Comparative genomic analysis of collagen gene diversity

Farhan Haq¹ · Nabeel Ahmed² · Muhammad Qasim³

Received: 14 March 2017 / Accepted: 6 February 2019 / Published online: 14 February 2019
© King Abdulaziz City for Science and Technology 2019

Abstract

Collagen gene family, comprising 30% of the total protein mass in mammals, is the major part of extracellular matrix. To understand the complexity of collagen gene family, detailed sequence, phylogenetic and synteny analyses of 44 collagen genes were performed. According to sequence analysis results, Fibril-associated collagen with interrupted triple helices (FACITs) were identified as the most recently evolved vertebrate-specific collagens while Fibril-forming collagens and Collagen VI, VII, XXVI, and XXVIII were the most ancient collagens, originating at the time of choanoflagellates. Network-forming collagens were entirely conserved from arthropods to homo sapiens, except one gene loss event. Of note, bird specific gene dispensability of COL1A1, COL3A1, COL5A3 and COL11A2 genes was observed in Fibril-forming collagens. According to phylogenetic analysis, gene duplications in collagen family occurred at variable time points during invertebrate to vertebrate evolution. However, majority of gene duplications in FACITs and network-forming collagens occurred at fish time point, suggesting large scale duplications at the root of vertebrate lineage. Lastly, synteny analysis identified 12 conserved blocks containing 27 collagen genes in vertebrate species. Interestingly, dysregulation of seven conserved blocks including block1 (COL11A1), block3 (COL3A1, COL5A2), block5 (COL6A5, COL6A6), block7 (COL1A2), block9 (COL4A1, COL4A2), block11 (COL6A1, COL6A2, COL18A1) and block12 (COL4A5, COL4A6) were also reported in different diseases including cancer. The current study revealed many critical insights into sequence, structural and functional diversity of collagen gene family. In future, by using this information we may be able to establish the clinical and pathological relevance of these conserved collagen blocks in different diseases.

Keywords Collagen · Domain · Duplication · Deletion · Phylogenetics · Synteny

Introduction

Collagens are ancient extracellular matrix proteins covering 30% of the total protein mass in mammals. Among all types of collagens, fibril collagens are the most abundant (Abedin and King 2010; King et al. 2008). A number of studies have been done on the sequence, structural and functional aspects of collagens (Brodsky and Ramshaw 1997; Byron et al.

2013; Exposito et al. 2010; Nagase and Fields 1996; Ricard-Blum 2011; Staatz et al. 1991). According to these studies, the first collagen gene was reported more than 600 million years ago in *Monosiga brevicollis*, the most closely related unicellular organisms with animals (Chu 2011; King et al. 2008). In present day vertebrate species, 44 collagen gene members are present suggesting excessive collagen diversity due to several small scale, large scale and whole genome duplication events occurring throughout the animal evolution (Abbasi 2008; Dennis and Eichler 2016; Meyer and Van de Peer 2005).

The complex collagen family is classified into different groups based on triple helix domain, length, and molecular weight (Chu 2011). These groups include Fibril collagens, Fibril-associated collagens with interrupted triple helices (FACITs), network-forming collagens, Collagens VI, VII, XXVI, and XXVIII, Membrane collagens and Multiplexins (Ricard-Blum 2011). The functional and evolutionary studies related to these groups have been done to understand

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13205-019-1616-9>) contains supplementary material, which is available to authorized users.

✉ Farhan Haq
farhan.haq@comsats.edu.pk

¹ Department of Biosciences, COMSATS University, Islamabad, Pakistan

² Department of Software Engineering, National University of Science and Technology, Islamabad, Pakistan

³ School of Medicine, AJOU University, Suwon, South Korea

their specific roles in different species (Boot-Handford and Tuckwell 2003; Byron et al. 2013; Egeblad et al. 2010; Gelse et al. 2003; Leitinger 2011; Rychel et al. 2006; Wada et al. 2006). Phylogenetic studies revealed that the abundance of Fibril-forming collagens in vertebrates is due to whole genome duplication (WGD) events which occurred at the time of vertebrate origin, the concept famously known as 2R hypothesis (Morvan-Dubois et al. 2003). On the contrary, some studies suggested that the abundance of collagens is due to the continuous small scale duplications and chromosomal rearrangements occurring throughout animal evolution (Abbasi 2008; Asrar et al. 2013). Yet, comprehensive phylogenetics-based classification of 44 collagen gene members is lacking.

