of interchain disulfide bonds. When Ca^{2+} is removed, the non-covalent assemblies dissociate into monomers.[8]

We additionally showed that the presence or absence of interchain disulfide-forming cysteine residues predicts the formation of stable, disulfide-linked homo- and heterotrimers for all fibrillar collagen C-Pro domains. Collagen strands that are known to homotrimerize have eight cysteines within their C-Pro domain, whereas those that are only found in heterotrimers have seven, with the second or third cysteine (C2 or C3) missing.[8] Crystal structures of C-Proα1(III)[9] (a naturally occurring homotrimer) and homotrimeric C-Proα1(I)[7] show that C2 and C3 form interchain disulfide bonds between the C-Pro monomers (Fig. 3-1B), suggesting that α2(I) and other collagen strands with seven cysteines cannot form homotrimers due to their inability to form interchain disulfide bonds.

Early work using mini collagens and a cell-free expression system suggested that simply substituting the serine in position two of proα2(I) with a cysteine was insufficient to generate disulfide-linked homotrimers.[10] However, our work demonstrated that the serine to cysteine substitution is actually sufficient to drive the formation of disulfide-linked C-Proα2(I) homotrimers, at least in the absence of the other collagen domains. This pattern held true for the other heterotrimer-forming collagen C-Pro domains. Inversely, substitution of C2 or C3 for serine in homotrimerizing C-Pro domains abolished their ability to form disulfide-linked homotrimers.[8]

Collectively, these data suggested a model by which chain-selective collagen assembly could occur. In Ca^{2+}-rich environments, such as the ER, Ca^{2+} templates non-covalent trimeric assembly of all stoichiometric configurations of collagen-I C-Pro domains. Subsequently, disulfide bonds form between adjacent subunits, providing a thermodynamic sink that stabilizes the trimers for triple helix formation (Fig. 3-1C).[8] This model would explain why 2:1 α1(I):α2(I) heterotrimers and α1(I) homotrimers are the only forms of collagen-I observed—they have sufficient cysteines to form disulfide bonds linking all three monomers, while 1:2 α2(I):α2(I) heterotrimers and α2(I) homotrimers do not.[8]

This model also helps explain the discrepancy between our data and the data obtained from the mini collagen system regarding the introduction of C2 in α2(I). The mini collagens were expressed in an *in vitro* translation system that required pre-translation treatment with Ca^{2+} chelators.[6, 10, 11] Thus, the absence of Ca^{2+} likely prevented non-covalent C-Pro association that precedes disulfide bond formation. Nevertheless, there remains the possibility that the N-Pro, telopeptides, or the truncated triple-helical domain present in the mini collagen system impact assembly.

A key implication of our model, although not yet experimentally tested, is that covalent linkage of three collagen strands ultimately governs the formation of triple helices. Notably, previous work in the mini-collagen system suggested that interchain disulfide bond formation is not universally required for collagen assembly, at least in the mini-collagens.[10] Moreover, a model based solely on disulfide bond formation begs the question of why the heterotrimer is the preferred form of collagen-I. The C-Proα1(I) homotrimer has an additional disulfide bond compared to the heterotrimer, implying that it is a more stable configuration. In addition, the α1(I) homotrimer triple helix was previously shown to be more thermally stable than the 2:1 heterotrimer.[12] Hence, the cysteine code on its own is not sufficient to fully explain patterns of collagen assembly beyond the ability of a given C-Pro domain to form a disulfide-linked trimer. Ca^{2+}-mediated dynamic assembly may play an even more important, not yet fully appreciated role.

In this chapter, we explore the applicability of our model in Figure 3-1C to full-length procollagen, where all domains are present. We found that, while substitution of the serine in position 2 of proα2(I) with cysteine does indeed lead to disulfide-linked trimer formation, this substitution does not yield stable, pepsin-resistant collagen triple helices. Further, substitution of the cysteine in position 2 of proα1(I) with serine abrogated disulfide-linked trimer formation, yet the resultant collagen assemblies were pepsin-resistant and therefore likely triple-helical. These results suggest that Ca^{2+}-mediated non-covalent assembly is critical for triple helix formation and that regions outside the C-Pro domain play unanticipated roles in governing collagen assembly patterns.

RESULTS

Covalent linkage of C-Proα2(I) is insufficient to drive formation of pepsin-resistant proα2(I) triple helices. Our previous work revealed that introduction of a cysteine at the C2 position of C-Proα2(I) drives formation of a covalently linked C-Proα2(I) homotrimer.[8] To determine whether this result is consistent in the context of the full-length proα2(I) protein, we expressed HA-tagged wild-type (WT) proα2(I) and S2C proα2(I) in HEK293T cells and assessed the secreted collagen using reducing and non-reducing immunoblots. We found that WT proα2(I) does not form disulfide-linked assemblies, as expected. Consistent with the C-Pro domain expressed in isolation, S2C proα2(I) forms disulfide-linked trimers (Fig. 3-2A).

We next assessed whether the proα2(I) variants form stable triple helices. Properly folded collagen triple helices are generally resistant to protease digestion, whereas misfolded or unfolded collagen strands are sensitive to protease digestion. This behavior is well-established using pepsin (Fig. 3-2B).[13] We precipitated collagen from the media of proα2(I) and S2C proα2(I)-expressing cells and subjected it to pepsin digestion at 4 °C overnight. We found that, in contrast to the properly folded collagen synthsized by primary fibroblasts, neither proα2(I) nor S2C proα2(I) formed substantive amounts of pepsin-resistant triple helices (Fig. 3-2C). This result indicates that covalent linkage of the C-Pro domain of proα2(I) is insufficient to produce stable, homotrimeric proα2(I) triple helices.

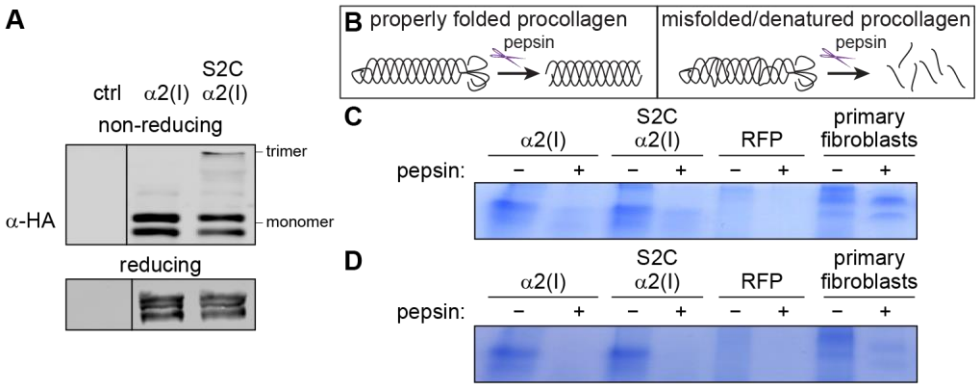


Figure 3-2: S2C α2(I) forms disulfide-linked trimers but does not effectively form pepsin-resistant helices.

(A) WT proα2(I) is expressed as a monomer, while S2C proα2(I) forms disulfide-linked trimers. (B) Properly folded collagen molecules are resistant to pepsin digestion, whereas misfolded collagen molecules will be digested by pepsin. (C) Both α2(I) and S2C α2(I) form little to no pepsin-resistant homotrimers compared to WT collagen secreted from primary fibroblasts. (D) Even at 30 °C, neither α2(I) nor S2C α2(I) form pepsin-resistant homotrimers.

Previous work suggests that α2(I) homotrimers formed by *in vitro* thermal renaturation have a melting temperature of only ~20 °C.[14] Hence, even if disulfide-linked S2C proα2(I) is able to template triple helix formation, the triple helix itself may not be able to fold at physiological temperatures. To assess this possibility, we cultured cells expressing WT proα2(I) and S2C proα2(I) at 30 °C. [Unfortunately, culturing cells at 25 °C or lower temperatures inhibited growth and substantially diminished collagen-I secretion (Fig. S3-1).] We subjected the secreted collagen produced at 30 °C to pepsin digestion overnight at 4 °C and found that both proα2(I) and S2C proα2(I) were still unable to form stable, pepsin-resistant homotrimers (Fig. 3-2D).

Covalent linkage of C-Proα1(I) is not required for formation of pepsin-resistant proα1(I) homotrimers. We next asked whether the reverse was true—would the absence of covalent linkage in proα1(I) prevent stable triple helix formation? We expressed WT proα1(I) and C2S proα1(I) and examined the secreted collagen. Proα1(I) formed the expected disulfide-linked trimers. However, we were quite surprised to find that C2S proα1(I) formed an apparent disulfide-linked dimer (Fig. 3-3A).

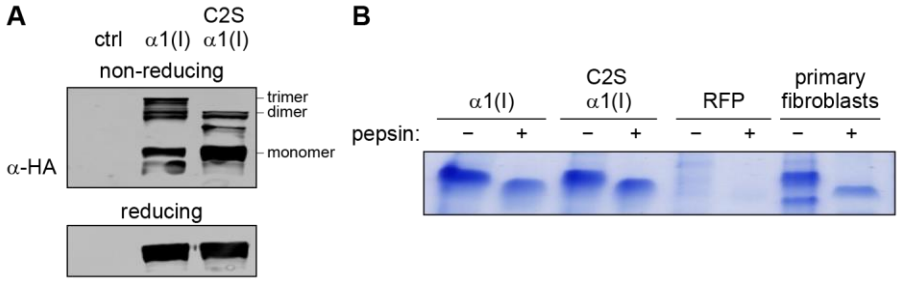


Figure 3-3: C2S α1(I) does not form disulfide-linked trimers yet is able to form pepsin-resistant triple helices.

(A) WT proα1(I) is expressed as disulfide-linked trimers, as expected. C2S proα1(I) forms an unexpected disulfide-linked dimer. (B) Both α1(I) and C2S α1(I) efficiently form pepsin-resistant homotrimers.

Removal of C2 from proα1(I) leaves C3 unpaired. However, given the location of C3 in a properly folded C-Proα1(I) monomer, it seems unlikely that C3–C3 disulfide bonds would be able to form in a properly folded trimer. The only other cysteines within the proα1(I) chain are located in the N-Pro domain. However, the crystal structure of the collagen-IIA N-Pro domain, which has 61% sequence identity with N-Proα1(I), including conservation of all ten cysteines, suggests that all ten N-Pro cysteines form intrachain disulfide bonds.[15] We did not observe disulfide-linked dimer formation of C2S C-Proα1(I) without the other domains, suggesting that the presence of other domains may help drive formation of an incorrect disulfide linkage within C-Pro or N-Pro domain.

Alternatively, N-Proα1(I) may have a different disulfide linkage pattern than N-Proα1(IIA). Further investigation is required to determine the origins of this disulfide-linked dimer.

Bearing in mind this unexpected observation, we precipitated WT proα1(I) and C2S proα1(I) from media and digested the proteins with pepsin. We found that both WT proα1(I) and C2S proα1(I) formed pepsin-resistant triple helices (Fig. 3-3B). This finding suggests that covalent linkage via the C2–C3 disulfide bonds in the C-Pro domain is not essential for homotrimeric α1(I) assembly. Likely, the triple-helical domain of α1(I) is able to self-assemble when the three strands are brought into close proximity, as has been observed during *in vitro* renaturation.[14, 16] This finding is consistent with what was previously observed with mini-proα1(III) in the mini collagen system.[11] Ca^{2+}-guided, non-covalent trimerization of the C-Pro domain is likely still essential to expedite this process and ensure correct register of the triple helix. Another possible explanation for this result is that the unexplained disulfide-linked dimer formed by C2S proα1(I) is somehow able to drive triple helix formation, despite misassembly.

