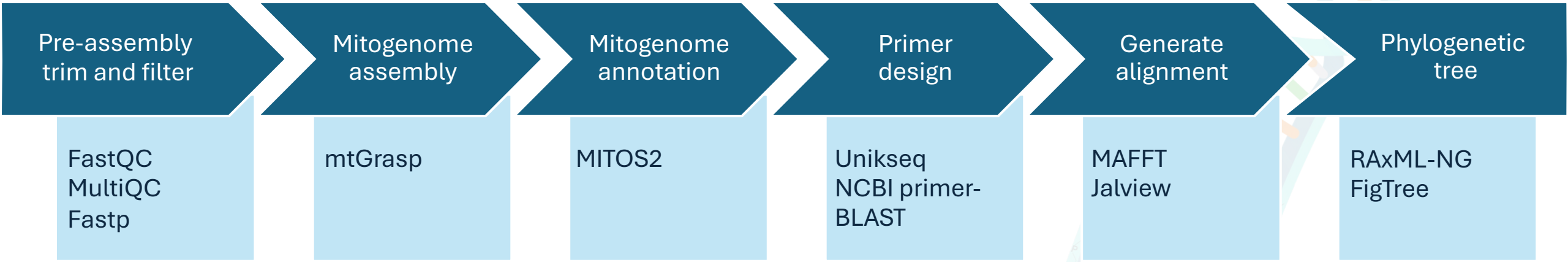
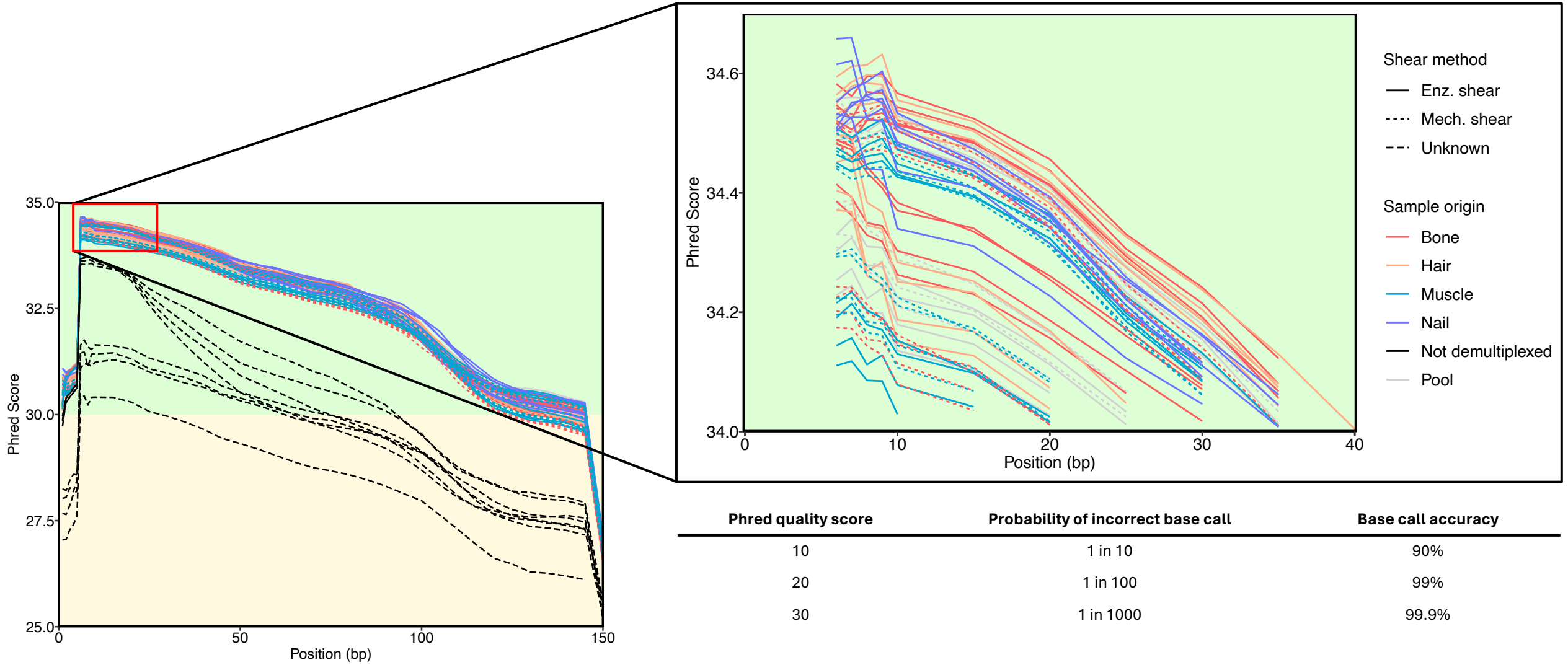
Black bear mitogenome assembly