In addition to excessive duplications, conservation of collagen gene members is also observed. These conserved syntenic blocks show resistance to sequence variation due to structural and functional conservation (Ehrlich et al. 1997; Kemkemer et al. 2009). The disruption of these syntenic blocks can interrupt the normal function of the genes, ultimately leading to different anomalies (e.g., cancer) (Huang et al. 2006; Kemkemer et al. 2009). In a previous study, conserved synteny of COL15A1 and its role in skeletal muscle cells was found in human, mouse and zebrafish (Bretaud et al. 2011). However, studies on interspecies species synteny analysis using collagen gene family are also limited. The identification of these conserved syntenic chromosomal blocks will be critical in understanding the structural and functional significance of collagen gene family.

In this study, we aimed to investigate the sequence, structural, functional and evolutionary significance of 44 collagen gene members. Therefore, firstly, occurrence of each collagen gene member was screened in different species ranging from fungi to humans. Secondly, gene family was divided into different groups based on their domain organization and phylogenetic analysis was performed to evaluate the duplication history of each collagen group. Thirdly, evolutionary conservation of collagen specific syntenic blocks in vertebrate lineage was identified. This study provided unprecedented insights into different evolutionary mechanisms responsible for structural and functional diversity of collagens.

Methods

Data collection

The protein sequence data of 44 collagen gene members was collected from primates, rodents, birds, fishes, chordates, nematodes, arthropods, choanoflagellates and fungi. The list species selected for the analyses include Human (*Homo sapiens*), Orangutan (*Pongo abelii*), Chimpanzee (*Pan troglodytes*),

Mouse (*Mus musculus*), Rat (*Rattus norvegicus*), Rabbit (*Oryctolagus cuniculus*), Chicken (*Gallus gallus*), Zebra Finch (*Taeniopygia guttata*), Turkey (*Meleagris gallopavo*), Zebrafish (*Danio rerio*), Tetraodon (*Tetraodon nigroviridis*), Fugu (*Takifugu rubripes*), *Ciona intestinalis*, *Caenorhabditis elegans* (*Caenorhabditis elegans*), Fruitfly (*Drosophila melanogaster*), *Monosiga brevicollis* and Yeast (*Saccharomyces cerevisiae*).

Identifying orthologs using BLASTP

The local alignment tool, BLASTP (Altschul et al. 1990) present at Ensembl genome browser (<http://ensembl.org/index.html>) was used to identify putative orthologs against human protein sequences. The orthologous sequences which were not available at Ensembl genome browser were searched in NCBI protein database (<https://www.ncbi.nlm.nih.gov>). The criteria for selecting an orthologs were high sequence identity and coverage as discussed in our previous study (Asrar et al. 2013). Additionally, for simpler organisms if the identity and coverage was low, specific collagen domains were also searched.

Sequence and phylogenetic analysis

Multiple sequence alignment of amino-acid sequences was done using CLUSTAL W (Thompson et al. 1994) with default parameters. Molecular Evolutionary Genetic Analysis (MEGA) v 6.06 tool was used for evolutionary analyses. Neighbor-Joining (NJ) method was used for constructing phylogenetic trees (Saitou and Nei 1987). Poisson correction method for amino-acid substitution implemented in MEGA was selected. Complete deletion option was selected to treat gaps in the alignments. The sequences which disrupted the whole phylogenetic trees were excluded from the analysis. Bootstrapping of the trees was done with 1000 replicates. Furthermore, to validate the phylogenetic trees generated by NJ, Maximum Likelihood method using Whelan and Goldman (WAG) model was used (Whelan and Goldman 2001).

Synteny analysis

Ensembl synteny viewer was used with default window size for Human, Mouse, Dog, Lizard and Zebrafish species. Collagen genes which were in proximity on similar chromosomes from fish to humans were grouped into one conserved block. Synteny plots were generated using CIRCOS tool.

