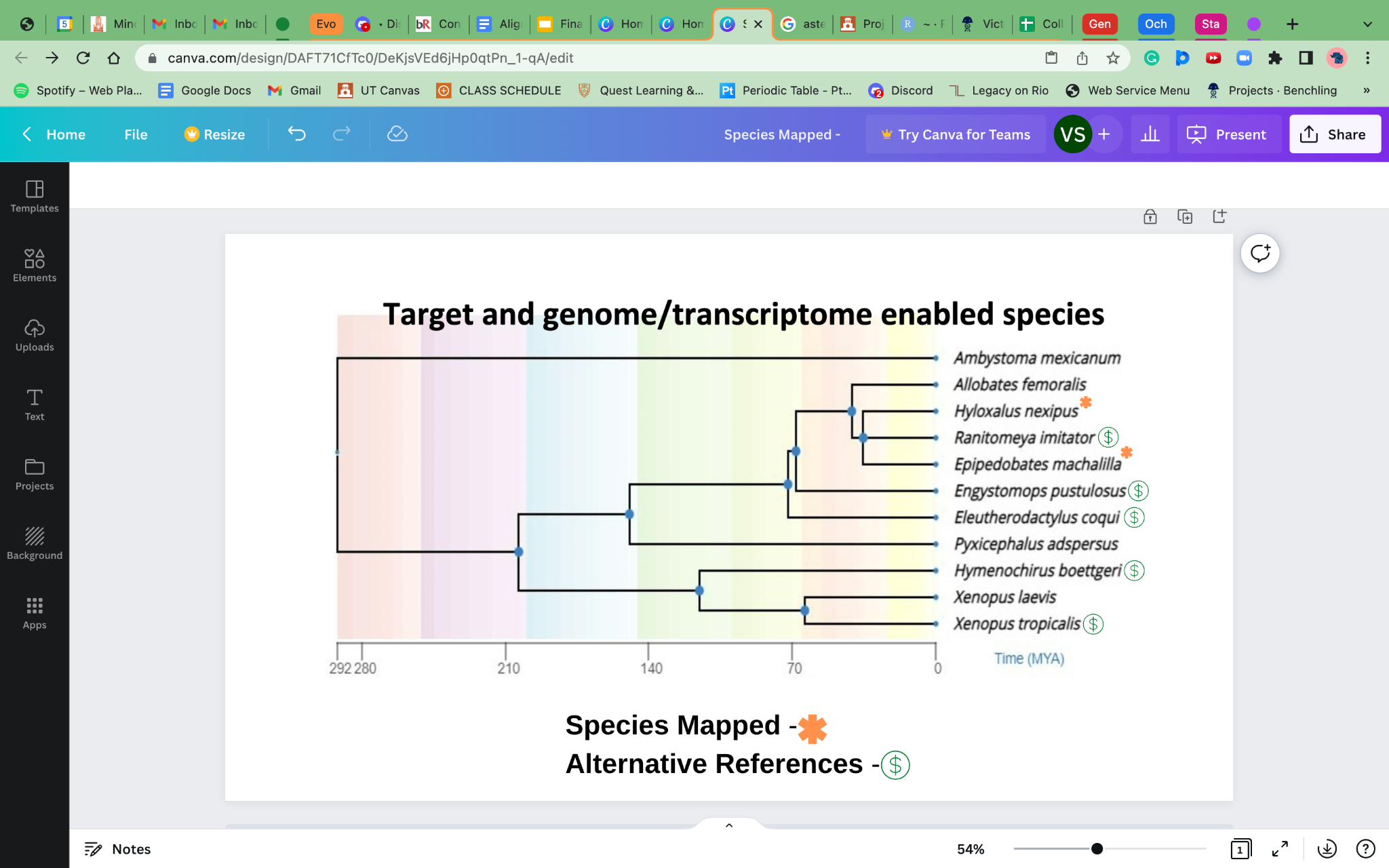
**Aligning multiple alternative reference transcriptomes using multiple alignment software for improvement in number of expressed genes relating to gastrulation**

