Published on Sep 15, 2023 by Arcadia Science **DOI**: 10.57844/arcadia-e367-8b55

Repeat expansions associated with human disease are present in diverse organisms

Some human proteins are encoded by genes with repetitive sequences, which, if they expand, damage the nervous system and cause disorders like Huntington's disease. We found animals with similar proteins that have more repeats than we've ever seen in healthy people.

Contributors (A-Z)

Prachee Avasthi, Feridun Mert Celebi, Seemay Chou, Rachel J. Dutton, Megan L. Hochstrasser, Elizabeth A. McDaniel, Kira E. Poskanzer, Taylor Reiter, Michael E. Reitman, Dennis A. Sun, Emily C.P. Weiss

Version 1 · Mar 31, 2025

Purpose

We wanted to explore human repeat expansion disorders, which are not well understood and have few effective therapeutic options. We hoped to provide clues into these disorders by exploring the taxonomic conservation of proteins that commonly contain pathogenic repeats and the range of repeat expansion variability in the organisms where they are found. In the long run, we think this understanding could

suggest appropriate organisms for mechanistic investigation of these disorders, and help inform therapeutic strategies.

Our first goal was to determine if other species have homologs of proteins with disease-related repeat expansions. If so, our second goal would be to determine if any homologs had more repeats than seen in healthy humans. We imagined that we might find species that exhibit some sort of pathogenic phenotype and could thereby serve as new disease models. Conversely, we might identify organisms that have large numbers of repeats but aren't afflicted by disease, which would suggest novel avenues for therapeutic investigation.

Using a combination of sequence- and structural-similarity searches, we identified ~400 homologous proteins that have longer repeats than found in healthy humans. We found that some groups of animals have multiple proteins with repeat expansions, including marsupials, bats, and shrews. While we don't currently plan to follow up on this work, we hope other scientists interested in neurodegeneration, DNA repair, and comparative biology build upon these findings.

- Access **data** from this pub, including tables of our similarity search hits and repeat-counting results, on Zenodo.
- All associated **code** is available in a series of GitHub repositories. See code for profiling the initial comparative results, assessing repeat length distribution in koala population sequencing data, and validating the expression of the identified homologs in RNA-seq data.

We've put this effort on ice! ☒

#TranslationalMismatch #LackingInfrastructure

We ended up mostly finding interesting repeat expansions associated with diseases that are rare, developmental, and bottlenecked by current experimental assays. Thus, further efforts would lack the translational potential to justify establishing new in-house assays and model systems at Arcadia to overcome these barriers.

Learn more about the Icebox and the different reasons we ice projects.

Background and goals

SHOW ME THE DATA: Access our repeat expansion homolog data, including tables of our similarity search hits and repeat-counting results (DOI: 10.5281/zenodo.10407307).

Simple DNA sequence repeats (e.g. CAGCAG) are widespread throughout the genomes of all eukaryotic organisms **[1]**. They have important roles in modulating gene expression and protein function **[2]**. Yet repeats also have a hidden danger: they are prone to mutations **[3]**. The number of repeats can expand over time, increasing with age **[4]** and across generations **[5]**. When these repeat expansions occur in protein-coding regions, they can cause devastating diseases.

Repeat expansions are associated with over 40 neurological disorders, including Huntington's disease, which profoundly damages the central motor centers of the brain and ultimately leads to cognitive impairment **[6]** and death. Many more expansion disorders are likely still undiscovered **[7]**. It's not fully known which repeat expansions will lead to human disease, or why repeat expansions primarily cause diseases specific to the nervous system **[8][9]**. Answering these questions could help us understand repeat expansion disorders, with the ultimate goal of creating better diagnostics and treatments. Current treatments for expansion disorders treat disease

symptoms (e.g. motor impairment) without addressing their root cause, leading to poor prognosis and clinical outcomes **[10]**.

Traditional animal models of neurodegeneration have been helpful for investigating disease processes *in vivo* and determining new therapeutic targets. However, their ability to predict clinically relevant treatments is poor, in part because they fail to accurately model human disease **[11]**. This is true for repeat expansion disorders which, despite approximately thirty years of extensive disease modeling and drug development using worms (*C. elegans*), flies (*D. melanogaster*), and mice (*M. musculus*) **[12][13][14]**, still lack adequate treatments **[7]**. Here, we looked for alternative ways to study repeat expansion disorders.

We reasoned that repeat expansions likely occur in similar proteins across organisms and hypothesized that human disease-associated repeat expansions (dREs) might occur in other species too. We hypothesized that if we discovered species with repeat expansions and phenotypes that mirror human disease, these species would provide a basis for natural disease models. Additionally, some species may have molecular mechanisms to compensate for repeat expansions, which would manifest as species with many repeats but without phenotypic effects. These species, if they exist, could provide insights into the factors required for repeat expansions to lead to pathology and the factors that prevent it, providing a basis for developing new therapeutics for repeat expansion disorders.

The approach

To demonstrate a proof-of-concept as efficiently as possible, we took a comparative approach (Figure 1). We used a published list of 60 disease-associated repeat loci from humans **[15]** and trimmed it down to just the 55 that occur in unique proteins. We then used a two-pronged strategy, using both sequence similarity with protein BLAST and structural similarity with Foldseek **[16]**, to find homologous proteins in other species. We searched for structural similarity on a subset (19/55) of proteins to confirm the utility of this approach and allow for iteration. After identifying structurally similar proteins, we downloaded the amino acid sequences for those proteins and profiled the repeats. We only analyzed repeats for query proteins that have a single repeating amino acid within coding regions (26/55 proteins). Finally, we compared the lengths of homolog repeats we found to the longest repeat length observed in healthy humans.

We describe detailed methods below — click here to skip straight to the results.