Results

The human genome screening for collagen gene family identified 44 gene members (Fig. 1). These gene members were divided into different groups. First group was of

fibril collagens, containing one major triple helix domain (Exposito et al. 2010). The genome screening identified seven gene members out of which COL1A, COL5A and COL11A had more than one paralogs (Fig. 1). Second group was fibril-associated collagens with interrupted triple helices (FACIT). Previously, FACITs were linked to the surface of fibril collagens (Shaw and Olsen 1991). The genome screening identified eight gene members out of which COL9A had more than one paralogs. network-forming collagens showed highly variable sequence lengths and triple helix domains (Sundaramoorthy et al. 2002). The genome screening identified 3 gene members out of which COL4A had six and COL8A had two paralogs (Fig. 1). Fourth group included COL6A, COL7A, COL26A and COL28A genes (Fig. 1). In this group triple helix domain is merged between several von Willebrand factor A domains (Veit et al. 2006). In this group, six paralogs were identified in COL6A gene. The last two groups included membrane collagens and multiplexins which had four (COL13A, COL17A, COL23A, COL25A) and two (COL15A, COL18A) gene members respectively (Fig. 1).

To find the occurrence of each collagen gene in vertebrates and invertebrate species, the protein sequences of all 44 genes were downloaded from Ensembl. The orthologous genes were screened using BLASTP in three primates, three rodents, three birds and three fish species. Furthermore, orthologous genes were screened in invertebrate species including chordates, nematodes, arthropods, choanoflagellates and fungi (Fig. 1). According to the results, primates and rodents had at least 40 collagen orthologs, except orangutan (Fig. 1). Interestingly, primates (except humans) showed less collagen genes compared to rodents. Birds and fish showed less than 40 collagen orthologs (Fig. 1). The invertebrate species including *Ciona intestinilis* (chordate) showed 11, *Caenorhabditis elegans* (nematode) showed 6, *Drosophila melanogaster* (arthropod) showed 3 and *Monosiga brevicollis* (choanoflagellate) showed 5 orthologous genes (Fig. 1). Consistent

with previous observations, no collagen gene member was found in *Saccharomyces cerevisiae* (Fungi).

Of note, based on domain information, origin of collagen genes showed different time points. FACITs was the most recent collagen group originating at root of vertebrate lineage (i.e., fish) (Fig. 1). The network-forming collagens, Membrane collagens and Multiplexins originated before protostome–deuterostome split (Fig. 1). The most ancient collagen groups were Fibril-forming collagens and Collagens VI, VII, XXVI, and XXVIII (Fig. 1). The human orthologs of these two groups were found in choanoflagellate (*Monosiga brevicollis*), most closely related unicellular organisms with animals. No collagen gene in *Saccharomyces cerevisiae* suggested that collagen was first originated at the time of choanoflagellate origin.

Conservation of 44 collagen genes

The protein sequence information also revealed many putative events like gene deletion, horizontal gene transfer, gene duplication as shown in Fig. 1. Interestingly, in Fibril-forming collagens, most of the gene loss events were found in bird species. For instance, gene loss events in COL1A1 (2 out of 3 bird species), COL3A1 (2 out of 3 bird species), COL5A3 (3 out of 3 bird species) and COL11A2 (3 out of 3 bird species) were identified (Fig. 1). Only 3 out of 11 collagen genes in this group showed conservation with no gene loss events (Fig. 1).

The FACITs were the most recently evolved group. Particularly, COL20A1 was the most recently evolved collagen gene out of all 44 human collagen genes (i.e., originating after fish–tetrapod split). In addition, only five gene loss events were observed in this group. Overall, 6 out of 10 genes in this group showed conservation with no gene loss events (Fig. 1). The network-forming collagens were evolved before protostome–deuterostome split. COL4A6 was the first network-forming collagen gene found in *D. melanogaster*. Of note, network-forming collagens were the most conserved out six collagen groups. Overall, 8 out of

Fig. 1 Gene screening of 44 collagen (COL) genes in 12 vertebrate and 5 invertebrate species. Red represents the gene is present while blue represents the gene is absent

9 genes in this group were totally conserved with no gene loss events (Fig. 1). Interestingly, the pattern of gene losses in Fibril-forming collagens, FACITs and network-forming collagens suggested that the gene which was lost once in animal phylogeny was prone to more gene losses compared to other genes. Collagens VI, VII, XXVI, and XXVIII were also found in *Monosiga brevicollis*, the most ancient organism of animal phylogeny containing collagen gene. Only 2 out of 8 genes in this group were totally conserved with no gene loss events (Fig. 1). Lastly, both Membrane collagens and Multiplexins were evolved before protostome–deuterostome split. Only 2 gene loss events were observed in Membrane collagens while only 1 gene loss event was observed in Multiplexins.