**Overarching Project**

We are interested in the gastrulation process, the reorganization of cell layers during early development, across the *Dendrobatid* species of interest, *Epipedobates machalilla* and *Hyloxalus nexipus,* and *Xenopus tropicalis. Xenopus* species have a faster rate of gastrulation when compared to *Dendrobatids,* and the *Dendrobatid* species themselves differ in gastrulation as well. Identifying the key differentially expressed genes could help us understand how the two *Dendrobatid* species differ in gastrulation and provide insight into how *Xenopus* and *Dendrobatids* differ in their gastrulation processes.

**Project Thesis**

Currently, the samples of *Epipedobates machalilla* and *Hyloxalus nexipus* are aligned to the *Xenopus tropicalis* genome which results in a large percentage of reads not aligned to the reference genome as there are over 200 million years of separation between the species. We are benchmarking multiple reference genomes/transcriptomes that are more closely related to our Dendrobatid species, as well as a more complete *X. tropicalis* genome, to determine which reference species will yield accurately annotated and quantified genetic expression in genes related to gastrulation. We will determine if an alternative reference genome will map significantly more genes related to gastrulation when compared to *X. tropicalis* by looking into both mapping percentages and annotations of the genes that are mapped.

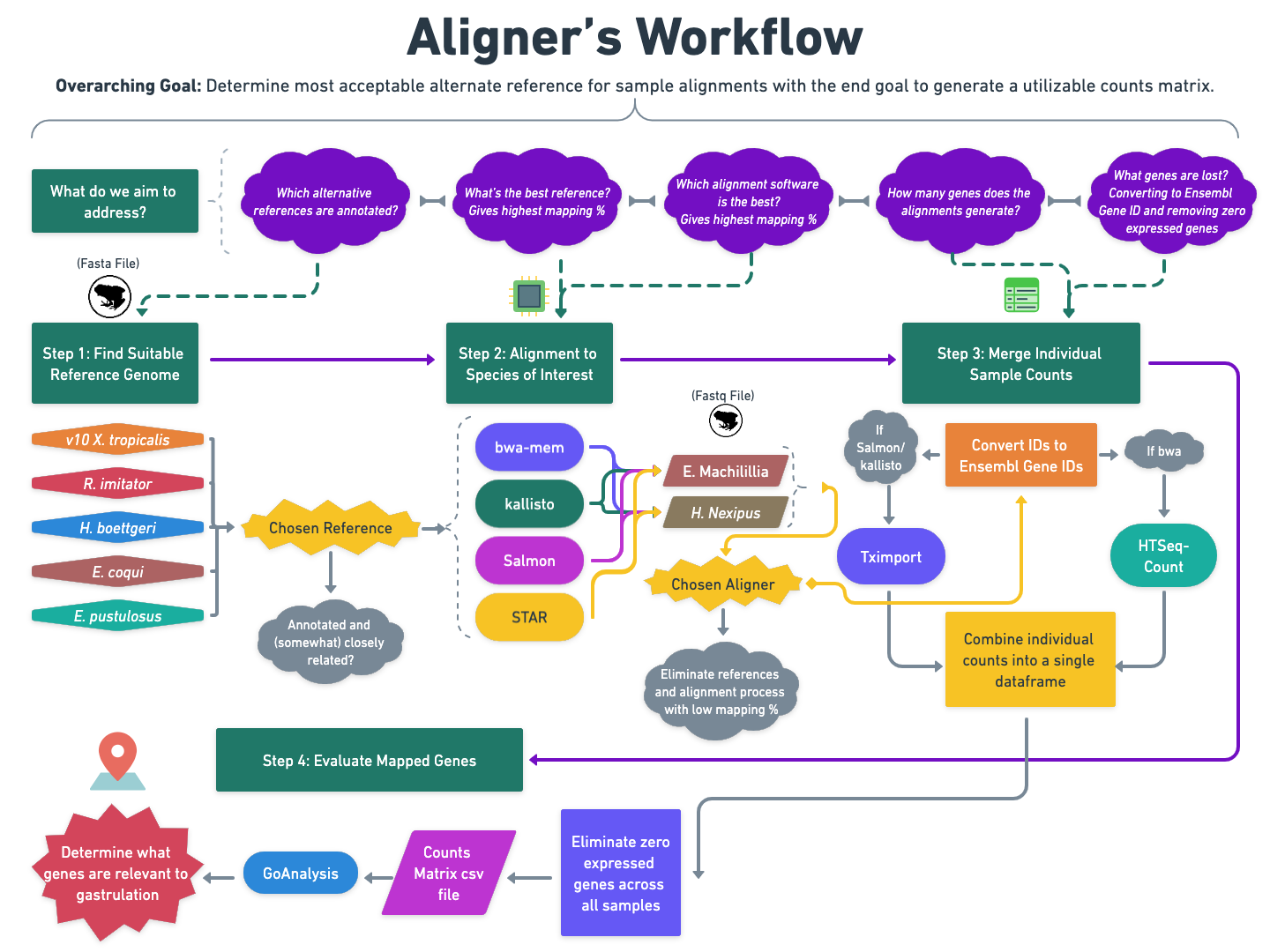


**Fig. 1** Phylogenetic tree of our Dendrobatid species of interest and the alternate reference species that we used.

**Methods**

*Ranitomeya imitator*, *Hymenochirus boettgeri*, *Eleutherodactylus coqui, Engystomops pustulosus, and Xenopus tropicalis* v.10 were chosen to be candidates for a usable alternate reference genome or transcriptome over the *X. tropicalis* reference that the Dendrobatidae were aligned to previously. These references were chosen because they are more closely related to *H. nexipus* and *E. machalilla* and have publicly available annotated transcriptomes. These newly chosen references are then used to map reads from *E. machalilla* and *H. nexipus* using bwa-mem (Heng, 2013), kallisto (Bray, 2016), and STAR (Dobin, 2013) alignment softwares (see **Fig. 2**). An alignment will be carried out for each reference and alignment software, which will then produce informative mapping percentages. These mapping percentages will be collected and compared among alignment approaches and reference species. The process that produces the highest mapping percentage and has the largest number of annotations will then be selected for downstream analysis. The remaining, alternate references and alignment softwares, will be eliminated from future analysis.