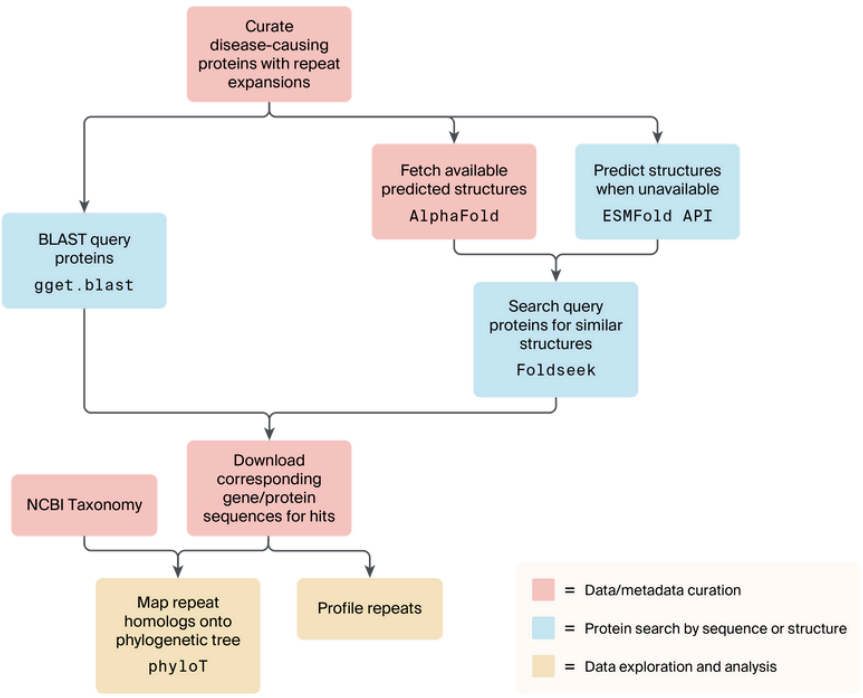

Figure 1

Schematized workflow for identifying homologs with repeat expansions.

Sequence homology

We used the gget package **[17]** using the `gget.blast` command in Python (version 3.11.4) to BLAST our proteins of interest against the non-redundant NCBI protein database with default search settings and a limit of 10,000 hits. We filtered our results with a sequence identity of 30% and a query coverage of 50%. After finding homologs, we filtered our results so each species had at most one homolog per queried protein.

Structural similarity

To find structurally similar proteins, we pulled human disease-related expansion protein AlphaFold structures of any size that were in the Protein Data Bank **[18][19]**. In cases where protein isoforms did not have an AlphaFold structure, we predicted

structures of the isoforms using an ESMFold API query **[20]** if they were shorter than 400 amino acids, or using ColabFold (version 1.5.2) with default settings **[21]** if they were larger than 400 amino acids. We used these PDB files to query the Foldseek web API **[16]** using the AlphaFold/UniProt50, AlphaFold/Swiss-Prot, and AlphaFold/Proteome databases (all version 4) with a maximum of 1,000 hits returned per database **[18][19]**. The scripts we used to query ESMFold and Foldseek are available in our GitHub repo (foldseek_apiquery.py and esmfold_apiquery.py).

Repeat length determination and comparison to humans

We used a custom-written script, developed with ChatGPT (GPT-3) and verified using test sequences for accuracy, to look for repetitive amino acid sequences in the homologs we found. TM-scores below 0.2 are considered to be unrelated proteins **[22]**; therefore, before repeat counting, we filtered Foldseek hits to keep only those with a > 0.2 TM-score against the query protein. For comparison to human repeats, we identified the longest repetitive stretch of whichever amino acid is linked to disease in the human homolog, regardless of its location. We then compared this length to the maximum repeat length in healthy humans based on the published list of disease-causing repeats **[15]**. COMP had no maximum listed and the PABPN1 limit was only relevant to the nucleotide, not amino acid repeats, so we sourced these limits from other references **[23][24]**. For the distribution of human androgen receptor repeat lengths (Figure 3, top left), we used data from the STRipy database.

When comparing the distribution of repeat lengths to humans, we took the longest repeat in each species. For ZIC3 and HOXD13 homologs, we noticed our searches returned homologs of ZIC2 and HOXA13, which have longer repeats and pathological limits in humans. Therefore, to avoid false positives, we excluded ZIC3 and HOXD13 from taxonomic tree visualizations. We used the seaborn package (version 0.12.2) in Python (version 3.11.4) to visualize the results.

Taxonomic tree and bar chart visualizations

To make taxonomic trees, we extracted lineage information for each NCBI taxid from the NCBI taxonomy table downloaded from the NCBI FTP site, as described in

"NCBI_taxid_to_lineage_and_barchart_tree_plotting.ipynb." We used the tidyverse (version 2.0.0) **[25]**, magrittr (version 2.0.3) **[26]**, and pacman (version 0.5.1) **[27]** packages in R (version 4.2.2) to analyze the data, to produce counts and average counts for the number of homologous proteins per taxonomic group and query protein, and to create a bar chart of the number of hit proteins per query protein. We used lineage information from NCBI taxonomy to create a taxonomic tree in phyloT (version 2) with phyloT database (version 2022.3); we used scientific names as node identifiers, expanded internal nodes, set the the "polytomy" option to "yes," and exported a Newick tree. We then uploaded Newick files to the iTOL (version 6.8) web server for visualization and formatting **[28][29]**.

Analysis of repeat length distribution in koala population sequencing data

To confirm that our results were not caused by an individual anomaly or genome assembly error, and to look at the distribution of repeat lengths in a natural population, we took advantage of previously existing koala population sequencing data **[30]**. We designed a pipeline to look at the repeat lengths in the koala RUNX2, FOXL2, ARX, and ZIC2 genes. Because only data-heavy BAM files of reads aligned to the koala reference genome were available (rather than individual genome assemblies), we used:

- s5cmd (version 2.2.2) **[31]** to download a single BAM file and associated indexing file from AWS
- SAMtools (version 1.17) **[32]** to extract the regions of the four genes of interest from the alignment based on their location in the koala reference genome (RefSeq assembly GCF_002099425.1)
- BEDtools (version 2.31.0) **[33]** to extract the reads from these extracted regions into per-gene FASTQ files

We removed the BAM and indexing files immediately after extraction to avoid storing BAM files locally. We repeated this process for all 430 koala samples. We then:

- assembled extracted reads for each gene using SPAdes (version 3.15.2) **[34]**

- predicted open reading frames (ORFs) and translated them with orfipy (version 0.0.4) **[35]**
- pulled the correct ORF out from the set of predictions using pattern matching to four amino acid sequences directly upstream of the expansion
- determined expansion lengths of the relevant amino acid (glutamine or alanine) for each gene and sample
- collected the results into a final table

The expansions we analyzed were around 60 base pairs (20 amino acids), relative to the 150 bp sequencing read length, suggesting that assembly error is unlikely to prevent us from accurately capturing expansions. We incorporated all of these steps, starting from data download, into a Snakemake pipeline **[36]**.

