Genotype-phenotype correlations in pathology caused by collagen type IV alpha 1 and 2 mutations

Marion Jeanne and Douglas B Gould

a - Genentech Inc, 1 DNA Way, South San Francisco, CA 94080, USA
b - Department of Ophthalmology, Department of Anatomy, Institute for Human Genetics, UCSF School of Medicine, San Francisco, CA 94143, USA

Correspondence to Marion Jeanne and Douglas B Gould: jeanne.marion@gene.com; GouldD@vision.ucsf.edu

http://dx.doi.org/10.1016/j.matbio.2016.10.003

Abstract

COL4A1 and COL4A2 are extracellular matrix proteins that form heterotrimers and are present in nearly all basement membranes in every organ. In the past decade, COL4A1 and COL4A2 mutations have been identified to cause a multi-system disorder for which penetrance and severity of constituent phenotypes can greatly vary. Here, we compare the outcomes of more than 100 mutations identified in patients and data from a murine allelic series to explore the presence of genotype–phenotype correlations – many of which are shared among other types of collagen. We find that there is a frequency bias for COL4A1 over COL4A2 mutations and that glycine (Gly) substitutions within the triple helical domain are the most common class of mutations. Glycine is most often replaced by a charged amino acid, however the position of the mutation, and not the properties of the substituting amino acid, appears to have a greater influence on disease severity. Moreover, the impact of position is not straightforward. Observations from a murine allelic series suggest that mutations in the NC1 domain may result in relatively mild phenotypes via a ‘quantitative’ mechanism similar to other types of collagens, however, this effect was not apparent in human reports. Importantly, other position-dependent effects had differential impacts depending on the phenotype of interest. For example, the severity of cerebrovascular disease correlated with an amino-to-carboxy severity gradient for triple-helical glycine substitutions whereas the penetrance and severity of myopathy and nephropathy appear to involve a functional sub-domain(s). Greater understanding of genotype–phenotype correlations and the interaction of consequences of different mutations will be important for patient prognosis and care and for developing mechanism-based therapeutics to treat individual components of this emerging syndrome.

1. Introduction

Type IV collagens are network-forming basement membrane collagens [1] and are encoded by six genes; collagen, type IV, alpha 1 (COL4A1) through 6 (COL4A6) in mammals [2,3]. COL4A1 and COL4A2 are present in almost all basement membranes, while the distributions of COL4A3, COL4A4, COL4A5 and COL4A6 are restricted to a smaller number of basement membranes present in the eye, ear and kidney [3,4,5]. In humans, COL4A1 and COL4A2 share a common locus at 13q34 where they are arranged head to head with a bi-directional promoter [6–9]. COL4A1 (52 exons) and COL4A2 (48 exons) encode proteins of 1669 and 1712 amino acids, respectively, which share 45% identity. COL4A1 and COL4A2 both contain the same three major domains: The non-collagenous (NC1) domain is a globular domain at the carboxy terminus that is required for the initiation of heterotrimer formation between one COL4A2 and two COL4A1 proteins (α1(IV)2α2(IV)) [10,11]. The majority of each COL4A1 and COL4A2 protein comprises a collageneous triple helical domain characterized by repetition of Gly-Xaa-Yaa motifs (where Xaa and Yaa are variable amino acids). In contrast to fibrillar collagens, COL4A1 and COL4A2 have more than 20 interruptions of the Gly-Xaa-Yaa repeats that
encompass protein interaction domains and confer local flexibility [12,13]. Following assembly in the endoplasmic reticulum, heterotrimers are transported to the Golgi where they are packaged for vesicular release from the cell [2,3]. Once they reach the extracellular space, heterotrimers participate in macromolecular network organizations whereby two heterotrimers associate via their NC1 domains, and four heterotrimers form anti-parallel lateral interactions via association of their 7S domains.

A targeted mutation of the Col4a1/Col4a2 locus on murine Chromosome 8 generated mice with null alleles for both Col4a1 and Col4a2 and provided the first insights into their biological roles [14]. This work showed that formation of the collagen IV network is dispensable for initial basement membrane assembly, but critical for its structural integrity. Mice homozygous for the Col4a1;Col4a2 null allele died between embryonic days (E)10.5–11.5 with disrupted basement membranes including Reichert’s membrane [14]. In contrast, heterozygous mice were described as having no overt phenotype. Soon after, heterozygous, semi-dominant pathogenic COL4A1 and COL4A2 mutations were identified in humans and in multiple mouse lines that model the pathology in patients [15–17]. Consistent with the broad distribution of α1(IV)α2(IV) heterotrimers in nearly all basement membranes, COL4A1 and COL4A2 mutations cause multisystem disorders with abnormalities in the vasculature, brain, eyes, kidneys and muscles being the most commonly reported to date [18–20]. Cerebrovascular disease (CVD) is one of the most notable consequences of COL4A1 and COL4A2 mutations and comprises a growing constellation of clinical manifestations including porencephaly, small-vessel disease, leukoencephalopathy, intracranial aneurysms and recurrent intracerebral hemorrhages (ICH). Other cerebral defects include calcification and cortical malformations similar to those observed in Walker-Warburg syndrome and muscle-eye-brain disease. Patients also frequently present with ocular disease (microphthalmia, cataracts, anterior segment dysgenesis, Axenfeld Rieger syndrome, glaucoma, optic nerve hypoplasia and retinal vascular tortuosity), kidney disease (hematuria and renal cysts of variable severity) and muscular defects (cramps, elevated serum creatine kinase level and muscular dystrophy). An excellent review of clinical consequences of COL4A1 and COL4A2 was recently published [20]. It is likely that the present observations are not exhaustive and that pathology in additional tissues and organs will be identified as studies advance.