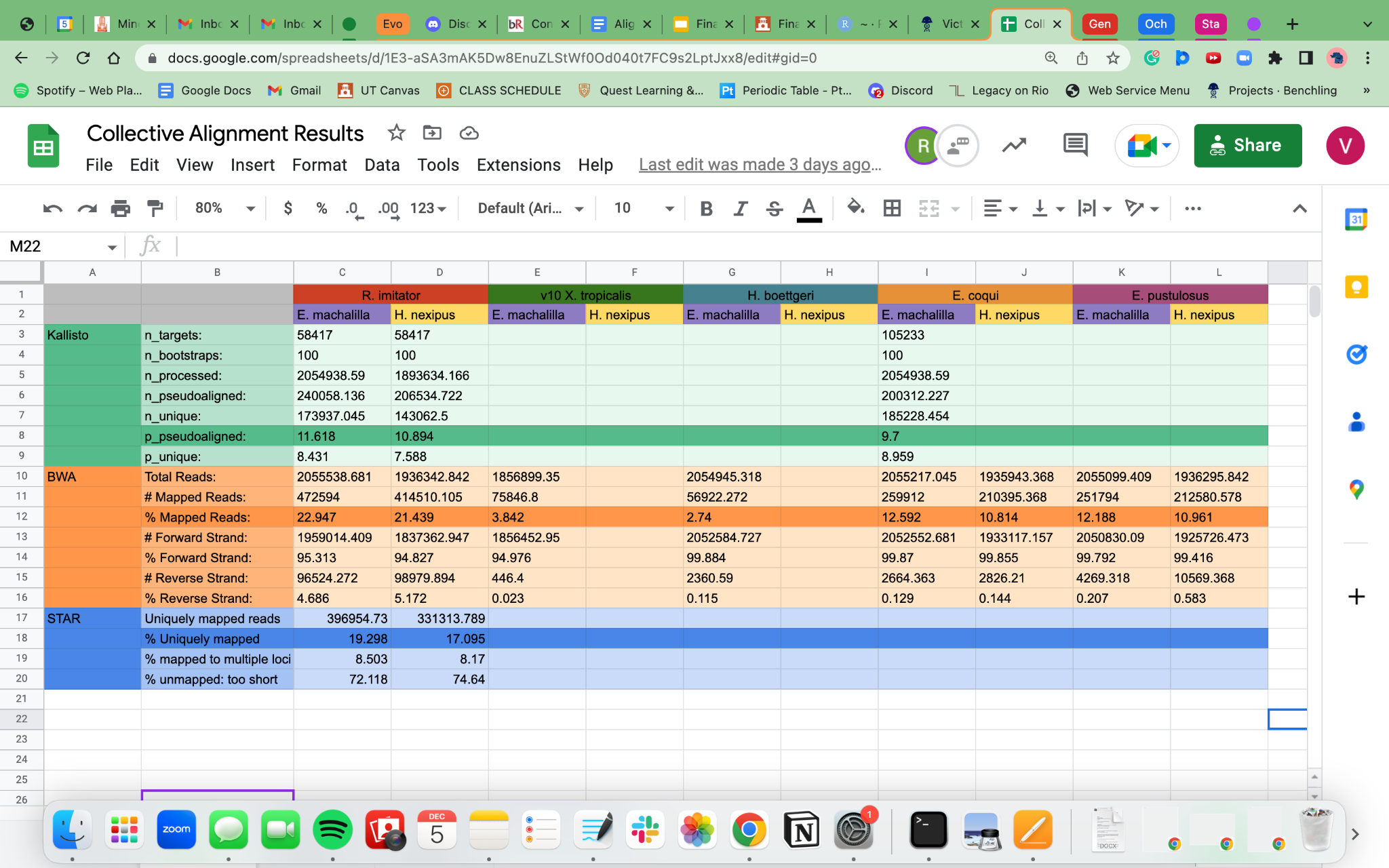
After we have selected the optimal reference transcriptome and alignment software, we created a counts matrix from them. We then used the Uniprot ID Mapping Software (David, 2008) to convert the reference’s annotations to more usable gene IDs including Ensembl ID (Hubbard, 2002), Xenbase ID (Fortriede, 2020), and Entrez ID (Maglott, 2007). Ensembl ID was chosen because of its popular gene database and its good compatibility. Xenbase was chosen due to its species specific gene database which may help in finding more relevant genes associated with gastrulation. Lastly, Entrez ID was chosen to work around the multiple animal gene problem that we faced with the reference annotations as they have a corresponding gene symbol. After converting the annotations to our chosen gene IDs, the quantities of relevant differentially expressed genes for each gene ID was produced and recorded.