Workflow summary – Steps and programs



Sample origin		Shear method	N samples	Total sequences (millions)	Average GC content (%)	Duplicate reads (%)	Average read length
	Bone	Enzymatic[1] Mechanical[2]	8 8	3.37 1.71	40.00 42.00	2.06 1.63	141.56 149.39
	Hair	Enzymatic[1]	8	1.92	46.00	1.31	129.12
	Muscle	Enzymatic[1] Mechanical[2]	8 8	13.94 1.64	40.75 41.00	3.44 1.60	148.30 149.46
	Nail	Enzymatic[1]	8	2.66	47.00	1.07	138.31
	Non-demultiplexed		8	1.72	42.25	13.75	145.15
	Pooled	Enzymatic[1] Mechanical[2]	8 8	2.99 2.95	40.88 41.63	2.08 2.28	147.77 149.07

Table 1. Sample overview. [1] NEBNext Ultra II FS PCR-Free Library Prep Kit. [2] Covaris - Treated with RNAse A.



Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%

Figure 1. Mean quality value across each base position in the raw reads. Each line represents a sample. Sample are color-coded according to their origin while line type represents shearing method. Value generated by FastQC (Andrews, 2010) and aggregated using MultiQC (Ewels *et al.*, 2016).

Key points

- **Quality is relatively similar** between samples regardless of origin and shear method with the exception of **non-demultiplex** samples for which the **quality** is relatively **bad**.

- Samples from **muscles** tissues generated **more sequences** than any other types of samples.

Classification	Requirement (MUST satisfy all conditions)
Reference-grade and complete	• Scaffold count = 1 • Circular • Longest sequence length >= 15kbp • No gaps • Scaffold begins with *tRNA-Phe* (start-site standardized and strand standardized) • Number of coding regions = 37
Reference-grade and near-complete	• Scaffold count = 1 • Linear • Longest sequence length >= 14,400bp • Scaffold begins with *tRNA-Phe* (start-site standardized and strand standardized) • Number of coding regions > 33
Not reference-grade	• Linear • Scaffold does not begin with *tRNA-Phe* (start-site not standardized and not strand standardized) • Scaffold count > 0
Failed assembly	• No scaffold produced

Notes on circular vs linear :

To assess whether the resulting mitogenome circular or not, *mtGrasp* attempts to computationally join the two flanking ends (start and end) of the assembled sequence either using *ABySS-Mergepairs* or *Sealer* depending on the availability of overlap between the start and end flanking regions.

Table 3. MtGrasp classification criteria of assembled mitogenomes.

Sample	Shear method	Contig length (bp)	Circular or Linear	Standardization status	Classification
gDNA-B-03-high_S5_L001	Enzymatic	16440	Linear	Non-standardized	Not reference grade
gDNA-B-03-high_S5_L002	Enzymatic	16179	Linear	Non-standardized	Not reference grade
gDNA-B-03-high_S5_L003	Enzymatic	16143	Linear	Non-standardized	Not reference grade
RNase-B-01-350bp_S3_L002	Mechanical	16066	Linear	Start site Strand Standardized	Near-complete
RNase-B-01-350bp_S3_L001	Mechanical	11520	Linear	Strand Standardized	Not reference grade
gDNA-H-04-low_S7_L001	Enzymatic	16480	Linear	Start site Strand Standardized	Near-complete
gDNA-H-04-low_S7_L002	Enzymatic	16842	Linear	Non-Standardized	Not reference grade
gDNA-H-04-low_S7_L003	Enzymatic	16588	Circular	Start site Strand Standardized	Complete
gDNA-H-04-low_S7_L004	Enzymatic	16612	Circular	Start site Strand Standardized	Complete
gDNA-M-05-high_S6_L001	Enzymatic	16618	Circular	Start site Strand Standardized	Complete
gDNA-M-05-high_S6_L002	Enzymatic	16618	Circular	Start site Strand Standardized	Complete
gDNA-M-05-high_S6_L003	Enzymatic	16658	Circular	Start site Strand Standardized	Complete
RNase-M-01-350bp_S4_L002	Mechanical	16618	Circular	Start site Strand Standardized	Complete
RNase-M-01-350bp_S4_L003	Mechanical	16624	Circular	Start site Strand Standardized	Complete
gDNA-N-03-low_S8_L001	Enzymatic	16269	Linear	Start site Strand Standardized	Near-complete
gDNA-N-03-low_S8_L002	Enzymatic	16200	Circular	Start site Strand Standardized	Complete
gDNA-N-03-low_S8_L003	Enzymatic	16305	Circular	Start site Strand Standardized	Complete
Bear-12112021-pool-350bp-S1_L002	Mechanical	16618	Circular	Start site Strand Standardized	Complete
Bear-12112021-pool-350bp-S1_L003	Mechanical	16590	Circular	Start site Strand Standardized	Complete
Bear-12112021-pool-S2_L002	Enzymatic	16630	Circular	Start site Strand Standardized	Complete
Bear-12112021-pool-S2_L003	Enzymatic	16618	Circular	Start site Strand Standardized	Complete