Swapping C-Pro domains does not alter assembly behavior of triple-helical domains. We next evaluated whether the assembly behavior of proα1(I), proα2(I), and their cysteine variants is dependent upon elements of the C-Pro domain other than the interstrand disulfide-forming cysteine residues. To do so, we created chimeric procollagen constructs in which the C-Pro domains of proα1(I) and proα2(I) are swapped: α1(I).C-Proα2(I) consists of the α1(I) protein from N-terminus to the C-Pro cleavage site followed by C-Proα2(I) and vice versa for α2(I).C-Proα1(I) (Fig. 3-4A).

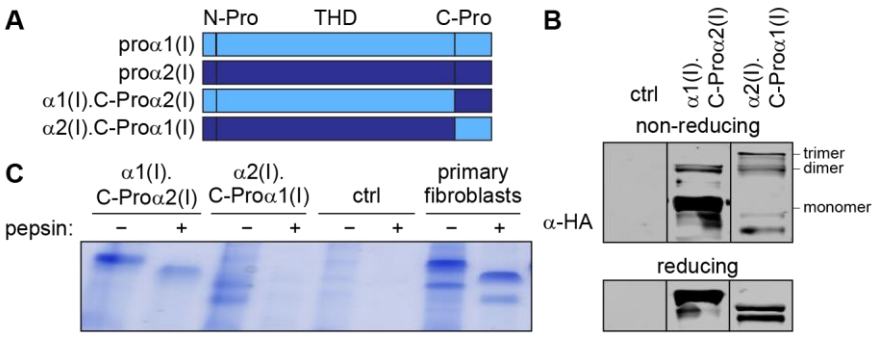


Figure 3-4: Chimeric collagens behave consistently with their N-Pro and triple-helical domains, rather than their C-Pro domains.

(A) We constructed chimeric collagens in which the C-Pro domains were swapped between proα1(I) and proα2(I). (B) Expression of α1(I).C-Proα2(I) and α2(I).C-Proα1(I) resulted in disulfide-linked dimers and trimers, respectively, following the patterns seen with C2S α1(I) and S2C α2(I). (C) α1(I).C-Proα2(I) formed pepsin-resistant triple helices, while α2(I).C-Proα1(I) did not, again consistent with C2S α1(I) and S2C α2(I).

We expressed HA-tagged α1(I).C-Proα2(I) and α2(I).C-Proα1(I) in HEK293T cells and examined the secreted collagen. As expected, α2(I).C-Proα1(I) formed disulfide-linked trimers. α1(I).C-Proα2(I) formed disulfide-linked dimers similar to C2S α1(I) but unlike WT proα2(I) (Fig. 3-4B). This result further suggests that some feature of the proα1(I) strand outside of the C-Pro domain can drive incorrect disulfide bond formation in either the N-Pro or C-Pro domains.

Pepsin digestion of the precipitated chimeras demonstrated once again that the α1(I) triple-helical domain can form pepsin-resistant helices without correct C2–C3 covalent linkage of the C-Pro domain, while the α2(I) triple-helical domain cannot (Fig. 3-4C). Other sequence elements within the two different C-Pro domains do not alter triple helix assembly patterns. This result is consistent with previous data obtained from α2(I).C-Proα1(I) chimeras expressed in insect cells.[17]

DISCUSSION

Our previous model for Ca^{2+}-guided and disulfide-mediated C-Pro assembly provided an explanation for the two species of collagen-I observed but did not comprehensively address all questions concerning collagen assembly patterns.[8] In particular, we hypothesized that interchain disulfide bond formation may be necessary to template triple helix assembly, which would explain why proα2(I) homotrimers and 1:2 proα1(I):proα2(I) heterotrimers are not observed despite their ability to non-covalently assemble.

Here, we directly investigate this hypothesis by expressing cysteine variants of full-length proα1(I) and proα2(I). We observed that interchain disulfide bond formation is not, in fact, essential for triple helix assembly. The C2S variant of proα1(I) and the α1(I).C-Proα2(I) chimera are both able to form pepsin-resistant helices, despite not being able to form the correct interchain disulfide bonds. Non-covalent association of proα1(I) appears to be sustained long enough to nucleate triple-helix folding and lead to the formation of pepsin-resistant helices.

One possible explanation for the ability of proα1(I) to form pepsin-resistant helices without covalent linkage is that non-covalent C-Pro association is sustained long enough to nucleate triple helix folding. Proline residues are critical for stabilizing collagen's triple helical structure, and those found in the Yaa position of the collagen Gly-Xaa-Yaa triplet are typically hydroxylated, which further stabilizes the triple helix.[3] The α1(I) chain has 236 proline residues, 49.6% of which are in the Yaa position, compared to α2(I), which has 203 proline residues, 46.3% of which are in the Yaa position. The higher proline and hydroxyproline content in proa1(I) may preorganize the proα1(I) strand such that it has higher innate trimerization potential and can more readily initiate triple helix assembly even during transient C-Pro association. Notably, the most C-terminal 75 amino acids of the α1(I) triple-helical domain, which has been shown to fold rapidly *in vitro* for α1(III), has five more Yaa position prolines than the same region of α2(I).[4]

The disulfide-linked dimer observed in C2S proα1(I) and α1(I).C-Proα2(I) raises questions about how proα1(I) is still able to form properly folded and aligned helices despite almost certainly containing an incorrect interchain disulfide bond. Whether the dimer itself drives triple helix folding or whether triple helix folding somehow induces dimer formation is thus far unknown. Proper alignment of the triple helix of C2S proα1(I) and α1(I).C-Proα2(I), exhibited by the pepsin-resistant fragment of the correct size, suggests that the C-Pro domains are still folded in a way that allows for nucleation and zipper-like folding. While formation of interchain disulfide bonds between two free C3 residues within the C-Pro domain is an obvious explanation, doing so would disrupt the existing interface between monomers, including the Ca^{2+} binding site, likely resulting in an entirely

different assembly. Alternatively, interchain disulfide formation or shuffling could occur after triple helix formation is initiated in either the C-Pro or N-Pro domains. This explanation appears more likely, given that WT proα2(I) does not form a disulfide-linked dimer but α1(I).C-Proα2(I) does. Proα1(III) has two additional cysteine residues in its N-Pro domain, which are not present in N-Proα1(I) or N-Proα1(IIA), that form interchain disulfide bonds after triple-helical folding is complete.[18] There remains the possibility that the N-Proα1(I) domain either forms different disulfide linkages than N-Proα1(IIA)[15] or can form an incorrect interchain disulfide bond after assembly.

Meanwhile, covalent linkage of proα2(I) into homotrimers via a S2C substitution or the C-Proα1(I) domain was insufficient to induce formation of a pepsin-resistant triple helix at physiological temperatures. Proline and hydroxyproline content likely also provide an explanation for this observation. The α2(I) chain may be unable to form a stable triple helix at physiological temperatures simply due to its primary structure, regardless of C-Pro domain stabilization. α2(I) homotrimers renatured *in vitro* have a melting temperature >10 °C lower than that of renatured α1(I) homotrimers.[14] In addition, 2:1 α1(I):α2(I) heterotrimers generally exhibit lower thermal stability and are faster to denature than α1(I) homotrimers.[12, 16, 19-21] A proline-rich region at the N-terminus of C-Proα2(I) has also been proposed to prevent α2(I) homotrimer formation,[22] though computational modeling suggests a similarly proline-rich sequence in the α1(I) C-telo domain would impact α1(I) homotrimerization.[23]

Alternatively, hydroxyproline-poor regions have been associated with local microunfolding sites that are more sensitive to protease cleavage, and these sites differ between the α1(I) homotrimer and 2:1 heterotrimer.[12] The α2(I) homotrimer may have additional and overall more microunfolding sites that render it susceptible to pepsin digestion even when properly folded into a triple helix. Orthogonal methods of measuring triple-helical structure may be more revealing of the true structure of S2C α2(I).

We note that the challenges associated with expressing collagen recombinantly required that our experiments be performed in HEK293T cells, which are not known to express collagen-I endogenously.[24] Our data demonstrate that we are able to form pepsin-resistant triple-helical proα1(I) homotrimers, suggesting these cells are able to appropriately modify and fold collagen-I. However, there remains the possibility that HEK293T cells do not have all of the biosynthetic machinery required for proper collagen folding and assembly.

With these caveats in mind, our results suggest that the ability of a given collagen strand to homotrimerize is not based entirely on the cysteine code, as previously suggested, but also on

regions outside the C-Pro domain. This finding indicates that further investigation of why the cysteine code is so well-conserved amongst the human fibrillar collagens and apparently evolutionarily tied to the appearance of collagen heterotrimers is necessary.[8]

One possible explanation for the loss of the C2 or C3 in heterotrimerizing collagens is that the absence of C2 or C3 may allow cells to maintain availability of the heterotrimerizing monomer. Our data show that S2C proα2(I) forms disulfide-linked trimers, which would trap proα2(I) strands in covalent complexes and make them unavailable for heterotrimer incorporation. Our initial work with isolated C-Pro domains showed that, even in the presence of excess C-Proα1(I), S2C C-Proα2(I) still forms disulfide-linked trimers.[8] Provided this result carries over with full-length proα2(I), the absence of C2 or C3 may be critical to ensure efficacious heterotrimer formation.

This explanation is not fully satisfying in that it does not explain why only some strands lack a cysteine while others don't, particularly if the cysteines are not required to template triple helix folding. Specifically, why loss of a cysteine is coupled with the less thermally stable collagen strand is unclear, as it leaves the α1(I) strand capable of forming highly irreversible, covalently-linked homotrimers with stable triple helices. The advantages of the proα1(I) being able to form such stable structures is particularly confounding, when the heterotrimer is the dominant and desired form. In fact, formation of proα1(I) homotrimers appears to be largely undesirable, as the absence of α2(I) in collagen tissues contribute to disease phenotypes.[25-28]

The mechanisms governing Nature's preference for the 2:1 α1(I):α2(I) heterotrimer despite the greater simplicity and stability (of both the covalently linked C-Pro domain and the triple-helical domain) of the α1(I) homotrimer have remained enigmatic through the decades of collagen research. Our results provide further insight into this fascinating molecular puzzle but emphasize the additional unknowns.

Heterotrimeric chain selection has generally been attributed to the C-Pro domain, but the specific interactions that are involved are only speculatively understood.[6, 7] As for the triple-helical domain, mixed results have been obtained *in vitro* as to whether α1(I):α2(I) heterotrimers or α1(I) homotrimer are preferentially formed.[12, 14] One study co-expressed a C3S variant of proα2(I), which lacks both interchain disulfide-forming cysteines, with WT proα1(I) and found that this variant was able to incorporate into heterotrimers.[29] This finding suggests that covalent linkage of C-Proα2(I) to C-Proα1(I) is also not essential for heterotrimeric triple helix formation. A natural follow-up using the constructs described in this chapter would be co-expression of the proα1(I) and proα2(I) variants to determine whether preference for the heterotrimer is driven primarily by

the C-Pro domain or other domains. Similarly, co-expression experiments could be used to examine whether a 1:2 α1(I):α2(I) heterotrimer is stable.