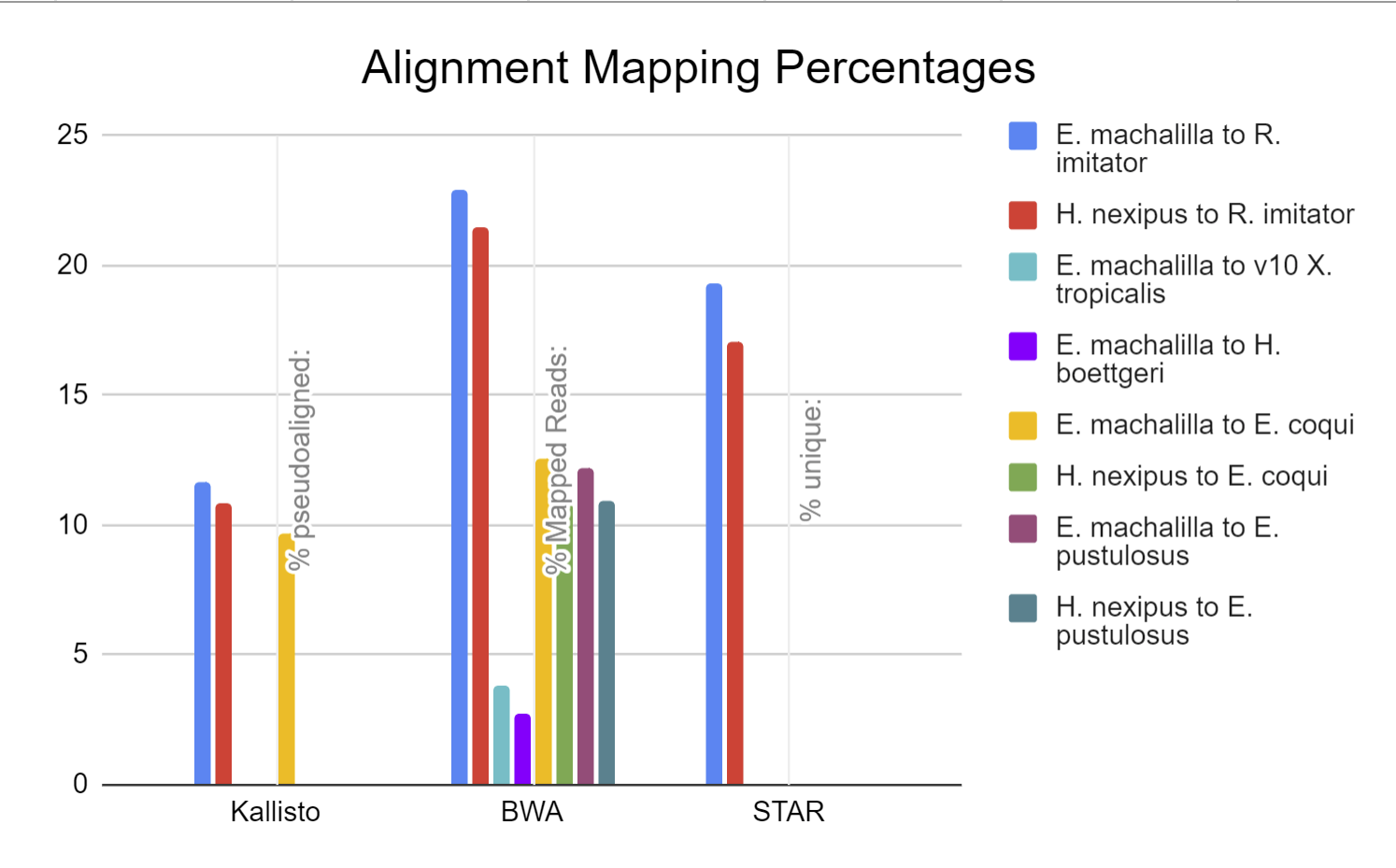
**Fig. 2** A detailed workflow chart illustrating the overall process of alignment.

**Results**

We collected all the mapping percentages produced from kallisto, STAR, and BWA-mem alignment software done on each reference species and put them in an organized spreadsheet (see Fig. 3). We also created a bar plot (see Fig. 4) to visualize the mapping percentages across our references and different alignment softwares. The overall highest mapped read is from bwa-mem using the *R. imitator* reference, with mapping percentages of *E. machalilla* and *H. nexipus* at 22.947% and 21.439%, respectively.

