Collaborators:

De novo assembly of the Atlantic herring (*Clupea harengus*) genome and comparison to a previously published assembly

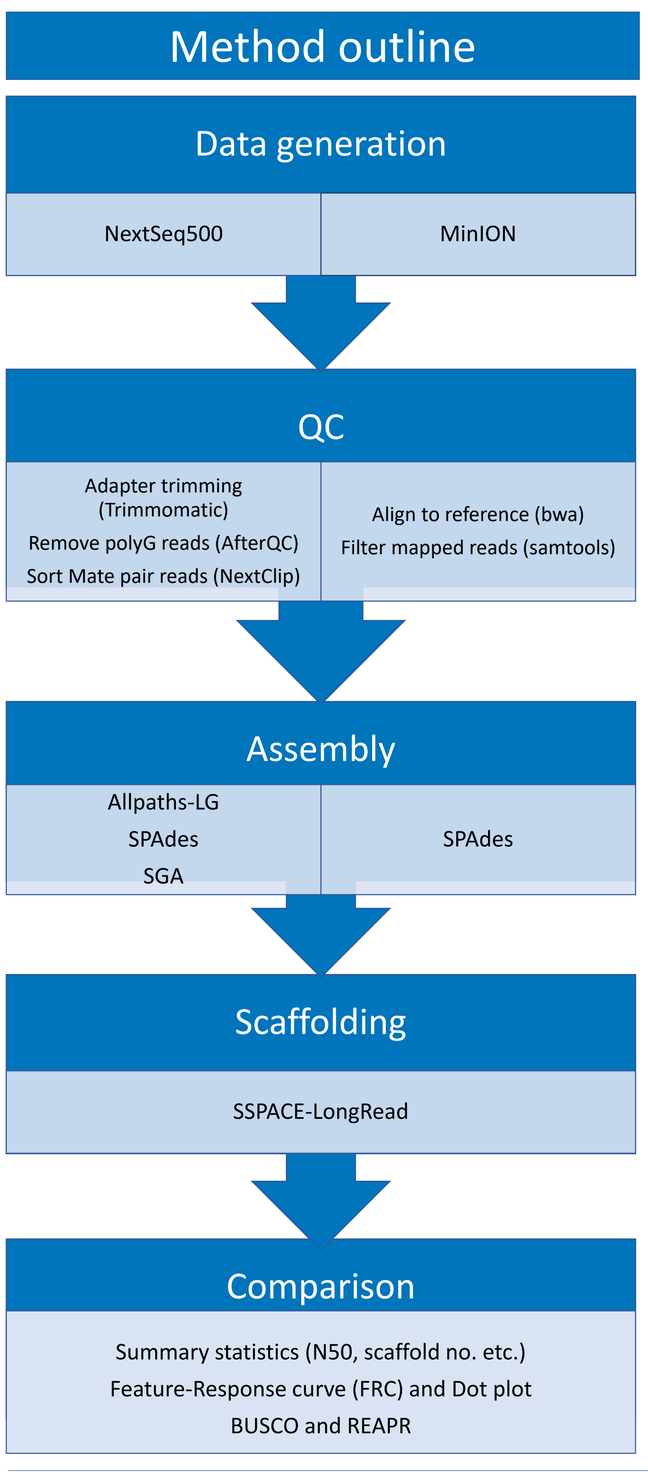
Sunnvør í Kongsstovu[1], Svein-Ole Mikalsen[2], Hannes Gislason[2], Hóraldur Joensen[2], Eydna í Homrum[3], Jan Arge Jacobsen[3], Debes H. Christiansen[4], Thomas Damm Als[5], Masa Roller[6], David Martín-Gálvez[6], Paul Flicek[6] and Hans Atli Dahl[7]

[1] Amplexa Genetics A/S, Hoyvíksvegur 51, FO-100 Tórshavn, Faroe Islands. [2] University of the Faroe Islands, Dept. of Natural Sciences, Nóatún 3, FO-100 Tórshavn. [3] Faroe Marine Research Institute, Nóatún 1, FO-100 Tórshavn, Faroe Islands. [4] Faroese Food and Veterinary Authority, National Reference Laboratory for Fish Diseases, Tórshavn, Faroe Islands. [5] Department of Biomedicine, Bartholins Allé 6, 8000 Aarhus C, Denmark. [6] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. [7] Amplexa Genetics A/S, Tolderlundsvej 3B, 2. DK-5000 Odense C, Denmark.

