GigaScience

Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing

--Manuscript Draft--

Manuscript Number:	GIGA-D-21-00273R1
Full Title:	Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing
Article Type:	Research
Funding Information:	Howard Hughes Medical Institute (institutional funds) — Dr. Erich D. Jarvis Rockefeller University (startup funds) — Dr. Erich D. Jarvis Max Planck Institute of Molecular and Cell Biology and Genetics (institutional funds) — Dr. Sylke Winkler Wellcome Trust (WT207492) — Dr. Iliana Bista Wellcome Trust (104640/Z/14/Z) — Dr. Iliana Bista Wellcome Trust (092096/Z/10/Z) — Dr. Iliana Bista
Abstract:	Studies in vertebrate genomics require sampling from a broad range of tissue types, taxa, and localities. Recent advancements in long-read and long-range genome sequencing have made it possible to produce high-quality chromosome-level genome assemblies for almost any organism. However, adequate tissue preservation for the requisite ultra-high molecular weight DNA (uHMW DNA) remains a major challenge. Here we present a comparative study of preservation methods for field and laboratory tissue sampling, across vertebrate classes and different tissue types. We find that no single method is best for all cases. Instead, the optimal storage and extraction methods vary by taxa, by tissue, and by down-stream application. Therefore, we provide sample preservation guidelines that ensure sufficient DNA integrity and amount required for use with long-read and long-range sequencing technologies across vertebrates. Our best practices generated the uHMW DNA needed for the high-quality reference genomes for Phase 1 of the Vertebrate Genomes Project (VGP), whose ultimate mission is to generate chromosome-level reference genome assemblies of all ~70,000 extant vertebrate species.
Corresponding Author:	Olivier Fedrigo The Rockefeller University New York, UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	The Rockefeller University
Corresponding Author's Secondary Institution:	
First Author:	Hollis A. Dahn
First Author Secondary Information:	
Order of Authors:	Hollis A. Dahn Jacquelyn Mountcastle Jennifer Balacco Sylke Winkler Iliana Bista Anthony D. Schmitt

	Olga Vinnere Pettersson
	Giulio Formenti
	Karen Oliver
	Michelle Smith
	Wenhua Tan
	Anne Kraus
	Stephen Mac
	Lisa M. Komoroske
	Tanya Lama
	Andrew J. Crawford
	Robert W. Murphy
	Samara Brown
	Alan F. Scott
	Phillip A. Morin
	Erich D. Jarvis
	Olivier Fedrigo
Order of Authors Secondary Information:	
Response to Reviewers:	Dear Dr. Hans Zauner, Thank you for the opportunity to revise and resubmit our article titled “Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing” to GigaScience. We greatly appreciate the effort put forth by you and the reviewers in providing feedback on our manuscript. The comments we received are insightful suggestions that we feel have improved the manuscript. We have been able to incorporate changes based on these contributions, as highlighted below. Comments from Reviewer 1: R. 1, Comment 1: In figure 2 the size distribution of DNA fragments is visualized from the different experiments. Most of the fragment distributions look like I would have expected them based on the work we did in the article cited as nr 25 in the reference list. However the muscle tissue from rats and the blood samples from the mouse and the frog indicates that there may be a misinterpretation in the article regarding the actual size distribution of fragments which needs to be looked in to. Response: The following comments indicate two major limitations to this methodology: the interpretable range of fragment sizes, and streaking artifacts. In the article we endeavor to interpret these results within the bounds of those limitations. We’ve added additional text and made figure modifications, detailed below, to clarify these caveats. R. 1, Comment 2: Starting with the mouse plots and especially the muscle one. There must either have been a physical shearing event that drastically reduced the size of DNA (using the terminology from ref 25 this would mean that physical shearing generated a characteristic fragment length of approximately 300-400 kb), or the lack of a sharp slope on the rightmost side of the ridgeline plot is due to the way the image was processed. All other animals got a peak on the rightmost side of the ridgeline plot and the agarose plug should, based on the referenced methods paper [7], generate megabase sized fragments which far exceed the size of the scale used in figure 2. I would presume these larger fragments would get stuck in or near the well which makes it easy to accidentally cut them out when doing the image analysis step which may explain their absence in the mouse samples. This leads me to the conclusion that the article is well designed to capture the impact of chemical shearing caused by different preservation methods but would benefit from evaluating whatever figure 2 properly covers the actual size distribution of fragments or only covers the portion of DNA fragments small enough to actually form bands on the PFGE gel with a substantial part of the DNA stuck in or near the

well.
Response: This is a correct indication of the limits of visualizing DNA fragment length distributions with current technologies, only fragments that fall within a given range can be viewed at once with any reasonable resolution. We targeted a range of fragment sizes that includes ideal sizes for long-read and long-range sequencing while also giving indications of degradation in the smaller end of the range. Fragments outside the target range can not be reliably interpreted. In PFGE, DNA fragments larger than the target range can be stuck in or near the well (right side of the plots). Unfortunately, in the well they then become indistinguishable from bright reflections off the edges of the well and are further obscured when streaking is present. We attempted early on to score samples for presence/absence of signal in or near the well, but found this to not be repeatable or informative using the gel images. Thus, we chose to omit the well itself and the space immediately around it from interpretation, from measurements for statistical testing, and from Fig. 2. We only interpret the portion of the PFGE gel where bands can form and where the standard can give us some indication of fragment size. The interpretable range of fragment sizes in these gels still offers important indicators of sample quality.
For clarity, the new version of Fig. 2 includes the well in each plot profile. The well peaks are cropped where they exceed the peak brightness of the rest of the lane. We do not consider the well brightness to be a reliable indicator of sample quality, but recognize that some readers may want to see the full pattern.
The following has also been added to the Fig. 2 legend: “Fluorescent stained DNA fragments are drawn with an electric current from the well at the right towards the left. Smaller fragments generally travel farther than larger fragments. The fragments that greatly exceed the targeted size range remain in the well and can not be reliably interpreted.”
“The well brightness is cropped where it exceeds the brightness of the rest of the gel lane.”
We have also added this text to the discussion section of the manuscript: “Additionally, we are only able to visualize DNA fragment size distributions within a certain range of sizes (approx. 40–400 kb for PFGE, 1.3–165 kb for FEMTO). Though we have targeted a size range that includes both ideal fragment sizes for long-read sequencing and fragments of lower molecular weight that may indicate degradation, fragments outside this range are not measured here.”

R. 1, Comment 3: The frog plot is a good example of how this may influence our interpretation of the ridgeline plots. If the extraction method generate high-quality DNA concentrated in the 300-400 kb range then there must be something very special with the frog DNA from blood as there is a continuous increase in the brightness all the way to the edge of the image. This implies that the sample contains a high amount of much larger DNA fragments than the other samples. I find this rather unlikely and if I saw this in my own data I would assume that we had a lot of very large DNA fragments that are out of scale for the gel electrophoresis but that in the case for the frog blood samples many of these fragments have been chemically sheared creating the "smeared" pattern we see in figure 2.
Response: Yes, the frog blood especially exhibits a “streaking” pattern in the gel where there is a strip of continuous brightness in the lane. This is another reason why we do not attempt to interpret the gel above where bands form in the lane. Before our initial submission, we performed a repeat run of the PFGE gels with streaking, but they produced identical results with streaks still present. Samples with this streaking pattern have performed well in past sequencing efforts, and it’s generally thought to be an indicator of high quality samples. However, barring more conclusive testing of this pattern, we do not attribute streaking as an indicator of quality in this manuscript.
We have added this section to the figure caption: “DNA fragments with lengths longer or shorter than peaks of the size standard can not be reliably interpreted due to lack of size reference and artifacts of gel electrophoresis as well as limitations of any type of gel electrophoresis to correctly size megabase-length fragments.”

R. 1, Comment 4: Dryad DOI doesn't work for me.
Response: It is possible that the reviewer was attempting to open the link provided in the cover letter, which has formatting mistakes related to the uploading process. A correct link was supplied to Dr. Zauner. Apologies for this inconvenience. Here is the correct and updated Dryad download link for reviewers:
https://datadryad.org/stash/share/uHgVucrNICiMT-

Y92O4M4Km3S4DyK3UJFA3qMJEbm4M

R. 1, Comment 5: Figure 1 - The meaning of x3 and x2 for the turtle should be described in the caption.
Response: The Figure 1 caption now has the added sentence “For the sea turtle samples, cells with numbers (x2 or x3) indicate conditions where samples from more than one individual were processed for comparison.”

R. 1, Comment 6: Figure 2 - Having the scale indicator (48.5. 145.5 etc) at the top as well as the bottom of each column would make it quicker to estimate the distribution of samples.
Response: Agreed. This has been added to Figure 2.

R. 1, Comment 7: The article completely omits Nanopore sequencing, is there a specific reason for why lessons here are not applicable to ONT?
Response: Nanopore equipment and expertise were not available at the time of this study, but will be part of further testing. We generally expect the same indicators of sample quality to correlate with successful sequencing in Nanopore sequencing as with other technologies, though it is not explicitly tested here. We’ve added a mention of Oxford Nanopore to the list of relevant technologies in the introduction to clarify: “Long-reads (generally > 10 kb; e.g. Pacific Biosciences or Oxford Nanopore), long-range molecules (generally > 50 kb; e.g. 10X Genomics linked reads), or optical mapping (> 150 kb; e.g. Bionano Genomics), and Hi-C proximity ligation (> 1 Mb; e.g. Arima Genomics) can span repeats thousands of base pairs in length [4], greatly improving assembly outcomes.”

R. 1, Comment 8: There is a very interesting paragraph starting with "The ambient temperature of the intended collecting locality should be a major consideration in planning field collections for high-quality samples. Here we test a limited number of samples at 37°C to". Even if the results were very poor information about the failed conditions would be appreciated. What tissues/animals did you use, did you do any preservation at all for the samples and did you measure the fragment length distribution anyway? Simply put, even if the DNA was useless for long read sequencing it is an interesting data point for the dynamics of DNA degradation and a valuable lesson for planning sampling in warm climates.
Response: The “limited number of samples” refers to the four mouse muscle samples reported with the rest of our results. No further samples were tested at 37°C in this study. We’ve modified that sentence for clarity as follows: “Here we test a limited number of samples at 37°C to resemble fieldwork conditions in warmer climates, resulting in no retention of workable amounts of uHMW DNA in any of these samples (4 mouse muscle samples; Fig. 2).”

Comments from Reviewer 2:
R. 2, Comment 1: Although the effectiveness of the tissue/preservative combination was only tested with the preparation of long range libraries, it would have been useful to select one or two cases for long range sequencing (PacBio or Oxford Nanopore) to explore the impact of the different QC parameters measured in this study.
Response: We agree that testing samples on long-read sequencing platforms would have been very useful. Unfortunately, the expense of long-read sequencing was prohibitive at the time. We do find that the results align generally with the experience of the Vertebrate Genomes Project. See response to the related comment #4 below for further details.

R. 2, Comment 2 (in text): space between quantity and unit symbol
Response: This change has been applied to the text.

R. 2, Comment 3 (in text): such as used twice in a sentence. please edit
Response: This sentence has been revised.

R. 2, Comment 4 (in text): Your work is a great contribution to the genomics field! However, DNA integrity and optimal QC parameters (Absorbance ratios at 260/230, 260/280, double stranded DNA proportion from a total gDNA prep) are not always predictors for Long Read Sequencing success. I am aware of the high cost that you would face if all these samples were sequenced, even one flowcell/sample in minION,

could cost a little fortune. But it would be fantastic if you could please indicate if any of the sample/preservation combinations showed consistent good LRSeqeuncing results.
Response: We concur that simply checking the DNA integrity and yield as QC parameters might not give a definitive indicator of sequencing success with long reads. Moreover, this particular paper has been concentrating on Vertebrata, a relatively small taxonomic group as compared to Arthropoda, Planta, etc.

Additionally, as new reagents for DNA preservation are constantly emerging on the market (e.g. Allprotect), sample collection committees within VGP, EBP, ERGA and other large reference genome sequencing initiatives will continue monitoring and studying the impact of those products on the sequencing outcome, but those would be subjects of other studies, publications, and public guidelines.

From the experience of several large PacBio sequencing providers, including the facilities involved in this publication, we know that the chemical purity of the HMW-DNA sample is at least as important as the molecule integrity. There is, however, no single definitive analytical parameter that has ever been defined for predicting the long-read sequencing outcome based, or for detection of any carry-over contaminants or significance to sequencing success. For this reason, chemical purity parameters were not within the scope of this manuscript, though they likely carry influence outside the variables manipulated in this study.

We have instead concentrated this study on the low-hanging fruit of integrity of HMW-DNA molecules and DNA yield, both of which robustly and intuitively influence sequencing success. From that point of view, our smaller sequencing tests and general experience outside the scope of this manuscript corroborate the results presented in the study. Two of these smaller tests are detailed below.

Recent test 1: Flash-frozen vs. EtOH-preserved reptile tissue
We compared the length of uHMW gDNA and performance on Bionano and PacBio continuous long reads of snap-frozen and EtOH-preserved nucleated blood from reptiles and found no significant changes in performance in Bionano molecule size and PacBio CLR sequencing subread length. The average length of unfiltered Bionano reads was longer when gDNA was extracted from flash-frozen tissue, though both treatments still returned results in an acceptable range. In general, Bionano optical mapping was working reliably for gDNA from EtOH-preserved tissues.
Some of the quality specifications are shown in the table in the attached "Response to reviewers" document.

Recent test 2:
We used mammalian kidney tissue that had been stored in Allprotect recently for PacBio HiFi and Hi-C (ARIMA protocol). Right after tissue extraction, the tissue was soaked in Allprotect at room temperature overnight, stored for about 1 year at -20C degrees, then shipped, and finally stored at -80 degrees before sequencing. The sample showed no indication of any contaminants on a Nanodrop spectra readout. Please see the PFGE gel image and summarized sequencing results in the attached "Response to reviewers" document.

R. 2, Comment 5 (in text): Did you load the same amount of DNA per PFGE lane for all samples?

Did you heat up the DNA sample + loading buffer before loading? 1-2 min at 65C followed by cooling at room temperature ~ 5 min before loading helps to prevent clumping.

Did you run a slice of each plug as a control for the DNA manipulation factor that could cause fragmentation while extracting the DNA from the plug?
Response:
Yes, we loaded approximately 100 ng of DNA per well.

No, we did not heat DNA and loading buffer prior to loading. DNA, loading buffer, and TE buffer were kept at room temperature before mixing and then loading. DNA was loaded after at least one week at room temperature, which allows for homogenization

of the sample and increases hydration of the DNA molecules in the sample. Although the original HMW DNA samples are often viscous, the addition of TE buffer and loading buffer dilutes them to the point that we find clumping is reduced.

No, each plug was carried through entirely for DNA extraction. Digesting the entire plug was integral to comparing DNA yield. The process to extract DNA from the plug is quite gentle (slow shaking for lysing and washing steps, Agarase digestion, drop dialysis) and the plug also protects the DNA.