As of April 2016, there were 93 and 12 mutations identified in human COL4A1 and COL4A2, respectively, and 13 and 3 in the corresponding murine genes (not including three targeted Col4a1 alleles). This sample size seemed appropriate to explore emerging genotype-phenotype correlations. Identifying genotype-phenotype correlations can be useful for patient management, genetic counseling, cohort stratifications for drug response or adverse events in clinical trials and for providing a better understanding of the biological function of distinct protein sub-domains. Equivalent analyses were performed with similar numbers of patients who have mutations in other types of collagens and these studies have provided valuable insight. The purpose of this review is to draw upon comparisons with other types of collagens and the current literature of human COL4A1 and COL4A2 mutations and mouse models to help identify potential genotype-phenotype correlations and understand their implications.

2. Genotype-phenotype correlations in patients with mutations in other types of collagens

Type IV collagens are part of the collagen superfamily comprising 28 members that have the defining feature of a triple helical domain [1]. Mutations in multiple types of collagens contribute to a diverse constellation of human diseases. There is great a depth of knowledge of disease mechanism for many types of collagens which can be instructive for understanding genotype-phenotype correlations for COL4A1 and COL4A2 mutations. However, despite commonalities, members of the collagen superfamily are structurally and functionally diverse and there is no universal disease mechanism. For example, type VI collagen mutations can cause muscular dystrophies that are notable for the breadth of variation in disease severity. The type VI collagens are non-fibrillar collagens and one of the many structural features that are notable for the breadth of variation in disease severity is the presence of the unique collagenous Gly residues in the triple helical domain that are contiguous in a single chain. Type VI collagen mutations can cause both Ulrich congenital muscular dystrophy (a severe neonatal or early childhood disease) and Bethlem myopathy (where ambulation is retained into adulthood), which were considered to be distinct diseases until molecular characterization revealed a genotype-phenotype relationship that influences disease expressivity. Dominant missense mutations of Gly residues in the triple helical domain are a common feature of type VI collagen mutations and the most severe cases cluster within a small region of ~15 amino acids near the amino terminus of the triple helical domain that is proposed to represent a functional sub-domain important for muscle function [21]. The type II collagens represent a similar example of phenotypic dichotomy attributable to genotype. Type II collagens are fibril forming α1(II)3 heterotrimeric collagen, with expression in chondrocytes and in the
ocular vitreous. COL2A1 mutations cause Stickler syndrome type I, characterized by high myopia, vitreoretinal degeneration, retinal detachment and cataracts. On the other end of the disease spectrum resulting from COL2A1 mutations are severe forms of skeletal dysplasia, chondrodysplasia and achondrogenesis. In a relatively clear example of genotype–phenotype correlations, the preponderance of mutations causing Stickler syndrome are highly disruptive mutations that likely result in functionally null alleles. In a published series of 188 Stickler syndrome type 1 probands, the predominant cause was loss-of-function mutations comprising splice site alterations, nonsense mutations, or insertion/deletions (indels). Only five of the 77 mutations that caused Stickler syndrome were Gly substitutions in the triple helical domain [22]. There is further evidence to suggest that the Gly substitutions that cause Stickler syndrome type 1 are near the amino terminal quarter of the protein. In contrast, Gly substitutions nearer the carboxy terminus – especially those for which Gly is replaced by a bulkier amino acid – are more often associated with severe forms of skeletal dysplasia, achondrogenesis or severe short stature phenotypes [22–25].

In a departure from distinct clinical entities, vascular Ehlers–Danlos syndrome (vEDS) is caused by dominant mutations in COL3A1. Type III collagen is fibril forming α1(III)3 heterotrimers and the triple helical domain consists of 343 strict repeats of the Gly-Xaa-Yaa motif. Typical consequences of vEDS are arterial dissections, bowel perforations and uterine ruptures [26,27]. A recent report [28] of genotype–phenotype associations in 146 vEDS index patients with molecular diagnosis found 126 distinct mutations that fit into five distinct categories: Gly substitutions in the triple helical domain (56.3%), splice-site and in-frame indels (28.6%), haploinsufficiency (5.5%), non-Gly substitutions in the triple helical domain (3.2%) and mutations outside of the triple-helical domain (6.3%). As a measure of relative severity the age at diagnosis was younger for patients with in-frame splice site mutations (25 years) than for those with Gly substitutions (34 years) and both were younger than for patients with other classes of mutations (45 years). Although there was a frequency bias for substitutions of Gly to more destabilizing amino acids (Aspartic acid, Glutamic acid, Glutamic acid, Valine), no correlation with disease severity was observed [28]. Instead, the position of the glycine substitutions appears to play a role with those nearer the carboxy terminus being more severe [27]. Moreover, the prognosis was better for patients who did not have Gly substitutions or in-frame coding mutations as they were generally spared from digestive complications. In a separate study of 630 index cases for which survival was the outcome measure, patients with triple helical glycine substitutions had significantly shorter survival than patients with null mutations, however, in this case the identity of the substituting amino acid did influence the outcome but the position of the mutation did not [29].

