# Deliverable 8.1

## Genomic datasets collected in task 8.2 deposited in public databases

**Date: 17/05/2017**

**HORIZON 2020 - INFRADEV**

**Implementation and operation of cross-cutting services and solutions**

**for clusters of ESFRI**

| | |
|---|---|
| **Grant Agreement number:** | **654008** |
| **Project acronym:** | **EMBRIC** |
| **Contract start date:** | **01/06/2015** |
| **Project website address:** | **www.embric.eu** |
| **Due date of deliverable:** | **31/05/2017 / month 24** |
| **Dissemination level:** | **Public** |

**Document properties**

| | |
|---|---|
| Partner responsible | USTAN |
| Author(s)/editor(s) | Ian A. Johnston, Christopher Hollenbeck, Daniel Garcia de la serrana and Elena Sarropoulou |
| Version | 1 |

**Abstract**

Deliverable D8.1 comprises the collection of genomic datasets which resulted out of task 8.2 which have been deposited in public databases (M24). Genomic datasets were generated using next generation sequencing technologies (NGS) for four finfish Atlantic salmon, gilthead sea bream, bluefin tuna and greater amberjack and one shellfish, the King scallop. Raw sequencing data have been submitted to, and are available from, were INSDC public databases.

# 1 Table of Contents

# 1. Introduction

WP8 focused on genomic resources and selective breeding in mariculture, an economically important domain that has, as yet, not fully benefitted from recent cutting-edge advances in marine biological sciences. Specifically, this WP focuses on producing genomic datasets and transcriptomes for genetic marker discovery either to fill identified gaps in the resource base or to provide material for Task 8.3 (workflows and pipelines for pedigree reconstruction and selection tools). All genomic datasets were submitted to the public databases of the International Nucleotide Sequence Database Collaboration (INSDC), provided by DDBJ, EMBL-EBI and NCBI.

# 2. Shellfish

The available genomic resources for shellfish species have been reviewed in order to identify potential gaps **(USTAN** and **XELECT)**. The results of this analysis will be disseminated through a review article in a major aquaculture journal (Hollenbeck C. and Johnston, I.A. 2017). This paper is now under review and includes extensive supplementary tables on the genomic resources available for all the major aquaculture species.

Several of the tasks in WP8 involve the King scallop (*Pecten maximus*) which is particularly poorly served in terms of existing genetic resources. The gaps filled included an annotated transcriptome of adductor muscle, hepatopancreas and gonad. We also carried out SNP discovery from genomic DNA involving sequencing of RAD-seq libraries.

### 2.1 Description of King scallop transcription (USTAN)

Task responsible: Dr. Garcia de la Serrana, D. and Prof. Johnston, I.A.

Total RNA was extracted from gill, abductor muscle, hepatopancreas and gonads of 5 scallops, pooled and sequenced by Miseq v4. Over 19 million reads were generated and assembled in 125,378 contigs with a mean length of 1133bp with over 12,000 successfully annotated.

### 2.2 Description on King scallop RAD-seq (USTAN)

Task responsible: Dr. Christopher Hollenbeck, Dr. Daniel Garcia de la serrana and Prof. Ian A. Johnston

DNA was extracted from muscle tissue of 230 individuals and used to construct double-digest restriction-site associated DNA (ddRAD) libraries, following (1). Libraries were sequenced with 150bp paired-end reads on an Illumina HiSeq 4000 DNA sequencing machine. Sequence data will be used for population genetics analysis of wild King scallops and development of genotyping assays for pedigree analysis in aquaculture (Task 8.3).

# 3. Finfish

An analysis of genomic resources and genetic tools highlighted the need for genomic datasets to develop pedigree construction panels using both SNP and haplotype markers. We decided to produce the data for a diploid genome (Gilthead sea bream, *Sparus aurata*) and a tetraploid genome (*Salmo salar*) which will be used in Task 8.3.

### 3.1 RAD-seq libraries for Atlantic salmon (USTAN and Xelect)

Task responsible: Luke Holman and Prof. Ian A. Johnston

DNA was extracted from 20-40mg of tissue from 102 Atlantic salmon. DNA samples were subject to single digestion RAD-seq protocol by Floragenex, Inc (Portland, USA). Briefly, samples were digested using Sbf1 restriction enzyme, individually barcoded with custom Floragenex adapters followed by PCR amplification of the fragments. Pooled libraries were sequenced over two lanes of the Illumina HiSeq 2000 platform (Eurofins Genomics, Ebersberg, Germany). Sequence data was de-multiplexed and trimmed to a length of 90 base pairs using custom Floragenex scripts and mapped to the Atlantic salmon genome (ICSASG_vhttp://www.ebi.ac.uk/ena/data/view/GCA_000233375.3)
using BOWTIE v.0.12.8 allowing up to three mismatches. SAMTOOLS (Li et al 2009) and custom Floragenex scripts were used for SNP calling and variants were output as a Variant Call Format (VCF) file.