Table 2. Mitogenome assembly summary output generated by mtGrasp (*Yang and Coombe, 2023*).

Key points

- While all **pooled** and **muscle** samples generated **reference-grade and complete mitogenome**, **bone** samples generated only incomplete and **poor quality mitogenome**.

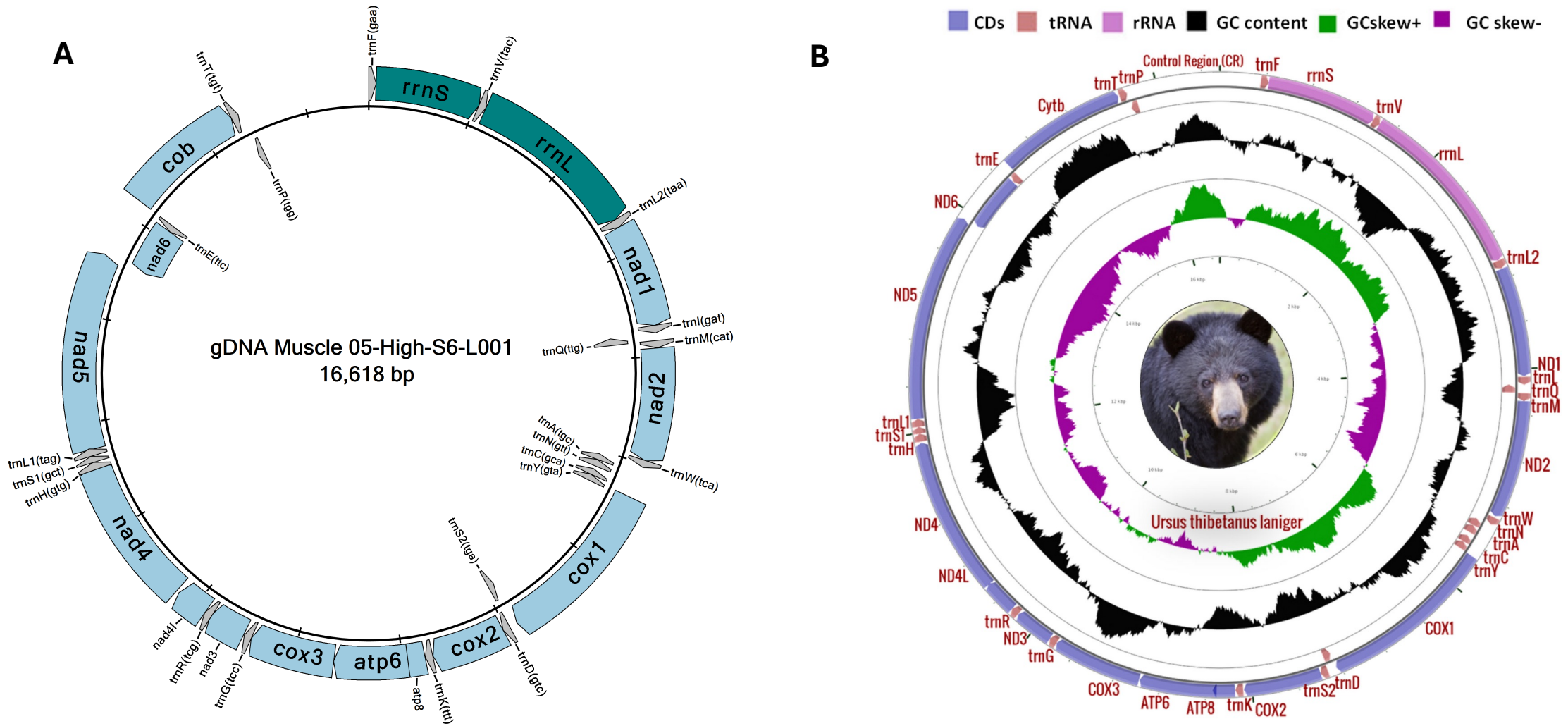


Figure 2. Reconstruction of the mitochondrial genome (**A**) of *Ursus americanus* using sample M-05-high-S6-L001. Protein-coding genes are shown in light blue, rRNA genes in green and tRNA in grey. Genome annotation was generated using MITOS2 (Donath *et al*., 2019) and circularMT (Goodman and Car, 2024) was used to draw the figure. (**B**) Mitochondrial genome of *Ursus thibetanus laniger* generated by Bit *et al.* (2021) for comparaisons.

Key points

- Assembled mitogenome of muscle sample 05-high-S6-L001 is **similar to other assembly from literature**.
- The mitogenome was **16,618 bp in length** and encoded the typical **13 protein-coding genes**, **2 ribosomal RNA genes** and **22 transfer RNA genes**.

Sample name	Start position	End position	Length	Protein coding gene
gDNA-B-03-high_S5_L001 – Non-Standardized	9286	9619	333	*cox3*
gDNA-H-04-low_S7_L001 – StartSite Strand Standardized Linear	8686	9019	333	*cox3*
gDNA-H-04-low_S7_L004 – StartSite Strand Standardized Circular	8686	9019	333	*cox3*
gDNA-M-05-high_S6_L001 – StartSite Strand Standardized Circular	8686	9019	333	*cox3*
gDNA-N-03-low_S8_L001 – StartSite Strand Standardized Linear	8686	9019	333	*cox3*
gDNA-B-03-high_S5_L001 – Non-Standardized	15138	15387	249	*cob*
gDNA-H-04-low_S7_L001 – StartSite Strand Standardized Linear	14538	14787	249	*cob*