Perhaps the richest source of information for collagen-related genotype–phenotype correlations comes from the type I collagens. Unlike collagen II and III that form homotrimers, collagen I (like collagen IV) form α1(I)α2(I) heterotrimers. Mutations in COL1A1 and COL1A2 cause Osteogenesis imperfecta (OI), a genetically and clinically heterogeneous disorder characterized by skeletal deformities and fragile bones that are susceptible to fracture. An early influential analysis of approximately 70 COL1A1 and COL1A2 mutations identified genotype–phenotype correlations that underscored the distinction between the quantitative and qualitative effects of mutations [30]. As larger sets of patients became available a clearer understanding emerged that appears to have broad applicability across many types of collagens [31]. Mutations were associated with milder forms of OI if they caused quantitative defects (mutations in the carboxyl-terminal domain that preclude the mutant protein from integrating into heterotrimers and frameshifts and null alleles that trigger loss of mRNA leading to a reduction in protein levels). In contrast, qualitative defects (such as Gly substitutions in the triple helical domain, exon skipping and rearrangements) responsible for protein structural defects were associated with more severe forms of OI [30,32]. Because mutations within the triple helical domain are common and associated with severe forms of disease they have been the subjects of greatest interest. A definitive study of more than 800 COL1A1 and COL1A2 Gly mutations and altered splice sites in the triple helical domain allowed the discovery and confirmation of a number of genotype–phenotype correlations [31]. Of the 832 independent mutations, 682 resulted in triple helical Gly substitutions (391 in COL1A1 and 291 in COL1A2) and 150 altered splice sites (102 in COL1A1 and 48 in COL1A2) [31]. COL1A1 splice site mutations often shifted the open reading frame resulting in absence of the mutant chain and manifested as quantitative mutations that were rarely severe (lethal). In contrast, triple helical glycine substitutions produced mutant proteins that incorporate into heterotrimers leading to important qualitative differences. Large-scale analyses of the effects of triple helical Gly substitutions produce a number of interesting genotype–phenotype correlations. Gly is represented by four codons: GGU, GGCG, GGA, GGG and transitions or transversions of either of the first two guanine nucleotides results in Gly substitutions to Alanine (Ala), Arginine (Arg), Aspartic acid (Asp), Cysteine (Cys), Glutamic acid (Glu), Serine (Ser), Tryptophan (Trp) or Valine (Val). One third of the Gly substitutions were lethal - especially when the substituting amino acid is
charged (Asp and Glu), or has a branched side chain (Arg and Val). When these substitutions were non-lethal they associated with severe outcomes compared to substitutions for Cys, Ser, or Ala. Furthermore, the influence of amino acid substitutions on disease severity was superimposed on the specific location of the mutation. For instance, when comparing only substitutions of Cys for Gly there was a relatively smooth phenotypic gradient from mild to severe in an amino-to-carboxy direction, suggesting a position effect [30]. While in a larger patient cohort no linear relationship emerged, with few exceptions, Gly substitutions within the first 200 amino acids of COL1A1 were non-lethal (on the mild end of the OI spectrum) whereas similar mutations closer to the carboxy-terminal end were often severe [31]. Additionally, there were two specific sub-domains associated with lethal phenotypes that align with major ligand binding regions and gap regions without reported mutations that contain putative integrin-binding sites, suggesting critical functions [31]. Mutations causing severe forms of OI were later described in a gap region, re-enforcing the hypothesis that these regions have important functional roles [33].

Notably, compared to COL1A1 mutations, COL1A2 mutations were identified much less frequently (291 vs 391 for Gly substitutions and 48 vs 102 for splice site mutations) and there were both similarities and differences in the outcomes [31]. Among patients with COL1A2 glycine substitutions, those with Asp, Glu and Arg – but not Val – tended to have worse outcomes. COL1A2 Gly substitutions were predominantly non-lethal, with the exceptions of exon skipping mutations in the carboxyl half of the protein and glycine substitutions that occurred in eight regularly spaced clusters [31]. The lethal clusters do not align with major ligand binding regions but were in regions suggested to be critical for interactions between proteoglycans and the collagen fibrils.

Overall, these studies revealed that mutations causing quantitative differences where the mutant allele is degraded or not incorporated into trimers are found less frequently and are associated with the mildest outcomes. Mutations that cause qualitative differences produce a wide range of phenotypes that depends on many factors including the gene involved and nature and location of the mutation. Analysis of COL1A1 demonstrated that, in addition to important functional subdomains, worse patient outcomes were associated with Gly substitutions for charged or branched amino acids and for Gly substitutions nearer the carboxy terminus. The position of mutations in COL1A2 appears to have less influence unless the occurred within one of the 8 regularly spaced clusters [31].

Mutations in the other type IV collagen isoforms also cause autosomal recessive (COL4A3 and COL4A4) or X-linked (COL4A5) Alport Syndrome and may offer insight into genotype–phenotype correlations [34]. The proteins encoded by these three genes form α3(IV)/α4(IV)/α5(IV) heterotrimers that are present in basement membranes of the glomerulus, cochlea and eye [10,35]. Alport Syndrome is characterized by progressive renal failure, sensorineural hearing loss, lenticulovenous, cataract and maculopathy [36,37]. Approximately 80–85% of Alport Syndrome cases are X-linked and caused by COL4A5 mutations [38,39]. Studies of large cohorts of males with X-linked Alport Syndrome reveal genotype–phenotype correlations that are often contradictory to those found in type I collagens. The consensus from studying over 1000 patients with COL4A5 mutations was that disease severity (age at onset for end stage renal failure) was greatest for truncating mutations, followed by splice variants and finally missense mutations [40–44]. Moreover, truncating or splicing mutations had two-fold greater odds of having ocular pathology compared to missense mutations [41]. This is in contrast to the concept that emerged from type I collagens where outcomes for qualitative mutations are worse than for quantitative mutations. Heterotrimer stoichiometry can offer an explanation for this observation since severe loss of function mutations cannot be compensated for by the homologous chromosome, leaving cells with a virtual absence of COL4A5. Position dependent effects were also reported, however, these too appear to be opposite to those reported in type I collagens. Mutations in the amino terminal signal peptide, regardless of type, led to end stage renal failure at younger ages than mutations in the triple helical or NC1 domains [41]. One study that only looked at triple helical glycine substitutions suggested that mutations in exons 1–20 (amino terminal) might have been less severe than those in exons 21–47 (carboxy terminal) [40] while a second, larger study found that mutations nearer the amino terminus were more severe [41]. Finally, a recent study suggested that missense substitutions for bulkier amino acids were associated with more severe outcomes [43].