R. 2, Comment 6 (in text): The picture shows embedded nuclei with or without crosslinker, but the same could be done with the extracted DNA from plugs
Response:
We absolutely agree that the conventional field-inversion PFGE instruments (e.g. BioRad CHEF) are much better suited to resolve sub-megabase size DNA fragments from in situ extractions. However, this specialized equipment is currently much less common as compared to a cheaper (albeit less precise) Pippin Pulse system from SAGE which is widely used at sequencing facilities.

R. 2, Comment 7 (in text): Perhaps analyzing a sample in a standard agarose gel might help evaluate fragmentation < 20 kbp.
Response: This is certainly something that could be incorporated into future testing as another metric of the smaller fragment sizes, perhaps as a cheaper alternative to FEMTO. Unfortunately, more testing at this point on the same extractions would not be useful; they are now several years old. Our subset of samples tested on the FEMTO Pulse system give a detailed perspective on <20 kb fragments.

R. 2, Comment 8 (in text): It might help to describe briefly how Arima prepared the nucleated blood Hi-C libraries, especially the samples preserved in DNAgard. I think this solution does contain an inhibitor of the crosslinking reaction, most likely a free amino group (from Tris buffer, for example). The Dovetail Genomics Omni-C kit protocol dilutes the nucleated blood preserved in EDTA tubes in 1 mL 1X PBS buffer and collects the cells by centrifugation before the crosslinking steps. Sorry I don't have access to the Arima protocol document to make a more informed comment.
Response: Our procedure for nucleated blood in a solution like ethanol (or DNAgard) is to pellet the cells, remove the supernatant, wash with 1X PBS containing 1% FBS, and then carry the washed pelleted cells into crosslinking and then Arima-HiC. Given our washing procedure, it seems less likely (although still possible) that residual tris is inhibiting the crosslinking reaction. Interestingly, DNAgard is a proprietary solution originally developed by Biomatrica, and so we were not able to find any resource that pertains to what the solution is actually composed of.
We've added a citation of the protocol document number and this note to the methods: "Briefly, standard protocol for nucleated blood in a solution like EtOH or DNAgard is to pellet the cells, remove the supernatant, wash with 1X phosphate buffered saline solution containing 1% Fetal Bovine Serum, and then carry the washed pelleted cells into crosslinking and then Arima-HiC.", and this to the discussion: "Though our washing protocol should minimize its effect, it is also possible that some unknown aspect of the DNAgard treatment of cells inhibited the crosslinking reaction, and Hi-C of unfixed cells would be expected to have low signal and high noise similar to degraded DNA."

Additional clarifications:

INSDC submission - We have removed the Hi-C sequence reads from the manuscript's associated Dryad repository and are in the process of uploading them to the publicly accessible Sequence Read Archive.
Analysis scripts - Two commented scripts have been added to the manuscript's associated Dryad repository. One contains statistical analysis of DNA yield and fragment length reported in the manuscript, and the other has basic bioinformatics and calculations based on the Hi-C reads reported in the manuscript.
The data availability section of the manuscript has also been updated to reflect these changes.

Please feel free to notify me of any further comments or questions. We look forward to

	your response. Thank you again for your hard work and for this opportunity. Sincerely, Olivier Fedrigo, Ph.D. Director Vertebrate Genome Laboratory
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be	Yes

either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	

Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing

Hollis A. Dahn[1†], Jacquelyn Mountcastle[2†], Jennifer Balacco[2], Sylke Winkler[3], Iliana Bista[4,5], Anthony D. Schmitt[6], Olga Vinnere Pettersson[7], Giulio Formenti[2], Karen Oliver[4], Michelle Smith[4], Wenhua Tan[3], Anne Kraus[3], Stephen Mac[6], Lisa M. Komoroske[8], Tanya Lama[8], Andrew J. Crawford[9], Robert W. Murphy[1], Samara Brown[2], Alan F. Scott[10], Phillip A. Morin[11], Erich D. Jarvis[2,12], Olivier Fedrigo[2*]

1 University of Toronto, Toronto, Ontario, Canada

2 The Rockefeller University, New York, New York, United States

3 Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Saxony, Germany

4 Tree of Life program, Wellcome Sanger Institute, Hinxton, Cambridgeshire, United Kingdom

5 University of Cambridge, Cambridge, Cambridgeshire, United Kingdom

6 Arima Genomics, Inc., San Diego, California, United States

7 National Genomics Infrastructure, SciLifeLab, Uppsala University, Uppsala, Sweden

8 University of Massachusetts Amherst, Amherst, Massachusetts, United States

9 Department of Biological Sciences, Universidad de los Andes, Bogotá, 111711, Colombia

10 Johns Hopkins University, Baltimore, Maryland, United States

11 Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, La Jolla, California, United States

12 Howard Hughes Medical Institute, Chevy Chase, Maryland, United States

*email: ofedrigo@rockefeller.edu

[†]co-first authors

Abstract

Studies in vertebrate genomics require sampling from a broad range of tissue types, taxa, and localities. Recent advancements in long-read and long-range genome sequencing have made it possible to produce high-quality chromosome-level genome assemblies for almost any organism. However, adequate tissue preservation for the requisite ultra-high molecular weight DNA (uHMW DNA) remains a major challenge. Here we present a comparative study of preservation methods for field and laboratory tissue sampling, across vertebrate classes and different tissue types. We find that no single method is best for all cases. Instead, the optimal storage and extraction methods vary by taxa, by tissue, and by down-stream application. Therefore, we provide sample preservation guidelines that ensure sufficient DNA integrity and amount required for use with long-read and long-range sequencing technologies across vertebrates. Our best practices generated the uHMW DNA needed for the high-quality reference genomes for Phase 1 of the Vertebrate Genomes Project (VGP), whose ultimate mission is to generate chromosome-level reference genome assemblies of all ~70,000 extant vertebrate species.

Keywords: long-read sequencing, genome assembly, tissue preservation, HMW DNA extraction

Introduction

The past two decades have seen genome sequencing become increasingly easy and affordable, driven by advancements in sequencing and computing technologies. Growing accessibility spurred the formation of large-scale consortia, such as the Genome 10K project (G10K), with the goal of generating genome assemblies for many species to enable new scientific discoveries and aid in conservation efforts [1]. However, initial efforts used short read sequencing (< 200 bp), such

as Illumina technology, which were later found to often result in genome assemblies that were highly fragmented, incomplete, and plagued with structural inaccuracies [1–3]. Subsequently, G10K initiated the Vertebrate Genomes Project (VGP), with the mission of producing high-quality, near-complete, and error-free genome assemblies of all ~70,000 extant vertebrate species [4]. By comparing sequencing data types and assembly algorithms, the VGP consortium determined that it was not possible to obtain high-quality reference assemblies at the chromosomal level without the complementary use of multiple long-read sequencing technologies. Long-reads (generally > 10 kb; e.g. Pacific Biosciences or Oxford Nanopore), long-range molecules (generally > 50 kb; e.g. 10X Genomics linked reads), or optical mapping (> 150 kb; e.g. Bionano Genomics), and Hi-C proximity ligation (> 1 Mb; e.g. Arima Genomics) can span repeats thousands of base pairs in length [4], greatly improving assembly outcomes. To take full advantage of these new sequencing and assembly methods, molecules of DNA need to be as long as possible.

While long-read and long-range (LR) data simplify and accelerate the assembly, they come with a major challenge: they require large amounts of very high-quality DNA. For short-read technologies, many nucleic acid isolation methods developed over the years, including the standard phenol-chloroform method [5] had been sufficient. LR technologies require relatively pure DNA in the 10 kb to 300 kb range. Additionally, the Hi-C method requires physical cross-linking of contacting DNA regions within the same chromosomes, thus requiring cell nuclei to be intact before processing and isolation of cross-linked DNA [4]. With Hi-C, 3D interactions within chromosomes serve to assemble contigs or short scaffolds into chromosomal-scale scaffolds. For LR technologies, only a few extraction methods are currently able to produce high molecular weight (HMW) DNA ranging from 45 to 150 kb or ultra-high molecular weight (uHMW) DNA which is over 150 kb long. These include bead-based (MagAttract HMW DNA Kit, Qiagen), high- salt [6], and agarose plug methods (Bionano Prep Soft/Fibrous Tissue Protocol, Bionano Genomics) [7]. More recently, a less laborious thermoplastic magnetic disks (Nanobinds) method was

developed by Circulomics [8]. Regardless of their capabilities, the performance of HMW and uHMW DNA extraction methods primarily depend on the type of sample and how it was collected, handled, and preserved.

The long-held “gold standard” in tissue preservation for high-quality DNA isolation has been flash-freezing tissues in liquid nitrogen directly after collection, followed by ultra-cold –80^{o}C long-term storage [9–14]. While liquid nitrogen is readily available in most laboratory setups, its limited availability in many fieldwork conditions can be an insurmountable hurdle. Indeed, a large portion of global biodiversity is located far from labs, and sampling such species will require long expeditions under rustic field conditions. Thus, transporting sufficient amounts of liquid nitrogen from the point of collection to the laboratory is often infeasible and the applicability of flash-freezing outside the lab environment is greatly limited [10,13,15]. Additional considerations specific to the studied species exacerbate the challenge of sample collection and preservation. DNA degradation is promoted by enzymes whose concentrations are likely to be tissue-specific and possibly species-specific. Small organisms provide little tissue, and preferred tissue types may be unavailable. Permitting restrictions also vary widely among species and among countries. Yet, methods for field sampling in non-model species for the purposes of LR sequencing remain anecdotal or unsubstantiated, as failed attempts are not published and very few preservation experiments have measured fragment sizes relevant to LR technologies [16,17]. Thus, methods that bridge the gaps between uHMW DNA, the lab, and field conditions still require benchmarking.

Here, we perform a series of benchmarking experiments to assess sample preservation methods under laboratory and simulated field conditions and compare the quality of uHMW DNA obtained. Specifically, we extract uHMW DNA from multiple tissue types of representative vertebrate species, which were collected under various preservation and temperature conditions. For each experimental sample, we evaluate the fragment length, yield, and purity of the uHMW DNA

extracted. Based on our findings, we propose a new set of guidelines for tissue preservation, ranging from best to minimally adequate practices for acquiring uHMW DNA from both laboratory and field collected samples, necessary for producing high-quality reference genome assemblies.

Results

In this study, we used the agarose plug method optimized by Bionano Genomics [7] across all species and preservation methods albeit with small protocol variations for fibrous tissues, soft tissues, and blood. We tested six preservation methods (**Fig. 1**): 1) flash frozen in liquid nitrogen, which served as the 'gold standard' and our point of reference; 2) 95% ethanol (EtOH), a long preferred method of field preservation of tissues [10,15,18]; 3) 20–25% dimethyl sulfoxide (DMSO) buffer (see Methods), which has been shown to be very effective at permeating tissues and preserving HMW DNA after long-term storage at ambient temperature [19,20]; 4) RNAlater Stabilization Solution (RNAlater; Invitrogen, Waltham, MA, USA), a commonly used preservative that also facilitates transcriptomics; 5) DNAgard tissue and cells (DNAgard; Biomatrica, San Diego, CA, USA), a commercial preservative designed for stabilizing DNA in tissues at room temperature; and 6) Allprotect Tissue Reagent (Allprotect; Qiagen, Hilden, Germany), another commercial preservative targeting stable room-temperature tissue preservation. We exposed preserved samples to different temperatures (4°C, room temperature, and 37°C) for various durations of time (6 hr to 5 months). We did so with up to 6 tissue types (muscle, blood, ovary, spleen, isolated red blood cells (RBCs), and whole-body) from 6 species representing five vertebrate lineages (a mammal, a bird, two turtles, an amphibian, and a bony fish; **Fig. 1**), for a total of 140 samples (**Table S1**). We assessed the fragment length distribution and DNA yield for each DNA sample. Statistical analyses were performed using linear models that included type of preservative, temperature/time treatment, vertebrate group, and tissue type as variables.

Fragment length distribution analysis. For extractions that yielded a detectable amount of DNA, we measured their fragment length distributions using at least one of two available techniques: Pulsed-field Gel Electrophoresis (PFGE) and the Agilent Femto Pulse system (FEMTO). PFGE was more informative for analyzing uHMW DNA molecules above 200 kb, due to greater dynamic range in molecular weight separation (**Fig. S1a**), whereas FEMTO was more useful for separating molecules within the 50–165 kb range (**Fig. S1b**). Overall, the agarose plug method yielded high-quality DNA concentrated in the 300–400 kb range (**Fig. 2**).

Temperature. From the linear modeling of both PFGE (**Fig. 2, Table S2**) and FEMTO results (**Figs. S2, S3**), we found that temperature treatment was the predictor with the strongest evidence of an effect on the proportion of DNA fragments above 145 kb for PFGE (DF = 6, LR Chisq = 36.62, p = 2.09e-06; **Fig. 3a**) and above 45 kb for FEMTO (DF = 8, LR Chisq = 44.80, p = 4.01e-07; **Fig. S4a**). Samples held at higher temperatures yielded a lower proportion of uHMW DNA, with flash-freezing performing best (**Fig. 3a**). However, samples refrigerated at 4°C for 6 hr following collection were statistically indistinguishable from flash-frozen samples (PFGE: z = 0.56, p = 1.00; FEMTO: z = 2.03, p = 0.48). Samples refrigerated at 4°C for longer periods of up to one week showed some signs of degradation, albeit not consistently across tissue types and species (**Figs. 2, S2, and S3**).

Preservation method. The predictor with the second strongest evidence of an effect on the proportion of DNA fragments above 145 kb or 45 kb was preservative treatment (PFGE: DF = 5, LR Chisq = 24.43, p = 0.0002, **Fig. 3b;** or FEMTO: DF = 6, LR Chisq = 25.01, p = 0.0003, **Fig. S4b,** respectively). In PFGE measurements, significant differences were found between DNAgard and EtOH preservation (z = 4.24, p = 0.001, **Fig. 3b**), with DNAgard generally performing poorer. Flash-freezing and EtOH performed better than the other preservation methods in PFGE, and albeit not statistically significant, they had the lowest standard deviation (**Fig. 3b**). Based on

PFGE, EtOH was slightly better than DMSO (**Fig. 3b**). Based on FEMTO, DMSO was slightly better than EtOH (**Fig. S4b**). Neither relationship showed significant differences in preservation. In FEMTO measurements, flash-frozen and DMSO-preserved samples showed significantly better preservation efficiency than RNAlater samples (vs. DMSO: $z = 3.42$, $p = 0.009$; vs. flash-frozen: $z = 3.50$, $p = 0.007$), tested on fish samples. Allprotect outperformed EtOH in room temperature mouse samples but underperformed in the refrigerated fish body set (**Figs. 2 and S3**).