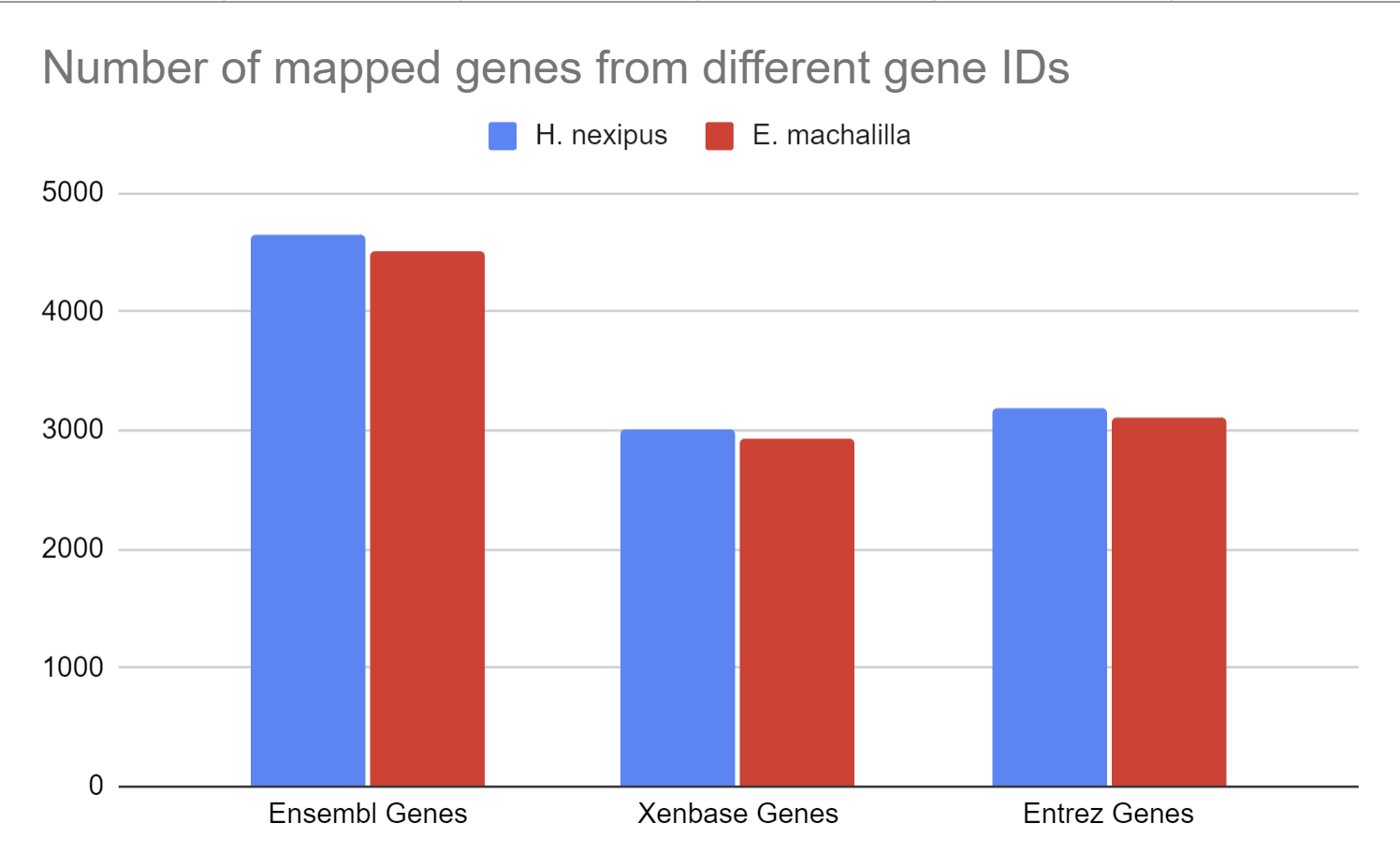


**Fig. 3** A spreadsheet of acquired alignment results across different alignment softwares and across alternate alignment references. The different reference species are placed at the top with each reference split between *E. machalilla* and *H. nexipus*. The three main alignment software one the side each contain their respective alignment statistics and the percentage of mapped reads are highlighted.