Validation of expression of homologs in brain and muscle tissues

From a list of species containing homologs of disease-causing repetitive genes, we queried for existing RNA sequencing datasets in the SRA for those species. We used the NCBI Entrez tools (version 19.2) **[37]** to first search in the SRA for all datasets matching the species of interest and gather the SRA run info. We then passed these run accessions to pysradb (version 2.2.0) **[38]** to access the metadata for each run. We filtered for SRA runs that were whole-tissue RNA-seq experiments from either brain or skeletal muscle tissues and with a minimum sequencing depth of 1 million reads.

We then created a workflow that automates downloading and processing data for mapping the RNA-seq experiments against the corresponding species genome. For each species, we downloaded the RefSeq genome and corresponding GTF annotation file. We downloaded each RNA-seq experiment with SRA-tools (version 3.0.6). We indexed each species' reference genome with STAR (version 2.7.11a) **[39]**, mapped corresponding RNA-seq experiments with STAR, sorted with SAMtools (version 1.18), and quantified gene counts with HTSeq (version 2.0.3) **[40]**. We then applied a threshold that if a gene was above the median count of reads in a sample, we counted it as "expressed." We then plotted the percentage of genes for a species in a sample type that we counted as expressed. For parsing and plotting expression results, we

used R (version 4.3.1) and packages tidyverse (version 2.0) **[25]** and ggpubr (version 0.6.0) **[41]**.

> All the **code** we generated and used for the pub is available in a series of GitHub repositories. See code for profiling the initial comparative results (DOI: 10.5281/zenodo.10403607), assessing repeat length distribution in koala population sequencing data (DOI: 10.5281/zenodo.10403617), and validating the expression of the identified homologs in RNA-seq data (DOI: 10.5281/zenodo.10403614).

Additional methods

We used ChatGPT to write some code and clean up other code.

The results

> **SHOW ME THE DATA**: Access our repeat expansion homolog data, including tables of our similarity search hits and repeat-counting results.

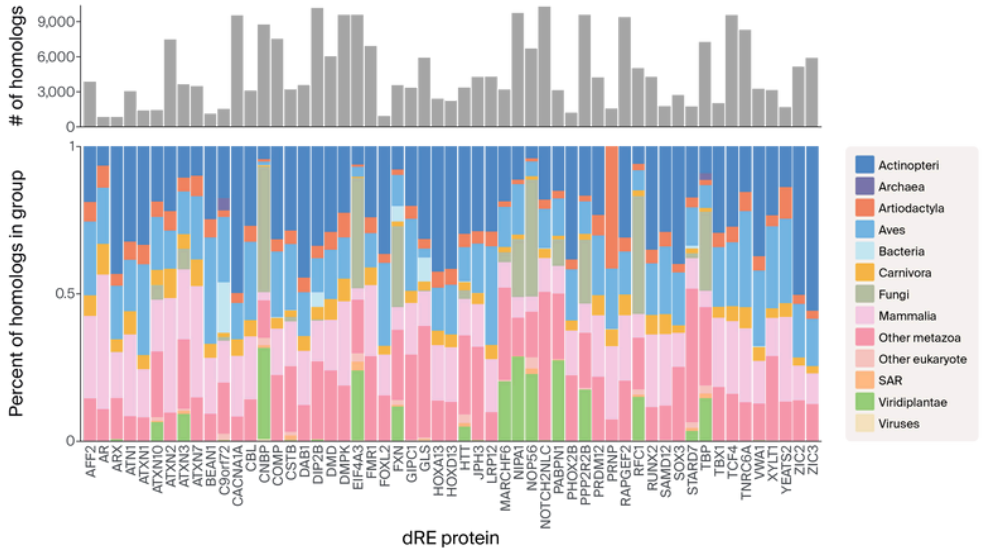

Figure 2

Homologs of human proteins with disease-associated simple sequence repeats.

The number (top row) and the taxonomic distribution of homologs identified using a combination of sequence- and structure-based searches for each query protein (x-axis).

Using a combination of sequence and structural similarity, we identified ~1,000–10,000 similar proteins of each of the 55 proteins we queried, which we describe here as “homologs” (Figure 2). Such homologs are widely distributed across metazoans. We also identified proteins, like FXN, that have homologs in fungi and plants but are highly divergent from humans (Figure 2, bright and muted green). This intrigued us because we chose to investigate these proteins for their connection to neurological disease. Our results suggest a subset of these proteins are widely conserved in species without a nervous system, suggesting repeat expansion may lead to different outcomes across species and cell types. Overall, we conclude that protein families underlying expansion disorders are not human-specific, but instead shared across species.

