The COL1A1 gene codes for collagen alpha (1) chain protein. The structural protein is associated with RefSeq accession # NP\_000079.2. Using this accession number a BLASTP search was performed with a words size of 6, an expect threshold of 10, the BLOSUM62 substitution matrix, gap extension cost of 1, gap existence cost of 11. Seven in-species protein hits resulted from this query [(AFD28984.1), (AAB94054.3), (AAH36531.1), (BAD92834.1), (CAA67261.1), (P02452.5), and (CAA98968.1)]. These hits are better described as variants with single point mutations than isoforms or homologues. Of the seven in-species hits most had 100% identity, several had one residue substitutions, and one had two residue substitutions. The seven in-species hits couldn’t possibly be pseudogenes because they are derived from a sequenced protein. Pseudogenes do no code for a protein. The BLAST hits from this query were highly conserved, the lowest percent identity hit was 85% with an E-value of 0.

A second BLAST was performed with reduced sensitivity by increasing the expect threshold to 12 and the maximum hits to 500. This query produced 24 in-species hits with several possible homologues. (EAW94630.1) is a COL1A1 isoform with perfect contiguous alignment to half of the amino acid sequence of NP\_000079.2. However, this hit is more likely the result of alternative splicing than a duplication event. Many in-species hits for collagen alpha-1(II) chain proteins are possible candidates of duplication event but this protein is ubiquitous among vertebrates and maximum parsimony suggests they are not paralogs. It is unlikely that many species experienced a similar duplication event.

Based on percent identity and e-value, the BLAST produced several possible orthologues. A possible vertebrate ortholog due to an e-value of 0.0 and 97% identity, is the bovine COL1A1 gene that codes for a protein connected to accession number P02453. Another possible vertebrate orthologue is the canine COL1A1 gene connected to accession number [NM\_001003090.1](http://www.ncbi.nlm.nih.gov/nuccore/50978773) with an e-value of 0.0.

There seem to be well over 100 possible orthologues with e-values of 0 and greater than 70% identity. The two BLASTs did not produce any possible plant, insect, or prokaryote homolog hits. The HomoloGene hits for the COL1A1 protein were significantly fewer in quantity than the BLAST hits. This is due to differences in the tools’ algorithms. Homologene is a more stringent search due to its global comparison rather than local alignment. HomoloGene compares sequences at the domain level, rather than the sequence level. This accounts for the fewer returned hits from the HomoloGene query.