**Fig. 4** A barplot to provide a visualization of the mapping percentages between the three alignment softwares. Each bar represents the alignment of one of our *Dendrobatid* species and one of our alternate reference species. The taller the bar, the higher the mapping percentage.

After we choose the alignment and and reference species based on mapping percentage, we generated a counts matrix using the reads. The counts matrix is generated from the mapped reads from bwa-mem using the *R. imitator* reference. The reference transcriptome contains a list of annotations, and this list of annotations are converted to Ensembl, Entrez, and Xenbase genes. Ensembl genes have the most amount of genes with 4,642 genes for *H. nexipus*, and 4,505 genes for *E. machalilla*. Xenbase resulted in 3,003 genes for *H. nexipus,* and 2,935 genes for *E. machalilla*. Entrez resulted in 3,196 genes for *H. nexipus,* and 3,117 genes for *E. machalilla*. A bar plot illustrating the number of genes between each gene ID for each species is made (See **Fig. 5**).



**Fig. 5** A barplot of the number obtained genes after gene ID conversion. The plot is divided between three groups representing a Gene database that we used for ID conversion. Each group is divided between our two *Dendrobatid* species with each bar representing the number of genes each gene conversion yielded for the *Dendrobatid* species.

**Discussion:**

The ultimate outcome of the entire project is to identify the genes that are related to and responsible for changes to gastrulation in terrestrial species. To do that, we must have a sufficient enough quality of mapped tag-seq reads from our samples in order to obtain a large amount of genes to work with. We are using multiple reference transcriptomes and alignment softwares to maximize our mapping percentage for downstream analysis. We chose references based on our phylogenetic model where we predict that species that are more closely related to our *Dendrobatids* will yield higher mapping percentages.

Of the aligned reads that we have generated, bwa-mem with the *R. imitator* reference transcriptome has the highest overall mapping percentage with 22.9% and 21.4% for *E. machalilla* and *H. nexipus* respectively. This outcome supports our phylogenetic model as the *R. imitator* species is most closely related to our dendrobatid species compared to the other alternate reference species. The *E. coqui* and *E. pustulosus* references yielded very similar mapping percentages with around 12% mapped to *E. machalilla* and 11% mapped to *H. nexipus*. This also fits our phylogenetic model as *E. coqui* and *E. pustulosus* have about 72 million years and 68 million years, respectively, of separation from our *Dendrobatids*. Additionally, *E. coqui* and *E. pustulosus* resulted in almost half the mapping percentage compared to *R. imitator* which is significant seeing that these reference species have over twice the number years of separation from our *Dendrobatids* when compared to *R. imitator*. The *X. tropicalis* and *H. boettgeri* references have the lowest overall mapping percentage with less than 4% mapped. Again, this result fits our phylogenetic model as they both have approximately over 200 million years of separation from our *Dendrobatids.* Overall, our mapping percentages have a good correlation with our phylogenetic model as references with less years of separation yields higher mapping percentages.

From our mapping results, the bwa-mem alignment software also seems to yield higher overall mapping percentages compared to our other alignment softwares which shows continuity with the results produced by the summer group. The Kallisto software seems to have the lowest mapping percentage overall which may be because the software is a pseudo aligner unlike our other alignment software. As a result, we stopped running kallisto and focused on bwa-mem and STAR. STAR seems to have a slightly lower mapping percentage when compared to bwa-mem, however more alignments using STAR should be done in order to verify this observation.