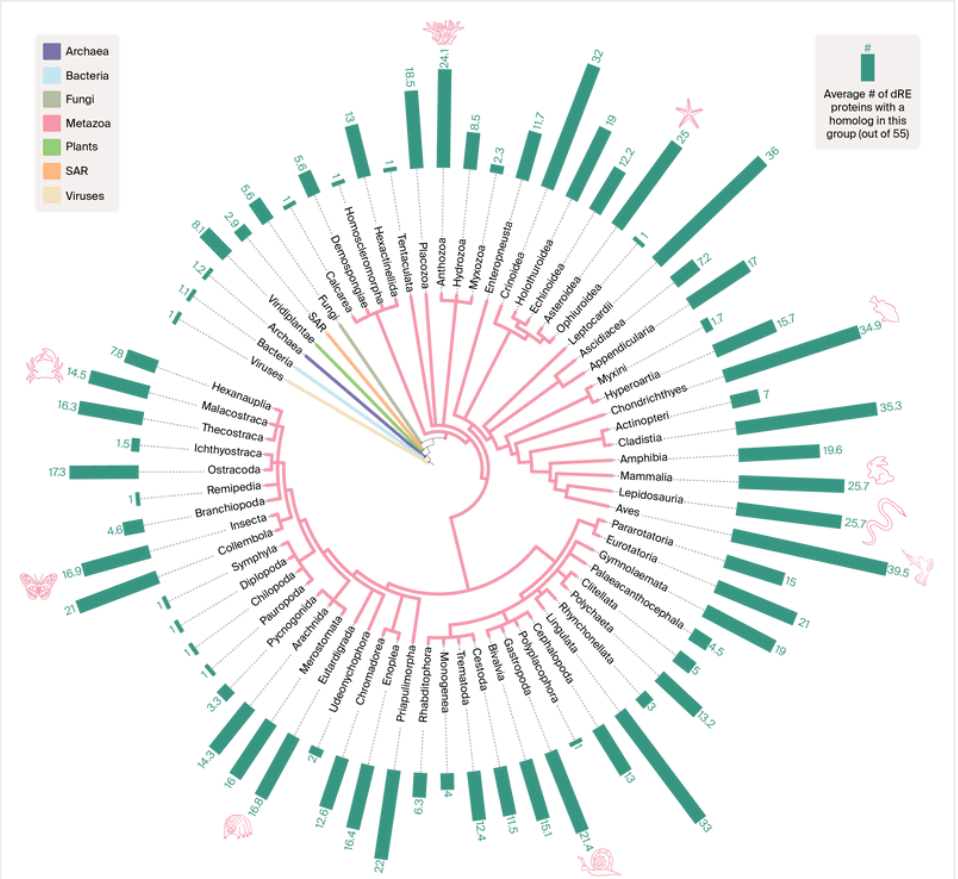

Figure 3

Taxonomic distribution of homologs of human proteins with disease-associated simple sequence repeats.

Taxonomic tree of groups with at least one homolog to a dRE protein, with branches colored by taxonomic group. Outer bars display the average number of query proteins for which the organisms in the taxonomic group had a hit. We've included a handful of organism silhouettes to provide a sense of what types of organisms have dRE homologs.

To consider our results through an evolutionary lens, we mapped homologs onto a taxonomic tree (Figure 3). We found that while metazoans have many homologs of our queried proteins, the average number of homologs varied widely by taxonomic group. The wide variety in homology across groups suggests that there may be important patterns of evolutionary loss and duplication that would help elucidate the origins and functions of repeat expansion proteins.

We next wanted to understand if there was natural variation in repeat lengths in the homologs we found, and particularly if there was any variation outside the range found in healthy humans. To do this, we assessed the repeat lengths in each homolog and compared them to the maximum length observed in healthy people. We used the amino acid sequence to look for repeats, and therefore only analyzed coding-region repeats that are not the result of insertion mutations (26/55 proteins). For example, the androgen receptor has a repeat length between 12 and 32 in the human population (Figure 3, A; top) and the maximum number of repeats in healthy humans is 40 (Figure 4, A; green dashed line). When compared to the repeat lengths in androgen receptor homologs (Figure 4, A; bottom), we saw most species have repeat lengths below the healthy human limit. However, using this methodology, we observed four species (Brandt's bat, Eurasian Badger, Short-eared elephant shrew, and White-tailed rat) with repeat lengths longer than we see in healthy humans.

This finding was not unique to the androgen receptor. We saw that most species have homologs with fewer repeats than the healthy human maximum. Yet, for each protein, we found a few species that have homologs with repeat lengths that match or exceed the healthy human limit (Figure 4, B). This is exciting because it suggests species may exist that have proteins with similar structures and mutations to the human homologs that cause nervous system disease.

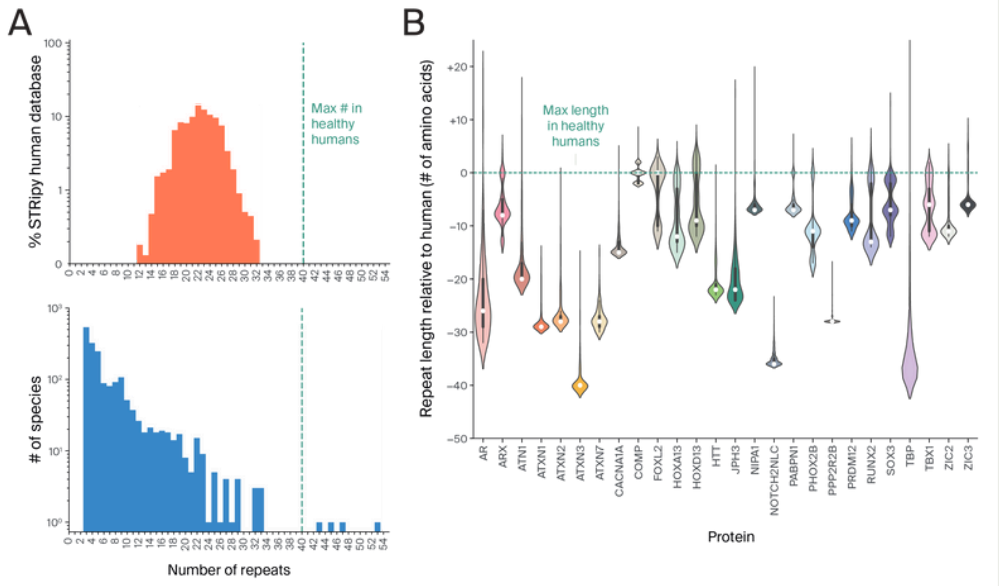

Figure 4

Distribution of amino acid repeat lengths for identified homologs relative to the healthy human limit.

(A) Distribution of repeat lengths in human population data (top) and distribution of the number of repeats across species with homologs (bottom) of the androgen receptor (AR) query protein. The green dashed line indicates the maximum number of repeats observed in healthy humans.