3. Genotype–phenotype correlations in patients with COL4A1 or COL4A2 mutations

This review includes 105 independent occurrences of COL4A1 or COL4A2 mutations of which, 93 were in COL4A1 (at 74 unique sites) and 12 in COL4A2 (at 10 unique sites) (Tables 1 and 2). All of the mutations described were dominant with equal representation in males (48%) and females (52%) and approximately equal numbers of inherited (47%) or sporadic (53%) incidences. Similar to other types of collagens, triple helical Gly substitutions
represented the most prevalent class of mutations (68 out of 93) and the most common substitutions were for charged amino acids; Arg (30), Glu (11), and Asp (10). The remaining substitutions were Ser (7), Val (5), Ala (3), Leu (1), and a termination codon (1). No substitutions for tryptophan or cysteine, which are the only other substitutions possible with a single nucleotide change of glycine codons, were reported. Splice site mutations within the triple helical domain constitute the second most prevalent class of mutations (8 out of 93). Three of the splice site mutations caused in-frame deletion of an entire exon (exon 21, 25 or 31) and two mutations resulted in nonsense-mediated decay of the transcripts with reduced COL4A1 expression while the consequences of the other three were not reported. Other classes of mutations were much less frequent including Yaa position substitutions (3), substitutions within the repeat interruptions (3) and frame-shifting indels (3). Eight mutations outside of the triple helical domain have also been reported, seven of which were located in the NC1 domain and one substitution mutation of the start methionine.

The low number of published COL4A2 mutations decreases the confidence for identifying broad trends, however, triple helical glycine substitutions (6) were also the most prevalent class of mutations where Gly was substituted for Arg (2), Glu (2) and Asp (2). One patient had a frame-shifting deletion and a second patient had a missense mutation, each occurring in the NC1 domain. There were also examples of triple helical Xaa position substitutions (4), however, three of these were identified independently in the same amino acid and this variant was also identified in controls raising the possibility that this variant might be non-pathogenic. If so, the total number of COL4A2 mutations could be as low as nine with six being Gly substitutions in the triple helical domain. The failure so far to identify mutations in the repeat interruptions could represent differences between COL4A1 and COL4A2 or simply reflect the low number of COL4A2 mutations discovered to date.

Overall, COL4A2 mutations appear to associate with milder phenotypes (later age-at-onset of CVD, and fewer systemic abnormalities reported). However, because of the limited number of COL4A2 mutations, we focus primarily on the 93 COL4A1 mutations for the investigation of genotype–phenotype correlations in patients. Fourteen COL4A1 mutations were presumed to reduce the amount of normal α1(IV)α2(IV) heterotrimers without producing abnormal heterotrimers and may therefore qualify as quantitative mutations: seven mutations in the NC1 domain, three frame-shifting indels, two frame-shifting splice mutations, one substitution of the start methionine, and one triple helical glycine substitution leading to a premature termination codon. Six out of the seven NC1 mutations and 12 out of the 14 total “quantitative” mutations were reported in patients with porencephaly and the remaining two patients had perinatal ICH. If one considers porencephaly and early age-at-onset as phenotypes on the severe end of the CVD spectrum, quantitative COL4A1 mutations appear to associate with relatively poor outcomes.

To compare the effects of qualitative measures, we focused on the 68 triple helical Gly substitutions in COL4A1. The frequencies of Arg, Glu, and Asp substitutions might suggest that the outcomes of

<table>
<thead>
<tr>
<th>Table 1. Distribution of COL4A1 and COL4A2 mutations.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Start codon</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Gly of Gly-Xaa-Yaa motif</td>
</tr>
<tr>
<td>Yaa of Gly-Xaa-Yaa motif</td>
</tr>
<tr>
<td>Xaa of Gly-Xaa-Yaa motif</td>
</tr>
<tr>
<td>Repeat Intermittences</td>
</tr>
<tr>
<td>Frameshift</td>
</tr>
<tr>
<td>Splice-site</td>
</tr>
<tr>
<td>NC1 domain</td>
</tr>
<tr>
<td>Missense</td>
</tr>
<tr>
<td>Nonsense</td>
</tr>
<tr>
<td>Duplication</td>
</tr>
<tr>
<td>Insertion</td>
</tr>
<tr>
<td>Totals</td>
</tr>
</tbody>
</table>
these mutations are more severe, however this conclusion would need to be corrected for codon usage and integrate other potential factors. In four pairs of instances the same glycine residue was independently substituted by distinct amino acids (G498V, G498D; G655R, G655E; G708R, G708V; G990E, G990V). However, in all cases, the substituting amino acids are either branched or charged and so might all be expected to have severe consequences and there were no obvious differences that could be attributed to the identity of the substituting amino acid. To further investigate potential differential effects of substituting amino acids, we compared the consequences of the 30 Arg substitutions to the Ser (7) and Ala (3) substitutions. No obvious correlation was observed when comparing substitutions to branched amino acids. To further investigate potential differences and there were no obvious differences that could be attributed to the identity of the substituting amino acids. For example, G670R caused porencephaly, whereas G708R caused ocular defects but no CVD and G696S had adult onset ICH. Of four independent occurrences of G755R mutations, none lead to porencephaly and instead all result in childhood or adult onset CVD. In contrast, two independent occurrences for the nearby G749S mutation caused porencephaly. Although these analyses indicate that substitutions to branched amino acids occur more frequently, there is no clear evidence that they are associated with more severe outcomes.