Introduction

The Atlantic herring (*Clupea harengus*) is one of the most abundant fish species in the world and is an important economical and nutritional resource. The Faroe Islands exported herring worth more than 400 million DKK in 2016, which amounts to 5% of the total value of exported goods that year[1].
The species is a highly migratory pelagic species with a complicated population structure. In order to keep the fisheries sustainable, knowledge of subpopulations, their migrations and mixing is important and would assist in the fight against illegal, unreported and unregulated fishing[2].
In 2016 Barrio *et al.* published the assembled herring genome from a herring caught in the Baltic sea[3]. Nevertheless as this is a species of such economical and nutritional importance we have undertaken a second assembly of the herring genome, using an Atlantic herring from the North sea. We also used different sequencing technologies and a different bioinformatical approach, in the hope of gaining further information about population structure and standing variation.
Here, we sequenced the herring genome with both Illumina and Oxford Nanopore Technologies platforms. The herring genome was assembled using the Illumina data and then scaffolded using the MinION long reads.
The basic assembly statistics were compared to the published assembly and the assembly completeness of both assemblies evaluated.

Method outline



Results

Table 1. Summary statistics for published assembly (Barrio *et al.*) and the assemblies generated during this study. V1 was assembled using the AllPaths-LG[4] software and the Illumina reads. V2 was generated by scaffolding V1 with MinION reads from two runs, using the SSPACE-LongRead[5] software. V3 was generated by closing gaps in V1 using GapFiller[6] software and scaffolded with four runs of MinION reads.

Assembly	No. Scaffolds	N50	L50	Assembly length	Gap length
Barrio *et al.*	6,915	1,860,920	113	807,711,962	82,755,438
This Study V1	15,378	175,084	993	702,694,152	177,906,258
This Study V2	12,219	219,865	820	716,584,296	191,669,007
This Study V3	10,354	262,663	719	729,006,009	186,023,247