Tissue type. Tissue type did not have a significant effect on fragment length overall (**Figs. 3c and S4c, Table S2**). However, muscle showed more variability than blood samples in uHMW DNA yield (> 145 kb). The RBCs samples showed the smallest proportion of degradation, while some muscle samples showed the highest degradation (**Fig. 3c**). In terms of variation between species, the mouse and fish samples showed a higher degree of degradation with respect to temperature treatment than the other species (**Figs. 2, S2, and S3**). It is unclear if this can be explained by a species-specific temperature sensitivity, or if it is caused by technical variation.

Interactions among variables. In terms of qualitatively assessing combinations of variables, storage in EtOH appeared to perform best at preserving uHMW DNA for all 4ºC refrigerated samples (**Fig. 2**). Notably, nucleated blood samples refrigerated with no added preservatives were stable for up to one week with no substantial signs of degradation (**Fig. 2**). An increased proportion of smaller DNA fragments was evident in refrigerated samples preserved using DNAgard, with the exception of turtle RBCs and muscle samples for which DNAgard results were equivalent to other preservation methods (**Fig. 2**). Fish body samples stored for 16 hr at 4°C showed notable degradation, but mouse spleen samples under the same treatment did not vary substantially from samples stored at 4°C for 6 hr (**Fig. 2**). Replicate sea turtle RBCs samples showed less variation within treatments for fragment size than for DNA yield (**Fig. S5a,b**).

Mouse muscle, fish muscle, and fish ovary samples showed considerable accumulation of smaller fragment sizes after one week at room temperature, where blood or muscle samples from other species did not show as dramatic an impact (**Figs. 2, S2, and S3**). However, fish muscle and ovary samples stored at room temperature for just one day still retained high proportions of uHMW DNA with marginal degradation (**Fig. S2**). For mouse muscle, DMSO, EtOH, or DNAgard did not seem to provide any added DNA protection against room temperature conditions (**Figs. 2 and S3**). At the same temperature conditions, mouse samples in Allprotect retained a non-negligible fraction of uHMW DNA, though with some degradation (**Figs. 2 and S3**). Overall, similar to the 4°C exposure, room temperature DMSO and EtOH samples performed relatively well, albeit showing some signs of degradation. Surprisingly, two samples left at room temperature for one week without any preservative (sea turtle RBCs and frog blood) were quite stable and yielded an appreciable fraction of uHMW DNA (**Fig. 2**). Additionally, sea turtle RBCs samples, when preserved with EtOH or even DNAgard and stored at room temperature for 5 months, yielded a large fraction of workable uHMW DNA (**Fig. 2**). This suggested that turtle RBCs may be viable for longer durations at room temperature. Additional replicates and further experimentation will be necessary to determine if the isolated RBCs tissue type or some biological difference in turtles is the key to this stability.

DNA yield. When the variables were tested individually, vertebrate group explained the least variance in DNA yield (3.69%, DF = 4, F = 3.25, p = 0.01; **Fig. 3d**); temperature treatment explained a similarly small proportion (7.35%, DF = 9, F = 2.88, p = 4.25e-3); preservative explained slightly more of the total variance (10.24%, DF = 6, F = 6.01, p = 1.73e-5; **Fig. 3e**); and tissue type explained the largest amount of variance (46.35%, DF = 5, F = 32.65, p = 2.20e-16; **Fig. 3f**). Both preservative and tissue type together explained 56.59% of the total variance (**Table S2**). Specifically, whole blood tended to generate the highest DNA yields, followed by spleen,

RBCs, whole-body, and ovary, while muscle generated relatively lower yield (**Fig. 3f**). In post-hoc tests, whole blood, RBCs, and ovary significantly outperformed muscle (vs. whole blood: $t = 11.75$, $p = 0.002$; vs. RBCs: $t = 8.36$, $p < 0.001$; vs. ovary: $t = 3.28$, $p = 0.01$), while the differences between muscle and whole body or spleen were not significant. Whole blood and RBCs also showed significantly higher yields than ovary samples (vs whole blood: $t = 3.89$, $p = 0.002$; vs. RBCs: $t = 3.36$, $p = 0.01$). Post-hoc comparisons of different temperature treatments or preservation reagents were not significant, possibly due to the higher variance influenced by the other variables of tissue type and species (**Fig. 3d-f**). Birds tended to have slightly better yields, with a marginally significant effect over non-avian reptiles ($t = 3.04$, $p = 0.02$).

Hi-C sequencing. The VGP is currently using Hi-C reads as a standard tool to generate chromosomal scale assemblies [4,21], as well as to phase haplotypes in some cases [22]. These chromosome interactions are captured *in situ* in the tissue before DNA is isolated and sequencing libraries made. To enable appropriate collection recommendations for use in this technology, we also explore the effect of tissue preservation on the quality of the Hi-C library preparation. Using a single species (zebra finch) we test a subset of tissue preservation methods (flash-frozen, 6 hr at 4°C, one week at room temperature) and tissue types (muscle, blood), with two replicates per treatment combination. These were processed to generate *in situ* Hi-C chromatin interactions maps against the VGP male reference genome [23,24].

We found that blood samples flash-frozen in EtOH yielded similar results compared to our flash-frozen positive control with no added preservative: 75–80% of all read-pairs were derived from *cis* interactions within the same chromosomes (**Fig. 4a**), and among them ~55–60% were derived from long-range (>15 kb) *cis* interactions. This indicates a high degree of useful long-range intra-chromosomal signal necessary for genome assembly. However, storage of blood in DNAgard resulted in the elimination of almost all *cis* interactions, down to ~10% total, across temperature

treatments (**Fig. 4a-c**), indicating largely random ligations and the loss of useful signal. Blood refrigerated for 6 hr maintained a high yield of long *cis* interactions, both when stored in EtOH and with no preservative. Blood samples stored at one week at room temperature in EtOH also yielded mostly long *cis* interactions similar to the flash-frozen treatments.

Overall, muscle and blood samples performed similarly across all treatments measured using Hi-C reads. They both yielded large amounts of long *cis* interactions (>15 kb) when flash-frozen or refrigerated at 4°C with no preservative or with EtOH (**Fig. 4a-b, d-e**). Muscle and blood samples also responded similarly to preservative treatments, with EtOH samples performing well across treatments and DNAgard samples underperforming across treatments (**Fig. 4**).

Discussion

During development of the assembly pipeline for the first set of VGP genomes [4], we tested various HMW and uHMW DNA extraction protocols compatible with several LR technologies, including the Qiagen MagAttract HMW DNA, the phenol-chloroform method [5], and the agarose plug protocol. The agarose plug method optimized by Bionano Genomics [7] was the most consistent method for producing a high yield of uHMW DNA suitable across all the LR technologies in the VGP pipeline. This method used agarose as a protective matrix to minimize DNA shearing during the extraction process and had long been shown to be an effective method for isolating megabase-size DNA from organisms including plants, animals, algae, and microbes [7]. In this study, we use only the agarose plug DNA extraction method.

Our study explored the effects of three variables –preservation method, tissue type, and storage temperature– in preserving the high-quality DNA required for generating chromosome-scale genome assemblies in six species representing five major vertebrate lineages. The results

identified promising alternatives to the standard flash-freezing method that is not easily performed in the field, particularly the preservation of samples in 95% ethanol (EtOH) or 20–25% DMSO-EDTA (DMSO) at 4°C.

We did not test all possible combinations of variables, which would require over 252 tests per species, but focused instead on the salient combinations of tissue types, reagents, and protocols that reflect real-world applications. There are also likely intervening stages of exposure to different temperatures, such as immediately post-mortem, that may have a considerable effect in hotter climates and are not simulated here. Additionally, we are only able to visualize DNA fragment size distributions within a certain range of sizes (approx. 40–400 kb for PFGE, 1.3–165 kb for FEMTO). Though we have targeted a size range that includes both ideal fragment sizes for long-read sequencing and fragments of lower molecular weight that may indicate degradation, fragments outside this range are not measured here. Despite these limitations, our results are consistent with samples from the over 136 species we have processed for the VGP to date (NCBI Bioproject PRJNA489243 as of July 13, 2021). We believe that the results presented here can inform the many logistical decisions of field researchers collecting samples from wild populations (**Fig. 5**).

Temperature exposure was the strongest predictor of fragment length distribution for these data. The potential of increased temperatures to destabilize DNA is well known, and samples exposed to higher temperatures for a longer period will allow for enzymatic activity that degrades DNA [25]. However, under certain conditions some samples stored at 4°C or even at room temperature show surprising viability. For example, samples preserved in EtOH and refrigerated for up to one week were nearly as good as flash-frozen samples. This is evident through high proportions of uHMW DNA molecules, though with some signs of degradation and variability across species and tissue types.

The ambient temperature of the intended collecting locality should be a major consideration in planning field collections for high-quality samples. Here we test a limited number of samples at 37°C to resemble fieldwork conditions in warmer climates, resulting in no retention of workable amounts of uHMW DNA in any of these samples (4 mouse muscle samples; Fig. 2). Thus, in hotter climates sample cooling or exploring alternative preservatives is critical. Options such as insulated boxes, ice packs, wet ice, dry ice, and electronic coolers should be considered for maintaining samples at low temperatures in the field. To minimize the time before storing in ultra-cold freezers, investigators might also choose to ship samples from the field to the lab before the conclusion of fieldwork. Further experimentation in conditions resembling warmer climates can more precisely define tolerable exposure intervals for sampling targeting uHMW DNA.

The “gold standard” for preserving samples for uHMW DNA extraction remains flash-freezing in liquid nitrogen before ultra-cold storage [9–14]. Our results highlight alternative preservation methods that are more readily available in the field. Liquid nitrogen can be challenging to acquire, contain, and transport in many fieldwork settings. Fortunately, samples preserved in EtOH or DMSO perform well with simple refrigeration. Although a small portion of DMSO samples failed (near-zero DNA extracted) for unclear reasons. In addition, these solutions consistently outperform the commercial preservatives RNAlater and DNAgard. Further, DNAgard is not suitable for maintaining long interaction distances for Hi-C library preparation. While these commercial reagents rely on mechanisms that were likely optimized for preserving lower molecular weight nucleic acids, they appear to be harmful to uHMW DNA and chromosomal 3D interactions. Preservatives that promote cell lysis may undermine the stability of DNA if they cannot adequately counter the increased exposure to sources of chemical degradation [14,25,26]. Though our washing protocol should minimize its effect, it is also possible that some unknown aspect of the DNAgard treatment of cells inhibited the crosslinking reaction, and Hi-C of unfixed cells would be expected to have low signal and high noise similar to degraded DNA. Of the three

commercial reagents tested, Allprotect shows the most promising results for preserving uHMW DNA, but more testing is necessary to better evaluate its performance relative to other preservatives and assess its compatibility with LR technologies.

In addition to popular commercial reagents, we evaluate some of the more commonly applied preservation methods today. EtOH has long been used for preserving samples for DNA analysis, and its proficiency at stabilizing specimens continues to be validated [12,18,27,28]. For example, Mulcahy et al. (2016) studied preservative effects on DNA integrity in white perch and blue crab muscle samples, using only a maximum of 45 kb DNA size resolution. Nevertheless, their finding that EtOH generally performs well as a DNA preservative agent is consistent with our results at this DNA size range. While EtOH is a compelling option, it comes with its own logistical considerations. EtOH can be problematic to transport on commercial flights or trains, or to ship in large quantities. Alternatively, DMSO benefits from fewer transport restrictions, but requires laboratory preparation prior to fieldwork and can be hazardous to handle. Commercial preservation reagents are usually more costly than EtOH or DMSO solutions, but are also under less restricted transport regulations.

The negative impact of DNAgard on Hi-C long-distance *cis* interactions is striking. This solution likely permeates the cell to inhibit nuclease activity, potentially affecting other protein integrity and impeding cross-linking. The increased fraction of inter-chromosomal interactions and decreased fraction of cis-interactions (> 15 kb) together are evidence of DNA degradation. These inter-chromosomal interactions are counter-productive noise with regard to chromosome-level scaffolding in that they erroneously provide scaffolding links between contigs derived from two different chromosomes. Our Hi-C data analysis also indicates, at least for birds, that EtOH storage of blood at 4°C or room temperature for one week or less tends to yield high-quality Hi-C chromatin interaction maps. Excluding samples in DNAgard, blood seems to be slightly more

resistant to reducing chromosome interactions than muscle when stored at 4°C or room temperature for one week, which would be a valuable feature for field collection.

Contrary to the differences in Hi-C performance, we did not find notable differences in DNA fragment length distributions between most tissue types. The exception is whole-body fish samples that were all significantly degraded, regardless of treatment. Potentially, this could owe to the larger mass of tissue taking longer to freeze through or infuse with preservative, hence allowing more time for degradation. However, we did observe substantial differences in total DNA yield, where blood and spleen samples tend to yield a larger amount of DNA while muscle samples produce the least. The comparatively lower DNA yield makes muscle samples a less practical choice in species where nucleated blood is available. Lower yield could also be costlier and more time consuming in the long run, as more DNA extractions would be required to achieve the necessary input amount. For species without nucleated blood (mammals), soft tissue samples such as the spleen outperform muscle in terms of yield. Note that low yield does not necessarily preclude muscle samples from usefulness, especially given they still perform well in terms of fragment length if appropriately collected and stored. We note that, as we demonstrated in a related study [29], blood is often not suitable for uHMW mitochondrial DNA extraction, while muscle tends to yield abundant mitochondrial DNA. This is an important consideration if the goal of collection is to sequence the mitochondrial genome.

Our study considers today's LR sequencing technologies and current DNA isolation protocols. Time will likely continue to yield new methods for preventing, assessing, and mitigating DNA degradation. Even since the outset of this study, promising new extraction methods have become available for uHMW DNA, such as Nanobind DNA extraction (Circulomics, Baltimore, MD, USA). Our comparisons focus on maximizing the quality of field-collected input material and we expect this to be largely independent of downstream extraction methods. Our results and experience

acquired with uHMW DNA and Hi-C data for more than 136 VGP genomes produced, yield guidelines for tissue type, preservatives, temperature, and other treatments necessary for generating high-quality genome assemblies from several vertebrate lineages, for laboratory and field collected samples (**Table 1**).

In planning biobanking for genomic purposes, another important strategy is to avoid or reduce the need for field-preserved samples. Seeking out animals already in captive collections and salvaging material reduces the methodological difficulty of preserving samples. Delaying blood collection, biopsy, or euthanasia of wild-caught specimens can also buy researchers time to move into more amenable preservation conditions such as a field station. However, this poses ethical challenges in the care of animals being held for days or weeks, and it is not feasible for larger animals.

Few studies have explored the effects of preservation methods on uHMW DNA integrity [17], but none that we are aware of have done so in as broad a set of field-relevant conditions as in the present study. Being able to collect samples well-suited for producing high-quality genome assemblies is a major undertaking. Our recommendations will enable many new high-quality sample collections and contribute to establishing a greater and more diverse array of vertebrate genomes from around the world.