Notably, the position of the mutation within the triple helical domain appeared to correlate with CVD severity. There are two Gly substitutions in the first quarter of the triple helical domain suggesting that Gly missense mutations near the amino-terminal end of the triple helical domain might cause milder phenotype. The 30 Arg substitutions occur at 20 unique sites allowing comparison of the effects of position for the same substituting amino acid. If one divides the triple helical domain roughly in half, 14 mutations are in the carboxy-terminal half (amino acid 805 and higher) and, of these, 12 caused porencephaly and two caused schizencephaly. In addition to early age-at-onset, these mutations were more often sporadic or de novo cases and were frequently associated with epileptic seizures and

Table 2. Clinical features associated with COL4A1 and COL4A2 mutations in patients.

<table>
<thead>
<tr>
<th>COL4A1</th>
<th>COL4A2</th>
<th>Phenotype</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>G498V</td>
<td>G498D</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G655R</td>
<td>G655E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G708R</td>
<td>G708V</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G990E</td>
<td>G990V</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G749S</td>
<td>G749R</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G670R</td>
<td>G670D</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G708R</td>
<td>G708D</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G696S</td>
<td>G696D</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G688D</td>
<td>G688E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G658D</td>
<td>G658E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G612D</td>
<td>G612E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G882D</td>
<td>G882E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G720D</td>
<td>G720E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G773R</td>
<td>G773E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G755R</td>
<td>G755E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G990R</td>
<td>G990E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G870R</td>
<td>G870E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G888R</td>
<td>G888E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G805R</td>
<td>G805E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G773R</td>
<td>G773E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G755R</td>
<td>G755E</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>G708R</td>
<td>G708E</td>
<td>None</td>
<td></td>
</tr>
</tbody>
</table>

Legend: Reported present | Reported absent | Not reported.
Genotype–Phenotype Correlations in Pathology Caused by Collagen Type IV alpha 1 and 2 Mutations

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Phenotype Correlations in Pathology Caused by Collagen Type IV alpha 1 and 2 Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>COL4A1</td>
<td></td>
</tr>
<tr>
<td>COL4A2</td>
<td></td>
</tr>
</tbody>
</table>

| Information for this Table was taken from the following references [15,17,20, 45–99] |

motor disorder and/or developmental delay. In contrast, of the 16 Gly to Arg mutations in the amino-terminal half of the protein, only five caused porencephaly, four mutations did not cause ICH/stroke, four mutations were identified in patients who were adults at the time of diagnosis and only four caused epileptic seizures.

In addition to modulating CVD severity, the location of the mutation may also influence the penetrance of other phenotypes. For instance, COL4A1 mutations cause ocular disease with high penetrance (72%), however the specific ocular defect depends on the position of the mutation. For example, mutations in the amino-terminal one-third of the protein caused retinal arteriolar tortuosity with complete penetrance, whereas mutations causing cataract, anterior segment dysgenesis, glaucoma and microphthalmia tended to occur closer to the carboxy terminus. Of interest, in the amino-terminal one-third of the protein there were mutations reported in six families with a clinical diagnosis of HANAC Syndrome (hereditary angioopathy with nephropathy, aneurysms, and muscle cramps) [51–53,112]. The mutations cluster in exons 24 and 25 within a 31 amino acid region (G498–G528) of the triple helical domain that encompasses several integrin-binding domains. While nephropathy, myopathy and aneurysms are not restricted to patients with mutations in this region, these phenotypes may be disproportionately represented in patients from these families. These patients also had retinal arteriolar tortuosity and heart disease but other features including porencephaly, cerebral calcification, cataract, glaucoma, microcephaly or epileptic seizures were not described. Notably, while three Gly substitutions within exon 25 were reported in patients with HANAC syndrome, a splice site mutation causing an in-frame deletion of this exon was found in a patient with bilateral diffuse white matter abnormalities, multiple clinically asymptomatic hemorrhages and finally an acute ICH but without aneurysms, myopathy or renal involvement [56]. Observations from patients with triple helical Gly substitutions support the existence of one or more functional subdomains within this domain that may associate with differential co-morbidities.

