Exploring the limitations of current sequencing technologies by investigating codfishes

Ole K. Tørresen[1], Helle T. Baalsrud[2], Siv N.K Hoff[1], Marius F. Maurstad[1], Ave Tooming-Klunderud[1], Morten Skage[1], Giada Ferrari[1], Spyros Kollias[1], Mariann Árnyasi[2], Sissel Jentoft[1], Kim Præbel[3], Kjetill S. Jakobsen[1]

[1]Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway. [2]Department of Animal and Aquacultural Sciences, Centre for Integrative Genetics (CIGENE), Norwegian University of Life Sciences, Ås, Norway
[3]Norwegian College of Fishery Science, Faculty of Biosciences, Fisheries and Economics, UiT, The Arctic University of Norway, Tromsø, Norway

Abstract

High quality genome assemblies of a multitude of species are crucial to facilitate our increased understanding of biology, biodiversity and conservation, but also as a foundation to fuel future biotechnology to provide humanity with food, medical treatment, drugs, vaccines, biofuels and biomaterials. To be able to generate these genome assemblies we are dependent on biological samples with high tissue quality to enable successful sequencing with different technologies. Several species might have difficulties in one or more of these steps. Codfishes (Gadidae) contain large amounts of short tandem repeats (STRs) in their genomes. This type of repeats has led to challenges in both sequencing and assembling these fishes. In the EBP-Nor project, we aim to sequence and assemble several of these. To overcome these difficulties, we are utilising several sequencing technologies, Oxford Nanopore and PacBio HiFi in addition to Hi-C. High quality genome assemblies combining these data are commonly created using assemblers such as hifiasm and verkko, before validation by, e.g., BUSCO and Flagger. To investigate the limitations of the sequencing technologies, we will annotate different features of the genomes, such as STRs and transposable elements, and analyse if the presence of these affect the sequencing depth. By utilising multiple species we can better understand if the limiting factor is biological, and not sample specific. These investigations are important to ensure that we are aware of the limitations of the current technologies so we can sequence and assemble all eukaryotic species successfully.

To investigate the effect of short tandem repeats (STRs) on sequencing data and the assembly process, we selected three species (chicken, snowy owl and hake) which had Oxford Nanopore, PacBio HiFi and Hi-C available. First, we ran GenomeScope2 on the PacBio HiFi reads (Fig 1-3). Hake shows a very strange plot compared to the other species, with no clear distribution. It is not clear how and why this pattern occurs.