Overall, the conversion from the annotations to our gene IDs resulted in a large amount of lost genes. From our counts matrix, the Ensembl Gene IDs contained the greatest number of genes with around 4,500 genes, which is greater than the 3,000 genes from Xenbase and Entrez. This however does not necessarily mean using Ensembl ID is the best choice as our Ensembl gene list contains genes from over 100 unique species. This is not ideal as two separate gene IDs from different species may inaccurately point to the same gene. As a result, our number of unique Ensembl genes may be inflated. Xenbase and Entrez genes gives us a better idea of the exact number of unique genes we can work with for Xenbase is species specific and Entrez has a corresponding gene symbol which helps with eliminating our multiple species problem in the annotations. The number of unique genes we have found, however, is not as much as we had hoped for. But the counts matrix generated using the original *X. tropicalis* reference transcriptome yielded a similar amount of genes with around 3,500 genes.

We would still want to run downstream analysis using the counts matrix generated from the *R. imitator* reference to see if the results are consistent with the *X. tropicalis* results. We also do not know exactly what genes are mapped, so a statistical analysis to determine if there are a significant amount of genes related gastrulation should be done on our data to identify viable genes for future analysis. Our mapped reads could have potentially captured important genes related to gastrulation that the *X. tropicalis* aligned reads did not. And a different software could also be used to translate our annotations to a gene database that may result in a lower amount of lost genes. The Entrez gene ID should also be converted to their corresponding gene symbols. Thus, the results produced from this project and the continuation of downstream analysis takes us one step closer to understanding the underpinnings of gastrulation.

**Citations:**

**BWA-mem**

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].

**kallisto**

Bray, N., Pimentel, H., Melsted, P. *et al.* Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34, 525–527 (2016). <https://doi.org/10.1038/nbt.3519>

**STAR**

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. <https://doi.org/10.1093/bioinformatics/bts635>

**Uniprot ID Mapping**

David, F. P., & Yip, Y. L. (2008). SSMap: a new UniProt-PDB mapping resource for the curation of structural-related information in the UniProt/Swiss-Prot Knowledgebase. *BMC bioinformatics*, *9*, 391. <https://doi.org/10.1186/1471-2105-9-391>

**Ensembl Gene**

T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, M. Clamp, The Ensembl genome database project, *Nucleic Acids Research*, Volume 30, Issue 1, 1 January 2002, Pages 38–41, <https://doi.org/10.1093/nar/30.1.38>

**NCBI Entrez Gene**

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic acids research*, *35*(Database issue), D26–D31. <https://doi.org/10.1093/nar/gkl993>

**Xenbase Gene**

Joshua D Fortriede, Troy J Pells, Stanley Chu, Praneet Chaturvedi, DongZhuo Wang, Malcom E Fisher, Christina James-Zorn, Ying Wang, Mardi J Nenni, Kevin A Burns, Vaneet S Lotay, Virgilio G Ponferrada, Kamran Karimi, Aaron M Zorn, Peter D Vize, Xenbase: deep integration of GEO & SRA RNA-seq and ChIP-seq data in a model organism database, *Nucleic Acids Research*, Volume 48, Issue D1, 08 January 2020, Pages D776–D782, <https://doi.org/10.1093/nar/gkz933>

**Data Availability:**

Alternate References:

The assemblies, annotations, and raw data are deposited in NCBI for v10 *X. tropicalis* ([BioProjects PRJNA577946](https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000004195.4/)), *E. coqui* ([BioProject PRJNA578591](https://www.ncbi.nlm.nih.gov/data-hub/genome/GCA_019857665.1/)), *E. pustulosus* ([BioProject PRJNA578590](https://www.ncbi.nlm.nih.gov/data-hub/genome/GCA_019512145.1/)), *H. boettgeri* ([BioProject PRJNA578589](https://www.ncbi.nlm.nih.gov/data-hub/genome/GCA_019447015.1/)), and *R. Imitator* ([BioProject PRJNA508211](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3499533)).