Ultimately, the data described here provide key refinements to the model we previously proposed for collagen assembly based on the C-Pro domain alone. Unlike previously suggested,[8] the ability of a given collagen strand to stably homo- or heterotrimerize is not based on the presence or absence of a single cysteine, but instead on other properties within collagen sequence. However, conservation of the cysteine code suggests that the cysteine residues must still be critical for assembly, potentially by preventing formation of unstable disulfide-linked trimers. Moreover, these data strengthen the importance of non-covalent C-Pro assembly, which we previously showed was guided by Ca^{2+} binding, and here is shown to be sufficient to template triple helix assembly of certain collagens. These explanations do not satisfactorily explain all aspects of collagen assembly, particularly the observed preference for the formation of 2:1 α1(I):α2(I) heterotrimers over more stable α1(I) heterotrimers. Thus, our work sets the stage for an abundance of follow-up studies with the ultimate goal of understanding and predicting collagen assembly at the molecular level.

MATERIALS AND METHODS

Cell Culture. HEK293T cells (ATCC) were cultured in complete DMEM media (Corning) supplemented with 10% fetal bovine serum (FBS), 100 IU penicillin, 100 μg/mL streptomycin, and 2 mM L-glutamine (Corning). Primary fibroblasts (GM05294, Coriell) were cultured in complete MEM media (Corning) supplemented with 15% fetal bovine serum (FBS), 100 IU penicillin, 100 μg/mL streptomycin, and 2 mM L-glutamine (Corning).

Vector Construction. Site-directed mutagenesis was performed on previously described PPT.HA.proα1(I).pENTR1A plasmids using the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent). Gibson assembly using NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs) was used to convert the FLAG tag from previously described PPT.FLAG.proα2(I).pENTR1A plasmids to an HA tag for consistency.[30] Domain swap chimeras were also created with Gibson assembly. Restriction sites within the native collagen gene were used to generate fragments when possible, to avoid PCR amplification of the highly repetitive, GC-rich triple-helical region. pENTR1A vectors were recombined into pcDNA-DEST40 vectors for expression in HEK293T cells.

Collagen construct expression. HEK293T cells were plated at a density of 7.5×10^5 cells/well in 6-well plates or 1.875×10^6 cells/60-mm dish the day before transfection. Cells were transfected using Lipofectamine 3000 (Thermo Fisher Scientific) according to the manufacturer's instructions. The following day, media was changed to fresh DMEM with 50 μM sodium ascorbate to induce procollagen expression and hydroxylation. Media was harvested for immunoblotting or precipitation approximately 24 h after ascorbate addition. For primary fibroblasts controls, cells were provided with fresh MEM with 200 μM sodium ascorbate 24 h prior to harvesting.

Immunoblotting. For non-reducing blots, media was treated with 100 mM iodoacetamide (VWR) for 30 min in the dark to prevent disulfide shuffling. Samples were then treated with 6× gel loading buffer (300 mM Tris, pH 6.8, 15% glycerol, 6% SDS, and 10% (w/v) bromophenol blue) and boiled for 10 min. For reducing blots, samples were directed treated with 6× gel loading buffer was supplemented with 167 mM DTT and boiled for 10 min. All samples were run on homemade 4/8% polyacrylamide gels. Immunoblots were probed with anti-HA primary antibody (ThermoFisher 26183, 1:10000) followed by 800CW or 680LT secondary antibodies (LI-COR). Images were obtained using an Odyssey infrared imager (LI-COR).

Pepsin digestions. Media from transfected 60-mm dishes was split into two equal aliquots and precipitated with 176 mg/mL ammonium sulfate or 25% (v/v) saturated ammonium sulfate overnight at 4 °C, rotating end-over-end. Samples were centrifuged at max speed for 30 min to pellet the protein, and supernatant was removed. The pellets were resuspended in 100 μL

of 0.5 M acetic acid either with or without 0.1 mg/mL pepsin (Amresco). Samples were digested overnight at 4 °C. Pepsin digestion was terminated by lyophilization. Lyophilized samples were then resuspended and run on homemade 8% SDS-PAGE gels, followed by staining with Coomassie Brilliant Blue R-250 (Ameresco).

SUPPORTING FIGURES

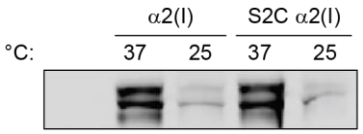


Figure S3-1: Proα2(I) and S2C proα2(I) secrete poorly at 25 °C.

REFERENCES

1. Mouw, J. K., Ou, G., Weaver, V. M., Extracellular matrix assembly: a multiscale deconstruction. *Nat. Rev. Mol. Cell Biol.* 15, 771-785 (2014).
2. Ricard-Blum, S., The collagen family. *Cold Spring. Harb. Perspect. Biol.* 3, a004978 (2011).
3. Shoulders, M. D., Raines, R. T., Collagen structure and stability. *Annu. Rev. Biochem.* 78, 929-958 (2009).
4. Bachinger, H. P., Bruckner, P., Timpl, R., Prockop, D. J., Engel, J., Folding mechanism of the triple helix in type-III collagen and type-III pN-collagen. Role of disulfide bridges and peptide bond isomerization. *Eur. J. Biochem.* 106, 619-632 (1980).
5. Engel, J., Prockop, D. J., The zipper-like folding of collagen triple helices and the effects of mutations that disrupt the zipper. *Annu. Rev. Biophys. Biophys. Chem.* 20, 137-152 (1991).
6. Lees, J. F., Tasab, M., Bulleid, N. J., Identification of the molecular recognition sequence which determines the type-specific assembly of procollagen. *EMBO J.* 16, 908-916 (1997).
7. Sharma, U., Carrique, L., Vadon-Le Goff, S., Mariano, N., Georges, R. N., Delolme, F., Koivunen, P., Myllyharju, J., Moali, C., Aghajari, N., Hulmes, D. J., Structural basis of homo- and heterotrimerization of collagen I. *Nat. Commun.* 8, 14671-14681 (2017).
8. DiChiara, A. S., Li, R. C., Suen, P. H., Hosseini, A. S., Taylor, R. J., Weickhardt, A. F., Malhotra, D., McCaslin, D. R., Shoulders, M. D., A cysteine-based molecular code informs collagen C-propeptide assembly. *Nat. Commun.* 9, 4206 (2018).
9. Bourhis, J. M., Mariano, N., Zhao, Y., Harlos, K., Exposito, J. Y., Jones, E. Y., Moali, C., Aghajari, N., Hulmes, D. J., Structural basis of fibrillar collagen trimerization and related genetic disorders. *Nat. Struct. Mol. Biol.* 19, 1031-1036 (2012).
10. Lees, J. F., Bulleid, N. J., The role of cysteine residues in the folding and association of the COOH-terminal propeptide of type-I and type-III procollagen. *J. Biol. Chem.* 269, 24354-24360 (1994).
11. Bulleid, N. J., Wilson, R., Lees, J. F., Type-III procollagen assembly in semi-intact cells: chain association, nucleation and triple-helix folding do not require formation of inter-chain disulphide bonds but triple-helix nucleation does require hydroxylation. *Biochem. J.* 317 (Pt 1), 195-202 (1996).
12. Kuznetsova, N. V., McBride, D. J., Leikin, S., Changes in thermal stability and microunfolding pattern of collagen helix resulting from the loss of alpha2(I) chain in osteogenesis imperfecta murine. *J. Mol. Biol.* 331, 191-200 (2003).
13. Bruckner, P., Prockop, D. J., Proteolytic enzymes as probes for the triple-helical conformation of procollagen. *Anal. Biochem.* 110, 360-368 (1981).
14. Tkocz, C., Kühn, K., The formation of triple-helical collagen molecules from alpha-1 or alpha-2 polypeptide chains. *Eur. J. Biochem.* 7, 454-462 (1969).
15. Xu, E. R., Blythe, E. E., Fischer, G., Hyvonen, M., Structural analyses of von Willebrand factor C domains of collagen 2A and CCN3 reveal an alternative mode of binding to bone morphogenetic protein-2. *J. Biol. Chem.* 292, 12516-12527 (2017).
16. Leikina, E., Mertts, M. V., Kuznetsova, N., Leikin, S., Type I collagen is thermally unstable at body temperature. *Proc. Natl. Acad. Sci. U. S. A.* 99, 1314-1318 (2002).
17. Myllyharju, J., Lamberg, A., Notbohm, H., Fietzek, P. P., Pihlajaniemi, T., Kivirikko, K. I., Expression of wild-type and modified proalpha chains of human type I procollagen in insect cells leads to the formation of stable [alpha1(I)]2alpha2(I) collagen heterotrimers and [alpha1(I)]3 homotrimers but not [alpha2(I)]3 homotrimers. *J. Biol. Chem.* 272, 21824-21830 (1997).

18. Bruckner, P., Bachinger, H. P., Timpl, R., Engel, J., Three conformationally distinct domains in the amino-terminal segment of type III procollagen and its rapid triple helix leads to and comes from coil transition. *Eur. J. Biochem.* 90, 595-603 (1978).
19. McBride, D. J., Jr., Choe, V., Shapiro, J. R., Brodsky, B., Altered collagen structure in mouse tail tendon lacking the alpha2(I) chain. *J. Mol. Biol.* 270, 275-284 (1997).
20. Miles, C. A., Sims, T. J., Camacho, N. P., Bailey, A. J., The role of the alpha2 chain in the stabilization of the collagen type I heterotrimer: a study of the type I homotrimer in oim mouse tissues. *J. Mol. Biol.* 321, 797-805 (2002).
21. Deak, S. B., van der Rest, M., Prockop, D. J., Altered helical structure of a homotrimer of alpha 1(I)chains synthesized by fibroblasts from a variant of osteogenesis imperfecta. *Coll. Relat. Res.* 5, 305-313 (1985).
22. McAlinden, A., Smith, T. A., Sandell, L. J., Ficheux, D., Parry, D. A., Hulmes, D. J., Alpha-helical coiled-coil oligomerization domains are almost ubiquitous in the collagen superfamily. *J. Biol. Chem.* 278, 42200-42207 (2003).
23. Malone, J. P., Alvares, K., Veis, A., Structure and assembly of the heterotrimeric and homotrimeric C-propeptides of type I collagen: significance of the alpha2(I) chain. *Biochemistry* 44, 15269-15279 (2005).
24. Thul, P. J., Akesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Bjork, L., Breckels, L. M., Backstrom, A., Danielsson, F., Fagerberg, L., Fall, J., Gatto, L., Gnann, C., Hober, S., Hjelmare, M., Johansson, F., Lee, S., Lindskog, C., Mulder, J., Mulvey, C. M., Nilsson, P., Oksvold, P., Rockberg, J., Schutten, R., Schwenk, J. M., Sivertsson, A., Sjostedt, E., Skogs, M., Stadler, C., Sullivan, D. P., Tegel, H., Winsnes, C., Zhang, C., Zwahlen, M., Mardinoglu, A., Ponten, F., von Feilitzen, K., Lilley, K. S., Uhlen, M., Lundberg, E., A subcellular map of the human proteome. *Science* 356 (2017).
25. Malfait, F., Symoens, S., Coucke, P., Nunes, L., De Almeida, S., De Paepe, A., Total absence of the alpha2(I) chain of collagen type I causes a rare form of Ehlers-Danlos syndrome with hypermobility and propensity to cardiac valvular problems. *J. Med. Genet.* 43, e36 (2006).
26. Nicholls, A. C., Osse, G., Schloon, H. G., Lenard, H. G., Deak, S., Myers, J. C., Prockop, D. J., Weigel, W. R., Fryer, P., Pope, F. M., The clinical features of homozygous alpha 2(I) collagen deficient osteogenesis imperfecta. *J. Med. Genet.* 21, 257-262 (1984).
27. Pace, J. M., Wiese, M., Drenguis, A. S., Kuznetsova, N., Leikin, S., Schwarze, U., Chen, D., Mooney, S. H., Unger, S., Byers, P. H., Defective C-propeptides of the proα2(I) chain of type I procollagen impede molecular assembly and result in osteogenesis imperfecta. *J. Biol. Chem.* 283, 16061-16067 (2008).
28. Pihlajaniemi, T., Dickson, L. A., Pope, F. M., Korhonen, V. R., Nicholls, A., Prockop, D. J., Myers, J. C., Osteogenesis imperfecta: cloning of a pro-alpha 2(I) collagen gene with a frameshift mutation. *J. Biol. Chem.* 259, 12941-12944 (1984).
29. Doyle, S. A., Smith, B. D., Role of the pro-alpha2(I) COOH-terminal region in assembly of type I collagen: disruption of two intramolecular disulfide bonds in pro-alpha2(I) blocks assembly of type I collagen. *J. Cell Biochem.* 71, 233-242 (1998).
30. DiChiara, A. S., Taylor, R. J., Wong, M. Y., Doan, N. D., Rosario, A. M., Shoulders, M. D., Mapping and exploring the collagen-I proteostasis network. *ACS Chem. Biol.* 11, 1408-1421 (2016).