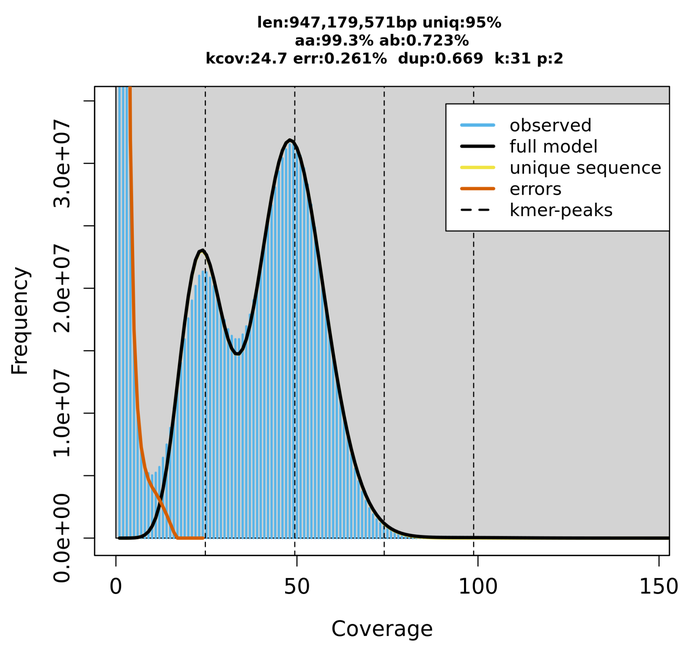

Figure 1. K-mer plot HiFi reads chicken

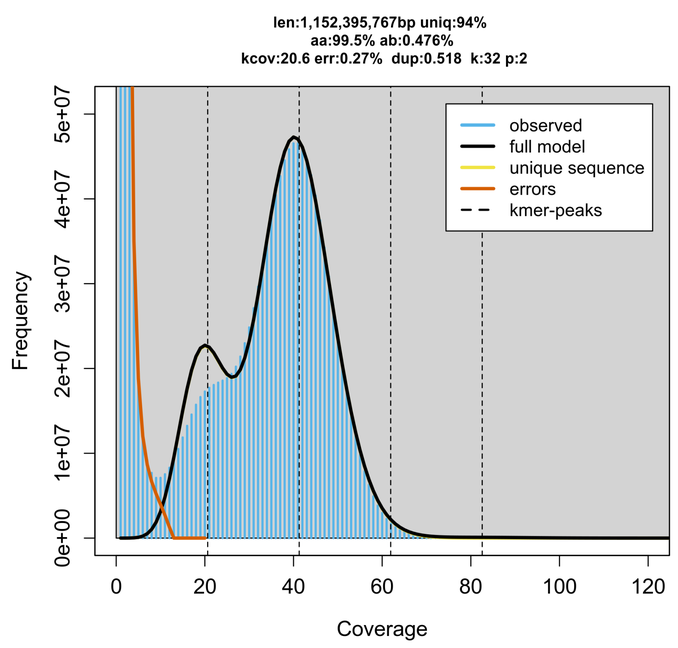

Figure 2. K-mer plot HiFi reads snowy owl

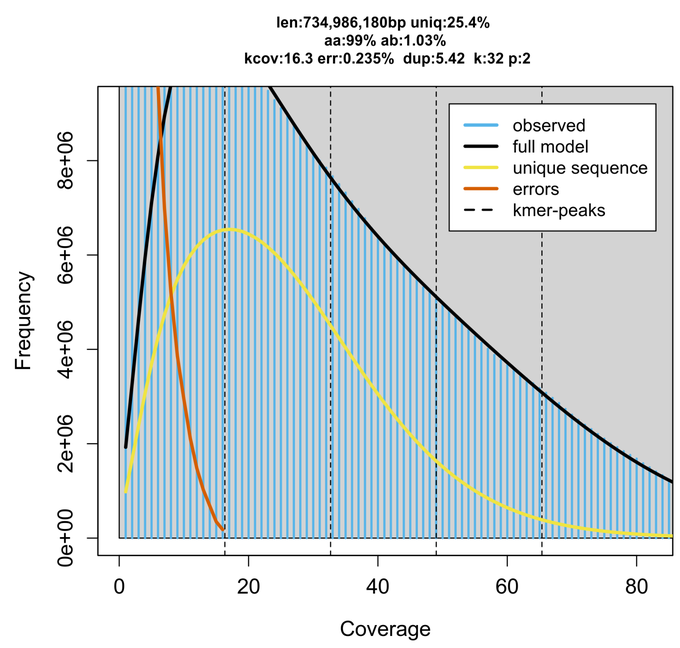

Figure 3. K-mer plot HiFi reads hake

We assembled these with hifiasm and scaffolded with YaHS. PacBio reads were mapped with the Flagger pipeline on the concatenated assemblies (ensuring that the reads map to the correct place). Mosdepth were run on the resulting BAM files to get mean coverage in 10k windows. ULTRA were run on the concatenated assemblies, and bedtools were used to get coverage of STRs (unit size 1-10 bp) in 10k windows across the assemblies. The statistics of the assemblies and ULTRA are shown in Table 1.

Species	PacBio coverage	ONT coverage	Assembly sizes (Mbp) (hap1/hap2)	N50 contig (Mbp) (hap1/hap2)	BUSCO complete (hap1/hap2)	Percent STRs
Chicken	51	104	1,009/1,057	38/36	90/92	2.5
Snowy owl	37	13	1,230/1,571	10/10	92/95	1.7
Hake	41	49	628/702	0.8/1	86/93	12.9

Table 1. Sequencing, assembly and short tandem repeat statistics of selected species.

To investigate the association between PacBio reads and STRs, we plotted the coverage vs STR content (in percentage) for all windows across the genome using R and ggplot2.

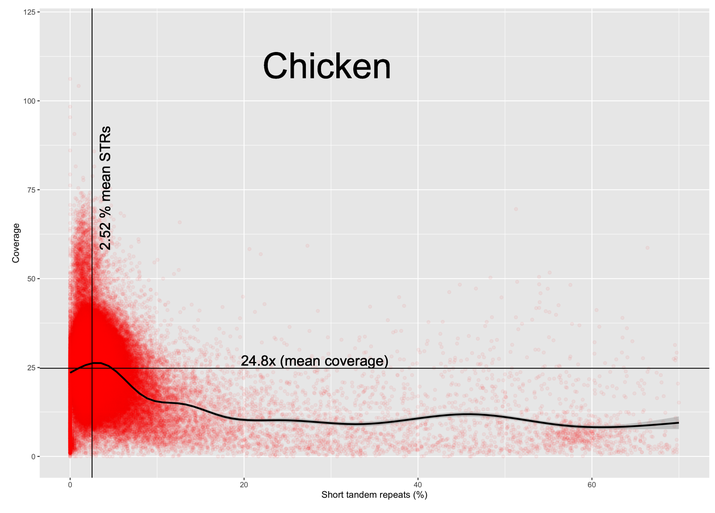

Figure 4. STRs vs coverage PacBio reads

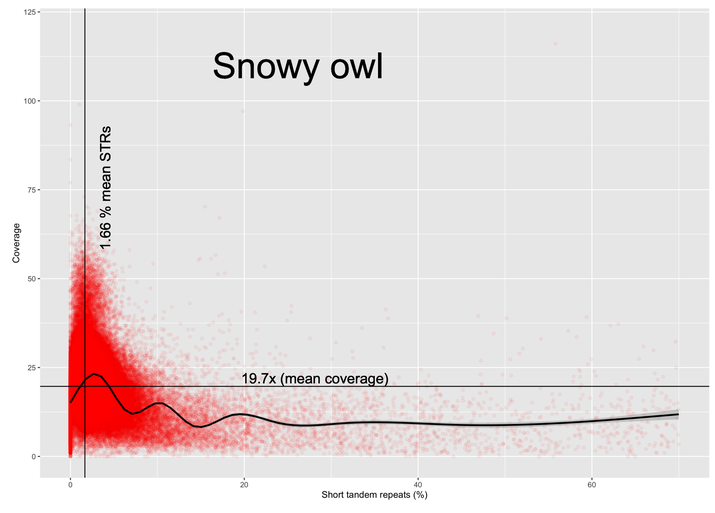

Figure 5. STRs vs coverage PacBio reads

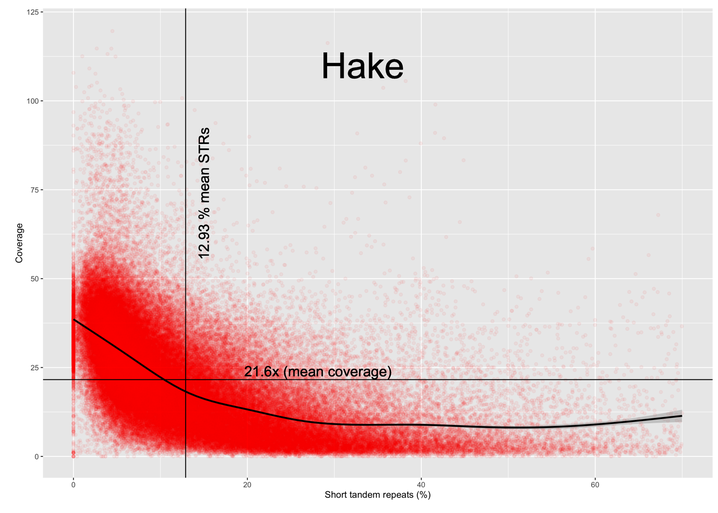

Figure 6. STRs vs coverage PacBio reads

In all species there is a trend toward lower coverage the higher amount of STRs in a window. However, due to its high STR content, hake has many more windows at higher STR content than the other species.

This is preliminary data, and we will extend these analyses to more species and genomic elements. For instance, we will look more careful into if there are specific STRs that are associated with lower coverage. Here we have not looked at length of the reads, but from earlier investigations we know there can be a strong association (reads are shorter where there are high density of STRs). Further, there might be types of transposable elements that could affect the sequencing of PacBio, so we will also annotate these and look at that association. Here we have not looked into what elements might affect ONT, but this is also an obvious analyses that will be done.

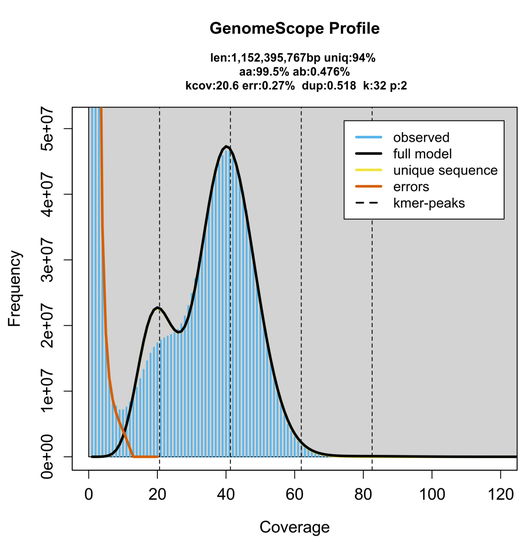

Figure 7. Get the poster here

Centre for Ecological and Evolutionary Synthesis

EARTH BIOGENOME NOR PROJECT

Figure 8. Picture of the first author

olekto@uio.no