(B) Distribution of repeat lengths across homologs relative to their length in humans. This plot only includes repeats that, for each individual query protein, are known to cause human disease (e.g. polyglutamine for AR). We've normalized the repeat length in each homolog to the maximum number of repeats found in healthy humans. For one species, Brandt's bat, the calculated TBP repeat length exceeds the y-bounds of the graph, so we've cut off the top of the violin plot for legibility.

Finally, we wanted to know if there were any particular species or taxonomic groups that would be the best candidates for finding compensatory mechanisms to repeat expansion-associated disease. We suspected that organisms with multiple proteins containing repeat expansions may have evolved mechanisms to avoid their deleterious effects. For each species, we asked how many homologs it has with

repeats that exceed the length found in healthy human variation. We found that, on average, some taxonomic groups, including rodents and bats, typically have 1–2 homologs with repeat expansions per species (Figure 5, A). These expansions are not restricted to a specific set of homologs, but are found across many of the homologs we investigated (Figure 5, B & Figure 6). In contrast, we found that on average, certain taxonomic groups, like marsupials (Diprotodontia), have three or more homologs with repeat expansions per species (Figure 5, A). These are limited to a small subset of homologs, primarily in the genes ARX, FOXL2, RUNX2, and ZIC2 (Figure 5, B, and Figure 6), all of which play important roles as developmental transcription factors. The presence of multiple genes containing expansions that would be pathogenic in humans could suggest that these proteins have different functions or interacting partners, developmental contexts, or cellular environments in marsupials that prevent them from being pathogenic. The apparent enrichment of “long” expansions in marsupials could also suggest that they’ve evolved mechanisms for preventing or dealing with toxic gain-of-function effects for these expansions.

To further validate our findings in marsupials, we analyzed publicly available population sequencing data for koalas **[30]**. We looked at the distribution of repeat expansion lengths in the ARX, FOXL2, RUNX2, and ZIC2 proteins. Of the 430 genomes we analyzed, only 10 koalas have expansion lengths in any of these four genes that differ from the reference genome, validating our initial finding. In the 10 cases that differ from the reference, these expansions are shorter than the reference expansion. In addition to the previous findings, we saw many COMP homologs, especially in birds and fish, with expansions longer than those seen in humans (Figure 5, B & Figure 6); however, these expansions are only marginally longer (by one amino acid), making us uncertain about the biological significance of this difference.

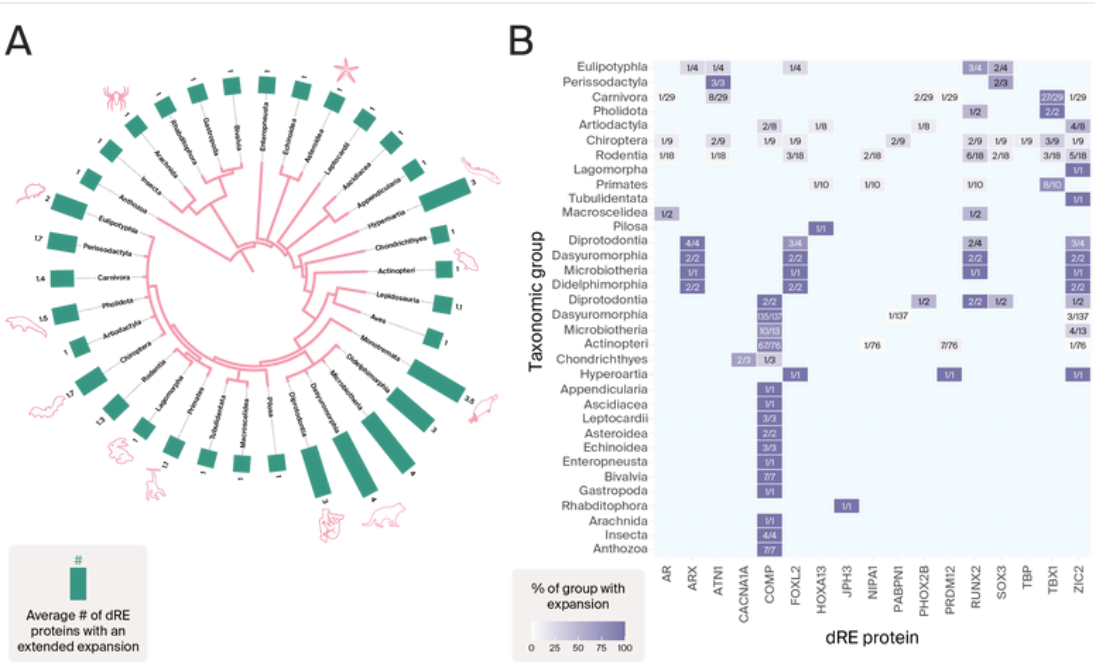

Figure 5

Taxonomic distribution of homologs with amino acid repeat lengths that exceed the maximum repeat length seen in healthy humans.

(A) Taxonomic tree of groups with at least one homolog to a dRE protein that contains a repeat longer than what is seen in healthy humans. Outer bars display the average number of query proteins for which the organisms in the taxonomic group had a homolog with a repeat length that exceeds this maximum. We've included a handful of organism silhouettes to provide a sense of what types of organisms have long repeat expansions. This plot only includes repeats that, for each individual query protein, are known to cause human disease (e.g. polyglutamine for human androgen receptor). The average does not take into account any organism in the group that did not have at least one homolog with a repeat length greater than the human maximum.

(B) Heatmap showing, per query gene, how many species (out of the total organisms in the group with at least one homolog with a repeat length greater than the healthy human maximum) had an extended repeat in this protein. Actual species numbers out of total species included in the group are shown as a fraction in the heatmap cell.