Methods

Sample collection. We collected samples from species representing major taxonomic classes of vertebrates, i.e. house mouse (*Mus musculus*), zebra finch (*Taeniopygia guttata*), Kemp's Ridley sea turtle (*Lepidochelys kempii*), painted turtle (*Chrysemys picta*), American bullfrog (*Rana catesbeiana*), and zebrafish (*Danio rerio*). All animal handling and euthanasia protocols were

approved by the Institutional Animal Care and Use Committees or equivalent regulatory bodies at the respective facilities: The Rockefeller University for the frog and bird samples; the Max Planck Institute for the mouse samples; the University of Toronto for the painted turtle samples; the Wellcome Sanger Institute for the fish samples; and the New England Aquarium rehabilitation facility for the sea turtle samples (**Table S1)**.

For this experiment, tissue samples were collected as available at facilities already handling the target species (**Fig. 1**). The tissue types collected per species are as follows: mouse, spleen and muscle; zebra finch, whole blood and muscle; sea turtle, isolated red blood cells (RBCs); painted turtle, whole blood and muscle; bullfrog, whole blood and muscle; zebrafish, whole body, ovary, and muscle. For all species except the sea turtle and the fish, samples originate from a single individual. In the sea turtle set, duplicate samples were obtained from three individuals. In the fish set tissue samples in some cases originated from different individuals, as their small body size does not allow for sufficient amounts of tissue from a single specimen.

Each taxon required a slightly different handling procedure. All samples except for those from sea turtles were sourced from captive individuals humanely euthanized in a laboratory setting with approved protocols cited below. All soft or fibrous tissue samples were collected in small 20–30 mg pieces until each 2 mL tube had roughly 50–100 mg total to allow for full penetration of the preservative. Mice were euthanized by CO_2 treatment in a GasDocUnit (Medres Medical Research GmbH, Cologne, Germany) following the instructions of the manufacturer (DD24.1-5131/451/8, Landesdirektion Sachsen). Skeletal muscle and spleen samples were then dissected and placed in standard cryotubes. Birds were euthanized via isoflurane overdose, and whole blood was collected into chilled sodium heparin-treated 1.5 ml microfuge tubes (IACUC #19101-H). Then 25–50 µL was immediately aliquoted into cryotubes. Sea turtle RBCs samples were collected from wild individuals undergoing medical treatment by drawing whole blood into 2 mL

sodium heparin-treated collection tubes and then spinning down to separate RBCs from plasma. RBCs were then aliquoted into sodium heparin-treated tubes. Painted turtle samples were collected from one individual euthanized via decapitation as part of another study (AUP 20012070). Painted turtle muscle samples were immediately taken from the pectoral girdle and whole blood was drawn from the heart before placement in standard cryotubes. Frog samples were sourced from one captive adult purchased from Rana Ranch in Twin Falls, Idaho, USA. The frog was euthanized using an intracoelomic injection with Euthasol™ or Fatal-Plus™ (pentobarbital and phenytoin) at a dosage of 100 mg/kg. After confirming that a deep plane of anesthesia was reached, the frog was rapidly and doubly pithed cranially and spinally, then decapitated (19085-USDA). Frog muscle tissue samples were immediately taken from the rear legs and blood was drawn from internal veins before placement in standard cryotubes. We extracted fish samples from multiple lab-raised individuals. To euthanize the fish, we used tricaine and then the brain was destroyed with a scalpel (PPL No.70/7606). We collected white muscle and ovary samples which were dissected out and placed into 2 ml cryotubes immediately after euthanasia. Fish whole-body samples were taken by removing the head, intestines, and swim bladder of individual fish and placing the remaining tissue into a cryotube.

Preservation treatments. A total of 140 freshly collected samples were subjected to different preservation and temperature treatments to test common preservation methods under simulated field or lab conditions (**Fig. 1**), with flash-frozen samples being used as baseline controls. Preservation method treatments refer to the preservative agent applied directly to the sample before ultra-cold (–80°C) storage; temperature treatments refer to the temperature exposed and the amount of time the sample remained at that temperature before ultra-cold storage.

All temperature treatments were applied immediately upon dissection of the material and placement into specimen tubes. Samples were exposed to temperature treatments of varying

lengths of time in refrigeration (4°C), room temperature (20–25°C), and elevated temperature in an incubator to simulate field conditions in a tropical climate (~37°C). All temperature conditions tested and the samples to which they were applied are as follows: control condition submerged in liquid nitrogen from dissection to ultra-cold storage (all tissue types and species), 6 hr at 4°C (frog blood and muscle, bird blood and muscle, painted turtle blood and muscle, sea turtle RBCs), 16 hr at 4°C (mouse spleen, fish whole body), 1 day at 4°C (fish ovary), 1 week at 4°C (mouse muscle, frog blood and muscle, bird blood and muscle, painted turtle blood and muscle), 1 day at room temperature (fish muscle and ovary), 1 week at room temperature (mouse muscle, frog blood and muscle, bird blood and muscle, painted turtle blood and muscle, sea turtle RBCs, fish muscle and ovary), 4 weeks at room temperature (fish muscle and ovary), 5 months at room temperature (sea turtle RBCs), and 1 week at 37°C (mouse muscle). Storage time at –80°C after treatment and before DNA extraction varied slightly between samples, but such variation is expected to have a negligible impact on sample quality.

The preservation methods tested here include flash-freezing in liquid nitrogen, no added preservative agent, 95% EtOH, 20–25% DMSO-EDTA (DMSO), DNAgard tissue and cells (DNAgard; cat. no. #62001-046, Biomatrica), Allprotect Tissue Reagent (Allprotect; cat. no. 76405, Qiagen), and RNAlater Stabilization Solution (RNAlater; cat. no. AM7021, Invitrogen). Our DMSO recipe was 20–25% DMSO, 25% 0.5 M EDTA, remaining 50–55% H2O, saturated with NaCl. Flash-freezing, EtOH, and DNAgard were tested on all included species and tissue types. DMSO was tested on all species and tissue types except sea turtle RBCs. No-preservative treatments were tested on bullfrog blood, bird blood, painted turtle blood, and sea turtle RBCs. Allprotect was tested on mouse spleen and muscle and fish body. RNAlater was tested on fish ovary and muscle samples.

To gain insights into variation within these treatments, isolated RBCs samples were collected from three different sea turtle individuals and processed separately as biological and technical replicates. The third replicate had insufficient material to test all treatments.

DNA extraction. We extracted DNA from all tissue samples using the agarose plug protocol as below at VGP data production hubs at the Rockefeller University, Wellcome Sanger Institute, and MPGI Max Planck Institute Dresden (**Table S1**). This method was established, at the time of this experiment, as standard protocol for long-read sequencing in all VGP projects [4]. From each tissue sample, a 30–40 mg piece was weighed and then processed using the Bionano Prep™ Animal Tissue DNA Isolation Fibrous Tissue Protocol (Bionano document number 30071) and Soft Tissue Protocol (Bionano document number 30077). Briefly, the fibrous tissue (muscle, whole) pieces were further cut into 3 mm pieces and fixed with 2% formaldehyde and Bionano Prep Animal Tissue Homogenization Buffer. Tissue was blended into a homogenate with a Qiagen Rotor-Stator homogenizer and embedded in 2% agarose plugs cooled to 43°C. Plugs were treated with Proteinase K and RNase A, and washed with 1X Bionano Prep Wash Buffer and 1X TE Buffer (pH 8.0). DNA was recovered with 2 µl of 0.5 U/µl Agarase enzyme per plug for 45 minutes at 43°C and further purified by drop dialysis with 1X TE Buffer. The soft tissue (spleen, ovary) pieces were further cut into 3 mm pieces and then homogenized with a tissue grinder followed by a DNA stabilization step with ethanol. The homogenate pellet was then embedded in 2% agarose plugs as in the fibrous tissue protocol above. For blood samples, DNA was extracted from whole blood or RBCs following the unpublished Bionano Frozen Whole Nucleated Blood Stored in Ethanol – DNA Isolation Guidelines. The ethanol supernatant was removed and the blood pellet was resuspended in Bionano Cell Buffer in a 1:2 dilution. For samples that freeze solidly at –80°C, tubes were thawed at 37°C for 2–4 minutes. The same Bionano guidelines for nucleated blood in ethanol were modified by adding 1–2 additional centrifugation steps at 5,000X g for 10 min prior to removing DNAgard supernatant and homogenizing blood cells in Bionano

Cell Buffer in a 1:2 dilution. All samples were mixed with 36 µl agarose and placed in plug molds following the animal tissue protocol.

Assessing sample purity and yield. All extractions had sufficient DNA yield to measure except one: mouse spleen tissue in DMSO. This sample congealed and solidified in such a way that no DNA could be extracted. To measure DNA yield and purity, we used both the fluorescence-based Broad Range Qubit® assay and absorbance-based Nanodrop One™. To measure yield, 2 µl aliquots of gDNA were taken from the top, middle, and bottom of each DNA sample and diluted in a Qubit Working Solution of 1:200 Dye Assay Reagent with BR Dilution Buffer. Sample concentrations were recorded on a Qubit 4 Fluorometer. The concentration of the top, middle, and bottom readings were averaged to estimate the concentration of each DNA sample. Spectrophotometry was then performed on a Nanodrop One to measure sample purity in terms of the 260/230 and 260/280 nm ratios.

Assessing sample fragment size distributions. Fragment length distributions of samples were measured with at least one of two available methods: Pulsed-field Gel Electrophoresis (PFGE) or the Agilent Femto Pulse system (FEMTO). PFGE was performed using the Sage Science™ Pippin Pulse gel system with the Lambda PFG Ladder (New England Biolabs). To quantify fragment length distribution from PFGE gel images, we compared the proportions of signal above and below 145 kb. This was done using the program ImageJ [30] following Mulcahy et al. (2016) based on the Gel Analysis tool in ImageJ. Further quantifying of the PFGE signal below 145 kb, such as the relative amount of low molecular weight DNA, was not robust due to compression or streaking obscuring smaller fragment patterns. Concise visualization of gel plot profiles was produced in the R package ggridges [31] with a custom Python script for piecewise linear scaling across different gels according to a common size standard. Grey-value intensity measured in ImageJ was scaled locally in each lane and cropped to the gel boundary such that, excluding the

well, the brightest value along the lane became 100 and the darkest became 0. Analysis of FEMTO outputs was carried out in the ProSize Data Analysis Software. First, each trace was assessed for signs of an unreliable run, including ladder quality, loading concentration, raised baseline, and unusual smear patterns. Runs with these hallmarks were not incorporated further. Because signals above 165 kb are not reliable on FEMTO, we only considered signals within the range of 1.3–165 kb. We then recorded the proportion of the sample measuring above 45 kb. Further visualization of FEMTO traces were made in the same manner as above with a custom python script and the R package ggridges, except scaling to a size standard was done in ProSize. Yields were insufficient for fragment size analysis from frog muscle in DMSO for one week at 4°C and 6 hr at 4°C and mouse spleen in DMSO for 16 hr at 4°C.

Statistical analysis. We used linear modeling in the R statistical package to explore the relative contribution of several factors to the variance in DNA yield and fragment length among tests. The three response variables (DNA yield per unit mass (yield), PFGE proportion > 145 kb (PFGE), FEMTO proportion > 45 kb (FEMTO)) were modeled separately. The data for each model were samples with those measurements, and all conditions had at least two replicates (yield: n = 139, PFGE: n = 102, FEMTO: n = 108). DNA yield was log-transformed using the natural logarithm to satisfy assumptions of normality and modeled with temperature, preservative, vertebrate group, and tissue type included as fixed effects. Homoscedasticity was checked after modeling and found to conform to assumptions. PFGE proportion and FEMTO proportion were modeled with quasibinomial error distributions with temperature, preservation method, and tissue type included as fixed effects. Vertebrate group was not included in the final fragment length models due to collinearity with tissue type. Post-hoc tests were done using the glht function of the R package multicomp to examine differences between the levels of each factor. Further model details including p-values and contingency tables are available in the supplementary materials (**Tables S2 and S3**).

Hi-C library preparation and sequencing. Because Hi-C methods require intact cell nuclei, we tested a subset of bird samples from our preservation experiments directly using the Arima-HiC platform. We tested blood and muscle samples in three different treatments: without preservatives, in EtOH, and in DNAgard. Each preservation method was subjected to three temperature treatments: immediately flash-frozen, 6 hr at 4°C, and one week at room temperature (20–25°C). After temperature treatment, each sample was moved to –80ºC. Blood with no preservative at room temperature for one week was excluded from this set. Two technical replicates of each sample were prepared and sequenced at Arima Genomics following their standard protocol (Arima Genomics, Doc A160177 v00). Briefly, standard protocol for nucleated blood in a solution like EtOH or DNAgard is to pellet the cells, remove the supernatant, wash with 1X phosphate buffered saline solution containing 1% fetal bovine serum, and then carry the washed pelleted cells into crosslinking and then Arima-HiC. We measured the performance of Arima-HiC runs by mapping the sequence reads to the zebra finch reference genome (GCA_003957565.1) to determine the proximity of ligated sequence pairs. Assessments were made based on the ratios of *cis* (intra-chromosome) to *trans* (inter-chromosome) read pairs as well as the total percentage comprised of long-distance (> 15 kb) cis pairs.

Data availability

Sample information, PFGE measurements, FEMTO measurements, and DNA yield data can be found in the supplemental materials. Raw FEMTO outputs, PFGE gel images, and analysis scripts are available on Dryad (doi:10.5061/dryad.000000041). Raw Hi-C read-pairs are publicly available on the NCBI Sequence Read Archive.

Acknowledgments

This research was supported by Howard Hughes Medical Institute Funds and Rockefeller University Startup funds to EDJ, institutional funds of the Max Planck Institute of Molecular Cell Biology and Genetics, and funds by the Wellcome Trust made out to the DNAP R&D team at Wellcome Sanger Institute. IB's time was supported by Wellcome grants WT207492 and 104640/Z/14/Z, 092096/Z/10/Z. Sampling was facilitated by Leslie Buck and Mouska Patang for the painted turtle and Brian Fabella for the bullfrog. Sea turtle sampling was conducted and generously facilitated by the New England Aquarium and Massachusetts Audubon Wellfleet Bay Wildlife Sanctuary authorized under USFWS permit TE01150C-1; sample transfer to LMK was permitted via a USFWS special authorization letter.

Author contributions

J.M., S.W., A.F.S., I.B., L.M.K., T.L., A.J.C., R.W.M., A.D.S., P.A.M., E.D.J., and O.F. initially conceptualized the study. H.A.D., J.M., J.B., S.W., A.F.S., S.M., O.V.P., I.B., K.O., M.S., W.T., A.K., L.M.K., E.D.J., and O.F., carried out data collection and preprocessing. H.A.D., J.B., G.F., and A.F.S. analyzed the data and produced the figures. The manuscript was drafted by H.A.D., J.M., J.B., G.F., E.D.J., and O.F., and all authors contributed to revisions.

Competing interests

The authors declare no competing interests.