A total of 452.9 million reads with a mean of 4.4 million reads per individual. Following mapping an average 56.7% of the reads were unambiguously mapped to the reference genome. The genomic data set generated was used for Task 8.3 and has already resulted in a published workflow (Holman et al. 2017).

**Publication**

Holman, L.E., Garcia de la serrana, D., Onoufriou, A., Hillestad, B. and Johnston, I.A. (2017.A workflow used to design low density SNP panels for parentage assignment and traceability

in aquaculture species and its validation in Atlantic salmon. Aquaculture In press available online 4th April 2017.
http://www.sciencedirect.com/science/article/pii/S0044848616312194

**3.2 RAD-seq libraries for Gilthead seabream (USTAN)**

Task responsible: Dr Christopher Hollenbeck and Prof. Ian A. Johnston

DNA was extracted from muscle tissue of 20 individuals and used to construct double-digest restriction-site associated DNA (ddRAD) libraries (1). Libraries were sequenced with 300bp paired-end reads on an Illumina MiSeq DNA sequencing machine. Raw sequence reads were assembled into 27,054 RAD-tags using the dDocent pipeline (2). After filtering, high-quality 4,139 SNPs were recovered and utilized for downstream design of SNP assays. This data set will be used in Task 8.3.

**3.3 Resequencing Atlantic salmon genome and Copy Number Variant detection (XELECT)**

Task responsible: Ms Alicia Bertolotti (Xelect and University of Aberdeen), Prof. Ian A. Johnston (USTAN and Xelect) and Dr Dan Macqueen (University of Aberdeen).

Task 8.3 required the preparation of genome sequences from Atlantic salmon.

Whole genome sequencing was carried out on DNA from nine Atlantic salmon individuals of farmed, Scottish origin using an Illumina NextSeq500. Each fish produced on average 221,158,064 paired-end raw reads (2x150bp) at about 10-15x coverage. Read quality was checked and trimmed using FastX (v0.0.13)(3).

Reads were then aligned to the 29 anchored chromosomes from the Atlantic salmon ICSASG_V2 reference genome using the Burrow-Wheelers Aligner (BWA) (v0.7.13)(4).

Computational prediction of CNVs was done using CNVnator (v0.3)(5). This programme uses the Read-Depth method, which relies on the depth of coverage of a genomic region being directly correlated to the copy number of that region. On average, 1,198 CNVs were identified per individual (n=9, median=1,116, SD=252.4); with 667 duplications and 531 deletions.

The overall length of CNVs per individual covered a total of 2.23% of the 29 Atlantic salmon chromosomes. This is likely to be an underestimate as the Atlantic salmon genome is highly

repetitive (60%). Additional CNV detecting programmes are currently being explored in order to combine results and extend our knowledge of the CNV landscape. Additionally, distribution of CNVs across chromosomes hint that CNV formation could be influenced by the rediploidization process following the salmonid-specific Whole Genome Duplication 95MYA (6). Higher CNV content was observed in regions where rediploidization was delayed (7) and duplicated genes have undergone the least divergence.

Of the currently known CNVs, 50% overlap with, or are within 1bp, of annotated protein-coding genes. 73% of these are duplications. This consistent with past reports that deletions are typically more deleterious and rarer around genes (8).

### 3.4 Sequencing of Bluefin tuna genome: generation of reference genome and Copy Number Variant detection (HCMR).

Task responsible: Jon Bent Kristoffersen, Viet Quoc Ha, Elena Sarropoulou

Whole genome sequencing was carried out on DNA from eight Bluefin tuna individuals. Fish was provided from **TunaTech,** Germany. Four individuals originated from Malta and four from Croatia. For high quality DNA extraction several protocols were tested. Some of the samples seemed to have partially degraded DNA, and multiple extractions attempts were needed to obtain DNA of sufficient quantity and quality.