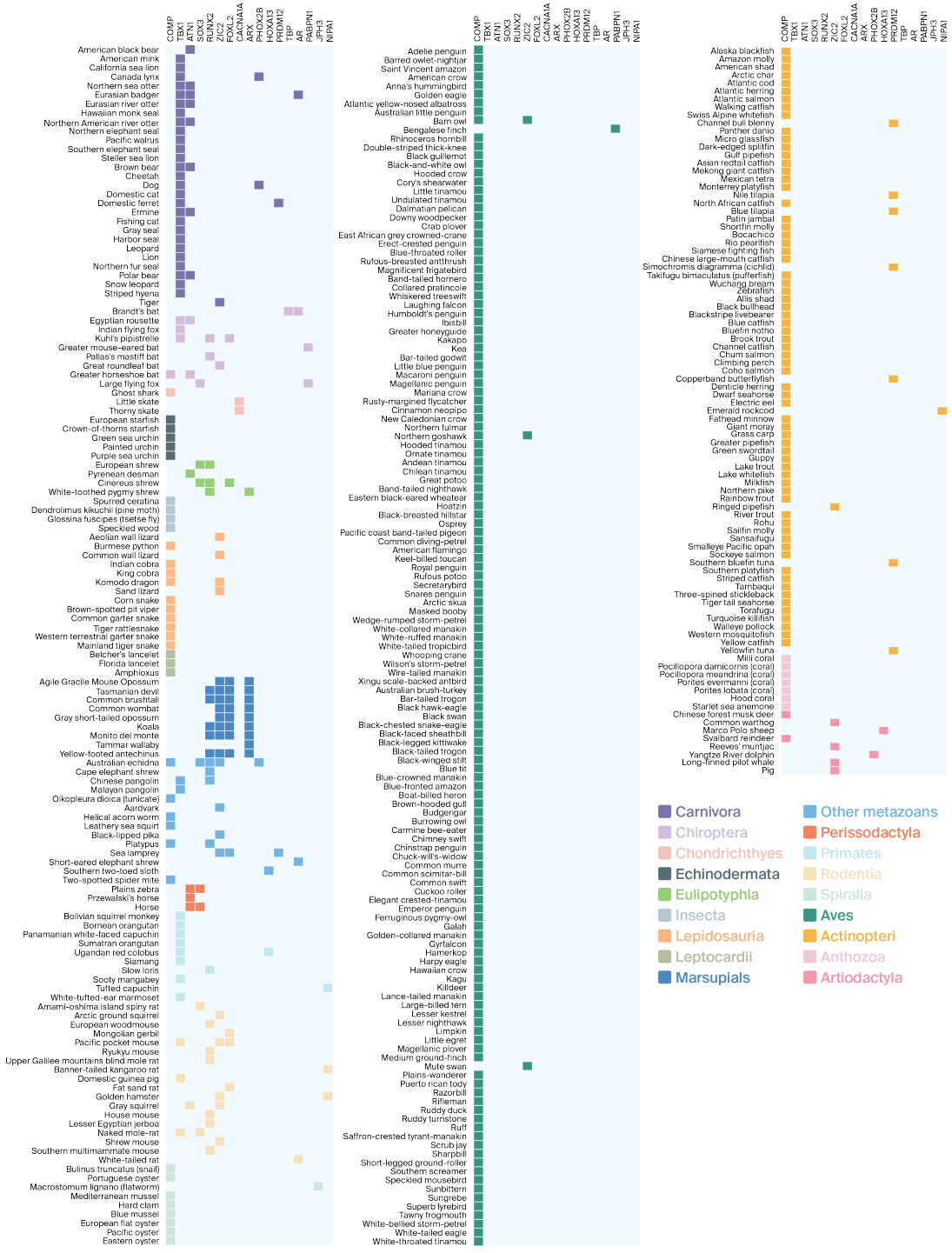

Figure 6

Species-level information for homologs with amino acid repeat lengths that exceed the maximum repeat length seen in healthy humans.

This plot only includes repeats that, for each individual query protein, are known to cause human disease (e.g. polyglutamine for human androgen

receptor).

From our preliminary comparative results, we wanted to validate if the homologs in the identified species are expressed in brain or muscle tissues, since most of the disease-causing loci are associated with neurodegenerative diseases. We were able to collect RNA-seq data from brain and/or muscle tissues from 42 species and focused on a subset for these preliminary checks, including the Australian echidna, common brushtail, European shrew, gray short-tailed opossum, koala, little skate, monito del monte, and naked mole rat (Figure 7). We set a conservative threshold where any gene with a read count higher than the median count in that sample was considered “expressed.” We found that for the most part, the identified homologs in these species are indeed expressed in brain and muscle tissues. These results are encouraging, suggesting that the identified homologs may have functional significance in these tissues and could be useful for downstream wet-lab experiments

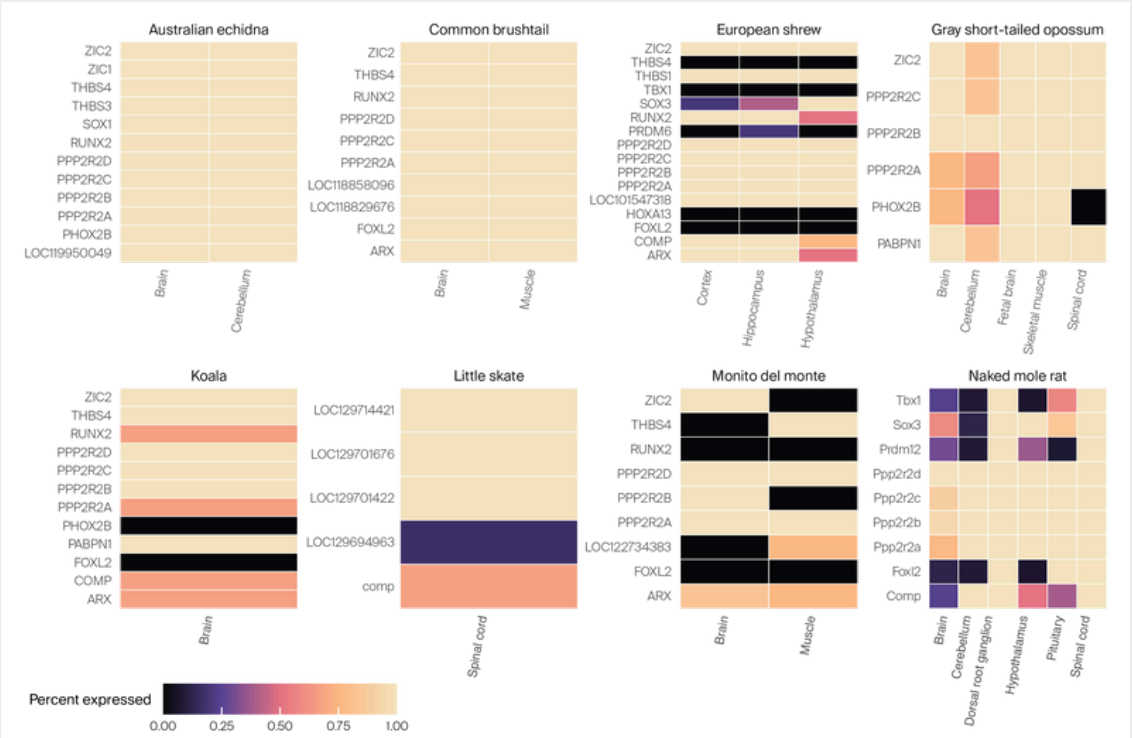

Figure 7

Expression of homologs in brain and muscle tissues in select species.

