Historical RNA expression profiles from the extinct Tasmanian tiger

Emilio Mármol-Sánchez[1,2*], Bastian Fromm[1,3], Nikolay Oskolkov[4], Zoé Pochon[2,5], Panagiotis Kalogeropoulos[1], Eli Eriksson[1], Inna Biryukova[1], Vaishnovi Sekar[1], Erik Ersmark[2,6], Björn Andersson[7], Love Dalén[2,6,8#*] and Marc R. Friedländer[1#*]

[1]Department of Molecular Biosciences, The Wenner-Gren Institute, Science for Life Laboratory, Stockholm University, Stockholm, Sweden. [2]Centre for Palaeogenetics, Stockholm, Sweden. [3]The Arctic University Museum of Norway, UiT, The Arctic University of Norway, Tromsø, Norway. [4]Department of Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Lund University, Lund, Sweden. [5]Department of Archaeology and Classical Studies, Stockholm University, Stockholm, Sweden. [6]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden. [7]Department of Cell and Molecular Biology, Karolinska Institute, Stockholm, Sweden. [8]Department of Zoology, Stockholm University, Stockholm, Sweden. [#]LD and MRF contributed equally. [*]Corresponding authors: MRF, LD and EMS.

Corresponding authors:

Marc R. Friedländer: marc.friedlander@scilifelab.se

Love Dalén: love.dalen@zoologi.su.se

Emilio Mármol-Sánchez: emilio.marmol.sanchez@gmail.com

Expression hotspots and rRNA annotation

We assessed whether certain clustered regions of the thylacine genome, harboring genes predicted to be highly expressed in living cells, showed signs of enriched expression. The thylacine genome assembly (Feigin et al. 2022) was divided into consecutive non-overlapping windows of 250 kbp across all annotated scaffolds, with shorter scaffolds considered as a single independent window. The selection of this window size was motivated by 250 kbp roughly representing the length of the longest scaffold excluding the seven chromosome-like scaffolds in the thylacine assembly. UMI-deduplicated reads mapped to each defined window were quantified. To establish a background distribution of reads across the thylacine genome, public thylacine DNA sequencing data were used (SRR5055304) (Feigin et al. 2017). Thylacine paired-end aDNA sequencing data was aligned to the thylacine assembly (Feigin et al. 2022) using Bowtie 2 v.2.4.2 aligner with end-to-end very sensitive specifications (Langmead and Salzberg 2012). RNA and DNA mapping distribution were compared to identify genome expression hotspots above expectations based on DNA mappings as reference. Mapped DNA reads were subsampled to match the library size of mapped RNA reads per tissue after UMI deduplication. The coverage function from SAMtools v.1.15.1 software (Li et al. 2009) was used to assess the breadth and depth of coverage per scaffold for RNA and DNA datasets.

RNA reads mapping to the top 20 highly expressed genomic windows were investigated for overlaps with annotated genes in the thylacine assembly (Feigin et al. 2022). Unassigned reads mapping to these genomic hotspots, which did not align to any previously annotated loci, were considered potential evidence of non-annotated highly expressed genes. To identify ribosomal RNAs (rRNAs), known to be the most abundant RNA molecules in metazoan cells (Westermann et al. 2012), unassigned reads from expression hotspots were aligned to annotated rRNA sequences in humans from the miRTrace database (Kang et al. 2018), and to the 18S rRNA sequence (GEDN01046116) from the Tasmanian devil in the SILVA database release 138.1 (Quast et al. 2013). RNA reads mapped to human or Tasmanian devil rRNAs indicated the presence of unannotated rRNA genes in the thylacine genome and were used for rRNA loci annotation. Novel rRNA gene boundaries were determined by merging clusters of aligned RNA sequences with more than 100 reads.

DNA contamination inference

We investigated the origin of reads mapped to intergenic regions to identify potential DNA contamination in RNA extracts and sequencing data. Specifically, we focused on mapped reads that did not overlap with any annotated loci in the thylacine assembly (Feigin et al. 2022), considering the novel annotations of missing rRNAs and microRNAs generated in this study.

The breadth and depth of coverage for each 250 kbp genomic window were calculated using intergenic mapped RNA sequences from skeletal muscle and skin tissues, and compared to thylacine DNA data (SRR5055304) (Feigin et al. 2017) as reference. To ensure a fair comparison, DNA reads were randomly subsampled to match the number of intergenic mapping reads obtained for each tissue. Additionally, differences in read length distribution between DNA and RNA reads were accounted for by adjusting the DNA breadth of coverage per window. This adjustment involved dividing the DNA breadth of coverage by a factor calculated based on the ratio between the DNA read length (100 bp) and the average intergenic RNA read length in skeletal muscle (23.6 bp) and skin (24.6 bp), respectively. Intergenic RNA reads mapped to genomic windows with a difference of up to 2-fold in breadth and depth of coverage compared to DNA data were considered as potential DNA contamination.

Exonic enrichment and exon-exon spanning reads identification

The distribution of RNA reads mapped to annotated exonic or intronic loci from genes with reliable expression evidence (breadth of coverage of at least 10%) in the thylacine assembly was compared with DNA sequencing data (SRR5055304). Exonic quantification of RNA reads was performed from spliced mature mRNAs transcriptome-wide alignments without intronic segments to avoid the loss of reads spanning exon-exon junctions. DNA reads mapping to intronic or exonic loci, as well as RNA reads mapping to intronic loci, were quantified based on genome-wide alignments. To allow for a fair comparison, exonic/intronic DNA mapping reads were randomly subsampled to match the library size of RNA reads mapping to exonic/intronic loci after UMI deduplication for each tissue. The difference in the number of observed (RNA) versus expected (DNA) reads mapping to exonic or intronic loci was expressed as a fold change ratio using DNA-based quantification as a reference. The significance of differences in the exonic/intronic proportion of mapped reads between RNA and DNA alignments was assessed with a two-proportions z-test using the *prop.test* R function.

Reads spanning exon-exon and exon-intron junctions were determined from those mapped to annotated protein-coding genes in the thylacine genome assembly. RNA reads were considered as spanning exon-exon and exon-intron junctions only if they had at least 4 nt overhangs from the respective junctions they overlapped with.

References

Feigin C, Frankenberg S, Pask A. 2022. A chromosome-scale hybrid genome assembly of the extinct Tasmanian tiger (*Thylacinus cynocephalus*). *Genome Biol Evol* **14**: evac048.

Feigin CY, Newton AH, Doronina L, Schmitz J, Hipsley CA, Mitchell KJ, Gower G, Llamas B, Soubrier J, Heider TN, et al. 2017. Genome of the Tasmanian tiger provides insights into the evolution and demography of an extinct marsupial carnivore. *Nat Ecol Evol* **2**: 182–192.

Kang W, Eldfjell Y, Fromm B, Estivill X, Biryukova I, Friedländer MR. 2018. MiRTrace reveals the organismal origins of microRNA sequencing data. *Genome Biol* **19**: 1–15.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590–D596.

Westermann AJ, Gorski SA, Vogel J. 2012. Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology* **10**: 618–630.