CHAPTER 4:

PERSPECTIVES

Decades of research have gone into elucidating the mechanisms of collagen folding and assembly. Yet the work in this thesis illustrates that we are a long way from understanding how cells are able to maintain proper biosynthesis of this complex protein and that this process of understanding is often two steps forward, one step back. The work described in Chapter 2 provides an answer to a major question that was addressed inconclusively 30 years ago and ignored ever since—what is the purpose of the highly-conserved collagen *N*-glycan? A new consideration for collagen proteostasis has allowed us and others to demonstrate that the formerly enigmatic collagen-I *N*-glycan is actually a key player in maintaining collagen proteostasis and biosynthesis. Progress in this area has opened up many new exciting avenues for further research. The work described in Chapter 3 builds upon recent work from us and others that has brought substantial clarity to an even more fundamental question of collagen biosynthesis—how is the type-I collagen heterotrimer preferentially formed? Our latest findings raise just as many questions as they answer and inspires appreciation for the complexity of Nature's molecular designs.

Recent advances in collagen research have unfolded alongside increased appreciation for proteostasis as a whole. The complexity of the ER proteostasis network and its influence on protein folding and quality control have become more and more evident for all proteins, collagen included. These developments have led to new paradigms in the collagenopathies,[1-3] including strategies for treating disease,[4] the discovery of new ER stress transducers involved in connective tissue development,[5-7] the discovery of new degradation pathways for collagens and non-collagenous proteins,[8-13] and more.

As described in earlier chapters, the collagen *N*-glycan is emerging as a key nexus in collagen proteostasis. Chapter 2 showed that the collagen-I *N*-glycan is critical for interacting with lectin chaperones to maintain collagen proteostasis when protein folding is challenged.[14] Separately, *N*-glycan processing has been shown to be key for directing misfolded or misassembled collagens to ERAD or macro-ER-phagy via calnexin-FAM134B, respectively.[9, 10] Emerging evidence that many of the key players involved in ERLAD processes also may be involved in ER-to-Golgi transport presents an appealing, but never before tested, hypothesis that the collagen *N*-glycan may be involved in making the key determination of whether a collagen molecule is folded and should be secreted or misfolded and should be degraded. Alternatively, Hsp47 also appears to be involved in many of these same processes, via interaction with ER cargo receptor TANGO1 and ER stress sensor IRE1.[15, 16] Future work examining how proteins are recognized as folded or misfolded and targeted to the correct fates are likely to reveal more about the involvement of these two elements of collagen proteostasis.

Thanks to new structural and biochemical work,[17-19] the mechanism of collagen assembly, particularly chain selection, has received renewed attention in recent years. Our earlier work discovering the importance of Ca^{2+}-mediated C-Pro assembly and addressing the role of the cysteine code has helped transform perspectives on this problem. However, the model derived from these data left key questions unanswered. Efforts towards clarifying these questions using full-length collagens add further intrigue, as discussed in Chapter 3. Our data show that certain collagens can form pepsin-resistant trimers even in the absence of covalent linkage in the C-Pro domain, suggesting that triple helix folding may not be as C-Pro-driven as previously thought. While the C-Pro domain is likely still involved in aligning three polypeptide strands for folding, the other domains appear to have some degree of innate trimerization that influence the assembly process. In addition, the formation of an unexpected disulfide-linked dimer by some α1(I) variants suggests these other domains may also impact C-Pro assembly. However, more work is required to clarify which domains contribute to assembly and how.

Critically, the most fundamental question about collagen type-specific assembly remains unanswered—how is the 2:1 α1:α2(I) heterotrimer favored over the α1(I) homotrimer? While the α1(I) homotrimer is more thermally stable and simpler on a molecular level,[20] it is rarely found in healthy tissue.[21] One possible mechanism for this bias in assembly is specific salt bridges or non-covalent interactions that form between the C-Pro domains of two different chains.[19] While modeling of the heterotrimeric C-Pro identified some putative interactions that impact assembly, the impact of these interactions on assembly have not been robustly tested. Another possibility is differences in the Ca^{2+}-binding interface. While the residues that directly coordinate Ca^{2+} are conserved between proα1(I) and proα2(I), the local environment may alter Ca^{2+}-binding affinities such that an α1(I):α2(I) interface is preferred. However, until the crystal structure of the heterotrimeric collagen-I C-Pro is solved, such a structural explanation will likely remain elusive. Moreover, the data presented in Chapter 3 suggest that the triple-helical domains might contribute to chain selection, particularly the thermodynamically instable α2(I) triple helix. Finally, another hypothesis to explain preferential heterotrimer formation may be a kinetic effect. Collagen assembly is generally presumed to proceed through a dimeric intermediate. The dimer may be either an α1(I):α1(I) homodimer or an α1(I):α2(I) heterodimer. If the former forms much more slowly than the latter, it may have a profound impact on assembly patterns. Our preliminary data examining this hypothesis are presented in Appendix A.

The experiments required to answer fundamental questions of collagen assembly are challenged by the complexity of manipulating and analyzing collagen and collagen expression systems. Collagen genes are highly repetitive and GC-rich, making them relatively unamenable

to standard molecular biology techniques, including high-profile CRISPR technologies that have transformed our ability to answer many other biological questions. The protein itself is also highly repetitive, difficult or impossible to synthesize recombinantly, fairly unstable at physiological temperatures, and tricky to characterize. Collagen does not have any enzymatic activity, thereby necessitating biophysical or tissue-level assays to characterize folding and function. In addition, heterogeneity in propeptide cleavage results in multiple species of collagen produced in experimental systems, including procollagen, pN-collagen (C-Pro cleaved), pC-collagen (N-Pro cleaved), and collagen without propeptides. These varied species can make characterization of a single species or total collagen difficult, requiring thoughtful insertion of epitope tags, antibodies that recognize different domains, and careful interpretation of data collected from a mixed collagen population.

The complexity of the collagen biosynthetic pathway necessitates more biologically relevant systems, such as primary cells or tissues, which can be challenging to work with and lead to large variations in results. Our lab and others have circumvented many of these issues in our C-Pro studies by expressing the C-Pro domain alone,[14, 17, 19, 22] while others have used mini collagen systems with a truncated triple-helical domain.[23] Work on the triple-helical domain often involves *in vitro* or acellular systems,[24] short model peptides prepared by chemical synthesis,[25] or other dramatic simplifications. While these systems are representative for some studies (Chapter 2), they often produce conflicting data and are not always consistent with data collected from more representative systems (Chapter 3).

Finally, collagen behavior can vary greatly between organisms, tissues, and types. Therefore, careful consideration is required when choosing experimental systems and interpreting data from collagen studies. Advances in gene editing, cell line engineering, stem cell technology, and organoid models[26-31] provide hope that fundamental studies of collagen biosynthesis will become more straightforward in the future.

REFERENCES

1. Forlino, A., Cabral, W. A., Barnes, A. M., Marini, J. C., New perspectives on osteogenesis imperfecta. *Nat. Rev. Endocrin.* 7, 540-557 (2011).
2. Forlino, A., Marini, J. C., Osteogenesis imperfecta. *Lancet* 387, 1657-1671 (2016).
3. Marini, J. C., Dang Do, A. N., "Osteogenesis Imperfecta" in Endotext, Feingold, K. R., Anawalt, B., Boyce, A., Chrousos, G., de Herder, W. W., Dhatariya, K., Dungan, K., Hershman, J. M., Hofland, J., Kalra, S., Kaltsas, G., Koch, C., Kopp, P., Korbonits, M., Kovacs, C. S., Kuohung, W., Laferrere, B., Levy, M., McGee, E. A., McLachlan, R., Morley, J. E., New, M., Purnell, J., Sahay, R., Singer, F., Sperling, M. A., Stratakis, C. A., Trence, D. L., Wilson, D. P., Eds. (South Dartmouth (MA), 2000).
4. Wong, M. Y., DiChiara, A. S., Suen, P. H., Chen, K., Doan, N. D., Shoulders, M. D., Adapting secretory proteostasis and function through the unfolded protein response. *Curr. Top. Microbiol. Immunol.* 414, 1-25 (2018).
5. Murakami, T., Saito, A., Hino, S., Kondo, S., Kanemoto, S., Chihara, K., Sekiya, H., Tsumagari, K., Ochiai, K., Yoshinaga, K., Saitoh, M., Nishimura, R., Yoneda, T., Kou, I., Furuichi, T., Ikegawa, S., Ikawa, M., Okabe, M., Wanaka, A., Imaizumi, K., Signalling mediated by the endoplasmic reticulum stress transducer OASIS is involved in bone formation. *Nat. Cell Biol.* 11, 1205-1211 (2009).
6. Symoens, S., Malfait, F., D'Hondt, S., Callewaert, B., Dheedene, A., Steyaert, W., Bachinger, H. P., De Paepe, A., Kayserili, H., Coucke, P. J., Deficiency for the ER-stress transducer OASIS causes severe recessive osteogenesis imperfecta in humans. *Orphanet J. Rare Dis.* 8, 154 (2013).
7. Ishikawa, T., Toyama, T., Nakamura, Y., Tamada, K., Shimizu, H., Ninagawa, S., Okada, T., Kamei, Y., Ishikawa-Fujiwara, T., Todo, T., Aoyama, E., Takigawa, M., Harada, A., Mori, K., UPR transducer BBF2H7 allows export of type II collagen in a cargo- and developmental stage-specific manner. *J. Cell Biol.* 216, 1761-1774 (2017).
8. Cinque, L., De Leonibus, C., Iavazzo, M., Krahmer, N., Intartaglia, D., Salierno, F. G., De Cegli, R., Di Malta, C., Svelto, M., Lanzara, C., Maddaluno, M., Wanderlingh, L. G., Huebner, A. K., Cesana, M., Bonn, F., Polishchuk, E., Hubner, C. A., Conte, I., Dikic, I., Mann, M., Ballabio, A., Sacco, F., Grumati, P., Settembre, C., MiT/TFE factors control ER-phagy via transcriptional regulation of FAM134B. *EMBO J.* 39, e105696 (2020).
9. Forrester, A., De Leonibus, C., Grumati, P., Fasana, E., Piemontese, M., Staiano, L., Fregno, I., Raimondi, A., Marazza, A., Bruno, G., Iavazzo, M., Intartaglia, D., Seczynska, M., van Anken, E., Conte, I., De Matteis, M. A., Dikic, I., Molinari, M., Settembre, C., A selective ER-phagy exerts procollagen quality control via a calnexin-FAM134B complex. *EMBO J.* 38, e99847 (2019).
10. Fregno, I., Fasana, E., Solda, T., Galli, C., Molinari, M., *N*-Glycan processing selects ERAD-resistant misfolded proteins for ER-to-lysosome-associated degradation. *EMBO J.* 40, e107240 (2021).
11. Reggio, A., Buonomo, V., Berkane, R., Bhaskara, R. M., Tellechea, M., Peluso, I., Polishchuk, E., Di Lorenzo, G., Cirillo, C., Esposito, M., Hussain, A., Huebner, A. K., Hubner, C. A., Settembre, C., Hummer, G., Grumati, P., Stolz, A., Role of FAM134 paralogues in endoplasmic reticulum remodeling, ER-phagy, and collagen quality control. *EMBO Rep.* 22, e52289 (2021).
12. Omari, S., Makareeva, E., Roberts-Pilgrim, A., Mirigian, L., Jarnik, M., Leikin, S., Lippincott-Schwartz, J., Noncanonical autophagy at ER exit sites in osteoblasts. *J. Bone Miner. Res.* 34, 230-230 (2019).
13. Omari, S., Makareeva, E., Roberts-Pilgrim, A., Mirigian, L., Jarnik, M., Ott, C., Lippincott-Schwartz, J., Leikin, S., Noncanonical autophagy at ER exit sites regulates procollagen turnover. *Proc. Natl. Acad. Sci. U.S.A.* 115, E10099-E10108 (2018).

14. Li, R. C., Wong, M. Y., DiChiara, A. S., Hosseini, A. S., Shoulders, M. D., Collagen's enigmatic, highly conserved *N*-glycan has an essential proteostatic function. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2026608118 (2021).
15. Ito, S., Nagata, K., Quality control of procollagen in cells. *Annu. Rev. Biochem.* 90, 631-658 (2021).
16. Ito, S., Nagata, K., Roles of the endoplasmic reticulum-resident, collagen-specific molecular chaperone Hsp47 in vertebrate cells and human disease. *J. Biol. Chem.* 294, 2133-2141 (2019).
17. DiChiara, A. S., Li, R. C., Suen, P. H., Hosseini, A. S., Taylor, R. J., Weickhardt, A. F., Malhotra, D., McCaslin, D. R., Shoulders, M. D., A cysteine-based molecular code informs collagen C-propeptide assembly. *Nat. Commun.* 9, 4206 (2018).
18. Bourhis, J. M., Mariano, N., Zhao, Y., Harlos, K., Exposito, J. Y., Jones, E. Y., Moali, C., Aghajari, N., Hulmes, D. J., Structural basis of fibrillar collagen trimerization and related genetic disorders. *Nat. Struct. Mol. Biol.* 19, 1031-1036 (2012).
19. Sharma, U., Carrique, L., Vadon-Le Goff, S., Mariano, N., Georges, R. N., Delolme, F., Koivunen, P., Myllyharju, J., Moali, C., Aghajari, N., Hulmes, D. J., Structural basis of homo- and heterotrimerization of collagen I. *Nat. Commun.* 8, 14671-14681 (2017).
20. Kuznetsova, N. V., McBride, D. J., Leikin, S., Changes in thermal stability and microunfolding pattern of collagen helix resulting from the loss of alpha2(I) chain in osteogenesis imperfecta murine. *J. Mol. Biol.* 331, 191-200 (2003).
21. Uitto, J., Collagen polymorphism: Isolation and partial characterization of α1(I)-trimer molecules in normal human skin. *Arch. Biochem. Biophys.* 192, 371-379 (1979).
22. Doan, N. D., Hosseini, A. S., Bikovtseva, A. A., Huang, M. S., DiChiara, A. S., Papa, L. J., 3rd, Koller, A., Shoulders, M. D., Elucidation of proteostasis defects caused by osteogenesis imperfecta mutations in the collagen-α2(I) C-propeptide domain. *J. Biol. Chem.* 295, 9959-9973 (2020).
23. Bulleid, N. J., Wilson, R., Lees, J. F., Type-III procollagen assembly in semi-intact cells: chain association, nucleation and triple-helix folding do not require formation of inter-chain disulphide bonds but triple-helix nucleation does require hydroxylation. *Biochem. J.* 317 (Pt 1), 195-202 (1996).
24. Bachinger, H. P., Bruckner, P., Timpl, R., Prockop, D. J., Engel, J., Folding mechanism of the triple helix in type-III collagen and type-III pN-collagen. Role of disulfide bridges and peptide bond isomerization. *Eur. J. Biochem.* 106, 619-632 (1980).
25. Jalan, A. A., Sammon, D., Hartgerink, J. D., Brear, P., Stott, K., Hamaia, S. W., Hunter, E. J., Walker, D. R., Leitinger, B., Farndale, R. W., Chain alignment of collagen I deciphered using computationally designed heterotrimers. *Nat. Chem. Biol.* 16, 423–429 (2020).
26. Craft, A. M., Rockel, J. S., Nartiss, Y., Kandel, R. A., Alman, B. A., Keller, G. M., Generation of articular chondrocytes from human pluripotent stem cells. *Nat. Biotechnol.* 33, 638-645 (2015).
27. Okada, M., Ikegawa, S., Morioka, M., Yamashita, A., Saito, A., Sawai, H., Murotsuki, J., Ohashi, H., Okamoto, T., Nishimura, G., Imaizumi, K., Tsumaki, N., Modeling type II collagenopathy skeletal dysplasia by directed conversion and induced pluripotent stem cells. *Hum. Mol. Genet.* 24, 299-313 (2015).
28. Kung, L. H. W., Sampurno, L., Yammine, K. M., Graham, A., McDonald, P., Bateman, J. F., Shoulders, M. D., Lamande, S. R., CRISPR/Cas9 editing to generate a heterozygous COL2A1 p.G1170S human chondrodysplasia iPSC line, MCRIi019-A-2, in a control iPSC line, MCRIi019-A. *Stem Cell Res* 48, 101962 (2020).
29. Lilianty, J., Bateman, J. F., Lamande, S. R., Generation of a heterozygous COL2A1 (p.G1113C) hypochondrogenesis mutation iPSC line, MCRIi019-A-7, using CRISPR/Cas9 gene editing. *Stem Cell Res* 56, 102515 (2021).

30. Lilianty, J., Nur Patria, Y., Stanley, E. G., Elefanty, A. G., Bateman, J. F., Lamande, S. R., Generation of a heterozygous COL2A1 (p.R989C) spondyloepiphyseal dysplasia congenita mutation iPSC line, MCRIi001-B, using CRISPR/Cas9 gene editing. *Stem Cell Res* 45, 101843 (2020).
31. Takeyari, S., Kubota, T., Ohata, Y., Fujiwara, M., Kitaoka, T., Taga, Y., Mizuno, K., Ozono, K., 4-Phenylbutyric acid enhances the mineralization of osteogenesis imperfecta iPSC-derived osteoblasts. *J. Biol. Chem.* 296, 100027 (2021).

APPENDIX A:

MONITORING ASSEMBLY OF THE COLLAGEN C-PRO DOMAIN

AUTHOR CONTRIBUTIONS

Rasia C. Li, Andrew S. DiChiara, and Matthew D. Shoulders designed research; R.C.L. performed research; R.C.L. and A.S.D. contributed reagents.

INTRODUCTION

One of the most compelling open questions concerning procollagen-I assembly is how cells preferentially form 2:1 α1(I):α2(I) heterotrimers rather than homotrimers. Proα1(I) homotrimers are likely more stable within the C-Pro domain, due to the third disulfide bond rather than the free cysteine in the heterotrimer,[1] and within the triple-helical domain.[2] Moreover, the C-Proα1(I) homotrimer crystal structure did not reveal any obviously unfavorable interactions in the homotrimer or suggest any favorable interactions for the heterotrimer.[3] The preference for the heterotrimer appears to be maintained in C-Pro expression systems,[1, 4] despite the data in Chapter 3 suggesting that triple-helical domains may be more involved in chain selection than previously thought.

One compelling hypothesis for preferential heterotrimer formation is a kinetic mechanism. Specifically, assembly of the procollagen is thought to occur through a dimer intermediate. Preferential heterotrimer formation could be explained by more rapid formation of an α1(I):α2(I) dimer compared to an α1(I):α1(I) dimer. Even in the absence of a dimer intermediate, heterotrimeric formation may be kinetically favorable.

In order to test this hypothesis, we first attempted to follow C-Pro assembly via *in vitro* refolding. Unfortunately, we found that the collagen-I C-Pro domain could not efficaciously refold *in vitro*. We next proposed examining C-Pro assembly *in cellulo* using a combination of pulse-chase and diagonal SDS-PAGE. This work is in progress and shows promise in its ability to help unveil the stepwise assembly of the C-Pro trimer.

RESULTS AND DISCUSSION

***In vitro* refolding.** *In vitro* refolding can be a powerful methodology to study protein folding by simplifying the environment in which folding takes place and allowing for direct monitoring of the target protein.[5, 6] It has been used to determine the folding pathway for many proteins, including key experiments that demonstrate the mechanism of collagen triple helix folding.[7] However, *in vitro* refolding can be challenging for large, multimeric proteins such as the C-Pro domain, due to the complexity of the folding process and the many empirical conditions that must be optimized for each protein.

The cysteines in the C-Pro domain allow for monitoring of the assembly process. By quenching thiols at various points during refolding, disulfide-linked intermediates can be trapped and characterized.[8] While this method does not allow for monitoring the actual folding process, it may be sufficient to monitor the kinetics of subunit assembly.

We began optimizing *in vitro* refolding conditions using homotrimeric C-Proα1(I) purified from mammalian cells, as it provides a simpler system than the heterotrimer. Following a previously published protocol for refolding of the C-Pro,[9] we denatured and reduced the purified protein, then dialyzed it against a redox-active buffer while maintaining denaturant to initiate disulfide shuffling. We then removed the denaturant by dialysis against redox-active buffer and took samples at various time points. These samples were immediately quenched with iodoacetamide, and assembly was monitored by non-reducing SDS-PAGE (Fig. A-1A).

We observed successful trimer reformation through a dimer intermediate between 4 to 6 h after removal of denaturant (Fig. A-1B). However, the overall yield of reassembled trimer was relatively low, and the vast majority of the protein remained in a monomeric state, before assembling into higher order oligomers and aggregates by 10 h. While many of the oligomers were disulfide-linked and collapsed into monomers upon reduction, a significant portion of the protein remained trapped in insoluble aggregates, indicating that both off-target disulfide formation and misfolding significantly affect trimer yield.