The occurrence of collagen genes based on protein sequence information suggested no specific pattern of collagen evolution. Several fish-specific duplication events were also identified (Electronic Supplementary Material ESM_1 and ESM_2).

Phylogenetic analysis and topology comparison approach

Here, we performed a comprehensive phylogenetic analysis of all 44 collagen genes using at least one representative from primates, rodents, birds, fish, chordates, nematodes, arthropods and choanoflagellates. Initially, phylogenetic trees were generated using Neighbor-Joining (NJ) method (Fig. 2) and reliability was evaluated using bootstrap (1000 pseudoreplicates). Furthermore, the phylogenetic trees were validated using Maximum Likelihood (ML) method and reliability was also evaluated using bootstrap (1000 pseudoreplicates). The trees generated by NJ and ML methods reconciled well with each other (Fig. 2, S1 and S2).

According to the phylogenetic trees, Fibril-forming collagen group showed 10, FACIT group showed 9, network-forming collagen group showed 8, Collagens VI, VII, XXVI, and XXVIII group showed 7, Membrane collagen group showed 3 and Multiplexin group showed 1 gene duplications (Electronic Supplementary Material ESM_1 and ESM_2). Next, tree topologies of all phylogenetic trees were constructed and compared with the topology of widely accepted whole genome duplication (WGD) topology [i.e., (AB) (CD)] (Fig. 3). The tree topologies of Fibril-forming collagen, network-forming collagen, Collagens VI, VII, XXVI, and XXVIII and Membrane collagen groups were not concordant with WGD topology (Fig. 3). Interestingly, FACIT group had no orthologous gene in invertebrates and showed excessive number of 8 gene duplications at fish time point suggesting fish-specific gene duplications. The Multiplexin group did not qualify for the analysis since it contained fewer number of collagen genes.

The phylogenetic history of Fibril collagens revealed 6 duplication events occurring before and 4 after invertebrate–vertebrate split (Fig. 3a). The phylogenetic history of FACITs showed all 8 duplications during vertebrate evolution (Fig. 3b). The phylogenetic history of network-forming collagens revealed that 3 duplication events occurred before and 5 occurred after invertebrate–vertebrate split (Fig. 3c). In Collagens VI, VII, XXVI, and XXVIII group all 7 duplications occurred before invertebrate–vertebrate split (Fig. 3d). Lastly, phylogenetic history of Membrane collagens revealed 1 duplication before and 2 after invertebrate–vertebrate split (Fig. 3e). These results showed asymmetrical pattern of gene duplications suggesting diversity of at least 4 out of 5 collagen groups did not arise exclusively at the root of vertebrate lineage (i.e., fish–tetrapod split).

Synteny analysis of 44 collagen genes on Human, Mouse, Dog, Lizard and Zebrafish

Next, we performed interspecies comparative analysis of chromosomal segments including 15 upstream and 15 downstream of collagen genes using the default window size of Ensembl synteny viewer for Human, Mouse, Dog, Lizard and Zebrafish species, to investigate the pattern of syntenic conservation of collagen genes in vertebrate lineage (Table 1).

Of note, we identified 27 out of 44 collagen genes showing synteny with at least one or two collagen paralogs. A total of 12 conserved syntenic blocks were identified (Table 1; Fig. 4). 3 syntenic blocks contained 3 collagen genes including COL6A1, COL6A2, COL18A1 at human chromosome 21q22.3, COL4A3, COL4A4, COL6A3 at human chromosome 2q36.3–q37.3 and COL9A1, COL19A1, COL21A1 at human chromosome 6p12.1–6q13. Rest of the 9 blocks contained 2 collagen genes (Table 1; Fig. 4 and ESM_3).