References

1. Koepfli KP, Paten B, Genome 10K Community of Scientists, O'Brien SJ. The Genome 10K Project: a way forward. Annu Rev Anim Biosci. 2015;3:57–111.

2. Ko BJ, Lee C, Kim J, Rhie A, Yoo D, Howe K, et al. Widespread false gene gains caused by duplication errors in genome assemblies. bioRxiv 2021,

http://dx.doi.org/10.1101/2021.04.09.438957

3. Kim J, Lee C, Ko BJ, Yoo D, Won S, Phillippy A, et al. False gene and chromosome losses affected by assembly and sequence errors. bioRxiv 2021, http://dx.doi.org/10.1101/2021.04.09.438906

4. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592:737–46.

5. Sambrook J, Russell DW. Purification of nucleic acids by extraction with phenol:chloroform. CSH Protoc. 2006; doi: 10.1101/pdb.prot4455

6. Lahiri DK, Nurnberger JI Jr. A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies. Nucleic Acids Res. 1991;19:5444.

7. Zhang M, Zhang Y, Scheuring CF, Wu C-C, Dong JJ, Zhang H-B. Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. Nat Protoc. 2012;7:467–78.

8. Zhang Y, Zhang Y, Burke JM, Gleitsman K, Friedrich SM, Liu KJ, et al. A Simple Thermoplastic Substrate Containing Hierarchical Silica Lamellae for High-Molecular-Weight DNA Extraction. Adv Mater. 2016;28:10630–6.

9. Frampton M, Sam D. Evaluation of Specimen Preservatives for DNA Analyses of Bees. J Hymenopt Res. 2008;17:195–200.

10. Kilpatrick CW. Noncryogenic preservation of mammalian tissues for DNA extraction: an assessment of storage methods. Biochem Genet. 2002;40:53–62.

11. Seutin G, White BN, Boag PT. Preservation of avian blood and tissue samples for DNA analyses. Canadian Journal of Zoology. 1991. p. 82–90. http://dx.doi.org/10.1139/z91-013

12. Reiss RA, Schwert DP, Ashworth AC. Field Preservation of Coleoptera for Molecular Genetic Analyses. Environmental Entomology. 1995. p. 716–9. http://dx.doi.org/10.1093/ee/24.3.716

13. Wong PB, Wiley EO, Johnson WE, Ryder OA, O'Brien SJ, Haussler D, et al. Tissue sampling methods and standards for vertebrate genomics. Gigascience. 2012;1:8.

14. Anchordoquy TJ, Molina MC. Preservation of DNA. Cell Preservation Technology. 2007. p. 180–8. http://dx.doi.org/10.1089/cpt.2007.0511

15. Camacho-Sanchez M, Burraco P, Gomez-Mestre I, Leonard JA. Preservation of RNA and DNA from mammal samples under field conditions. Mol Ecol Resour. 2013;13:663–73.

16. Mulcahy DG, Macdonald KS 3rd, Brady SG, Meyer C, Barker KB, Coddington J. Greater than kb: a quantitative assessment of preservation conditions on genomic DNA quality, and a proposed standard for genome-quality DNA. PeerJ. 2016;4:e2528.

17. Zhang Y, Broach J. Abstract 5125: A novel method for isolating high-quality UHMW DNA from 10 mg of freshly frozen or liquid-preserved animal and human tissue including solid tumors. Molecular and Cellular Biology / Genetics. 2019. http://dx.doi.org/10.1158/1538-7445.am2019-5125

18. Srinivasan M, Sedmak D, Jewell S. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. Am J Pathol. 2002;161:1961–71.

19. Oosting T, Hilario E, Wellenreuther M, Ritchie PA. DNA degradation in fish: Practical solutions and guidelines to improve DNA preservation for genomic research. Ecol Evol. 2020;10:8643–51.

20. Michaud CL, Foran DR. Simplified Field Preservation of Tissues for Subsequent DNA

Analyses. Journal of Forensic Sciences. 2011. p. 846–52. http://dx.doi.org/10.1111/j.1556-4029.2011.01771.x

21. Bista I, McCarthy SA, Wood J, Ning Z, Detrich HW Iii, Desvignes T, et al. The genome sequence of the channel bull blenny, (Günther, 1861). Wellcome Open Res. 2020;5:148.

22. Kronenberg ZN, Rhie A, Koren S, Concepcion GT, Peluso P, Munson KM, et al. Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. Nat Commun. 2021;12:1935.

23. Balakrishnan CN, Edwards SV, Clayton DF. The Zebra Finch genome and avian genomics in the wild. Emu - Austral Ornithology. 2010. p. 233–41. http://dx.doi.org/10.1071/mu09087

24. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, et al. The genome of a songbird. Nature. 2010;464:757–62.

25. Klingström T, Bongcam-Rudloff E, Pettersson OV. A comprehensive model of DNA fragmentation for the preservation of High Molecular Weight DNA. bioRxiv 2018, http://dx.doi.org/10.1101/254276

26. Elmore S. Apoptosis: A Review of Programmed Cell Death. Toxicologic Pathology. 2007. p. 495–516. http://dx.doi.org/10.1080/01926230701320337

27. Doyle JJ, Dickson EE. Preservation of plant samples for DNA restriction endonuclease analysis. TAXON. 1987. p. 715–22. http://dx.doi.org/10.2307/1221122

28. Evans RK, Xu Z, Bohannon KE, Wang B, Bruner MW, Volkin DB. Evaluation of Degradation Pathways for Plasmid DNA in Pharmaceutical Formulations via Accelerated Stability Studies. Journal of Pharmaceutical Sciences. 2000. p. 76–87. http://dx.doi.org/10.1002/(sici)1520-6017(200001)89:1<76::aid-jps8>3.0.co;2-u

29. Formenti G, Rhie A, Balacco J, Haase B, Mountcastle J, Fedrigo O, et al. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. Genome Biol. 2021;22:120.

30. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nature Methods. 2012. p. 671–5. http://dx.doi.org/10.1038/nmeth.2089

31. Claus O. Wilke (2021). ggridges: Ridgeline Plots in “ggplot2”. R package version 0.5.3. https://cran.r-project.org/package=ggridges

Figure 1. Experimental design for benchmarking tissue preservation.

Graphical visualization of samples and treatments used in this study. Rows denote preservative treatments and columns temperature treatments. Colors indicate different types of tissue samples (see legend at top right). For the sea turtle samples, cells with numbers (x2 or x3) indicate conditions where samples from more than one individual were processed for comparison. All samples were transferred to –80°C after the specified temperature treatment, e.g. ‘6 hr 4C’ means stored at 4°C for 6 hours before transfer to –80°C. Abbreviations are as follows: RBCs, isolated red blood cells; EtOH, 95% ethanol; DMSO, a mix of 20–25% dimethyl sulfoxide, 25% 0.5 M EDTA, and 50–55% H2O; DNAgard, DNAgard tissue and cells cat. no. #62001-046, Biomatrica; Allprotect, Allprotect Tissue Reagent cat. no. 76405, Qiagen); RNAlater, RNAlater Stabilization Solution cat. no. AM7021, Invitrogen; FF, flash-frozen in liquid nitrogen immediately upon dissection; 6hr, six hours; 1d, one day; 1wk, one week; 5mon, five months; RT, room temperature (20–25°C). Samples were collected from these species: house mouse (*Mus musculus*), zebra finch (*Taeniopygia guttata*), Kemp’s Ridley sea turtle (*Lepidochelys kempii*), painted turtle (*Chrysemys picta*), American bullfrog (*Rana catesbeiana*), and zebrafish (*Danio rerio*).

Figure 2. Pulsed-field gel electrophoresis (PFGE) measurements of uHMW DNA comparing different sample temperature and storage times.

PFGE traces are visualized as overlapping ridgeline plots. Fluorescent stained DNA fragments are drawn with an electric current from the well at the right towards the left. Smaller fragments generally travel farther than larger fragments. The fragments that greatly exceed the targeted size range remain in the well and can not be reliably interpreted. Each ridgeline plot corresponds to a gel lane and a single DNA extract with brightness converted to a plot profile. The x-axis denotes molecule length scaled via piecewise linear scaling to match across gels of different lengths with a common size standard (Lambda PFG Ladder, New England Biolabs). The x-axis is the same in both columns. The y-axis of each plot is brightness scaled proportionally in each gel lane from just below the well to just beyond the 48.5 kb ladder peak such that the relatively intense brightness of the well itself is excluded from scaling. The well brightness is cropped where it exceeds the brightness of the rest of the gel lane. DNA fragments with lengths longer or shorter than peaks of the size standard can not be reliably interpreted due to lack of size reference and artifacts of gel electrophoresis as well as limitations of any type of gel electrophoresis to correctly size megabase-length fragments. Colors represent different sample preservation methods, as indicated in the legend at bottom right. All samples were transferred to –80°C after the specified temperature treatment, e.g. ‘6hr 4C’ means stored at 4°C for 6 hours before transfer to –80°C. Abbreviations are as follows: RBCs, isolated red blood cells; EtOH, 95% ethanol; DMSO, a mix of 20–25% dimethyl sulfoxide, 25% 0.5 M EDTA, and 50–55% H2O; DNAgard, DNAgard tissue and cells cat. no. #62001-046, Biomatrica; Allprotect, Allprotect Tissue Reagent cat. no. 76405, Qiagen); FF, flash-frozen in liquid nitrogen immediately upon dissection; 6hr, six hours; 1d, one day; 1wk, one week; 5mon, five months; RT, room temperature (20–25°C). Three additional samples were tested, but produced insufficient DNA for fragment length analysis: frog muscle in DMSO for one week at 4°C and 6 hr at 4°C and mouse spleen in DMSO for 16 hr at 4°C. For

measurements based on the FEMTO pulse instrument and additional tissue types, see **Figs. S2, S3**.

Figure 3. Testing the effect on two measures of uHMW DNA quality. Distributions of sample groups are overlaid with results of linear modeling of fragment length (n = 102, **a-c**) and DNA yield (n = 139, **d-f**). Shown are univariate scatterplots overlain with box plots indicating the median, quartiles, and full range of individual observations. Fragment length was quantified here as the proportion of pulsed-field gel electrophoresis (PFGE) signal above 145 kb, and was modeled in a generalized linear model with temperature (**a**), preservative (**b**), and tissue type (**c**) as predictors. DNA yield per input mass was log-transformed and modeled with temperature (**d**), preservative (**e**), tissue type (**f**), and vertebrate group as predictors. Significant relationships from post-hoc comparisons are shown as connecting bars with significance levels: **** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Sample sizes for each factor are given along the x-axis.

Figure 4. Hi-C platform benchmarking of bird samples. Stacked bar plots denoting proportions of Hi-C reads mapped to the zebra finch reference genome involving different chromosomes (*trans*), on the same chromosome but less than 15 kb apart (*cis* < 15 kb), and on the same chromosome and greater than 15 kb apart (*cis* > 15 kb). Tested samples include blood samples (**a**-**c**), and muscle samples (**d**-**f**). The desirable outcome is to have much greater proportions of Hi-C reads being long-range *cis* pairs, which reflects an efficient capture of long-range interactions needed for genome scaffolding and haplotype phasing. Hi-C data was generated by Arima Genomics following their standard protocol.

Figure 5. Considerations for collection of tissues for long-read sequencing of non-model organisms.

General representation of a sequencing pipeline and considerations that may directly or indirectly affect the quality of sequencing output. Stars indicate particular sources of variation manipulated in this study. Several logistical aspects need to be considered prior to sample collection for uHMW DNA isolation with the goal of producing reference-quality genomes. The collector needs to identify what tissue types can be collected from the target species, what preservation methods and cold storage are available, and how quickly samples can be transported to a –80ºC ultra-cold freezer.

Table 1: Sample collection guidelines for generating high-quality genomes.

Compiled here are guidelines based on the best-performing protocols tested in this study and broadly in the Phase 1 VGP genomes.

Tissue selection	Tissues listed in decreasing preference. Multiple tissue types should be collected when possible.	
	Fish	soft tissues; muscle; body with head, digestive tract, and swim bladder removed
	Amphibians	blood, muscle
	Birds	blood, muscle
	Non-avian Reptiles	blood/isolated red blood cells, muscle
	Mammals	soft tissues like spleen, muscle
Preservation		
	Ideal:	**Flash freezing or short-term refrigeration before deep freeze** Blood or tissue specimens in 95% EtOH or 20-25% DMSO-EDTA can be stored at 4°C or on ice for up to 6 hours after dissection with little to no decrease in sample quality relative to immediate flash freezing.
	Good:	**Mid-term refrigeration before deep freeze** Samples in 95% EtOH or 20-25% DMSO-EDTA can be stored for longer periods on ice/4°C of up to one week with minimal potential decrease in sample quality.

Acceptable:	**Mid-term room temperature storage before deep freeze** Blood in 95% EtOH can be stored at room temperature (20–25°C) for up to one week with some potential decrease in DNA quality, most likely yielding extracts still within acceptable parameters for current long-read sequencing platforms. This condition is less likely to yield acceptable results with tissue samples.