Library construction for the four individuals from Malta and the four individuals from Croatia was performed with Illumina's DNA PCR-free LT (Set A) kit. An insert size of 350 bp was chosen, which is suitable for HiSeq X sequencing. Each library has a unique 6 base index sequence, and the indexes were chosen so as to have high base diversity in the index read. Mechanical DNA shearing was done with a Covaris S220. Sample 2c-29 had partially sheared DNA in every DNA extraction. To compensate, the shearing time for this sample was reduced by 20 %. Library construction followed the

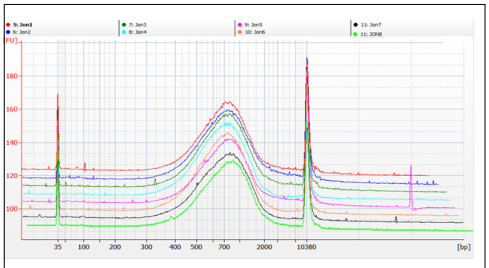

**Figure 1**. Bioanalyzer traces for the libraries. Since these are PCR-free libraries, the libraries are partially single-stranded and migrate slower than fully double-stranded DNA. Hence the peak size is at around 750 bp, even though the true size is around 470 bp.

manufacturer's instructions. The size distributions of the finished libraries were checked with a high-sensitivity DNA chip in Bioanalyzer (**Fig 1**), and were as expected. The library concentrations were checked by Qubit (for total amount of DNA) and qPCR (for amount with Illumina adapters annealed at both ends), and were found to be excellent (**Table 1**), all well above the minimum acceptable of 4 nM, and also above the recommended minimum of 10 nM.

**Table 1.** Overview of the samples and libraries. Under sample ID, the "a", "b", "c" etc. denotes successive DNA extraction events. The colors of the index bases indicate the different fluorescent colors seen by the sequencer. Every column should have close to an equal balance of the two colors for best results.

| | | | Qubit | qPCR | | |
|---|---|---|---|---|---|---|
| **Sample ID** | **Area** | **sex** | **ng/ul** | **nM** | **Adapter** | **Index sequence** |
| 8a | Malta | m | 8.64 | 17.9 | AD002 | C G A T G T |
| 2c | Malta | f | 11.2 | 18.9 | AD004 | T G A C C A |
| 7c | Malta | f | 9.6 | 17.5 | AD005 | A C A G T G |
| 9d | Malta | m | 13.9 | 24.3 | AD006 | G C C A A T |
| CR-c | Croatia | - | 6.54 | 11.8 | AD007 | C A G A T C |
| BFT2-4thMay | Croatia | - | 9.86 | 17.7 | AD016 | C C G T C C |
| BFT1-29thMay | Croatia | - | 9.92 | 19.0 | AD018 | G T C C G C |
| 2c-29 | Croatia | - | 11.1 | 22.2 | AD019 | G T G A A A |

Each fish produced on average 575,186,435 paired-end raw reads (2x150bp). Read quality was checked and trimmed using Trimmomatic v0.33 software (Bolger *et al.* 2014).

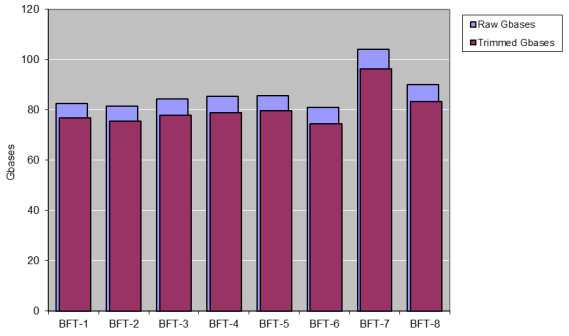


**Figure 2.** Graphical presentation of raw GB and trimmed GB obtained out of 8 bluefin tuna individuals.

**3.5 Transcriptome and genome study in the Greater Amberjack Seriola dumerilii (HCMR).**

Task responsible: Elisavet Kaitetzidou, Elena Sarropoulou

This study was performed in collaboration with the National funding project KRHPIS 2011 – 2015. The present work was performed in collaboration with the national project KRHPIS and valuable molecular resources have been generated for the Greater Amberjack (S.dumerilii) in form of a first draft genome comprising 45, 909 scaffolds as well as gonad and muscle reference transcriptome. Comparative mapping to *Seriola quinqueriadiata* (Aoki et al., 2014, 2015) as well as to the model fish species Medaka (*O.latipes*) allowed the generation of *in silico* groups comprising 44 % of the obtained scaffolds and 53 % of the obtained transcripts. Furthermore first insights were obtained for sex determination and differentiation as well as first hints towards the physiological cause of slow growers are gained. **Tables 2 and 3** provide sequence read information of raw, trimmed and aligned reads to the generated reference genome. The publication of the work is under preparation.