Interestingly, in comparison with other vertebrate species, zebrafish genome showed intense fish-specific chromosomal rearrangements and duplications. In addition, many paralogs of collagen genes were present in zebrafish genomes which may have lost during fish–tetrapods split (Electronic Supplementary Material ESM_1 and ESM_2). The identification of these conserved syntenic blocks of collagen genes might reflect some structural and functional relationships between each other.

Discussion

Complex genomic events including duplications, deletions, chromosomal rearrangements and breakage tend to obscure the exact mechanism of genome shaping of present day vertebrate species (Abbasi 2015). However, comparative

Fig. 2 Phylogenetic trees of FACITs using NJ and ML methods. The phylogenetic history of FACITs generated by NJ and ML methods reconcile well with each other. **a** Poisson correction method was used for NJ tree generation. Scale bar value was 0.2. **b** Whelan And Gold-

man model was used for ML tree generation. Scale bar value was 0.5. All positions containing gaps and missing data were eliminated. Only bootstrap values > 95% were included. The details about the methodology are mentioned in electronic supplementary material

genomics has provided unprecedented insights into the different evolutionary mechanisms responsible for this diversity (Alföldi and Lindblad-Toh 2013). In this study, comparative genomic analysis of 44 collagen genes using 17 invertebrate and vertebrate species revealed many important insights into the functional conservation, evolutionary conservation and diversification of collagen gene family. According to the analysis, network-forming collagens was the most conserved group in animal phylogeny. Fibril-forming collagens and Collagens VI, VII, XXVI, and XXVIII were the most ancient groups in animal phylogeny. Of note, the emergence of FACITs at the root of vertebrate lineage suggested vertebrate specific roles of these collagens. Phylogenetic analysis revealed different duplication patterns of each collagen

group including small scale gene duplications, rearrangements and genome duplications. Of note, synteny analysis using vertebrate species revealed 12 conserved syntenic blocks containing 27 collagen gene members suggesting functional conservation of these blocks.

The emergence of new genes, either by de novo or by duplications, plays key role in the structural, functional and phenotypic evolution of a genome (Conrad and Antonarakis 2007). On the other hand, gene loss events due to genetic variations are also contributing to genomic and phenotypic diversity (Albalat and Cañestro 2016). In this study, a continuous number of gene losses and duplications were observed. Interestingly, gene which was lost once in animal phylogeny was prone to more gene losses compared to other

Fig. 3 The NJ tree topologies of Fibril-forming collagens, FACITs, network-forming collagens, Collagens VI, VII, XXVI, and XXVIII and Membrane collagens. The tree topologies showed multiple duplication patterns

Table 1 12 Syntenic blocks containing at least two collagen gene members

Sr#	Gene groups	Chromosomal arm
BLOCK1	COL24A1, COL11A1	1p21.1–p22.3
BLOCK2	COL8A2, COL16A1	1p34.3–35.2
BLOCK3	COL3A1, COL5A2	2q32.2
BLOCK4	COL4A3, COL4A4, COL6A3	2q36.3–q37.3
BLOCK5	COL6A5, COL6A6	3q22.1
BLOCK6	COL9A1, COL19A1, COL21A1	6p12.1–6q13
BLOCK7	COL1A2, COL28A1	7p21.3–7q21.3
BLOCK8	COL14A1, COL22A1	8q24.12–3
BLOCK9	COL4A1, COL4A2	13q34
BLOCK10	COL9A3, COL20A1	20q13.33
BLOCK11	COL6A1, COL6A2, COL18A1	21q22.3
BLOCK12	COL4A5, COL4A6	Xq22.3

conserved genes. However, the proportion of gene duplication was comparatively very high compared to gene loss events, suggesting a need of gene dosage compared to gene dispensability (Albalat and Cañestro 2016).

The huge difference of gene number between invertebrates (i.e., 5–11 genes) and vertebrates (i.e., 37–44 genes) suggested that collagen gene diversity took place due to WGD events which might have occurred at the root of vertebrate lineage (Kasahara 2007). However, tree topologies generated from phylogenetic trees revealed multiple duplication patterns. According to the results, majority of gene

duplications in Fibril-forming collagen and Collagens VI, VII, XXVI, and XXVIII groups occurred in invertebrate lineage, suggesting variable duplication time points before invertebrate–vertebrate split. Conversely, majority of gene duplications in FACITs (i.e., all 8 duplications) and network-forming collagens (i.e., 5 out of 8 duplications) occurred within vertebrate lineage, suggesting a major duplication event before fish–tetrapods split. However, apart from many duplications at fish time point, tree topologies of FACITs and network-forming collagens were not consistent with 2R topology. Therefore, dissimilar evolutionary patterns observed in these six collagen groups suggested multiple duplication mechanisms (including small/large scale duplications and WGD) occurring throughout animal evolution (Hakes et al. 2007).

The conservation of a gene implies a conserved biological function vital for an organism's stability (Bejerano et al. 2004). Phylogenetic analysis is, however, suitable method to determine the orthologs/paralogs present in multiple genomes but is not sufficient enough to determine functional relevance between these orthologs/paralogs (Christoffels et al. 2004; Montpetit et al. 2003). So, the linear order of 30 genes surrounding each collagen gene (15 upstream and 15 downstream) was also evaluated and 12 conserved blocks containing more than one collagen gene were identified in vertebrate genomes (ESM_3). Interestingly, collagen genes present in one single conserved block were found to be associated with different diseases (Conrad and Antonarakis 2007; Egeblad et al. 2010). For instance, genetic variations

Fig. 4 Circos plots of syntenic BLOCK4 and BLOCK11. BLOCK4 contains COL4A3, COL4A4 and COL6A3 genes while BLOCK11 contains COL6A1, COL6A2 and COL18A1 genes. Chromosomal representations for human, mouse, dog and zebrafish are hs, mm, cf and dr respectively

in COL4A1/COL4A2 of block9 were found to be associated with cerebral small vessel disease and hemorrhagic stroke (Rannikmäe et al. 2015). Deletions in COL4A5/COL4A6 of block12 were found to be associated is leiomyoma and nephropathy (Oohashi et al. 2011). RNA dysregulation of COL6A5/COL6A6 of block5 was found to be associated with myopathy (Sabatelli et al. 2011). Interestingly, single nucleotide polymorphisms in block11 containing three genes (i.e., COL6A1/COL6A2/COL18A1) were also found to be associated with OPLL, Ossification of the Posterior Longitudinal Ligament (Tanaka et al. 2003). Therefore, we suggest that the association of these 12 conserved blocks with different hereditary diseases is worth investigation.

In addition, genetic alterations (including DNA copy number variations) are now forming a major class of cancer-related genomic events (Conrad and Antonarakis 2007; Shahid et al. 2016a). For instance, segmental duplication of chromosomal region containing COL18A1 gene of block11 and somatic mutations in COL6A5/COL6A6 of block5 were found to be associated with breast cancer progression (Dorman et al. 2014; Hwang et al. 2010). The expression levels of COL1A2 (block7), COL3A1 (block3), COL6A3 (block4) were found to be associated with the prognosis of glioma patients (Shahid et al. 2016b). Furthermore, dysregulation of COL11A1 (block1) and COL5A2 (block3) were found

to be associated with chemotherapy response of ovarian cancer patients (Matondo et al. 2017). Therefore, based on our analysis, we suggest that the combinatorial effect of the genes present in one conserved block can be associated with different hereditary and somatic diseases including cancer. It will be, therefore, interesting to identify segmental duplications/deletions in these 12 conserved blocks and evaluate their association with the different clinicopathological outcome of cancer patients. Above all, this is the first study which covered sequence, structural and evolutionary aspects of 44 collagen genes.

Acknowledgements Special thanks to the teams behind UCSC, NCBI and Ensembl genome browsers, for making all the data publicly available for the analyses.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest.