Figure 1

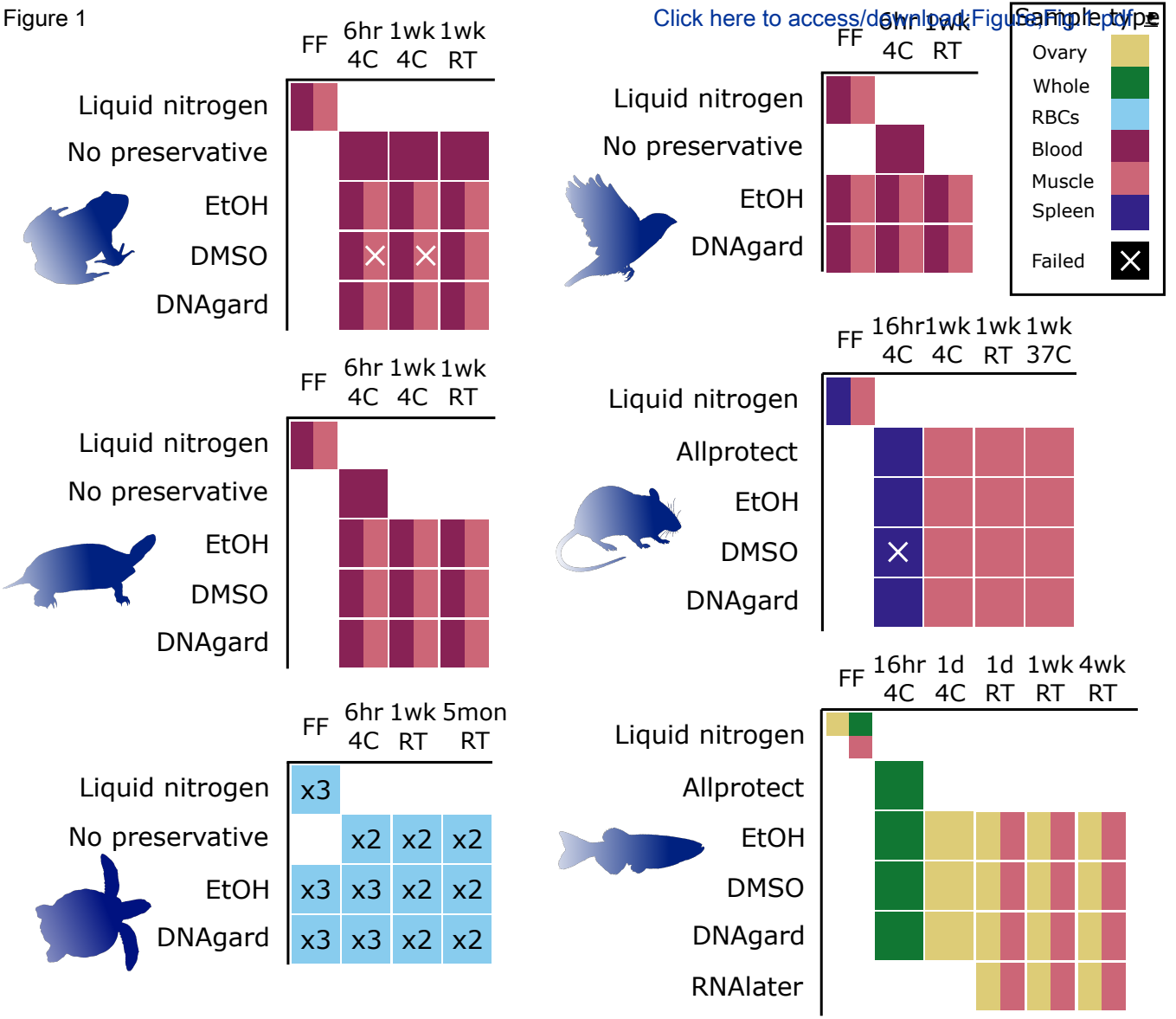
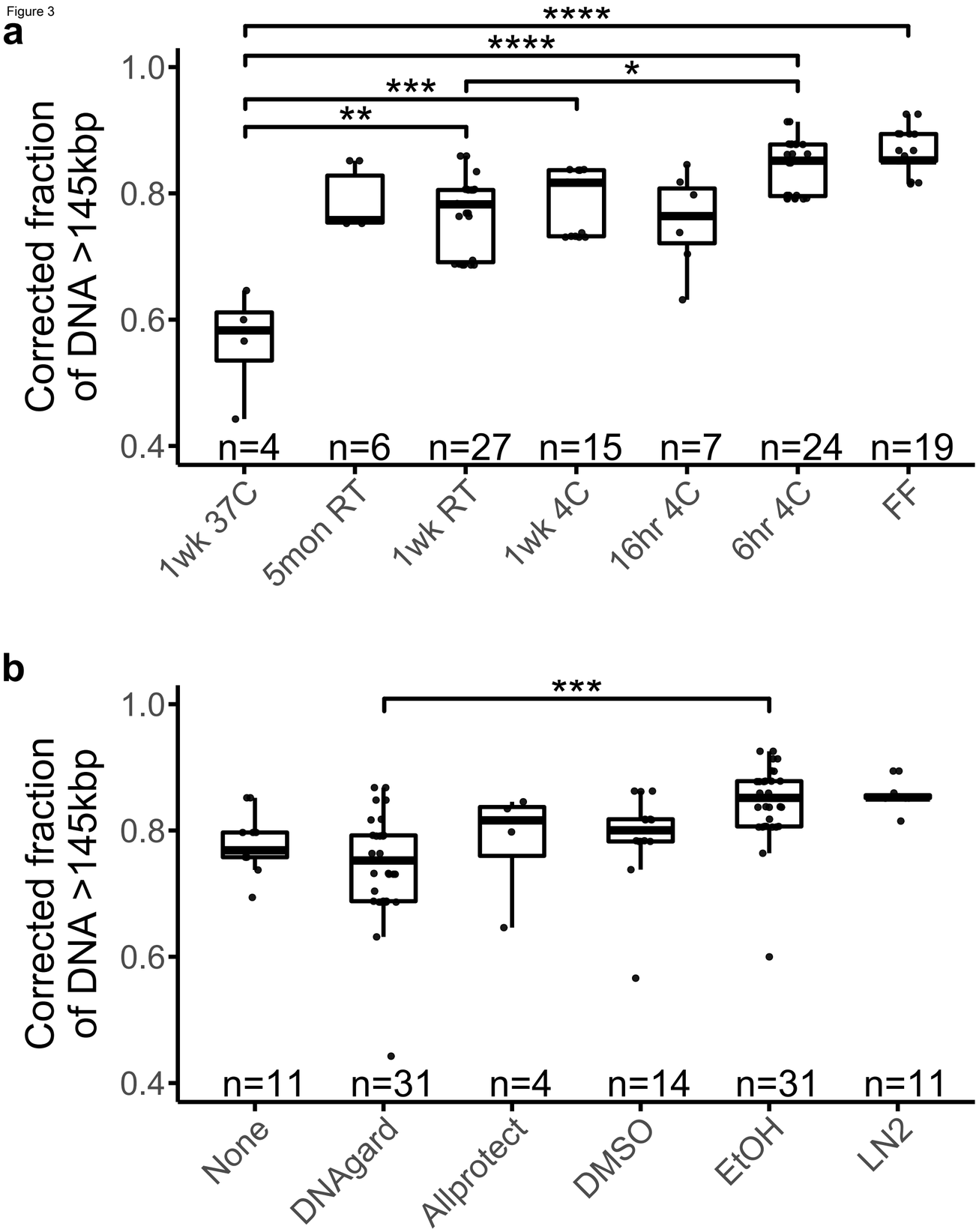
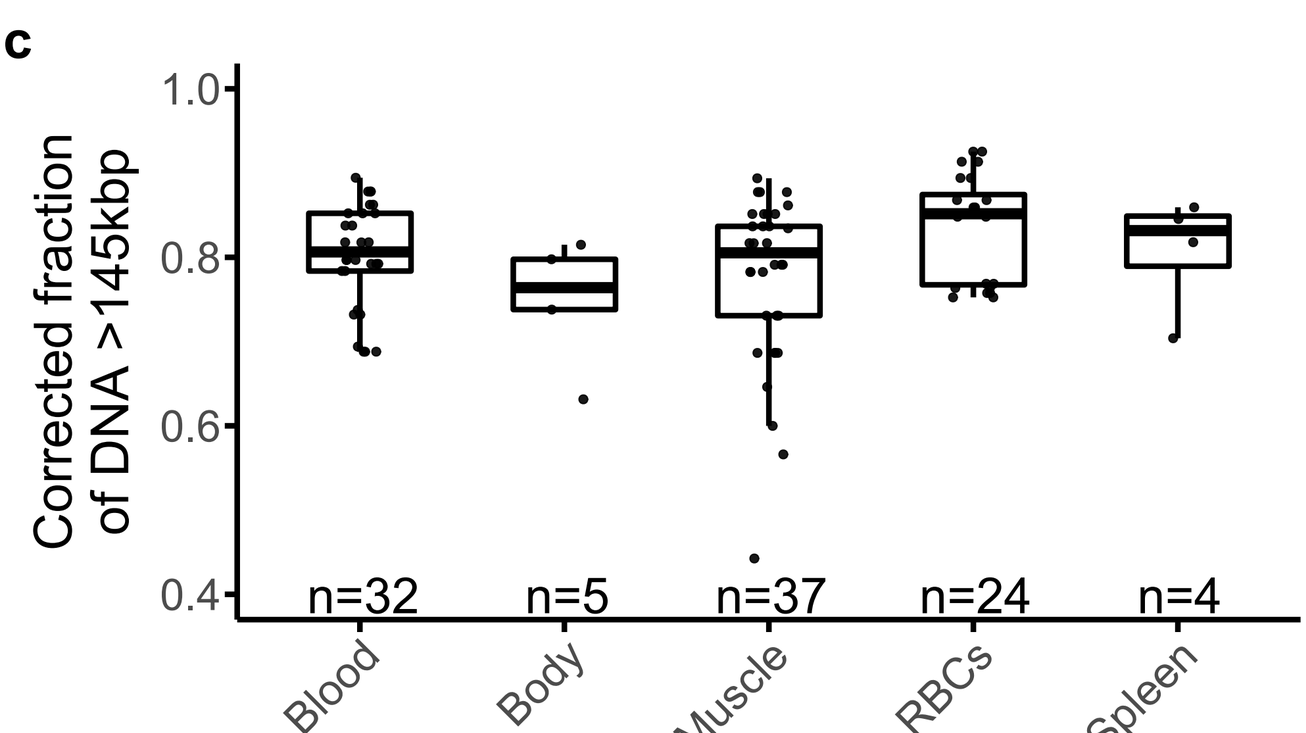
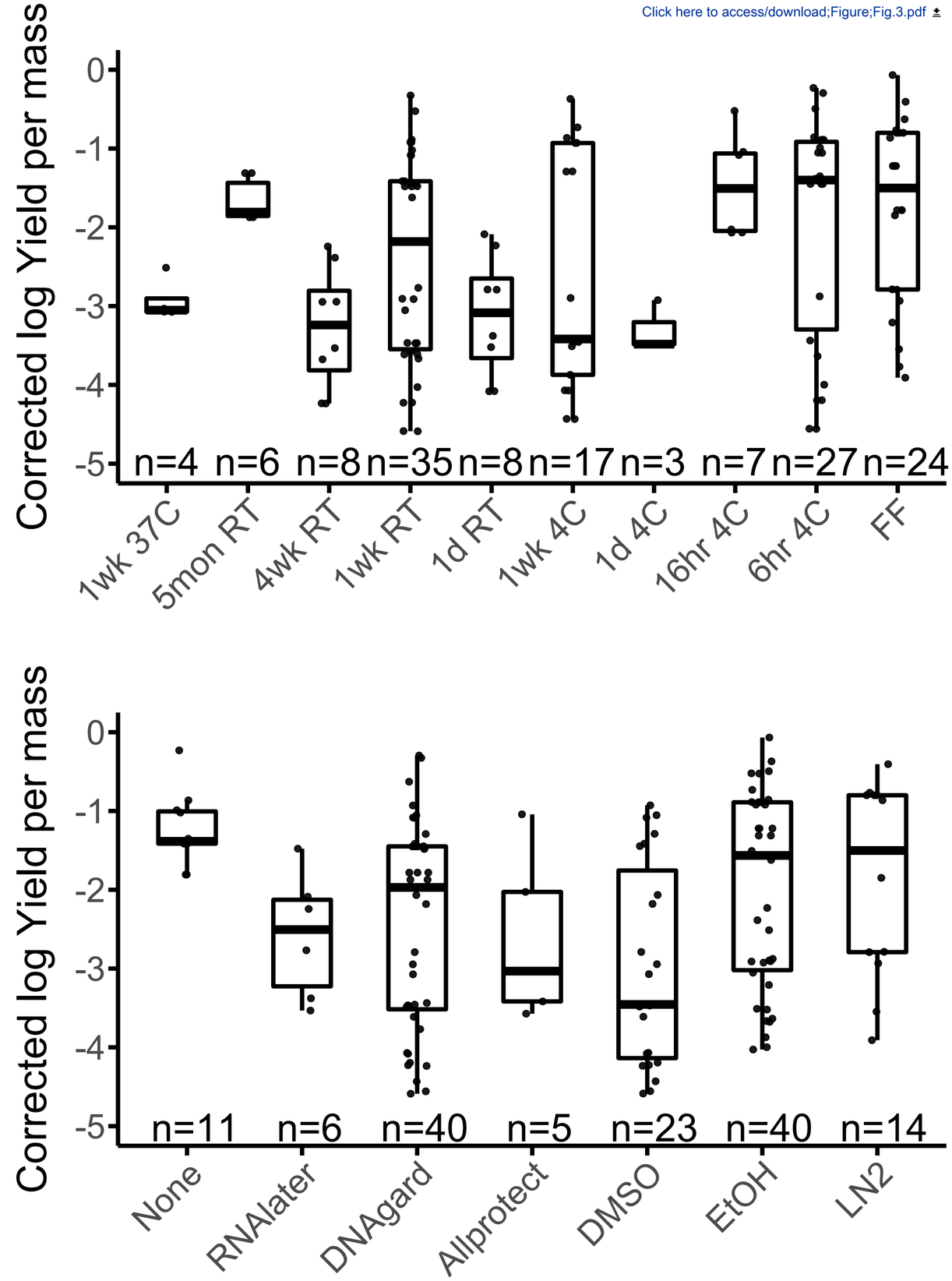