We collected brain and muscle RNA-seq experiments and mapped to the reference genome of each species. We considered genes “expressed” in that tissue if the count of reads mapping to the gene was higher than the median read count. For each species, we show the percentage of the specific tissue type samples in which we consider the gene expressed.

Overall, we identified species with homologs to human proteins that contain repeats longer than any seen in healthy humans. We hypothesize that these species may have functional challenges associated with repeat expansion disorders, or have evolved molecular mechanisms to compensate for repeat expansions. While we did not look into compensatory mechanisms in this work, we think these species could provide a fruitful basis for disease models and new therapies for expansion disorders.

Key takeaways

In this project, we wanted to learn whether other organisms have natural occurrences of repeat expansions associated with human diseases. We took a comparative

approach and found that most human proteins with disease-associated repetitive genome sequences have homologs across metazoans.

We next found that some species have repeats that are longer than ever found in healthy humans. This suggests other species have proteins that look very similar to those that cause human disease. We don't know what the functional effects of these proteins are and our findings suggest three possibilities:

1. Some repeat-expanded homologs in other species naturally lead to pathology that mirrors human disease. These species could be good natural models for human repeat expansion disorders.
2. Some repeat-expanded homologs do not lead to pathology because of compensatory mechanisms, which could serve as a starting point for identifying therapeutics.
3. Some repeat-expanded homologs do not lead to pathology for some other reason, perhaps due to differences in overall cellular context or physiology.

Finally, we found that rodents, bats, and marsupials might be good starting points for further investigation because they have multiple homologs with repeat expansions.

Our results suggest that other species have naturally occurring repeat expansions similar to those that cause disease in humans. We conclude that there are likely species that could be investigated as natural models of human expansion diseases, or as sources of therapeutics.

Limitations

This project was a quick proof-of-principle intended to determine if repeat expansions, similar to those found in human disease, occur naturally in some species. To pursue this goal as efficiently as possible, we limited our search to coding-region repeats and used amino acid sequences to characterize repeat lengths. This strategy provided promising initial results, but cannot determine whether there are repeats at the nucleotide level, nor can it quantify repeat lengths in non-coding regions, which account for ~50% of the disease-related repeat expansions for which we identified homologs. It would complement these initial findings to count nucleotide repeats in

both the homologs we analyzed and the homologs of non-coding repeat expansion proteins.

Our strategy was also limited by the quality of genome assemblies from which the gene and protein sequences originated. Short-read sequencing technologies cannot fully resolve simple sequence repeats longer than 250 bp. Genomes sequenced with high-coverage long-reads such as PacBio or Nanopore can be used to span long repeat units. However, depending on the repeat type, errors in assembly such as irreversibly collapsing repeats in the assembly graph and fragmentation, can still occur **[42]**. Additionally, there are far fewer high-quality genomes sequenced with long-read technologies than those with short-read draft genomes due to the economic cost of long-read sequencing. Therefore, there are likely cases of false negatives in our results due to genome sequencing and assembly methods for the corresponding homologs. While we would not put a lot of stock in the absolute value of the expansion count, as this may vary by individuals and could be impacted by assembly errors, we do think that the presence of an expansion is likely a true signal rather than a false positive. Indeed, many of our protein homolog hits come from genomes generated with long-read sequencing technologies. We manually investigated 31 of our hits with the longest repeat lengths and found that the majority (18/31, 58%) came from long-read genomes. We conclude that long-read sequencing data was a crucial resource for this approach and that continued analysis will benefit from ongoing efforts to increase long-read datasets across species **[43]**.

Next steps

We've iced this project because it lacks the translational potential to justify experimental next steps at Arcadia. Repeat expansion disorders are rare, and many occur developmentally, which makes the translational path forward challenging. For us, further experiments would require developing in-house assays that work in diverse species and improve upon existing options with unclear translational relevance (e.g. protein aggregation). While we don't currently plan to make this investment, we think future experimental next steps could include:

1. Investigate the structure and *in vitro* aggregation properties of repeat-expanded homologs to determine how they might be useful for disease modeling.

2. Heterologously express repeat expansion homologs in human cells to investigate whether they have pathogenic morphological and physiological effects.
3. Heterologously express repeat-expanded human proteins in the cells of species with natural repeat expansions to determine if some species are resistant to disease-relevant repeat expansion.
4. Perform comparative genetics and transcriptomics of species with repeat-expanded homologs to identify innovations that help overcome the toxic effects of repeat expansion.

Currently, we don't believe that assays of repeat expansion protein properties can be done in a way that is unbiased by our incomplete understanding of mechanism and also has clear translational relevance for human disease. Additionally, we believe heterologous expression experiments will be challenging to interpret based on the differences introduced by expressing proteins across species. These caveats present a substantial bottleneck that we are not prepared to overcome at this time. We hope that others surmount the experimental challenges to pursue these promising avenues and explore repeat-associated human disease from a fresh angle.