gDNA-H-04-low_S7_L004 – StartSite Strand Standardized Circular	14538	14787	249	*cob*
gDNA-M-05-high_S6_L001 – StartSite Strand Standardized Circular	14538	14787	249	*cob*
gDNA-N-03-low_S8_L001 – StartSite Strand Standardized Linear	14538	14787	249	*cob*

Table 4. Unique sequence identified by *unikseq* (Allison *et al.*, 2023).

Cytochrome-c oxidase (*cox3*) gene

Forward primer Sequence (5'->3')	Forward primer Length	Forward primer Tm	Forward primer GC%	Reverse primer Sequence (5'->3')	Reverse primer Length	Reverse primer Tm	Reverse primer GC%	Product length
AGCCCTCTCAGCCCTTCTTA	20	59.96	55	GGTGGCCCTGAAAGGTACTC	20	60.04	60	152
GAGAGTACCTTTCAGGGCCAC	21	60.07	57.14	CTGGAGTTGGTGCTAGGCTT	20	59.67	55	142
GCCCTCTCAGCCCTTCTTATG	21	60.2	57.14	TCTGGAGTTGGTGCTAGGCT	20	60.55	55	272
AGCCCTCTCAGCCCTTCTTAT	21	60.34	52.38	GGTGCTAGGCTTGAGTGGTA	20	59.1	55	264
GAGAGTACCTTTCAGGGCCA	20	58.72	55	GAGTTGGTGCTAGGCTTGAGT	21	60	52.38	139
TCGCAGGATTTTTCTGAGCC	20	58.55	50	GGTCAGCATGCTCCCAGTTC	20	61.03	60	70
GAGCCCTCTCAGCCCTTCTTA	21	60.97	57.14	TGTGGTGGCCCTGAAAGGTA	20	61.06	55	156
CCACTCAAGCCTAGCACCAA	20	59.96	55	GGCACTTCTAGTGGATTTAGGGG	23	60.43	52.17	84
ACACCCCTGTTGTCCAAAAA	20	57.76	45	CTCCCAGTTCTGGAGTTGGTG	21	60.27	57.14	129
GCCTAGCACCAACTCCAGAA	20	59.68	55	GCACTTCTAGTGGATTTAGGGGA	23	59.55	47.83	75

Table 5. Primer pair options 1 generated by primer BLAST (Ye *et al.*, 2012) using longest (333 bp) unique sequence identified by *unikseq* (Allison *et al.*, 2023). Sequence identified by *unikseq* corresponds to a region of the mitochondrial *cox3* gene. Primer pairs with one or more primer sequence also found in other bear species are highlighted in grey.