Figure 3







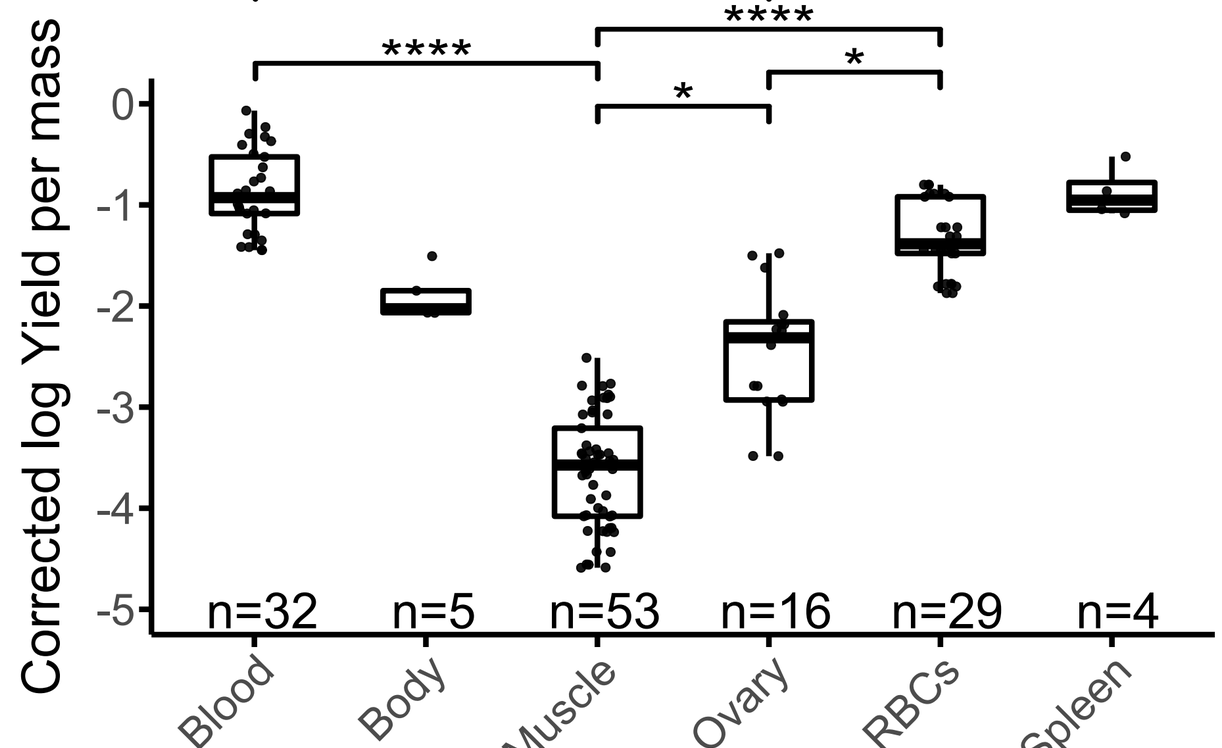


Figure 4

Click here to access/download;Figure;Fig.4.pdf

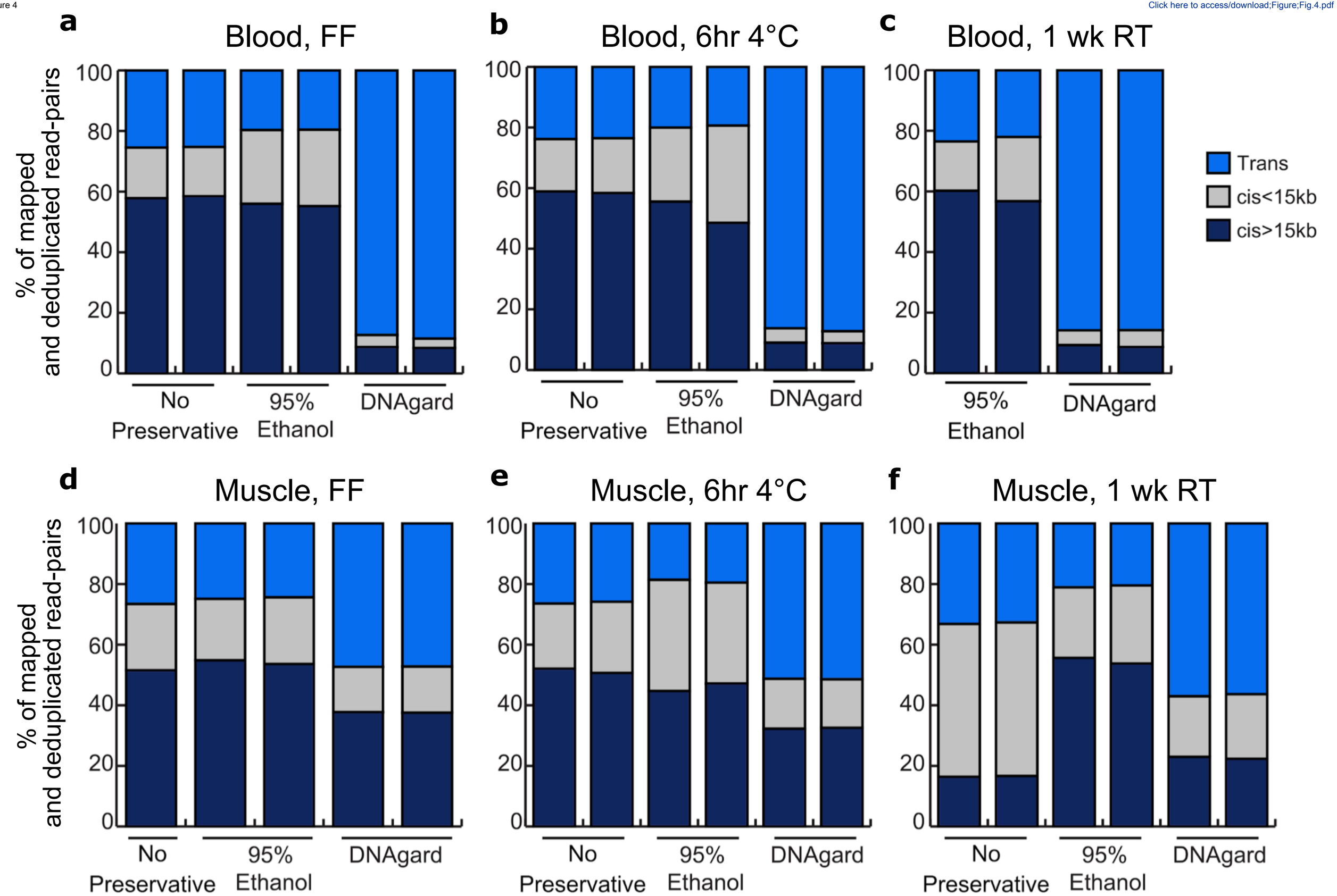
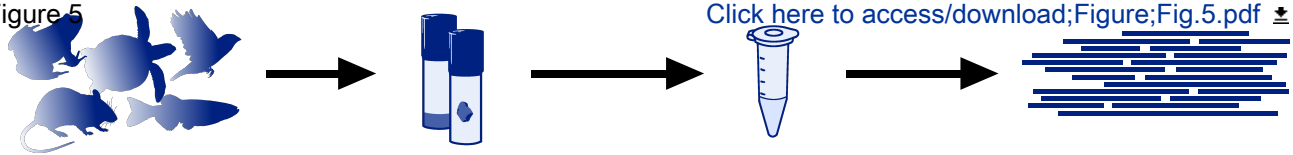


Figure 5

Click here to access/download;Figure;Fig.5.pdf



Sample Collection

Biology
- tissue type ★
- size limitations
- genome size

Collection
- lethal vs. non-lethal
- euthanasia method
- time between death and collection
- contamination prevention

Preservation
- preservative solution ★
- dry vs. wet preservation
- tissue surface area
- sample to preservative ratio

Sample Storage and Transport

Short-term
i.e. field conditions
- field storage temperature ★
- time in field storage ★
- temperature fluctuations
- agitation in transport

Long-term
- storage temperature
- time in storage
- freeze/thaw cycles
- preservative replacement

DNA Extraction and Handling

Extraction
- extraction technology
- input sample amount
- cleaning steps
- yield amount

DNA Storage and transport
- shipping conditions
- storage temperature

general in-lab handling technique

Quality Control

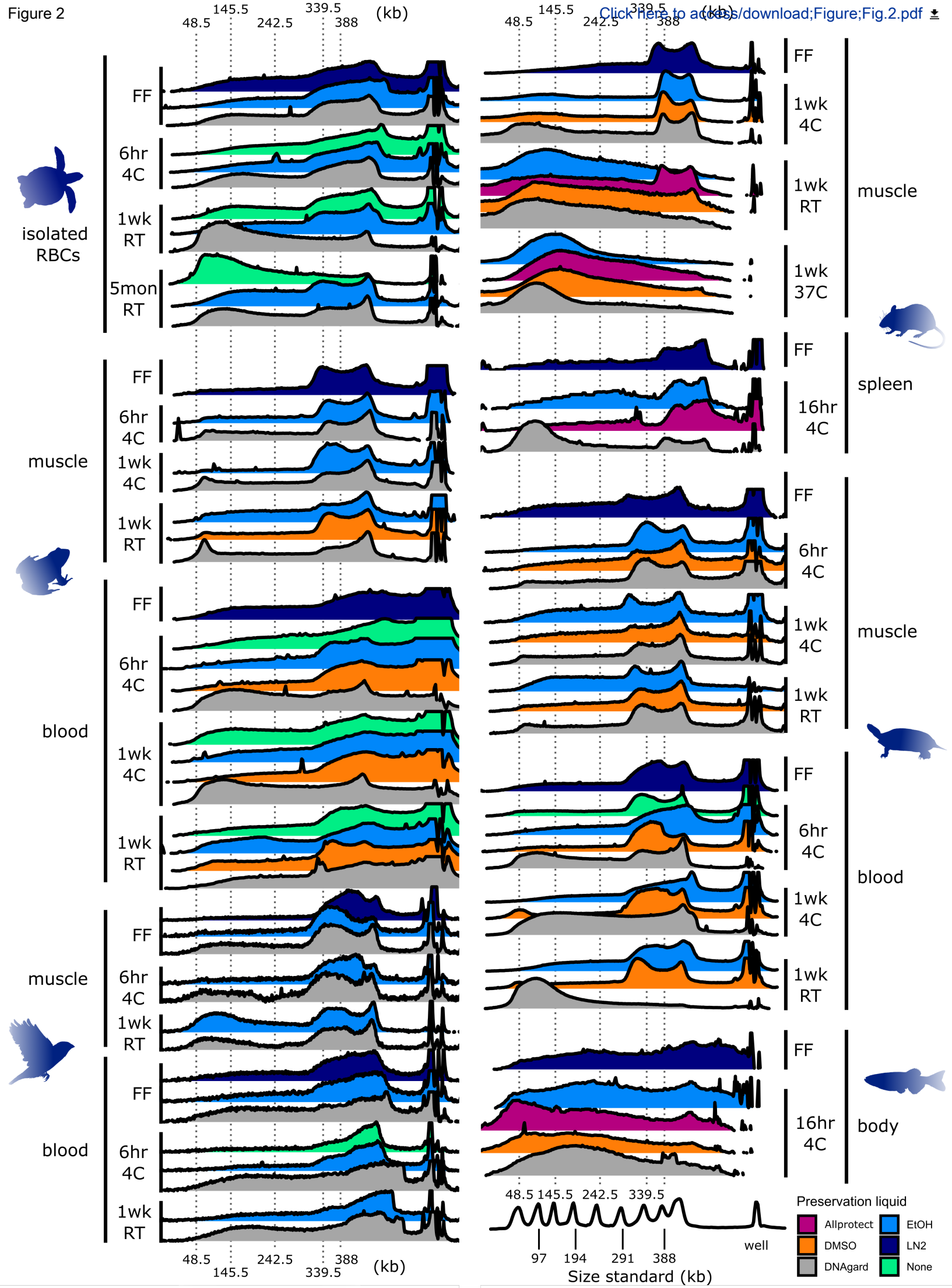
Assessment method
- fragment length assessment technology
- robustness of measurements

Sequence Generation

Pre-processing
informed by quality control
- size selection
- labeling
- library prep

Sequencing method
- technology
- platform
- parameter choice

Figure 2



Supplementary Table 1

Supplementary Table 2

Click here to access/download

Supplementary Material

Supplementary Table 2.xlsx

Supplementary Table 3

Click here to access/download

Supplementary Material

Supplementary Table 3.xlsx

Supplementary Figure 1

Supplementary Figure 2

Supplementary Figure 3

Click here to access/download
Supplementary Material
Fig.S3_supplemental material.pdf

Supplementary Figure 4

Click here to access/download

Supplementary Material

Fig.S4_supplemental material.pdf

Supplementary Figure 5

Click here to access/download
Supplementary Material
Fig.S5_supplemental material.pdf

SCIENCE FOR THE BENEFIT OF HUMANITY

Dr. Olivier Fedrigo, Ph.D.
Vertebrate Genome Laboratory
The Rockefeller University
Box 366
1230 York Avenue
New York, New York 10065

Tel: (212) 327-8216 | Fax: (212) 327-8276
Email: ofedrigo@rockefeller.edu
Website: https://vertebrategenomelab.org/

Dear Dr. Hans Zauner,

Thank you for the opportunity to revise and resubmit our article titled “Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing” to GigaScience. We greatly appreciate the effort put forth by you and the reviewers in providing feedback on our manuscript. The comments we received are insightful suggestions that we feel have improved the manuscript. We have been able to incorporate changes based on these contributions, as highlighted below.

Comments from Reviewer 1:

R. 1, Comment 1: *In figure 2 the size distribution of DNA fragments is visualized from the different experiments. Most of the fragment distributions look like I would have expected them based on the work we did in the article cited as nr 25 in the reference list. However the muscle tissue from rats and the blood samples from the mouse and the frog indicates that there may be a misinterpretation in the article regarding the actual size distribution of fragments which needs to be looked in to.*

Response: The following comments indicate two major limitations to this methodology: the interpretable range of fragment sizes, and streaking artifacts. In the article we endeavor to interpret these results within the bounds of those limitations. We’ve added additional text and made figure modifications, detailed below, to clarify these caveats.

R. 1, Comment 2: *Starting with the mouse plots and especially the muscle one. There must either have been a physical shearing event that drastically reduced the size of DNA (using the terminology from ref 25 this would mean that physical shearing generated a characteristic fragment length of approximately 300-400 kb), or the lack of a sharp slope on the rightmost side of the ridgeline plot is due to the way the image was processed. All other animals got a peak on the rightmost side of the ridgeline plot and the agarose plug should, based on the referenced methods paper [7], generate megabase sized fragments which far exceed the size of the scale used in figure 2. I would presume these larger fragments would get stuck in or near the well which makes*

it easy to accidentally cut them out when doing the image analysis step which may explain their absence in the mouse samples. This leads me to the conclusion that the article is well designed to capture the impact of chemical shearing caused by different preservation methods but would benefit from evaluating whatever figure 2 properly covers the actual size distribution of fragments or only covers the portion of DNA fragments small enough to actually
form bands on the PFGE gel with a substantial part of the DNA stuck in or near the well.
Response: This is a correct indication of the limits of visualizing DNA fragment length distributions with current technologies, only fragments that fall within a given range can be viewed at once with any reasonable resolution. We targeted a range of fragment sizes that includes ideal sizes for long-read and long-range sequencing while also giving indications of degradation in the smaller end of the range. Fragments outside the target range can not be reliably interpreted. In PFGE, DNA fragments larger than the target range can be stuck in or near the well (right side of the plots). Unfortunately, in the well they then become indistinguishable from bright reflections off the edges of the well and are further obscured when streaking is present. We attempted early on to score samples for presence/absence of signal in or near the well, but found this to not be repeatable or informative using the gel images. Thus, we chose to omit the well itself and the space immediately around it from interpretation, from measurements for statistical testing, and from Fig. 2. We only interpret the portion of the PFGE gel where bands can form and where the standard can give us some indication of fragment size. The interpretable range of fragment sizes in these gels still offers important indicators of sample quality.
For clarity, the new version of Fig. 2 includes the well in each plot profile. The well peaks are cropped where they exceed the peak brightness of the rest of the lane. We do not consider the well brightness to be a reliable indicator of sample quality, but recognize that some readers may want to see the full pattern.
The following has also been added to the Fig. 2 legend: “Fluorescent stained DNA fragments are drawn with an electric current from the well at the right towards the left. Smaller fragments generally travel farther than larger fragments. The fragments that greatly exceed the targeted size range remain in the well and can not be reliably interpreted.”
“The well brightness is cropped where it exceeds the brightness of the rest of the gel lane.”
We have also added this text to the discussion section of the manuscript: “Additionally, we are only able to visualize DNA fragment size distributions within a certain range of sizes (approx. 40–400 kb for PFGE, 1.3–165 kb for FEMTO). Though we have targeted a size range that includes both ideal fragment sizes for long-read sequencing and fragments of lower molecular weight that may indicate degradation, fragments outside this range are not measured here.”