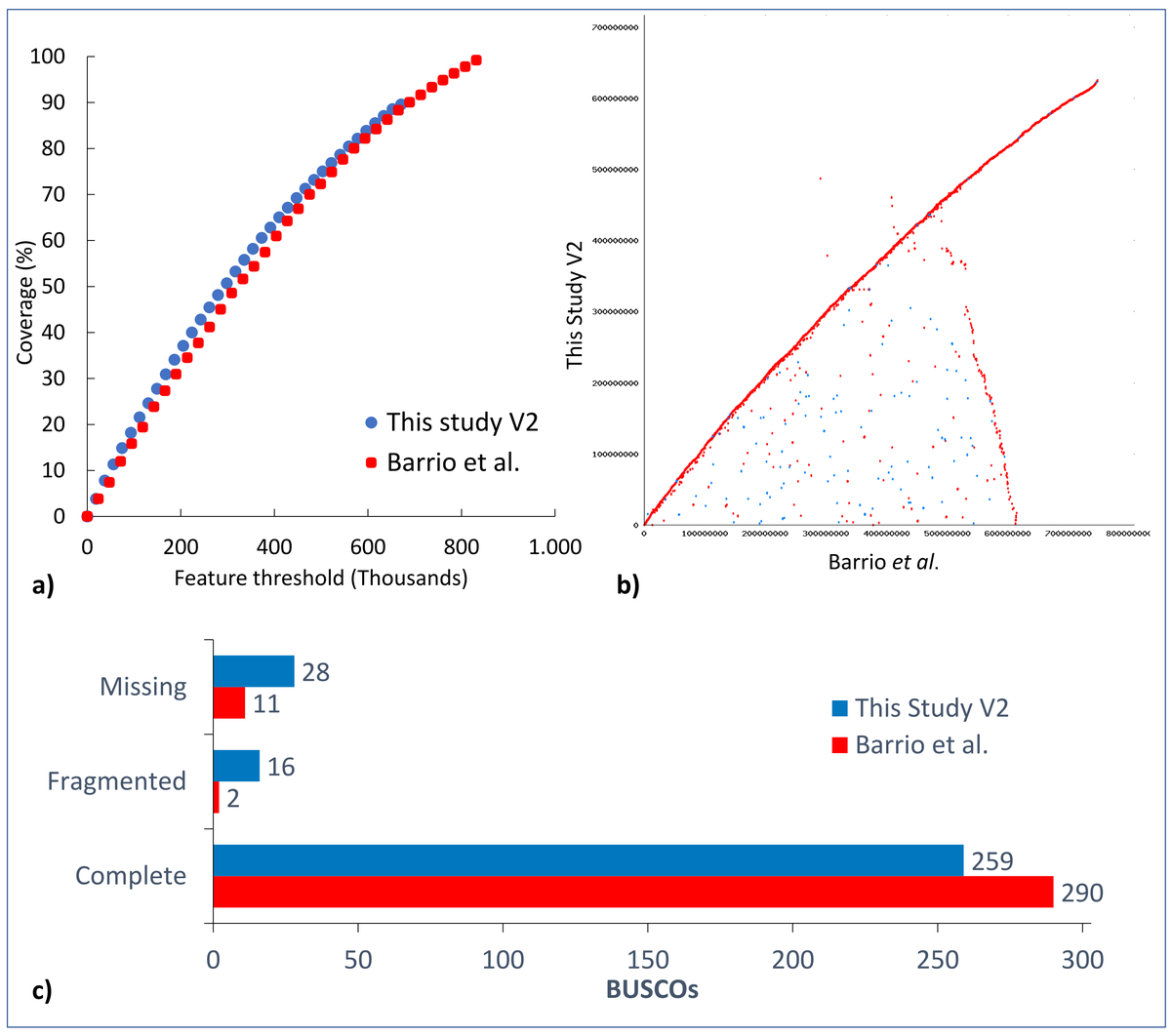


Figure 1. Comparison of Barrio *et al*. herring assembly and assembly V2 from this study **a)** Comparison of features in assembly with a Feature-Response curve (FRC). The FRC was generated using FRC_align[7] software. **b)** Structural variation between assemblies shown as a dotplot generated with MUMmer3[8]. **c)** Assembly completeness assessed using the BUSCOv3[9] software. Bar chart shows the number of complete, fragmented and missing Benchmarking Universal Single-Copy Orthologs (BUSCOs) in each assembly.

Table 2. Assembly errors and warnings called by REAPR[10]. Errors are calculated by coverage and fragment coverage distribution (FCD). Subgroups for warnings are reported by REAPR but are not included her for simplicity.

Errors and warnings	This Study V2	Barrio et al.
Error free bases	59.45%	11.42%
Errors	6,085	6,719
FCD errors within a contig	558	1,637
FCD errors over a gap	5,104	5,068
Low fragment coverage within a contig	89	6
Low fragment coverage over a gap	334	8
Warnings	509,349	2,160,429

Table 3. Summary statistics for broken assemblies generated by REAPR based on errors in table 2.

Assembly	No. Scaffolds	N50	L50	Assembly length
Barrio *et al.*	50,731	45,390	3,993	792,634,161
This study V2	23,172	105,152	1,784	706,219,224

Discussion

Comparing assemblies is complicated and a single comparative analysis often does not include enough parameters for a fair comparison. Therefore, we decided to compare the herring assembly from this study to a published assembly using the five different approaches below.

Summary statistics:
According to the summary statistics the assemblies generated in this study are more fragmented than the previously published herring assembly.

FRC:
The FRCs for the V2 and the previously published assembly are very similar. The V2 FRC is slightly steeper, indicating less assembly errors are present in the V2 assembly compared to the Barrio *et al.* assembly.

Dot plot:
There are no macroscopic structural variations between V2 assembly from this study and the previously published assembly. However, there is some unexplained noise in the graph.

BUSCO:
The V2 assembly has less completed BUSCOs and more fragmented and missing BUSCOs. This suggests that the Barrio *et al.* assembly is more complete than the V2 assembly. This makes sense as the Barrio *et al*. assembly has a longer total length and shorter gaps. The missing and fragmented BUSCOs in the V2 assembly could be in these gaps.

REAPR:
REAPR reported less errors and warnings and more error free bases for the V2 assembly, compared to Barrio *et al.* assembly. REAPR outputs broken assemblies that are broken based on the assembly errors called. The broken V2 assembly was less fragmented than the broken Barrio *et al.* assembly. This suggests that even though the original V2 assembly is more fragmented than the published assembly, when assembly errors are taken into account, it has better summary statistics.

Conclusion:
Our V2 assembly of the herring genome has considerably lower N50 value than that of Barrio *et al.*, but several assembly parameters suggest that our assembly is of higher quality. Thus, the results show that our new assembly is a good representation of the Atlantic herring genome. The V3 assembly has even better summary statistics than V2 but the other parameters also need to be evaluated.

References

[1] Hagstovan http://statbank.hagstova.fo/pxweb/fo/H2/H2__UH/
[2] Nielsen, E.E., *et al*. *Nature communications* 3 (2012): 851.
[3] Barrio, A.M., *et al*. *Elife* 5 (2016): e12081.
[4] Butler, J., *et al*. *Genome Research* 18 (2008):810–20.
[5] Boetzer, M. and Pirovano, W. *BMC Bioinformatics* (2014)15:211.
[6] Boetzer, M. and Pirovano, W. *Genome biology* 13.6(2012): R56.
[7] Vezzi, F. https://github.com/vezzi/FRC_align.
[8] Kurtz, S., *et al. Genome biology* 5.2 (2004): R12.
[9] Simão, FA., *et al. Bioinformatics* 31.19 (2015): 3210-3212.
[10] Hunt, M., *et al. Genome biology* 14.5 (2013): R47.

Financial support: