Incorporating Ultra-Long ONT reads in VGP High Quality Reference Genome Assemblies

Nivesh Jain[1†], Tatiana K. Tilley[1†], Linelle Ann L. Abueg[1], Jennifer Balacco[1], Nadolina M. Brajuka[1], Giulio Formenti[1], Erich D. Jarvis[1,2], and Olivier Fedrigo[1]

(1)Vertebrate Genome Lab, The Rockefeller University, New York, NY (2)Howard Hughes Medical Institute, New York, NY

† indicates authors contributed equally to the work

ABSTRACT: The Vertebrate Genomes Project (VGP) aims to create an open-access genome repository cataloging at least one high-quality, near-gapless, chromosome-level, phased, and annotated reference genome assembly for all ~70,000 extant vertebrate species. Technologies are rapidly developing and in the short time since the formation of the VGP consortium, the pipeline utilizes a combination of long-read and long-range technologies. Our group, the Vertebrate Genome Lab (VGL), is one of the main data production and assembly hubs for the VGP. Samples acquired from an extensive network of collaborators are bio-banked into the collection before undergoing a variety of specialized preparation techniques. These include high molecular weight DNA extraction methods as well as quality control to assess fragment length and appropriate yield in preparation for PacBio sequencing and Bionano optical mapping. PacBio high fidelity (HiFi) sequences are used to create the initial assembly, and then optical maps and Hi-C sequences are used for scaffolding, which increase the contiguity of the assembly (VGP assembly pipeline v2.1). We at VGL are currently optimizing our Oxford Nanopore Technologies (ONT) workflow for a wide variety of non-model organisms that have not been previously sequenced on this platform, from DNA extraction to library preparation and sequencing, aiming to maximize N50, long reads (>100kb), and data output. This necessitates working with an extremely diverse set of samples in terms of taxonomy, tissue type, and quality. As such, it is challenging to formulate one workflow that produces dependable results. We present results from varying extraction methods and library preparation modifications.

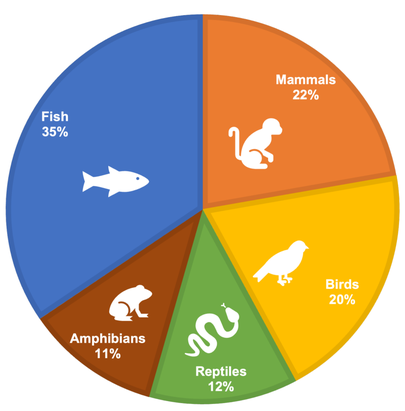
These ultra-long ONT reads will be used to support existing genome assemblies generated at VGL. The final chromosome-level assembly is manually curated and immediately made public before submission to the National Center for Biotechnology Information (NCBI) for annotation. These high-quality genomes are an essential resource for groups working in conservation, comparative genomics, non-model organism systems, and zoonotic epidemiology.

THE VERTEBRATE GENOMES PROJECT

We are here!

PHASE 1 Orders 268 | **PHASE 2** Families ~1,100 | **PHASE 3** Genera ~10,000 | **PHASE 4** Species ~70,000

ORDERS IN PHASE I



The **Vertebrate Genomes Project (VGP)** is an international and multidisciplinary project whose goal is to generate high quality reference genome assemblies for all ~70,000 extant vertebrate species. The project is divided into four phases based on systemic classifications. Phase I, which includes 261 orders, began in 2017 and is almost completed as of December 2022.

METHODOLOGY

1. DNA ISOLATION

We tested the **Monarch® HMW DNA Extraction Kit for Blood (**New England BioLabs, USA) with modifications for UHMW DNA, using samples of **blood in ethanol** from the **Olive Python** (*Liasis olivaceus*). This kit uses a **glass bead capture technology** to give larger DNA fragments.

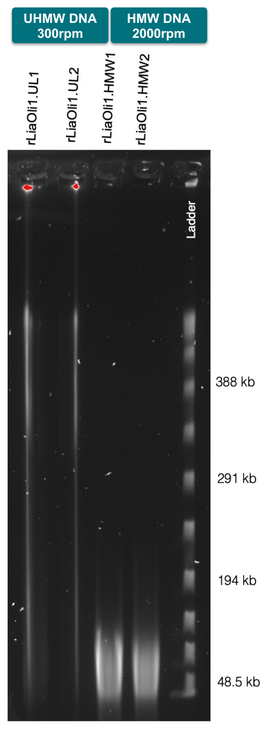
World of Reptiles.

The only major modification for UHMW extraction is the agitation speed at which the cells are lysed on the thermomixer, which is **2000rpm/10min./56°C** for the standard HMW protocol and **300rpm/10min./ 56°C** for the protocol modified for UHMW DNA.

2. QUALITY CONTROL (CONCENTRATION AND FRAGMENT LENGTH)

We used the **Qubit™ dsDNA Broad-Range Assay Kit** (Invitrogen™, USA) to quantify the DNA extractions with some slight modifications. **Triplicate readings** were taken for each extraction and the samples were **sonicated** before the addition of dye and buffer solution in order to shear the UHMW and HMW DNA for relatively more accurate concentration readings. UHMW and HMW DNA is **difficult to quantify**, and to accurately determine the final mass of eluted DNA is almost not possible due to a **high CV%** (as visible in the table below).

Extraction ID	Top Conc. (ng/µL)	Middle Conc. (ng/µL)	Bottom Conc. (ng/µL)	Average Conc. (ng/µL)	CV%	Mass (µg)	Volume (µL)
rLiaOli1.BL1.UL1	3.33	1.98	157.0	54.1	164.71	41.12	750
rLiaOli1.BL1.UL2	14.9	1.39	119.0	45.1	142.71	9.02	750
rLiaOli1.BL1.UL3	102.0	56.8	19.4	59.4	69.63	11.88	200
rLiaOli1.BL1.UL5	45.1	238.0	42.5	108.53	103.31	21.71	200
rLiaOli1.BL1.UL6	43.4	26.8	26.0	32.07	30.63	6.41	200
rLiaOli1.BL1.UL7	127.0	79.8	53.0	86.6	43.3	5.63	100
rLiaOli1.BL1.UL8	28.80	18.60	13.50	20.30	38.38	4.06	200
rLiaOli1.BL1.UL9	10.40	27.70	51.50	29.87	69.09	5.97	200



A **pulsed-field gel** electrophoresis system is used to assess fragment length – here, we demonstrate the increased fragment length of the UHMW DNA compared to the HMW DNA. Thus, the longer **UHMW** DNA isolation protocol was used **upstream of ONT Ultra-Long (SQK-ULK001) Library Prep.**

ACKNOWLEDGEMENTS

Oxford NANOPORE Technologies · NEW ENGLAND BioLabs Inc. · wellcome sanger institute · Darwin TREE of LIFE · CBG Max Planck Institute of Molecular Cell Biology and Genetics · EARTH BIOGENOME PROJECT · Galaxy · B10K · NCBI · African BioGenome Project · GENOME 10K · BAT 1K

@niveshjain1 | @tilley_tatiana | @genomewarriors | @genomeark

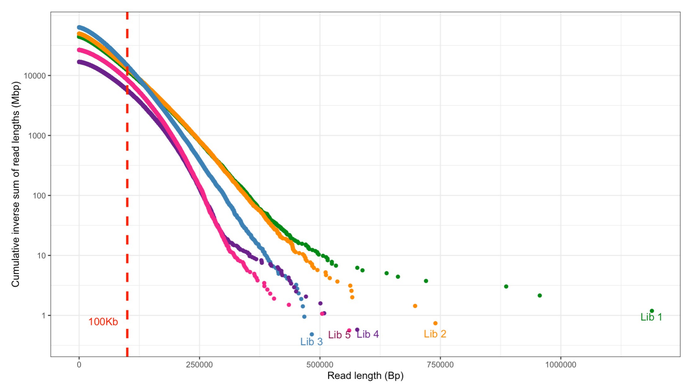
DOWNSTREAM PROCESSING AND ONT SEQUENCING

1. LIBRARY PREPARATION AND SEQUENCING STATS

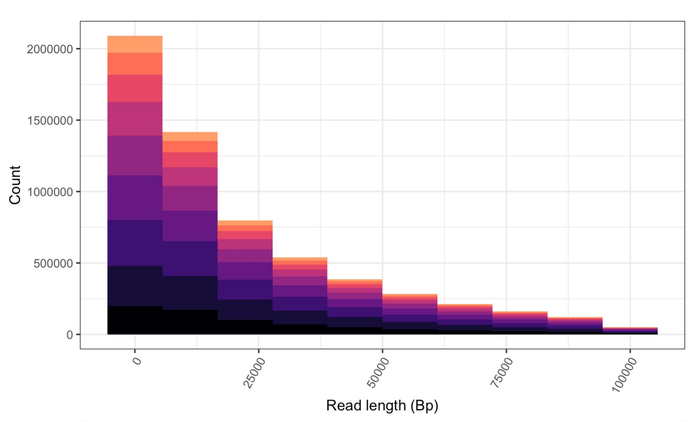
We prepared **ONT Ultra-Long libraries** using UHMW DNA extracted with the the NEB Monarch® HMW DNA Extraction Kit. We modified the protocol by **increasing the total adapter ligation time** to 20 minutes at room temperature (originally 5 minutes) and 10 minutes at 75°C (originally 5 minutes). In addition, we periodically mixed the samples with a wide-bore pipette tip throughout the adapter ligation step.

ONT Library	Estimated DNA Input	Estimated Bases	Data Produced	Reads Generated	Estimated N50
rLiaOli1 Library 1	40 µg	55.44 Gb	423.57 GB	2.38 M	60.76 kb
rLiaOli1 Library 2	20 µg	60.94Gb	465.12 GB	2.47 M	58.81 kb
rLiaOli1 Library 3	33 µg	76.65 Gb	584.95 GB	2.99 M	57.45 kb
rLiaOli1 Library 4	5 µg	21.4	163.56	663.12 k	72.77
rLiaOli1 Library 5 (a load of Lib 4 on another FC)	10 µg	32.57	247.3	1.07 M	70.75

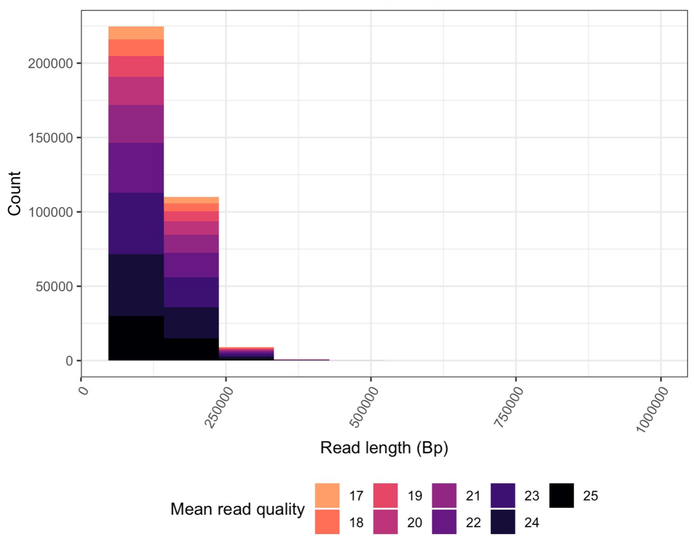
2. SEQUENCING RUN ASSESSMENT



Inverse cumulative sum of read lengths for the five datasets of ONT UL reads (Lib 1-5). For example, in Lib 1, the cumulative sum of read lengths greater than or equal to 100 kb is approximately 12.3 Gb.



Histogram of read lengths below 100 kb filtered for quality. Reads with an average quality (Phred score) between the first and ninth decile were retained and are represented in the graph.



Histogram of read lengths above 100 kb filtered for quality. There are approximately 7,850 reads exceeding 250Kb, but they are not visible due to the scale of the count data. In both figures, quality does not appear to be particularly associated with read length.

LIMITATIONS

Several factors prohibit performing ONT Ultra-Long sequencing on every VGP species. These samples are often collected in conditions where **optimal sample preservation is not possible**, such as in areas that are remote or lack sufficient infrastructure to allow for ideal sample collection and thus result in lower yield and/or shorter fragments during DNA isolation. Additionally, many VGP samples are precious and limited, meaning there is not enough DNA available to use for ONT UL sequencing as a **high DNA input required** - lower inputs will generate shorter strands due to the increased ratio of transposase to DNA. Aside from sample-related limiting factors, UHMW DNA is inherently **difficult and time-consuming to homogenize**, and requires longer elution and incubation times to preserve longer fragments. Additionally, performing **size selection** on ultra-long libraries is not as simple as with shorter libraries, although ONT has available a Short Fragment Eliminator kit that may improve N50 which has not been tested herein.

In conclusion, ONT Ultra-Long sequencing with VGP samples is only possible when a large amount of high-quality sample exists for a given species.