References

Abbasi AA (2008) Are we degenerate tetraploids? More genomes, new facts. *Biol Direct* 3:50

- Abbasi AA (2015) Diversification of four human HOX gene clusters by step-wise evolution rather than ancient whole-genome duplications. *Dev Genes Evol* 225:353–357
- Abedin M, King N (2010) Diverse evolutionary paths to cell adhesion. *Trends Cell Biol* 20:734–742
- Albalat R, Cañestro C (2016) Evolution by gene loss. *Nat Rev Genet* 17:379–391
- Alföldi J, Lindblad-Toh K (2013) Comparative genomics as a tool to understand evolution and disease. *Genome Res* 23:1063–1068
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
- Asrar Z, Haq F, Abbasi AA (2013) Fourfold paralogy regions on human HOX-bearing chromosomes: role of ancient segmental duplications in the evolution of vertebrate genome. *Mol Phylogenet Evol* 66:737–747
- Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. *Science* 304:1321–1325
- Boot-Handford RP, Tuckwell DS (2003) Fibrillar collagen: the key to vertebrate evolution? A tale of molecular incest. *BioEssays News Rev Mol Cell Dev Biol* 25:142–151
- Bretnaud S, Pagnon-Minot A, Guillon E, Ruggiero F, Le Guellec D (2011) Characterization of spatial and temporal expression pattern of Col15a1b during zebrafish development. *Gene Expr Patterns* 11:129–134
- Brodsky B, Ramshaw JAM (1997) The collagen triple-helix structure. *Matrix Biol* 15:545–554
- Byron A, Humphries JD, Humphries MJ (2013) Defining the extracellular matrix using proteomics. *Int J Exp Pathol* 94:75–92
- Christoffels A, Koh EGL, Chia J-M, Brenner S, Aparicio S, Venkatesh B (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol* 21:1146–1151
- Chu M-L (2011) Structural proteins: genes for collagen. In: eLS. Wiley, Chichester. <https://doi.org/10.1002/9780470015902.a0005023.pub2>
- Conrad B, Antonarakis SE (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genom Hum Genet* 8:17–35
- Dennis MY, Eichler EE (2016) Human adaptation and evolution by segmental duplication. *Curr Opin Genet Dev* 41:44–52
- Dorman SN, Viner C, Rogan PK (2014) Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. *Sci Rep* 4:7063
- Egeblad M, Rasch MG, Weaver VM (2010) Dynamic interplay between the collagen scaffold and tumor evolution. *Curr Opin Cell Biol* 22:697–706
- Ehrlich J, Sankoff D, Nadeau JH (1997) Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics* 147:289–296
- Exposito J-Y, Valcourt U, Cluzel C, Lethias C (2010) The fibrillar collagen family. *Int J Mol Sci* 11:407–426
- Gelse K, Pöschl E, Aigner T (2003) Collagens—structure, function, and biosynthesis. *Adv Drug Deliv Rev* 55:1531–1546
- Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL (2007) All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol* 8:R209
- Huang X, Godfrey TE, Gooding WE, McCarty KS, Gollin SM (2006) Comprehensive genome and transcriptome analysis of the 11q13 amplicon in human oral cancer and synteny to the 7F5 amplicon in murine oral carcinoma. *Genes Chromosomes Cancer* 45:1058–1069
- Hwang K-T, Chung JK, Jung IM, Heo SC, Ahn YJ, Ahn HS, Chang MS, Kim J-A, Han W, Noh D-Y (2010) COL18A1 as the candidate gene for the prognostic marker of breast cancer according to the analysis of the DNA copy number variation by array CGH. *J Breast Cancer* 13:37
- Kasahara M (2007) The 2R hypothesis: an update. *Curr Opin Immunol* 19:547–552
- Kemkemmer C, Kohn M, Cooper DN, Froenicke L, Högel J, Hameister H, Kehrer-Sawatzki H (2009) Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. *BMC Evol Biol* 9:84
- King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I et al (2008) The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451:783–788
- Leitinger B (2011) Transmembrane collagen receptors. *Annu Rev Cell Dev Biol* 27:265–290
- Matondo A, Jo YH, Shahid M, Choi TG, Nguyen MN, Nguyen NNY, Akter S, Kang I, Ha J, Maeng CH et al (2017) The Prognostic 97 Chemoresponse Gene Signature in Ovarian Cancer., The Prognostic 97 Chemoresponse Gene Signature in Ovarian Cancer. *Sci Rep Sci Rep* 7:7, 9689–9689
- Meyer A, Van de Peer Y (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays* 27:937–945
- Montpetit A, Wilson MD, Chevrette M, Koop BF, Sinnett D (2003) Analysis of the conservation of synteny between Fugu and human chromosome 12. *BMC Genom* 4:30
- Morvan-Dubois G, Le Guellec D, Garrone R, Zylberberg L, Bonnaud L (2003) Phylogenetic analysis of vertebrate fibrillar collagen locates the position of zebrafish alpha3(I) and suggests an evolutionary link between collagen alpha chains and hox clusters. *J Mol Evol* 57:501–514
- Nagase H, Fields GB (1996) Human matrix metalloproteinase specificity studies using collagen sequence-based synthetic peptides. *Pept Sci* 40:399–416
- Oohashi T, Naito I, Ueki Y, Yamatsuji T, Permpoon R, Tanaka N, Naomoto Y, Ninomiya Y (2011) Clonal overgrowth of esophageal smooth muscle cells in diffuse leiomyomatosis-Alport syndrome caused by partial deletion in COL4A5 and COL4A6 genes. *Matrix Biol J Int Soc Matrix Biol* 30:3–8
- Rannikmäe K, Davies G, Thomson PA, Bevan S, Devan WJ, Falcone GJ, Traylor M, Anderson CD, Battey TWK, Radmanesh F et al (2015) Common variation in COL4A1/COL4A2 is associated with sporadic cerebral small vessel disease. *Neurology* 84:918–926
- Ricard-Blum S (2011) The collagen family. *Cold Spring Harb Perspect Biol* 3:a004978
- Rychel AL, Smith SE, Shimamoto HT, Swalla BJ (2006) Evolution and development of the chordates: collagen and pharyngeal cartilage. *Mol Biol Evol* 23:541–549
- Sabatelli P, Gara SK, Grumati P, Urciuolo A, Gualandi F, Curci R, Squarzone S, Zamparelli A, Martoni E, Merlini L et al (2011) Expression of the collagen VI $\alpha 5$ and $\alpha 6$ chains in normal human skin and in skin of patients with collagen VI-related myopathies. *J Invest Dermatol* 131:99–107
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425
- Shahid M, Choi TG, Nguyen MN, Matondo A, Jo YH, Yoo JY, Nguyen NNY, Yun HR, Kim J, Akter S et al (2016a) An 8-gene signature for prediction of prognosis and chemoresponse in non-small cell lung cancer. *Oncotarget* 7:86561–86572
- Shahid M, Cho KM, Nguyen MN, Choi TG, Jo YH, Aryal SN, Yoo JY, Yun HR, Lee JW, Eun YG et al (2016b) Prognostic value and their clinical implication of 89-gene signature in glioma. *Oncotarget* 7:51237–51250
- Shaw LM, Olsen BR (1991) FACIT collagens: diverse molecular bridges in extracellular matrices. *Trends Biochem Sci* 16:191–194

- Staatz WD, Fok KF, Zutter MM, Adams SP, Rodriguez BA, Santoro SA (1991) Identification of a tetrapeptide recognition sequence for the alpha 2 beta 1 integrin in collagen. *J Biol Chem* 266:7363–7367
- Sundaramoorthy M, Meiyappan M, Todd P, Hudson BG (2002) Crystal structure of NC1 domains structural basis for type IV collagen assembly in basement membranes. *J Biol Chem* 277:31142–31153
- Tanaka T, Ikari K, Furushima K, Okada A, Tanaka H, Furukawa K-I, Yoshida K, Ikeda T, Ikegawa S, Hunt SC et al (2003) Genomewide linkage and linkage disequilibrium analyses identify COL6A1, on Chromosome 21, as the locus for ossification of the posterior longitudinal ligament of the spine. *Am J Hum Genet* 73:812–822
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680
- Veit G, Kobbe B, Keene DR, Paulsson M, Koch M, Wagener R (2006) Collagen XXVIII, a novel von willebrand factor a domain-containing protein with many imperfections in the collagenous domain. *J Biol Chem* 281:3494–3504
- Wada H, Okuyama M, Satoh N, Zhang S (2006) Molecular evolution of fibrillar collagen in chordates, with implications for the evolution of vertebrate skeletons and chordate phylogeny. *Evol Dev* 8:370–377
- Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18:691–699