Using phylogeny to identify orthologs

Background:
I want to run an experiment to measure gene (transcript) expression in the frog *Eleutherodactylus coqui*. To do so, I need to correctly identify orthologs in *E. coqui*, and I do this by making a phylogenetic tree. It is important to establish orthology (identify the genes correctly) before spending more time and resources on this experiment. Most importantly, moving forward without knowing what the gene is could lead to me to make incorrect evolutionary conclusions. *E. coqui* has no genome yet.

E. coqui differs from most other frogs- it has no tadpole stage, and hatches from a terrestrial egg as a miniature adult. Although about a third of all frogs develop without a tadpole, we don't know a lot about how this developmental mode (called direct development) evolved. My research focuses on understanding the mechanistic basis of development in *E. coqui*, so that there is a basis for comparison to frogs that go through metamorphosis.

In order to do my experiment, I need a control gene. I will compare gene expression levels to this control gene. I chose the gene a-actinin4, an actin binding gene involved in cytoskeletal structure, because I thought it would be present in *E. coqui* and it should not change substantially in transcript expression. I cloned a fragment of what I think is a-actinin4 from *E. coqui*. But how do I know that the sequence that I hold in my hand is *actually* a-actinin4, and not another gene that has a similar sequence? I use bioinformatic tools, and make a phylogenetic tree to determine what the sequence I cloned actually is.

"Ecoqui_pred_actn4" means that this is the predicted sequence for actn4 in *E. coqui*. I used this to design primers, but what I cloned doesn't completely match this sequence. "Ecoqui_actn_clone" is the sequence I cloned.

1. Do a reciprocal blast search. A reciprocal blast search searches the National Center for BioInformatics Database (NCBI) for what gene looks the most similar, and is a "quick and dirty" way to know what your gene looks like.
    a. Navigate to: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins
    b. Copy the text following the carrot (>) of the "Ecoqui_actn_clone" and paste this sequence into "Enter query sequence".
    c. Click "Blast" at the bottom of the page.
    d. What are the top two hits, or matches?
    e. What do all the matches have in common? Are they mostly one gene?
2. Navigate to http://phylogeny.fr. Click phylogeny analysis on the top bar, and go down to "Advanced". Name your analysis, and click "Create workflow".
    a. On the next page, copy the entire contents of the file "a-actinin orthologs.fasta" into the data input text box, and click go. Enter your email address if you would like to return to this.

   b. Scroll down, and under the Gblocks: Curation section, click the boxes next to "allow gaps in final blocks" and "allow gap positions within the final blocks". Click submit.
   c. To be able to view this tree in other phylogenetic tree viewers, click the "Newick" hypertext next to "Export", just under the tree. You can come back later and play with the viewer in phylogeny.fr (it's quite good), but I have found it slow in the past two weeks. So we will use the tree viewer at the Interactive Tree of Life. See the wikipedia page for a full list of tree viewers with different capabillities! https://en.wikipedia.org/wiki/List_of_phylogenetic_tree_visualization_software
3. Navigate to: https://itol.embl.de
4. Scroll down to Annotate and click Upload a tree. Name your tree, and paste the contents of the file "actn orthologs tree newick" into the Tree text box.
   a. On the right hand side, play with different ways of visualizing the tree in the control panel. Eventually, toggle "ignore" next to Branch lengths.
   b. Look for three main groups in the tree- actn1, actn3, and actn4. These are paralogs: a-actinin genes that resulted from duplication a long time ago. Actn3 in alligator is more closely related to actn3 in sea turtles, than it is to actn1 in alligator.
5. **To which other sequences is the "Ecoqui_actn4_clone" sequence mostly closely related?**
6. **What does this suggest to you about the sequence I cloned?**
7. **Which two genes are most closely related to each other?**