References

1 Bagshaw ATM. (2017). Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes. https://doi.org/10.1093/gbe/evx164

2 Wright SE, Todd PK. (2023). Native functions of short tandem repeats. https://doi.org/10.7554/elife.84043

3 Malik I, Kelley CP, Wang ET, Todd PK. (2021). Molecular mechanisms underlying nucleotide repeat expansion disorders. https://doi.org/10.1038/s41580-021-00382-6

4 Kacher R, Lejeune F-X, Noël S, Cazeneuve C, Brice A, Humbert S, Durr A. (2021). Propensity for somatic expansion increases over the course of life in Huntington disease. https://doi.org/10.7554/elife.64674

5 Carpenter NJ. (1994). Genetic Anticipation. https://doi.org/10.1016/s0733-8619(18)30071-9

6 Jimenez-Sanchez M, Licitra F, Underwood BR, Rubinsztein DC. (2016). Huntington's Disease: Mechanisms of Pathogenesis and Therapeutic Strategies. https://doi.org/10.1101/cshperspect.a024240

7 Depienne C, Mandel J-L. (2021). 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? https://doi.org/10.1016/j.ajhg.2021.03.011

8 Paulson HL, Bonini NM, Roth KA. (2000). Polyglutamine disease and neuronal cell death. https://doi.org/10.1073/pnas.210395797

9 Hannan AJ. (2018). Tandem repeats mediating genetic plasticity in health and disease. https://doi.org/10.1038/nrg.2017.115

10 Ellerby LM. (2019). Repeat Expansion Disorders: Mechanisms and Therapeutics. https://doi.org/10.1007/s13311-019-00823-3

11 Ransohoff RM. (2018). All (animal) models (of neurodegeneration) are wrong. Are they also useful? https://doi.org/10.1084/jem.20182042

12 Rudich P, Lamitina T. (2018). Models and mechanisms of repeat expansion disorders: a worm's eye view. https://doi.org/10.1007/s12041-018-0950-8

13 Ueyama M, Nagai Y. (2018). Repeat Expansion Disease Models. https://doi.org/10.1007/978-981-13-0529-0_5

14 Kaye J, Reisine T, Finkbeiner S. (2021). Huntington's disease mouse models: unraveling the pathology caused by CAG repeat expansion. https://doi.org/10.12703/r/10-77

15 Shi Y, Niu Y, Zhang P, Luo H, Liu S, Zhang S, Wang J, Li Y, Liu X, Song T, Xu T, He S. (2023). Characterization of genome-wide STR variation in 6487 human genomes. https://doi.org/10.1038/s41467-023-37690-8

16 van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. (2023). Fast and accurate protein structure search with Foldseek. https://doi.org/10.1038/s41587-023-01773-0

17 Luebbert L, Pachter L. (2023). Efficient querying of genomic reference databases with *gget*. https://doi.org/10.1093/bioinformatics/btac836

18 Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N,

Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. (2021). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. https://doi.org/10.1093/nar/gkab1061

19 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. (2021). Highly accurate protein structure prediction with AlphaFold. https://doi.org/10.1038/s41586-021-03819-2

20 Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. https://doi.org/10.1126/science.ade2574

21 Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. (2022). ColabFold: making protein folding accessible to all. https://doi.org/10.1038/s41592-022-01488-1

22 Zhang Y. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. https://doi.org/10.1093/nar/gki524

23 Delot E. (1999). Trinucleotide expansion mutations in the cartilage oligomeric matrix protein (COMP) gene. https://doi.org/10.1093/hmg/8.1.123

24 Cruz-Aguilar M, Guerrero-de Ferran C, Tovilla-Canales JL, Nava-Castañeda A, Zenteno JC. (2017). Characterization of *Pabpn1* Expansion Mutations in a Large Cohort of Mexican Patients with Oculopharyngeal Muscular Dystrophy (Opmd). https://doi.org/10.1136/jim-2016-000184

25 Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. (2019). Welcome to the Tidyverse. https://doi.org/10.21105/joss.01686

26 Bache S, Wickham H (2022). magrittr: A Forward-Pipe Operator for R. https://magrittr.tidyverse.org, https://github.com/tidyverse/magrittr.

27 Rinker TW, Kurkiewicz D. (2019). pacman: Package Management for R. http://github.com/trinker/pacman

28 Letunic I, Bork P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. https://doi.org/10.1093/nar/gkab301

29 Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. (2006). Toward Automatic Reconstruction of a Highly Resolved Tree of Life. https://doi.org/10.1126/science.1123061

30 Hogg CJ, Silver L, McLennan EA, Belov K. (2023). Koala Genome Survey: An Open Data Resource to Improve Conservation Planning. https://doi.org/10.3390/genes14030546

31 Peak. s5cmd. https://github.com/peak/s5cmd

32 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. (2021). Twelve years of SAMtools and BCFtools. https://doi.org/10.1093/gigascience/giab008

33 Quinlan AR, Hall IM. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. https://doi.org/10.1093/bioinformatics/btq033

34 Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. (2020). Using SPAdes De Novo Assembler. https://doi.org/10.1002/cpbi.102

35 Singh U, Wurtele ES. (2021). orfipy: a fast and flexible tool for extracting ORFs. https://doi.org/10.1093/bioinformatics/btab090

36 Köster J, Rahmann S. (2012). Snakemake—a scalable bioinformatics workflow engine. https://doi.org/10.1093/bioinformatics/bts480

37 Sayers E. A General Introduction to the E-utilities. (2009). https://www.ncbi.nlm.nih.gov/books/NBK25497/

38 Choudhary S. (2019). pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive. https://doi.org/10.12688/f1000research.18676.1

39 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. (2012). STAR: ultrafast universal RNA-seq aligner. https://doi.org/10.1093/bioinformatics/bts635

40 Putri GH, Anders S, Pyl PT, Pimanda JE, Zanini F. (2022). Analysing high-throughput sequencing data in Python with HTSeq 2.0. https://doi.org/10.1093/bioinformatics/btac166

41 Kassambra A. (2023). ggpubr: 'ggplot2' Based Publication Ready Plots. https://rpkgs.datanovia.com/ggpubr/

42 Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ, Anisimova M, Jakobsen KS, Linke D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-

level challenges for genome and protein databases. https://doi.org/10.1093/nar/gkz841

43 The Darwin Tree of Life Project Consortium. (2022). Sequence locally, think globally: The Darwin Tree of Life Project. https://doi.org/10.1073/pnas.2115642118