Cytochrome b (*cob*) gene

Forward primer Sequence (5'->3')	Forward primer Length	Forward primer Tm	Forward primer GC%	Reverse primer Sequence (5'->3')	Reverse primer Length	Reverse primer Tm	Reverse primer GC%	Product length
GGATCTGAGGGGGCTTTTCT	20	59.08	55	TGGACTGCTGCTAGTGTCAA	20	58.95	50	100
TGAGGGGGCTTTTCTGTGAA	20	59.15	50	TAGGTGGACTGCTGCTAGTG	20	58.81	55	99
GCCCTGAGGCCAAATATCCT	20	59.52	55	AGAAAAGCCCCCTCAGATCC	20	59.08	55	109
CTGCCCTGAGGCCAAATATC	20	58.39	55	TTCACAGAAAAGCCCCCTCA	20	59.15	50	116
GAATGGATCTGAGGGGGCTT	20	59.15	55	GGACTGCTGCTAGTGTCAAG	20	58.28	55	103
AGGATATGTCCTGCCCTGAG	20	58.27	55	AAAGCCCCCTCAGATCCATT	20	58.69	50	118
TGCCCTGAGGCCAAATATCC	20	59.81	55	AAAAGCCCCCTCAGATCCAT	20	58.69	50	108
AGCCACCGCATTCATAGGATA	21	59.02	47.62	GGCTGATAGGAGGTTGGTGA	20	58.8	55	85
TCACCAACCTCCTATCAGCC	20	58.8	55	GAATCGTGTCAGAGTTGCCT	20	57.91	50	98
ATGGATCTGAGGGGGCTTTT	20	58.69	50	AGGTGGACTGCTGCTAGTG	19	59.02	57.89	105

Table 6. Primer pair options 1 generated by primer BLAST (Ye *et al.*, 2012) using second-longest (249 bp) unique sequence identified by *unikseq* (Allison *et al.*, 2023). Sequence identified by *unikseq* corresponds to a region of the mitochondrial *cob* gene. Primer pairs with one or more primer sequence also found in other bear species are highlighted in grey.

Mitochondrially encoded ATP synthase membrane subunit 6 (*atp6*) gene

Forward primer Sequence (5'->3')	Forward primer Length	Forward primer Tm	Forward primer GC%	Reverse primer Sequence (5'->3')	Reverse primer Length	Reverse primer Tm	Reverse primer GC%	Product length
ATCTTCTGGGTCTGTTGCCA	20	58.93	50	ATAGCAACTGTGCCTGTCCA	20	59.31	50	97
CACACTCATTCACACCCACC	20	58.76	55	GATAGCAACTGTGCCTGTCC	20	58.63	55	80
GCTCGACCAATCTTCTGGGT	20	59.75	55	TGCCTGTCCACAAGGGAATA	20	58.63	50	96
ACACTCATTCACACCCACCA	20	59.16	50	GTCCGATAGCAACTGTGCCT	20	60.11	55	83
CTCATTCACACCCACCACAC	20	58.76	55	AGTCCGATAGCAACTGTGCC	20	60.11	55	81
TGGCTCGACCAATCTTCTGG	20	59.75	55	CCTGTCCACAAGGGAATAGC	20	57.96	55	96
GGTCTGTTGCCACACTCATT	20	58.39	50	TAGCAACTGTGCCTGTCCAC	20	60.25	55	88
AAAGGACAAACCTGAGCACT	20	57.26	45	TGGTGGGTGTGAATGAGTGT	20	59.16	50	97
TCGACCAATCTTCTGGGTCT	20	58.06	50	GCCTGTCCACAAGGGAATAG	20	57.96	55	93
CGACCAATCTTCTGGGTCTG	20	57.98	55	TGCCTGTCCACAAGGGAAT	19	58.84	52.63	93

Table 7. Primer pair options 1 generated by primer BLAST (Ye *et al.*, 2012) using third-longest (225 bp) unique sequence identified by *unikseq* (Allison *et al.*, 2023). Sequence identified by *unikseq* corresponds to a region of the mitochondrial *atp6* gene. Primer pairs with one or more primer sequence also found in other bear species are highlighted in grey.

Key points

- Many options for primer design targeting different protein coding genes from the mitogenome.

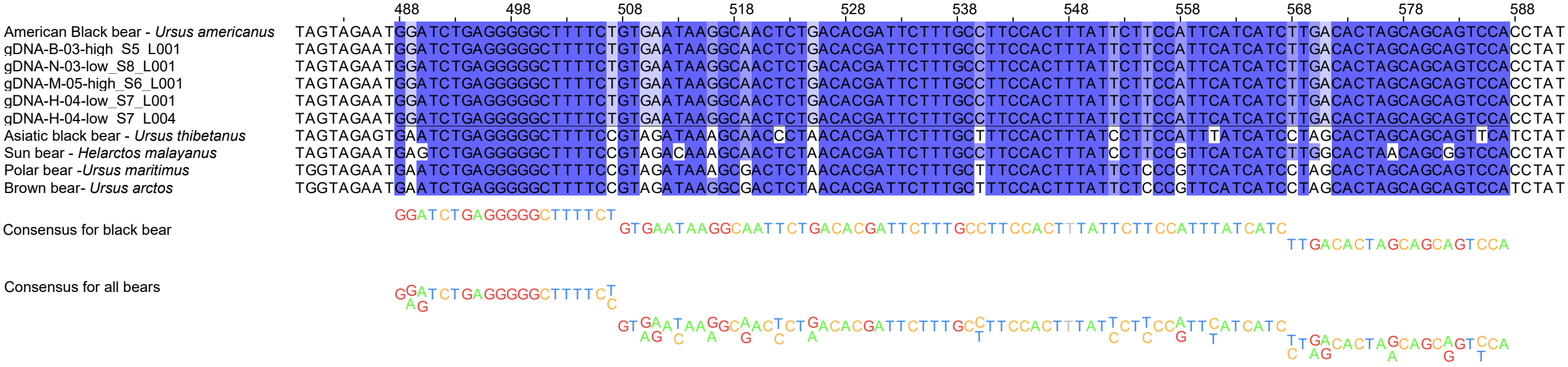


Figure 3. MAFFT alignment (Katoh *et al.*, 2013) of a specific region of the cob gene (488 - 587) identified as a potential target sequence by *unikseq* (Allison *et al.*, 2023) and primer BLAST (Ye *et al.*, 2012). Nucleotides are colored according to the percentage of nucleotides in each columns that agree with the consensus for all sequences. Figure was generated using Jalview (Waterhouse *et al.*, 2009).

Key points

- Alignment provides a visual of **differences in nucleotides** in the *cob* DNA region targeted by the primer pair GGATCTGAGGGGGCTTTTCT and TGGACTGCTGCTAGTGTCAA between different species of bear.