**Table 2.** RNA sequencing data of four female and four male individuals submitted to Illumina HiSeq sequencing.

| | Raw reads | Trimmed reads | After PhiX removal - Used for DE | % of raw reads used for DE | Tophat2 align | |
|---|---|---|---|---|---|---|
| F1 | 17,268,356 | 14,425,361 | 14,391,074 | 83.34% | 10,210,127 | 70.9% |
| F2 | 17,083,394 | 13,542,326 | 13,516,020 | 79.12% | 9,601,965 | 71.0% |
| F3 | 23,331,846 | 19,330,707 | 19,265,030 | 82.57% | 13,835,238 | 71.8% |
| F4 | 20,580,574 | 16,881,832 | 16,834,901 | 81.80% | 12,066,021 | 71.7% |
| M1 | 13,860,606 | 11,838,653 | 11,805,696 | 85.17% | 8,603,874 | 72.9% |
| M2 | 15,315,961 | 12,898,261 | 12,868,488 | 84.02% | 9,205,543 | 71.5% |
| M3 | 15,086,732 | 12,036,225 | 11,997,908 | 79.53% | 8,623,374 | 71.9% |
| M4 | 13,297,840 | 10,635,465 | 10,600,787 | 79.72% | 7,548,436 | 71.2% |

**Table 3.** RNA sequencing data of fast/normal and slow growers submitted to Illumina MiSeq sequencing.

| | Raw reads | Trimmed reads | After PhiX removal - Used for DE | % of raw reads used for DE | Tophat2 align | |
|---|---|---|---|---|---|---|
| Fast1 | 2434221 | 2164923 | 2148193 | 88.25% | 914312 | 42.56% |
| Fast2 | 2335103 | 2004168 | 1990068 | 85.22% | 921965 | 46.33% |
| Fast3 | 2944122 | 2649012 | 2627657 | 89.25% | 1265753 | 48.17% |
| Fas4 | 2912235 | 2616056 | 2600521 | 89.30% | 1304009 | 50.14% |
| Slow1 | 2088813 | 1843495 | 1830095 | 87.61% | 868946 | 47.48% |
| Slow2 | 2263688 | 1941004 | 1925396 | 85.06% | 985902 | 51.21% |
| Slow3 | 2497198 | 2197131 | 2178922 | 87.25% | 958565 | 43.99% |
| Slow4 | 2155458 | 1892706 | 1877917 | 87.12% | 869033 | 46.28% |

# 4. Genomic datasets collected in task 8.2 deposited in public databases

**4.1 King scallop transcriptome.**
Accession number PRJEB17629 in the EBI Short Read Achieve (SRA).

http://www.ebi.ac.uk/ena/data/view/PRJEB17629

**4.2 King Scallop RAD-seq.**
EBI Accession No. PRJEB17629

http://www.ebi.ac.uk/ena/data/view/PRJEB17629

**4.3 Atlantic salmon RAD-seq.**
EBI Short Read Archive (SRA) Accession number PRJEB17687.

http://www.ebi.ac.uk/ena/data/view/PRJEB17687

**4.4 Gilthead seabream RAD-seq.**
Raw sequence data has been submitted to the European Nucleotide Archive under accession number PRJEB19745.

http://www.ebi.ac.uk/ena/data/view/PRJEB19745

**4.5 Genome sequences from 9 farmed Atlantic salmon**
EBI Accession numbers. Bioproject: PRJNA378201

http://www.ebi.ac.uk/ena/data/view/PRJNA378201

Biosamples:

1. SAMN06480819. sample ID XA01;
   http://www.ebi.ac.uk/ena/data/view/SAMN06480819
2. SAMN06480820. sample ID XA61;
   http://www.ebi.ac.uk/ena/data/view/SAMN06480820
3. SAMN06480821. sample ID XA79;
   http://www.ebi.ac.uk/ena/data/view/SAMN06480821
4. SAMN06480822. sample ID XE32;
   http://www.ebi.ac.uk/ena/data/view/SAMN06480822
5. SAMN06480823. sample ID XE06;
   http://www.ebi.ac.uk/ena/data/view/SAMN06480823

6. SAMN06480824. sample ID XE09; http://www.ebi.ac.uk/ena/data/view/SAMN06480824
7. SAMN06480825. sample ID XH03; http://www.ebi.ac.uk/ena/data/view/SAMN06480825
8. SAMN06480826. sample ID XH82; http://www.ebi.ac.uk/ena/data/view/SAMN06480826
9. SAMN06480827. sample ID XH96; http://www.ebi.ac.uk/ena/data/view/SAMN06480827

**4.6 Genome sequences from 8 Bluefin tuna from Malta and Croatia**

NCBI SRP (study) accession: SRP105339

http://www.ebi.ac.uk/ena/data/view/SRP105339

Bioproject: PRJNA384315

http://www.ebi.ac.uk/ena/data/view/PRJNA384315

Biosamples:

1. SAMN06834835 BFT-1; http://www.ebi.ac.uk/ena/data/view/SAMN06834835
2. SAMN06834836 BFT-2; http://www.ebi.ac.uk/ena/data/view/SAMN06834836
3. SAMN06834837 BFT-3; http://www.ebi.ac.uk/ena/data/view/SAMN06834837
4. SAMN06834838 BFT-4; http://www.ebi.ac.uk/ena/data/view/SAMN06834838
5. SAMN06834839 BFT-5; http://www.ebi.ac.uk/ena/data/view/SAMN06834839
6. SAMN06834840 BFT-6; http://www.ebi.ac.uk/ena/data/view/SAMN06834840
7. SAMN06834841 BFT-7; http://www.ebi.ac.uk/ena/data/view/SAMN06834841
8. SAMN06834842 BFT-8; http://www.ebi.ac.uk/ena/data/view/SAMN06834842

**4.7 Genome sequences from 2 Greater Amberjack and 16 transcriptome sequences**

NCBI SRP (study) accession: SRP105319

http://www.ebi.ac.uk/ena/data/view/SRP105319

Bioproject: PRJNA384295

http://www.ebi.ac.uk/ena/data/view/PRJNA384295

Biosamples:

1. SAMN06841032; http://www.ebi.ac.uk/ena/data/view/SAMN06841032
2. SAMN06841033; http://www.ebi.ac.uk/ena/data/view/SAMN06841033
3. SAMN06841034; http://www.ebi.ac.uk/ena/data/view/SAMN06841034
4. SAMN06841035; http://www.ebi.ac.uk/ena/data/view/SAMN06841035
5. SAMN06841036; http://www.ebi.ac.uk/ena/data/view/SAMN06841036

6. SAMN06841037; http://www.ebi.ac.uk/ena/data/view/SAMN06841037
7. SAMN06841038; http://www.ebi.ac.uk/ena/data/view/SAMN06841038
8. SAMN06841039; http://www.ebi.ac.uk/ena/data/view/SAMN06841039
9. SAMN06841040; http://www.ebi.ac.uk/ena/data/view/SAMN06841040
10. SAMN06841041; http://www.ebi.ac.uk/ena/data/view/SAMN06841041
11. SAMN06841042; http://www.ebi.ac.uk/ena/data/view/SAMN06841042
12. SAMN06841043; http://www.ebi.ac.uk/ena/data/view/SAMN06841043
13. SAMN06841044; http://www.ebi.ac.uk/ena/data/view/SAMN06841044
14. SAMN06841045; http://www.ebi.ac.uk/ena/data/view/SAMN06841045
15. SAMN06841046; http://www.ebi.ac.uk/ena/data/view/SAMN06841046
16. SAMN06841047; http://www.ebi.ac.uk/ena/data/view/SAMN06841047
17. SAMN06841048; http://www.ebi.ac.uk/ena/data/view/SAMN06841048
18. SAMN06841049; http://www.ebi.ac.uk/ena/data/view/SAMN06841049

# 5. References:


1. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res [Internet]. 2011 Jun [cited 2015 Mar 23];21(6):974–84.
2. Aoki, J., W. Kai, Y. Kawabata, A. Ozaki, K. Yoshida et al., 2014 Construction of a radiation hybrid panel and the first yellowtail (Seriola quinqueradiata) radiation hybrid map using a nanofluidic dynamic array. BMC Genomics 15: 165.
3. Aoki, J., W. Kai, Y. Kawabata, A. Ozaki, K. Yoshida et al., 2015 Second generation physical and linkage maps of yellowtail (Seriola quinqueradiata) and comparison of synteny with four model fish. BMC Genomics 16: 406.
4. Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, et al. Copy number variation of individual cattle genomes using next-generation sequencing. 2012;778–90.
5. Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.
6. Hannon GJ. FASTX-Toolkit [Internet]. 2010.
7. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
8. Lien S, Koop BF, Sandve SR, Miller JR, Matthew P, Leong JS, et al. The Atlantic salmon genome provides insights into rediploidization. 2016;(6020).
9. Macqueen DJ, Johnston I a. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc Biol Sci [Internet]. 2014;281(1778):20132881.
10. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one*, **7**, e37135.
11. Puritz JB, Hollenbeck CM, Gold JR (2014) dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, **2**, e431.