**Github and project organization**

I am reading that your project seeks to address two major questions/aims. I would make a separate directory for each analysis and this will keep you on track and ensures that you have proper files to work with. Then for each directory, you create a README.md file which describes the work flow of your analysis.

Also, all primary sequence files should be in fasta formats except for gene lists that can be saved as a list. This will ease sequence alignment.

I see your README.md is empty but it’s meant to describe the workflow with the codes you used for every stage of the analysis. I guess you might be having these somewhere but not to forget them please document as soon as the returned result makes sense.

**Methods**

**Testing for Homology**

While downloading the sequences, did you BLAST to look for Homology to the COL1A1 and COL1A2 sequences or isolated the sequences pretty randomly? If by BLAST what were the threshold? It would also be meaningful to specify if you downloaded the whole gene sequences or only the protein coding nucleotide sequences only (CDS)? This would help to eliminate noise.

I would only download the protein nucleotide coding sequence (ORF) and or trim the files to make them have equal lengths. *Be sure to include an appropriate outgroup where you will root the tree.*

**Building a phylogenetic relationship**

I am reading that you used Maximum likelihood to build phylogenetic relationships to your genes of interest however I am curious of the model of nucleotide substitution you used and why? There are several models such as GTRGAMMA, GTRCAT, and each has unique strengths to average over the data and given hypothesis. I would revisit Computer lab 4 <https://isu-molphyl.github.io/EEOB563-Spring2018/computer_labs/lab4> and Assignment 5 <https://isu-molphyl.github.io/EEOB563-Spring2018/assignments/assignment5.pdf> for clarity. It will guide you how best to pick the models and how they relate to each other.

When you mention of bootstrapping, it is very meaningful to specify the number of times/replicates and or iterations (all these mean the same thing) since these will be returned on branch lengths and you have to tell the reader what these numbers do mean.

I copied this for you just to be clear of what I am talking about

**raxmlHPC -m GTRGAMMA -p 12345 -b 12345 -# 100 -s dna.phy -n T14**We need to tell RAxML that we want to do bootstrapping by providing a bootstrap random number seed via -b 12345 and the number of bootstrap replicates we want to compute via -# 100. Note that, RAxML also allows for automatically determining a sufficient number of bootstrap replicates, in this case you would replace -# 100 by one of the bootstrap convergence criteria -# autoFC, -# autoMRE, -# autoMR, -# autoMRE\_IGN.

**ALSO,** you can build strict, majority and majority rule extended and see how they relate to one another. You can present the one that makes sense to you and put the rest in supplemental data.

**Comparing multiple COL1A1 and COL1A2 sequences**

I see you used percent identity matrix to infer homology but I think it would be more than I deal to construct phylogeny by any method such as Maximum likelihood or Neighbor joining to see if the percent identity is consistent with the species phylogeny. Then you would make a solid conclusion out of this.

**Tree visualization**

You can view your trees via Fig tree (<http://tree.bio.ed.ac.uk/software/figtree/>) , Dendroscope (<http://danielhuson.github.io/dendroscope3>) or via online newick tree viewer (<https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwjZ-YnzisDaAhXH5oMKHXUACgUQFggnMAA&url=http%3A%2F%2Fetetoolkit.org%2Ftreeview%2F&usg=AOvVaw3doI0PfkFMwmBsxr8Uk2CV>)

Then your trees will not only look neat but will be easy to view by the reader.

**In General**,

I see your project really have greater impacts. You need to fix one or two things and you will be good to go.