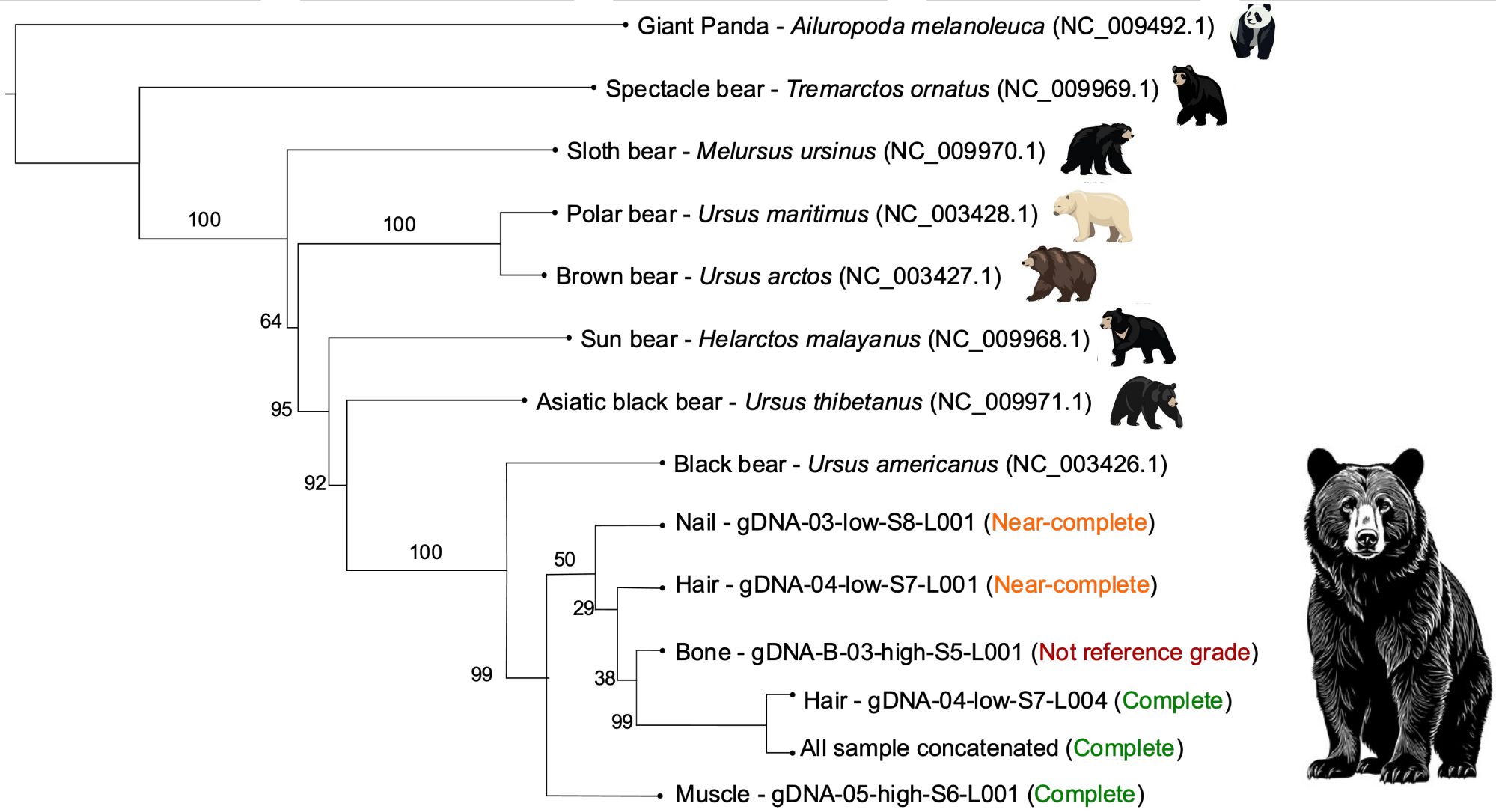


Figure 4. Maximum likelihood phylogenetic tree of family *Ursidae* including provided samples. Sequences were aligned using MAFFT (Katoh and Standley, 2013) after which tree was inferred using RAxML-NG maximum likelihood with 999 bootstraps (Kozlov *et al.*, 2019). FigTree (Rambaut, 2010) was used to visualise and edit tree.

Key points

- Phylogenetic placement of provided samples is **consistent with literature**.

General conclusion

- While the quality of all samples was relatively similar, **muscle** and **pooled samples generated higher quality assembled mitogenomes.**

- We found **many different primer options** which target different protein coding genes from the bear mitogenome. Effectiveness of different options should be tested with PCR.

References

Allison, M. J., Warren, R. L., Lopez, M. L., Acharya-Patel, N., Imbery, J. J., Coombe, L., Yang, C. L., Birol, I., & Helbing, C. C. (2023). Enabling robust environmental DNA assay design with “unikseq” for the identification of taxon-specific regions within whole mitochondrial genomes. *Environmental DNA*, 5, 1032–1047. https://doi.org/10.1002/edn3.438

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data.

Bit A, Thakur M, Singh SK, Joshi BD, Singh VK, Sharma LK, Tripathy B, Chandra K. (2021) Assembling mitogenome of Himalayan Black Bear (U. t. laniger) from low depth reads and its application in drawing phylogenetic inferences. *Scientific Report*, 11(1), 730. https://doi.org/10.1038/s41598-020-76872-y

Ewels P, Magnusson M, Lundin S, Käller M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 32(19), 3047-8. https://doi.org/10.1093/bioinformatics/btw354

Goodman, S. J. et Carr, I. M. (2024). Drawing mitochondrial genomes with circularMT. *Bioinformatics*, 40(7). https://doi.org/10.1093/bioinformatics/btae450

Katoh K, Standley DM. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772-80. https://doi.org/10.1093/molbev/mst010

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. et Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21), 4453-4455. https://doi.org/10.1093/bioinformatics/btz305

Matthias, B., Alexander, D., Frank, J., Fabian, E., Catherine, F., Guido, F., Joern, P., Martin, M. et Peter, F. S. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, 69(2), 313-319. https://doi.org/10.1016/j.ympev.2012.08.023

Rambaut, A. (2018) Figtree ver 1.4.4. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh

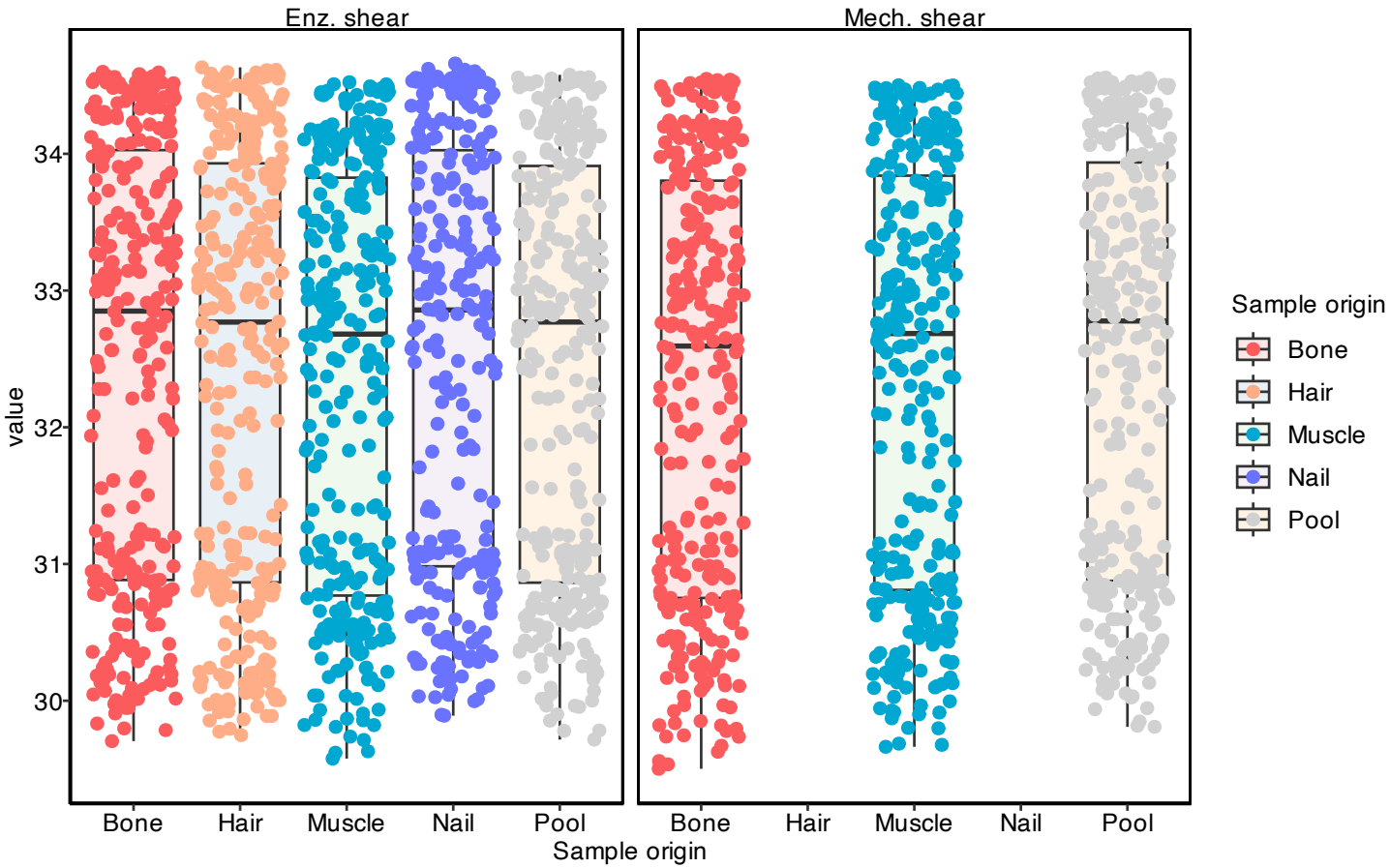
Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M., Barton, G.J. (2009) Jalview version 2: A Multiple Sequence Alignment and Analysis Workbench. *Bioinformatics,* 25(9), 1189-1191. https://doi.org/10.1093/bioinformatics/btp033

Yang, L. (2023). Streamlined high throughput assembly and standardization of reference-grade animal mitochondrial genomes (T). University of British Columbia

Ye, J., Coulouris, G., Zaretskaya, I. *et al.* (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13, 134. https://doi.org/10.1186/1471-2105-13-134

Black bear mitogenome assembly - Supplementary

Quality based on sample origin and shearing method



1. Shapiro test to evaluate if distribution of value is similar to normal distribution
   - Assumption of normality not met ($p < 0.05$)
   - Non-parametric tests used
2. Kruskal-Wallis test to compare all samples
   - Not significant ($p > 0.05$)
3. Wilcoxon test to compare shear method
   - Not significant ($p > 0.05$)

mtGrasp (A) circularization and (B) standardization workflow