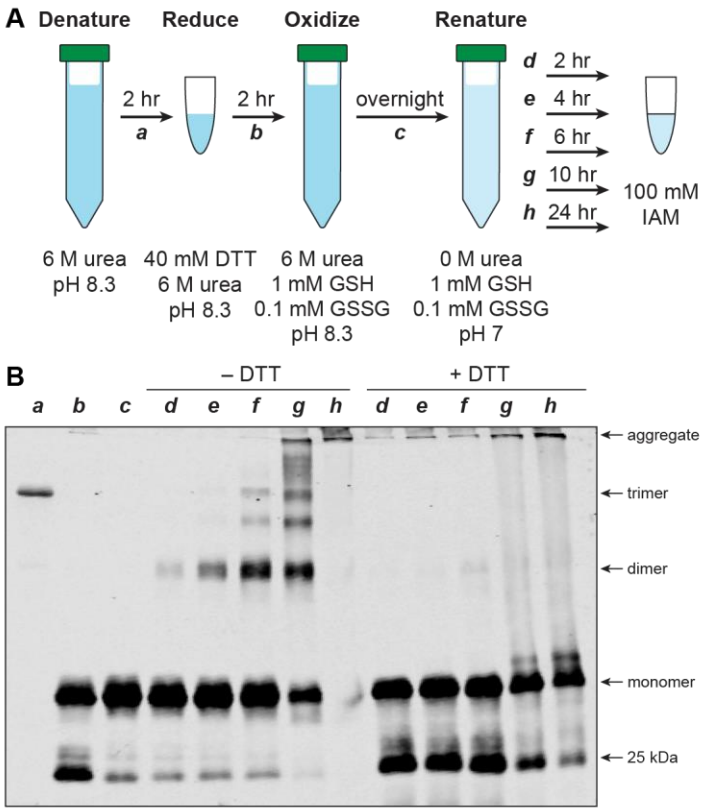


Figure A-1: *In vitro* refolding of the C-Pro domain shows disulfide-dependent reassembly. Refolding began by denaturing purified Colα1(I) C-Pro homotrimer by dialysis into 6 M urea (*a*), then reducing with 40 mM DTT (*b*). Denaturant was removed by dialysis against redox active buffer (*c*). Refolding was initiated by dialysis into redox buffer with no denaturant (d–h). Samples were removed at each of the points specified and quenched with 100 mM iodoacetamide before being run on an SDS-PAGE gel.

Confusingly, the reducing samples on SDS-PAGE revealed an unknown band that runs at approximately 25 kDa (Fig. A-1B). We confirmed that this band is derived from C-Proα1(I) by both Western blotting and mass spectrometry but could not identify its origins. An intriguing result of this species being present in the refolding is that it appears to be able to incorporate into a disulfide-linked species, as seen by the band that runs slightly lower than the trimer.

We tested a number of conditions to improve folding yield and reduce oligomer formation, guided by existing literature on *in vitro* refolding. Conditions tested included altering pH, redox buffer composition, denaturant, protein concentration, among others.[10] We found that while basic, slightly reducing environments were best suited for disulfide-linked trimer formation, they also enhanced disulfide-linked oligomer formation.

While many other conditions could still be tested, the difficulty we encountered improving yield suggests that the C-Pro domain cannot fold unassisted *in vitro* and in fact requires the ER

proteostatic environment for *in vivo* folding. Addition of purified chaperones, such as BiP or PDI, has been shown to increase *in vitro* refolding yield for many proteins, including antibody fragments that share many features with the C-Pro (e.g., *N*-glycosylation, disulfide bonds, and oligomeric assembly).[11-14] In fact, PDI has been shown to accelerate disulfide bond formation for reduced collagen strands *in vitro*.[15, 16]

One key challenge to testing additional *in vitro* refolding conditions is that the use of SDS-PAGE to assess refolding is low-throughput and does not report on correct structure formation, merely the presence or absence of a disulfide-linked species. Due to the C-Pro domain's lack of substrate binding or enzymatic activity, traditional methods of monitoring refolding by assessing activity or affinity to a substrate cannot be used. Biophysical methods, such as SEC or CD, would provide structural information, but data obtained from these methods can be difficult to deconvolute.

***In vivo* assembly monitoring.** It is not clear whether the *in vitro* refolding patterns we observed are to any extent reflective of the correct C-Pro assembly pattern. Thus, we also sought to examine the C-Pro assembly mechanism and the kinetics of trimer formation *in vivo*. We previously generated C-Pro expression constructs that allow for clear separation of C-Proα1(I) and C-Proα2(I) using SDS-PAGE by addition of an XTEN linker to C-Proα1(I).[4] We co-expressed the two constructs in HEK293T cells and examined the intracellular fraction by non-reducing SDS-PAGE. We then excised the relevant band and reduced the protein in-gel, before running the sample on a second SDS-PAGE gel. This method, known as diagonal SDS-PAGE, allows for the determination of the composition of disulfide-linked species.[17, 18] In our preliminary work, we have observed that this assay works well for the C-Pro domain. We are able to separate the heterotrimer into C-Proα1(I) and C-Proα2(I) (Fig. A-2). Unfortunately, due to apparent loss of signal after the second SDS-PAGE, we have thus far been unable to separate the dimer in the reducing dimension.

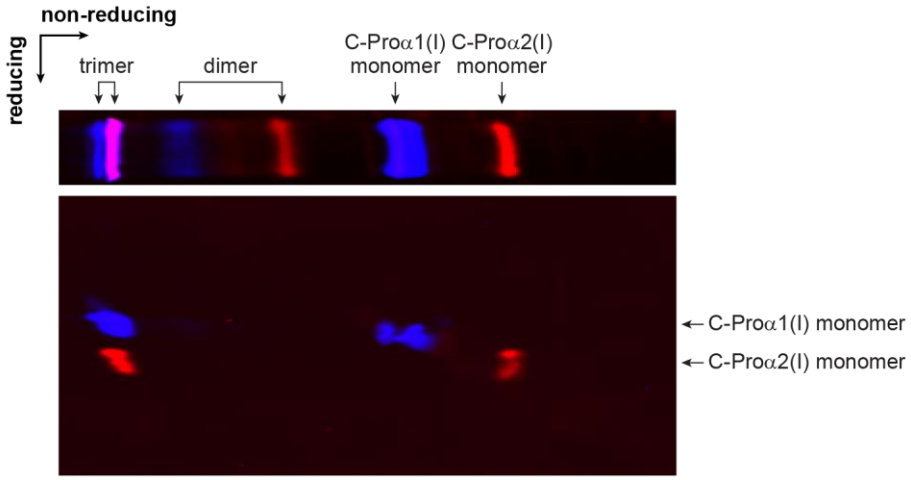


Figure A-2: Diagonal SDS-PAGE allows for separation of the components of hetero-oligomers.

Non-reducing lysate from cells expressing XTEN-C-Proα1(I) and C-Proα2(I) was run on an SDS-PAGE gel. The band shown above was excised, reduced, then rotated 90°, and run on a second gel. The two gels in combination show that we are clearly able to separate C-Proα1(I) and C-Proα2(I) from heterooligomers.

The steady-state gel in Figure A-2 suggests that we produce both homo- and heterotrimer, potentially a result of expressing more C-Proα1(I) than C-Proα2(I). Based on prior work,[1] optimizing the ratio of the two constructs will allow us to produce mostly heterotrimer. We also observe two distinct dimer species, apparently both homodimeric C-Proα1(I) and C-Proα2(I). Homodimeric C-Proα2(I) in particular is clearly an off-target product, indicating that assembly can involve disulfide shuffling. In the future, we plan to apply a pulse-chase protocol to monitor assembly with higher resolution and in a more quantitative manner to determine a kinetic mechanism for C-Pro assembly.

METHODS AND MATERIALS

Cell Culture. HEK293T cells (ATCC) were cultured in complete DMEM media (Corning) supplemented with 10% fetal bovine serum (FBS), 100 IU penicillin, 100 μg/mL streptomycin, and 2 mM L-glutamine (Corning). Freestyle 293F cells (Gibco) were cultured in 293 Freestyle expression media (Gibco).

Vector Construction. C-Pro expression vectors for both C-Pro purification and transient expression have been previously described. The vectors used include PPT.HA.His.HRV-3C cleavage site.C-Pro1A1.pcDNA3.1,[1, 4] PPT.FLAG.XTEN.C-Pro1A1.pcDNA3.1, and PPT.HA.C-Pro1A2.pcDNA3.1.[4]

C-Pro expression and purification. Expression and purification of the C-Pro domain from Freestyle 293F cells has been previously described.[1] In brief, 6×His-tagged C-Proα1(I) was transfected using 293Fectin according to manufacturer's instructions. Media was harvested 6 days after transfection and purified via Ni-NTA chromatography. The eluted protein was cleaved with His-tagged HRV-3C protease (Pierce) to remove the tags, then subject to a second Ni-NTA column. The cleaved protein was dialyzed into 20 mM bis-Tris propane (BTP), 150 mM NaCl, and 2 mM $CaCl_2$ and further purified by size exclusion chromatography on a Bio-Rad SEC 650 column.

***In vitro* refolding.** Purified C-Proα1(I) was denatured by dialysis against 6 M urea, 20 mM BTP, pH 8.3, 150 mM NaCl, and 2 mM $CaCl_2$ for 2 h at room temperature. The protein was then reduced with 40 mM dithiothreitol for 2 h at 37 °C. The reduced protein was dialyzed overnight against the above buffer with addition of 1 mM reduced glutathione (GSH) and 0.1 mM oxidized glutathione (GSSG). The following day, the dialysis buffer was changed to 20 mM BTP, pH 7.0, 150 mM NaCl, and 2 mM $CaCl_2$ with no urea and fresh GSH and GSSG at the same concentrations. Dialysis proceeded at room temperature, and samples were removed at the indicated time points. The samples were immediately quenched with addition of iodoacetamide (IAM) to a final concentration of 100 mM.

Diagonal SDS-PAGE experiments. HEK293T cells were transfected with a 2:1 ratio of PPT.FLAG.XTEN.C-Pro1A1.pcDNA3.1 and PPT.HA.C-Pro1A2.pcDNA3.1 using Lipofectamine 3000, according to the manufacturer's instructions. Media was changed 1 day after transfection, and cells were harvested 2 days after transfection. Cells were collected by trypsinization, washed with PBS, and then lysed in Triton-X lysis buffer (50 mM Tris, pH 7.5, 200 mM NaCl, 1% Triton-X, 1 mM EDTA, 1 mM EGTA, 1.5 mM phenylmethylsulfonyl fluoride (PMSF; Sigma), and protease inhibitor (Pierce)) for 20 min on ice. Lysates were clarified by centrifugation, then subjected to 100 mM iodoacetamide treatment for 30 min in the dark.

After running the non-reduced lysates on SDS-PAGE, the relevant lane was excised. The gel slide was then incubated in 125 mM Tris, pH 6.8, 0.1% SDS, and 30 mM DTT for ~2 h at room temperature. The gel slide was washed with water, then laid on top of a second 12% SDS-PAGE gel and sealed in place with 4% polyacrylamide.

Immunoblotting. Samples were boiled in 6× gel loading buffer (300 mM Tris, pH 6.8, 15% glycerol, 6% SDS, and 10% (w/v) bromophenol blue) for 10 min. Samples were then run on homemade 4/12% polyacrylamide gels and transferred to nitrocellulose for antibody probing. Western blots were probed with the following primary antibodies: anti-C-Proa1(I) (Sigma HPA008405, 1:1000), anti-HA (ThermoFisher 26183, 1:10000), or anti-FLAG (Agilent 200474, 1:1000), followed by 800CW or 680LT secondary antibodies (LI-COR). Images were obtained using an Odyssey infrared imager (LI-COR).

REFERENCES

1. DiChiara, A. S., Li, R. C., Suen, P. H., Hosseini, A. S., Taylor, R. J., Weickhardt, A. F., Malhotra, D., McCaslin, D. R., Shoulders, M. D., A cysteine-based molecular code informs collagen C-propeptide assembly. *Nat. Commun.* 9, 4206 (2018).
2. Kuznetsova, N. V., McBride, D. J., Leikin, S., Changes in thermal stability and microunfolding pattern of collagen helix resulting from the loss of alpha2(I) chain in osteogenesis imperfecta murine. *J. Mol. Biol.* 331, 191-200 (2003).
3. Sharma, U., Carrique, L., Vadon-Le Goff, S., Mariano, N., Georges, R. N., Delolme, F., Koivunen, P., Myllyharju, J., Moali, C., Aghajari, N., Hulmes, D. J., Structural basis of homo- and heterotrimerization of collagen I. *Nat. Commun.* 8, 14671-14681 (2017).
4. Doan, N. D., Hosseini, A. S., Bikovtseva, A. A., Huang, M. S., DiChiara, A. S., Papa, L. J., 3rd, Koller, A., Shoulders, M. D., Elucidation of proteostasis defects caused by osteogenesis imperfecta mutations in the collagen-α2(I) C-propeptide domain. *J. Biol. Chem.* 295, 9959-9973 (2020).
5. Rudolph, R., Lange, C., "Chapter 2.2 Strategies for the Oxidative in vitro Refolding of Disulfide-bridge-containing Proteins" in Oxidative Folding of Peptides and Proteins. (The Royal Society of Chemistry, 2009), 10.1039/9781847559265-00192, pp. 192-219.
6. Jaenicke, R., Lilie, H., Folding and association of oligomeric and multimeric proteins. *Adv. Protein Chem.* 53, 329-401 (2000).
7. Bachinger, H. P., Bruckner, P., Timpl, R., Prockop, D. J., Engel, J., Folding mechanism of the triple helix in type-III collagen and type-III pN-collagen. Role of disulfide bridges and peptide bond isomerization. *Eur. J. Biochem.* 106, 619-632 (1980).
8. Mamathambika, B. S., Bardwell, J. C., Disulfide-linked protein folding pathways. *Annu. Rev. Cell. Dev. Biol.* 24, 211-235 (2008).
9. Doege, K. J., Fessler, J. H., Folding of carboxyl domain and assembly of procollagen I. *J. Biol. Chem.* 261, 8924-8935 (1986).
10. Mayer, M., Buchner, J., Refolding of inclusion body proteins. *Methods Mol. Med.* 94, 239-254 (2004).
11. Mayer, M., Kies, U., Kammermeier, R., Buchner, J., BiP and PDI cooperate in the oxidative folding of antibodies in vitro. *J. Biol. Chem.* 275, 29421-29425 (2000).
12. Yao, Y., Zhou, Y., Wang, C., Both the isomerase and chaperone activities of protein disulfide isomerase are required for the reactivation of reduced and denatured acidic phospholipase A2. *EMBO J.* 16, 651-658 (1997).
13. Buchner, J., Brinkmann, U., Pastan, I., Renaturation of a single-chain immunotoxin facilitated by chaperones and protein disulfide isomerase. *Biotechnology. (N. Y.)* 10, 682-685 (1992).
14. Lilie, H., McLaughlin, S., Freedman, R., Buchner, J., Influence of protein disulfide isomerase (PDI) on antibody folding in vitro. *J. Biol. Chem.* 269, 14290-14296 (1994).
15. Koivu, J., Myllyla, R., Interchain disulfide bond formation in types I and II procollagen. Evidence for a protein disulfide isomerase catalyzing bond formation. *J. Biol. Chem.* 262, 6159-6164 (1987).
16. Forster, S. J., Freedman, R. B., Catalysis by protein disulphide-isomerase of the assembly of trimeric procollagen from procollagen polypeptide chains. *Biosci. Rep.* 4, 223-229 (1984).
17. Yano, H., Wong, J. H., Lee, Y. M., Cho, M. J., Buchanan, B. B., A strategy for the identification of proteins targeted by thioredoxin. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4794-4799 (2001).
18. Samelson, L. E., Diagonal gel electrophoresis. *Curr. Protoc. Immunol.* Chapter 8, Unit 8 6 (2001).

APPENDIX B:

MAPPING DISULFIDE BONDS IN THE COLLAGEN C-PRO DOMAIN

Portions of this chapter have been adapted from:

DiChiara, A. S., Li, R. C., Suen, P. H., Hosseini, A. S., Taylor, R. J., Weickhardt, A. F., Malhotra, D., McCaslin, D. R., Shoulders, M. D., A cysteine-based molecular code informs collagen C-propeptide assembly. *Nat. Commun.* 9, 4206 (2018).

AUTHOR CONTRIBUTIONS

Rasia C. Li and Matthew D. Shoulders designed research; R.C.L. performed research and analyzed data.

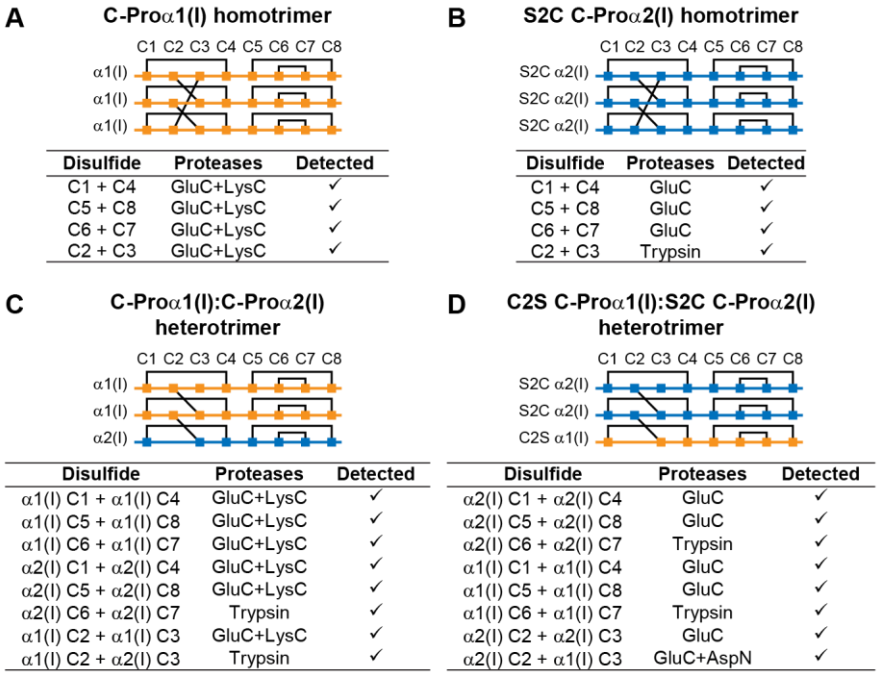
INTRODUCTION

The collagen-I C-Pro domain is disulfide-rich, with eight cysteines on each of the two α1(I) chains and seven cysteines on the single α2(I) chain. The second cysteine in α2(I) is replaced with a serine.[1] Early analyses of the disulfide bonding pattern in the collagen-I C-Pro were performed using cyanogen bromide (CNBr), which cleaves C-terminal to methionine residues. This work concluded that intrachain disulfide bonds were formed between C5–C8 and C6–C7, while interchain disulfide bonds were formed between the first four cysteine residues.[2, 3] However, experimental data later suggested that C1 and C4 in fact form an intrachain disulfide bond that is critical for folding and secretion.[4] While the structure of the native C-Pro heterotrimer is unknown, the crystal structure of homotrimeric C-Proα1(I) confirms that C1 and C4 in C-Proα1(I) form an intrachain disulfide bond, leaving C2 and C3 to form interchain disulfide bonds covalently linking all three chains of the C-Pro domain (Fig. B-1A).[5] Although the disulfide network in the heterotrimeric C-Pro domain is not definitively established, it seems likely to follow the same overall structure, with C2–C3 interchain disulfide bonds connecting all three strands and a free C3 in one of the C-Proα1(I) strands (Fig. B-1C).

The ability to detect disulfide bond linkages is a potentially useful methodology for examining the folding and assembly pathway of the C-Pro domain by determining the order of disulfide formation and for determining how the disulfide bond network is affected in disease-causing C-Pro variants, particularly those that involve substitution of one of the cysteines. Advances in mass spectrometry methods have made it possible to detect and characterize disulfide-linked peptides.[6] Here, we describe early-stage efforts to map the disulfide bond network in wild-type C-Pro assemblies and in various disease-causing mutants.

RESULTS AND DISCUSSION

Confirmation of the disulfide bond map of C-Proα1(I). The crystal structure of homotrimeric C-Proα1(I) shows that C1–C4, C5–C8, and C6–C7 form intrachain disulfide bonds, while C2–C3 form interchain disulfide bonds.[5] Therefore, the homotrimer presented an ideal model for optimization and proof of concept. To prepare samples for mass spectrometry, we first treated the purified protein with iodoacetamide to block any free cysteines. We found that it was necessary to remove the *N*-glycan using PNGase to allow for detection of the peptide containing C6, as the relevant amino acids are only five residues apart. For the C-Proα1(I) homotrimer, digestion with a combination of endoproteinase GluC, which cleaves C-terminal to glutamate residues, and LysC, which cleaves C-terminal to lysine residues, was sufficient to generate peptides that contained only one cysteine. Samples were run on LC-MS, and the data were manually searched for peptide masses corresponding to the theoretical mass of the expected disulfide-linked peptides. We were able to detect all four disulfide linkages in the C-Proα1(I) homotrimer using this method (Fig. B-1A).



Disulfide	Proteases	Detected
C1 + C4	GluC+LysC	✓
C5 + C8	GluC+LysC	✓
C6 + C7	GluC+LysC	✓
C2 + C3	GluC+LysC	✓

Disulfide	Proteases	Detected
C1 + C4	GluC	✓
C5 + C8	GluC	✓
C6 + C7	GluC	✓
C2 + C3	Trypsin	✓

Disulfide	Proteases	Detected
α1(I) C1 + α1(I) C4	GluC+LysC	✓
α1(I) C5 + α1(I) C8	GluC+LysC	✓
α1(I) C6 + α1(I) C7	GluC+LysC	✓
α2(I) C1 + α2(I) C4	GluC+LysC	✓
α2(I) C5 + α2(I) C8	GluC+LysC	✓
α2(I) C6 + α2(I) C7	Trypsin	✓
α1(I) C2 + α1(I) C3	GluC+LysC	✓
α1(I) C2 + α2(I) C3	Trypsin	✓

Disulfide	Proteases	Detected
α2(I) C1 + α2(I) C4	GluC	✓
α2(I) C5 + α2(I) C8	GluC	✓
α2(I) C6 + α2(I) C7	Trypsin	✓
α1(I) C1 + α1(I) C4	GluC	✓
α1(I) C5 + α1(I) C8	GluC	✓
α1(I) C6 + α1(I) C7	Trypsin	✓
α2(I) C2 + α2(I) C3	GluC	✓
α2(I) C2 + α1(I) C3	GluC+AspN	✓

Figure B-1: Hypothesized disulfide linkages detected in all homo- and heterotrimeric species.

(A) The C-Proα1(I) disulfide map was confirmed by detection of all four known disulfide linkages. (B) The S2C C-Proα2(I) disulfide map was shown to resemble the C-Proα1(I) map. (C) The wild-type heterotrimer disulfide map contains the same linkages as the homotrimers with exception of the free C3 in one C-Proα1(I) strand. (D) The mutant heterotrimer formed by C2S C-Proα1(I) and S2C C-Proα2(I) exhibits the same linkages as wild-type heterotrimer.

Mapping the S2C C-Proα2(I) homotrimer. Wild-type C-Proα2(I) lacks a single cysteine at the second position when compared to C-Proα1(I). We previously demonstrated that substitution of a cysteine for the serine at that position allowed S2C C-Proα2(I) to form disulfide-linked homotrimers, which is not possible for wild-type C-Proα2(I).[1] This homotrimer would presumably form the same disulfide linkages as C-Proα1(I). We sought to confirm this hypothesis using the method developed for C-Proα1(I). However, the different primary sequence of S2C C-Proα2(I), specifically the absence of either a glutamate or lysine between C1 and C2, necessitated the use of a different set of proteases to produce detectable peptides. GluC digestion in phosphate buffer was used to cleave at both glutamate and aspartate residues, the latter of which is present between C1 and C2. Under these conditions, all linkages other than C6–C7 were detected. The C6–C7 peptide was likely not detected owing to its relatively large size. However, an independent trypsin digest permitted detection of the C6–C7 linkage (Fig. B-1B). Detection of the C2–C3 linkage in S2C C-Proα1(I) supports our model that both C2 and C3 are required for the formation of stable C-Pro homotrimers by allowing for disulfide linkage of all three strands.[1]

Mapping the heterotrimers. While the disulfide bond network within the 2:1 C-Proα1(I):C-Proα2(I) heterotrimer is not definitively known, it is presumed to resemble that of the C-Proα1(I) homotrimer based on overall sequence homology and experimental data. We applied our methodology to purified heterotrimeric C-Pro, using the optimized protease conditions. Our results confirmed that C1–C4 form intrachain disulfide bonds on both C-Proα1(I) and C-Proα2(I) (Fig. B-1C), resolving the previous discrepancy regarding the linkage of these two cysteines. Detection of the free C3 on C-Proα1(I) also confirmed this linkage pattern. Notably, none of the interchain disulfide bonds proposed by the initial CNBr mapping were detected.[2]

We were also curious about the disulfide-linkage pattern in the mutant heterotrimer observed upon co-expression of S2C C-Proα2(I) and C2S C-Proα1(I). These two variants have swapped cysteine patterns and form a heterotrimer when co-expressed.[1] Thus, we expect that the overall disulfide linkage patterns would be conserved. We indeed found this pattern to be the case, with detection of all proposed linkages and the free cysteine at position 3 of C-Proα1(I) (Fig. B-1D).

Determining disulfide bond formation in disease variants. A variety of mutations within the collagen-I C-Pro domain can lead to osteogenesis imperfecta (OI).[7] Many of these variants exhibit impaired folding and secretion, leading to disease phenotypes. We were curious whether the impaired folding of these variants would lead to disruptions in the disulfide bond

network. Therefore, we expressed four OI-causing C-Proα1(I) variants alongside wild-type C-Proα1(I) and performed disulfide mapping on the homotrimer variants. We were particularly interested in the D1277H variant, as it exhibits defective assembly, forming primarily dimers and oligomers.[1, 8]

Variant	C1 + C4	C5 + C8	C6 + C7	C2 + C3	C2 + C2	C3 + C3	% Correct
WT	✓	✓	✓	✓	×	✓	88.7%
G1272V	✓	✓	✓	✓	×	✓	89.4%
W1275R	✓	✓	✓	✓	×	×	100%
A1286D	✓	✓	✓	✓	×	✓	64.4%
D1277H	✓	✓	✓	✓	×	✓	0.99%

Figure B-2: OI variants within C-Proa1(I) are able to disrupt disulfide bonding.
While correct disulfide bonds were detected in all variants, incorrect interchain disulfide bonds between C3–C3 were detected in three of four variants along with wild-type C-Proa1(I). The percentage of correct C2–C3 bonds relative to all interchain disulfides may suggest the extent of misfolding by each of these variants.

Using the protocol optimized for wild-type C-Proα1(I), we detected all four correct disulfide linkages in all variants (Fig. B-2). We then searched the data for incorrect disulfide linkages. In particular, we looked at potentially incorrect pairings of the interchain disulfide bonds that may result in the disulfide-linked dimers and oligomers observed for D1277H. In three of four variants as well as wild-type, we found C3–C3 linkages. We calculated the abundance of each peptide detected for the correct and incorrect linkages and calculated the percentage of the correct C2–C3 linkage relative to all interchain linkages detected (C2–C3 + C3–C3). Of the five variants, wild-type, G1272V, and W1275R maintained high relative levels of the correct C2–C2 linkage. The D1277H variant exhibited almost exclusively incorrect C3–C3 linkages, supporting its observed misassembly. The A1286D variant surprisingly exhibited the correct linkage just over half of the time, despite forming almost exclusively disulfide-linked trimer (see Chapter 2, Fig. S2-1D).

Limitations of the method. The presence of incorrect C3–C3 linkages in wild-type C-Proα1(I) led us to search all of our data sets for other off-target disulfide linkages. We found various off-target disulfide linkages in all data sets. These incorrect linkages were all detected at approximately an order of magnitude lower abundance than the correct ones. However, owing to differences in ionization potential among the different peptides, we viewed the current experimental setup as not quantitative enough to make definitive claims about which disulfides are the most abundant in a single sample.

One strategy to account for ionization differences between peptides is to use peptide standards to quantify the absolute abundance of a given peptide in a sample.[9, 10] However, owing to the number of unique disulfide-bound peptides that can be formed (99 for the heterotrimer alone), compounded by the number of digestion conditions and disease variants used, synthesizing peptide standards for each possibility would be prohibitively complex. Another relatively feasible approach would be to use HPLC coupled with UV or fluorescence detection to determine relative abundance of individual peptides within a sample.[6]

In addition, our current workflow requires calculating theoretical masses for each disulfide-linked peptide and searching the data manually. More sophisticated mass spectrometry strategies are becoming increasingly available for identifying disulfide-linked peptides and offer many options for generating more high throughput and accurate methods of determining the disulfide bond patterns in the collagen C-Pro domain.[6] These studies should be followed-up in future work.

METHODS AND MATERIALS

Cell Culture. Freestyle 293F cells (Gibco) were cultured in 293 Freestyle expression media (Gibco).

Vector Construction. C-Pro expression vectors for both C-Pro purification have been previously described.[1] The vectors used include: PPT.HA.His.HRV-3C site.C-Pro1A1.pcDNA3.1, PPT.FLAG.His.HRV-3C site.S2C C-Pro1A2.pcDNA3.1, PPT.HA.HRV-3C site.C-Pro1A1.pcDNA3.1, PPT.HA.His.HRV-3C site.C-Pro1A2.pcDNA3.1, PPT.HA.His.HRV-3C site.C2S C-Pro1A1.pcDNA3.1, and PPT.HA.HRV-3C site.S2C C-Pro1A2.pcDNA3.1. Expression vectors for OI variants were generated using the QuikChange XL II site-directed mutagenesis kit (Agilent).

C-Pro expression and purification. Expression and purification of the C-Pro domain from Freestyle 293F cells has been previously described.[1] In brief, the relevant constructs were transfected using 293Fectin according to manufacturer's instructions: For production of homotrimeric C-Proα1(I), PPT.HA.His.HRV-3C site.C-Pro1A1.pcDNA3.1. For production of homotrimeric S2C C-Proα2(I), PPT.FLAG.His.HRV-3C site.S2C C-Pro1A2.pcDNA3.1. For production of wild-type heterotrimer: PPT.HA.HRV-3C site.C-Pro1A1.pcDNA3.1 + PPT.HA.His.HRV-3C site.C-Pro1A2.pcDNA3.1. For production of mutant heterotrimer: PPT.HA.His.HRV-3C site.C2S C-Pro1A1.pcDNA3.1 + PPT.HA.HRV-3C site.S2C C-Pro1A2.pcDNA3.1. Media was harvested 6 days after transfection and purified via Ni-NTA chromatography. The eluted protein was cleaved with HRV-3C protease to remove the tags, then subject to a second Ni-NTA column. The cleaved protein was dialyzed into 20 mM bis-Tris propane (BTP), 150 mM NaCl, and 2 mM $CaCl_2$ and further purified by size exclusion chromatography on a Bio-Rad SEC 650 column.

Disulfide mapping by mass spectrometry. Purified C-Pro was treated with 55 mM iodoacetamide for 1 h in the dark. The *N*-glycan was removed using PNGase F following manufacturer's instructions (NEB), followed by removal of the enzyme by centrifugation over a 50 kDa MWCO filter. The protein was precipitated with trichloroacetic acid, and the resultant pellet was washed with acetone. The pellet was resuspended in 8 M urea, then diluted to 0.38 M urea in the appropriate buffer. The protein was digested with the indicated protease overnight at 37 °C. Proteolyzed samples were quenched by addition of formic acid to a final concentration of 5% and applied to Protea C18 SpinTips. The eluent from the SpinTips was evaporated completely by speed vacuum centrifugation. For double digestions, the peptides were resuspended in the second buffer then subject to protease digestion and desalting using Protea C18 SpinTips again. The final sample was resuspended in 0.1% formic acid. Samples were injected onto an EASY-

nLC 1000 with MS data acquired on a Thermo QExactiv mass spectrometer. MS spectra were manually searched for the predicted masses of each set of disulfide-linked peptides.

The protease conditions used were as follows: Trypsin digestions were performed at a 1:40 substrate:enzyme ratio in 100 mM ammonium bicarbonate, pH 6.4. Endoproteinase GluC (Pierce) was used at a 1:20 substrate:enzyme ratio in either 100 mM ammonium acetate, pH 4.0 (glutamate only) or PBS, pH 6.8 (glutamate and aspartate). rLys-C (Promega) was used at a ratio of 1:27 in 25 mM Tris, pH 7.0, 1 mM EDTA. Asp-N (Promega) was used as a ratio of 1:27 in 50 mM Tris, pH 7.0, 0.5 mM $ZnCl_2$.

REFERENCES

1. DiChiara, A. S., Li, R. C., Suen, P. H., Hosseini, A. S., Taylor, R. J., Weickhardt, A. F., Malhotra, D., McCaslin, D. R., Shoulders, M. D., A cysteine-based molecular code informs collagen C-propeptide assembly. *Nat. Commun.* 9, 4206 (2018).
2. Koivu, J., Identification of disulfide bonds in carboxy-terminal propeptides of human type I procollagen. *FEBS Lett.* 212, 229-232 (1987).
3. Olsen, B. R., "The carboxyl propeptides of procollagen: structure and functional considerations" in New Trends in Basement Membrane Research*,* Kuhn, K. S., H.; Timpl, R., Ed. (Raven Press, New York, 1982), pp. 225-236.
4. Lees, J. F., Bulleid, N. J., The role of cysteine residues in the folding and association of the COOH-terminal propeptide of type-I and type-III procollagen. *J. Biol. Chem.* 269, 24354-24360 (1994).
5. Sharma, U., Carrique, L., Vadon-Le Goff, S., Mariano, N., Georges, R. N., Delolme, F., Koivunen, P., Myllyharju, J., Moali, C., Aghajari, N., Hulmes, D. J., Structural basis of homo- and heterotrimerization of collagen I. *Nat. Commun.* 8, 14671-14681 (2017).
6. Weinfurtner, D., "Chapter 1.4: Analysis of Disulfide Bond Formation in Therapeutic Proteins" in Oxidative Folding of Proteins: Basic Principles, Cellular Regulation and Engineering. (The Royal Society of Chemistry, 2018), 10.1039/9781788013253-00081, pp. 81-98.
7. Symoens, S., Hulmes, D. J. S., Bourhis, J. M., Coucke, P. J., De Paepe, A., Malfait, F., Type I procollagen C-propeptide defects: study of genotype–phenotype correlation and predictive role of crystal structure. *Hum. Mutat.* 35, 1330-1341 (2014).
8. Li, R. C., Wong, M. Y., DiChiara, A. S., Hosseini, A. S., Shoulders, M. D., Collagen's enigmatic, highly conserved *N*-glycan has an essential proteostatic function. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2026608118 (2021).
9. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* 389, 1017-1031 (2007).
10. Bronsema, K. J., Bischoff, R., van de Merbel, N. C., Internal standards in the quantitative determination of protein biopharmaceuticals using liquid chromatography coupled to mass spectrometry. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 893-894, 1-14 (2012).