There are a number of limitations that demand approaching genotype–phenotype correlations in patients with caution. Most of the mutations occurred only once and to label the ‘severity’ of a mutation based on a sample size of one risks oversimplification of many complex genetic, environmental and stochastic interactions that are integrated in the phenotype. Even in cases of recurrent mutations, the conclusions are not simple. There were eight amino acids that had multiple
hits and G755R and G773R were each reported in four independent studies. While the outcomes for G755R were comparable (all patients had ICH, leukoencephalopathy and cataracts with an age of onset between 10 and 21 years, and did not cause porencephaly or epilepsy) those for G773R were less so (three mutations caused porencephaly or severe stroke whereas one mutation cause only very mild white matter changes). Genetic context differences could explain some of the discordance in the expressivity resulting from a given mutation. However, phenotypic differences within families who are more closely matched genetically, including probands with porencephaly born to asymptomatic parents who carried mutations, were also observed. There are multiple possible explanations for these cases including genetic modifiers, mosaicism, biological variation or environmental influences.

4. Genotype–phenotype correlations in Col4a1 and Col4a2 mutant mice

Studies of mouse models with Col4a1 and Col4a2 mutations are a powerful way to simplify the genetic and environmental influences and complement patient studies. In addition to the Col4a1;Col4a2 null mutation [14], and a Col4a1 conditional mutation [101,110], there are presently 17 mutations reported in Col4a1 and Col4a2 that cause pathology in multiple tissues and organs [16,17,100,104] (Fig. 1). Because of procedural differences between research groups, notably the focus and depth of analysis, it is difficult to directly compare graded levels of relative severities for each mutation. Importantly, the genetic context in which the mutations are studied also has a significant effect on the penetrance and/or severity of pathology [18,80,101,111]. For example, Col4a1 mutant mice maintained on a pure C57BL/6J background have severe ocular dysgenesis, myopathy and ICH that are all suppressed in mutant mice that were crossed for a single generation to the CAST/EiJ inbred strain [15,18,80,101,111]. Notably, the 129SvEvTac inbred strain also suppressed anterior segment dysgenesis, but not ICH, indicating that genetic context can have differential effects on distinct phenotypes and suggesting heterogeneity in the pathogenic mechanisms underlying different Col4a1-related phenotypes [101].

To systematically investigate potential genotype–phenotype correlations, nine Col4a1 and one Col4a2 mutations [100] were uniformly backcrossed to the C57BL/6J genetic background [101,102]. The allelic series comprised seven missense mutations of glycine residues in the triple helical domain (six in COL4A1, one in COL4A2), one missense mutation

![Fig. 1. Summary of reported phenotypes for murine Col4a1 and Col4a2 mutations. (A) Schematic diagram illustrating murine mutations reported in the literature [14–18, 80, 100–111]. (B) List of Col4a1 and Col4a2 mutations and the phenotypes that have been reported for each.](image-url)
in the NC1 domain of COL4A1, and a point mutation of splice acceptor site that results in the deletion of exon 41 ($\Delta$ex41) but retains the open reading frame (Fig. 2A). By controlling environmental factors and the genetic context, studies using this allelic series demonstrated allelic heterogeneity between mutations. The severity of ICH was compared between strains of mutant mice aged for 7–9 months and different classes of mutations that implicated three potential genotype–phenotype correlations were identified [101,102]. First was a domain effect whereby the NC1 domain mutation ($\text{Col4a1}^{S1582P}$), which presumably reflects a quantitative mutation (Fig. 2B), caused less severe ICH than the triple helical domain mutations (Fig. 2C). Second, for point mutations within the triple helical domain, there was a position effect whereby mutations nearer the carboxy termini (where heterotrimer winding is initiated) caused more severe ICH than mutations nearer the amino termini. Third, there appeared to be a ‘class effect’ whereby $\text{Col4a1}^{\Delta\text{ex41}}$ was more severe than missense mutations even when compared to a glycine point mutation within exon 41 ($\text{Col4a1}^{G1180D}$). The pathogenicity resulting from $\text{COL4A1}$ and $\text{COL4A2}$ mutations is generally attributed to impaired secretion and concomitant intracellular accumulation and extracellular deficiency of mutant heterotrimers. Notably, ICH severity in this allelic series correlated with the relative levels of heterotrimer accumulation between mutations (Fig. 2B and C) [101,102].

An interesting contrast was observed when myopathy was quantified in this same allelic series [102]. Similar to the previous trends, mice with a presumed quantitative mutation in the NC1 domain ($\text{Col4a1}^{S1582P}$) had little or no phenotype and myopathy resulting from the $\text{Col4a1}^{\Delta\text{ex41}}$ mutation was more severe than that resulting from the ‘position-matched’ point mutation ($\text{Col4a1}^{G1180D}$) (Fig. 2D). However, in contrast to the trends observed in ICH, there was no apparent position effect for mutations within the triple helical domain. Instead, $\text{Col4a1}^{G394V}$, which was among the mutations with the

![Fig. 2. Allelic heterogeneity for heterotrimer biosynthesis, intracerebral hemorrhage and myopathy. (A) Schematic diagram illustrating the mutations in the allelic series. All mutations are in the triple helical domain, except $\text{COL4A1}^{S1582P}$, which is in the C-terminal NC1 domain. $\text{COL4A1}^{\Delta\text{ex41}}$ is a splice site mutation causing deletion of exon 41. (B) Western blot quantification demonstrating that the mutation in the NC1 domain ($\text{COL4A1}^{S1582P}$) had relatively low levels of intracellular heterotrimer accumulation and that there is a position-dependent graded severity of heterotrimer accumulation for mutations within the triple helical domain whereby mutations nearer the NC1 domain had higher levels of heterotrimer accumulation. (C) Quantification of intracerebral hemorrhage revealed that $\text{COL4A1}^{\Delta\text{ex41}}$ mutation leads to the most severe phenotype and that point mutations in the triple helical domain nearer the C-terminus tended to cause more (or more severe) hemorrhages. (D) Quantification of non-peripheral nuclei revealed that the $\text{Col4a1}^{G394V}$ mutation, which is in an integrin-binding domain, causes the most severe myopathy. For all groups, $n \geq 5$. Comparisons between wild-type ($\text{Col4a1}^{+/+}$) and mutant mice were performed using Student’s t-test; *, $p < 0.05$; **, $p < 0.01$. For comparison among the different strains, a one-way ANOVA followed by a Tukey’s post hoc test was performed; †, $p < 0.05$. Panels A, B and D are modified from Kuo et al. (2014) [102] and Panel C is modified from Jeanne et al. (2015) [101].
least intracellular accumulation and mildest ICH, had the most severe myopathy demonstrating an interesting dichotomy for this mutation. Col4a1^{G394V} occurs near a putative integrin-binding site adjacent to the cluster of HANAC mutations, in a region that may represent a functional sub-domain with an important role in muscle development and function [102].

Kidney abnormalities (albuminuria and thin or split glomerular basement membranes) were previously reported in Col4a1^{+/-\Delta ex41} mice [15] but had not been examined in other strains from the allelic series. Here we show that hematuria shares a profile that resembles ICH for the Col4a1 mutations whereby the Col4a1^{51582F} mutation had almost no phenotype, the Col4a1^{\Delta ex41} mutation was worse than Col4a1^{G1160D} and mutations nearer the carboxy termini were more severe than those nearer the amino termini (Fig. 3A). These data may suggest that hematuria has a vascular origin or that hematuria and ICH share pathogenic mechanisms. When we measured albuminuria the results were similar with the notable exception that the Col4a2 mutation (Col4a2^{G646D}) was disproportionately severe compared to Col4a1 mutations (Fig. 3B). This observation might suggest that COL4A2 is more important for nephropathy or that this mutation occurs within a functional sub-domain important for renal pathology. In support of the latter alternative, the Col4a1^{G658D} mutation, which is nearby, lead to the most severe albuminuria among the Col4a1 point mutations.

Another potential genotype–phenotype correlation was reported in an allelic series of three mutations: Col4a1^{G627W}, Col4a1^{K950E} and Col4a1^{G1046D} (also called Bru, Raw and Svc, respectively) [16]. The authors reported that the Gly substitutions (Col4a1^{G627W} and Col4a1^{G1046D}) caused relatively more severe ocular and renal phenotypes than a Lys substitution affecting the Yaa position amino acid in the Gly-Xaa-Yaa motif (Col4a1^{K950E}), which only causes a retinal vascular phenotype. Col4a1^{G627W} was identified in a different genetic background, nonetheless, Col4a1^{K950E} and Col4a1^{G1046D} were studied in similar genetic contexts and the position of the mutation within the Gly-Xaa-Yaa repeat is a plausible explanation for the mild phenotype(s) in Col4a1^{K950E} mice. An initial report of these alleles suggested that the glomerular basement membrane was unremarkable but Bowman’s capsule had basement membrane defects and hypertrophic cells [16] which differed from the ultrastructural glomerular basement membrane abnormalities [15], albuminurea and hematuria described for the other alleles. More recently, a detailed characterization of renal pathology caused by Col4a1^{K950E} and Col4a1^{G1046D} mutations described extensive pathology and potentially distinct involvement of glomerular and tubular defects [107]. Notably, differences in focus, techniques and study depth underscore the challenges of comparing genotype–phenotype data between research groups. A targeted allele that is analogous to a human HANAC-causing mutation (Col4a1^{G498V}) was also recently published [104] and while the relative severities of the phenotypes are difficult to compare in isolation, this report also

Fig. 3. Relative measures of hematuria and albuminuria in mice from an allelic series of Col4a1 and Col4a2 mutations. We collected urine and measured levels of hematuria and albuminuria using colorimetric dipsticks and a clinical chemistry autoanalyzer, respectively (described in Gould et al. [15]). A) The proportion of animals of each genotype with no, mild, moderate or severe hematuria are indicated as percentages. Sample sizes for each genotype are indicated at the top. B) We measured albuminuria by calculating the ratio of albumin to creatinine (mg/g) and indicated the proportion of animals of each genotype with no, mild (<30), moderate (30–60) or severe (>60) albuminuria as a percentage. Sample sizes for each genotype are indicated at the top.
includes a thorough characterization of renal pathology and supports the existence of one or more amino-terminal sub-domain(s) that disproportionately influence the severity of myopathy and nephropathy. Despite a limited number of available mutations, findings from studies using murine models of Col4a1 mutations, including two allelic series, provide insights into disease mechanisms and support the conclusion that there are genotype—phenotype correlations. The data suggest that 1) mutations in the NC1 domain (and likely other ‘quantitative’ mutations) may associate with milder outcomes than those in the triple-helical domain; 2) Gly substitutions in the triple helical domain have position-dependent effects on biosynthesis that correlate with CVD, but not other phenotypes; 3) in frame, exon-skipping mutations within the triple helical domain may be disproportionately severe compared to Gly substitutions 4) Yaa position substitutions may be less severe than Gly substitutions 5) one or more functional sub-domains near the amino-termini may be involved in myopathy and possibly nephropathy. The fact that genetic background and genotype—phenotype correlations had differential effects on different phenotypes suggest that there is mechanistic heterogeneity.

5. Conclusions and perspectives

Mutations in COL4A1 and COL4A2 have been studied for nearly a decade and, compared to type I collagen mutations underlying OI and COL4A5 mutations underlying Alport Syndrome, the number of patients from which to draw genotype—phenotype correlations is relatively limited. However, this retrospective analysis suggests the existence of genotype—phenotype correlations for COL4A1 and COL4A2 mutations that are consistent with observations for some other types of collagens. In both humans and mice, there is a frequency bias for mutations in COL4A1 over COL4A2. This inequity likely reflects a combination of ascertainment bias and true biological differences. Mutations in COL4A1 were reported several years before mutations in COL4A2 and routine COL4A2 screening in patients lagged. As mutation detection shifts from candidate gene sequencing to exome analyses this bias should disappear. However, ascertainment bias alone does not explain the discrepancies in the numbers of mutations in each gene. The genes are roughly equal in size and would be expected to acquire mutations at equivalent rates in patients and during random chemical mutagenesis. It is possible that milder outcomes for COL4A2 mutations may explain the frequency difference in mutations identified. It is interesting to note however that in the murine allelic series, the Col4a2 mutation essentially behaved like a position-matched Col4a1 mutation with the exception of one phenotype (albuminurea) for which this mutation was more severe.

As it is the case for other types of collagens, Gly substitutions were the most common class of mutation in patients. Mutations in 50 out of 437 triple helical Gly residues in COL4A1 and six out of 428 triple helical Gly residues for COL4A2 have been identified. While substitutions for branched amino acids occurred more frequently, the identity of the substituting amino acid did not appear to have a strong influence on disease severity. Instead, there appear to be important positional influences; mutations located near the amino terminus of COL4A1 are associated with milder pathology, and the severity of CVD tends to be worse for Gly substitutions that are closer to the carboxyl end of the triple helical domain. Finally, mutations that clustered near the region containing integrin-binding domains appear to be associated with a greater likelihood for patients to have myopathy and nephropathy suggesting the existence of a functional subdomain influencing the penetrance of distinct phenotypes.

In findings from the murine allelic series the mutation with presumed quantitative effects (S1582P in the NC1 domain) had milder effects on heterotrimer biosynthesis, ICH and myopathy than most mutations with presumed qualitative effects (mutations in the triple helical domain). In contrast, in patients, qualitative COL4A1 mutations do not appear to be associated with more severe outcomes than quantitative mutations. As there are biases toward identification of mutations with severe outcomes in early studies, it is possible that these classes of mutations will diverge as more patients are reported with a broader spectrum of mutations. An additional caveat is that the molecular consequences of most mutations were inferred and not tested, which could lead to erroneous classification. Importantly, although porencephaly and early age-at-onset CVD were considered to be more severe than adult age-at-onset CVD, determining the relative outcome severities is subjective and remains a pervasive challenge. There is the additional challenge in integrating multiple co-morbid phenotypes, some of which appear to be mechanistically distinct. In contrast, characterizations of OI severity relies on well-established criteria for bone fragility that are more quantitative and focused on a relatively defined outcome. Finally, the roles of different environmental factors (including anti-coagulants, surgical delivery, and exercise) [15,101] must all be considered as they may contribute to misleading conclusions.

In contrast to patient studies, the allelic series of mutant mice comprises less saturation of mutations across the protein or for each class of mutation. For example, there is only one presumed qualitative mutation in the NC1 domain. Although the effect was
consistent across phenotypes (and consistent with the absence of phenotypes in mice heterozygous for Col4a1; Col4a2 null alleles) these findings are based upon observations for a single allele and it would be overly simplistic to presume that mutations across the entire domain will behave equally. However, these resources are powered by standardized environments, genetic homogeneity and large sample sizes for an individual mutation which all improve reproducibility. Continued efforts to characterize pathology in Col4a1 and Col4a2 mutant mice will improve our understanding of the phenotypic spectrum of Col4a1- and Col4a2-related disease. Better definition of the genotype–phenotype correlations and patient outcomes will help to understand risks including co-morbidities or age-at-onset and will ultimately impact therapeutic approaches, genetic counseling and patient care. This information will ideally come from large, prospective, longitudinal studies to detail the natural histories of patients for whom there are biochemical data, and quantitative outcome measures for multiple phenotypes. Together, these and other advances may reveal pathways that can be targeted therapeutically to reduce, delay or prevent aspects of this syndrome.

Acknowledgements

We thank Dr. Cassandre Labelle-Dumais for the comments on the manuscript and assistance with the figures. This study was funded in part by an American Heart Association post-doctoral fellowship to Dr. Jeanne (13POST1420019), National Institute of Neurological Disease and Stroke research grant to Dr. Gould (NS083830) and support to the Department of Ophthalmology from the National Eye Institute (core grant EY02162) and Research to Prevent Blindness (unrestricted grant).

Received 12 August 2016; Accepted 10 October 2016
Available online 26 October 2016

Abbreviations used:

COL4A1, collagen type IV alpha 1; COL4A2, collagen type IV alpha 2; CVD, cerebrovascular disease; HANAC, hereditary angiopathy with nephropathy, aneurysms and muscle cramps; ICH, intracerebral hemorrhages; I. O., Osteogenesis imperfecta; NC1, non collagenous 1; VEDS, vascular Ehlers–Danlos syndrome; Ala, Alanine; Arg, Arginine; Asp, Aspartic acid; Cys, Cysteine; Glu, Glutamic acid; Gly, Glycine; Lys, Lysine; Ser, Serine; Trp, Tryptophan; Val, Valine; WWS, Walker-Warburg Syndrome.

References

null