R. 1, Comment 3: *The frog plot is a good example of how this may influence our interpretation of the ridgeline plots. If the extraction method generate high-quality DNA concentrated in the 300-400 kb range then there must be something very special with the frog DNA from blood as there is a continuous increase in the brightness all the way to the edge of the image. This implies that the sample contains a high amount of much larger DNA fragments than the other samples. I find this rather unlikely and if I saw this in my own data I would assume that we had a lot of very large DNA fragments that are out of scale for the gel electrophoresis but that in the case for the frog blood samples many of these fragments have been chemically sheared creating the "smeared" pattern we see in figure 2.*

Response: Yes, the frog blood especially exhibits a “streaking” pattern in the gel where there is a strip of continuous brightness in the lane. This is another reason why we do not attempt to interpret the gel above where bands form in the lane. Before our initial submission, we performed a repeat run of the PFGE gels with streaking, but they produced identical results with streaks still present. Samples with this streaking pattern have performed well in past sequencing efforts, and it’s generally thought to be an indicator of high quality samples. However, barring more conclusive testing of this pattern, we do not attribute streaking as an indicator of quality in this manuscript.
We have added this section to the figure caption: “DNA fragments with lengths longer or shorter than peaks of the size standard can not be reliably interpreted due to lack of size reference and artifacts of gel electrophoresis as well as limitations of any type of gel electrophoresis to correctly size megabase-length fragments.”

R. 1, Comment 4: *Dryad DOI doesn't work for me.*

Response: It is possible that the reviewer was attempting to open the link provided in the cover letter, which has formatting mistakes related to the uploading process. A correct link was supplied to Dr. Zauner. Apologies for this inconvenience. Here is the correct and updated Dryad download link for reviewers:
https://datadryad.org/stash/share/uHgVucrNICiMT-Y92O4M4Km3S4DyK3UJFA3qMJEbm4M

R. 1, Comment 5: *Figure 1 - The meaning of x3 and x2 for the turtle should be described in the caption.*

Response: The Figure 1 caption now has the added sentence “For the sea turtle samples, cells with numbers (x2 or x3) indicate conditions where samples from more than one individual were processed for comparison.”

R. 1, Comment 6: *Figure 2 - Having the scale indicator (48.5. 145.5 etc) at the top as well as the bottom of each column would make it quicker to estimate the distribution of samples.*
Response: Agreed. This has been added to Figure 2.

R. 1, Comment 7: *The article completely omits Nanopore sequencing, is there a specific reason for why lessons here are not applicable to ONT?*
Response: Nanopore equipment and expertise were not available at the time of this study, but will be part of further testing. We generally expect the same indicators of sample quality to correlate with successful sequencing in Nanopore sequencing as with other technologies, though it is not explicitly tested here. We've added a mention of Oxford Nanopore to the list of relevant technologies in the introduction to clarify: "Long-reads (generally > 10 kb; e.g. Pacific Biosciences or Oxford Nanopore), long-range molecules (generally > 50 kb; e.g. 10X Genomics linked reads), or optical mapping (> 150 kb; e.g. Bionano Genomics), and Hi-C proximity ligation (> 1 Mb; e.g. Arima Genomics) can span repeats thousands of base pairs in length [4], greatly improving assembly outcomes."

R. 1, Comment 8: *There is a very interesting paragraph starting with "The ambient temperature of the intended collecting locality should be a major consideration in planning field collections for high-quality samples. Here we test a limited number of samples at 37°C to". Even if the results were very poor information about the failed conditions would be appreciated. What tissues/animals did you use, did you do any preservation at all for the samples and did you measure the fragment length distribution anyway? Simply put, even if the DNA was useless for long read sequencing it is an interesting data point for the dynamics of DNA degradation and a valuable lesson for planning sampling in warm climates.*
Response: The "limited number of samples" refers to the four mouse muscle samples reported with the rest of our results. No further samples were tested at 37°C in this study. We've modified that sentence for clarity as follows: "Here we test a limited number of samples at 37°C to resemble fieldwork conditions in warmer climates, resulting in no retention of workable amounts of uHMW DNA in any of these samples (4 mouse muscle samples; Fig. 2)."

Comments from Reviewer 2:
R. 2, Comment 1: *Although the effectiveness of the tissue/preservative combination was only tested with the preparation of long range libraries, it would have been useful to select one or two cases for long range sequencing (PacBio or Oxford Nanopore) to explore the impact of the different QC parameters measured in this study.*

Response: We agree that testing samples on long-read sequencing platforms would have been very useful. Unfortunately, the expense of long-read sequencing was prohibitive at the time. We do find that the results align generally with the experience of the Vertebrate Genomes Project. See response to the related comment #4 below for further details.

R. 2, Comment 2 (in text): *space between quantity and unit symbol*
Response: This change has been applied to the text.

R. 2, Comment 3 (in text): *such as used twice in a sentence. please edit*
Response: This sentence has been revised.

R. 2, Comment 4 (in text): *Your work is a great contribution to the genomics field! However, DNA integrity and optimal QC parameters (Absorbance ratios at 260/230, 260/280, double stranded DNA proportion from a total gDNA prep) are not always predictors for Long Read Sequencing success. I am aware of the high cost that you would face if all these samples were sequenced, even one flowcell/sample in minION, could cost a little fortune. But it would be fantastic if you could please indicate if any of the sample/preservation combinations showed consistent good LRSeqeuncing results.*
Response: We concur that simply checking the DNA integrity and yield as QC parameters might not give a definitive indicator of sequencing success with long reads. Moreover, this particular paper has been concentrating on Vertebrata, a relatively small taxonomic group as compared to Arthropoda, Planta, etc.

Additionally, as new reagents for DNA preservation are constantly emerging on the market (e.g. Allprotect), sample collection committees within VGP, EBP, ERGA and other large reference genome sequencing initiatives will continue monitoring and studying the impact of those products on the sequencing outcome, but those would be subjects of other studies, publications, and public guidelines.

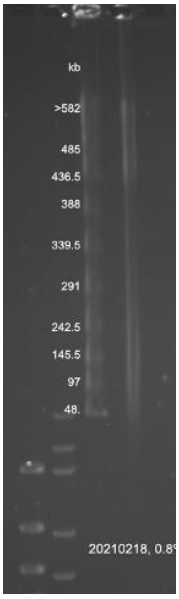
From the experience of several large PacBio sequencing providers, including the facilities involved in this publication, we know that the chemical purity of the HMW-DNA sample is at least as important as the molecule integrity. There is, however, no single definitive analytical parameter that has ever been defined for predicting the long-read sequencing outcome based, or for detection of any carry-over contaminants or significance to sequencing success. For this reason, chemical purity parameters were not within the scope of this manuscript, though they likely carry influence outside the variables manipulated in this study.

We have instead concentrated this study on the low-hanging fruit of integrity of HMW-DNA molecules and DNA yield, both of which robustly and intuitively influence sequencing success. From that point of view, our smaller sequencing tests and general experience outside the scope of this manuscript corroborate the results presented in the study. Two of these smaller tests are detailed below.

Recent test 1: Flash-frozen vs. EtOH-preserved reptile tissue
We compared the length of uHMW gDNA and performance on Bionano and PacBio continuous long reads of snap-frozen and EtOH-preserved nucleated blood from reptiles and found no significant changes in performance in Bionano molecule size and PacBio CLR sequencing subread length. The average length of unfiltered Bionano reads was longer when gDNA was extracted from flash-frozen tissue, though both treatments still returned results in an acceptable range. In general, Bionano optical mapping was working reliably for gDNA from EtOH-preserved tissues.
Some of the quality specifications are shown in the table below.

Preservation	Temperature	Number of species	PFGE size (kb)	Bionano average unfiltered molecule size (kb)	PacBio average subread length (kb)
Liquid N2 only	Flash-frozen	2	40 to > 400	168	21,7
EtOH	Flash-frozen	4	40 to > 400	130	21,4

Recent test 2:
We used mammalian kidney tissue that had been stored in Allprotect recently for PacBio HiFi and Hi-C (ARIMA protocol). Right after tissue extraction, the tissue was soaked in Allprotect at room temperature overnight, stored for about 1 year at -20C degrees, then shipped, and finally stored at -80 degrees before sequencing. The sample showed no indication of any contaminants on a Nanodrop spectra readout. Please see the PFGE gel image and summarized sequencing results below.



Preservation	Temperature	Length of gDNA	PacBio HiFi reads	HiC performance
Allprotect	-20C, 1 year	50 kb to > 400 kb	ccs yield 17,9 Gb (average fragment length 15,6 kb)	ok, 23% duplicated reads

R. 2, Comment 5 (in text): *Did you load the same amount of DNA per PFGE lane for all samples?*

Did you heat up the DNA sample + loading buffer before loading? 1-2 min at 65C followed by cooling at room temperature ~ 5 min before loading helps to prevent clumping.

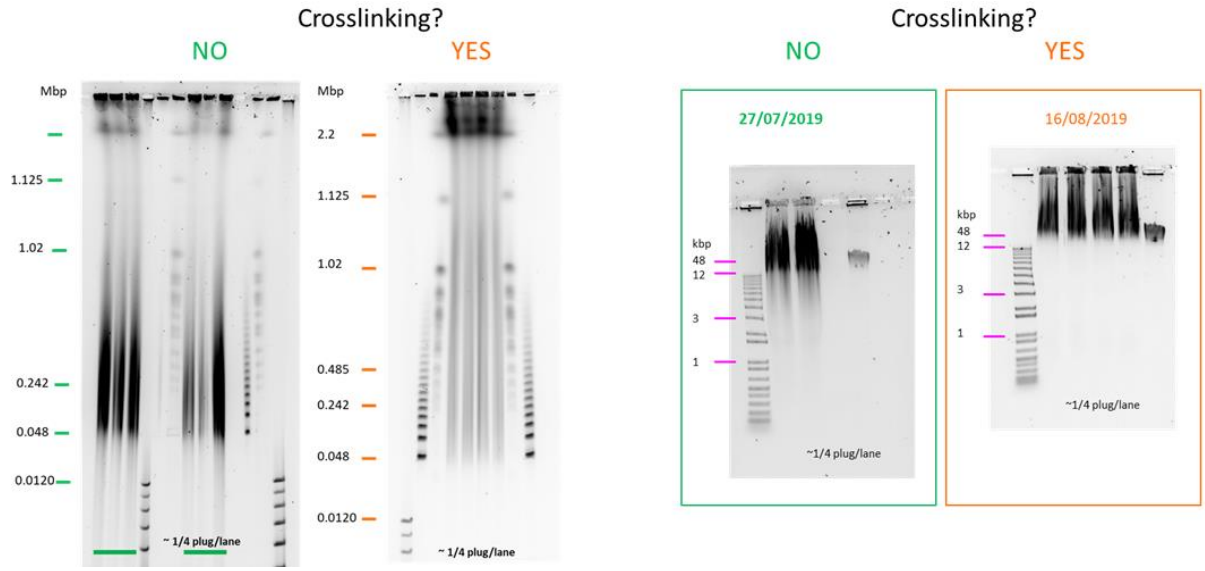
Did you run a slice of each plug as a control for the DNA manipulation factor that could cause fragmentation while extracting the DNA from the plug?
Response:
Yes, we loaded approximately 100 ng of DNA per well.

No, we did not heat DNA and loading buffer prior to loading. DNA, loading buffer, and TE buffer were kept at room temperature before mixing and then loading. DNA was

loaded after at least one week at room temperature, which allows for homogenization of the sample and increases hydration of the DNA molecules in the sample. Although the original HMW DNA samples are often viscous, the addition of TE buffer and loading buffer dilutes them to the point that we find clumping is reduced.

No, each plug was carried through entirely for DNA extraction. Digesting the entire plug was integral to comparing DNA yield. The process to extract DNA from the plug is quite gentle (slow shaking for lysing and washing steps, Agarase digestion, drop dialysis) and the plug also protects the DNA.

R. 2, Comment 6 (in text):



The picture shows embedded nuclei with or without crosslinker, but the same could be done with the extracted DNA from plugs

Response:

We absolutely agree that the conventional field-inversion PFGE instruments (e.g. BioRad CHEF) are much better suited to resolve sub-megabase size DNA fragments from *in situ* extractions. However, this specialized equipment is currently much less common as compared to a cheaper (albeit less precise) Pippin Pulse system from SAGE which is widely used at sequencing facilities.

R. 2, Comment 7 (in text): *Perhaps analyzing a sample in a standard agarose gel might help evaluate fragmentation < 20 kbp.*

Response: This is certainly something that could be incorporated into future testing as another metric of the smaller fragment sizes, perhaps as a cheaper alternative to FEMTO. Unfortunately, more testing at this point on the same extractions would not be

useful; they are now several years old. Our subset of samples tested on the FEMTO Pulse system give a detailed perspective on <20 kb fragments.

R. 2, Comment 8 (in text): *It might help to describe briefly how Arima prepared the nucleated blood Hi-C libraries, especially the samples preserved in DNAgard. I think this solution does contain an inhibitor of the crosslinking reaction, most likely a free amino group (from Tris buffer, for example). The Dovetail Genomics Omni-C kit protocol dilutes the nucleated blood preserved in EDTA tubes in 1 mL 1X PBS buffer and collects the cells by centrifugation before the crosslinking steps. Sorry I don't have access to the Arima protocol document to make a more informed comment.*
Response: Our procedure for nucleated blood in a solution like ethanol (or DNAgard) is to pellet the cells, remove the supernatant, wash with 1X PBS containing 1% FBS, and then carry the washed pelleted cells into crosslinking and then Arima-HiC. Given our washing procedure, it seems less likely (although still possible) that residual tris is inhibiting the crosslinking reaction. Interestingly, DNAgard is a proprietary solution originally developed by Biomatrica, and so we were not able to find any resource that pertains to what the solution is actually composed of.
We've added a citation of the protocol document number and this note to the methods: "Briefly, standard protocol for nucleated blood in a solution like EtOH or DNAgard is to pellet the cells, remove the supernatant, wash with 1X phosphate buffered saline solution containing 1% Fetal Bovine Serum, and then carry the washed pelleted cells into crosslinking and then Arima-HiC.", and this to the discussion: "Though our washing protocol should minimize its effect, it is also possible that some unknown aspect of the DNAgard treatment of cells inhibited the crosslinking reaction, and Hi-C of unfixed cells would be expected to have low signal and high noise similar to degraded DNA."

Additional clarifications:

INSDC submission - We have removed the Hi-C sequence reads from the manuscript's associated Dryad repository and are in the process of uploading them to the publicly accessible Sequence Read Archive.
Analysis scripts - Two commented scripts have been added to the manuscript's associated Dryad repository. One contains statistical analysis of DNA yield and fragment length reported in the manuscript, and the other has basic bioinformatics and calculations based on the Hi-C reads reported in the manuscript.
The data availability section of the manuscript has also been updated to reflect these changes.

Please feel free to notify me of any further comments or questions. We look forward to your response.

Thank you again for your hard work and for this opportunity.

Sincerely,

Olivier Fedrigo, Ph.D.
Director
Vertebrate Genome Laboratory